Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires

le 28 juin 2024

Par : Ba Minh Nghi NGUYEN

Titre : Mesure du phénomène de l'anti-sélection en assurance santé à travers la théorie de l'utilité aléatoire : Applications à partir de la base Open Damir

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : FORSIDES France
Signature :

Directeur de Mémoire en entreprise :
Nom : Charlène FUSIS
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-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)

Secrétariat :

Signature du responsable entreprise

Bibliothèque :

Signature du candidat

Résumé

L'anti-sélection est un problème majeur ayant un impact significatif sur le domaine de l'assurance, notamment sur le marché de l'assurance santé individuelle. Les conséquences qu'il engendre sont sévères pour les assureurs, allant jusqu'au point où le marché devient déséquilibré et les assureurs peuvent être contraints d'arrêter les produits d'assurance exposés à l'anti-sélection en raison des pertes continuelles. Cependant, il est très difficile de traiter l'anti-sélection, d'une part, car sa présence dans le portefeuille ne peut pas être directement mesurée, et d'autre part, en raison de l'asymétrie d'information inévitable.

En assurance santé, notamment en France, des produits complémentaires santé à adhésion facultative sont proposés à différents niveaux de couverture. La faculté d'adhésion et la forte concurrence sur le marché font des produits d'assurance complémentaire santé individuelle des cibles faciles pour l'anti-sélection.

Ce mémoire vise à étudier ce phénomène pour les produits d'assurance complémentaire santé individuelle en utilisant la théorie de l'utilité aléatoire et tente de fournir un guide d'utilisation des modèles de choix discrets sur un portefeuille d'assurance complémentaire santé. En utilisant les définitions des coefficients d'anti-sélection par niveau de couverture contractuelle, nous démontrons que l'anti-sélection peut être atténuée grâce à des stratégies de tarification bien élaborées.

Pour atteindre cet objectif, nous avons construit un jeu de données simulé reflétant un portefeuille d'assurance réel. Notons que l'anti-sélection est étroitement liée aux préférences des assurés en matière de couverture complémentaire santé. Nous avons donc déployé des modèles de choix discrets pour modéliser les comportements des assurés. Enfin, nous avons appliqué ces modèles pour prédire les variations de l'anti-sélection sur différents contrats d'assurance et analyser les tendances de la souscription.

Mots-clés : *Anti-sélection*, *Assurance santé individuelle*, *Tarification* ,*Open Damir*, *Théorie d'utilité aléatoire.*

Abstract

Adverse selection is a major problem with a significant impact on the insurance industry, particularly in the individual health insurance market. Its consequences are severe for insurers, reaching the point where the market becomes unbalanced and insurers may be forced to discontinue insurance products exposed to adverse selection due to continuous losses. However, addressing adverse selection is very challenging, firstly because its presence in the portfolio cannot be directly measured, and secondly due to the inevitable information asymmetry.

In health insurance, especially in France, voluntary supplementary health insurance products are offered at different coverage levels. The voluntary nature of enrollment and the high competition in the market make individual health insurance supplementary products easy targets for adverse selection.

This thesis aims to study this phenomenon for individual health insurance supplementary products using random utility theory and attempts to provide a guide for using discrete choice models on a health insurance supplementary portfolio. By using definitions of adverse selection coefficients by contractual coverage level, we demonstrate that adverse selection can be mitigated through well-designed pricing strategies.

To achieve this goal, we constructed a simulated dataset reflecting a real insurance portfolio. It is worth noting that adverse selection is closely related to insured preferences for health insurance coverage. Therefore, we deployed discrete choice models to model insured behaviors. Finally, we applied these models to predict variations in adverse selection across different insurance contracts and analyze subscription trends.

Keywords : *Adverse selection*, *Individual health insurance*, *Pricing*, *Open Damir*, *Random utility theory*.

Note de Synthèse

Anti-sélection en assurance santé et sa problématique

Depuis longtemps, l'anti-sélection a été considérée comme un problème moral nuisant au marché de l'assurance. Dans le domaine de l'assurance santé, notamment sur le marché de la complémentaire santé individuelle, ce problème est encore plus accentué en raison de la nature des préférences individuelles en matière de risques, qui sont souvent non observables, et du fait que les assureurs ont du mal à contrôler l'asymétrie d'information en faveur des assurés. Même en proposant différents niveaux de remboursement comme incitation à l'auto-sélection, ces efforts ne suffisent pas à résoudre le problème de l'anti-sélection, ce qui entraîne une augmentation des coûts pour les contrats offrant une couverture élevée et incite les assurés à opter pour des couvertures moins étendues, créant ainsi une spirale négative pour ces contrats. Dans ce contexte, il est crucial pour les assureurs de comprendre les motivations des souscripteurs de complémentaire santé et de pouvoir mesurer la distorsion de la sinistralité réelle dans leur portefeuille par rapport à leur sinistralité ciblée, afin de pouvoir ajuster leur stratégie de tarification.

Ce mémoire vise à proposer un guide méthodologique économétrique pour aborder le problème de l'anti-sélection dans un portefeuille santé individuel. L'objectif principal est de modéliser et de comprendre les préférences des assurés, ainsi que de quantifier la différence de sinistralité causée par la distorsion dans la répartition des assurés sur trois niveaux de couverture du portefeuille. À partir de cette base, les modèles sont utilisés pour évaluer l'impact d'un changement de tarif sur la souscription des assurés du portefeuille, ce qui conduit à une analyse des changements dans la distribution des niveaux de couverture.

Méthodologie générale de l'analyse de l'anti-sélection

Le mémoire repose sur la méthode présentée dans la figure 1, illustrant une réponse à la problématique du mémoire. Le cadre du mémoire s'inscrit dans le contexte d'un assureur fictif proposant différents contrats de complémentaire santé individuelle au sein de portefeuilles fictifs.

Ce mémoire utilise la base de données publique Open Damir de 2021 pour construire les primes d'assurance et les sinistres agrégés par segment d'assuré sur trois niveaux de couverture, définis par la tranche d'âge, la région et le sexe. De plus, des statistiques nationales sont utilisées pour créer deux bases d'assuré avec un échantillon de 50 000 lignes, en ajoutant la variable de niveau de vie - le revenu. Les indices d'anti-sélection mesurent la distorsion dans le niveau de sinistralité réelle des contrats par rapport à leur niveau moyen sur l'ensemble du portefeuille. Ces coefficients d'anti-sélection sont pertinents, car ils permettent de revenir aux probabilités de choix, liées à l'utilisation du modèle de choix discrets. Les modèles sont ensuite entraînés et testés pour valider chaque spécification. Les probabilités prédites par le modèle sur la base désagrégée sont agrégées pour formuler les coefficients d'anti-sélection. Une analyse spécifique de l'état de l'anti-sélection est ainsi menée, avec l'impact d'un changement de tarification.

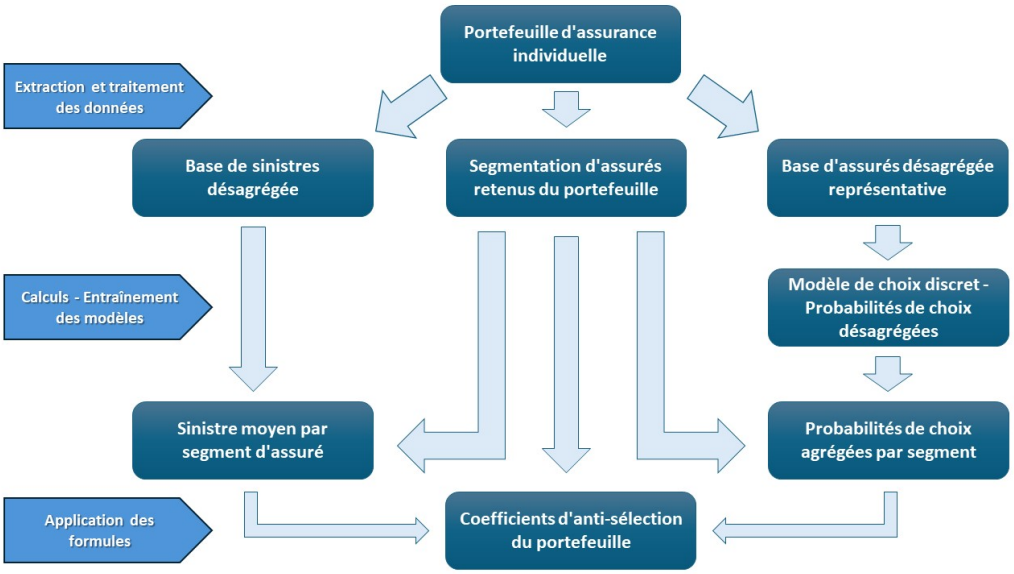


FIGURE 1 : Stratégie de modélisation de l'anti-sélection grâce au modèle de choix discrets

La base Open Damir et la construction de segments d'assurés

La base de données Open Damir répertorie les remboursements des soins effectués par tous les régimes d'assurance maladie, à l'exception d'une grande partie des frais hospitaliers dans le secteur public. Elle contient des détails sur les actes médicaux, les bénéficiaires des soins et les professionnels de la santé, tout en préservant leur anonymat grâce à l'agrégation des informations. Les données de l'année 2021 sont utilisées dans le cadre de ce mémoire. Il existe des remboursements des régimes spéciaux comme celui de la complémentaire santé solidaire CSS (CMUC) ou du régime d'Alsace-Lorraine. Le mémoire sépare ces types de remboursements ainsi que le traitement des lignes de remboursement incohérentes dans le but de créer une base de données prête à appliquer les formules de remboursement de complémentaire santé. Au cours du traitement de la base, trois variables sont choisies pour leur pertinence dans la compréhension des différences de consommation de soins entre les profils : la tranche d'âge (la modalité d'assuré de moins de 20 ans est éliminée), la région et le sexe.

Construction de la base de sinistre moyen et des tarifs de complémentaire santé

Le mémoire présente une fiche de garanties benchmark de Forsides avec trois niveaux de couverture : minimum, moyen et maximum. Elle est composée de plusieurs postes de garantie souvent retrouvés dans les grilles de garantie des complémentaires santé sur le marché :

- Soin courant : Consultations généralistes ; Consultations spécialistes ; Actes d'imagerie médicale ; Actes techniques médicaux ; Analyses et examens de laboratoire ; Honoraires paramédicaux ; Pharmacie remboursée par RO (Régime Obligatoire) ; Matériel médical.
- Optique : Monture ; Verres simples & monture ; Verres complexes ou très complexes & monture ; Lentilles acceptée par RO ; Lentilles refusée par RO (Régime Obligatoire) ; Chirurgie rétractive.
- Dentaire : Soins ; Inlay-Onlay ; Parodontologie ; Prothèses ; Orthodontie ; Implantologie.
- Aides Auditives : Appareil auditif remboursé par RO (Régime Obligatoire).
- Hospitalisation : Forfait journalier ; Chambre particulière en Hospitalisation/Psychiatrie ; Frais de séjour ; Honoraires ; Frais d'accompagnement ; Transport.

- Autres : Cure thermale ; Médecine douce ; Soins à l'étranger.

Les garanties sont appliquées aux lignes de la base Open Damir de l'année 2021 après le traitement de la base, créant ainsi trois nouvelles variables : le remboursement du contrat de niveau minimum, de niveau moyen et de niveau maximum. L'agrégation de ces variables de segmentation construit donc la base de sinistre agrégé par segment.

La tarification d'un contrat complémentaire santé consiste à additionner tous les tarifs des postes de garantie. La formule de tarification d'un poste de garantie par la méthode d'expérience se passe par :

$$\text{Sinistre moyen} = \text{Prime pure} = \frac{\text{Montant total de sinistres passées}}{\text{Exposition totale au risque}}$$

Ce mémoire utilise en particulier cette méthode en raison du caractère agrégé des remboursements de la base Damir. Les sinistres agrégés par segment correspondent au montant de sinistre passé, et l'exposition totale au risque est assimilée au nombre de personnes sur chaque segment dans la base de population de l'INSEE entre 2021 et 2022. Les primes pures annuelles fictives retenues pour les 3 niveaux de couverture sont finalement segmentés en 182 segments, compte tenu de la différence considérable de la sinistralité moyenne entre les deux classes homme-femme (voir l'exemple de la figure 2).

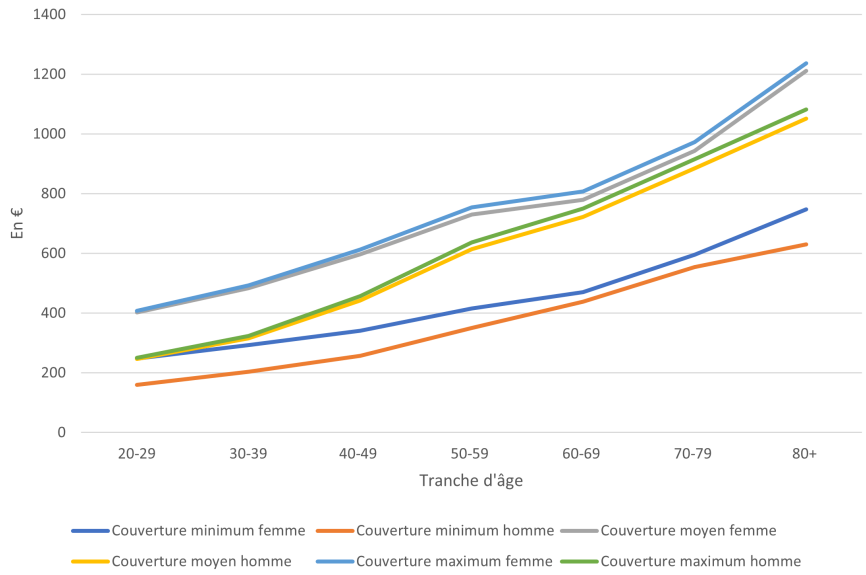


FIGURE 2 : Primes pures par tranche d'âge en région Provence-Alpes-Côte d'Azur et Corse, illustrant la différence entre la sinistralité de l'homme et de la femme

Construction de la base d'assurés

En utilisant les statistiques nationales de la DREES en 2016 sur les contrats complémentaires santé individuels (présentées dans BARLET et al. (2019)), définis en trois niveaux de gamme (entrée de gamme, gamme moyenne et haut de gamme), deux portefeuilles d'assurés avec le choix de niveau de couverture sont générés à partir de la déformation de la proportion des contrats suivant l'avis d'experts sur les trois variables de segment. Le premier portefeuille est basé seulement sur trois variables de segmentation de la base de sinistres, alors que le deuxième prend en compte les niveaux de vie par segment ainsi qu'un proxy de qualité de remboursement des contrats (illustré par la figure 3).

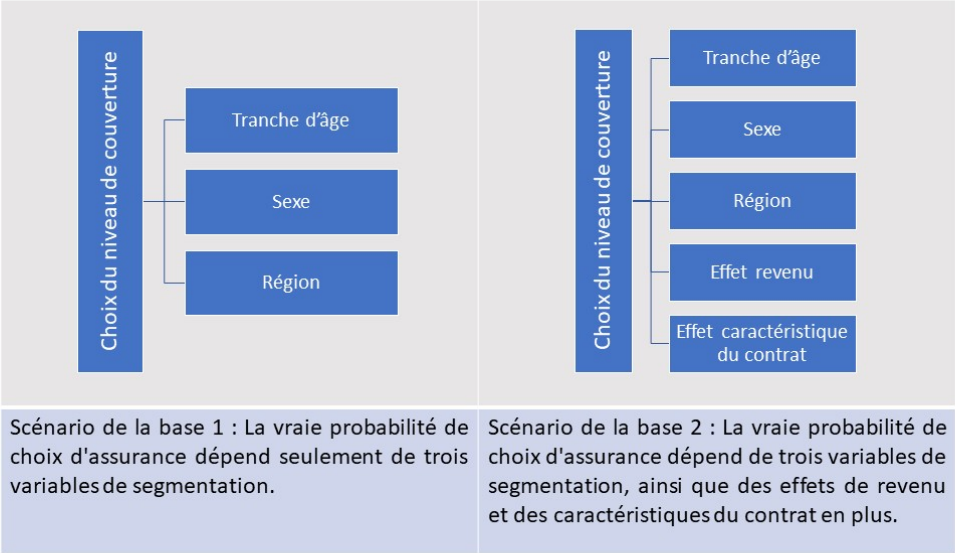


FIGURE 3 : Différence de la structure de relation de deux bases de données

Théorie de l'utilité aléatoire et modèle de choix discret

Le mémoire se concentre sur la modélisation et l'analyse des préférences des assurés pour les trois niveaux de couverture lors de leur souscription à une complémentaire santé individuelle. La théorie de l'utilité aléatoire, développée depuis les années 1960, fournit un cadre d'analyse probabiliste utile pour expliquer les préférences pour les biens en fonction des caractéristiques du bien et de l'individu considéré. Les choix d'alternatives sont mutuellement exclusifs et chaque individu évalue chaque choix avec une quantité d'utilité associée, en choisissant celui avec la plus grande utilité parmi les choix disponibles.

Les préférences des individus sont souvent imparfaites, ne reflétant pas toujours la meilleure utilité au sens de l'utilité espérée. Elles sont donc modélisées par une quantité d'utilité aléatoire, notée U_{ni}, pour l'individu n sur le choix i, avec $n \in \{1, \ldots, N\}$ et $i \in \{1, \ldots, I\}$. N et I représentent respectivement le nombre de personnes dans la base de données et le nombre d'alternatives dans le panier de choix. Chaque U_{ni} est décomposé en deux parties : l'utilité déterministe V_{ni} et le résidu aléatoire inobservable ϵ_{ni}, tel que $U_{ni} = V_{ni} + \epsilon_{ni}$.

Les composants déterministes sont davantage décomposés sous l'hypothèse d'additivité de l'utilité marginale :

$$V_{ni} = V_X(X_n) + V_Z(Z_i) + V_{X,Z}(X_n, Z_i),$$

où :

- V_{ni} représente la composante déterministe de l'utilité de l'alternative i pour l'individu n,
- $V_X(X_n))$ est la portion d'utilité liée aux caractéristiques de l'individu n,
- $V_Z(Z_i)$ désigne la portion d'utilité de l'alternative i liée aux attributs de cette dernière, et
- $V_{X,Z}(X_n, Z_i)$ représente la portion d'utilité résultant des interactions entre les attributs de l'alternative i et les caractéristiques de l'individu n.

Pour chaque panier de choix disponible $C_n = \{1, \dots, I_n\}$ de l'individu n, les composantes aléatoires ϵ_{ni}, $i \in C_n$ sont supposées indépendantes et identiquement distribuées selon la loi de Gumbel II. Le modèle de choix discret de base, appelé modèle logit multinomial (MNL), exprime les probabilités de choix comme suit :

$$P(i|\mathcal{C}_n) = P_n(i|\mathcal{C}_n) = \frac{e^{V_{ni}}}{\sum_{j=1}^{I_n} e^{V_{nj}}}.$$

Plusieurs variantes de ce modèle sont considérées pour assouplir l'hypothèse IANP restrictive du modèle, en permettant une hétérogénéité inobservable dans les préférences des individus. Ces modèles sont testés par rapport au modèle simple à l'aide du test du modèle emboîté afin de vérifier la présence de l'hypothèse IANP.

L'entraînement du modèle sur les deux portefeuilles fictifs a conduit aux mêmes conclusions : le modèle MNL est préféré et il n'y a pas d'hétérogénéité inobservable dans les données simulées (l'hypothèse IANP est donc respectée). L'effet de revenu, la diminution marginale de l'utilité du revenu ainsi que l'effet caractéristique du contrat sont ainsi retrouvés par le modèle sur le deuxième portefeuille, tandis que le premier portefeuille les rejette. Le modèle retenu pour le deuxième portefeuille dans ce mémoire est appelé Logit_ASR_PS :

$\forall i \in \{1,2,3\}, V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n = j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n = k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n = l]} + \beta_{PLAF} PLAF_{ni} + \beta_{FIN} FIN_{ni}$,

où $\forall i \in \{1,2,3\}, \begin{cases} FIN_{ni} = Box\_Cox(Niv\_de\_vie_n - Prime_{ni}, \lambda) \\ PLAF_{ni} = \frac{Prime_{ni}}{Prime_{n2}} \end{cases}$.

Coefficients d'anti-sélection globaux des niveaux de couverture

Sous l'hypothèse de séparabilité de l'aléa moral et de l'anti-sélection dans le cadre des contrats individuels, la différence entre les sinistralités moyennes réelles et les sinistralités moyennes ciblées donne la mesure de l'anti-sélection. Le coefficient global est défini pour un niveau de couverture $i \in \{1,2,3\}$ comme le quotient de ces deux sinistralités (par exemple, pour la couverture minimum $i = 1$) :

$$Coef\_global_1 = \frac{Sin\_moy_1}{Sin\_portef\_moy_1}.$$

En passant à la limite lorsque la taille du portefeuille tend vers l'infini, le coefficient converge en probabilité vers :

$$Coef\_global_1 \xrightarrow[N \to +\infty]{\mathcal{P}} \frac{\frac{\sum_{a=1}^{A} \%^a E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1)}{\sum_{a=1}^{A} \%^a E[\theta^a|C^a=1]\mathbb{P}(C^a=1)}}{\frac{\sum_{a=1}^{A} \%^a (E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1) + E[g_1(S^a)|C^a=2]\mathbb{P}(C^a=2) + E[g_1(S^a)|C^a=3]\mathbb{P}(C^a=3))}{\sum_{a=1}^{A} \%^a (E[\theta^a|C^a=1]\mathbb{P}(C^a=1) + E[\theta^a|C^a=2]\mathbb{P}(C^a=2) + E[\theta^a|C^a=3]\mathbb{P}(C^a=3))}}.$$

Résultats du calcul des coefficients d'anti-sélection

La combinaison de la sinistralité moyenne et des probabilités de choix agrégées prédites par le modèle, calculées sur 182 segments ($A = 182$), permet d'illustrer le niveau d'anti-sélection existant dans les deux portefeuilles (voir le tableau 1).

Les coefficients globaux semblent confirmer la présence d'anti-sélection. Cependant, un coefficient global sur l'ensemble des postes de garantie risque de masquer certaines informations liées à l'anti-sélection. Pour mieux comprendre, deux postes spécifiques ont été évalués : les consultations chez un médecin généraliste et les prothèses auditives hors 100% Santé, dont les résultats sont présentés dans le tableau 1.

Base d'assuré 1	Entrée de gamme	Moyenne gamme	Haut de gamme
Prothèse auditive	57.393%	100.755%	167.603%
Consultation généraliste	98.312%	100.872%	101.072%
Ensemble de contrat	84.511%	101.610%	117.838%
Base d'assuré 2	Entrée de gamme	Moyenne gamme	Haut de gamme
Prothèse auditive	70.710%	98.705%	166.339%
Consultation généraliste	99.974%	98.652%	101.370%
Ensemble de contrat	90.057%	99.833%	117.811%

TABLE 1 : Comparaison des coefficients d'anti-sélection sur les postes de garantie

Concernant les consultations chez un médecin généraliste, il n'a pas été observé de forte anti-sélection, ce qui suggère que ce poste de garantie ne soit pas significativement affecté par ce phénomène. Cependant, cette conclusion semble en contradiction avec des études antérieures qui ont identifié un risque moral associé aux niveaux de remboursement différents pour les médecins. Cette discordance pourrait être due à la simplification des données, qui n'a pas pris en compte la distinction entre les médecins adhérant à OPTAM/OPTAM-CO.

En revanche, pour les prothèses auditives, une forte anti-sélection a été constatée. Les assurés optent souvent pour des couvertures plus élevées, ce qui entraîne des sinistres nettement plus élevés que la moyenne du portefeuille. Cette tendance s'explique par le fait que les jeunes, qui sont plus susceptibles d'opter pour une couverture minimale, ont une sinistralité plus faible, tandis que les personnes âgées, qui optent souvent pour une couverture maximale, ont une sinistralité plus élevée. Ainsi, la composition démographique du portefeuille influence fortement les sinistres dans cette catégorie.

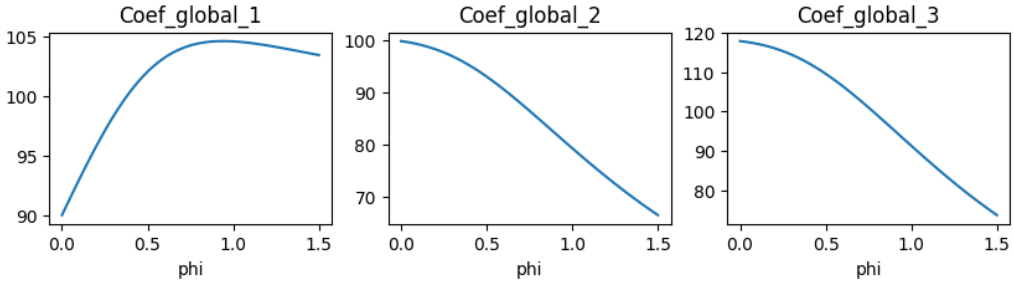
Impact du changement de tarification sur l'état d'anti-sélection

L'analyse a montré une forme d'anti-sélection où les individus à haut risque souscrivent en masse aux niveaux de couverture élevés, tandis que les individus à faible risque se concentrent sur le niveau minimum de couverture, entraînant ainsi une hausse de la sinistralité sur les contrats plus généreux et potentiellement une spirale de la mort pour ces contrats. En se concentrant sur la deuxième base, où l'effet du salaire est considérablement élevé, accompagné d'une diminution marginale de l'utilité du revenu ($\beta_{FIN} = 1.03$ et $\lambda = 0.68$), l'introduction de majorations sur les contrats à haut niveau de couverture dissuadera les individus à haut risque de souscrire. Les majorations consistent à multiplier les primes actuelles proposées aux assurés par un coefficient de majoration. Ce mémoire propose deux façons de majorer les primes, comme indiqué dans la table 2 : une approche homogène pour les contrats de niveau moyen et maximum ou une approche non homogène.

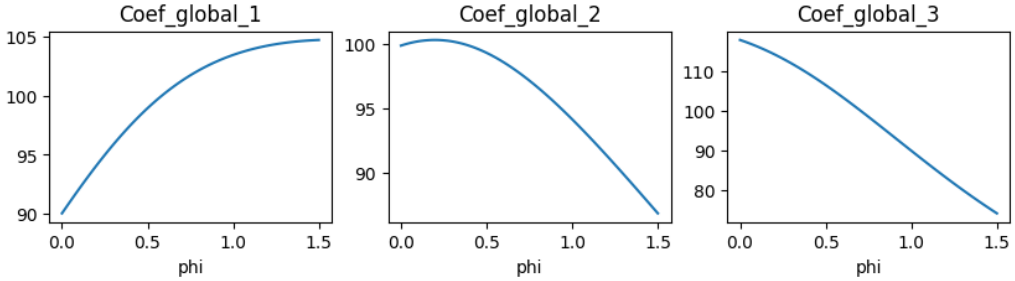
Niveau de couverture	Stratégie homogène	Stratégie non homogène
Minimum	1	1
Moyen	$1 + \phi$	$1 + 0.5 \times \phi$
Maximum	$1 + \phi$	$1 + \phi$

TABLE 2 : Coefficient de majoration des primes proposées selon les niveaux de couverture

L'application des coefficients de majoration prouve son effet d'améliorer l'équilibre de sinistralité des contrats (voir les figures 4). En effet, la distorsion de sinistres baisse rapidement avec l'augmentation de ϕ pour les deux niveaux de couverture les plus élevés, tandis qu'elle se stabilise vers une légère hausse de sinistralité pour le niveau de couverture minimum. La méthode d'ajustement non homogène des primes donne de meilleurs résultats concernant le transfert de profils à haut risque entre les couvertures minimum et moyenne.



(a) Scénario avec l'augmentation homogène : Changement des coefficients d'anti-sélection selon ϕ



(b) Scénario avec l'augmentation non-homogène : Changement des coefficients d'anti-sélection selon ϕ

FIGURE 4 : L'évolution de coefficient anti-sélection en fonction de ϕ pour deux méthodes de majoration

D'autre part, l'agrégation des segments des primes d'assurance semble encourager la souscription de contrats de meilleure couverture pour les hauts risques, ce qui aggrave l'état d'anti-sélection. Une analyse du mouvement lié au passage d'une tarification totalement segmentée à une tarification unique confirme cette conclusion et souligne l'importance des variables telles que le sexe, la région et la tranche d'âge lorsqu'elles sont utilisées pour tarifer les contrats complémentaires santé, bien que l'utilisation du sexe à des fins de tarification soit interdite dans la réalité.

Conclusion

En se concentrant sur un portefeuille individuel, le phénomène d'anti-sélection pourrait être séparé et mesuré par deux facteurs : la préférence pour le risque et les niveaux de risque. L'étude des dépenses sur la base Open Damir permet de dégager les profils à haut risque ainsi que les postes de garantie très diversifiés en termes de risque. D'autre part, l'approche par modèle de choix discret confirme la pertinence de sa capacité de modélisation de référence ainsi que ces coefficients interprétables. Néanmoins, cette méthode reste très dépendante d'hypothèses supposées (parfois difficilement testables) lors de la spécification de la formule des utilités. L'hétérogénéité inobservable demeure un défi pour les assureurs qui cherchent à la comprendre, car elle constitue un problème endogène au choix des assurés en matière d'assurance et éventuellement de leurs dépenses en santé. Des limitations liées au manque de données réelles et au temps de calcul des modèles ont été identifiées. Des pistes de recherche future ont été proposées, notamment en termes de modélisation des comportements des assurés et d'exploitation des données disponibles.

Synthesis note

Adverse selection in health insurance and its issues

For a long time, adverse selection has been considered a moral hazard problem harming the insurance market. In the field of health insurance, particularly in the individual supplementary health insurance market, this problem is even more pronounced due to the nature of individual risk preferences, which are often unobservable, and because insurers struggle to control the information asymmetry in favor of policyholders. Even by offering different reimbursement levels as an incentive for self-selection, these efforts are not sufficient to solve the adverse selection problem, leading to increased costs for contracts offering extensive coverage and encouraging policyholders to opt for less comprehensive coverage, thus creating a negative spiral for these contracts. In this context, it is crucial for insurers to understand the motivations of supplementary health insurance subscribers and to be able to measure the distortion of actual claims experience in their portfolio compared to their targeted claims experience, in order to adjust their pricing strategy.

This thesis aims to propose an econometric methodological guide to address the adverse selection problem in an individual health insurance portfolio. The main objective is to model and understand policyholders' preferences, as well as to quantify the difference in claims experience caused by the distortion in the distribution of policyholders across three coverage levels in the portfolio. Based on this foundation, the models are used to assess the impact of a rate change on policyholders' subscription in the portfolio, leading to an analysis of changes in the distribution of coverage levels.

General methodology for analyzing adverse selection

The thesis relies on the method outlined in Figure 5, which illustrates a response to the thesis problem. The thesis framework is set in the context of a fictitious insurer offering various individual supplementary health insurance contracts within fictitious portfolios.

This thesis utilizes the 2021 public Open Damir database to construct insurance premiums and aggregated claims by insured segment across three coverage levels, defined by age range, region, and gender. Additionally, national statistics are used to create two insured databases with a sample of 50,000 rows, adding the variable of the standard of living - income. Adverse selection indices measure the distortion in the actual claims level of contracts compared to their average level across the entire portfolio. These adverse selection coefficients are relevant as they allow for a return to choice probabilities, linked to the use of discrete choice models. The models are then trained and tested to validate each specification. The predicted probabilities by the model on the disaggregated database are aggregated to formulate the adverse selection coefficients. A specific analysis of the state of adverse selection is thus conducted, along with the impact of a change in pricing.

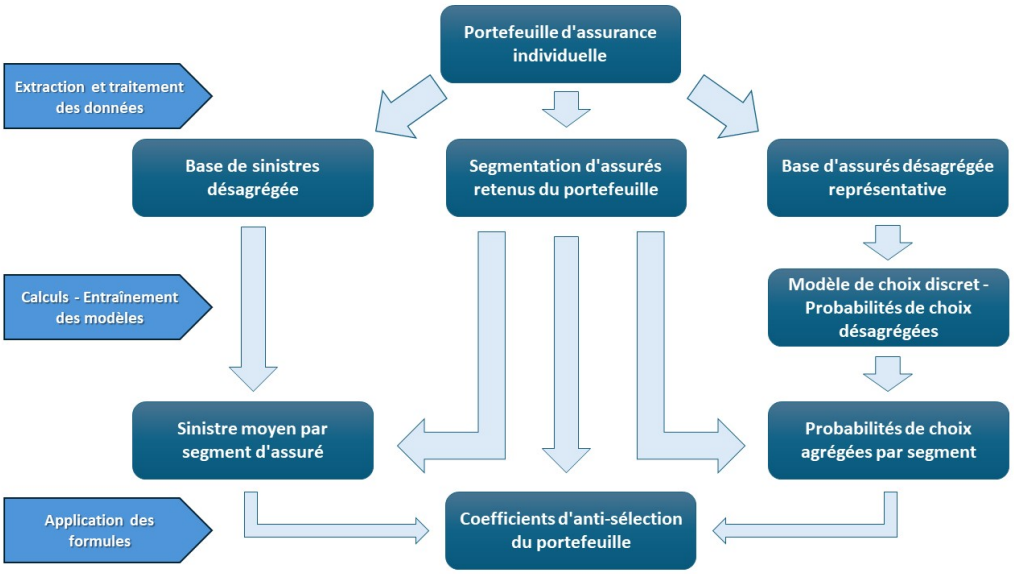


Figure 5: Adverse selection modeling strategy using discrete choice model

Open Damir database and construction of insured segments

The Open DAMIR database records reimbursements for medical care provided by all health insurance schemes, except for a large portion of hospital expenses in the public sector. It contains details on medical procedures, care recipients, and healthcare professionals, while preserving their anonymity through data aggregation. The data from the year 2021 are used in this thesis. There are reimbursements from special schemes such as the complementary solidarity health insurance CSS (CMUC) or the Alsace-Lorraine scheme. The thesis separates these types of reimbursements as well as the treatment of inconsistent reimbursement lines in order to create a database ready to apply complementary health insurance reimbursement formulas. During the processing of the database, three variables are chosen for their relevance in understanding differences in healthcare consumption among profiles: age group (the category of insured individuals under 20 years old is eliminated), region, and gender.

Construction of the average claims database and supplementary health insurance premiums

The thesis presents a benchmark coverage sheet from Forsides with three coverage levels: minimum, medium, and maximum. It consists of several coverage items commonly found in supplementary health insurance grids on the market:

- Common Care: General practitioner consultations; Specialist consultations; Medical imaging procedures; Medical technical procedures; Laboratory analyses and examinations; Paramedical fees; Pharmacy items reimbursed by RO (Compulsory Scheme); Medical equipment.
- Optics: Frame; Single lenses & frame; Complex or very complex lenses & frame; Lenses accepted by RO; Lenses refused by RO (Compulsory Scheme); Refractive surgery.
- Dental: Care; Inlay-Onlay; Periodontology; Prostheses; Orthodontics; Implantology.
- Hearing Aids: Hearing aid reimbursed by RO (Compulsory Scheme).

- Hospitalization: Daily allowance; Private room in Hospitalization/Psychiatry; Stay costs; Fees; Accompanying expenses; Transportation.
- Others: Thermal cure; Alternative medicine; Care abroad.

The coverage is applied to the lines of the Open Damir database for the year 2021 after database processing, thus creating three new variables: the reimbursement for minimum level coverage, medium level coverage, and maximum level coverage. The aggregation of these segmentation variables constructs the aggregated claims database by segment.

The pricing of a supplementary health insurance contract consists of adding up all the rates of coverage categories. The pricing formula for a coverage category by the experience rating method is:

$$\text{Average claim} = \text{Pure premium} = \frac{\text{Total past claims amount}}{\text{Total exposure to risk}}$$

This thesis specifically employs this method due to the aggregated nature of the reimbursements in the Damir database. The aggregated claims by segment correspond to the amount of past claims, and the total exposure to risk is assimilated to the number of individuals in each segment in the population database from INSEE between 2021 and 2022. The annual pure premiums retained for the 3 levels of coverage are finally segmented into 182 segments, considering the considerable difference in average claims between the two gender classes (see the example in Figure 6).

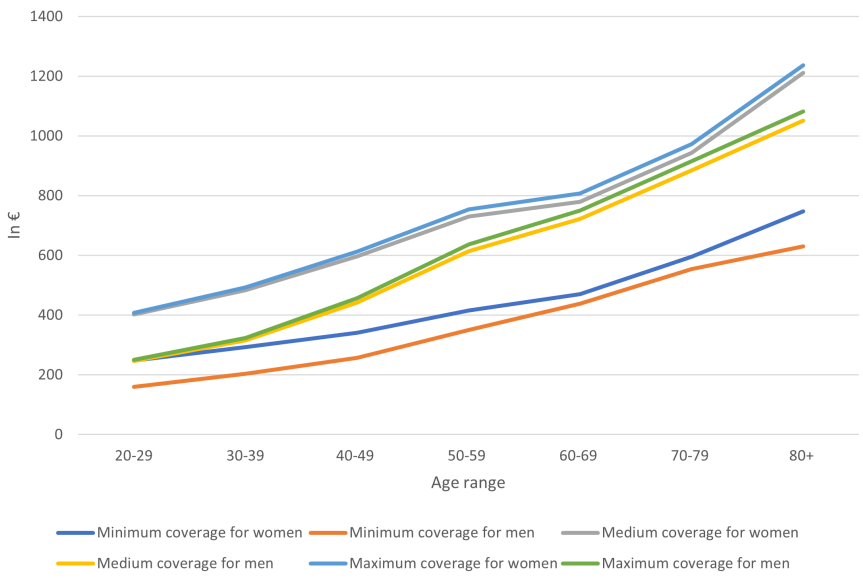
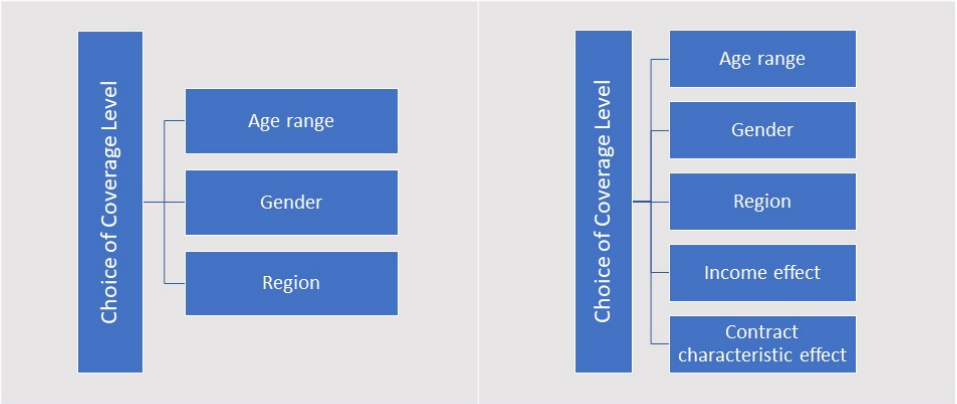


Figure 6: Pure premiums per age group in the Provence-Alpes-Côte d'Azur and Corsica regions, illustrating the difference in claims between men and women.

Construction of the Insured Database

Using national statistics from the DREES in 2016 on individual supplementary health insurance contracts (presented in Barlet et al. (2019)), defined in three levels of coverage (entry-level, mid-range, and high-end), two portfolios of insured individuals with coverage level choices are generated based on the adjustment of the proportion of contracts according to expert advice on the three segmentation variables. The first portfolio is based solely on three segmentation variables from the claims database,

while the second takes into account income levels per segment as well as a proxy for the quality of reimbursement of the contracts (illustrated in Figure 7).



Scenario for Base 1: The actual probability of insurance choice depends only on three segmentation variables.

Scenario for Base 2: The actual probability of insurance choice depends on three segmentation variables, as well as income effects and contract characteristics.

Figure 7: Difference in the relational structure of two databases

Random Utility Theory and Discrete Choice Model

The thesis focuses on modeling and analyzing the preferences of policyholders for the three levels of coverage when subscribing to individual supplementary health insurance. Random utility theory, developed since the 1960s, provides a useful probabilistic framework for explaining preferences for goods based on the characteristics of both the good and the individual. Choices of alternatives are mutually exclusive, and each individual evaluates each choice with an associated utility, selecting the one with the greatest utility among the available choices.

Individual preferences are often imperfect, not always reflecting the best utility in terms of expected utility. They are therefore modeled by a random utility quantity, denoted U_{ni}, for individual n on choice i, with $n \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, I\}$. N and I represent respectively the number of individuals in the database and the number of alternatives in the choice set. Each U_{ni} is decomposed into two parts: the deterministic utility V_{ni} and the unobservable random residue ϵ_{ni}, such that $U_{ni} = V_{ni} + \epsilon_{ni}$.

The deterministic components are further decomposed under the assumption of additivity of marginal utility:

$$V_{ni} = V_X(X_n) + V_Z(Z_i) + V_{X,Z}(X_n, Z_i),$$

where:

- V_{ni} represents the deterministic component of the utility of alternative i for individual n,
- $V_X(X_n)$ is the portion of utility related to the characteristics of the individual n,

- $V_Z(Z_i)$ denotes the portion of utility of an alternative i related to its attributes, and
- $V_{X,Z}(X_n, Z_i)$ represents the portion of utility resulting from interactions between the attributes of alternative i and the characteristics of individual n.

For each available choice set $C_n = \{1, \dots, I_n\}$ of individual n, the random components ϵ_{ni}, $i \in C_n$ are assumed to be independently and identically distributed according to the Gumbel II distribution. The basic discrete choice model, called the multinomial logit model (MNL), expresses choice probabilities as follows:

$$P(i|\mathcal{C}_n) = P_n(i|\mathcal{C}_n) = \frac{e^{V_{ni}}}{\sum_{j=1}^{I_n} e^{V_{nj}}}.$$

Several variants of this model are considered to relax the restrictive IIA assumption of the model, allowing for unobservable heterogeneity in individuals' preferences. These models are tested against the simple model using the nested model test to check for the presence of the IIA assumption.

Training the model on the two fictitious portfolios led to the same conclusions: the MNL model is preferred and there is no unobservable heterogeneity in the simulated data (thus the IIA assumption is respected). The income effect, the marginal utility decrease of income, as well as the contract characteristic effect, are thus identified by the model in the second portfolio, while the first portfolio rejects them. The model chosen for the second portfolio in this thesis is called Logit_ASR_PS:

$\forall i \in \{1,2,3\}, V_{ni} = \beta_{i,const} + \sum_{j \in \text{Age}} \beta_{i,j} \mathbb{1}_{[age_n = j]} + \sum_{k \in \text{Gender}} \beta_{i,k} \mathbb{1}_{[gender_n = k]} + \sum_{l \in \text{Region}} \beta_{i,l} \mathbb{1}_{[region_n = l]} + \beta_{PLAF} PLAF_{ni} + \beta_{FIN} FIN_{ni}$,

where $\forall i \in \{1,2,3\}, \begin{cases} FIN_{ni} = \text{Box\_Cox}(Income_n - Premium_{ni}, \lambda) \\ PLAF_{ni} = \frac{Premium_{ni}}{Premium_{n2}} \end{cases}$.

Global Adverse Selection Coefficients for Coverage Levels

Under the assumption of separability of moral hazard and adverse selection in individual contracts, the difference between actual average claims and targeted average claims provides a measure of adverse selection. The global coefficient is defined for a coverage level $i \in \{1,2,3\}$ as the ratio of these two claims (for example, for the minimum coverage $i = 1$):

$$Coef\_global_1 = \frac{Sin\_avg_1}{Sin\_portfolio\_avg_1}.$$

As the portfolio size tends to infinity, the coefficient converges in probability to:

$$Coef\_global_1 \xrightarrow[N \to +\infty]{\mathcal{P}} \frac{\frac{\sum_{a=1}^{A} \%^a E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1)}{\sum_{a=1}^{A} \%^a E[\theta^a|C^a=1]\mathbb{P}(C^a=1)}}{\frac{\sum_{a=1}^{A} \%^a (E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1) + E[g_1(S^a)|C^a=2]\mathbb{P}(C^a=2) + E[g_1(S^a)|C^a=3]\mathbb{P}(C^a=3))}{\sum_{a=1}^{A} \%^a (E[\theta^a|C^a=1]\mathbb{P}(C^a=1) + E[\theta^a|C^a=2]\mathbb{P}(C^a=2) + E[\theta^a|C^a=3]\mathbb{P}(C^a=3))}}.$$

Results of Adverse Selection Coefficients Calculation

The combination of average claims and predicted aggregated choice probabilities by the model, calculated over 182 segments ($A = 182$), illustrates the level of adverse selection existing in the two portfolios (see Table 3).

The global coefficients seem to confirm the presence of adverse selection. However, a global coefficient across all coverage categories may obscure some information related to adverse selection. To

Insured Database 1	Entry Level	Mid-Range	High End
Hearing Prosthesis	57.393%	100.755%	167.603%
General Practitioner Consultation	98.312%	100.872%	101.072%
Overall Contract	84.511%	101.610%	117.838%
Insured Database 2	Entry Level	Mid-Range	High End
Hearing Prosthesis	70.710%	98.705%	166.339%
General Practitioner Consultation	99.974%	98.652%	101.370%
Overall Contract	90.057%	99.833%	117.811%

Table 3: Comparison of Adverse Selection Coefficients on Coverage Categories

better understand this, two specific coverage categories were evaluated: consultations with a general practitioner and hearing aids outside 100% Health, the results of which are presented in Table 3.

Regarding consultations with a general practitioner, no strong adverse selection was observed, suggesting that this coverage category is not significantly affected by this phenomenon. However, this conclusion seems to contradict previous studies that have identified moral hazard associated with different reimbursement levels for doctors. This discrepancy could be due to the simplification of data, which did not take into account the distinction between doctors adhering to OPTAM/OPTAM-CO.

In contrast, for hearing aids, significant adverse selection was observed. Insured individuals often opt for higher coverage, resulting in claims significantly higher than the portfolio average. This trend is explained by the fact that younger individuals, who are more likely to opt for minimal coverage, have lower claims, while older individuals, who often opt for maximal coverage, have higher claims. Thus, the demographic composition of the portfolio strongly influences claims in this category.

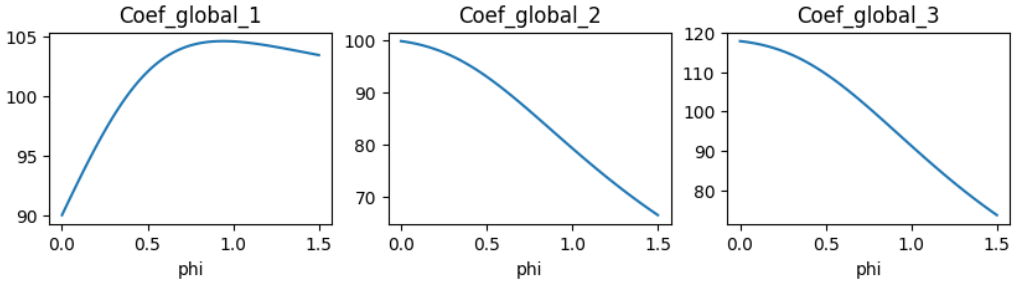
Impact of changes in premium rating on the state of adverse selection

The analysis has shown a form of adverse selection where high-risk individuals subscribe massively to high coverage levels, while low-risk individuals concentrate on the minimum coverage level, resulting in an increase in claims on more generous contracts and potentially a death spiral for these contracts. Focusing on the second database, where the income effect is significantly high, accompanied by a marginal decrease in income utility ($\beta_{FIN} = 1.03$ and $\lambda = 0.68$), introducing surcharges on high coverage contracts will discourage high-risk individuals from subscribing. The surcharges consist of multiplying the current premiums offered to policyholders by a surcharge coefficient. This paper proposes two ways to increase premiums, as shown in Table 4: a homogeneous approach for medium and maximum coverage contracts or a non-homogeneous approach.

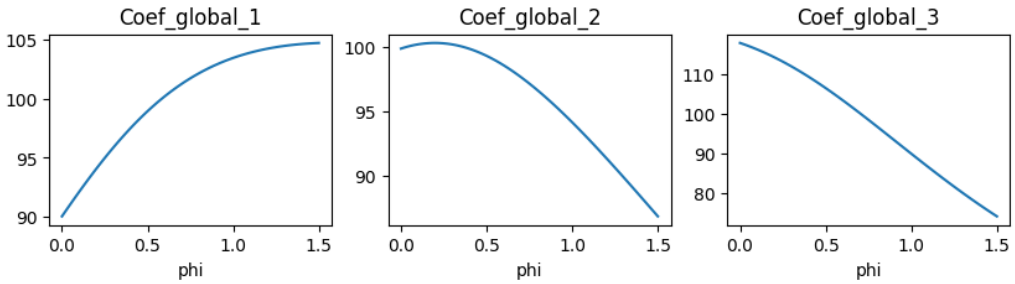
Coverage Level	Homogeneous Strategy	Non-homogeneous Strategy
Minimum	1	1
Medium	$1 + \phi$	$1 + 0.5 \times \phi$
Maximum	$1 + \phi$	$1 + \phi$

Table 4: Surcharge Coefficients for Proposed Premiums by Coverage Levels

The application of surcharge coefficients proves its effectiveness in improving the claims balance of the contracts (see Figure 8). Indeed, the claim's distortion decreases rapidly with the increase of ϕ for the two highest coverage levels, while it stabilizes towards a slight increase in claims for the minimum coverage level. The non-homogeneous premium adjustment method yields better results regarding the transfer of high-risk profiles between minimum and medium coverage levels.



(a) Scenario with homogeneous surcharge: Change in anti-selection coefficients according to ϕ



(b) Scenario with non-homogeneous surcharge: Change in anti-selection coefficients according to ϕ

Figure 8: The evolution of the adverse selection coefficient as a function of ϕ for two surcharge methods.

On the other hand, the aggregation of insurance premium segments seems to encourage the subscription to better coverage contracts for high-risk individuals, thus exacerbating the state of adverse selection. An analysis of the movement, associated with transitioning from fully segmented pricing to uniform pricing, confirms this conclusion and underscores the importance of variables such as gender, region, and age group when used to price supplementary health insurance contracts, even though the use of gender for pricing is prohibited in reality.

Conclusion

Focusing on an individual portfolio, the phenomenon of adverse selection could be separated and measured by two factors: risk preference and levels of risk. Studying expenses based on the Open Damir dataset helps identify high-risk profiles as well as guarantees with varying levels of risk. On the other hand, the discrete choice model approach confirms the relevance of its modeling capacity as well as these interpretable coefficients. However, this method remains highly dependent on assumed (sometimes difficult to test) hypothesis when specifying utility formulas. Unobservable heterogeneity remains a challenge for insurers seeking to understand it, as it constitutes an endogenous problem to insureds' choices in insurance and potentially their healthcare spending. Limitations related to the lack of real data and the computational time of models have been identified. Future research directions have been proposed, particularly in terms of modeling insureds' behaviors and utilizing available data.

Remerciements

Je tiens à exprimer mes sincères remerciements à toutes les personnes qui ont contribué à la réalisation de ce mémoire, ainsi qu'à FORSIDES France pour l'opportunité de stage enrichissant.

Je souhaite adresser mes remerciements les plus chaleureux à Charlène FUSIS, senior manager chez Forsides, pour avoir proposé ce sujet très intéressant, pour son encadrement agréable et sa disponibilité tout au long de la rédaction de ce mémoire.

Je tiens également à remercier Quentin GUIBERT, mon tuteur académique, pour son accompagnement durant le stage et ses précieux conseils pour la rédaction de ce mémoire.

Je souhaite accorder une mention spéciale de remerciement à Binh TRINH pour son écoute et son soutien moral pendant le stage.

Un grand merci également à Alexis CHASSAGNE, Benoît REGENT-KLOECKNER et Edouard VIROT pour leurs conseils et relectures sur cette étude pendant mon séjour chez FORSIDES France. Je profite également de cette occasion pour exprimer ma gratitude à tous les collaborateurs de FORSIDES France avec qui j'ai eu le plaisir de travailler, pour leur aide dans mon intégration rapide dans l'équipe et pour l'ambiance conviviale.

Enfin, je remercie ma famille, mes amis Mengru et Guillaume pour leur aide à la relecture, ainsi que pour leur écoute et leurs précieux conseils.

Table des matières

Introduction

L'assurance complémentaire santé constitue un élément essentiel du système de protection sociale en France, voire à l'échelle mondiale. Toutefois, les assureurs de ce secteur, confrontés à une concurrence intense, font face à un défi de taille : l'anti-sélection. Ce phénomène résulte de l'asymétrie d'information, où les assurés choisissent leurs contrats en fonction de leurs besoins spécifiques liés à leur niveau de risque, que l'assureur ne peut pas connaître. Cela crée un déséquilibre entre les coûts supportés par l'assureur et les primes qu'il perçoit. Une telle pratique peut entraîner une spirale de tarification à la hausse, rendant l'assurance complémentaire santé inaccessible à de nombreuses personnes et compromettant la viabilité de l'offre d'assurance en l'absence de mutualisation. À titre d'exemple, l'assureur Alan a dû cesser la souscription de son produit complémentaire santé individuelle en raison de l'augmentation de la sinistralité due à l'anti-sélection (Alan (2020)).

Ce mémoire cherche à répondre à la question suivante : comment élaborer, à travers la tarification, un modèle à partir d'un portefeuille d'assurance complémentaire santé permettant de quantifier et d'analyser l'impact de l'anti-sélection ? Nous adoptons une approche économique basée sur la préférence de choix pour aborder ce défi complexe en assurance santé. Notre parcours de recherche est jalonné de plusieurs étapes essentielles, chacune contribuant à la compréhension et à la résolution de ce problème majeur. Une des principales contributions de ce mémoire est l'introduction de la théorie de l'utilité aléatoire pour étudier la dynamique de souscription, servant de modèle principal à notre approche.

Outre la construction du modèle, ce mémoire présente l'utilisation de la base de données publique Open Damir, une base de données de remboursement de la Sécurité Sociale couvrant toute la France. Cette base est devenue cruciale pour les assureurs en santé en raison de sa portée nationale et de sa richesse en informations sur la consommation de l'assurance santé par rapport à leurs propres bases de données. Son utilisation garantit une tarification objective, pouvant servir de référence pour évaluer les offres de complémentaire santé.

Cette étude est structurée en cinq chapitres principaux. Le premier chapitre offre une introduction au système de l'assurance santé en France, ainsi qu'à la tarification de l'assurance santé à différents niveaux de couverture. Le deuxième chapitre analyse les dépenses médicales en France à l'aide de la base de données Open Damir, en proposant une tarification des trois niveaux de couverture selon différents segments de la population. Le troisième chapitre présente deux indices mesurant la présence de l'anti-sélection en assurance santé et construit deux scénarios de portefeuilles d'assurance complémentaire santé individuelle basés sur des données nationales. Dans le quatrième chapitre, l'étude introduit la théorie de l'utilité aléatoire, les modèles de choix discrets et leur application sur les portefeuilles fictifs afin d'explorer les préférences des assurés. Enfin, le cinquième chapitre applique les modèles présentés précédemment sur les deux portefeuilles générés, permettant de tirer des conclusions sur l'anti-sélection et d'évaluer l'impact de la tarification à court terme.

Chapitre 1

Système de santé et d'assurance en France

1.1 Histoire de la Sécurité Sociale en France

En France, le système de protection sociale date du 18e siècle. En effet, d'après les informations disponibles sur le site de la Sécurité Sociale La Sécurité Sociale (sans date[b]) et Assurance Maladie (2023d), la loi sur les accidents du travail en 1898 a posé les bases de ce système. Puis en 1945, la création de la Sécurité Sociale a fusionné toutes les anciennes assurances (maladie, retraite...) afin d'assurer à chacun et sa famille la subsistance dans des conditions décentes. Un an après sa création, la Sécurité Sociale intégrait des risques professionnels dans sa fonction d'indemnisation et de prévention des risques. En 1967, une réorganisation conduit à créer trois branches autonomes au sein du régime général de la Sécurité Sociale, dont la création de la caisse nationale de l'assurance maladie des travailleurs salariés (Cnamts). En 1994, l'Assurance Maladie - Risques professionnels acquiert une certaine autonomie et une gestion financière séparée. Par la suite, en 1996, une réforme profonde transforme la Caisse nationale d'assurance maladie et met en place un régime universel d'Assurance Maladie.

En 1998, la carte Vitale voit le jour dans le but de moderniser le processus de remboursement des assurés et simplifier leurs démarches administratives. Cette carte électronique individuelle d'assuré social constitue une preuve d'affiliation à un régime d'Assurance Maladie et atteste des droits spécifiques dont bénéficie chaque personne en fonction de sa situation. Bien qu'elle ne contienne aucune information médicale, elle regroupe toutes les données administratives nécessaires pour le remboursement des soins et la prise en charge hospitalière.

En 2000, la création de la Couverture Maladie Universelle (CMU) marque un tournant majeur dans le système de santé. La CMU de base représente l'abandon du système assurantiel bismarckien. Désormais, la seule condition requise pour être pris en charge par l'Assurance maladie est de résider en France. Puis quatre ans plus tard, en 2004, l'Assurance Maladie fait l'objet d'une réforme visant à améliorer son efficacité et sa gestion des soins. Par la suite, en 2005, la branche Risques Professionnels de l'Assurance Maladie signe sa première convention d'objectifs et de gestion (COG) avec l'État, renforçant ainsi sa coopération. L'année 2015 marque un jalon significatif, avec la célébration des 70 ans de la Sécurité Sociale, témoignant de son importance et de son évolution dans le temps. Au cours de cette même année, l'Assemblée nationale adopte le projet de loi santé instaurant la généralisation du tiers payant (dispense d'avance de frais) à tous les assurés.

En 2016, la CMU évolue vers la Protection Universelle Maladie (PUMa), élargissant ainsi la couverture médicale pour les travailleurs et résidents en France. En 2018, l'intégration de l'Assurance Maladie au régime social des indépendants et à la Sécurité Sociale des étudiants favorise une meilleure cohérence et un accès équitable aux soins pour tous. En 2019, la transition de la CMU-C et de l'aide

au paiement de la complémentaire santé vers la Complémentaire Santé Solidaire (CSS) renforce la protection des personnes à faibles revenus.

En 2021, l'Assurance Maladie joue un rôle essentiel dans la gestion de la pandémie de COVID-19. Des mesures du dispositif "Aller vers" sont mises en place pour faciliter la vaccination des personnes fragiles, isolées ou présentant des facteurs de risques médicaux tels que les ALD et l'immunodépression. Grâce à ce dispositif, environ 300 000 personnes ont été accompagnées vers la vaccination. De plus, en 2022, le lancement de Mon Espace Santé offre aux assurés un accès sécurisé à leurs données de santé et facilite les échanges avec les professionnels de santé. Tout au long de ces étapes, les réformes et mesures adoptées ont renforcé le système de santé et amélioré la prise en charge des assurés.

1.2 Assurance maladie et complémentaire santé

1.2.1 Régime général de la Sécurité Sociale

D'après Assurance Maladie (2023c) et La Sécurité Sociale (sans date[a]), le système de Sécurité Sociale en France est composé de deux régimes principaux, à savoir le régime général et le régime agricole, ainsi que de plusieurs régimes spéciaux. Le régime général couvre la majorité de la population, notamment les travailleurs salariés, les travailleurs indépendants et les résidents bénéficiant de droits au titre de la résidence. Le régime agricole est spécifiquement dédié aux exploitants et aux salariés agricoles. Quant aux régimes spéciaux, ils englobent des professions spécifiques telles que les marins, les mineurs, les employés de la SNCF, de la RATP, d'EDF-GDF, ainsi que ceux de l'Assemblée nationale, du Sénat, des clercs et employés de notaires.

Le régime général est composé de 5 branches, chacune couvrant les grands risques et gérant le recouvrement des cotisations :

- La branche Maladie, qui couvre les conséquences financières d'une maladie, d'un accident du travail ou d'une maladie professionnelle, ainsi que d'une maternité ou paternité,
- La branche Famille,
- La branche Recouvrement,
- La branche Vieillesse,
- La branche Autonomie.

Dans le cadre de ce mémoire, nous nous concentrons sur la branche de l'Assurance maladie. Cette branche est responsable de la prise en charge des dépenses de santé des assurés et garantit l'accès aux soins. Elle joue un rôle important dans la prévention des maladies et contribue à la régulation du système de santé en France. La branche maladie du régime général est gérée par la Caisse nationale de l'Assurance Maladie et son réseau qui comprend les caisses primaires d'assurance maladie (CPAM), les caisses générales de Sécurité Sociale (CGSS) dans les départements d'outre-mer, les directions régionales du service médical (DRSM), les caisses d'assurance retraite et de la santé au travail (Carsat) ainsi que les unions de gestion des établissements de caisse d'assurance maladie (Ugecam). Pour le régime agricole, c'est la MSA (Mutualité sociale agricole) qui est chargée de la gestion de la branche maladie.

En 2022, d'après La Sécurité Sociale (2023),les dépenses de l'assurance maladie représentent 40% des dépenses totales du régime de base de la Sécurité Sociale (voir Figure 1.2). Parmi ces 40% de dépenses financières de l'Assurance Maladie dans le cadre de l'objectif national des dépenses d'assurance maladie – ONDAM, les deux postes les plus importants sont les soins de ville (honoraires des professionnels de santé libéraux, etc.) et les établissements de santé, qui représentent respectivement

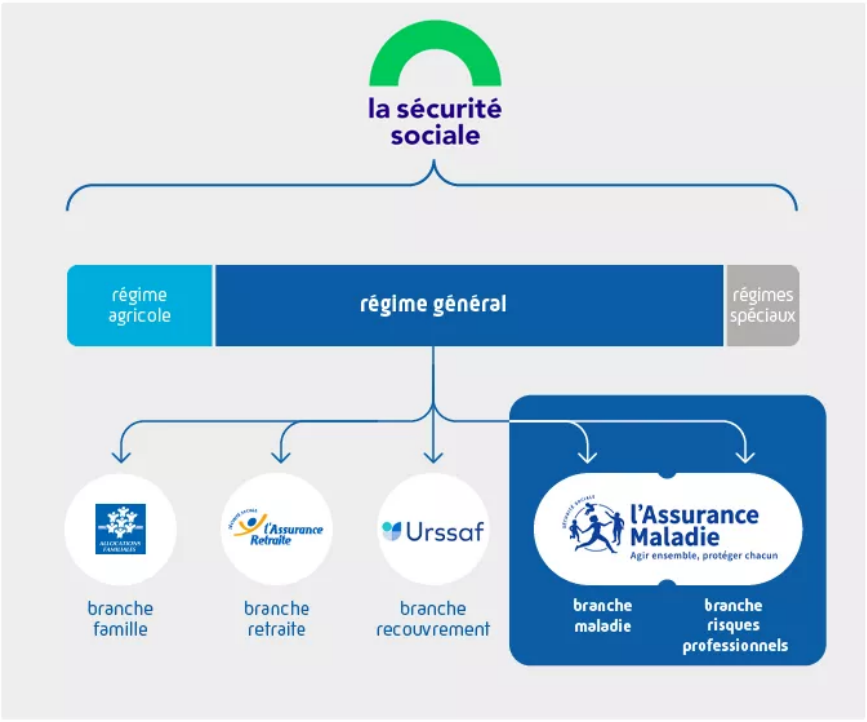


FIGURE 1.1 : Organigramme de la Sécurité Sociale

Source : ASSURANCE MALADIE (2023c)

43.5% et 39.8% des dépenses (voir Figure 1.3). À savoir l'objectif national de dépenses d'assurance maladie (ONDAM) fixe un plafond de dépenses pour les soins de santé en ville, à l'hôpital (public ou privé) et dans les centres médico-sociaux.

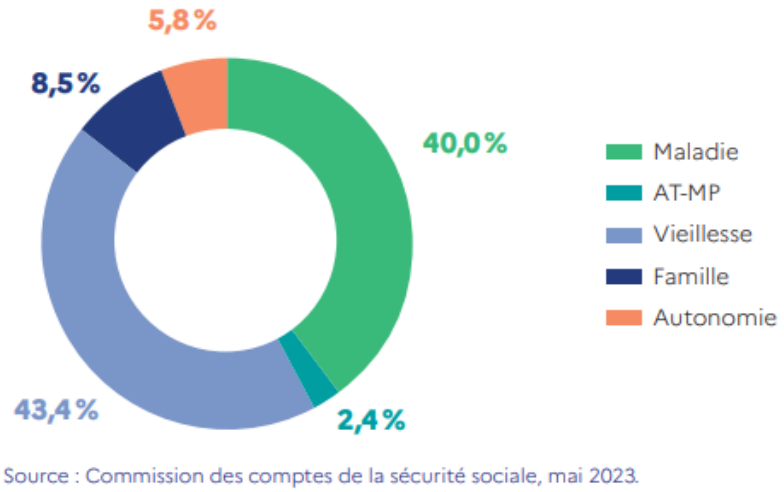


FIGURE 1.2 : Part de chaque branche dans les dépenses des régimes de base en 2022

Source : LA SÉCURITÉ SOCIALE (2023)

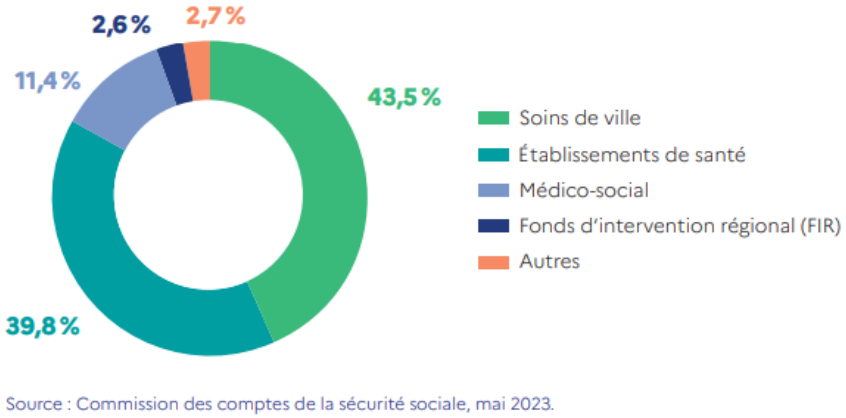
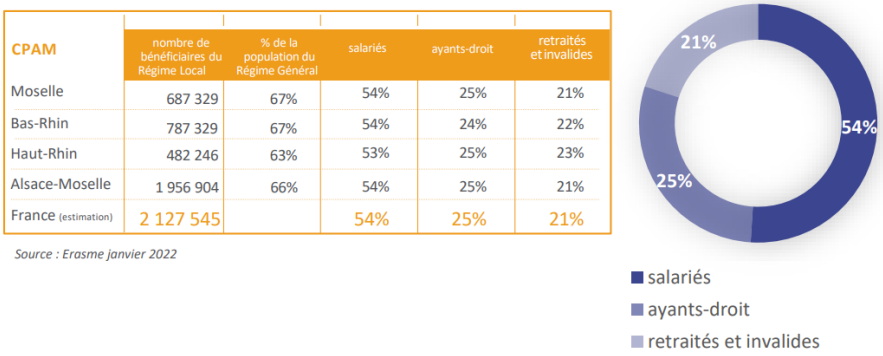


FIGURE 1.3 : Dépenses de santé financées par l'assurance maladie dans le cadre de l'Ondam en 2022

Source : LA SÉCURITÉ SOCIALE (2023)

1.2.2 Régime local Alsace-Moselle

Le régime local de la Sécurité Sociale en vigueur en Alsace-Moselle est considéré comme l'un des régimes spéciaux. Il est géré par la Caisse Primaire d'Assurance Maladie d'Alsace-Moselle (CPAM) et offre des prestations de Sécurité Sociale plus étendues que celles du régime général français, notamment en ce qui concerne le remboursement des frais médicaux et les indemnités journalières en cas d'arrêt de travail. Ce régime présente une particularité par rapport au régime général de la Sécurité Sociale : il fonctionne à la fois comme un régime obligatoire, avec les remboursements régis par le Code de la Sécurité Sociale, et comme une partie de versement complémentaire au-delà du régime général, dans un esprit de solidarité envers la population affiliée afin de compenser le taux de cotisation élevée sans l'engagement des employeurs.



CPAM	nombre de bénéficiaires du Régime Local	% de la population du Régime Général	salariés	ayants-droit	retraités et invalides
Moselle	687 329	67%	54%	25%	21%
Bas-Rhin	787 329	67%	54%	24%	22%
Haut-Rhin	482 246	63%	53%	25%	23%
Alsace-Moselle	1 956 904	66%	54%	25%	21%
France (estimation)	2 127 545		54%	25%	21%

Source : Erasme janvier 2022

FIGURE 1.4 : Statistique sur le régime local d'Alsace Moselle 2021

Source : RÉGIME LOCAL D'ASSURANCE MALADIE D'ALSACE MOSELLE (2022)

Selon RÉGIME LOCAL D'ASSURANCE MALADIE D'ALSACE MOSELLE (2022), il a été constaté que 92% des bénéficiaires de la Sécurité Sociale dans les départements de Moselle, Bas-Rhin et Haut-Rhin sont affiliés à une Caisse Primaire d'Assurance Maladie (CPAM) de ces départements. Le Régime Local compte un total de 2,1 millions de bénéficiaires (figure 1.4), comprenant des salariés, des chômeurs, des invalides, des retraités et les membres de leur famille à leur charge. Parmi ces bénéficiaires, on compte 1,6 million d'assurés, ce qui représente 75% du total, et 523 000 ayants-droit, soit 25% du total. La

majorité des bénéficiaires (plus de la moitié) sont des actifs, tandis que 21% sont des retraités. Il est également important de noter qu'en Alsace-Moselle, 66% des assurés relevant du Régime Général sont également bénéficiaires du Régime Local.

1.2.3 La complémentaire santé

Malgré le fait d'être couverte par le régime obligatoire de l'Assurance Maladie, le reste à charge pour les assurés après les remboursements de la Sécurité Sociale seule demeure élevé, ce qui peut conduire à des renoncements aux soins.

Un contrat complémentaire santé, souscrit par un individu ou une entreprise, a pour objectif de compléter la couverture offerte par la Sécurité Sociale. Il vise à prendre en charge les frais médicaux non remboursés ou partiellement remboursés par la Sécurité Sociale, tels que les consultations médicales, les médicaments, les soins dentaires, les frais d'hospitalisation, etc.

Ce contrat est proposé par des organismes d'assurance privés ou des mutuelles et peut être souscrit de manière individuelle, pour soi-même et éventuellement sa famille, ou de manière collective, dans le cadre d'un contrat de groupe proposé par l'employeur à ses salariés.

Le contrat complémentaire santé définit les garanties, les plafonds de remboursement, les exclusions et les modalités de remboursement. Les cotisations à payer pour bénéficier de ce contrat varient en fonction de l'âge, du niveau de garantie choisi et de la composition de la famille. Il est donc essentiel d'étudier attentivement les différentes offres disponibles sur le marché, de comparer les garanties et les tarifs avant de souscrire à un contrat complémentaire santé.

La Sécurité Sociale et les organismes complémentaires forment les structures principales de financement pour la consommation de soins et de biens médicaux de la France (avec 96% de la population couverte par une complémentaire santé suivant les statistiques globales en 2023 de France Assureurs) selon la figure 1.5

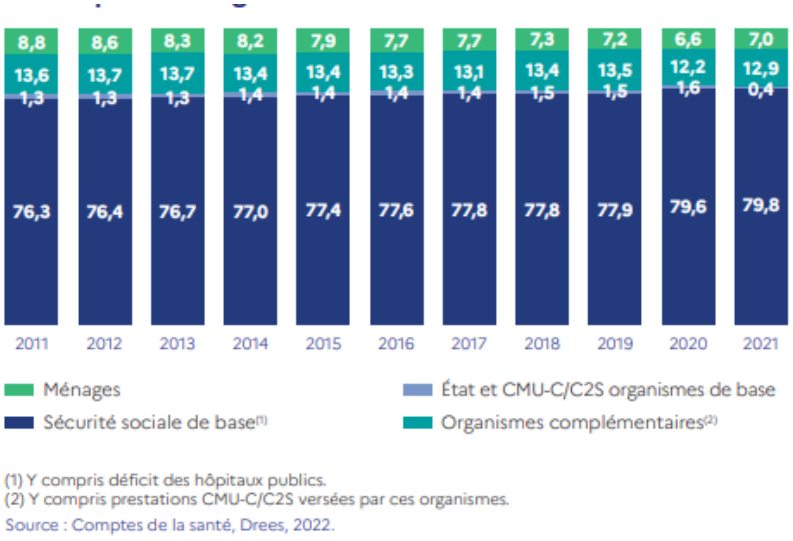


FIGURE 1.5 : Évolution de la structure du financement de la consommation de soins et de biens médicaux en pourcentage

Source : de WILLIENCOURT (2022)

1.2.4 Les réformes dans le domaine de l'assurance santé

Résumées dans de Williencourt (2022), plusieurs réformes ont été menées afin de donner plus de droit et de protection à l'assuré :

- La loi Évin, également connue sous le nom de loi n°89-1009 du 31 décembre 1989, établit des règles protectrices pour les assurés dans le domaine de la prévoyance/santé collective. Ces règles doivent être prises en compte lors de la négociation des accords entre les organismes assureurs et les assurés. De plus, les anciens salariés peuvent bénéficier de la continuité de leur couverture complémentaire dans certaines conditions, conformément à l'article 4 de la loi Évin. Ce dispositif, appelé "portabilité", s'applique à la fois à la santé et à la prévoyance.

- La réforme des contrats responsables du 1er avril 2015 vise à encourager la responsabilisation des assurés en matière de dépenses de santé. Les contrats d'assurance santé dits "responsables" doivent respecter des critères spécifiques, tels que des plafonds sur les remboursements et la non-prise en charge de certaines dépenses jugées excessives. L'objectif est de maîtriser les coûts de la santé et d'inciter les assurés à adopter des comportements de prévention. Cette réforme a pour effet de limiter les garanties offertes par les contrats et de renforcer le rôle des complémentaires santé dans le financement des soins.

- Loi sur la généralisation de la complémentaire santé entrée en vigeur le 1er janvier 2016 (ANI) : Cette réforme a rendu obligatoire la souscription d'une complémentaire santé pour tous les salariés du secteur privé via un contrat de groupe proposé par leur employeur. L'objectif était de garantir une couverture santé complémentaire à tous les travailleurs, en complément de l'Assurance maladie.

- La loi du 14 juillet 2019 permet aux assurés de résilier leur contrat de complémentaire santé à tout moment après un an de souscription, sans frais ni pénalité. Le décret du 24 novembre 2020 précise les contrats concernés, qui couvrent les risques liés à la santé. Depuis le 1er décembre 2020, les assurés peuvent bénéficier de ce droit de résiliation, y compris pour les contrats en cours avant cete date.

- La réforme du "100% Santé", mise en place par le décret n° 2019-21 du 11 janvier 2019, vise à lutter contre la hausse du renoncement aux soins pour les soins dentaire, l'optique et l'audiologie. Son objectif est de garantir un accès aux équipements d'optique, aides auditives et soins prothétiques dentaires sans reste à charge pour les patients. Cette réforme propose des paniers spécifiques de soins et d'équipements, qui sont entièrement remboursés par la Sécurité Sociale et les organismes complémentaires dans le cadre des contrats solidaires et responsables. Les équipements du panier "Classe I" sont entièrement pris en charge, tandis que ceux du panier "Classe II" ont des tarifs libres et une prise en charge limitée par les organismes complémentaires. De plus, la réforme établit différents paniers de soins prothétiques dentaires, avec des limites de facturation et des tarifs encadrés, afin d'offrir des options de traitement adaptées aux besoins des patients.

1.3 Les différents organismes proposant des contrats de complémentaire santé

1.3.1 Les institutions de prévoyance

Les institutions de prévoyance sont des entités régies par le code de la Sécurité Sociale, opérant en tant que sociétés de personnes à but non lucratif. Elles utilisent leurs excédents financiers pour

améliorer leurs services, offrir de nouvelles garanties et renforcer la sécurité de leurs engagements. Dotées d'une gestion paritaire, elles réunissent des représentants des salariés et des entreprises au sein de leur Conseil d'administration. Ces institutions couvrent les principaux risques de la vie tels que la maladie, l'incapacité ou l'invalidité, et le décès, en proposant diverses catégories d'assurances définies dans le code de la Sécurité Sociale. En parallèle de leurs activités d'assurance, elles mènent également des actions sociales.

1.3.2 Les mutuelles

Les mutuelles sont régies par le code de la mutualité, où le livre II couvre les organismes pratiquant des opérations d'assurance et de capitalisation, principalement dans le domaine de la complémentaire santé. Une mutuelle est une société de personnes de droit privé à but non lucratif, similaire aux institutions de prévoyance, qui réinvestit ses excédents financiers pour améliorer les services offerts à ses adhérents et renforcer leur protection.

1.3.3 Les sociétés d'assurances

Les entreprises d'assurance (dont les sociétés d'assurance et de réassurance de droit français) sont régies par le Code des Assurances, qui constitue le cadre juridique et réglementaire applicable à leur activité. Ce code vise à encadrer et réguler l'ensemble des opérations d'assurance, y compris la création et le fonctionnement des entreprises d'assurance.

D'après de Willencourt (2022), il est observé que la répartition des contrats individuels et collectifs en matière de santé varie selon le type d'organisme. Les contrats individuels sont généralement souscrits par des particuliers, tandis que les contrats collectifs sont souscrits par des personnes morales, telles que les employeurs, pour couvrir un groupe de personnes physiques, souvent des salariés. Les institutions de prévoyance se spécialisent principalement dans les contrats collectifs, qui représentent 87% de leurs cotisations, tandis que les mutuelles sont largement orientées vers les contrats individuels, qui représentent 67% de leurs cotisations. Les entreprises d'assurance occupent une position intermédiaire, avec 54% des cotisations provenant de contrats collectifs. Il convient de noter que la part des contrats collectifs continue de croître en 2021.

1.3.4 Le dispositif Complémentaire Santé Solidaire (CSS)

La protection universelle maladie (PUMA) a été mise en place le 1er janvier 2016 afin d'assurer une prise en charge continue des frais de santé pour toute personne travaillant ou résidant de manière stable en France. Cette protection permet de maintenir l'affiliation à son régime d'assurance maladie même en cas de changement de situation ou de perte d'activité, évitant ainsi les interruptions de couverture.

Cependant, certains frais restent à la charge de l'assuré, tels que la part complémentaire, le forfait journalier en cas d'hospitalisation, la participation forfaitaire et les franchises médicales. Deux dispositifs ont été mis en place pour améliorer la prise en charge de ces dépenses de santé pour les personnes à faibles revenus : la couverture maladie universelle complémentaire (CMU-C) et l'aide à l'acquisition d'une assurance complémentaire santé (ACS).

En France, la CMU-C offre une couverture santé complète et gratuite aux résidents réguliers dont les revenus sont inférieurs à un certain seuil. Ce dispositif assure une prise en charge intégrale des dépenses de santé, sans nécessité d'avancer les frais.

L'ACS permet aux personnes dont les revenus sont légèrement supérieurs aux plafonds de la CMU-C de souscrire plus facilement à un contrat d'assurance complémentaire santé. Elle offre une aide financière d'une durée d'un an pour financer ce contrat. Cette aide est uniquement utilisable pour souscrire à l'un des contrats de complémentaire santé homologués par l'État en raison de leur bon

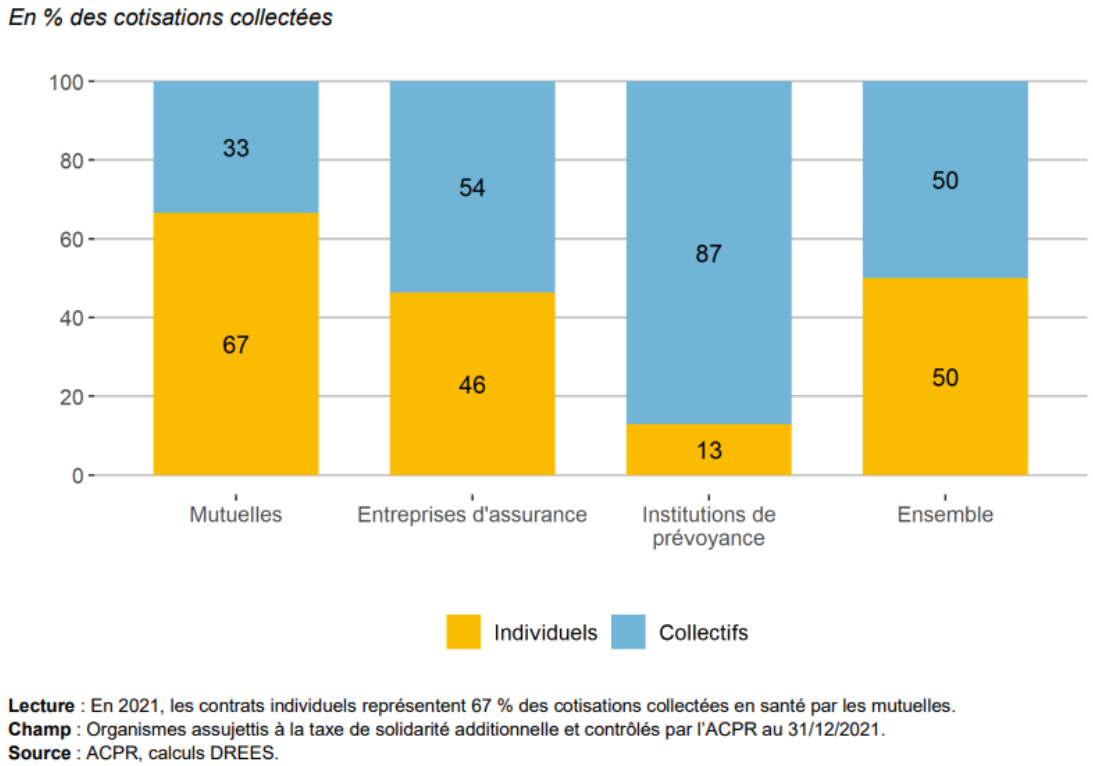


FIGURE 1.6 : Part des contrats individuels et collectifs dans l'ensemble des cotisations collectées en santé par les différents types d'organismes en 2021

Source : de WILLIENCOURT (2022)

rapport qualité-prix.

À partir du 1er novembre 2019, l'ACS et la CMU-C ont été regroupées en un seul dispositif appelé la Complémentaire santé solidaire (CSS ou C2S), géré selon le choix de l'assuré : soit par une caisse d'assurance maladie, soit par un organisme complémentaire (mutuelle, société d'assurance, institution de prévoyance) inscrit sur la liste des organismes complémentaires gestionnaires de la Complémentaire santé solidaire. La CSS constitue une assistance financière visant à prendre en charge les frais médicaux des individus à revenu modeste. Son accessibilité dépend de la situation et des ressources financières de chaque bénéficiaire. Grâce à la CSS, dans la plupart des cas, l'assuré n'a pas à payer les tickets modérateurs ni les participations forfaitaires, par exemple :

- Les consultations médicales, les soins dentaires, les services infirmiers, la kinésithérapie, l'hospitalisation, etc.
- Les médicaments délivrés en pharmacie.
- Les dispositifs médicaux tels que les pansements, les cannes ou les fauteuils roulants.
- La plupart des frais liés aux lunettes, aux prothèses dentaires ou auditives.

Étant donné que la CSS présente des caractéristiques similaires aux contrats complémentaires proposés par les assureurs, elle ne fait pas partie du champ d'étude de ce mémoire, car son financement et sa tarification relèvent de la responsabilité de la Sécurité Sociale.

1.4 Le principe de remboursement par le régime obligatoire de l'Assurance Maladie et la complémentaire santé

1.4.1 Cas général

Les remboursements sont réalisés en principe indemnitaire, ce qui signifie que le montant remboursé à l'assuré ne pourra jamais être supérieur à la dépense réelle. En ce qui concerne la Sécurité Sociale, une dépense (notée *Depense*) liée à une consommation médicale peut se décomposer en plusieurs éléments (voir figure 1.7) :

- La base de remboursement de la Sécurité Sociale (BR) est définie pour chaque acte médical consommé.
- Le dépassement correspond à la différence entre la dépense réelle et la base de remboursement.

En ce qui concerne la base de remboursement mentionnée précédemment, la Sécurité Sociale rembourse souvent moins que la totalité de la BR, avec un taux de remboursement de la base de remboursement (noté $Taux_{SS}$). Par exemple, pour les consultations/visites chez les médecins généralistes/spécialistes, le remboursement de la Sécurité Sociale est généralement de 70% de la BR. Ainsi, on obtient : $Remboursement_{SS} = BR \times Taux_{SS}$.

Il existe une partie du remboursement réel qui est déduite par la Sécurité Sociale, sauf dans certains cas particuliers comme pour les bénéficiaires de la CSS. Cette déduction se présente sous la forme d'une franchise (notée $Franchise$) ou d'une participation forfaitaire (généralement 1€ pour les actes de consultation ou 24€ pour les actes lourds dépassant 120€). Le montant réel remboursé sera donc : $Remboursement_{Reel_SS} = BR \times Taux_{SS} - Franchise$.

Le ticket modérateur est le montant restant à charge après réduction du remboursement réel de la Sécurité Sociale. En association avec les dépassements définis précédemment, il constitue le montant restant à charge remboursable par les organismes complémentaires (noté RAC_{OC}). Les participations forfaitaires ne sont jamais prises en charge par les contrats qualifiés de "responsables" (qui seront abordés plus tard dans ce chapitre). Ainsi, la partie remboursable après le remboursement de la Sécurité Sociale sera calculée comme suit : $RAC_{OC} = Depense - BR \times Taux_{SS}$ ou $RAC_{OC} = Depense - BR \times Taux_{SS} + Franchise$ dans le cas d'un contrat non responsable.

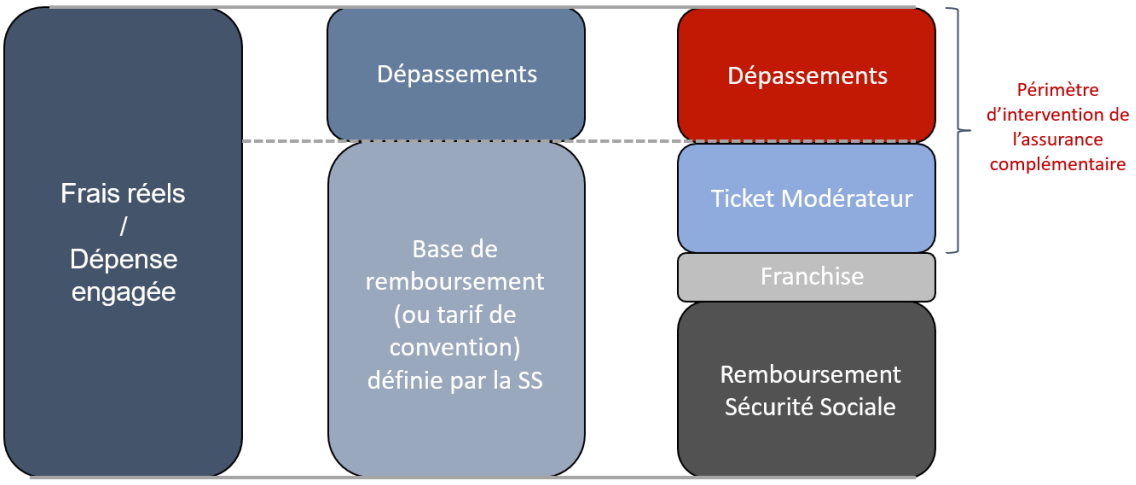


FIGURE 1.7 : Décomposition de la dépense médicale remboursable par la Sécurité Sociale

Les complémentaires santé définissent ensuite les conditions de remboursement variant selon les contrats. Elles remboursent une partie des restes à charge après remboursement de la Sécurité Sociale (noté ROC le montant remboursé réel par la complémentaire santé) jusqu'à un plafond défini par le contrat. On peut citer, par exemple, les définitions du plafond ainsi que le calcul du ROC dans le cas d'échéance à travers le tableau 1.1.

Définition du plafond du contrat	Formule pour calculer ROC
En pourcentage de la BR (noté $Taux_{plafond}$) défini par la Sécurité Sociale en incluant le remboursement de la Sécurité Sociale	$ROC = Min[(Taux_{plafond} - Taux_{SS}) \times BR; RAC_{OC}]$
En pourcentage de la BR (noté $Taux_{plafond}$) défini par la Sécurité Sociale en excluant le remboursement de la Sécurité Sociale	$ROC = Min[Taux_{plafond} \times BR; RAC_{OC}]$
En pourcentage du plafond (noté $Taux_{plafond}$) mensuel de la Sécurité Sociale (PMSS)	$ROC = Min[Taux_{plafond} \times PMSS; RAC_{OC}]$
En montant en € (noté $Plafond_{euro}$)	$ROC = Min[Plafond_{euro}; RAC_{OC}]$

TABLE 1.1 : Différentes définitions du plafond de remboursement et la formule pour ROC

Pour illustrer un exemple simple, un assuré bénéficie du régime obligatoire et d'une complémentaire santé minimum qui rembourse uniquement le ticket modérateur, soit un remboursement total de 100% de la BR y compris le remboursement de Sécurité Sociale. Cet assuré va chez un médecin généraliste dont le prix de la consultation est :

$$Depense = Prix_{Consultation} = 28€\,.$$

La Sécurité Sociale lui remboursera 70% de la base de remboursement ($BR = 25€$ pour une consultation médecin généraliste). En tenant compte de la participation forfaitaire de $1€$ non prise en charge par la Sécurité Sociale, ce qui est réellement remboursé par la Sécurité Sociale s'écrit donc :

$$Rembousement_{Reel\_SS} = BR \times Taux_{SS} - Franchise = 25 \times 70\% - 1 = 16.5€\,.$$

Si son contrat complémentaire santé est responsable, ce qui l'empêche le remboursement de la franchise $1e$, le reste à charge remboursable par la complémentaire santé est égale à :

$$RAC_{OC} = Depense - BR \times Taux_{SS} - 1 = 28 - 25 \times 70\% - 1 = 9.5€\,.$$

Sachant que le contrat prévoit de rembourser le ticket modérateur (donc 30% de la BR), l'assuré sera remboursé le montant :

$$ROC = Min[(Taux_{plafond} - Taux_{SS}) \times BR - 1; RAC_{OC}] = Min[(100\% - 70\%) \times 25 - 1; 9.5] = 6.5€\,.$$

Son reste à charge finale vaut :

$$RAC_{assure} = 28 - 16.5 - 6.5 = 5€\,.$$

1.4.2 Différents facteurs impactant le niveau de remboursement de la Sécurité Sociale

La Sécurité Sociale sépare les médecins en 3 secteurs différents : le secteur 1, le secteur 2 et le secteur 3. La différence entre ces secteurs affecte la base de remboursement de la Sécurité Sociale ainsi que le mode de tarification des médecins.

Pour les médecins du secteur 1, leurs tarifs doivent correspondre aux bases de remboursement fixées par la Sécurité Sociale. Ils ne peuvent pas pratiquer de dépassement d'honoraires, sauf dans certains cas (par exemple, en cas d'exigence des patients pour une consultation en dehors des horaires habituels).

Le secteur 2, également appelé "conventionné à honoraires libres", permet aux médecins de fixer librement le tarif de leurs consultations, dans des limites modérées. Les patients sont remboursés sur la base du tarif fixé par la convention médicale. Les mutuelles santé prennent en charge les dépassements d'honoraires selon un taux contractuellement défini. L'Optam et l'Optam-CO sont des options de pratique tarifaire introduites dans la convention médicale de 2016 pour améliorer l'accès aux soins :

- En adhérant à l'OPTAM, les médecins bénéficient de majorations et d'actes spécifiques, ainsi que d'une prime proportionnelle à leur activité.
- En adhérant à l'OPTAM-CO, les médecins spécialisés ont accès à des majorations et à des tarifs majorés pour leurs actes techniques, en modulant leurs dépassements d'honoraires. Les patients bénéficient d'un meilleur remboursement pour ces actes.

Les visites chez les médecins du secteur 1 ou du secteur 2 adhérant à OPTAM ou OPTAM-CO bénéficient de remboursements plus avantageux dans les contrats complémentaires santé ainsi que de la Sécurité Sociale. D'après La Sécurité Sociale (2022), le tarif de base de la Sécurité Sociale d'une consultation s'élève à 25€ pour les médecins généralistes du secteur 1 et un montant dépassement honoraire encadré à la charge de l'assuré pour les médécins du secteur 2 qui adhèrent au dispositif de pratique tarifaire maîtrisée (OPTAM). Pour les médecins du secteur 2 qui n'ont pas adhéré à ce dispositif, la base de remboursement diminue à 23€.

Le secteur 3, étant en dehors du système conventionnel, permet aux médecins de fixer librement leurs tarifs. Les bases de remboursement de l'Assurance Maladie pour les consultations de ces médecins sont très faibles (de 0,43€ à 0,61€ pour une consultation de médecine générale, de 0,85€ à 1,22€ pour une consultation chez un spécialiste).

Un autre facteur important qui impacte le niveau de remboursement des assurés est le respect du parcours de soin et la déclaration du médecin traitant. Pour un assuré/patient, le médecin traitant est celui qui le soigne régulièrement, l'oriente dans le parcours de soins, gère son dossier médical, assure une prévention personnalisée et établit le protocole de soins en cas d'affection de longue durée. Un parcours de soins respecté est lorsque l'assuré, en cas de besoin de soin, consulte d'abord son médecin déclaré (c'est-à-dire le médecin traitant) et éventuellement est orienté vers d'autres professionnels de santé. Un parcours de soins non respecté entraînerait, selon Assurance Maladie (2023a), une baisse de remboursement de l'Assurance Maladie (par exemple, de 70% à 30% de la base de remboursement pour une consultation chez un médecin généraliste n'étant pas le médecin traitant), alors que souvent la complémentaire santé considère que la Sécurité Sociale rembourse déjà le montant correspondant au respect du parcours de soins donc elle ne rembourse pas la majoration liée au non-respect de parcours de soin.

1.4.3 Le dispositif de Tiers-payant

Selon Assurance Maladie (2023b), le tiers payant est un système qui permet à un assuré de ne pas avancer les frais de santé lorsqu'il consulte un professionnel de santé. Au lieu de payer la totalité des frais médicaux au moment de la consultation, l'assuré présente sa carte d'assurance maladie ou éventuellement sa carte de complémentaire santé, et les frais sont directement pris en charge par la Sécurité Sociale et/ou l'organisme d'assurance complémentaire.

Le tiers payant peut être total ou partiel. Dans le tiers payant intégral, l'assuré n'a qu'à payer la partie dépassement du remboursement total s'il existe, tandis que dans le tiers payant partiel (sans la prise en charge de la complémentaire santé), l'assuré doit payer en plus la partie non prise en charge par

l'Assurance Maladie, telle que le ticket modérateur ou la franchise. Afin de généraliser le tiers payant en totalité, depuis le 1er janvier 2017, le tiers payant est obligatoire pour les assurés bénéficiant du CSS (Contrat de Solidarité et de Stabilité) ou les soins pris en charge au titre de la maternité ou d'une affection de longue durée (ALD). De plus, le tiers payant sur la part obligatoire peut également être proposé à tous les patients. Ainsi, le tiers payant facilite l'accès aux soins en évitant à l'assuré d'avancer les sommes d'argent nécessaires pour la consultation ou les dépenses liées à la pharmacie, aux examens médicaux, etc. Cela peut être particulièrement utile dans les situations où les frais de santé peuvent être élevés.

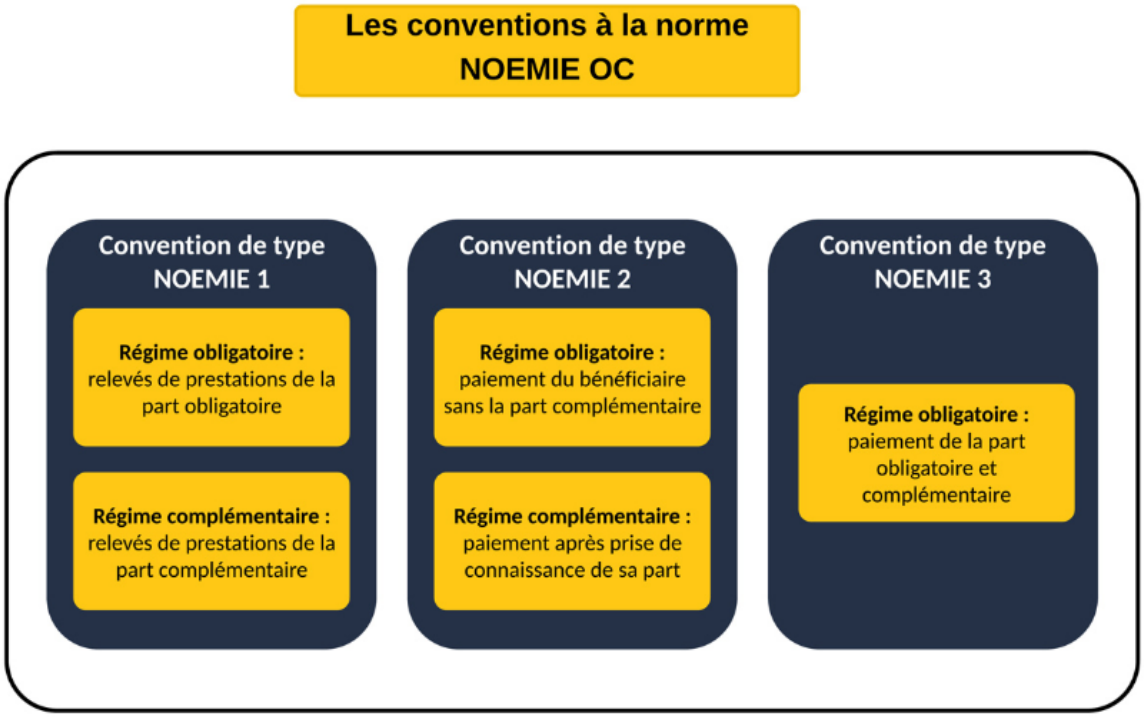


FIGURE 1.8 : Les différentes conventions NOEMIE pratiquées par la Sécurité Sociale et les organismes d'assurance complémentaire

Source : SIMPLICIA (2020)

L'accord sur la norme NOEMIE OC (Norme Ouverte d'Échange entre la Maladie et les Intervenants Extérieurs pour les organismes complémentaires) permet une communication efficace entre les caisses de l'Assurance Maladie Obligatoire (AMO) et les organismes complémentaires santé (AMC) en matière de remboursement des factures émises en tiers payant, par la codification et la transmission des actes médicaux, des paiements forfaitaires ou des régularisations de paiement. D'après SIMPLICIA (2020), il existe trois types de convention NOEMIE qui différencient la pratique de remboursement :

- NOEMIE 1 : La convention de type NOEMIE 1 prévoit que le régime obligatoire (CPAM) envoie uniquement les informations relatives au paiement de la part obligatoire. Ainsi, le relevé de prestations de la part obligatoire est pris en charge par le régime obligatoire, tandis que le relevé de prestations de la part complémentaire est à la charge du régime complémentaire.
- NOEMIE 2 : Dans la convention de type NOEMIE 2, le régime obligatoire effectue le paiement directement au bénéficiaire, sans prendre en compte la part complémentaire. L'organisme

complémentaire procède à son propre paiement une fois qu'il a connaissance de la part à sa charge.

- NOEMIE 3 : La convention de type NOEMIE 3 est la plus courante, car le régime obligatoire effectue immédiatement le paiement des deux parties, à la fois la part obligatoire et la part complémentaire.

1.5 Tarification en assurance IARD

1.5.1 Notion de risque - Segmentation de risque

L'assurance dommage, également connue sous le nom d'IARD (Incendie, Accidents et Risques Divers), est un secteur essentiel de l'assurance qui couvre les risques liés aux biens et aux responsabilités. Lorsqu'un assuré souscrit à un contrat d'assurance, il est indemnisé de manière forfaitaire ou indemnitaire en cas de sinistre, en échange du paiement d'une prime d'assurance. Pour assurer ces risques, les compagnies d'assurance établissent des primes en fonction de l'évaluation des risques encourus par les assurés. L'idée centrale de la tarification est que la prime demandée à chaque assuré représente en moyenne la valeur de son risque. Cette approche actuarielle permet aux assureurs de gérer les coûts des sinistres tout en offrant une protection adéquate aux assurés. Dans cette introduction, nous explorerons les mécanismes de tarification en assurance dommage, en mettant l'accent sur la notion de risque et son évaluation par les compagnies d'assurance pour déterminer les primes.

La notion de risque est fondamentale dans la tarification de l'assurance. Le risque représente la probabilité qu'un événement indésirable se produise, entraînant ainsi des charges pour l'assureur. Dans le contexte de l'assurance, les risques peuvent être liés à des accidents, des dommages matériels, des responsabilités civiles, etc.

Pour bien tarifier les contrats d'assurance, les compagnies d'assurance doivent segmenter les risques en profils de risque homogènes. Cela signifie regrouper les assurés ayant des caractéristiques similaires qui présentent des niveaux de risque comparables. En segmentant les risques de manière homogène, les assureurs peuvent évaluer plus précisément les probabilités de sinistres et déterminer des primes adéquates pour chaque groupe d'assurés.

En résumé, la tarification d'assurance repose sur une évaluation rigoureuse des risques et la segmentation de ces risques en profils homogènes. Cette approche permet aux assureurs de proposer des primes justes et équilibrées tout en garantissant la viabilité financière de leurs activités d'assurance.

En particulier, la tarification des contrats complémentaire santé fait partie du domaine de l'assurance IARD. Sa particularité sera détaillée plus tard.

1.5.2 Modèle collectif - Méthode de Coût x Fréquence

Modélisation du risque d'un contrat assurance

Examinons maintenant un contrat d'assurance qui garantit une indemnisation $\phi(B)$ à chaque fois qu'un sinistre causant un dommage financier B à l'assuré se produit (ϕ est une fonction mesurable de $\mathbb{R}_+ \to \mathbb{R}_+$ définie par le contrat d'assurance). L'approche de fréquence x sévérité nous permet d'écrire le sinistre total durant la vie du contrat comme :

$$X = \begin{cases} \sum_{k=1}^{N} B_k & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

Avec les hypothèses et notations :

- N est le nombre de sinistres produit pendant la période du contrat. N est une variable aléatoire discrète à valeur dans $\mathbb{N}$

- $(B_k)_k$ sont des variables aléatoires de même loi de B et indépendantes ($B_k \overset{\text{iid}}{\sim} B$) présentant les montants de dommage à chaque survenance du sinistre, $B_k \in \mathbb{R}_+ \forall k \in \{1,..,N\}$.
- La fréquence N étant indépendante de tous les montants de sinistre B_k

Sous ces hypothèses définies ci-dessus, X est une variable aléatoire dépendant de B et N. On peut calculer l'espérance de X à l'aide de l'espérance conditionnelle :

$$E[X] = E\left[\sum_{k=1}^{N} B_k\right] = E\left[E\left[\sum_{k=1}^{N} B_k \middle| N\right]\right] = E\left[\sum_{k=1}^{N} E[B_k|N]\right] = E\left[\sum_{k=1}^{N} E[B_k]\right] = E\left[\sum_{k=1}^{N} E[B]\right] = E[N]E[B].$$

Notons $B'_k = \phi(B_k), \forall k \in \{1,..,N\}$ et $B' = \phi(B)$ et supposons que ϕ est mesurable, on a :

$$(B'_k)_k \overset{\text{iid}}{\sim} B'_k \text{ et } N \perp B'_k, \forall k \in \{1,..,N\}$$

De la même façon, on arrive à calculer l'indemnité totale de l'assureur :

$$X' = \begin{cases} \sum_{k=1}^{N} \phi(B_k) & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

Si B' est intégrable et sa variance est finie, on aura $E[X'] = E[N]E[B']$.

Le principe de prime pure en assurance propose que l'assureur va charger l'assuré un montant égale à l'espérance de ses indemnités, c'est-à-dire l'indemnité moyenne pour ce contrat en contrepartie de ses indemnités en cas de sinistre. On note la prime pure $\pi = E[X']$.

Dans la suite des calculs, le sinistre que l'assureur porte est X' et non X.

Modélisation du risque d'un portefeuille d'assurance pour un type de risque

Considérons un assureur qui propose le même contrat d'assurance défini précédemment à n assurés venant souscrire chez lui. On se place dans le cas où l'assureur a bien segmenté la population de sorte que le risque de chaque personne souscrivant à ce contrat soit le même et qu'on suppose que chaque individu est indépendant (i.e $X'_i \overset{\text{iid}}{\sim} X', \forall i \in \{1,..,n\}$).

On note S'_n la charge globale sur tous les contrats soucrits. L'indemnité moyenne par contrat est calculée comme suit :

$$\bar{X}'_n = \frac{1}{n} S'_n = \frac{1}{n} \sum_{i=1}^{n} X'_i.$$

En tenant compte du caractère iid des risques des contrats, grâce à la loi des grands nombres, on obtient :

$$\bar{X}'_n \xrightarrow[n\to+\infty]{L^2} E[X'].$$

Ce qui veut dire que plus le nombre de personnes qui soucrit au contrat est grand, plus la prime moyenne que l'assureur doit faire payer à chaque assuré tend vers leur propre risque moyen, et qui est encore plus sûr lorsque la variance de $\bar{X}'_n$ tend vers zéro :

$$\text{Var}(\bar{X}'_n) = \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X'_i) = \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} X'_i) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X'_i] = \frac{1}{n}\text{Var}[X'_i] \xrightarrow[n\to+\infty]{} 0.$$

Cette caractéristique intéressante est due à l'effet de la mutualisation des risques individuels dans le portefeuille. Dans ce cas, on peut avoir $\bar{X}\prime_n \approx \pi$ et donc il est un estimateur sans biais de la prime individuelle π.

1.5.3 Chargement et taxe des contrats d'assurance

Notons bien que cette convergence de la prime moyenne par contrat n'entraîne pas la convergence totale des sinistres vers la prime totale sur l'ensemble du contrat. Autrement dit, on n'aura pas la convergence :

$$\sum_{i=1}^{n}[X_i'] \xrightarrow[n\to+\infty]{\mathbb{P}} nE[X'].$$

L'assureur ne peut donc pas être certain que les sinistres seront entièrement contrôlés. En prenant seulement l'espérance de son risque comme prime technique d'assurance, l'assureur a presque 50% de chance d'être en perte sur l'ensemble du portefeuille. Afin de réduire la probabilité d'être en ruine, il est nécessaire d'ajouter une marge de risque au-dessus de la prime actuarielle obtenue à partir du modèle de tarification. Pour plus de détails sur la démonstration, le lecteur pourrait préférer à Dutang (2021) qui formule correctement la nécessité de l'ajout de la marge de risque positive pour diminuer la probabilité de ruine lorsque le portefeuille est grand.

En dehors de cette marge de risque, des frais d'acquisition et de gestion de contrat, des rémunérations des agents intermédiaires ainsi que la taxe sur les contrats d'assurance s'ajoutent, et on obtient la prime commerciale, c'est-à-dire la prime réellement demandée à l'assuré. La prime commerciale est calculée comme suivant la figure 1.9.

Prime pure :
π

Prime pure
+ Marge technique ($\alpha\%$ ***de prime pure***) **:**
$\pi * \left(1 + \frac{\alpha}{100}\right)$

Prime pure
+ Marge technique
+ Chargement de gestion/d'acquisition/d'administration ($\beta\%$ ***de prime de commerce hors taxe***) **:**
$\pi * \frac{1+\frac{\alpha}{100}}{1-\frac{\beta}{100}}$

Prime pure
+ Marge technique
+ Charges de gestion/d'acquisition/d'administration
+ Charge fiscale ($\gamma\%$ ***de prime hors taxe***) **:**
$\pi * \frac{1+\frac{\alpha}{100}}{1-\frac{\beta}{100}} * (1 + \frac{\gamma}{100})$ **= Prime commerciale**

Figure 1.9 : Ajout des chargements et taxe sur le prime pure

1.5.4 La tarification sur un portefeuille existant

La tarification sur un portefeuille existant est une étape cruciale dans le domaine de l'assurance. Lorsqu'une compagnie d'assurance propose des contrats à ses assurés, elle doit évaluer régulièrement le risque associé à son portefeuille actuel. Cette évaluation permet de déterminer si les primes d'assurance actuelles sont suffisantes pour couvrir les risques et les coûts associés aux prestations à verser aux assurés. Une méthode couramment utilisée par les compagnies d'assurance pour effectuer cette évaluation est la formule de prime égale à la somme des sinistres divisée par l'exposition au risque :

$$\text{Prime moyenne} = \frac{\text{Montant total de sinistres passées}}{\text{Exposition totale au risque}}.$$

Pour mettre en œuvre cette formule, l'assureur commence par collecter et analyser les données des sinistres passés sur une période donnée. Les sinistres correspondent aux demandes de prestations reçues par les assurés et aux montants remboursés par l'assureur suite à ces demandes.

En parallèle, l'exposition au risque est évaluée en fonction du nombre d'assurés présents dans le portefeuille et de la durée pendant laquelle ils ont été couverts par l'assurance. Cette mesure de l'exposition au risque permet de prendre en compte le nombre d'assurés et la période pendant laquelle ils ont été exposés aux risques couverts par l'assurance.

Une fois les données des sinistres et de l'exposition au risque disponibles, l'assureur calcule la prime moyenne par assuré en divisant la somme des sinistres par l'exposition au risque. Cette prime moyenne est ensuite utilisée pour ajuster les primes individuelles des assurés dans le portefeuille.

Effectivement, la formule de prime présentée est en réalité une autre version de l'approche coût moyen x fréquence. En reprenant les expressions du coût moyen et de la fréquence, soit :

- Coût moyen $= \frac{\text{Montant total de sinistres passées}}{\text{Nombre de sinistres passées}}$.
- Fréquence $= \frac{\text{Nombre de sinistres passées}}{\text{Exposition totale au risque}}$.

Prenons la formule Prime pure = Coût moyen x Fréquence, on obtient :

$$\text{Prime pure} = \text{Coût moyen} \times \text{Fréquence} = \frac{\text{Montant total de sinistres passées}}{\text{Nombre de sinistres passées}} \times \frac{\text{Nombre de sinistres passées}}{\text{Exposition totale au risque}}.$$

Ce qui simplifie à :

$$\text{Prime pure} = \frac{\text{Montant total de sinistres passées}}{\text{Exposition totale au risque}}.$$

Ainsi, on constate que la formule de prime pure obtenue correspond effectivement à une version simplifiée de l'approche coût moyen x fréquence, où l'on calcule directement la prime moyenne par assuré en utilisant les montants totaux de sinistres passés et l'exposition totale au risque. Cette approche permet d'ajuster les primes en fonction des sinistres passés et de l'exposition au risque de l'ensemble du portefeuille, tout en prenant en compte les données réelles pour évaluer le niveau de risque et adapter les tarifs d'assurance en conséquence.

1.6 Tarification du produit complémentaire santé

Un contrat complémentaire santé est destiné à couvrir un large spectre de risques liés à la santé des assurés. Il est généralement conçu pour rembourser les frais médicaux payés par les assurés, mais peut également inclure des services ou dispositifs de prévention proposés par l'assureur pour améliorer l'accompagnement des assurés.

1.6.1 Famille de garantie d'un contrat complémentaire santé

Dans le cadre de l'illustration des méthodes de tarification en assurance santé, seuls les risques principaux couverts par les assureurs sont étudiés. D'après Dgccrf (Direction Générale De La Concurrence, De La Consommation Et De La Répression Des Fraudes) (sans date), un contrat en complémentaire santé donne accès aux postes généralement couverts, tels que :

- Frais d'hospitalisation médicale ou chirurgicale, actes et frais de chirurgie ;
- Forfait journalier hospitalier et, éventuellement, supplément en chambre particulière ;

- Consultations et visites de médecins généralistes ou spécialistes ;
- Frais pharmaceutiques ;
- Frais d'analyse et de laboratoire ;
- Actes de pratique médicale courante et actes d'auxiliaires médicaux ;
- Actes d'électroradiologie, de neuropsychiatrie, d'obstétrique ;
- Frais d'optique : montures, verres et lentilles ;
- Frais de soins et de prothèses dentaires ;
- Frais d'orthopédie et de prothèses.

Sachant que chaque poste de garantie présente un risque pour les assureurs, ces risques sont généralement modélisés séparément en supposant qu'ils ne soient pas corrélés, puis sommés au final pour obtenir une prime totale, ce qui n'est pas très réaliste en pratique. Un contrat complémentaire santé est entièrement caractérisé par une grille de garantie, dont l'indemnité de chaque poste de garantie est prévue par une modalité de remboursement sous forme de plafond ou limite de montant de remboursement. Selon chaque contrat complémentaire santé, les plafonds de remboursement peuvent être pris en compte en incluant ou en excluant les remboursements de la Sécurité Sociale sur le même acte médical. Dans notre cas, nous nous intéressons au plafond de remboursement en incluant les remboursements de la Sécurité Sociale. Par exemple :

$$\text{Consultation médecin généraliste : Plafond de remboursement de l'assureur} = 150\%BR,$$

dont :

- La base de remboursement pour cet acte s'élève à : $BR = 25€$;
- Et la part de remboursement de Sécurité Sociale s'élève à $70\%BR = 17.5€$;
- Ainsi, le Ticket Modérateur vaut : $Ticket\_moderateur = 30\%BR = 7.5€$;
- Le Dépassement remboursable est calculé par : $Depassement = 50\%BR = 12.5€$;
- Finalement, le montant maximum remboursable par l'assureur vaut : $Ticket\_moderateur + Depassement = 7.5 + 12.5 = 20€$;

1.6.2 Grille de garantie Benchmark

Pour cette étude, nous disposons d'une grille de garantie benchmark en interne du cabinet de conseil Forsides France, en figure 1.2, qui couvre trois types de contrats complémentaires santé modaux, basés sur les contrats les plus souscrits par les assurés. Cette grille est établie en se basant sur les distinctions effectuées dans Barlet et al. (2019). Les trois types de contrats inclus dans cette grille sont les suivants : les contrats à garanties hautes, offrant une couverture étendue et complète pour les soins médicaux, les contrats à garanties moyennes, proposant une couverture équilibrée, et les contrats à garanties minimales (panier de soin), qui couvrent les besoins essentiels en matière de santé. Cette grille de garantie Benchmark nous permet de comparer et d'évaluer différentes offres de contrats complémentaires santé.

Dans cette grille de garantie benchmark, on peut parfois trouver des plafonds de remboursement exprimés en euros (€) ou en pourcentage des frais réels (%FR). Cette approche est tout à fait cohérente avec la réalité, car il arrive parfois que la base de remboursement soit inférieure au montant réellement

		Minimum	Moyen	Maximum
SOINS COURANTS				
Consultations généralistes				
	Adhérent à l'OPTAM	100% BR	300% BR	400% BR
	Non adhérent à l'OPTAM	100% BR	275% BR	300% BR
Consultations spécialistes				
	Adhérent à l'OPTAM	100% BR	300% BR	400% BR
	Non adhérent à l'OPTAM	100% BR	275% BR	300% BR
Actes d'imagerie médicale				
	Adhérent à l'OPTAM	100% BR	300% BR	400% BR
	Non adhérent à l'OPTAM	100% BR	275% BR	300% BR
Actes techniques médicaux				
	Adhérent à l'OPTAM	100% BR	300% BR	400% BR
	Non adhérent à l'OPTAM	100% BR	275% BR	300% BR
Analyses et examens de laboratoire	Analyses médicales	100% BR	250% BR	300% BR
Honoraires paramedicaux	Auxiliaire médicaux	100% BR	250% BR	300% BR
Pharmacie remboursée par RO		100% BR	250% BR	300% BR
Materiel medical	Appareillage/orthopédie dépassement honoraire	100% BR	250% BR	300% BR
OPTIQUE				
Monture		x	250 €	350 €
Verres simples & monture		100% BR	300 €	400 €
Verres complexes ou très complexes & monture		100% BR	450 €	600 €
Lentilles acceptée par RO		47.38 €	250 €	400 €
Lentilles refusée par RO		x	300 €	350 €
Chirurgie refractive		x	300 €	400 €
DENTAIRE				
Soins		100% BR	275% BR	350% BR
Inlay-Onlay		100% BR	350% BR	450% BR
Parodontologie		x	150 €	300 €
Prothèses		100% BR	350% BR	450% BR
Orthodontie		100% BR	200% BR	300% BR
Implantologie		x	250€ / implant (max 3 par an)	450€ / implant (max 3 par an)
AIDES AUDITIVES				
Appareil auditif remboursé par RO		700€ / 2 oreilles	1400€ / 2 oreilles	1800€ / 2 oreilles
HOSPITALISATION				
Forfait journalier		100% FR	100% FR	100% FR
Chambre particulière en				
	Hospitalisation	x	170€ / jour	200€ / jour
	Psychiatrie	x	x	x
Frais de séjour				
	Adhérent à l'OPTAM	100% FR	100% FR	100% FR
	Non adhérent à l'OPTAM	100% FR	100% FR	100% FR
Honoraires				
	Adhérent à l'OPTAM	100% BR	300% BR	400% BR
	Non adhérent à l'OPTAM	100% BR	275% BR	300% BR
Frais d'accompagnement	Enfants	x	70€ / jour	100€ / jour
Transport		100% BR	200% BR	300% BR
AUTRES				
Cure thermale		100% BR	200% BR	300% BR
Médecine douce		30€ / an	150€ / an	200€ / an
Soins à l'étranger	Remboursés par le RO	100% BR	200% BR	300% BR

Contrats responsables : le régime complémentaire rembourse au minimum le TM de tous les actes remboursés par le régime obligatoire.

TABLE 1.2 : Grille de garantie benchmark disponible pour trois types de couverture ordonnées (Source : Forsides France)

payé (par exemple, pour les verres ou montures en dehors du panier 100% Santé, la base de remboursement peut être inférieure à 5€ alors que les frais réels s'élèvent à 200-300€). Notre grille de garantie indique ainsi que les remboursements en pourcentage de la base de remboursement (BR) ou en pourcentage des frais réels (FR) incluent également la part de la Sécurité Sociale. En revanche, les remboursements en euros viennent en complément du remboursement de la Sécurité Sociale, c'est-à-dire que ces remboursements ne déduisent pas la part déjà remboursée par la Sécurité Sociale.

1.6.3 Contrat responsable - Contraintes de tarification

Selon les informations disponibles concernant les contrats responsables dans Dgccrf (Direction Générale De La Concurrence, De La Consommation Et De La Répression Des Fraudes) (sans date), la majorité des contrats proposés sur le marché sont qualifiés de "responsables" car ils répondent à des conditions de prise en charge définies par la réglementation. Ces contrats offrent diverses prestations considérées comme "responsables", visant à encourager les assurés à suivre un parcours de soins coordonnés et à limiter les dépenses de santé publique.

Les prestations d'un contrat "responsable" incluent :

- Le remboursement à 100% de la base de remboursement de la Sécurité Sociale pour les soins courants, tels que les consultations médicales et les médicaments dont le service médical rendu est majeur et remboursés à 65% ;
- Le remboursement intégral du forfait journalier hospitalier sans limitation de durée ;
- Le remboursement à 100% de la base de remboursement de la Sécurité Sociale pour les soins dentaires courants, comme les consultations, les détartrages ou les traitements de caries ;
- Le remboursement à 100% de la base de remboursement de la Sécurité Sociale pour les frais optiques. Pour les frais optiques dépassant le tarif conventionnel, certaines complémentaires santé peuvent proposer une prise en charge limitée, notamment une paire de lunettes tous les 2 ans au maximum, des montures jusqu'à 150 €, et des limites minimales et maximales selon la complexité de l'équipement ;

En revanche, les contrats responsables n'incluent pas les éléments suivants dans leur prise en charge :

- La participation forfaitaire de 1€ pour chaque acte de consultation réalisé par un médecin de ville, dans un établissement ou centre hospitalier, dans la limite de 50 € par an et par personne ;
- Les franchises médicales laissées à la charge de l'assuré pour les médicaments et les transports sanitaires, plafonnées à 50 € par an et par personne ;
- La majoration de la participation de l'assuré pour le non-respect du parcours de soins coordonnés ;
- Les dépassements d'honoraires lorsque l'assuré consulte un spécialiste sans passer par un médecin traitant désigné, comme prévu par la loi ;

Ces contrats responsables présentent des avantages fiscaux et sociaux, notamment une exonération des charges sociales pour les cotisations versées par l'employeur dans la limite d'un plafond spécifique. Ils offrent également une fiscalité avantageuse pour les travailleurs non-salariés et permettent aux travailleurs salariés de déduire la part salariale de leur impôt sur le revenu. Il est important de noter que la taxe sur les cotisations est réduite à 13,27% pour les contrats responsables, comparativement à 20,27% pour les contrats non responsables.

En outre, il existe également des contrats dits "solidaires" qui ne demandent pas d'informations médicales à la souscription et n'ajustent pas leurs tarifs en fonction de l'état de santé de l'assuré.

En résumé, le contrat complémentaire santé responsable représente une option avantageuse pour les assurés et les employeurs, en proposant une couverture adéquate et des avantages fiscaux, tout en encourageant les assurés à suivre un parcours de soins coordonnés.

1.7 Problèmes assurantiels liés au comportement des assurés

L'homme étant un individu conscient, cette conscience peut créer un biais psychologique lorsqu'un individu est confronté à un choix risqué. Un individu est dit avoir une aversion au risque s'il n'apprécie pas les résultats du hasard et cherche à éviter ce hasard ou à en réduire les conséquences, même au détriment partiel de son bien-être. Cela étant dit, la demande d'assurance repose en grande partie sur cette caractéristique humaine, car l'assurance vise à réduire et à éviter les conséquences financières auxquelles sont exposés les assurés.

Les fondements théoriques de la tarification présentés ci-dessus semblent solides et donnent l'impression de bien fonctionner sur le marché de l'assurance. Cependant, en réalité, le marché souffre du problème de l'asymétrie de l'information entre l'assuré et l'assureur, en partie influencé par les réglementations et les lois en place. En cas d'information parfaite, les assureurs connaissent parfaitement le niveau de risque de chaque individu assuré et peuvent tarifer correctement leur prime de risque. En l'absence d'information, certains phénomènes d'asymétrie de l'information peuvent survenir, en faveur des assurés ou des assureurs. Trois phénomènes d'asymétrie de l'information peuvent être observés sur le marché de l'assurance : l'anti-sélection, l'aléa moral et la sélection avantageuse.

1.7.1 Problème d'aléa moral

Ce phénomène, appelé ex-post, découle du fait que les assureurs n'ont pas d'informations sur le comportement des assurés après la souscription. Il désigne la tendance des assurés à adopter un comportement plus risqué lorsqu'ils sont couverts par l'assurance que lorsqu'ils ne le sont pas. Un tel comportement peut obliger les assureurs à augmenter les primes d'année en année pour compenser l'augmentation de la sinistralité, ce qui peut inciter les assurés à faible risque à résilier leur contrat. Le principe de mutualisation est alors mis à mal, et la tarification continue d'être sous-évaluée.

Pour remédier à ce problème, il existe certaines solutions, comme la mise en œuvre de la tarification sur la base des sinistres passés de l'assuré (bonus-malus en assurance auto) ou la mise en place de franchises d'assurance pour dissuader les assurés de prendre des risques excessifs. En assurance santé, il n'est généralement pas possible d'appliquer la tarification en fonction des sinistres passés, car la loi interdit l'utilisation des dépenses en santé du passé de l'assuré comme critère de tarification. Les franchises, comme celle de 1€ pour les consultations instaurée par la Sécurité Sociale, sont un exemple des efforts déployés pour réduire la surconsommation des soins.

Selon Albouy et Crepon (2007), le marché de l'assurance santé complémentaire en France est particulièrement sujet à ce problème, notamment sur le marché des contrats collectifs. En effet, les contrats collectifs sont négociés par l'employeur, et si les garanties du contrat sont très élevées et que les assurés sont bien remboursés, les assurés des contrats collectifs ont tendance à consommer davantage pour profiter de la couverture qui leur est offerte involontairement.

1.7.2 Problème d'anti-sélection et segmentation de la population

Ce phénomène, qui est souvent décrit dans Rothschild et Stiglitz (1976), se manifeste lorsque les assurés possèdent des informations sur leur propre risque et décident d'utiliser cet avantage pour souscrire à une assurance à un tarif inférieur à leur prime de risque réelle, ou décident de ne pas souscrire s'ils jugent la prime proposée trop élevée. Ce comportement peut conduire à des hausses de

sinistralité et obliger les assureurs à augmenter les primes, ce qui peut entraîner une spirale de hausse des primes jusqu'à l'effondrement du marché.

Il existe de nombreuses solutions pour contrer ce comportement stratégique des assurés, telles que l'introduction de délais de carence, la segmentation de la population et la segmentation de l'offre. Les délais de carence sont des clauses des contrats d'assurance qui définissent une période, entre la date de souscription du contrat et l'activation de la couverture, pendant laquelle les sinistres déclarés ne sont pas remboursés. Il s'agit d'un moyen très efficace pour l'assurance contre les punaises de lit, car il empêche les individus déjà infestés de souscrire à un contrat et de déclarer immédiatement un sinistre.

La segmentation des niveaux de contrat selon Rothschild et Stiglitz (1976) est un moyen efficace, car elle permet aux assurés de choisir le niveau de couverture qui correspond le mieux à leurs besoins. Cela évite l'anti-sélection, car les assurés ayant différents niveaux de risque et de volonté de payer choisissent le niveau de couverture qui leur convient. On parle dans ce cas de phénomène d'auto-sélection.

Selon Denuit et Charpentier (2005), la segmentation des assurés est une pratique essentielle dans le domaine de l'assurance. En répartissant les assurés en groupes homogènes en fonction de leurs caractéristiques de risque, les assureurs peuvent proposer des primes plus précises et équitables. Cela permet de mieux refléter le risque individuel de chaque assuré, évitant ainsi l'anti-sélection, où les assurés les plus risqués cherchent à s'assurer en masse, ce qui pourrait déséquilibrer les coûts pour les assureurs. En offrant des primes différenciées en fonction des caractéristiques de risque, l'assureur peut attirer un large éventail de clients tout en maintenant un équilibre financier. Cependant, il est crucial de veiller à ce que cette segmentation soit basée sur des critères pertinents et éthiques pour garantir une tarification équitable et éviter toute discrimination injuste. Sur le marché de l'assurance santé facultatif offrant différents choix, Marquis (1992) démontre que l'anti-sélection est réduite lorsque l'assureur propose des primes variables en fonction de facteurs démographiques, alors qu'un seul tarif par niveau de couverture est très susceptible d'entraîner une anti-sélection assez forte pour éradiquer les contrats de haut niveau de couverture.

En assurance santé, sur le marché de la complémentaire santé individuelle, on observe clairement ce cas en raison de la loi qui empêche les assureurs de poser des questions concernant l'état de santé de l'assuré, ainsi que l'interdiction de l'utilisation des tests génétiques. L'assureur est donc en obligation d'accepter l'assuré qui a une meilleure connaissance de sa santé que l'assureur lui-même et qui choisit les options très bien remboursées, perçues comme avantageuses et moins chères pour eux. Selon la littérature, l'anti-sélection est beaucoup plus présente sur le marché individuel que sur le marché collectif. La comparaison effectuée par Genier (1998) entre la consommation de soins des personnes couvertes par une assurance complémentaire à adhésion obligatoire et celles qui ont choisi d'adhérer volontairement à une telle assurance s'avère plus convaincante pour établir la présence d'anti-sélection. Ces deux groupes de la population sont tous assurés complémentairement, ce qui signifie que leur utilisation des services de santé peut être influencée par l'effet potentiel de l'aléa moral. Cependant, la principale différence réside dans le fait que les premiers ne sont a priori pas choisis en fonction de leur niveau de risque (car leur adhésion est imposée), tandis que les seconds le sont peut-être de manière volontaire. Par conséquent, c'est la possible auto-sélection qui pourrait expliquer les variations dans l'accès aux soins médicaux entre ces deux groupes.

1.7.3 Sélection avantageuse

La sélection avantageuse est un phénomène en faveur des assureurs, car elle se produit lorsque les assurés à faible risque cherchent à se couvrir davantage. Selon Einav et Finkelstein (2011) : "La sélection avantageuse survient lorsque les personnes prêtes à payer le plus cher pour une assurance sont celles qui sont les plus prudentes face au risque (et ont donc le coût attendu le plus bas). En effet, il est naturel de penser que dans de nombreux cas, les individus qui accordent une grande importance à l'assurance sont également susceptibles de prendre des mesures pour réduire leurs coûts

attendus : conduire plus prudemment, investir dans des soins préventifs de santé, etc.". En fonction de la proportion de cette population dans la population globale, ce phénomène peut neutraliser l'effet de l'anti-sélection sur les contrats de haut niveau de couverture, car les assurés à faible risque se mutualisent avec ceux à haut risque et maintiennent un niveau de sinistralité moyen moins élevé que s'il n'y avait pas ce phénomène.

Chapitre 2

Tarification avec Open Data : Cas de la base Open Damir

2.1 Open Data en santé et la base Open Damir

2.1.1 Système de données publiques en France

D'après la présentation générale sur le site officiel de l'Assurance Maladie, le système national des données de santé (SNDS) a été créé en France par la loi de modernisation du système de santé en 2016. Il vise à développer l'utilisation des données de santé pour analyser et améliorer la santé de la population.

Selon Assurance Maladie (2023e),le SNDS est basé sur le Système national d'information interrégimes de l'Assurance Maladie (Sniiram), qui regroupe les données de remboursement de l'Assurance Maladie et les données des hôpitaux. Il intègre également de nouvelles composantes telles que les données sur les causes médicales de décès, les données relatives au handicap et les données liées à la Covid-19 extraites des bases Vaccin Covid et SI-DEP.

Les composantes principales du SNDS sont mises à disposition par la Caisse nationale de l'Assurance Maladie (Cnam) sur son portail, ces produits de restitution par objectifs sont présentés dans la figure 2.1. Le SNDS permet d'avoir une vision complète du parcours de soins de l'ensemble de la population sur une période maximale de 20 ans.

Les différentes bases de données qui composent le SNDS sont les suivantes :

- Le Sniiram : Il regroupe les informations pseudonymisées sur les remboursements effectués par l'Assurance Maladie pour les soins du secteur libéral. Il contient des données sur les bénéficiaires, les consommations de soins, les médicaments délivrés, les actes médicaux, etc.
- Le PMSI (Programme de médicalisation des systèmes d'information) : Il permet d'analyser l'activité médicale des établissements hospitaliers. Il recueille des informations administratives et médicales sur les séjours hospitaliers, les diagnostics, les actes médicaux, etc.
- Les bases de données sur les causes médicales de décès : Elles sont utilisées pour établir la statistique nationale des causes médicales de décès en France. Elles contiennent des informations sur les maladies ou affections à l'origine du décès.
- Les données relatives au handicap : Elles proviennent des maisons départementales des personnes handicapées et contiennent des informations sur les demandes d'accompagnement formulées pour les personnes handicapées.
- Les données issues des bases Vaccin Covid et SI-DEP : Elles comprennent des informations sur la vaccination contre la Covid-19 et les dépistages réalisés.

Le SNDS est continuellement enrichi avec de nouvelles bases de données, telles que l'enquête statistique EpiCov sur l'état de santé lié à l'épidémie de Covid-19, la base HEPATHER sur les patients atteints d'hépatite B ou C, et la base OSCOUR sur les données de passage aux urgences.

L'accès aux données du SNDS est réglementé et réservé à certains organismes ayant des missions d'intérêt public, telles que les organismes de recherche, les agences de santé, les autorités publiques, etc. Ces organismes doivent respecter des règles strictes de confidentialité et de protection des données.

FIGURE 2.1 : Les bases de données en santé publique produites par SNDS

Source : SNDS (SYSTÈME NATIONAL DES DONNÉES DE SANTÉ) (sans date)

2.1.2 Présentation de la base Damir

La mise en place de l'Open Data

Depuis 2009, des bases de données Open Data ont été créées pour surveiller les dépenses d'assurance maladie. Ces bases de données se déclinent en cinq séries distinctes : la base complète Open Damir, qui couvre l'ensemble des régimes d'assurance maladie, les données nationales sur les dépenses du régime général de l'Assurance Maladie pour les soins de ville, les dépenses du régime général de l'Assurance Maladie pour les établissements de santé privés, ainsi que les données spécifiques par CPAM sur les dépenses du régime général de l'Assurance Maladie pour les soins de ville et les établissements de santé privés.

Le principal défi pour rendre les bases Open Data accessibles au public réside dans la confidentialité des données, conformément à l'Article 47 de la loi Santé. Cet article stipule que "Les données du

système national des données de santé qui sont mises à disposition du public sont traitées de manière à prendre la forme de statistiques agrégées ou de données individuelles structurées de manière à rendre impossible l'identification directe ou indirecte des personnes concernées".

Dans le cadre d'une réflexion sur l'ouverture en open data de la base de données Damir (Caillol (2015)), la CnamTS a organisé, en collaboration avec la mission Etalab, un hackathon le 26 janvier 2015. L'objectif de cet événement était d'expérimenter la mise à disposition et l'exploitation de données anonymes du Sniiram (Système National d'Information Inter-Régimes de l'Assurance Maladie) par des acteurs issus de divers secteurs du domaine de la santé. La base Damir est accessible sur le portail Sniiram aux utilisateurs habilités, et son principal objectif est de permettre une analyse macro-économique de l'évolution des dépenses d'assurance maladie. Plus précisément, la base Open Damir est constituée d'une série mensuelle de dépenses, incluant les remboursements de soins, les indemnités journalières, les aides financières, etc., à partir de janvier 2009, avec une publication régulière de nouvelles données chaque année. Les données de la base Open Damir sont disponibles sur le site de l'Assurance Maladie (voir Assurance Maladie (2024)).

Les caractéristiques de la base Damir

Les dépenses dans la base Damir sont caractérisées par 42 variables qui portent sur les informations du patient, du professionnel de santé, de l'établissement de soins et de la caisse de Sécurité Sociale. En plus de ces 42 variables explicatives, 13 indicateurs ont été créés pour mesurer le montant des dépenses, et ils sont classés en indicateurs préfiltrés et indicateurs bruts en fonction de leur utilisation :

- Les indicateurs préfiltrés permettent de différencier les dépenses liées aux remboursements du régime obligatoire et aux remboursements complémentaires pour les mêmes soins provenant des régimes spéciaux, de la CSS (ex CMU-C), etc.
- Les indicateurs bruts sont les mêmes pour tous les remboursements liés aux mêmes soins.

Ces variables sont classifiées en 2 groupes d'indicateurs, à savoir les "Indicateurs bruts" et les "Indicateurs préfiltrés", et réparties en 6 axes d'analyse, à savoir :

- Période
- Prestation
- Organisme de prise en charge
- Bénéficiaire des soins
- Professionnel de santé exécutant
- Professionnel de santé prescripteur

Elles sont détaillées dans le tableau des variables présenté dans le tableau A.1 en annexe.

Depuis 2009, certains changements ont eu lieu dans la codification des variables de la base Damir, notamment :

- Avant 2015, les régions étaient codifiées sous forme de zones d'études et d'aménagement du territoire (ZEAT) avec 9 zones (variable BEN_RES_ZEAT). Depuis 2015, elles sont codifiées en grande région avec 19 régions (variable BEN_RES_REG) (voir figure A.2 en annexe).
- Depuis avril 2021, le CMU-C a été remplacé par le CSS dans la base Damir. Bien que le nom des variables soient resté inchangé (BEN_CMU_TOP,...), cela n'a pas de profonds changements pour la compréhension de la base.

- Chaque année, de nouveaux actes sont introduits ou retirés dans la variable "Nature de prestation" (PRS_NAT) de la base Damir, telle que des actes liés à la vaccination contre la Covid-19 (base 2021).

Pour une ligne de prestation donnée dans la base Open Damir, les indicateurs de dépense ne correspondent pas à la prestation d'un seul individu, mais à l'ensemble des prestations des individus ayant les mêmes variables caractéristiques (la même valeur prise par toutes les variables des 6 axes, y compris la période de soins et les autres modalités de prise en charge). La base Open Damir est agrégée, ce qui signifie qu'il n'est pas possible d'identifier les dépenses au niveau individuel.

De plus, il convient de noter que la base Damir ne comprend que les prestations payées par la Sécurité Sociale, à l'exception du dispositif CSS (Complémentaire Santé Solidaire). Il est important de souligner que les dépenses liées aux hospitalisations dans les établissements publics ne sont pas incluses dans la base, car la plupart de ces prestations ne sont pas fournies par la CNAM (Caisse Nationale d'Assurance Maladie).

Pour le but de la tarification basée sur la base Damir, nous avons décidé de traiter les données de l'année 2021, car il s'agit des données les plus récentes au moment de la rédaction de ce mémoire. De plus, nous ne faisons pas de distinction entre la CMU-C et la CSS, et avons choisi la CSS pour les représenter toutes les deux dans la suite du mémoire.

2.1.3 Statistique globale de Open Damir et le traitement des données

Méthodologie générale

La base de données Damir constitue une ressource précieuse pour l'analyse des dépenses de santé. Cependant, pour mener une étude rigoureuse et pertinente, il est crucial de mettre en place une méthodologie adéquate. Cette introduction présente les principaux éléments de cette méthodologie, en se focalisant sur la délimitation des données pertinentes, le traitement des données aberrantes et l'interprétation des résultats.

Dans un premier temps, il est essentiel de définir la périphérie des données d'intérêt, en se concentrant spécifiquement sur les dépenses de santé et en excluant les informations relatives à la prévoyance ou aux aides gouvernementales. Cette délimitation préalable des variables pertinentes permet de cibler l'étude sur les aspects spécifiques des dépenses de santé, tout en minimisant la complexité des données. Ainsi, une sélection judicieuse des variables clés à analyser facilite l'obtention de résultats significatifs.

Ensuite, le traitement des données aberrantes constitue une étape cruciale pour garantir l'intégrité des résultats. Les valeurs aberrantes, qui sont des observations extrêmes ou incohérentes, peuvent fausser l'analyse et conduire à des conclusions erronées. Par conséquent, des méthodes statistiques appropriées doivent être appliquées pour détecter et traiter ces valeurs aberrantes. Cette approche permet d'assurer la fiabilité des résultats en excluant les données inexploitables et en préservant la qualité de l'analyse.

Enfin, lors de l'interprétation des données, il est essentiel de comprendre comment les modalités des variables sont codées et comment elles sont liées au remboursement du régime général de santé. Une attention particulière est accordée à l'identification des données associées aux remboursements des régimes spéciaux et de la Caisse de Sécurité Sociale (CSS). Cette analyse permet de déceler les schémas et les tendances spécifiques à chaque régime de remboursement, offrant ainsi une compréhension approfondie de l'impact de ces variables sur les dépenses de santé.

En résumé, cette méthodologie de traitement de la base de données Damir repose sur la délimitation préalable des données pertinentes, le traitement rigoureux des données aberrantes et l'interprétation minutieuse des modalités des variables en lien avec les remboursements du régime général de santé. Cette approche méthodologique solide permet d'obtenir des résultats fiables et d'effectuer une analyse approfondie des dépenses de santé dans la base de données Damir, contribuant ainsi à une meilleure compréhension des enjeux liés à ces dépenses.

Variables socio-économiques et la périphérie de l'étude

Pour une prestation donnée, les bénéficiaires des soins sont caractérisés par un maillage de l'âge, du sexe, de la région de résidence ainsi que les sous-catégories comme la qualité de bénéficiaire ou le statut de bénéficiaire de CSS(les tableaux de modalités sont présentés dans la table A.3 en annexe). La figure 2.1 présente un exemple de ligne simplifiée de la base Damir.

Variables de la base Damir	Exempe ligne 1	Exempe ligne 2
Année et mois de remboursement	2021-07	2021-11
Tranche d'âge du bénéficiare	30-39 ans	0-20 ans
Région de résidence du bénéficiaire	Île de France	Normandie
Sexe du bénéficiaire	Homme	Femme
Qualité du bénéficiaire	Assuré	Enfant d'assuré
Bénéficiaire CSS	Non	Oui
Année de soins	2021	2021
Mois de soins	4	5
Nature de prestation	CONSULTATION MEDECINE GENERALE	CONSULTATION MEDECINE GENERALE
Type de remboursement	TICKET MODERATEUR CSS	PRESTATION DE REFERENCE
Quantité d'acte	2	3
Montant de la dépense	50 €	75 €
Base de remboursement	50 €	75 €
Taux de remboursement	30%	70%
Montant remboursé	15 €	52.50 €

TABLE 2.1 : Exemple de lignes simplifiées dans la base Damir

Parmi les prestations présentes dans la base Damir, certaines données concernent des prestations en dehors du régime général, par exemple :

- Les prestations liées aux accidents du travail (AT) ou aux maladies professionnelles (MP).
- Les indemnités journalières ou les indemnités de congé maternité/paternité.
- Les prestations issues du niveau de remboursement le plus élevé du régime local de l'Alsace-Moselle.
- Le dispositif CSS (anciennement CMU-C avant avril 2021) rembourse les tickets modérateurs ou les dépassements aux assurés ou aux professionnels de santé via le tiers payant, et ces remboursements sont codés dans la base Damir.
- Les prestations liées aux dépenses des bénéficiaires affectés par les maladies ALD (Affections de Longue Durée) et donc remboursées totalement par le régime général.
- Les aides de l'État (aide au logement, aide financière en cas d'accident, etc.) et les régimes spéciaux (régime de protection sociale pour les notaires, régime des travailleurs SNCF, etc.).

Notre étude sur la base Open Damir en 2021 est restreinte aux assurés du régime général (non CSS, non-régime local Alsace-Moselle), aux dépenses et remboursements du régime de base sans supplément, avec le statut de bénéficiaire légitime (assuré, son conjoint et ses enfants). La section de prétraitement des données en annexe A.2 consiste à rendre les données conformes à ce cadre d'étude.

2.2 Tarification sur la base Open Damir

Depuis sa mise à disposition du public, la base Open Damir est devenue de plus en plus populaire dans le domaine de l'assurance santé, ce qui se manifeste par le nombre croissant de mémoires rédigés par des actuaires qui étudient cette base, notamment Fotsing (2018), Huynh (2021), et Yerle (2020). La plupart de ces mémoires explorent la base Damir en tant que source de sinistres pour pratiquer la tarification des contrats complémentaires santé.

Dans le but d'illustrer la tarification à partir de l'Open Data, nous suivons les étapes définies par Huynh (2021) pour élaborer des tarifs pour nos contrats complémentaires santé avec différents niveaux de couverture. L'application de la grille de garantie disponible dans la base Open Damir revêt une importance capitale lorsqu'il s'agit de tarifer les contrats complémentaires santé. Cette grille permet aux assureurs d'évaluer et de comparer les différentes offres de contrats en fonction des garanties et prestations spécifiques proposées par la Sécurité Sociale. En effet, la Base Open Damir fournit une source fiable et exhaustive d'informations sur les remboursements effectués par la Sécurité Sociale, ce qui permet aux assureurs de prendre des décisions éclairées lors de la conception de contrats d'assurance parfaitement adaptés aux besoins de leurs assurés. Dans cette section, nous explorerons en détail comment l'application de la grille de garantie Benchmark fournit une base solide pour le processus de tarification des contrats complémentaires santé, en se basant exclusivement sur les remboursements effectués par la Sécurité Sociale.

2.2.1 Classification des actes de la base Damir

Répartition des actes médicaux en famille de poste de garantie

Pour appliquer efficacement les formules de tarification à la base Damir, il est essentiel de classer les actes par famille de garantie. Cette classification permettra une analyse précise et cohérente des données, garantissant ainsi des résultats fiables et pertinents lors de l'évaluation des contrats complémentaires santé. Chaque prestation de la Sécurité Sociale est identifiable grâce à son nom d'acte (Nature prestation "PRS_NAT") dans la base Damir. Certaines prestations peuvent être codées en utilisant la Classification Commune des Actes Médicaux (CCAM) ou, de manière encore plus optimale, en utilisant la norme NOEMIE OC, comme expliqué dans les parties précédentes. Le recours à la norme NOEMIE OC facilite grandement la classification des actes des prestations dans la base Damir, tant pour les différents postes de garantie que pour les remboursements aux assurés.

En particulier, selon le mémoire de Huynh (2021) sur la tarification des contrats complémentaires santé avec la base Damir, tous les actes de la variable "PRS_NAT" sont correctement classés dans leur famille de garantie, correspondant aux familles de garanties courantes dans les contrats complémentaires santé, y compris notre grille de garantie. Pour cette raison, nous avons décidé de réutiliser sa classification en effectuant une validation supplémentaire sur les actes déjà classifiés, ainsi qu'une classification pour les nouveaux actes non classifiés.

Adaptation de la grille de garantie benchmark

Un processus de synchronisation minutieux a été entrepris pour aligner deux grilles de garantie distinctes, comme illustré dans la figure 2.2. Il s'agit de la grille issue du mémoire de Huynh (2021) et de celle que nous avons à notre disposition, présentée au chapitre 1. L'objectif est d'associer les plafonds de garantie aux différents niveaux de couverture des familles de garantie pour chaque acte. Pour ce faire, nos familles et sous-familles de garantie devaient se rapprocher autant que possible de la classification de Huynh (2021). Ce processus a été réalisé en accordant une attention particulière aux actes liés à l'offre 100% Santé en optique, dentaire et auditif, tout en tenant compte des contraintes imposées par les contrats responsables.

En particulier, en ce qui concerne les actes qui entraînent un différentiel de remboursement en

ID	Plafond couverture minimum	Type de remboursement	Plafond couverture moyen	Type de remboursement	Plafond couverture maximum	Type de remboursement	Nom du poste de garantie
1	100	BR	276.4016052	BR	305.6064207	BR	Consultations généralistes/Consultations spécialistes
2	100	BR	276.4016052	BR	305.6064207	BR	Actes techniques médicaux
4	100	BR	250	BR	300	BR	Honoraires paramedicaux
5	100	BR	250	BR	300	BR	Analyses et examens de laboratoire
6	100	BR	276.4016052	BR	305.6064207	BR	Actes d'imagerie médicale
7	100	BR	200	BR	300	BR	Transport
8	100	BR	250	BR	300	BR	Materiel medical
9	100	BR	250	BR	300	BR	Materiel medical
10	100	BR	200	BR	300	BR	Pharmacie remboursée par RO
12	100	FR	100	FR	100	FR	Frais de séjour
13	100	BR	276.4016052	BR	305.6064207	BR	Honoraires Hospitalisation
14	100	FR	100	FR	100	FR	Forfait journalier
15	0	EUR	170	EUR	200	EUR	Chambre particulière en Hospitalisation
16	0	EUR	70	EUR	100	EUR	Frais d'accompagnement
18.1	100	FR	100	FR	100	FR	Prothèses dentaires/Inlay-Onlay 100% Santé
18	100	BR	350	BR	450	BR	Prothèses dentaires/Inlay-Onlay
19	100	BR	275	BR	350	BR	Soins dentaires
20	0	EUR	150	EUR	300	EUR	Parodontologie
21	0	EUR	250	EUR	450	EUR	Implantologie
22	100	BR	200	BR	300	BR	Orthodontie
24.1	100	FR	100	FR	100	FR	Monture 100% Santé
24	0	EUR	250	EUR	350	EUR	Monture
25.1	100	FR	100	FR	100	FR	Verres 100% Santé
25	100	BR	300	EUR	400	EUR	Verres simples & monture
25.5	100	BR	450	EUR	600	EUR	Verres complexes ou très complexes & monture
26	0	EUR	300	EUR	400	EUR	Chirurgie refractive
27	47.38	EUR	250	EUR	400	EUR	Lentilles acceptée par RO
27.5	0	EUR	300	EUR	350	EUR	Lentilles refusée par RO
29.1	100	FR	100	FR	100	FR	Appareil auditif 100% Santé
29	700	EUR	1400	EUR	1800	EUR	Appareil auditif remboursé par RO/ 2 Oreilles
29.5	350	EUR	700	EUR	900	EUR	Appareil auditif remboursé par RO/ 1 Oreille
32	100	BR	200	BR	300	BR	Cure thermale
33	30	EUR	150	EUR	200	EUR	Médecine douce

TABLE 2.2 : L'alignement de deux classifications des postes de garantie

fonction de l'affiliation des professionnels de la santé à OPTAM/OPTAM-CO, il n'existe pas de codification spécifique permettant de distinguer les prestations des médecins affiliés à OPTAM dans la base Damir. Pour résoudre cette question, une solution a été mise en place pour regrouper les plafonds de remboursement liés aux actes réalisés par ces médecins. Cette solution consiste à pondérer les plafonds de remboursement en fonction de la proportion des médecins adhérant à OPTAM/OPTAM-CO parmi l'ensemble des médecins. Selon TRANTHIMY (2020), en 2020, on recensait 17 219 médecins affiliés à OPTAM/OPTAM-CO, sur un total de 307 130 médecins en France. Ainsi, le plafond de remboursement commun pour les actes, qu'ils impliquent ou non des médecins affiliés à OPTAM, peut être défini comme suit :

$$Plafond_{commun} = Plafond_{OPTAM} \times \frac{17219}{307130} + Plafond_{non\_OPTAM} \times \frac{(307130 - 17219)}{307130}$$

Ceci représente une solution simple étant donné l'absence de distinction claire des remboursements liés aux médecins OPTAM/OPTAM-CO. Nous devons considérer que tous les médecins sont du même secteur 2 avec OPTAM/OPTAM-CO, avec une possibilité de dépassement d'honoraires. Cependant, les remboursements des complémentaires sont en réalité très hétérogènes en raison de l'adhésion à OPTAM/OPTAM-CO et de la proportion de médecins de secteur 1, de secteur 2 OPTAM et de secteur 2 non-OPTAM, qui varient potentiellement en fonction de la région ou des spécialités. C'est une limite à prendre en compte pour les assureurs s'ils veulent étudier la sinistralité des postes concernant les médecins de secteur 1 ou 2 à travers la base Damir.

Lors de cette étape, le concept du 100% Santé ainsi que les différences de garanties pour une même famille de garantie ont été introduits en se basant sur les actes classifiés dans le mémoire de HUYNH (2021). Les postes concernés sont les suivants :

- Verre : Les verres simples et complexes ont été reclassifiés en deux sous-postes avec des garanties différentes. Les actes correspondant aux verres 100% Santé sont classés dans le sous-poste "Verres 100% Santé" et bénéficient d'un remboursement intégral du reste à charge après la Sécurité Sociale.
- Monture : Seules les montures du panier A ont été classées dans le sous-poste "Monture 100% Santé" avec un remboursement intégral.
- Prothèse auditive : Les actes ont été séparés en 3 sous-postes, notamment les prothèses 100% Santé bénéficiant d'un remboursement intégral des frais restants, les prothèses pour 2 oreilles et les prothèses pour chaque oreille.
- Prothèse dentaire : Étant donné que la grille de garantie ne présente pas de différence de remboursement pour le panier tarif libre ou le panier tarif maîtrisé, il a été décidé de classer les prothèses dentaires et les inlays/onlays en deux parties : normale et 100% Santé, avec une prise en charge intégrale du reste à charge.

Ces classifications permettent de prendre en compte les spécificités du 100% Santé et d'assurer un remboursement adéquat des prestations pour les assurés concernés.

Une fois la grille de garantie adaptée à la base Damir, il est nécessaire de fiabiliser la base afin que la tâche de tarification corresponde au tarif de la population la plus générale possible, c'est-à-dire le tarif applicable aux assurés du régime général de la Sécurité Sociale.

2.2.2 Retraitement de données aberrantes

Avant d'appliquer la grille de garantie benchmark à la base Open Damir pour tarifier les contrats complémentaires santé, il est essentiel de traiter les données aberrantes spéciales de cette base. En raison du caractère exhaustif de la base Open Damir, il peut exister des cas atypiques où les remboursements enregistrés par la Sécurité Sociale semblent incohérents par rapport aux montants réels payés par les assurés. Ces situations particulières peuvent être dues à des facteurs tels que des dépassements d'honoraires importants, des actes médicaux rares ou complexes, ou encore des erreurs dans la saisie des données. Afin d'obtenir une tarification précise et pertinente des contrats complémentaires santé, il est donc nécessaire de nettoyer la base de données en identifiant et en traitant ces données aberrantes. Ce processus permettra d'obtenir des informations plus fiables et représentatives des remboursements réels effectués par les assureurs, et ainsi d'appliquer la grille de garantie de manière plus cohérente et adaptée aux différents profils d'assurés.

Les traitements élémentaires retenus à la première étape de traitement de la base Open Damir

À cette étape, nous avons établi une procédure de traitement des données selon le principe suivant : la complémentaire santé ne rembourse que les restes à charge après le remboursement de la Sécurité Sociale issue du régime général. De plus, elle ne rembourse que les prestations pour lesquelles les patients sont identifiables, en excluant toutes les régularisations présentes dans la base Damir. Les étapes d'élimination des données sont partiellement décrites dans la section précédente et sont désormais réalisées en utilisant les poids correspondants calculés dans la base Damir, comme récapitulé dans la figure 2.3.

Après cette étape, la base est tout d'abord nettoyée afin d'avoir seulement les informations sur les dépenses de l'assuré du régime général de la Sécurité Sociale ainsi que les remboursements du régime obligatoire pour ces dépenses. Avant de plonger dans les méthodes de tarification directement sur la base Damir, il convient d'effectuer encore une fois la vérification de l'exactitude et la certitude des données afin de fiabiliser les procédures de tarification.

Variable	Mode de traitement	But	Pourcentage de ligne éliminé par rapport au nombre de ligne total	Pourcentage de ligne éliminé par rapport au dépense total (Indicateur préfiltré)	Pourcentage de ligne éliminé par rapport au nombre d'acte total (Indicateur préfiltré)
Age bénéficiaire	Supprimer des lignes de l'âge inconnu	Rendre les prestations identifiables à l'assureur	0.136%	0.100%	1.400%
Sexe bénéficiaire	Supprimer des lignes du sexe inconnu	Rendre les prestations identifiables à l'assureur	0.000%	0.000%	0.000%
Région bénéficiaire	Supprimer des lignes de l'âge inconnu	Rendre les prestations identifiables à l'assureur	6.234%	4.126%	5.005%
Indicateur bénéficiaire C2S	Supprimer des lignes concernant les bénéficiaires de la C2S	Rendre les prestations éligibles au remboursement des assureurs	48.235%	12.547%	8.156%
Type de remboursement	Supprimer les lignes donc le type de remboursement ne correspond pas à des prestations issue du régime générale	Rendre les prestations éligibles au remboursement des assureurs	45.802%	Non évaluable	Non évaluable
Type d'enveloppe de la prestation	Supprimer les lignes donc le remboursement fait partie de l'enveloppe de prestation C2S ou régime locale Alsace-Moselle	Rendre les prestations éligibles au remboursement des assureurs	4.682%	0.535%	0.026%
Indicateur de dépense préfiltré	Supprimer les lignes de régulation (Dépense négative)	Ne pas prendre en compte les régulations de la base Damir	26.611%	-0.652%	20.096%
Indicateur de nombre d'acte préfiltré	Supprimer les lignes de régulation (Nombre d'acte négative)	Ne pas prendre en compte les régulations de la base Damir	13.042%	-0.616%	-0.329%
Indicateur de montant remboursé réel de la Sécurité Sociale	Supprimer les lignes de régulation et des franchises, participation forfaitaire (Montant remboursé négatif)	Ne pas prendre en compte les régulations, les participations forfaitaires lors de la tarification (Comme il est déjà prise en compte dans les prestations de référence)	24.981%	0.616%	0.329%
Base de remboursement	Supprimer les lignes de régulation et des franchises (Base de remboursement négative)	Ne pas prendre en compte les régulations, les participations forfaitaires lors de la tarification (Comme il est déjà prise en compte dans les prestations de référence)	24.582%	0.794%	23.595%

TABLE 2.3 : Traitement préliminaire de la base Damir avec les poids d'importance des données dans la base

Problème de différence entre le remboursement théorique et réel de la Sécurité Sociale dans la base Damir

Au cours de l'analyse des données de la base Damir, nous avons identifié un problème potentiel lié à l'indicateur de montant réel remboursé, qui diffère de l'indicateur théorique sur le montant qui devrait être remboursé. Cette divergence peut être préoccupante car elle peut affecter la précision des résultats de tarification. Pour remédier à cette situation, nous allons mener des investigations approfondies pour comprendre les raisons de cette différence et corriger les données concernées. Pour formuler le problème correctement, on suppose qu'a priori le montant remboursé écrit sur chaque ligne de la base Damir doit être égale au produit de la base de remboursement et du taux de remboursement de la Sécurité Sociale écrit sur la même ligne. Or il se trouve que sur certaine ligne de la base, ces deux montants diffèrent (on utilise l'indicateur préfiltré dans notre étude) :

$$FLT\_REM\_MNT \neq PRS\_REM\_BSE \times PRS\_REM\_TAU$$

Le tableau 2.4 récapitule les proportions de lignes présentant ce problème dans la base Damir, selon le poste de garantie existant dans la base. On peut constater en particulier que les postes de lentilles ou de soins dentaires manifestent une forte incohérence, ce qui peut s'expliquer par le fait que les dépenses réelles des assurés étaient inférieures au montant remboursable par la Sécurité Sociale. Par principe d'équité, la Sécurité Sociale ne peut pas rembourser plus que ce qui a été réellement dépensé. Étant donné que le nombre de lignes de données présentant ce problème est trop important pour être éliminé de la base, nous considérons que l'indice de remboursement réel sur chaque ligne de la base Damir est celui qui reflète la réalité et sera utilisé pour le calcul des remboursements des assureurs dans le cadre de la tarification.

ID	Poste de garantie	Somme de dépenses des lignes concernant (en €)	Somme de nombre d'acte des lignes concernant	Nombre de ligne concernant dans la base	% Dépenses total	% Nombre d'acte total	% Nombre de ligne total
1	Consultations généralistes/Consultations spécialistes	1132941936	31910954	472930	9%	6%	2%
2	Actes techniques médicaux	2256611740	61479478	2258392	43%	63%	44%
4	Honoraires paramedicaux	4421815654	338295034	6380460	26%	15%	20%
5	Analyses et examens de laboratoire	2753263200	105489726	3464273	33%	34%	36%
6	Actes d'imagerie médicale	3958336966	99165543	3318037	51%	67%	51%
7	Transport	656511168	31331519	644481	14%	3%	12%
8	Materiel medical / Petit appareillage	2232931357	46412501	2187740	28%	42%	31%
9	Materiel medical / Grand appareillage	393823.56	1024	175	0%	0%	0%
10	Pharmacie remboursée par RO	9694930909	2844653391	28783782	32%	60%	55%
12	Frais de séjour	3444629708	14031773	349811	28%	43%	43%
13	Honoraires Hospitalisation	325373781.7	6759934	403834	8%	30%	26%
14	Forfait journalier	845587.5	43499	3690	0%	1%	1%
15	Chambre particulière en Hospitalisation	11255229.38	541985	24728	22%	35%	34%
18.1	Prothèses dentaires/Inlay-Onlay 100% Santé	53119204 8.9	772467	67407	19%	9%	12%
18	Prothèses dentaires/Inlay-Onlay	1406196500	1601216	56319	42%	22%	11%
19	Soins dentaires	1539830614	44991935	768770	61%	77%	49%
20	Parodontologie	7493456.19	156011	3159	20%	24%	5%
21	Implantologie	8877647.58	12971	964	2%	1%	2%
22	Orthodontie	92777916.7	1236720	60746	5%	28%	34%
24.1	Monture 100% Santé	363886.41	12257	772	2%	2%	1%
24	Monture	93786701.25	622148	22022	5%	5%	4%
25.1	Verres 100% Santé	2585453.87	34306	4427	2%	2%	1%
25	Verres simples & monture	197571466.8	1084045	82986	4%	4%	4%
25.5	Verres complexes ou très complexes & monture	3920995.49	19035	3915	2%	2%	2%
27	Lentilles	27903972.66	220930	43962	94%	92%	85%
29.1	Appareil auditif 100% Santé	1364118.76	1465	317	0%	0%	0%
29	Appareil auditif remboursé par RO/ 2 Oreilles	9631049.12	134967	4859	13%	17%	5%
29.5	Appareil auditif remboursé par RO/ 1 Oreille	5769109.25	3767	513	1%	1%	1%
32	Cure thermale	125750396.3	507134	59378	68%	49%	41%
33	Médecine douce	1012013.11	36727	3398	87%	89%	79%

TABLE 2.4 : Présentation par poste de garantie du problème de différence entre le remboursement théorique et réel de la Sécurité Sociale dans la base Damir

Problème des restes à charge négatifs

Une fois que l'indicateur principal du montant remboursé par la Sécurité Sociale a été choisi, il est essentiel de vérifier que le reste à charge après le remboursement de la Sécurité Sociale, c'est-à-dire le montant que les complémentaires santé peuvent éventuellement rembourser dans notre cas, est positif ou nul. Lorsque la différence entre la variable de dépense "FLT_PAI_MNT" et le montant de remboursement de la Sécurité Sociale "FLT_REM_MNT" est négative, cela indique une incohérence, car cela va à l'encontre du principe de l'indemnisation par l'assurance. De plus, il se peut qu'il existe une autre ligne dans la base Damir pour corriger cette situation. Néanmoins, dans notre analyse, nous excluons toutes les lignes ayant un reste à charge négative en raison de leur très faible impact sur l'ensemble de la base. Cette décision est justifiée par les proportions très faibles, comme indiqué dans le tableau 2.5, où toutes les proportions de volume d'actes, de dépenses totales et de nombre de lignes concernées sont inférieures à 1%.

2.2.3 Calcul de la prime d'assurance par segment de population

Cette partie est dédiée à la méthodologie de calcul des primes d'assurance pour chaque segment de la population. Nous commencerons par expliquer en détail comment nous avons réalisé le calcul de remboursement de l'assureur, en analysant les données ligne par ligne dans la base Damir. Ensuite, nous présenterons la base de population de l'INSEE que nous avons utilisée pour notre analyse, ainsi que la maille de segmentation retenue. Enfin, nous partagerons les résultats obtenus concernant les primes d'assurance, en prenant en compte les ajustements des charges nécessaires pour une évaluation complète et précise. Notons que tous les calculs ont été réalisés sur la base Damir pour l'année 2021.

Calcul de remboursement d'assureur ligne par ligne

Par rapport à l'approche du Coût x Fréquence définie dans les sections précédentes, la méthode de tarification sur un portefeuille existant est retenue pour évaluer un contrat complémentaire santé basé sur la base Damir, en raison de l'agrégation des lignes dans la base sur plusieurs remboursements. Pour

ID	Poste de garantie	Somme de dépenses des lignes concernant (en €)	Somme de nombre d'acte des lignes concernant	Nombre de ligne concernant dans la base	% Dépenses total	% Nombre d'acte total	% Nombre de ligne total
1	Consultations généralistes/Consultations spécialistes	9316060.69	124622	5533	0.0722%	0.0243%	0.0263%
2	Actes techniques médicaux	346871.47	3796	604	0.0066%	0.0039%	0.0117%
4	Honoraires paramedicaux	2219465.39	194660	2860	0.0132%	0.0087%	0.0090%
5	Analyses et examens de laboratoire	416323.06	16187	213	0.0050%	0.0052%	0.0022%
6	Actes d'imagerie médicale	1662130.96	20747	1848	0.0214%	0.0140%	0.0283%
7	Transport	635798.17	37140	473	0.0138%	0.0038%	0.0087%
8	Materiel medical / Petit appareillage	2319188.56	24474	559	0.0286%	0.0219%	0.0079%
9	Materiel medical / Grand appareillage	16501.46	77	15	0.0104%	0.0253%	0.0267%
10	Pharmacie remboursée par RO	3769624.46	428801	3059	0.0123%	0.0091%	0.0059%
12	Frais de séjour	11148572.6	37007	746	0.0903%	0.1126%	0.0919%
13	Honoraires Hospitalisation	152762.3	1353	240	0.0039%	0.0060%	0.0156%
14	Forfait journalier	12652.56	350	11	0.0006%	0.0040%	0.0030%
15	Chambre particulière en Hospitalisation	1757597.79	64580	1534	3.4053%	4.1660%	2.1242%
18.1	Prothèses dentaires/Inlay-Onlay 100% Santé	481	28	10	0.0000%	0.0003%	0.0017%
18	Prothèses dentaires/Inlay-Onlay	1663.38	48	10	0.0000%	0.0007%	0.0020%
19	Soins dentaires	20245.13	911	59	0.0008%	0.0016%	0.0038%
20	Parodontologie	227.1	20	3	0.0006%	0.0031%	0.0049%
21	Implantologie	135.47	3	1	0.0000%	0.0003%	0.0016%
22	Orthodontie	3981.9	177	44	0.0002%	0.0040%	0.0245%
24.1	Monture 100% Santé	0	0	0	0.0000%	0.0000%	0.0000%
24	Monture	0	0	0	0.0000%	0.0000%	0.0000%
25.1	Verres 100% Santé	14	2	1	0.0000%	0.0001%	0.0003%
25	Verres simples & monture	2	2	1	0.0000%	0.0000%	0.0000%
25.5	Verres complexes ou très complexes & monture	0	0	0	0.0000%	0.0000%	0.0000%
27	Lentilles	77.2	6	2	0.0003%	0.0025%	0.0039%
29.1	Appareil auditif 100% Santé	0	0	0	0.0000%	0.0000%	0.0000%
29	Appareil auditif remboursé par RO/ 2 Oreilles	379.98	86	8	0.0005%	0.0110%	0.0089%
29.5	Appareil auditif remboursé par RO/ 1 Oreille	0	0	0	0.0000%	0.0000%	0.0000%
32	Cure thermale	1671.55	11	4	0.0009%	0.0011%	0.0028%
33	Médecine douce	0	0	0	0.0000%	0.0000%	0.0000%

TABLE 2.5 : Problème des restes à charge négatifs

chaque niveau de garantie donné, les remboursements de l'assureur sont produits synthétiquement pour 100% des actes réalisés dans la base, puis agrégés par famille de garantie pour obtenir une vision globale de l'ordre de grandeur des remboursements pour les contrats d'entrée de gamme, de gamme moyenne et haut de gamme pour l'ensemble de la population française. Nous posons donc l'hypothèse que les remboursements produits synthétiquement appartiennent à l'ensemble de la population en France. Cette hypothèse de calcul sur l'ensemble des prestations dans la base est fondée sur la statistique selon laquelle plus de 96% de la population française dispose d'une complémentaire santé en 2019 (voir PIERRE et ROCHEREAU (2022)), ce qui nous semble plausible de considérer que presque 100% de la population est couverte par une complémentaire santé.

Suite aux traitements précédents, nous introduisons une formule de calcul du remboursement de l'assureur pour une ligne de la base Damir, afin de produire des montants de sinistres hypothétiques respectant les principes d'indemnisation assurantielle. Il est important de noter qu'une ligne de prestation agrégée regroupe plusieurs consommations de soins, et nous supposons que toutes les consommations dans une même ligne présentent les mêmes caractéristiques (prix, base de remboursement, taux de remboursement de la Sécurité Sociale, etc.).

En prenant en compte cette hypothèse, nous procédons au calcul des remboursements avec les indices agrégés pour une ligne particulière de la base Damir, comme suit :

$$\begin{aligned} &Remboursement\_assureur = \\ &\qquad Min(Max(0; Depense - Remboursement\_SS); Montant\_maximum\_remboursable), \end{aligned} \quad (2.1)$$

avec :

- $Depense$ correspond à l'indicateur de dépense agrégée préfiltrée de la ligne, c'est-à-dire "FLT_PAI_MNT".
- $Remboursement\_SS$ correspond à l'indicateur de remboursement réel de la Sécurité Sociale agrégé préfiltré de la ligne, c'est-à-dire "FLT_REM_MNT".
- $Max(0; Depense - Remboursement\_SS)$ garantit que le reste à charge ne soit pas négatif pour le calcul du remboursement.

Le montant maximum remboursable ($Montant\_maximum\_remboursable$) dépend du mode de remboursement, et nous distinguons trois cas généraux :

- Remboursement en %BR : $Montant\_maximum\_remboursable = Max(0; Base\_remboursement\_SS \times Taux\_plafond_{\%BR} - Remboursement\_SS)$.Où $Base\_remboursement\_SS$ représente la base de remboursement agrégée indiquée sur chaque ligne de la base.
- Remboursement en %FR : $Montant\_maximum\_remboursable = Max(0; Depense - Remboursement\_SS) \times Taux\_plafond_{\%FR}$.
- Remboursement en € : $Montant\_maximum\_remboursable = Montant\_plafond_{Euro}$.

Ces formules permettent de calculer le remboursement de l'assureur en fonction des montants de dépenses et de remboursements réels de la Sécurité Sociale agrégés sur chaque ligne de la base Damir, tout en respectant les principes d'indemnisation de l'assurance. Le remboursement de l'assureur est donc calculé pour toutes les lignes de la base Damir traitée, en utilisant les trois niveaux de garantie pour chaque famille de garantie. Chaque ligne de la base est classifiée dans un poste de garantie en fonction de l'indice "Nature de Prestation" (PRS_NAT). Pour le poste de lentilles remboursées par la Sécurité Sociale, nous vérifions le montant remboursé par la Sécurité Sociale et, en fonction de ce montant (nul ou strictement positif), les lignes sont séparées en "Lentilles remboursables par la Sécurité Sociale" et "Lentilles non remboursables par la Sécurité Sociale".

Base de population INSEE - segment retenu

Parmi les variables décrivant les caractéristiques des assurés, nous avons décidé de retenir les variables Tranche d'âge, Sexe et Région de résidence comme critères de segmentation de la population présente dans la base Damir. Cette décision est basée sur des tests d'indépendance réalisés sur ces trois variables dans le chapitre 8 du mémoire de Fotsing (2018). Pour associer la segmentation choisie à une exposition totale au risque (c'est-à-dire le nombre d'assurés couverts dans la base Damir), nous avons opté pour la population issue des données INSEE (Insee (2023a)), cette population est estimée au premier janvier de chaque année. En effet, la base Damir ne contient pas d'information sur l'effectif des assurés couverts pour chaque ligne de prestation. Notre hypothèse est donc que la base Damir représente l'historique des sinistres d'un grand portefeuille (considéré ici comme l'ensemble de la France), et que chaque personne en France est couverte par notre complémentaire santé synthétique. Ainsi, l'exposition totale est supposée égale à la population française réelle en France pour l'année 2021, et le coefficient d'exposition est supposé valoir 1 pour chaque assuré.

La base INSEE est maillée de la manière la plus fine, par sexe, par tranche d'âge quinquennale de 0 à 95+ ans, et par département français, et est estimée à la date du premier janvier de chaque année. Afin de pouvoir la faire correspondre avec la segmentation de Damir, il a été nécessaire d'agréger la population des départements au sein d'une même région présente dans Damir, ainsi que des tranches d'âge de la base. Il existe ainsi 8 tranches d'âge x 13 régions x 2 sexes, soit 208 segments dans la base Damir. Après le regroupement de la population au 01/01/2021, nous avons obtenu la population associée à chaque segment. Pour prendre en compte la variation de la population entre le 1er janvier 2021 et le 1er janvier 2022, nous avons pris la moyenne des effectifs de ces deux dates pour chaque segment. Soit a un segment d'assuré :

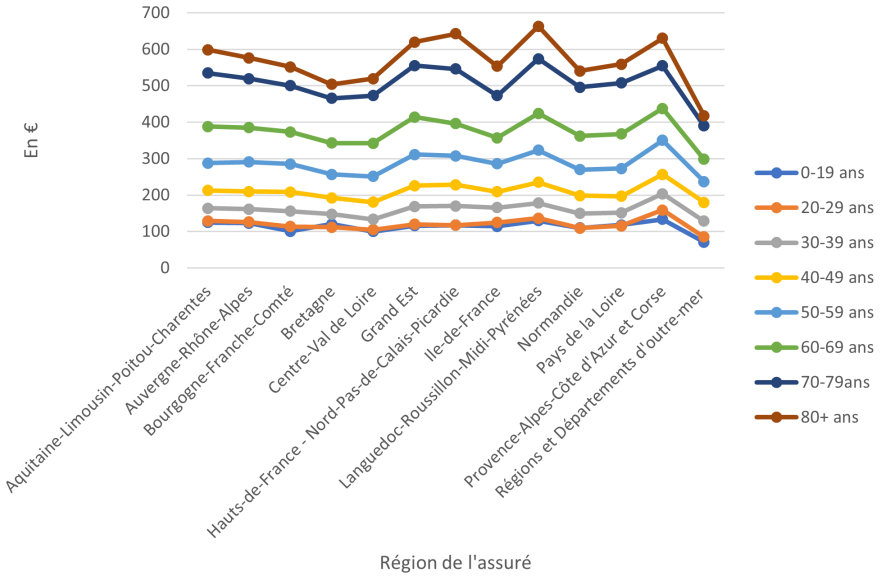
$$Population\_segment\_a_{2021} = \frac{Population\_segment\_a_{01/01/2021} + Population\_segment\_a_{01/01/2022}}{2}$$

Résultat des primes obtenues

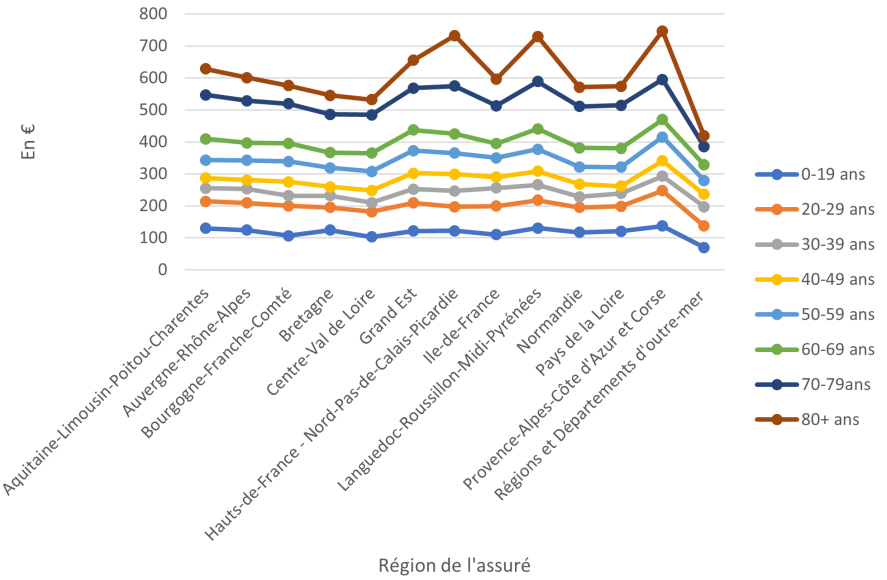
Les remboursements synthétiques ont été agrégés par segment de la population et par identité de la prestation de garantie. En les divisant par la population du segment considéré, on dispose de la prime

technique correspondant au risque de chaque profil assuré par prestation de garantie ainsi que par contrat souscrit (prime totale de prestations). Plus précisément, soit $Remboursement\_assureur^{a}_{i,m}$ la somme des remboursements d'assureur pour le segment a déduits de la base Damir pour la prestation m de niveau de garantie i en 2021, nous calculons la prime technique ou sinistre moyen pour le segment a déduits de la base Damir pour la prestation m de niveau de garantie i en 2021 comme suit :

$$Prime\_pure^{a}_{i,m} = Sinistre\_moyen^{a}_{i,m} = \frac{Remboursement\_assureur^{a}_{i,m}}{Population\_segment\_a_{2021}}$$



(a) Primes techniques pour les hommes



(b) Primes techniques pour les femmes

FIGURE 2.2 : Variation de la prime pour le niveau de couverture minimum en fonction de la tranche d'âge et de la région de résidence (profil homme-femme)

La figure 2.2 nous donne une vision globale des contrats responsables de couverture minimum, car la

grille de garantie minimum respecte les contraintes de tarification des contrats responsables. Ainsi, on observe une augmentation de la prime pure en fonction de la tranche d'âge : plus l'assuré est âgé, plus sa prime est élevée. Cette observation s'explique facilement par le fait que la population diminue de plus en plus à mesure que l'âge évolue, et dans la base Damir, les personnes âgées consomment davantage que les jeunes. On remarque bien que les primes annuelles pour les femmes sont considérablement plus grandes que celles des hommes de l'ordre de plusieurs dizaines d'euros. Cette différence peut être attribuée à une fréquence d'utilisation des soins plus importante chez les femmes, sachant que ces deux profils ont le même niveau de remboursement (les remboursements de soin sont bornés par le contrat responsable).

En vue d'une analyse globale, on peut constater que les variations liées à l'âge sont très marquées pour l'ensemble des postes de garantie. Cependant, en examinant chaque poste de garantie individuellement, on peut confirmer que l'impact de l'âge est beaucoup moins clair pour certains d'entre eux. Prenons, par exemple, le poste d'analyse médicale. La figure 2.3 montre qu'il n'y a pas d'effet significatif de l'âge sur ce poste, car les montants moyens des sinistres ne varient que de moins de 6 euros entre les tranches d'âge. De plus, pour n'importe quelle tranche d'âge, cette famille de garantie ne présente que peu de dépassements par rapport à la base de remboursement, ce qui se traduit par des montants moyens très proches pour les trois niveaux de garantie.

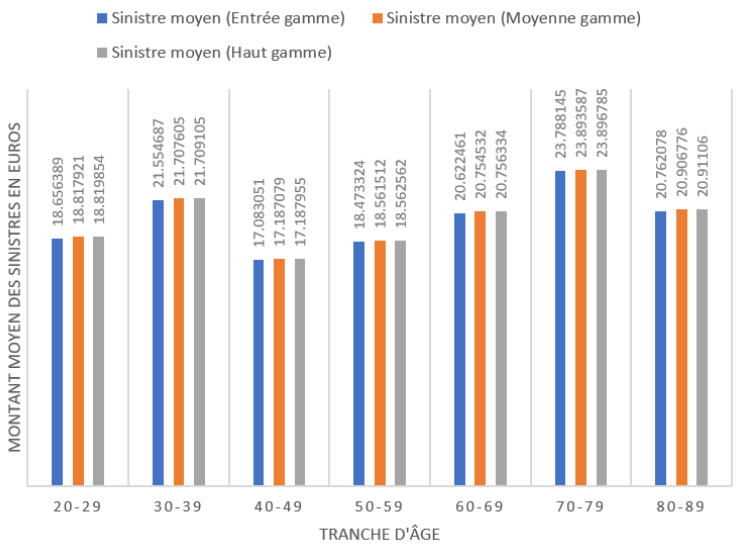
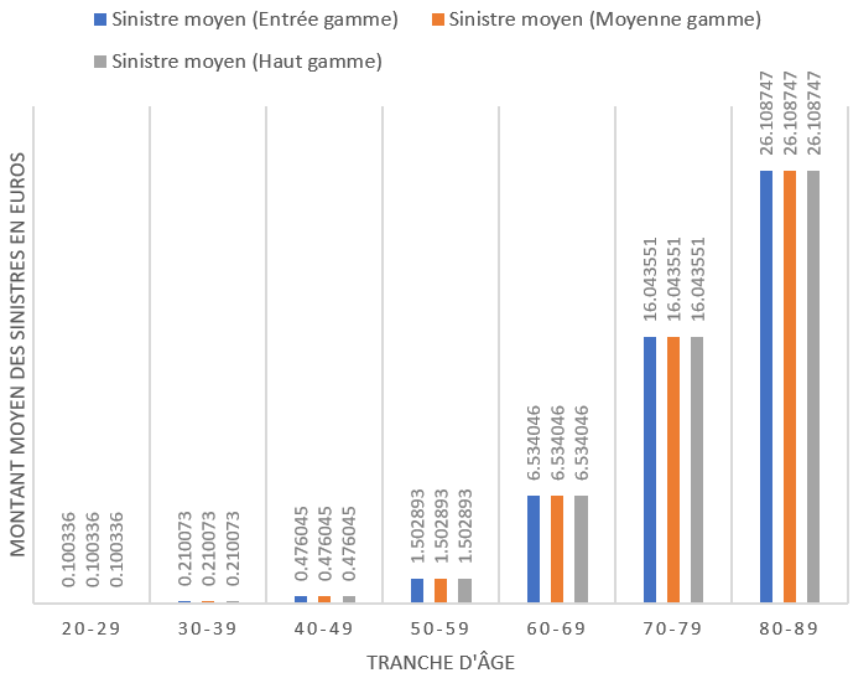


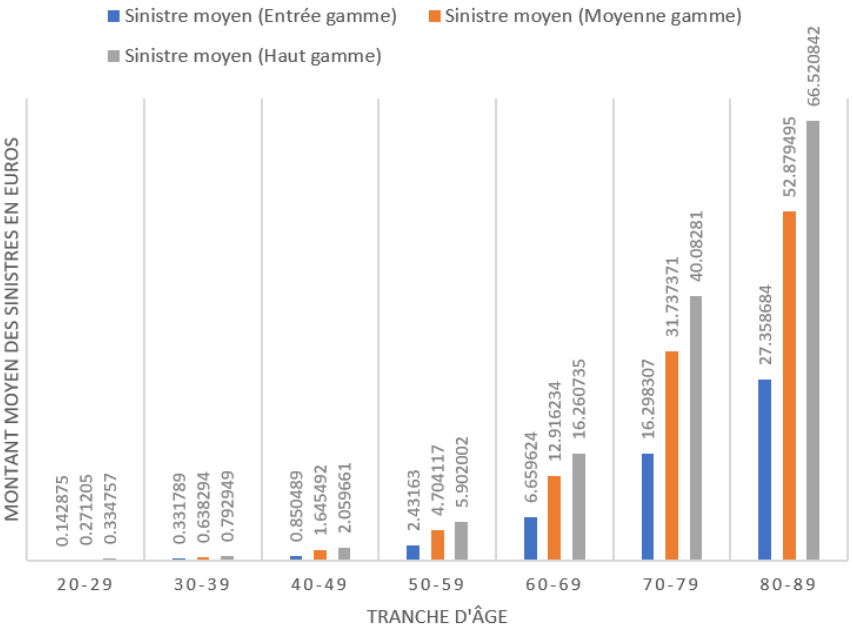
FIGURE 2.3 : Coût moyen d'un sinistre par tranche d'âge sur le poste de garantie d'analyse médicale

En observant le poste d'Appareil Auditif (figure 2.4), qui est segmenté en appareils 100% Santé et hors panier 100%, on peut constater que la consommation varie fortement en fonction de la tranche d'âge pour les deux paniers. Pour le panier hors 100% Santé, les appareils facturés dépassent souvent largement les plafonds de remboursement des contrats d'entrée et de moyenne gamme. L'âge est donc un facteur explicatif de la forte consommation d'appareils auditifs, et cette consommation augmente significativement avec l'âge. Cependant, il est important de noter que la consommation, même expliquée par l'âge, ne croît pas nécessairement avec l'âge, comme c'est le cas pour le poste de garantie Optique. Les montants moyens des sinistres par tranche d'âge (figure 2.5) montrent que la consommation augmente avec l'âge jusqu'à un certain point, puis diminue. Pour le panier libre, les profils de haute consommation se situent chez les 50-59 ans et les 60-69 ans, tandis que pour le panier 100% Santé, ce sont les assurés de la tranche d'âge 70-79 ans qui consomment le plus en moyenne.

Les primes pures ne peuvent pas être demandées directement aux assurés, car elles doivent être augmentées des charges et taxes. La taxe appliquée sur un contrat complémentaire santé responsable



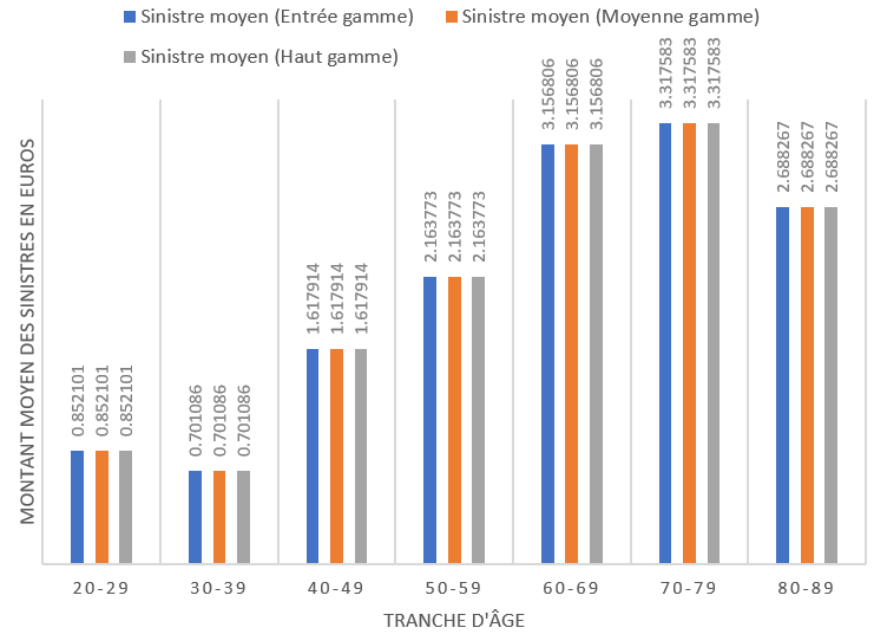
(a) Panier 100% Santé



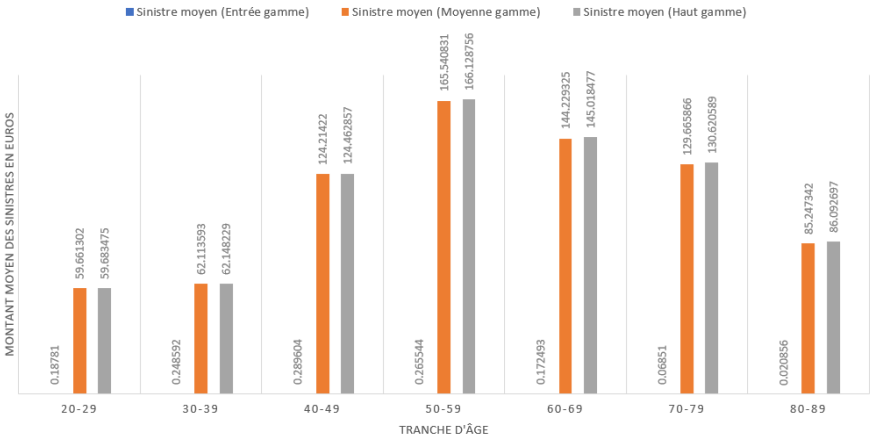
(b) Panier hors 100% Santé

FIGURE 2.4 : Coût moyen d'un sinistre par tranche d'âge sur les postes de garantie d'appareil auditif

s'appelle « Contribution de solidarité additionnelle » (CSA) et s'élève à 13,27%, selon BARLET et al. (2019). En outre, un taux de chargement moyen de 20% sur les primes commerciales, tel que mentionné par MALAKOFF HUMANIS (2020), est également ajouté. Cela permet d'obtenir la prime commerciale



(a) Panier 100% Santé



(b) Panier hors 100% Santé

FIGURE 2.5 : Coût moyen d'un sinistre par tranche d'âge sur les postes de garantie d'optique

proposée aux assurés :

$$Prime\_commerce = Prime\_pure \times \frac{1 + 13.27\%}{1 - 20\%}$$

Cependant, dans la suite du mémoire, nous avons choisi d'appliquer le taux technique comme tarif commercial des contrats visés, afin de maintenir un tarif neutre par rapport aux effets de la taxe et des charges.

2.2.4 Les tarifs retenus

Selon notre cadre de contrats individuels, constatant une différence moyenne de sinistralité entre les hommes et les femmes (illustrée par la figure 2.6, où les sinistralités moyennes de la couverture minimale diffèrent entre les hommes et les femmes), nous avons choisi de tarifer en fonction du sexe.

Bien que la différenciation tarifaire en fonction du sexe soit interdite, nous avons choisi de l'adopter comme les primes proposées aux assurés du portefeuille dans les prochains chapitres afin de présenter l'impact du sexe sur l'effet du prix dans les modèles entraînés, ainsi que de quantifier cet impact lors d'un changement de tarification. Cela constitue donc un aspect différent de notre tarification fictive par rapport aux tarifs dans la réalité et permettra d'étudier hypothétiquement le comportement des assurés sous l'effet de la tarification par sexe.

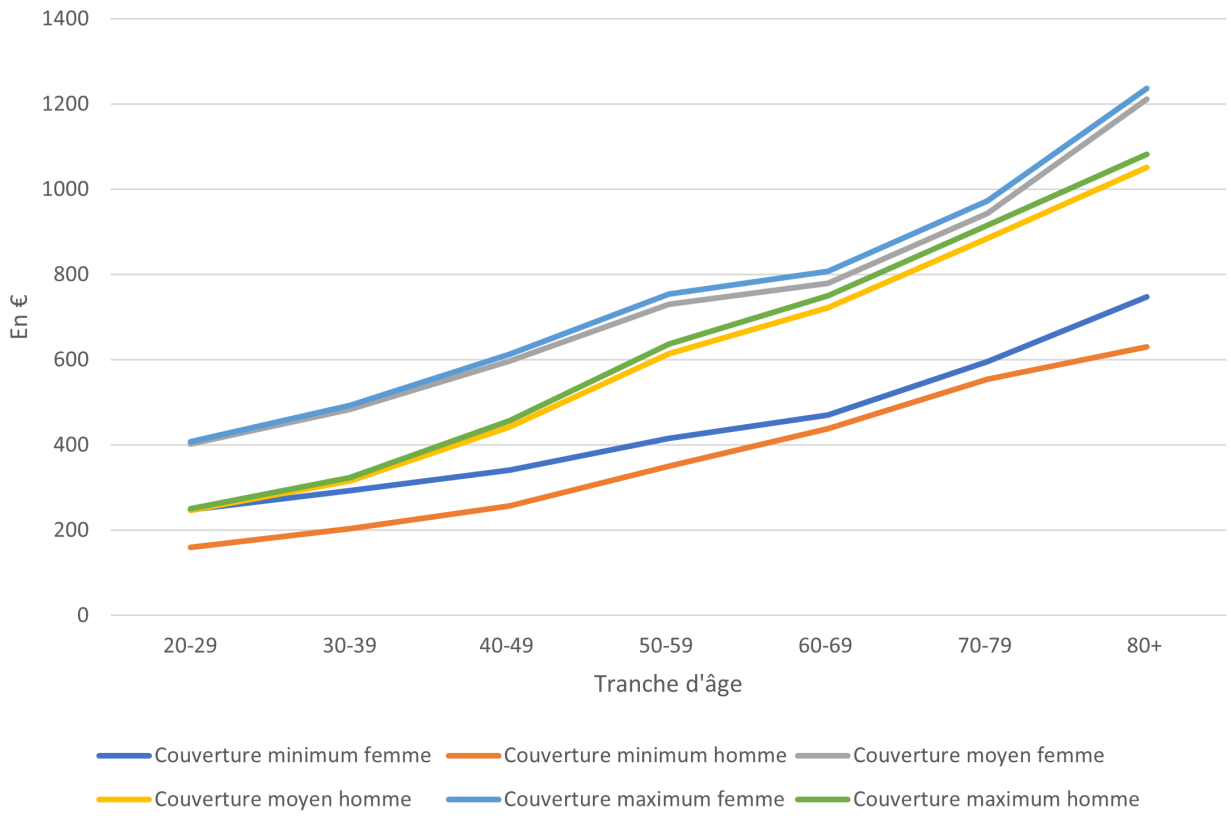


FIGURE 2.6 : Primes pures par tranche d'âge en région Provence-Alpes-Côte d'Azur et Corse, illustrant la différence entre la sinistralité de l'homme et de la femme

Nos tarifs pour les trois niveaux de couverture reposent sur trois variables discriminatoires : la tranche d'âge, le sexe et la région de résidence des bénéficiaires. Compte tenu de l'inadéquation de la tranche d'âge 0-20 ans pour la proposition de contrats complémentaires santé (la plupart des assurés de cette tranche d'âge sont encore rattachés aux contrats complémentaires santé de leurs parents), nous avons décidé de ne pas proposer notre contrat complémentaire santé à ce segment de la population.

Les tarifs retenus (présentés dans la table A.7) seront utilisés dans le prochain chapitre comme les tarifs des contrats complémentaires santé individuels proposés aux assurés en France relevant du régime général, sans être bénéficiaires de la CSS ni du régime local. Ces tarifs sont dits "de l'assuré seul", car les autres personnes telles que les enfants de l'assuré, les conjoints, etc., ne sont pas couvertes.

Chapitre 3

Construction de portefeuilles fictifs avec anti-sélection

L'objectif de ce chapitre, inspiré par la littérature, est de construire des portefeuilles fictifs d'assurance complémentaire santé individuelle comportant de l'anti-sélection à partir de données publiques, ainsi que d'utiliser les tarifs issus de la base Open Damir. Nous commençons par approfondir le phénomène, puis nous procédons à la construction d'un indicateur pour mesurer le degré de l'anti-sélection existant dans un portefeuille.

3.1 Littérature sur l'anti-sélection en assurance complémentaire santé

3.1.1 Anti-sélection dans les littératures modernes

Dans la littérature classique de Rothschild et Stiglitz (1976), le phénomène d'anti-sélection est défini dans une situation où il existe plusieurs profils de risque que les assureurs ne parviennent pas à distinguer, ce qui conduit à l'application d'un seul tarif et à une spirale de mort. La solution proposée est de proposer plusieurs offres d'assurance adaptées à chaque profil de risque afin d'atteindre un équilibre réalisé par l'auto-sélection des assurés selon leur risque. Cette définition de l'anti-sélection est unidimensionnelle en fonction du profil de risque.

Cependant, en réalité, même en présence d'offres d'assurance avec différents niveaux de couverture, comme en assurance santé, les assureurs peuvent encore être sujets à des pertes causées par l'anti-sélection en raison de l'asymétrie d'information sur la tarification et la souscription. D'un côté, nous pouvons donner une explication classique où les profils à haut risque recherchent une bonne couverture (Bakker et van Vliet (1993)). Mais d'un autre côté, les preuves incidentes d'anti-sélection (Cutler et Zeckhauser (1998), Geruso et al. (2023), Powell et Goldman (2021)) dans les marchés avec plusieurs niveaux de couverture montrent que l'existence de contrats avec différents niveaux de remboursement n'est pas forcément efficace pour éviter l'anti-sélection.

Nous pouvons donc adopter une compréhension de l'anti-sélection où une personne a non seulement connaissance de son profil de risque mais aussi une référence pour l'assurance (voir Alessie et al. (2020) pour l'explication du bidimensionnel de l'anti-sélection), ce qui les incite à souscrire à différents niveaux de couverture. En même temps, les assureurs ne sont pas en mesure de distinguer les différents profils de risque dans leurs tarifs (par exemple, l'interdiction de tarification par sexe en France ou par les tests médicaux). L'anti-sélection dans ce cas se traduit simplement par une mauvaise allocation des profils de risque dans le portefeuille, avec une composition de hauts risques plus importante que prévu lors de la tarification, ce qui entraîne des sinistres plus importants que prévu. Couplé avec le fait que les assurés à haut risque tendent à souscrire à de bonnes couvertures, nous pouvons dire que l'anti-sélection affecte surtout les contrats d'assurance généreuse en raison du manque de mutualisation avec les profils

à faible risque, ce qui entraîne la spirale de la mort pour les contrats généreux, comme l'exemple de l'assurance santé à Harvard (Cutler et Zeckhauser (1998)). Un tel phénomène de souscription non proportionnelle des profils de risques dans une complémentaire santé, où les bas risques optent pour d'autres contrats moins coûteux tandis que les hauts risques restent, est également appelé la sélection favorable, comme décrit par Newhouse et al. (2012) pour les contrats d'assurance santé aux États-Unis.

Selon Legal (2008) sur les liens entre la demande d'assurance santé et le risque moral, l'augmentation des dépenses chez les personnes assurées avec une couverture généreuse peut résulter à la fois de l'aléa moral ex-post (où les gens essaient de consommer plus parce que leur assurance le permet) et de l'anti-sélection (où les gens consomment beaucoup conformément à leurs anticipations de dépenses de santé avant de souscrire). Il est important de noter que dans le cas de l'aléa moral ex-post, peu importe le profil de risque de l'assuré, le comportement sera similaire et il y aura des dépenses significatives sur tout le portefeuille, tandis que l'anti-sélection passe par la réalisation des hautes dépenses anticipées avant de souscrire aux couvertures généreuses, et donc plutôt sur certains profils en particulier. Pour identifier ces anticipations de dépenses menant au choix du niveau de couverture, nous devons analyser les facteurs courants qui impactent la demande d'assurance.

3.1.2 Effets des variables socio-économiques sur la demande d'assurance

Les littératures sur la demande d'assurance ont trouvé des preuves de la dépendance des préférences en matière d'assurance en fonction de variables socio-économiques telles que l'âge, le sexe, les facteurs géographiques, le niveau d'éducation, l'état de santé, la composition familiale, etc. Nous présentons ici quelques résultats importants, à la fois en dehors de la France et en France, liés à nos trois facteurs de tarification, à savoir l'âge, le sexe et le facteur géographique.

En ce qui concerne l'âge des assurés, les études menées par Liu et Boes (2022) (sur l'assurance santé en Suisse) et Van de Ven et Van Praag (1981) (sur l'assurance santé aux Pays-Bas) constatent une augmentation de la demande d'assurance, c'est-à-dire un niveau de couverture plus élevé, avec l'âge, ce qui est cohérent avec l'augmentation de la consommation médicale chez les personnes âgées.

En ce qui concerne la variable sexe, les études menées par Gorter et Schilp (2012) (sur le marché hollandais), Kalouguina et Wagner (2020) (sur le marché suisse) et Liu et Boes (2022) ont trouvé que les femmes choisissent plus souvent des contrats avec moins de franchise, c'est-à-dire plus généreux. Cela peut s'expliquer possiblement par deux raisons : d'une part, les femmes consomment généralement plus de soins médicaux que les hommes, d'autre part, les femmes sont peut-être plus averses au risque.

L'étude menée par Kalouguina et Wagner (2020) a également révélé que le facteur géographique entraîne des différences dans la demande d'assurance, avec les régions rurales étant plus susceptibles de choisir des niveaux de couverture plus bas. Elle a également constaté que le facteur de taille de la famille et le niveau d'éducation sont négativement corrélés à la demande d'assurance, une observation confirmée par Van de Ven et Van Praag (1981).

Par ailleurs, une étude sur le marché de l'assurance santé en Australie menée par Cameron et al. (1988) postule que tous les facteurs mentionnés ci-dessus sont corrélés à la demande d'assurance.

En France, Legal (2008) a constaté que l'âge et le sexe sont très corrélés avec le niveau de couverture de plusieurs postes de garantie. Plus l'assuré est âgé, plus la couverture choisie sera grande, et les femmes souscrivent davantage que les hommes à des couvertures élevées. L'importance de ces deux variables dans le choix de l'assurance santé est ainsi confirmée par Franc et al. (2010a), Saliba et Ventelou (2007), et Legal (2009). Les facteurs géographiques sont pris en compte dans l'étude de Franc et al. (2010a), où certaines modalités géographiques sont déterminantes de la demande de surcomplémentaire, ainsi que dans l'étude de Saliba et Ventelou (2007), sous la forme de différences de prime entre la région de l'assuré et Paris, mettant en évidence l'importance de cet effet géographique.

À la fin du chapitre 2, nous avons pris en compte trois variables de segmentation de la population, à savoir l'âge, le sexe et la région de l'assuré, pour la tarification. Cette approche reflète bien notre

précision dans l'analyse de chaque facteur de risque susceptible d'influencer le choix d'assurance, où le niveau de risque présente une hétérogénéité marquée en fonction de ces trois critères. Nous supposons donc l'hypothèse de corrélation entre l'anti-sélection et les variables socio-économiques.

Hypothèse : La demande d'assurance et son implication dans le problème de l'anti-sélection sont liées à des facteurs socio-économiques.

3.1.3 Facteurs de risque santé et revenu sur la demande d'assurance santé

Comme expliqué précédemment, nous notons qu'il y a deux facteurs qui expliquent la demande d'assurance santé, notamment l'anti-sélection (Alessie et al. (2020)) : le type de risque en santé et la préférence pour le risque/l'assurance. Nous décomposons le risque dans la préférence pour l'assurance en deux composantes majeures : la préférence pour le capital santé et la préférence pour le risque monétaire (ou l'aversion au risque monétaire). Alors, comment est défini le premier terme ?

En économie comportementale appliquée au domaine de la santé, plus particulièrement dans le choix des traitements médicaux, un individu est supposé avoir un certain capital santé au début de sa vie, qui diminue au cours du temps en parallèle à son vieillissement (voir Majnoni D'Intignano (2013)). Les soins médicaux sont un moyen de restaurer le capital santé avant la baisse causée par une maladie. Un individu a naturellement une aversion au risque de santé, c'est-à-dire une incitation à choisir un traitement médical maintenant plutôt que de voir sa santé se détériorer après un choc sanitaire imprévu. En fonction de son aversion au risque, Bien (2001) définit la prime de risque santé qu'un individu est prêt à payer pour obtenir un état de santé certain lorsqu'il est confronté à des chocs de santé aléatoires. On peut introduire un effet d'hétérogénéité lié à la restauration du capital santé, qui consiste à dire que les individus, pour un même choc de santé, cherchent à avoir des traitements de meilleure qualité ou plus accessibles, ce qui augmente la perte financière et incite l'individu à souscrire à des contrats de couverture plus élevée que d'habitude. Nous parlons ici de l'effet "access motive" (Schneider (2004)).

Dans plusieurs études sur la demande d'assurance santé (Geruso et al. (2023), Liu et Boes (2022), Kalouguina et Wagner (2020), Gorter et Schilp (2012), Cameron et al. (1988), Schneider (2004), Alessie et al. (2020), Saliba et Ventelou (2007)), les revenus sont fortement corrélés avec la demande d'assurance et affectent surtout l'aversion au risque monétaire. Geruso et al. (2023), Van de Ven et Van Praag (1981) et Alessie et al. (2020) observent une tendance à la baisse du niveau de couverture (ou de la quantité d'assurance achetée) lorsque les revenus sont élevés. En effet, l'aversion au risque monétaire diminue lorsque les revenus augmentent, ce qui est tout à fait normal puisque ceux qui ont des hauts revenus sont indifférents entre la perte financière subie en cas de sinistre avec une faible couverture et l'achat d'une assurance plus importante. En revanche, pour les personnes à faibles revenus, elles sont plutôt incitées à acheter de l'assurance car le risque financier en cas de sinistre est important et stressant pour elles. D'autre part, certains auteurs trouvent une corrélation positive entre le niveau de couverture souscrit et le salaire, par exemple Kalouguina et Wagner (2020) trouve que moins un assuré est riche, moins il souscrit à un niveau de couverture élevé et plus un assuré est riche, plus il souscrit à un niveau de couverture élevé dans le cas de Cameron et al. (1988) et Geruso et al. (2023). Franc et al. (2010b) et Grignon, Kambia-Chopin et al. (2009) mettent en évidence l'importance de la variable salaire ou revenu dans la demande d'assurance et donc sur la détermination de l'utilité de l'assurance perçue par l'assuré. Il convient également de souligner que les caractéristiques spécifiques des contrats d'assurance, telles que les limites de couverture ou les franchises applicables à chaque produit, exercent une influence significative sur l'utilité globale ressentie par l'assuré.

En France, les études de Saliba et Ventelou (2007), Legal (2009) et Franc et al. (2010a) semblent bien confirmer l'importance de l'effet revenu sur le choix de couverture d'assurance, ce qui confirme notre intérêt pour son utilisation par la suite.

Par rapport au profil de risque, nous avons des raisons de croire que cette variable impacte for-

tement la demande d'assurance, comme dans la théorie classique de l'anti-sélection : ceux en moins bonne santé cherchent à se couvrir davantage que ceux en bonne santé. Certaines études portant sur la demande d'assurance et la consommation de soins ou l'état de santé (comme proxy de risque) sont en faveur de cette hypothèse de corrélation, comme Geruso et al. (2023), Liu et Boes (2022), Kalouguina et Wagner (2020) et Van de Ven et Van Praag (1981). Cependant, la corrélation positive entre le niveau de couverture et les dépenses de santé est possiblement un signe d'aléa moral. Il est donc nécessaire de prendre en compte les cas inverses de cette hypothèse, par exemple Cameron et al. (1988) suggère que l'état de santé affecte plutôt la consommation de soins que la décision d'assurance, contrairement aux préférences pour les risques.

En synthétisant la littérature présentée, nous mettons l'hypothèse suivante :

Hypothèse : La demande d'assurance dépend des niveaux de risque ainsi que des préférences pour le risque. Elle est notamment fortement corrélée au facteur de revenu ou au niveau de richesse.

À partir de l'hypothèse sur les facteurs influençant le choix de l'assurance, nous pouvons récapituler les principaux facteurs qui contribuent au phénomène d'anti-sélection, résultant du comportement stratégique des individus au moment de la souscription des contrats d'assurance :

- Niveau de risque et préférence pour les risques : Ces facteurs influencent directement la fréquence des sinistres individuels, leur anticipation de dépenses et constituent une asymétrie d'information car les assurés connaissent mieux leur niveau de risque que les assureurs.

- Préférence pour le risque santé : Ce facteur détermine le degré d'aversion au risque lié au capital santé de l'individu, et son hétérogénéité influence la variation du coût des sinistres sur le même capital santé et en cas du même choc de santé. En général, on peut dire que ceux qui sont en mauvaise santé et plus averses au risque sanitaire réalisent plus de dépenses en santé et donc supportent des coûts d'assurance plus élevés.

- Préférence pour le risque monétaire : Ce facteur détermine le degré d'aversion au risque lié au capital monétaire de l'individu, et son hétérogénéité influence la variation de perception du coût des sinistres sur sa richesse. Cette préférence est très corrélée avec la variable de revenu, car les riches ont tendance à avoir moins d'aversion au risque monétaire que les pauvres.

3.1.4 La particularité de l'anti-sélection en assurance santé en France

L'hypothèse sur l'asymétrie d'information sur le choix de couverture en santé

Ericson et Sydnor (2017), dans leur étude sur le niveau de couverture des contrats d'assurance sur le marché de l'assurance santé aux États-Unis, ont abordé le problème de la confusion des assurés lors du choix de leur contrat d'assurance. Cette confusion, causée par la complexité des remboursements très différents offerts par les contrats d'assurance complémentaire santé, ainsi que par le manque d'information sur le système médical, a réduit la capacité des assurés à anticiper les coûts des soins liés à leur santé. Cette confusion a donc conduit à des choix aléatoires et non optimaux pour les individus. Les auteurs, dans leur étude sur un marché non réglementé (c'est-à-dire sans segmentation, etc.), où la prime technique reflète le coût moyen du risque de chaque contrat, ont également constaté que cette répartition aléatoire des assurés en situation de confusion réduisait l'effet de l'anti-sélection. En effet, plus le nombre de personnes confuses souscrivant à un contrat d'assurance augmente, plus le sinistre moyen de ce contrat tend à se rapprocher du sinistre moyen de ce plan pour l'ensemble de la population, ce qui le rend moins élevé que s'il n'était souscrit que par des personnes bien informées stratégiquement. En général, on pourrait dire que plus les gens comprennent bien le système de santé-assurance et leur propre risque de santé, plus ils sont capables d'anticiper leurs dépenses, renforçant ainsi l'effet de l'anti-sélection. Seuls les individus présentant un risque élevé et une forte aversion au risque bénéficient de cette situation, car leur risque est mutualisé avec ceux qui présentent un risque

moindre, et leur coût marginal est supérieur à la prime à payer. Ces mêmes auteurs ont étudié le cas d'un marché de l'assurance avec davantage de réglementation et un ajustement des primes en fonction du risque (c'est-à-dire où il n'est pas possible de tarifer en fonction de l'état de santé). Ils ont constaté que l'effet d'ajustement des primes atténuait l'effet de l'anti-sélection, tandis que l'effet de la confusion au moment du choix du contrat annulait cet effet.

Pour appliquer ces analyses au marché de l'assurance santé en France, il convient de prendre en compte certaines caractéristiques importantes du système :

- Les organismes complémentaires segmentent de plus en plus leurs tarifs des contrats individuels, ce qui permet d'obtenir une prime de risque plus proche de la sinistralité moyenne des assurés, renforçant ainsi l'effet de l'auto-sélection des assurés sur les niveaux de couverture en fonction de leurs anticipations.

- La Sécurité Sociale et le gouvernement jouent un rôle important en établissant une couverture de base qui prend en charge la grande majorité des dépenses nécessaires. En effet, les dispositifs du "100% Santé" et des contrats responsables permettent à l'ensemble de la population, même en souscrivant au contrat minimum de couverture, de se couvrir contre les risques de santé de manière assez complète, à condition de ne pas exiger des soins de haute qualité.

- Les remboursements et l'offre de soins sont partiellement encadrés (complètement lorsque les médecins sont en secteur 1) par les plafonds de remboursement (BR). Le fait que les garanties des complémentaires santé viennent compléter le remboursement de base est assez compréhensible, et les assurés peuvent anticiper facilement leurs dépenses de santé.

Nous posons donc une hypothèse sur le choix d'assurance santé des Français :

Hypothèse : En France, la proportion de choix de couverture en assurance complémentaire santé résultant d'une confusion des assurés est négligeable, tandis que l'effet de l'asymétrie d'information en faveur des assurés est significatif. En effet, les assurés choisissent l'assurance en fonction de leurs anticipations.

Non seulement nous posons l'hypothèse sur la rationalité des assurés sociaux en France, nous pouvons ainsi pointer les conséquences de risque moral sur la sinistralité des contrats complémentaire santé. Comme la Sécurité Sociale a mis en place une couverture universelle pour presque toutes les dépenses médicales des Français, et que les complémentaires santé viennent combler la partie nécessaire pour les soins de base à tous les niveaux de couverture, la différence de coût des sinistres réside principalement dans le produit médical lui-même. Saliba et Ventelou (2007) dit le fait que les assurés, étant définis haut risque comme très mauvaise santé, sont en général exemptés de ticket modérateur et donc ne pèsent pas lourdement pour les complémentaires santé. Les auteurs ont ainsi expliqué la non-corrélation entre l'état de santé et le choix de souscription d'assurance par l'influence des dépenses concernant les soins auxiliaires, sachant que ces soins se distinguent par leur prix, leur qualité et leur facilité d'accès. L'étude de Plantier (2021) utilise un système de deux équations de Tobit pour estimer à la fois la probabilité de souscrire à une complémentaire santé et la consommation médicale, afin de montrer la liaison entre les préférences pour le risque de santé et monétaire. Dans son étude, elle a trouvé l'effet de sélection dû à l'hétérogénéité des préférences individuelles plutôt qu'au niveau de risque. Avec ses raisonnements, nous posons par la suite une autre hypothèse sur les risques moraux en assurance santé en France :

Hypothèse : En France, la différence de sinistralité entre les contrats d'assurance complémentaire santé ou la préférence pour les niveaux de couverture vient principalement de la préférence pour le risque santé des assurés. Cela se traduit en termes d'anti-sélection par le fait que les assurés anticipent des dépenses pour des soins plus coûteux ou plus confortables que les soins de base, puis souscrivent à des contrats d'assurance complémentaire santé de bonne couverture permettant de réaliser ces soins.

En d'autres termes, si les assurés ne cherchent pas à se soigner par des services plus chers comme des lunettes de marque ou des médecins de secteur 2, ils n'auront pas besoin de souscrire à des contrats

plus généreux que celui remboursant les tickets modérateurs et bénéficiant du dispositif "100% Santé". Pour la suite, il est donc important de regarder les qualités des contrats d'assurances (sur plusieurs caractères comme le niveau de remboursement, le nombre de soins remboursés, etc.) pour examiner les motivations de souscription des assurés.

Hypothèse de séparabilité entre l'anti-sélection et l'aléa moral dans le marché individuel

Sur les littératures de marché de l'assurance santé étrangères, le fait que les individus aient un choix d'assurance quand il y a plusieurs offres avec des niveaux de remboursement différents, à travers leur employeur ou les assureurs privés ou publics, est normalement accompagné de l'implication que la magnitude du phénomène d'anti-sélection est amplifiée tandis que celui de l'aléa moral sera réduit. En effet, l'étude de Alessie et al. (2020), Remmerswaal et al. (2019) et van Vliet (2004) sur le marché de l'assurance aux Pays-Bas montre que l'effet des contrats avec franchise volontaire (c'est-à-dire réduisant la quantité d'assurance) réduit l'effet de l'aléa moral et incite à l'effet d'anti-sélection. De plus, l'étude de Powell et Goldman (2021) sur les différences de générosité de contrat et Gardiol et al. (2005) sur les différences de niveaux de franchise nous donne un poids plus important à l'anti-sélection par rapport à l'aléa moral dans le contexte de l'assurance avec un choix de niveau de couverture.

En France, même avec un système de l'assurance sociale à deux étages, ce qui est différent de la plupart des autres systèmes, il existe un marché d'assurance santé individuel avec différents niveaux de couverture possibles, contrairement au marché collectif qui propose un seul niveau de couverture pour tous les assurés couverts. Nous pouvons donc déduire l'existence du phénomène de l'anti-sélection (Caussat et Glaude (1993b), Albouy et Crepon (2007)) ainsi qu'une réduction de la magnitude du phénomène d'aléa moral. En effet, Albouy et Crepon (2007) disent que l'aléa moral existe surtout et presque uniquement dans le cadre de contrats collectifs où les salariés ne connaissent que leur niveau de garantie et anticipent les charges médicales après la souscription du contrat par leur employeur. Logiquement, les personnes qui ne sont pas couvertes par les contrats collectifs (souvent généreux) ou la CMUC (le minimum de remboursement) auront moins de propension à choisir une couverture généreuse s'ils n'ont pas déjà connaissance de leurs habitudes de consommation de soins. Par exemple, si leur médecin de famille est de secteur 2 avec un dépassement important, ou si la densité des médecins de secteur 1 est élevée près de chez eux, alors l'assuré aura clairement envie de se couvrir contre ces dépassements en choisissant des contrats généreux. La décision d'assurance dans ce cas est donc une conséquence des préférences en santé et l'effet de l'aléa moral ne joue pas un grand rôle dans l'augmentation de la sinistralité. Nous émettons donc une hypothèse importante pour la réalisation de notre mémoire :

Hypothèse : L'effet de l'aléa moral est considéré comme beaucoup plus faible et assimilé à l'anti-sélection sur le marché de l'assurance à adhésion facultative. Presque toutes les conséquences de l'asymétrie d'information dans le marché de l'assurance santé individuelle peuvent être attribuées au phénomène de l'anti-sélection.

Les postes incitent à l'anti-sélection

Certaines études sur le problème de l'anti-sélection en assurance santé en France ont souligné que les postes moins bien remboursés par la Sécurité Sociale, tels que l'optique, la dentisterie ou l'audition, présentaient des effets d'anti-sélection marqués. En effet, Legal (2008), Wang (2015), Valdiguie (2017) et Weiss (2017) donnent une importance à l'effet de l'anti-sélection sur les postes optique, dentaire et auditif. Il est plausible de dire que les assurés anticipent déjà l'achat de lunettes avant de souscrire à un contrat et préfèrent acheter des produits onéreux afin d'être mieux remboursés. Les postes de médecin généraliste et spécialiste présentent quant à eux moins de problèmes d'anti-sélection. On pourrait également dire que les postes d'hospitalisation comme les frais de séjour et d'ambulance

ne présentent pas, à priori, de phénomène d'anti-sélection car les assurés ne peuvent pas anticiper leur risque d'hospitalisation normalement.

D'autre part, le dispositif du "100% Santé" devrait combler ces lacunes en matière de remboursement, en imposant une limite à la fréquence de consommation et en liant la surconsommation de ces postes à la qualité des appareils ou des services de santé. Ainsi, identifier l'anti-sélection parmi les postes Optique/Dentaire/Auditif revient à identifier le recours à l'achat de lunettes/prothèses/soins dentaires en dehors du 100% Santé ainsi que leur variation de coût, comme cela montre le plus clairement les préférences des assurés. Nous posons ensuite une hypothèse sur les postes de garantie présentant anti-sélection :

Récapitulatif des hypothèses retenues sur l'anti-sélection

En parcourant la littérature sur l'anti-sélection en marché de l'assurance santé, nous présentons le tableau 3.1 résumant les hypothèses permettant de caractériser l'anti-sélection en France avec les informations à disposition des assureurs.

Hypothèse
Utilisation des variables socio-économiques pour quantifier la demande en assurance est plausible
Corrélation entre l'anti-sélection et la préférence pour les risques ainsi que le niveau de risque sont forts
Rationalité des assurés dans le choix de l'assurance limite l'effet de souscription par hasard
Effet de la consommation de soins de confort expliquant la haute sinistralité des contrats généreux
L'anti-sélection prime sur l'aléa moral dans le cadre du marché de la complémentaire santé individuelle

TABLE 3.1 : Récapitulatif des hypothèses posées sur la modélisation de l'anti-sélection dans le cadre du mémoire

En fin de compte, on constate que la détection de l'anti-sélection revient à identifier l'anticipation des dépenses médicales très probables des assurés avant qu'ils ne souscrivent à un contrat d'assurance. Cette anticipation est fortement influencée par la perception financière de la situation de l'assuré ainsi que par ses préférences en matière de santé.

Sous les hypothèses présentées ci-dessus, le recours à des méthodes de choix discret au chapitre 5 mettra en évidence cet effet de préférence pour la santé et l'anticipation des dépenses que les assureurs ne sont pas en mesure de connaître. Le travail important lors de la modélisation consistera donc à intégrer les attributs liés à la préférence pour la santé ainsi qu'à l'aversion au risque de perte financière, en exploitant les données et en choisissant des variables de manière adéquate pour refléter l'hétérogénéité au sein de la population à travers le modèle. On voit bien qu'un des proxies de la préférence pour la santé est le plafond de remboursement des postes (voir PLANTIER (2021)) ou l'historique de consommation médicale. La section suivante permettra de développer un cadre théorique lié à un vrai portefeuille d'assurance complémentaire santé individuelle.

3.2 Mesurer l'anti-sélection dans un portefeuille d'assurance

Après avoir analysé les facteurs causant du phénomène d'anti-sélection, on peut procéder à des calculs d'écart de sinistre réel et ciblé et mesurer l'impact du phénomène de l'anti-sélection sur un portefeuille des compagnies d'assurance. Les sections suivantes illustrent comment on peut incorporer les modèles de choix discret dans l'évaluation du phénomène d'asymétrie d'information posant un vrai problème aux organismes complémentaires santé.

3.2.1 Présentation de la base d'assurés dans un portefeuille

Afin de généraliser la démarche de la méthode, on suppose qu'un assureur donné, après la période de souscription des contrats complémentaires santé individuels par exemple de trois niveaux (l'entrée de gamme, milieu de gamme, haut de gamme), dispose à la fin de l'année d'une base d'assurés contenant des informations liées à chaque assuré (tableau 3.2) dans son portefeuille ainsi qu'une base de sinistres liée au remboursement des frais médicaux des contrats (tableau 3.3).

ID	Âge	Sexe	Région	Revenu	Secteur d'activité	Exposition	Contrat
1	35	Homme	IDF	45,000 €	Industriel	0.4	Basse gamme
2	29	Femme	Bretagne	60,000 €	Pharmacie	0.8	Basse gamme
3	42	Femme	Normandie	74,000 €	Assurance	0.97	Moyen gamme
4	77	Homme	Occitanie	53,000 €	Retraite	1	Haut gamme

TABLE 3.2 : Exemple d'un portefeuille existant des assureurs proposé les produits complémentaires santé : la base d'information d'assuré

ID	Code d'acte	Payement	BR	Taux BR	Franchise	Reste à charge	Contrat	Remboursement effectué
1	Consultation médecin généraliste	25 €	25 €	70%	1	8.5 €	Basse gamme	7.5 €
4	IRM	160 €	69 €	70%	0	111.7 €	Haut gamme	111.7 €
3	Auditive 100% Santé	900 €	640 €	70%	0	452.0 €	Moyen gamme	452.0 €

TABLE 3.3 : Exemple d'un portefeuille existant des assureurs proposé les produits complémentaires santé : la base de sinistre

La base assuré est supposée de contenir les informations importantes :

- Identifiant permettant à lier son dossier au remboursement à lui.
- Âge (par tranche d'âge possible si le produit d'assurance est segmenté par tranches d'âge).
- Sexe (supposé simple entre Homme et Femme).
- Adresse ou région d'habitation.
- Niveau du contrat que chaque assuré souscrit donc trois niveaux (il est possible de souscrire à un seul contrat).
- Exposition égale à la fraction de période qu'il reste dans le portefeuille et d'un période d'une année, ceci étant compris entre $[0; 1]$
- Salaire ou les autres informations liées au secteur d'activité, etc. Ses informations étant facultatives comme les assureurs ont des difficultés de demande plus des informations aux assurés.

La base de sinistres doit comporter des variables suivantes :

- Identifiant d'assuré dans la base assuré.

- Code d'acte de soin.
- Montant total des frais médicaux.
- Base de remboursement de la Sécurité Sociale.
- Taux de remboursement de la Sécurité Sociale.
- Niveau de couverture du contrat correspond au contrat souscrit par l'assuré.
- Montant remboursé par l'organisme complémentaire santé.

Les éléments listés au-dessus correspondent au portefeuille globalement géré par l'assureur, où trois niveaux de contrat sont proposés à tous assurés et les garanties sont invariables pour chaque assuré, seules les primes proposés peuvent être différentes d'un segment assuré à un autre.

3.2.2 Segmentation du portefeuille : loi de consommation

Lors de l'analyse de la base assurée, il est important et très souvent nécessaire de segmenter la population en segments homogènes liés à la région, au genre ou à des tranches d'âge (le regroupement par tranche d'âge permet de concentrer un plus grand nombre d'assurés que le décompte par âge spécifique) et de calculer empiriquement la répartition des niveaux de contrats complémentaires santé pour chaque segment de la population. Une telle démarche d'analyse des différences de taux de souscription et de consommation entre les segments est essentielle pour détecter l'anti-sélection dans le portefeuille d'assurés. Pour tenir compte des informations disponibles dans la base Damir, nous nous restreignons, dans cet exemple, au cas d'un segment défini par la base Damir, comprenant l'âge de l'assuré, le sexe de l'assuré et la région de résidence de l'assuré, car la consommation varie clairement avec les modalités de ces variables. Ensuite, on se restreint à notre exemple des contrats complémentaires santé répartis sur 3 niveaux de couverture pour le reste du mémoire : le niveau de couverture minimum, moyen et maximum.

Pour un cas plus général avec plusieurs segmentations, dans un portefeuille de N assurés, nous considérons A segments au total présents dans le portefeuille : $\forall a \in \{1,\dots,A\}, \text{Card}(a) = n_a$ et $\sum_{a=1}^{A} n_a = N$. Soit $a \in \{1,\dots,A\}$ un segment de la population avec un nombre d'assurés n_a appartenant au segment a, il faut répartir le maillage de segments de telle façon que la taille de chaque segment soit suffisamment grande pour appliquer la loi des grands nombres. Ainsi, les segments trop petits doivent être regroupés afin de ne pas exposer le cas de segmentations sur une même loi de probabilité très dispersée.

Soit $n \in \{1,..,N\}$ un assuré donné appartenant au segment $a \in \{1,..,A\}$. Nous notons S_n une variable aléatoire positive représentant l'ensemble des frais médicaux qu'il a consommé lors de son adhésion au contrat complémentaire santé pendant un an. Soit $\theta_n, \theta_n \in]0,1]$ une variable aléatoire décrivant l'exposition de l'assuré n au portefeuille et $C_n, C_n \in \{1,2,3\}$ son choix de niveau de couverture du contrat complémentaire santé. Nous faisons l'hypothèse que les trois variables S_n, θ_n, C_n sont iid et de même loi que S_a, θ_a, C_a pour tous les assurés du même segment $a : (S_n, \theta_n, C_n)_{n\in\{1,..,n_a\}} \overset{\text{iid}}{\sim} \mathcal{L}(S_a, \theta_a, C_a), \forall a \in \{1,..,A\}$. Nous supposons ainsi que S, θ sont intégrables pour tout a et qu'ils le sont conditionnement par C. L'idée derrière ces hypothèses est qu'elles nous permettent de modéliser la charge médicale moyenne sur toutes les postes de garantie d'un segment de population donné. La dépendance entre S et C (de même pour θ et C) est une hypothèse très importante pour souligner l'effet du choix stratégique du niveau de couverture sur la consommation (de même pour la durée de souscription) de chaque segment. Dans ce cas, on peut sous-entendre que S, θ sont des lois continues de mélange discret conditionnées à la variable C.

De plus, on suppose que le remboursement de l'organisme complémentaire d'un contrat de niveau couverture $i \in \{1,2,3\}$ (3 niveaux avec 1 correspondant au niveau entré de gamme, 2 pour moyen de

gamme et 3 pour haut de gamme) peut être représenté par des fonctions $g_i : \mathbb{R}^+ \to \mathbb{R}^+$ mesurables satisfaisant deux conditions suivantes :

- $\forall x \in \mathbb{R}^+, 0 \leq g_1(x) \leq g_2(x) \leq g_3(x) \leq x$.
- $\forall i \in \{1, 2, 3\}, g_i(S)$ est intégrable et $g_i(S)$ est intégrable conditionnellement à C.

La deuxième condition peut être déduite de la première par l'inégalité $E(g_i(S)) \leq E(S) < +\infty, \forall i \in \{1, 2, 3\}$, sachant que S est intégrable avec ou sans conditionnement par C. Les sinistres de l'assureur sont donc définis par la variable aléatoire $g_i(S)$ correspondant au remboursement à l'assuré souscrivant au contrat i sur une année pour leur consommation S.

Le portefeuille peut être dès maintenant modélisé par les segments de la population. Chaque segment a est donc représenté par une loi de consommation médicale liée à la variable S_a, à l'exposition du segment θ_a et à la tendance à souscrire aux trois niveaux de couverture C_a. La fin de cette section démontre donc la formule de deux coefficients disant coefficient d'anti-sélection du portefeuille.

3.2.3 La construction d'un coefficient d'anti-sélection

La difficulté d'obtenir les informations sur l'anticipation des dépenses de santé rend impossible la mesure directe de l'asymétrie d'information entre les assureurs et les assurés. Il est donc important de passer de la mesure aux conséquences produites par ce phénomène, incluant l'augmentation de la charge moyenne de sinistres réels par rapport à la charge moyenne ciblée pour certains contrats d'assurance, ou le niveau de couverture dans notre cas, car ils sont souscrits majoritairement par des personnes plus risquées ou anticipant plus de charges que prévu par les contrats. Cette différence de charges de sinistres est liée non seulement à l'anti-sélection, mais aussi au problème de l'aléa moral, où l'anticipation des soins médicaux augmente après que les assurés ont pris connaissance de leur niveau de couverture. En appliquant l'hypothèse de séparabilité entre l'anti-sélection et l'aléa moral, nous sommes convaincus que dans le cadre du contrat complémentaire santé individuel, l'effet de l'anti-sélection l'emporte sur celui de l'aléa moral, car les individus sont censés idéalement connaître leur niveau de risque et les garanties de chaque contrat proposé.

Selon une telle hypothèse posée pour pouvoir distinguer l'effet de l'anti-sélection, nous décidons que le rapport de sinistre réel et anticipé dont les assureurs ont dans leur portefeuille après une année est bel et bien une mesure du phénomène de l'anti-sélection. On en déduit ainsi que la variable S, en conditionnant à C, est la charge médicale totale d'un assuré qu'il a anticipé partiellement à dépenser, plus la partie aléatoire venant de sa morbidité pendant l'année.

Nous introduisons deux coefficients d'anti-sélection pour chaque niveau de couverture démontrés par la suite :

- Coefficient d'anti-sélection relatif du portefeuille pour le niveau de couverture $i \in \{1, 2, 3\}$ et le segment $a \in \{1, .., A\}$: $Coef\_relative_i^a$.
- Coefficient d'anti-sélection global hypothétique du portefeuille pour le niveau de couverture $i \in \{1, 2, 3\}$: $Coef\_global_i$.

Coefficient d'anti-sélection relatif du portefeuille

Ce coefficient relatif est une mesure locale de la différence de sinistre moyen entre un segment a donné et tous les segments pour un contrat $i \in 1, 2, 3$. Sans perte de généralité, on prend le contrat $i = 1$ et le segment a'. On note $Sin\_total_1^{a'}$ et $Expo\_total_1^{a'}$ respectivement le sinistre total et l'exposition totale des assurés dans le segment a' ayant choisi la couverture de niveau 1. On note $seg(a)$ l'ensemble des assurés du segment a, $Card(seg(a)) = n_a \forall a \in \{1, .., A\}$.

Soit $n \in seg(a'), (S_n^{a'}, C_n^{a'}, \theta_n^{a'}) \overset{\text{iid}}{\sim} \mathcal{L}(S_{a'}, C_{a'}, \theta_{a'})$, On calcule le sinistre moyen avec la formule de tarification par sinistre passé :

$$Sin\_moy_1^{a'} = \frac{Sin\_total_1^{a'}}{Expo\_total_1^{a'}} = \frac{\sum_{n\in seg(a')} g_1(S_n^{a'})\mathbb{1}_{[C_n^{a'}=1]}}{\sum_{n\in seg(a')} \theta_n^{a'}\mathbb{1}_{[C_n^{a'}=1]}} = \frac{\frac{1}{n_{a'}}\sum_{n\in seg(a')} g_1(S_n^{a'})\mathbb{1}_{[C_n^{a'}=1]}}{\frac{1}{n_{a'}}\sum_{n\in seg(a')} \theta_n^{a'}\mathbb{1}_{[C_n^{a'}=1]}}.$$

Soit $\%_N^a = \frac{n_a}{N}$ le pourcentage des assurés du segment a dans le portefeuille. On note bien que $n_a = \sum_{\text{assuré}} \mathbb{1}_{[\text{assuré}\in \text{seg}(a)]}$ est une variable aléatoire et que $\mathbb{1}_{[\text{assuré}\in \text{seg}(a)]} \overset{\text{iid}}{\sim} \mathcal{L}(a)$.

En supposant que la composition du segment de la population du portefeuille reste stable même si la taille du portefeuille augmente, au sens où quand $N \to +\infty$, nous avons $n_a \xrightarrow[N\to+\infty]{} +\infty$, pour tout $a \in \{1, \dots, A\}$. Dans ce cas, on obtient, par la loi des grands nombres :

$$\%_N^a = \frac{n_a}{N} \xrightarrow[N\to+\infty]{\mathcal{P}} \%^a.$$

Et dans le cas extrême où le portefeuille est la France entière, nous avons $\%^a$ bien approximé par le pourcentage du segment a dans la population française. En conséquence, on peut supposer que $Card(\text{seg}(a)) = n_a = \%_N^a \times N, \forall a \in \{1, \dots, A\}$. On réécrit la formule de $Sin\_moy_1^a$ en utilisant la propriété de convergence en probabilité qui s'applique sur les applications continues, et le produit de deux suites de variables aléatoires convergeant en probabilité pour le quotient de deux suites convergentes en probabilité, issues de la loi des grands nombres :

$$Sin\_moy_1^{a'} = \frac{\frac{1}{\%_N^{a'}\times N}\sum_{n\in seg(a')} g_1(S_n^{a'})\mathbb{1}_{[C_n^{a'}=1]}}{\frac{1}{\%_N^{a'}\times N}\sum_{n\in seg(a')} \theta_n^{a'}\mathbb{1}_{[C_n^{a'}=1]}} \xrightarrow[N\to+\infty]{\mathcal{P}} \frac{E[g_1(S^{a'})\mathbb{1}_{[C^{a'}=1]}]}{E[\theta^{a'}\mathbb{1}_{[C^{a'}=1]}]}.$$

On cherche à développer davantage certains termes dans la formule ci-dessus :

Pour $E[g_1(S^{a'})\mathbb{1}_{[C^{a'}=1]}]$, par définition de l'espérance conditionnelle, on obtient l'égalité :

$$E[g_1(S^{a'})\mathbb{1}_{[C^{a'}=1]}] = E[g_1(S^{a'})|C^{a'}=1]\mathbb{P}(C^{a'}=1).$$

De même façon, on obtient :

$$E[\theta^{a'}\mathbb{1}_{[C^{a'}=1]}] = E[\theta^{a'}|C^{a'}=1]\mathbb{P}(C^{a'}=1).$$

Comme $E[g_1(S^{a'})|C^{a'}=1]$ est l'espérance de remboursement moyen de l'assureur pour l'assuré de segment a' ayant choisi la couverture 1, nous pourrons l'approximer par la loi de grande nombre :

$$E[g_1(S^{a'})|C^{a'}=1] \approx \frac{\sum_{n\in seg(a')} g_1(S_n^{a'})\mathbb{1}_{[C_n^{a'}=1]}}{n_1^{a'}} = \frac{Sin\_total_1^{a'}}{n_1^{a'}}.$$

Où $n_1^{a'}$ est le nombre de personnes de segment a' ayant souscrit le contrat 1. De même façon, nous calculons $E[\theta^{a'}|C^{a'}=1]$ en utilisant la loi de grande nombre :

$$E[\theta^{a'}|C^{a'}=1] \approx \frac{\sum_{n\in seg(a')} \theta^{a'}\mathbb{1}_{[C_n^{a'}=1]}}{n_1^{a'}} = \frac{Expo\_total_1^{a'}}{n_1^{a'}}.$$

En remplaçant $E[g_1(S^{a'})\mathbb{1}_{[C^{a'}=1]}]$ et $E[\theta^{a'}\mathbb{1}_{[C^{a'}=1]}]$ dans la formule pour calculer $Sin\_moy_1^a$, on obtient :

$$Sin\_moy_1^{a'} \xrightarrow[N\to+\infty]{\mathcal{P}} \frac{E[g_1(S^{a'})|C^{a'}=1]\mathbb{P}(C^{a'}=1)}{E[\theta^{a'}|C^{a'}=1]\mathbb{P}(C^{a'}=1)} = \frac{E[g_1(S^{a'})|C^{a'}=1]}{E[\theta^{a'}|C^{a'}=1]}.$$

Autrement dit, $Sin\_moy_i^a$ tendra en probabilité vers le rapport du sinistre moyen des assurés du segment a' ayant choisi la couverture i et l'exposition moyenne de ces mêmes personnes.

Maintenant, nous nous intéressons à la quantité appelée sinistre moyen sur tous les segments du niveau de couverture 1 : Sin_moy_1. Il est calculé en utilisant les mêmes arguments que les calculs ci-dessus sur la stabilité de la composition du segment et l'approximation par la loi des grands nombres.

$$Sin\_moy_1 = \frac{Sin\_total_1}{Expo\_total_1} = \frac{\sum_{a=1}^{A}\sum_{n\in seg(a)} g_1(S_n^a)\mathbb{1}_{[C_n^a=1]}}{\sum_{a=1}^{A}\sum_{n\in seg(a)} \theta_n^a\mathbb{1}_{[C_n^a=1]}} = \frac{\sum_{a=1}^{A}\frac{\%_N^a}{\%_N^a\times N}\sum_{n\in seg(a)} g_1(S_n^a)\mathbb{1}_{[C_n^a=1]}}{\sum_{a=1}^{A}\frac{\%_N^a}{\%_N^a\times N}\sum_{n\in seg(a)} \theta_n^a\mathbb{1}_{[C_n^a=1]}}$$

$$\xrightarrow[N\to+\infty]{\mathcal{P}} \frac{\sum_{a=1}^{A}\%^a E[g_1(S^a)\mathbb{1}_{[C^a=1]}]}{\sum_{a=1}^{A}\%^a E[\theta^a\mathbb{1}_{[C^a=1]}]} = \frac{\sum_{a=1}^{A}\%^a E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1)}{\sum_{a=1}^{A}\%^a E[\theta^a|C^a=1]\mathbb{P}(C^a=1)}$$

$$\approx \frac{\sum_{a=1}^{A}\%_N^a\frac{Sin\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)}{\sum_{a=1}^{A}\%_N^a\frac{Expo\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)}.$$

Par construction du coefficient anti-sélection relative au segment a' pour le niveau de couverture 1 que nous notons $Coef\_relative_i^{a'}$, nous calculons cette quantité par la formule :

$$Coef\_relative_1^{a'} = \frac{Sin\_moy_1^{a'}}{Sin\_moy_1} \xrightarrow[N\to+\infty]{\mathcal{P}} \frac{\frac{E[g_1(S^{a'})|C^{a'}=1]}{E[\theta^{a'}|C^{a'}=1]}}{\frac{\sum_{a=1}^{A}\%^a E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1)}{\sum_{a=1}^{A}\%^a E[\theta^a|C^a=1]\mathbb{P}(C^a=1)}} \approx \frac{\frac{Sin\_total_1^{a'}}{Expo\_total_1^{a'}}}{\frac{\sum_{a=1}^{A}\%_N^a\frac{Sin\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)}{\sum_{a=1}^{A}\%_N^a\frac{Expo\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)}}.$$

Pour un portefeuille de grande taille (avec un N très grand), nous pouvons bien approximer ce coefficient d'anti-sélection. Cette quantité nous indique si les assurés d'un segment consomment en moyenne plus ou moins que tous ceux qui ont pris le même contrat. En l'appliquant aux profils des tranches d'âge différents, nous pourrions observer la différence systématique entre les jeunes et les vieux, où les vieux consomment généralement plus que les jeunes. Cependant, cette différence semble relativement faible lors de l'application à d'autres variables que la tranche d'âge, puisque les autres segmentations que la tranche d'âge n'ont pas autant de grands écarts de sinistralité entre les modalités. Par conséquent, mesurer les coefficients relatifs n'apporte pas toutes les informations nécessaires. C'est la raison pour laquelle il faut une autre mesure prenant en compte tous les profils de risque, que l'on appelle le coefficient global hypothétique du portefeuille.

Coefficient d'anti-sélection global hypothétique du portefeuille

Ce coefficient, comme son nom l'indique, prend en compte tous les profils au sein d'un segment pour donner une mesure globale sur tous les assurés du portefeuille. Nous le définissons comme le rapport entre le sinistre moyen causé par tous ceux qui ont choisi la couverture i et le sinistre moyen causé si tous les assurés avaient hypothétiquement choisi la même couverture. La motivation derrière cela repose sur l'hypothèse que les dépenses médicales aléatoires liées à la morbidité de la population sont couvertes même par les contrats de niveau de couverture minimum, alors que la différence d'anticipation est systématiquement liée à la différence de qualité de soins utilisés et donc aux frais médicaux élevés. Nous pourrons effectuer la comparaison entre le sinistre moyen mutualisé sur tous les profils et le sinistre moyen mutualisé sur le profil de type luxueux, car cela permet de quantifier l'asymétrie d'information sur le choix stratégique. De plus, les assureurs disposent de l'historique des frais médicaux et des remboursements de la Sécurité Sociale, donc ils pourraient facilement effectuer les calculs de remboursement à un niveau de couverture donné.

Comme dans la section précédente, nous prenons la couverture 1 sans perte de généralité. Notons $Sin\_portef\_total_1$ et $Expo\_portef\_total_1$ respectivement le sinistre total et l'exposition totale sur le contrat 1 si tous les assurés dans le portefeuille avaient hypothétiquement choisi le même niveau de couverture 1. Nous conservons le même calcul de remboursement que ceux ayant choisi la couverture 1 et réappliquons le niveau de couverture 1 à toutes les dépenses médicales causées par ceux qui n'ont pas choisi la couverture 1. Nous procédons ensuite aux calculs du sinistre moyen du portefeuille entier si tous les assurés dans le portefeuille avaient hypothétiquement choisi le même niveau de couverture 1 :

$$Sin\_portef\_moy_1 = \frac{Sin\_portef\_total_1}{Expo\_portef\_total_1} = \frac{\sum_{n=1}^{N} g_1(S_n)}{\sum_{n=1}^{N} \theta_n}$$

$$= \frac{\sum_{a=1}^{A} \sum_{n\in seg(a)} (g_1(S_n^a)\mathbb{1}_{[C_n^a=1]} + g_1(S_n^a)\mathbb{1}_{[C_n^a=2]} + g_1(S_n^a)\mathbb{1}_{[C_n^a=3]})}{\sum_{a=1}^{A} \sum_{n\in seg(a)} (\theta_n^a\mathbb{1}_{[C_n^a=1]} + \theta_n^a\mathbb{1}_{[C_n^a=2]} + \theta_n^a\mathbb{1}_{[C_n^a=3]})}$$

$$= \frac{\sum_{a=1}^{A} \frac{\%_N^a}{\%_N^a \times N} \sum_{n\in seg(a)} (g_1(S_n^a)\mathbb{1}_{[C_n^a=1]} + g_1(S_n^a)\mathbb{1}_{[C_n^a=2]} + g_1(S_n^a)\mathbb{1}_{[C_n^a=3]})}{\sum_{a=1}^{A} \frac{\%_N^a}{\%_N^a \times N} \sum_{n\in seg(a)} (\theta_n^a\mathbb{1}_{[C_n^a=1]} + \theta_n^a\mathbb{1}_{[C_n^a=2]} + \theta_n^a\mathbb{1}_{[C_n^a=3]})}$$

$$\xrightarrow[N\to+\infty]{\mathcal{P}} \frac{\sum_{a=1}^{A} \%^a (E[g_1(S^a)\mathbb{1}_{[C^a=1]}] + E[g_1(S^a)\mathbb{1}_{[C^a=2]}] + E[g_1(S^a)\mathbb{1}_{[C^a=3]}])}{\sum_{a=1}^{A} \%^a (E[\theta^a\mathbb{1}_{[C^a=1]}] + E[\theta^a\mathbb{1}_{[C^a=2]}] + E[\theta^a\mathbb{1}_{[C^a=3]}])}$$

$$= \frac{\sum_{a=1}^{A} \%^a (E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1) + E[g_1(S^a)|C^a=2]\mathbb{P}(C^a=2) + E[g_1(S^a)|C^a=3]\mathbb{P}(C^a=3))}{\sum_{a=1}^{A} \%^a (E[\theta^a|C^a=1]\mathbb{P}(C^a=1) + E[\theta^a|C^a=2]\mathbb{P}(C^a=2) + E[\theta^a|C^a=3]\mathbb{P}(C^a=3))}$$

$$\approx \frac{\sum_{a=1}^{A} \%_N^a (\frac{Sin\_total_1^a}{n_1^a}\mathbb{P}(C^a=1) + \frac{Sin\_total_{2\to1}^a}{n_2^a}\mathbb{P}(C^a=2) + \frac{Sin\_total_{3\to1}^a}{n_3^a}\mathbb{P}(C^a=3))}{\sum_{a=1}^{A} \%_N^a (\frac{Expo\_total_1^a}{n_1^a}\mathbb{P}(C^a=1) + \frac{Expo\_total_2^a}{n_2^a}\mathbb{P}(C^a=2) + \frac{Expo\_total_3^a}{n_3^a}\mathbb{P}(C^a=3))}.$$

Dans le dernier calcul, en utilisant la loi des grands nombres, nous calculons $Sin\_total_{2\to1}^a$ et $Sin\_total_{3\to1}^a$ comme les sinistres totaux si les assurés du segment a qui ont souscrit aux contrats de niveau de couverture 2 et 3 avaient effectivement souscrit au contrat de niveau de couverture 1. Cela repose sur l'hypothèse que les assureurs ont accès à toutes les informations de la Sécurité Sociale.

Le coefficient global $Coef\_global_1$ s'est exprimé comme suivant :

$$Coef\_global_1 = \frac{Sin\_moy_1}{Sin\_portef\_moy_1}.$$

En passant à la limite lorsque N tend vers l'infini, nous obtenons :

$$Coef\_global_1 \xrightarrow[N\to+\infty]{\mathcal{P}} \frac{\frac{\sum_{a=1}^{A} \%^a E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1)}{\sum_{a=1}^{A} \%^a E[\theta^a|C^a=1]\mathbb{P}(C^a=1)}}{\frac{\sum_{a=1}^{A} \%^a (E[g_1(S^a)|C^a=1]\mathbb{P}(C^a=1)+E[g_1(S^a)|C^a=2]\mathbb{P}(C^a=2)+E[g_1(S^a)|C^a=3]\mathbb{P}(C^a=3))}{\sum_{a=1}^{A} \%^a (E[\theta^a|C^a=1]\mathbb{P}(C^a=1)+E[\theta^a|C^a=2]\mathbb{P}(C^a=2)+E[\theta^a|C^a=3]\mathbb{P}(C^a=3))}}$$

$$\approx \frac{\frac{\sum_{a=1}^{A} \%_N^a \frac{Sin\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)}{\sum_{a=1}^{A} \%_N^a \frac{Expo\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)}}{\frac{\sum_{a=1}^{A} \%_N^a (\frac{Sin\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)+\frac{Sin\_total_{2\to1}^a}{n_2^a}\mathbb{P}(C^a=2)+\frac{Sin\_total_{3\to1}^a}{n_3^a}\mathbb{P}(C^a=3))}{\sum_{a=1}^{A} \%_N^a (\frac{Expo\_total_1^a}{n_1^a}\mathbb{P}(C^a=1)+\frac{Expo\_total_2^a}{n_2^a}\mathbb{P}(C^a=2)+\frac{Expo\_total_3^a}{n_3^a}\mathbb{P}(C^a=3))}}.$$

Ce coefficient dépend fortement du grand numérateur qui relie trois principaux facteurs : la sinistralité, le nombre d'assurés par segment et la probabilité de souscription au contrat d'assurance. Le

numérateur est d'autant plus grand si le contrat contient principalement des profils à hautes dépenses médicales en moyenne, avec une forte anticipation de souscription et peu d'assurés dans les segments de base dépense en moyenne. En raison de l'effet de mutualisation des différents profils, l'impact de la forte consommation médicale par anticipation devrait effectivement être moins important pour tous les assurés du portefeuille (le grand dénominateur). Par conséquent, en découlant de notre hypothèse selon laquelle la différence principale de dépenses est liée aux préférences médicales, on observerait normalement de grands coefficients d'anti-sélection (supérieurs à 1) sur les contrats de haut niveau de couverture et de faibles coefficients (inférieurs à 1) sur les contrats de bas niveau de couverture.

Les probabilités $\mathbb{P}(C^a = 1)$, $\mathbb{P}(C^a = 2)$, $\mathbb{P}(C^a = 3)$ sont en effet des probabilités agrégées de choix de couverture pour le segment a. Elles représentent donc la demande d'assurance de chaque segment de la population. Elles restent constantes par rapport à la taille du portefeuille N, mais elles sont influencées par d'autres facteurs liés aux préférences des assurés au moment de la souscription. Guidés par les préférences de consommation, nous nous trouvons dans une application importante des modèles de choix discrets, à savoir la prédiction des changements de la demande d'assurance lors de l'introduction de modifications des attributs liés au contrat d'assurance. Un tel changement affectera également la relation entre les variables S et θ et C, mais dans le but d'analyser le comportement d'anti-sélection simple, on peut toujours utiliser l'approximation par la loi des grands nombres sur le sinistre passé.

3.3 Génération du portefeuille pour l'étude

L'absence d'un portefeuille d'assurance peut représenter un défi de taille pour notre étude car nous ne sommes pas en mesure d'observer les vrais comportements des assurés qui sont très subtils et complexes. Cependant, cette lacune peut être surmontée en générant notre propre base d'assurés à partir de données statistiques nationales et en tirant parti des avis d'experts. En recueillant des informations précises sur les tendances nationales en matière d'assurance, les besoins du marché et les comportements des consommateurs, cette approche proactive nous permettra de construire progressivement notre portefeuille d'assurance.

Comme discuté précédemment, l'étude de l'anti-sélection en assurance complémentaire santé individuelle peut être abordée de manière plus générale en évaluant un portefeuille de contrats individuels. Afin de garantir l'applicabilité générale de cette approche, nous supposons que notre portefeuille d'assurance, assimilé à la base Damir que nous avons traitée, comporte des contrats complémentaires santé généraux offerts à tous les secteurs d'activité, à toutes les tranches d'âge (sauf pour ceux de moins de 20 ans) et dans toutes les régions de la France. Nous modéliserons ainsi trois niveaux de couverture correspondant aux trois produits les plus représentatifs dans un portefeuille d'assurance complémentaire santé (entrée de gamme, gamme intermédiaire et haut de gamme). Les tarifs proposés aux assurés sont segmentés en fonction de trois variables discriminantes de la base Damir, afin de fournir les tarifs les plus adaptés au profil de l'assuré. De plus, nous supposons que l'assureur propose tout d'abord un tarif technique, c'est-à-dire une prime technique pour chaque niveau de couverture. Cette partie de ce chapitre vise à simuler des portefeuilles fictifs dans lesquels nous définirons la structure de comportement des assurés. Pour plus de rigueur, ce modèle sera de type préférence révélée, dont l'application pratique est fortement recommandée pour une base de données ressemblant à nos bases fictives. Cependant, il convient d'être prudent lors de l'inclusion de variables hypothétiques, car cela pourrait créer des incohérences avec la réalité. Il est donc essentiel de choisir les variables à considérer dans le modèle avec discernement et de s'appuyer sur les principaux résultats de la littérature abordée précédemment.

3.3.1 Méthode de génération du portefeuille

Cette section a pour objectif d'illustrer l'application d'une méthode d'analyse basée sur des modèles de choix discrets à un portefeuille d'assurance. Elle comprend trois phases : la génération des données, l'entraînement des modèles, la validation de leur ajustement aux données, ainsi que la mise en œuvre d'une politique d'ajustement des primes pour atténuer le problème de l'anti-sélection.

La phase de génération des données requiert une expertise approfondie en matière de comportement des assurés ainsi que le déploiement d'un modèle visant à établir une structure de corrélation aussi cohérente que possible entre les variables prises en compte. Grâce aux données statistiques nationales de la DREES en 2016 sur les contrats complémentaires santé individuels (présentées dans Barlet et al. (2019)), définis en trois niveaux de gamme (entrée de gamme, gamme moyenne et haut de gamme), nous disposons d'une répartition générique de la part des niveaux de couverture générale (33% pour l'entrée de gamme, 56% pour la gamme moyenne et 11% pour le haut de gamme) que nous ajustons en fonction de la segmentation de la population, en nous basant sur les avis d'experts en matière de consommation médicale en France, notamment via la base Damir, ainsi que sur des études portant sur la couverture de l'assurance complémentaire santé en France.

Une fois cette base établie, nous cherchons à intégrer les effets du revenu/niveau de vie des individus ainsi que leur attitude envers les couvertures des contrats, en utilisant les données démographiques de l'INSEE. Étant donné que toutes ces variables sont indépendantes les unes des autres, nous déterminons une structure mathématique de type utilité aléatoire (abordée dans le prochain chapitre) qui impose une corrélation entre les variables dans les données. Il y aura au total deux portefeuilles fictifs, définis sur deux structures de corrélation de plus en plus complexes, mais correspondant de plus en plus à la réalité.

3.3.2 Génération des portefeuilles fictifs : Intégration des données externes

Pour rapprocher un portefeuille d'assurance de celui détenu par un assureur, la base d'assurés simulée est constituée d'une base de données individuelle englobant diverses variables socio-démographiques cruciales telles que l'âge, le sexe, et la région de résidence. Ces variables revêtent une grande importance, car de nombreux assureurs utilisent une segmentation tarifaire en fonction de ces critères. De plus, cette base contient des informations relatives au choix de la couverture de la complémentaire santé, distinguant ainsi trois niveaux de couverture : minimum, moyen ou maximum. Il est à noter que chaque assuré ne peut opter que pour une seule catégorie de couverture. Les assurés de moins de 20 ans ont été exclus de cette base, car il est supposé que la plupart d'entre eux sont rattachés à la complémentaire santé de leurs parents. En ce qui concerne les primes d'assurance, elles ont été obtenues à partir de la base Damir, impliquant l'utilisation des trois variables démographiques clés de cette base, à savoir l'âge (divisé en 7 tranches d'âge), le sexe (homme ou femme), et la région de résidence (répartie en 13 régions distinctes). Cette base d'assurés simulée servira de fondement essentiel pour l'analyse et la modélisation de l'anti-sélection en assurance complémentaire santé. Nous avons décidé de générer une base d'assurés de 50 000 individus. Cette taille a été choisie afin de garantir de bonnes propriétés statistiques tout en évitant de créer une base de données excessivement volumineuse.

Avant de générer le portefeuille, il faut donc établir les relations entre les variables dans la base. Dans ce cas, nous déterminons les probabilités qu'un individu soit inclus dans la base, c'est-à-dire la proportion des assurés dans le portefeuille après toutes les souscriptions. Pour chaque individu considéré, son choix de couverture d'assurance sera simulé directement à partir de nos probabilités de souscrire aux trois niveaux de contrat, dérivées des statistiques nationales et adaptées à leur segment de population.

Proportion d'assurés dans la base

Puisque le découpage de la segmentation est identique à celui de la base Open Damir, nous utilisons la base démographique INSEE présentée à la section 2.2.3, qui contient trois variables : l'âge, le sexe, et la région, soit 7 tranches d'âge, 2 sexes et 13 régions, pour un total de 182 segments. En divisant la population de chaque segment par la somme de la population, nous définissons la répartition des assurés de notre portefeuille, similaire à celle de la base Damir. Afin d'obtenir un nombre de données suffisant pour l'estimation des modèles sans le surcharger, nous conservons 50 000 assurés, donc 50 000 lignes.

Association des statistiques nationales au profil d'assuré particulier

Comme nos statistiques nationales datent de 2016, nous utilisons la pyramide d'âge de la population française en 2016 comme référence. Étant donné que la proportion des contrats dans les trois types de couverture est en moyenne la même pour l'ensemble de la population, il est logique de l'associer à la répartition démographique du groupe d'âge le plus représenté en France en 2016. En se basant sur la pyramide des âges en 2016 (3.1), nous pouvons facilement observer que la tranche d'âge de 40 à 49 ans était la plus importante en termes de population. Nous associons donc les statistiques nationales à la tranche d'âge de 40 à 49 ans, sans distinction entre hommes et femmes. En prenant également en compte la prédominance de l'Île-de-France en tant que région la plus peuplée, nous affectons finalement cette proportion au profil des assurés âgés de 40 à 49 ans en Île-de-France. Ensuite, nous ajustons ces probabilités de base pour chaque segment considéré dans le portefeuille.

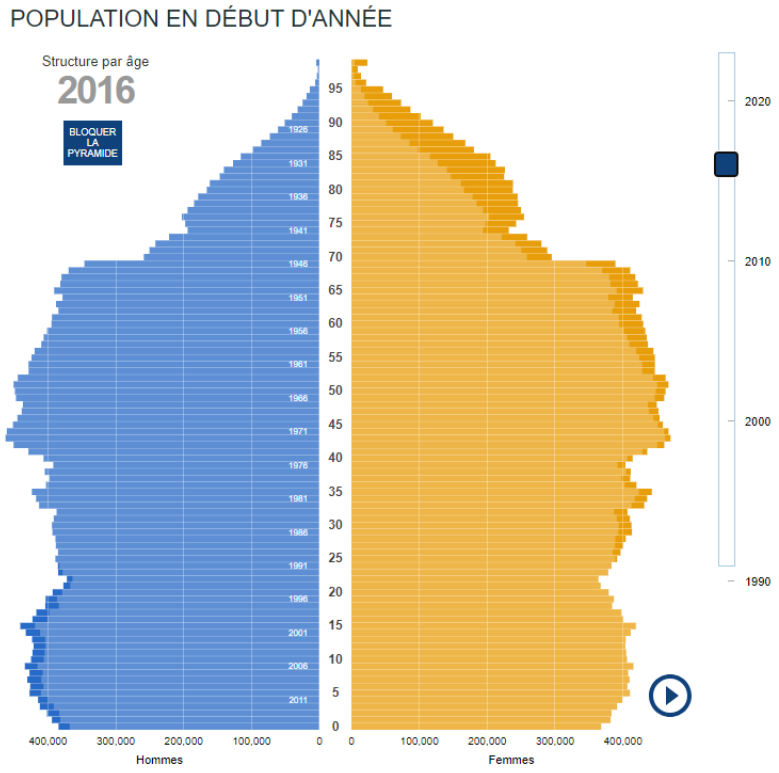


FIGURE 3.1 : Pyramide d'âge de la France entière 2016

Source : INSEE (sans date)

Proportion de choix en fonction du niveau de sinistre dans la base Damir

Grâce à des analyses sur la sinistralité générale de la base Damir au chapitre 2, nous tirons les enseignements suivants :

- La sinistralité totale de chaque couverture (c'est-à-dire la prime moyenne par personne pour l'ensemble des postes dans la grille de garantie) a tendance à augmenter avec l'âge de façon presque linéaire, voire exponentielle. On peut donc dire qu'en croissant avec l'âge, les probabilités de souscrire aux couvertures moyenne et maximale augmentent de manière hétérogène : jusqu'à la tranche d'âge de 60 ans, il est plus probable de souscrire à la couverture moyenne qu'à la couverture maximale, et à l'inverse pour les tranches d'âge de 70 ans, 80 ans et plus, où la couverture maximale présente moins d'écart de sinistralité à partir de la tranche d'âge de 60 ans.

- En ce qui concerne les régions, nous constatons une disparité de sinistralité par région. Les régions les plus urbanisées ont tendance à consommer davantage que les régions rurales du centre de la France. Les probabilités de souscription aux différentes couvertures seront donc ajustées en fonction de la région, où les régions à faible sinistralité auront une probabilité plus élevée de souscrire à la couverture minimale, tandis que pour les régions où la sinistralité est supérieure à celle de l'Île-de-France, la probabilité tendra davantage vers les couvertures moyenne ou maximale.

- En ce qui concerne le sexe, on observe une différence significative de sinistralité entre les deux sexes pour les tranches d'âge jeunes (moins de 50 ans), mais cette différence diminue avec l'âge. Nous ajustons donc les probabilités en augmentant (ou diminuant) les probabilités de souscrire aux contrats de couverture moyenne et maximale pour les femmes (ou les hommes) par rapport aux probabilités en fonction de la région et de l'âge. La différence entre les probabilités pour les hommes et les femmes diminue significativement avec l'âge.

Toutes les analyses ci-dessus résultent de l'information tirée de la base Open Damir ainsi que de l'avis d'experts sur le sujet. Nous avons ainsi procédé manuellement à une série de déformations des statistiques nationales en fonction de l'âge, puis de la région, et enfin du sexe. Les probabilités retenues sont présentées aux graphiques 3.2, 3.3, 3.4.

La base d'assurés en intégrant l'effet du revenu et des caractéristiques du contrat

En évaluant les options de couverture choisies par les assurés, il est essentiel de considérer l'impact que la souscription d'une couverture complémentaire santé peut avoir sur le capital monétaire des individus. C'est ce qui peut potentiellement influencer profondément les assurés lorsqu'ils envisagent les options à souscrire après Legal (2008). En réalité, cette information n'est pas toujours disponible pour les assureurs lors de la souscription, car elle est souvent jugée non pertinente pour les produits complémentaires santé. Par conséquent, cela peut entraîner une forte asymétrie d'information entre l'assuré et l'assureur. Une façon simple de surmonter cette difficulté est de considérer que les individus disposent d'un capital monétaire moyen par classe de population (dans notre cas, les classes correspondent au maillage mentionné ci-dessus), correspondant à leur revenu net ou au niveau de vie des ménages de la même classe, obtenu à partir de données Open Data. Le niveau de vie, également appelé "revenu disponible équivalent" par Insee (2021), se calcule en prenant le revenu total du ménage et en le divisant par un indicateur appelé "nombre d'unités de consommation" (UC). Tous les individus au sein d'un même ménage ont le même niveau de vie.

Pour illustrer notre notion de niveau de vie moyen par classe, l'INSEE dispose de données sur le niveau de vie des personnes en France de 1996 à 2019 (figure 3.5). Nous choisissons les niveaux de vie en 2019, car ce sont les données les plus récentes. Bien que ce niveau de vie soit commun à tous

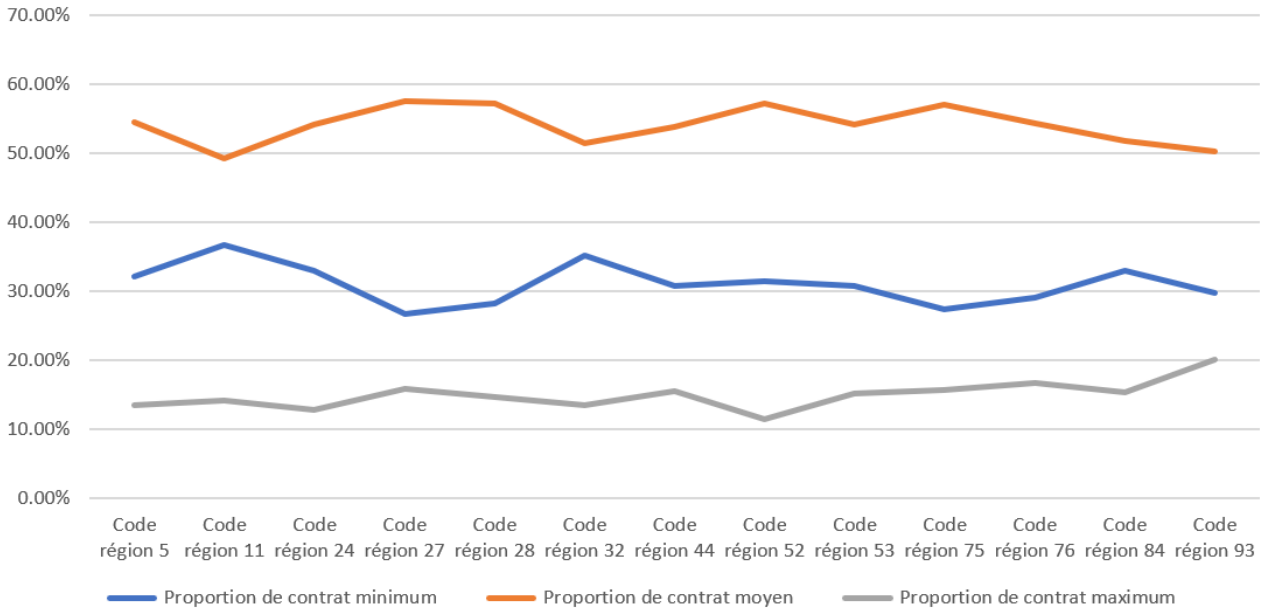


FIGURE 3.2 : Proportion des choix de niveaux de couverture en fonction de la région

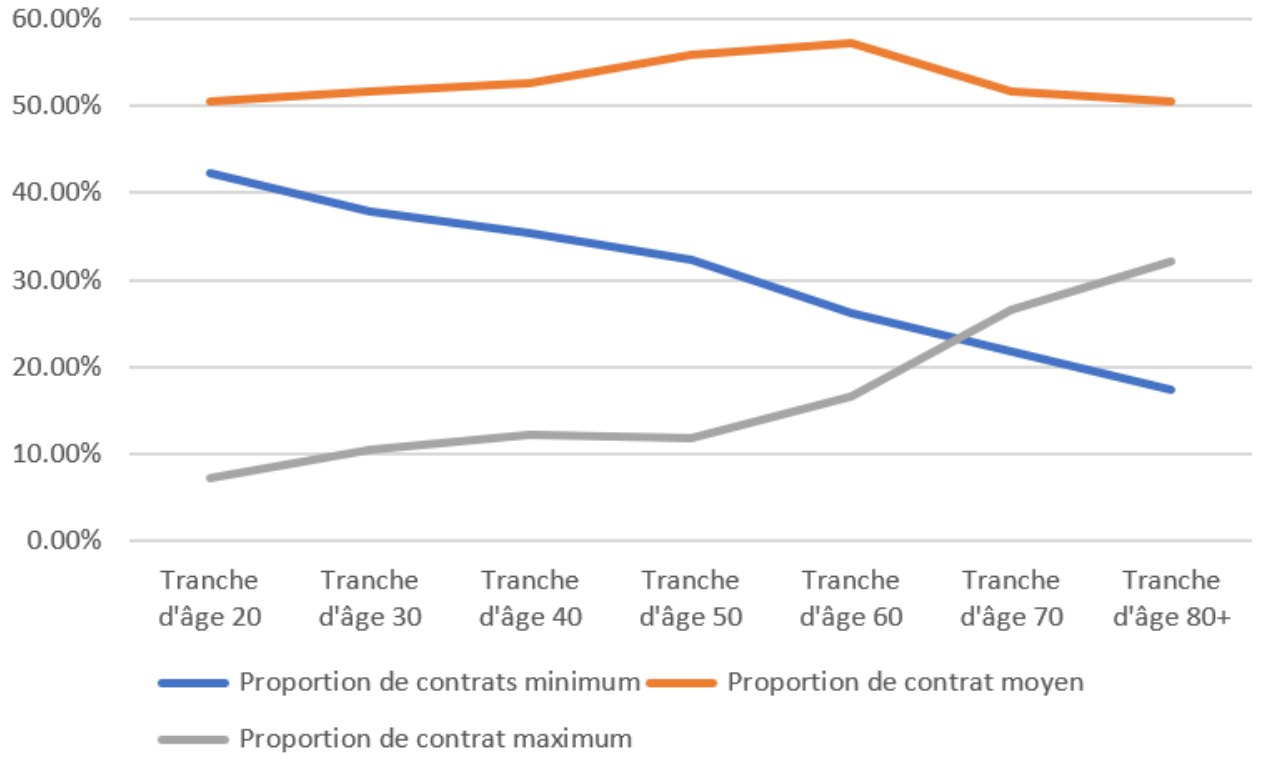


FIGURE 3.3 : Proportion des choix de niveaux de couverture en fonction de l'âge

les individus en France, pour affiner les niveaux de vie par âge et par sexe, nous utilisons une autre source d'information : l'INSEE - Salaire net horaire moyen selon la catégorie socioprofessionnelle, le sexe et l'âge en 2019 (voir INSEE (2023c). Nous procédons comme suit :

- Nous calculons le revenu moyen de toutes les régions en tenant compte de la population de

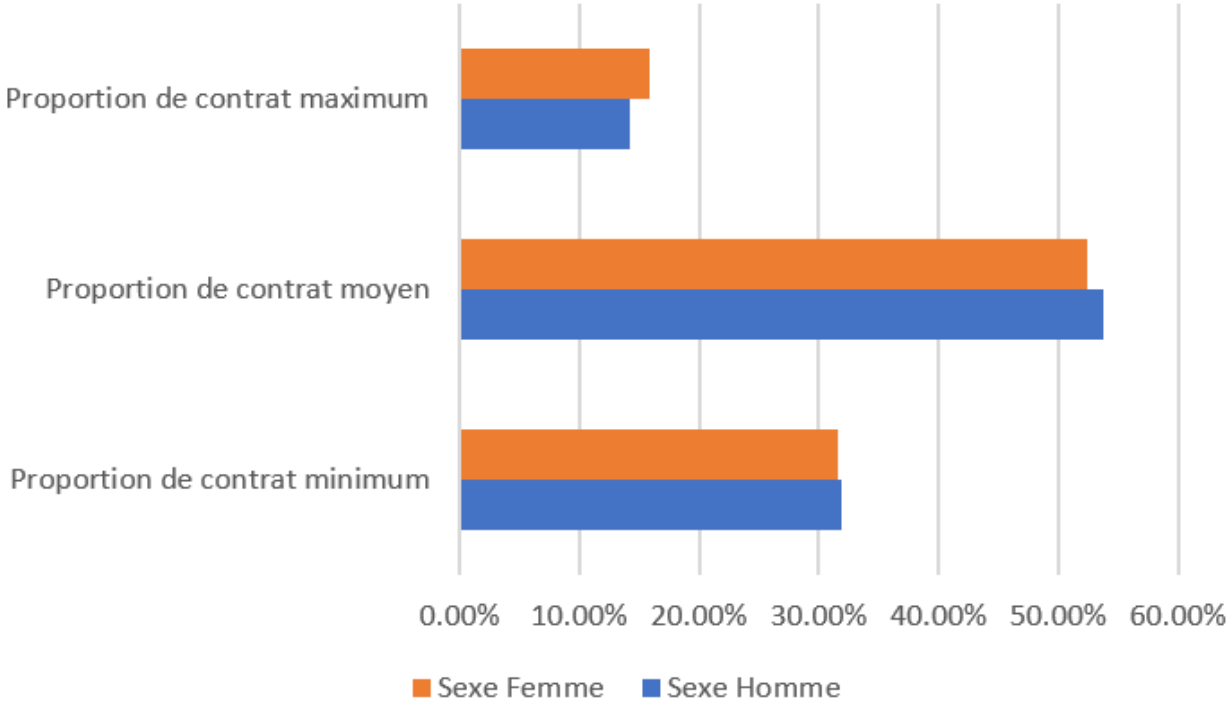


FIGURE 3.4 : Proportion des choix de niveaux de couverture en fonction du sexe

chaque région, puis déduisons le niveau de vie moyen national pour chaque région en utilisant le rapport entre le salaire net moyen de ces régions et celui de la France entière, en regroupant les départements d'outre-mer (DOM-TOM) ainsi que la région de la Côte-d'Azur et de la Corse (3.6).

- Ensuite, nous multiplions les niveaux de vie par tranche d'âge et par région obtenus avec les poids de revenu homme/femme au sein de la même région (3.7).

Ainsi, nous obtenons les revenus correspondant au maillage de la base Open Damir, ce qui rend possible l'application de la variable salaire au modèle de choix discret. Chaque individu possède un niveau de vie moyen du segment (tranche d'âge, sexe, région) dans la nouvelle base.

La première base d'assurés obtenue précédemment est basée uniquement sur les déformations des probabilités de souscription suivant des proxies de risque observables (profil de consommation médicale par âge, sexe et région), alors qu'elle n'est pas liée à l'effet de salaire (c'est-à-dire les probabilités de souscription déformées selon le profil de salaire). Dans ce cas, le risque pour la santé prévaut donc sur le revenu, où peu importe comment les revenus varient, leur poids sur les probabilités de choix reste négligeable. Comment peut-on prendre en compte cette variable de manière significative dans les données ? La réponse à cette question est par la simulation contrefactuelle. Pour pouvoir le faire, nous avons choisi au préalable une structure de comportement qui est présentable mathématiquement, et donc de définir une structure de modèle logit multinomial (présentée au chapitre 4) prenant en compte cette variable de revenu (avec un coefficient assez grand) sans que les comportements (probabilités de choix de couverture) des assurés dans la première base ne soient changés de manière considérable. La présentation du modèle utilisé pour la génération de cette deuxième base sera faite lors de l'entraînement du modèle au chapitre 5 (section 5.1.2).

Des études sur la couverture santé, telles que celles menées par FRANC et al. (2010a) et LEGAL (2008), ont montré que la connaissance des taux de remboursement des postes présentant une forte

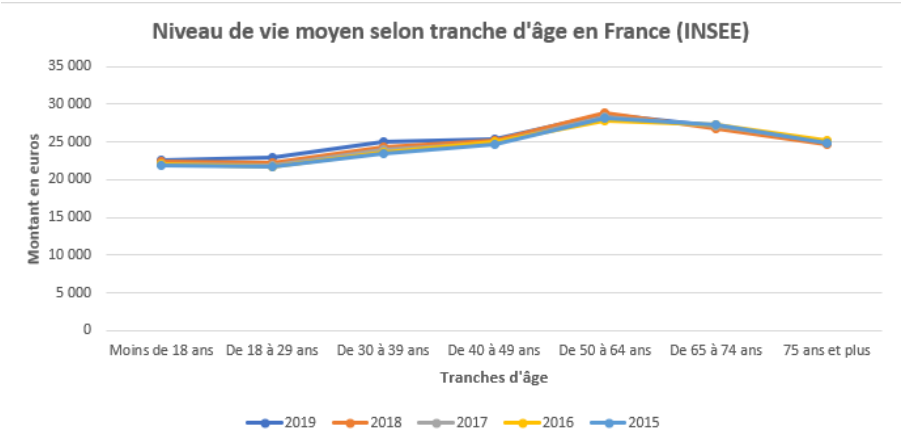


FIGURE 3.5 : Niveau de vie moyen de la France 2015-2019

Source : INSEE (2023b)

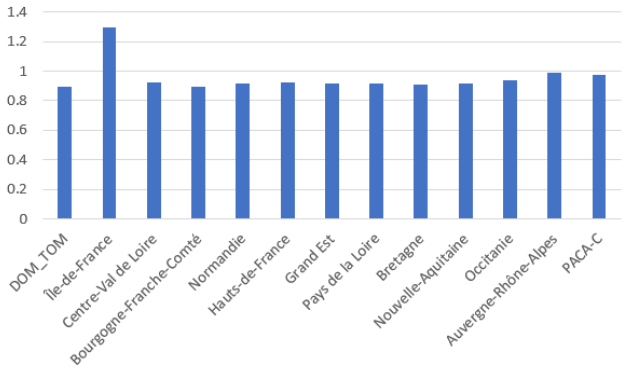


FIGURE 3.6 : Coefficient de disparité de salaire par région en France

Source : INSEE (2023c)

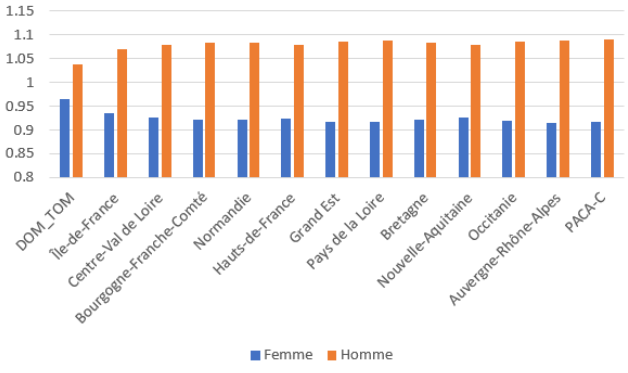


FIGURE 3.7 : Coefficient de disparité de salaire par sexe pour chaque région en France

Source : INSEE (2023c)

anti-sélection impacte le choix de couverture des individus lors de la souscription. Cette source d'anti-sélection, influencée non seulement par le taux de morbidité des assurés, mais aussi par leur préférence

en matière de qualité des soins médicaux face au même risque, devrait être essentiellement prise en compte dans notre cas et nous permettrait de le distinguer des autres secteurs de l'assurance. Le modèle générant la deuxième base d'assurés prendra en compte le rapport de prime technique d'assurance de chaque couverture par rapport à celui de la couverture moyenne, comme un proxy de la perception par les assurés du remboursement moyen des options, sous l'hypothèse qu'ils connaissent bien leur risque. Il est à noter que les probabilités lors de la génération du choix individuel dans les deux bases ne comportent pas d'aléa découlant de l'hétérogénéité des individus dans la base, car tous les individus de même profil ont le même comportement, c'est-à-dire une incitation à souscrire.

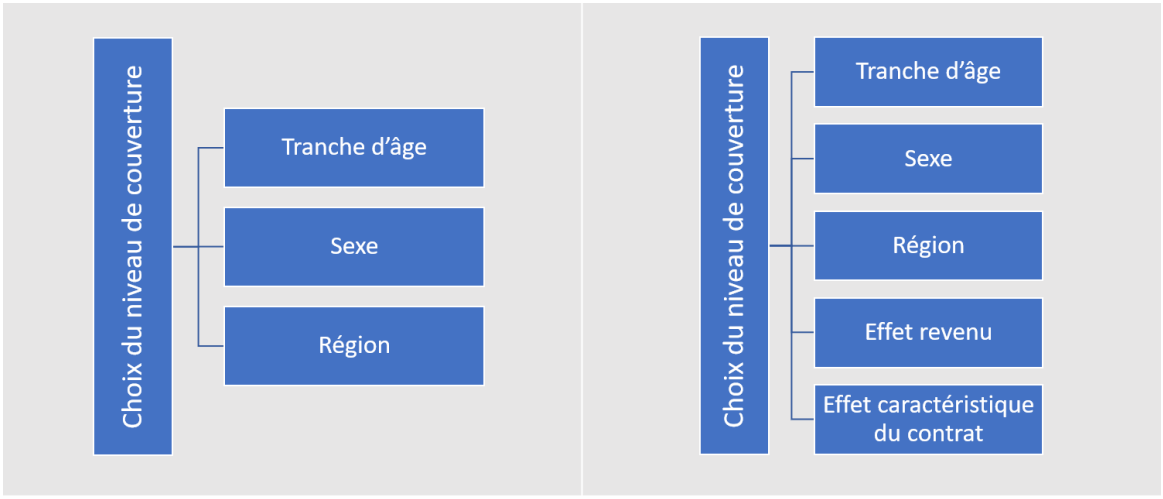
3.3.3 Récapitulatif sur les variables de chaque base

Afin de pouvoir appliquer le modèle, nous effectuons une transformation en variable binaire pour chaque caractère socio-démographique considéré. Nous obtenons ainsi 2 variables indicatrices pour le sexe, 7 variables indicatrices pour les tranches d'âge ainsi que 13 variables indicatrices représentant les 13 régions. Pour chaque assuré, nous associons leur prime technique, c'est-à-dire leur tarif proposé lors de la souscription, ainsi que le niveau de vie à disposition (qui a également été ajouté à la première base pour déterminer les probabilités en prenant en compte l'effet du revenu). Les deux bases d'assurés générées comportent les mêmes variables explicatives et se différencient par leur structure de relation causale avec la variable choix d'assurance (voir figure 3.8). Nous détaillons les variables des deux bases dans la table 3.4.

Variable	Description de la variable
Tranche d'âge d'assuré (AGE_BEN_SNDS)	
Age_20	Tranche d'âge 20-29 ans ($\mathbb{1}_{AGE\_BEN\_SNDS=20}$)
Age_30	Tranche d'âge 30-39 ans ($\mathbb{1}_{AGE\_BEN\_SNDS=30}$)
Age_40	Tranche d'âge 40-49 ans ($\mathbb{1}_{AGE\_BEN\_SNDS=40}$)
Age_50	Tranche d'âge 50-59 ans ($\mathbb{1}_{AGE\_BEN\_SNDS=50}$)
Age_60	Tranche d'âge 60-69 ans ($\mathbb{1}_{AGE\_BEN\_SNDS=60}$)
Age_70	Tranche d'âge 70-79 ans ($\mathbb{1}_{AGE\_BEN\_SNDS=70}$)
Age_80	Tranche d'âge 80+ ans ($\mathbb{1}_{AGE\_BEN\_SNDS=80}$)
Région d'habitation d'assuré (BEN_RES_REG)	
R5	Régions et Départements d'outre-mer ($\mathbb{1}_{BEN\_RES\_REG=5}$)
R11	Ile-de-France ($\mathbb{1}_{BEN\_RES\_REG=11}$)
R24	Centre-Val de Loire ($\mathbb{1}_{BEN\_RES\_REG=24}$)
R27	Bourgogne-Franche-Comté ($\mathbb{1}_{BEN\_RES\_REG=27}$)
R28	Normandie ($\mathbb{1}_{BEN\_RES\_REG=28}$)
R32	Hauts-de-France-Nord-Pas-de-Calais-Picardie ($\mathbb{1}_{BEN\_RES\_REG=32}$)
R44	Grand Est ($\mathbb{1}_{BEN\_RES\_REG=44}$)
R52	Pays de la Loire ($\mathbb{1}_{BEN\_RES\_REG=52}$)
R53	Bretagne ($\mathbb{1}_{BEN\_RES\_REG=53}$)
R75	Aquitaine-Limousin-Poitou-Charentes ($\mathbb{1}_{BEN\_RES\_REG=75}$)
R76	Languedoc-Roussillon-Midi-Pyrénées ($\mathbb{1}_{BEN\_RES\_REG=76}$)
R84	Auvergne-Rhône-Alpes ($\mathbb{1}_{BEN\_RES\_REG=84}$)
R93	Provence-Alpes-Côte d'Azur et Corse ($\mathbb{1}_{BEN\_RES\_REG=93}$)
Sexe d'assuré (BEN_SEX_COD)	
Homme	Sexe masculin ($\mathbb{1}_{BEN\_SEX\_COD=1}$)
Femme	Sexe feminin ($\mathbb{1}_{BEN\_SEX\_COD=2}$)
Prime technique d'assurance annuelle (sinistre moyen par niveau de couverture)	
Prime_min	Prime demandé pour le contrat de couverture minimum en €

Prime_moy	Prime demandé pour le contrat de couverture moyen en €
Prime_max	Prime demandé pour le contrat de couverture maximum en €
Niveau de vie d'assuré (budget moyen par segment de la population)	
Niv_de_vie	Revenu net total divisé par unité de consommation du ménage (INSEE) en €
Choix de couverture en complémentaire santé réelle d'assuré	
Choix	Indicateur de couverture choisi par assuré après avoir été proposé 3 offres ($\mathbb{1}_{couverture\_minimum=1} + 2 \times \mathbb{1}_{couverture\_moyen=1} + 3 \times \mathbb{1}_{couverture\_maximum=1}$)

TABLE 3.4 : Description des variables de la base d'assurés 1 et 2



Scénario de la base 1 : La vraie probabilité de choix d'assurance dépend seulement de trois variables de segmentation.

Scénario de la base 2 : La vraie probabilité de choix d'assurance dépend de trois variables de segmentation, ainsi que des effets de revenu et des caractéristiques du contrat en plus.

FIGURE 3.8 : Différence de la structure de relation de deux bases de données

Chapitre 4

Théorie de l'utilité aléatoire et la demande en assurance santé

Dans le chapitre 3, nous avons élaboré un indice d'anti-sélection basé sur les proportions de choix de couverture dans l'ensemble du portefeuille. L'analyse de l'anti-sélection consiste à examiner les probabilités de choix de couverture des assurés lors de la souscription du contrat d'assurance, intrinsèquement liées à l'activité de souscription d'assurance. Comme l'ont souligné Marquis et Holmer (1986), l'anti-sélection se réfère à la tendance des individus à souscrire à des polices d'assurance en fonction de leur propre perception des risques qu'ils encourent. Dans le contexte de la santé, cela signifie que les individus sont plus susceptibles de choisir des polices d'assurance qui correspondent à leurs besoins médicaux anticipés, ce qui peut créer des déséquilibres dans le groupe assuré. Les auteurs expliquent que "adverse selection[...]occurs when the insurer cannot perfectly predict the probability distribution of an individual's health care expenditures" (Marquis et Holmer (1986)). En d'autres termes, les assureurs peuvent être confrontés à des coûts plus élevés que prévu si un grand nombre d'assurés à haut risque souscrivent à des polices spécifiques.

Afin de mieux appréhender comment les individus prennent des décisions en matière d'assurance et comment l'anti-sélection influe sur ces choix, il faut chercher à comprendre la motivation de leur décision. Mathématiquement, on peut expliquer de telles décisions grâce à la formulation d'une quantité appelée l'utilité, plus souvent entendue et utilisée sous le nom de théorie de l'utilité espérée. Or, dans ce mémoire, nous avons choisi de travailler avec la théorie de l'utilité aléatoire en raison de son adaptation au contexte du choix d'assurance, en particulier dans le contexte de l'assurance santé, comme d'autres littératures sur le choix de l'assurance santé l'ont démontré : Powell et Goldman (2021), Keane (2004), Becker et Zweifel (2008), etc.

Les modèles de choix discret, tels que les modèles logit multinomiaux et les modèles mixtes logit, sont des outils statistiques couramment utilisés pour modéliser les décisions des individus face à un ensemble d'alternatives. Ils permettent de capturer la complexité des préférences individuelles envers les différents plans d'assurance en fonction de divers attributs, tels que les primes, les franchises et les niveaux de couverture. En modélisant la probabilité de choix des individus, on peut mieux appréhender comment l'anti-sélection influe sur la décision de souscription à une assurance particulière. Les chercheurs peuvent ainsi estimer la prévalence de l'anti-sélection sur le marché de l'assurance en examinant les schémas de souscription et en identifiant les préférences des individus à haut risque par rapport à ceux à faible risque. Les résultats obtenus à partir de ces modèles fournissent des informations quantitatives essentielles pour évaluer l'ampleur de l'anti-sélection et prendre des mesures appropriées pour atténuer ses effets néfastes sur le marché de l'assurance.

À travers ce chapitre, nous introduisons des modèles économétriques permettant à la fois de prédire les choix de couverture sous forme de probabilité et de faciliter une analyse et une interprétation simples de la dynamique sous-jacente.

4.1 Les décisions d'assurance : du phénomène complexe à l'utilité aléatoire

4.1.1 De la théorie dérivée du domaine psychologique aux modèles très connus en économétrie

La théorie de l'utilité aléatoire, qui a émergé dans les années 1960, incarne l'évolution d'une idée originellement ancrée dans la psychologie vers un domaine de recherche économique dynamique hautement prisé au sein de la communauté économique. Comme l'a souligné Luce (1959), elle offre "une représentation novatrice des décisions humaines, prenant en compte les éléments psychologiques qui échappent à la rationalité stricte". Cette approche novatrice a graduellement gagné en popularité grâce à sa capacité à appréhender avec précision la complexité des prises de décision humaines, et elle s'est trouvée des applications variées, dont certaines revêtent une importance considérable dans le domaine de l'assurance.

Initialement conçue pour capturer les comportements de choix qui ne se conforment pas strictement à la rationalité économique, la théorie de l'utilité aléatoire a été adoptée par des psychologues désireux de proposer des modèles plus adaptés pour rendre compte des variations d'humeur et des états mentaux changeants influençant les choix individuels. Cette approche, bien qu'ayant pris racine dans la psychologie, a progressivement suscité un vif intérêt parmi les économistes, enthousiasmés par sa capacité à mieux représenter la diversité des décisions humaines.

Comme Marschak (1950) l'a évoqué dans les années 1950, cette théorie permet de "saisir l'essence même des préférences individuelles dans un monde de choix complexes". Au fil du temps, cette théorie a été appliquée dans divers domaines, tels que la différenciation des produits, les interactions sociales et les jeux non coopératifs. Dans le contexte particulier de l'assurance, elle a introduit une perspective innovante en intégrant des éléments aléatoires dans les décisions de choix d'assurance. Cette approche permet ainsi de tenir compte des fluctuations émotionnelles et des variations de préférences individuelles, autant de facteurs qui varient d'un individu à l'autre.

Au cœur de cette approche résident des hypothèses essentielles concernant les préférences stochastiques et les comportements influencés par des états d'esprit changeants. La modélisation basée sur l'utilité aléatoire propose ainsi une représentation plus réaliste des décisions individuelles, en reconnaissant l'importance des facteurs psychologiques dans le processus de choix.

Un des aspects les plus frappants de cette théorie réside dans son efficacité avérée dans le domaine de la sélection des agents, notamment en matière d'assurance. En introduisant des compétitions basées sur les performances individuelles et en tenant compte des incertitudes dans les évaluations, cette approche a permis le développement de mécanismes de sélection plus sophistiqués, mieux adaptés aux situations où l'incertitude joue un rôle crucial.

Comme McFadden l'a déclaré en 1965 (selon McFadden (2000)), cette théorie permet "d'élucider les mécanismes complexes sous-jacents aux choix individuels et aux comportements économiques". En somme, la théorie de l'utilité aléatoire, dont les origines remontent aux travaux de psychologues, a acquis une immense popularité parmi les économistes en raison de sa capacité à expliquer les choix complexes et imparfaitement rationnels. En constante évolution depuis ses débuts, elle a trouvé de diverses applications, dont l'assurance, où elle a grandement contribué à enrichir la compréhension des mécanismes de sélection des agents.

4.1.2 Les concepts dans la théorie de l'utilité aléatoire

Avant d'approfondir le concept des modèles, cette section aborde les hypothèses dans lesquelles la théorie peut être appliquée. Les concepts économiques définissent le cadre d'application de ce mémoire, comme l'expérimentation de choix discrets et le type d'étude sera abordé en annexes B.1, B.2.

Hypothèse de rationalité économique

Dans le contexte des modèles de choix discrets en micro-économie, une question fondamentale réside dans l'hypothèse de rationalité qui sous-tend les décisions individuelles. L'article Billot et Thisse (1995), examine de manière approfondie cette question en explorant différentes hypothèses de rationalité et en mettant en perspective l'hypothèse de rationalité instrumentale parfaite.

L'hypothèse de rationalité instrumentale parfaite, considérée comme le fondement de l'approche micro-économique traditionnelle, repose sur deux postulats essentiels.

Hypothèse : Chaque individu est doté d'une relation de préférence complète et transitive, permettant de comparer et de classer les actions possibles. Par conséquence, chaque individu choisit toujours l'action qui est la plus préférée parmi toutes les options disponibles.

Cependant, l'article souligne que cette hypothèse de rationalité parfaite peut être remise en question, ce qui ouvre la voie à des modélisations plus nuancées du comportement individuel. Plusieurs interprétations sont examinées, chacune éloignant l'approche traditionnelle de la rationalité instrumentale parfaite :

- Prise en compte de l'information limitée : Une première manière de s'éloigner de la rationalité instrumentale parfaite est de reconnaître que l'individu pourrait ne pas avoir accès à toutes les informations nécessaires pour prendre des décisions parfaitement rationnelles. Cette limitation d'information peut conduire à des choix qui ne sont pas strictement optimaux.

- Fluctuations dans les évaluations : L'article évoque également la possibilité que l'état d'esprit d'un individu puisse fluctuer lors du processus d'évaluation des actions possibles. Ces fluctuations peuvent être influencées par des facteurs internes (comme l'humeur) et externes (comme le contexte). Ainsi, un individu peut évaluer et classer les actions différemment en fonction de son état d'esprit, conduisant à des choix variables.

- Incertitude sur les préférences : Une autre interprétation suggère que l'individu peut ne pas connaître parfaitement ses propres préférences. Des préférences mal définies ou cachées pourraient entraîner des choix qui ne correspondent pas nécessairement à une maximisation de l'utilité dans le sens traditionnel.

- Erreurs d'évaluation : L'article aborde également la possibilité que les individus puissent se tromper lors de l'évaluation des actions possibles. Étant donné la complexité de l'évaluation de multiples attributs, il est concevable que des erreurs surviennent, conduisant à des choix qui ne reflètent pas pleinement la rationalité.

L'article Billot et Thisse (1995) présente une analyse critique et nuancée des hypothèses de rationalité appliquées aux modèles de choix individuels discrets. En explorant diverses interprétations, l'article remet en question l'hypothèse de rationalité instrumentale parfaite, offrant ainsi une perspective éclairante sur les comportements réels et les éventuelles limites de la rationalité dans le processus de prise de décision individuelle.

4.1.3 Contexte d'application du modèle de choix discret dans le mémoire.

En se basant sur une compréhension des concepts économiques techniques utilisés pour évaluer la demande des consommateurs, notre motivation réside dans l'évaluation de la demande en assurance au sein d'une étude basée sur les Préférences Révélées.

Ce mémoire se concentre principalement sur les données recueillies sur un portefeuille d'assurance pour une année donnée, ce qui exclut le traitement des données en panel.

4.2 Modèle de choix discrets - Familles de modèles appliquées en économie

4.2.1 Capturer l'authentique hétérogénéité inter-individuelle : Une recherche guidée par l'essai

L'évolution des modèles de choix discrets a été guidée par l'impératif de capturer avec précision l'hétérogénéité inhérente aux préférences individuelles, conduisant à des avancées transformatrices. Initialement, les modèles supposaient une homogénéité, mais à mesure que le domaine progressait, des modèles sophistiqués et adaptables ont émergé pour aborder les divers goûts inhérents aux contextes de prise de décision du monde réel. Le modèle de Logit Conditionnel de McFadden, mis en évidence dans l'article de McFadden (1973), a marqué une étape initiale dans la reconnaissance de l'hétérogénéité en permettant la variation des coefficients entre les individus dans un cadre de choix fixe.

Le chemin vers une flexibilité accrue devient évident avec l'introduction du modèle de Logit Multinomial (MNL), également discuté dans le même article. Bien qu'il ait pris en compte des paramètres d'utilité spécifiques à chaque individu, le modèle MNL a maintenu des coefficients fixes, limitant sa capacité à capturer pleinement les complexités des préférences. En s'appuyant sur cette base, les modèles ultérieurs visaient à mieux encapsuler l'hétérogénéité. Le modèle Logit Emboîté (NL), organisant les choix de manière hiérarchique pour tenir compte des effets de cluster de choix en garantissant l'hypothèse d'Indépendance des Alternatives Non Pertinente (IANP) au sein de chaque cluster, permettait des représentations plus authentiques de la prise de décision dans des contextes spécifiques.

Poursuivant cette trajectoire, d'autres modèles influents ont considérablement contribué à capturer l'hétérogénéité. Le modèle Probit Multinomial MNP, discuté par Train (2009), et le modèle de Logit de loi mélange continue M-MNL, présenté par McFadden et Train (2000a), sont apparus comme des alternatives cruciales au modèle MNL, permettant des représentations plus nuancées des préférences. Dans le même ordre d'idées, le modèle de la Valeur Extrême Généralisée (GEV) a introduit le concept d'utilité dérivée de distributions de valeurs extrêmes, offrant un cadre plus large pour capturer l'hétérogénéité. Étant donné que les modèles MNL et MNP sont assez classiques pour la littérature, le modèle de Robit Multinomial MNR, développé par Krueger et al. (2023), suppose une distribution de Student pour les erreurs et établit ainsi une autre approche de la modélisation en permettant de modifier la queue de distribution.

Une avancée substantielle a été réalisée avec l'introduction des modèles de Classe Latente, qui ont reconnu la possibilité de segmenter les individus en classes distinctes basées sur des préférences communes, permettant ainsi d'accueillir à la fois l'hétérogénéité et l'homogénéité au sein de classes distinctes. Le modèle Logit de Classe Latente (LC-MNL), servant d'exemple, classifiait les individus en différents groupes, chacun caractérisé par ses paramètres d'utilité uniques. Cette approche nuancée permettait aux individus d'appartenir à diverses classes latentes avec des motifs de préférence différents.

L'article de Krueger et al. (2020) marque une avancée remarquable dans la modélisation de l'hétérogénéité. En utilisant le processus de Dirichlet comme une distribution de mélange non paramétrique, le modèle de Mélange Logit Multinomial de Processus de Dirichlet (DPM-MNL) est apparu. Se distinguant des modèles antérieurs, le modèle DPM-MNL n'exige pas la pré-définition du nombre de composantes de mélange, offrant une solution basée sur les données pour capturer l'hétérogénéité des préférences. Ce virage vers une modélisation à dimension infinie souligne la reconnaissance croissante de la nature multifacette des préférences individuelles dans le domaine.

Des tendances motivées par les méthodes de type apprentissage automatique se sont ainsi développées de plus en plus dans le domaine. L'idée principale est d'utiliser des algorithmes de type boîte noire pour apprendre les hétérogénéités inobservables grâce à leur capacité à capturer les tendances dans les données de façon non paramétrique, à l'inverse de la spécification paramétrique des méthodes de choix discret classiques. Sifringer et al. (2020) a développé les modèles Logit multinomial d'apprentissage L-MNL et Logit emboîté d'apprentissage L-NL partiellement joints par les réseaux de neurones pour

apprendre spécifiquement une partie d'utilité importante qui ne peut pas être spécifiée par l'expertise, tandis que le modèle de Classe Latente avec processus gaussien (GP-LC-MNL) de Sfeir et al. (2022) vise à incorporer la méthode probabiliste non paramétrique dans la détection des classes latentes inobservables.

En résumé, l'évolution de la modélisation des choix discrets a été guidée par la nécessité de capturer avec précision la diversité des préférences individuelles. Allant des premières tentatives pour incorporer des variations au développement de modèles complexes tels que le DPM-MNL, et des extensions ultérieures comprenant le Probit, le Logit Mixte, le GEV, le GMNL, l'hétérogénéité d'échelle et les modèles de Classe Latente, le domaine s'est continuellement efforcé de créer des modèles qui reflètent fidèlement les dynamiques multifacettes de la prise de décision humaine.

Dans le cadre de ce mémoire, nous retenons les modèles suivants que nous présenterons aux lecteurs en raison de leur utilité pratique, tels que le modèle Logit Multinomial, le Probit, le Logit Mixte, le Logit Hétérogène et le modèle de Classe Latente.

4.2.2 Notion d'utilité dans le contexte de choix discret

Soit une situation dans laquelle un individu doit choisir un choix i parmi un ensemble fini de choix $C = \{1, .., I\}$ ($i \in C$, $Card(C) > 1$). Selon l'hypothèse de rationalité instrumentale parfaite, chaque individu, lorsqu'il est confronté à un choix spécifique, est supposé accorder un poids à ce choix dans son processus de sélection. Cette quantité est généralement notée U_i et représente l'utilité du choix i pour l'individu, reflétant ainsi le bénéfice moral perçu par cet individu. L'utilité peut dépendre des caractéristiques de l'individu, des choix disponibles, de l'environnement ou de l'interaction de ces éléments. En termes simples, un individu préfère le choix qui lui procure la plus grande utilité parmi toutes les options.

Dans le contexte de l'assurance, la notion d'utilité espérée est couramment utilisée car elle fournit une référence pour les choix d'assurance dans un avenir incertain (comme la somme des utilités pondérées par les probabilités de tous les scénarios possibles). Il est important de noter que les caractéristiques concaves et croissantes en termes de ressources économiques d'une fonction d'utilité, dans le cadre de la théorie de l'utilité espérée, découlent du phénomène d'aversion au risque financier des agents économiques. Prenons un exemple : choisir entre ne rien choisir et choisir une loterie entraînant les conséquences $x_1, x_2, ..., x_M$ dans M scénarios possibles sur le capital financier Cap_fin avec les probabilités de réalisation $p(k)$ ($k \in \{1, .., M\}$). Un agent adverse au risque financier préfèrera ne rien choisir plutôt que de choisir la loterie, ce qui se traduit une définition de U telle que :

$$E(U_{rien}) = U(Cap\_fin) > E(U_{lotterie}) = \sum_{k=1}^{M} p(k) \times U(Cap\_fin + x_k).$$

Cependant, la notion d'utilité espérée peut parfois être en incohérence avec les choix réels des agents économiques. Influencés par les émotions, les biais cognitifs et d'autres facteurs psychologiques, les individus ne suivent pas toujours parfaitement le comportement prédit par cette théorie. Un exemple évident est que le même individu, confronté à la même situation et aux mêmes choix, peut aboutir à des décisions différentes. Cela s'explique par des informations spécifiques à l'individu que les économistes n'ont pas pu observer ou intégrer dans la théorie de l'utilité espérée.

C'est pourquoi la théorie de l'utilité aléatoire, conçue pour quantifier de manière probabiliste les variations endogènes et inobservables des agents économiques, est beaucoup plus largement utilisée dans la modélisation du comportement humain en pratique. Nous posons ainsi l'hypothèse suivante pour travailler dans un cadre d'utilité aléatoire.

Hypothèse : Chaque individu est doté d'informations privées ou de raisonnements personnels sur ses préférences de choix, qui sont privés à l'observation. Cela impacte fortement le choix en dehors du sens de l'utilité espérée, ce qui rend leurs choix aléatoires pour les modélisateurs. Leur utilité évaluée

sur chaque choix est ainsi considérée comme aléatoire.

4.2.3 Formulation d'utilité aléatoire

Décomposition en partie déterministe et aléatoire

On se place dans un contexte de choix, donc il y a un ensemble de N individus en face de I choix. Comme son nom l'indique, l'utilité du choix perçue par l'individu n ($n \in \{1, .., N\}$) pour l'alternative i ($i \in \{1, .., I\}$) est une quantité aléatoire. Cependant, dans le contexte de la modélisation, cette utilité U_{ni} peut être décomposée en deux parties distinctes : une composante déterministe observable V_{ni} destinée aux modélisateurs, et une partie aléatoire d'utilité ϵ_{ni} due à des effets endogènes non observables de l'individu. Cette décomposition s'exprime de la manière suivante :

$$U_{ni} = V_{ni} + \epsilon_{ni}.$$

La partie déterministe de l'utilité peut dépendre des caractéristiques de l'alternative, des caractéristiques de l'individu et de leurs interactions. Elle peut être divisée en deux sous-composantes : celles liées exclusivement aux attributs des alternatives, et celles liées aux caractéristiques de l'individu. De plus, il existe une troisième sous-composante reflétant les interactions entre les attributs des alternatives et les caractéristiques de l'individu. Il est courant de supposer que les utilités déterministes sont additives en utilité marginale, on établit cette hypothèse :

Hypothèse : L'utilité déterministe est additive comme une somme des utilités marginales des effets et variables.

Sous l'hypothèse de l'additivité d'utilité marginale, la partie déterministe de l'utilité V_{ni} peut être formulée comme suit :

$$V_{ni} = V_X(\mathbf{X_n}) + V_Z(\mathbf{Z_i}) + V_{X,Z}(\mathbf{X_n}, \mathbf{Z_i}),$$

où :

- V_{ni} représente la composante déterministe de l'utilité de l'alternative i pour l'individu n,
- $V_X(\mathbf{X_n})$) est la portion d'utilité liée aux caractéristiques de l'individu n,
- $V_Z(\mathbf{Z_i})$ désigne la portion d'utilité de l'alternative i liée aux attributs de cette dernière
- $V_{X,Z}(\mathbf{X_n}, \mathbf{Z_i})$ représente la portion d'utilité résultant des interactions entre les attributs de l'alternative i et les caractéristiques de l'individu n.

Chacune de ces sous-composantes est abordée en détail. Il est important de noter qu'il est possible de modéliser au niveau de l'unité domestique, alors que l'utilité décrite ici n'est plus individuelle mais s'applique à l'ensemble du ménage. Dans ce cas, il est nécessaire de prendre en compte les effets endogènes de l'interaction entre les individus au sein d'un même ménage.

Partie déterministe liée aux attributs des alternatives

Cette partie de l'utilité est associée exclusivement aux attributs spécifiques des alternatives. Ces attributs influencent l'utilité de chaque alternative pour l'ensemble de la population ciblé. Les attributs considérés incluent des caractéristiques mesurables susceptibles d'influencer les préférences et les choix des individus parmi les alternatives, notamment les attributs de prix, les quantités d'offre, le nombre de bénéfices pour certains choix, etc. Ces mesures varient d'une alternative à l'autre pour un même individu, ainsi qu'entre individus. Par exemple, le coût des lunettes d'entrée de gamme à haut de gamme pour les enfants et les adultes. La fonction d'utilité associée à cette sous-composante pourrait être formulée de la manière linéaire suivante :

$$V_Z : \mathbb{R}^F \to \mathbb{R}, V_Z(\mathbf{Z_i}) = \gamma_1 \times Z_{i1} + \gamma_2 \times Z_{i2} + ... + \gamma_k \times Z_{ik} + ... + \gamma_F \times Z_{iF},$$

où :

- γ_k est le paramètre qui définit la direction et l'importance de l'effet de l'attribut k ($k \in \{1, .., F\}$) sur l'utilité de l'alternative i
- Z_{ik} est la valeur de l'attribut k pour l'alternative i

Partie déterministe liée aux caractéristiques de l'individu

Les différences dans les "biais" d'évaluation entre les individus peuvent être prises en compte en intégrant des variables personnelles et domestiques dans les modèles de choix discrets. Les variables couramment utilisées à cet effet comprennent :

- Le revenu du ménage ou le salaire par personne,
- Le sexe,
- L'âge,
- La région où vit l'individu,
- Le nombre de travailleurs dans le ménage,
- Le nombre d'individus dans le ménage.

Dans certains cas, ces variables sont combinées. Par exemple, le revenu du ménage peut être divisé par le nombre d'individus du même ménage pour indiquer le niveau de vie disponible pour chaque membre du ménage.

Cette approche suppose la composante "biais" de la fonction d'utilité sommable, qui prend souvent une forme linéaire en variables et peut être exprimée comme une combinaison linéaire des caractéristiques pour un effet de fond déterministe différent résultant de la classe sociale de l'individu :

$$V_X : \mathbb{R}^M \to \mathbb{R}, V_X(\mathbf{X_n})) = \beta_{i0} \times \mathrm{ASC}_i + \beta_{i1} \times X_{1n} + \beta_{i2} \times X_{2n} + ... + \beta_{im} \times X_{mn} + ... + \beta_{iM} \times X_{Mn},$$

où :

- β_{im} est le paramètre qui définit la direction et l'ampleur du biais supplémentaire résultant de l'augmentation de la caractéristique m ($m \in \{1, .., M\}$) du décideur (le cas $m = 0$ représente la constante spécifique à l'alternative),
- X_{mn} est la valeur de la caractéristique m pour l'individu n.

Partie déterministe définie par les interactions entre les attributs des alternatives et les caractéristiques de l'individu

La dernière composante de l'utilité prend en compte les différences dans la manière dont les attributs sont évalués par différents individus. Par exemple, dans le cas d'un choix de mode de transport, certaines personnes peuvent accorder plus d'importance au temps de trajet que d'autres en raison de leurs caractéristiques personnelles. Cela peut être représenté en introduisant une interaction entre les attributs de l'alternative et les caractéristiques de l'individu. Par exemple, le temps de trajet peut être pondéré différemment en fonction du sexe de l'individu ou de son revenu. Cette sous-composante de l'utilité pourrait être formulée comme suit :

$$V_{X,Z} : \mathbb{R}^M \times \mathbb{R}^F \to \mathbb{R}, V_{X,Z}(\mathbf{X_n}, \mathbf{Z_i}) = \sum_{k=1}^{F} \sum_{m=1}^{M} \delta'_{km} \times Z_{ik} \times X_{mn},$$

avec δ'_{km} le paramètre qui définit la force de l'interaction entre l'attribut k de l'alternative et la caractéristique m de l'individu n.

La non-linéarité dans l'évaluation de l'utilité marginale des variables et l'effet de revenu

Il est important de noter que le choix de modéliser l'interaction entre les variables prises en compte dépend fortement de la connaissance du modélisateur. Pour mieux comprendre les méthodes de modélisation de la partie fixe d'utilité basée sur les caractéristiques observables, les lecteurs pourraient s'intéresser au Bierlaire (sans date).

Une connaissance importante du comportement humain pour les économistes lors de la phase de conception de la forme d'utilité déterministe est l'effet de revenu (ou le reste de somme après le paiement pour un choix) et la diminution de l'utilité marginale du revenu, c'est-à-dire la diminution de l'utilité marginale du revenu, avec la valeur du numéraire diminuant avec l'augmentation du revenu (abordé dans Delle Site et al. (2022)). En effet, face à une situation où le choix peut affecter la richesse (monétaire ou patrimoniale) *a posteriori* de l'individu, certaines personnes ont tendance à être indifférentes aux petites variations de leur richesse lorsque celle-ci est très élevée. Par exemple, pour ceux qui gagnent plus de 12 000 € par mois, le choix d'acheter des lunettes coûtant 500 € leur pèse financièrement beaucoup moins qu'une personne gagnant le salaire minimum défini par l'État (le SMIC). De plus, entre deux paires de lunettes coûtant respectivement 100 € et 200 €, et deux paires de lunettes haut de gamme coûtant 1 150 € et 1 250 €, les gens ont tendance à ignorer la différence de prix des deux paires de lunettes haut de gamme, même si les différences de prix sont les mêmes, soit 100 €. Il en résulte donc que l'importance du coût du choix évalué par l'individu diminue avec sa richesse au début, ou peut être faussée par la grandeur du prix. Suivant Cherchi et al. (2004) et Gaudry et al. (1989) pour le modèle Logit Box-Cox, le modélisateur peut choisir d'intégrer cet effet $H(cout_i, richesse_n)$ linéairement dans V_{ni} à travers d'un coefficient β_{cout_ni} sous plusieurs formes, notamment deux formes très utilisées :

- Sous forme d'un coût baissant le niveau d'utilité : Exemple de l'intéraction souvent utilisé $H(cout_i, richesse_n) = \frac{cout_i}{richesse_n}$, mais le paramètre β_{cout_ni} devrait être négatif en raison de la représentation d'un coût diminuant l'utilité.

- Sous forme de richesse restante après le choix de l'alternative i : $H(cout_i, richesse_n) = \Phi(richesse_n - cout_i)$ avec Φ une fonction concave croissante. Dans cette représentation, le paramètre β_{cout_ni} est supposé positif car l'individu est d'autant plus content que sa richesse résiduelle est grande.

Des familles de transformations concaves croissantes couramment utilisées dans la littérature sont :

- Fonction linéaire par morceaux : Transformation la plus adaptée à la réalité, car toute fonction peut être approximée par une fonction linéaire par morceaux, mais elle demande un processus d'identification compliqué.

- Transformation Box-Cox avec paramètre α : Famille de transformations pour $x > 0$, incluant la fonction logarithme naturel dans la limite où α tend vers 0 et la fonction linéaire lorsque $\alpha = 1$. Dans notre cas, le paramètre α doit être compris entre 0 et 1 pour assurer la concavité.

$$Box - Cox(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(x) & \text{si } \lambda = 0 \end{cases}$$

- Fonction utilité exponentielle avec paramètre α : une forme de fonction utilité populaire dans les modèles économiques, concave lorsque $\alpha \geq 0$

$$u(x) = \begin{cases} \frac{1 - e^{-ax}}{a} & \text{si } a \neq 0 \\ x & \text{si } a = 0 \end{cases}$$

- Fonction utilité power-exponential : Saha (1993) a introduit la fonction d'utilité expo-puissance définie où $\theta > 1$ et $\alpha\beta > 0$

$$u(x) = \theta - e^{-\beta x^{\alpha}}.$$

En supposant une telle hypothèse de non-linéarité dans le problème que nous modélisons, il a été inévitable de mener un travail de recherche et de solliciter l'expertise, ainsi que de réaliser des tests pour vérifier cette hypothèse. Nous aborderons dans le chapitre 4 la mise en place de la transformation Box-Cox et la vérification de sa linéarité à travers le paramètre λ.

4.2.4 Modèle de Luce et l'hypothèse IANP

Maximum d'utilité aléatoire et modèle de Luce

En évaluant tout le panier de choix $C = \{1, .., I\}$, l'individu n choisira le choix idéal $i' \in C$ maximisant son utilité accordée au choix $U_{ni'}$:

$$i' \in C, i' = argmax_{j\in[1,..I]}(U_{nj}) = argmax_{j\in[1,..,I]}(V_{nj} + \epsilon_{nj}) \qquad (*)$$

Or, comme les éléments aléatoires $\epsilon_{nj,j\in[1,..,I]}$ sont inobservables par le modélisateur, on n'est pas en capacité de prédire directement son choix mais plutôt évalué les probabilités que l'utilité de chaque choix est maximisé, avec des connaissances ou hypothèses supposées sur la distribution de la partie aléatoire de l'utilité.

Il existe un type de modèle simple évaluer les probabilités de choix assez proche de modèle de choix discret mais basant sur le poid déterministe des choix, on parle de modèle Plackette-Luce. Soit ν_i le poid accordé au choix i ($i \in [1, .., I]$), on peut calculer la probabilité que l'individu choisi l'option i comme suite :

$$P(i) = \frac{\nu_i}{\sum_{j=1}^{I} \nu_j}.$$

Probabilité dérivée du modèle d'utilité aléatoire

En suivant le cours du Professeur Bierlaire (Bierlaire (sans date)), on peut formuler mathématiquement la probabilité de choisir l'option i à partir du modèle d'utilité aléatoire général, en tenant compte du fait que l'aspect aléatoire réside uniquement dans ϵ.

Soit $\mathcal{C}_n \subset C$ l'ensemble des choix disponibles pour l'individu n parmi N individus, on note $Card(\mathcal{C}_n) = I_n$. Étant donné que V_{ni} dépend des caractéristiques observables $\mathbf{X_n}$ et $\mathbf{Z_i}$, on peut considérer que $\mathcal{C}_n$ englobe toutes ces informations.

Par principe de maximisation d'utilité aléatoire de l'équation (*), on obtient la probabilité :

$$P(\text{choix idéal} = i | i \in \mathcal{C}_n) = P(i|\mathcal{C}_n) = P(U_{ni} \geq U_{nj}, \text{ pour tout } j = 1, \ldots, I_n),$$

où :

$$U_{ni} = V_{ni} + \epsilon_{ni}.$$

Ici, $\epsilon_{\mathbf{n}}$ représente le vecteur des I_n termes d'erreur : $\epsilon_{\mathbf{n}} = (\epsilon_1, \ldots, \epsilon_{I_n})$. Si ϵ_n est une variable aléatoire multivariée avec une fonction de répartition cumulative (CDF) notée $F_{\epsilon_{\mathbf{n}}}(\epsilon_1, \ldots, \epsilon_{I_n})$ et une fonction de densité de probabilité (pdf) $f_{\epsilon_{\mathbf{n}}}(\epsilon_1, \ldots, \epsilon_{I_n})$ définie comme :

$$f_{\epsilon_{\mathbf{n}}}(\epsilon_1, \ldots, \epsilon_{I_n}) = \frac{\partial^{I_n} F_{\epsilon_{\mathbf{n}}}}{\partial\epsilon_1 \cdots \partial\epsilon_{I_n}}(\epsilon_1, \ldots, \epsilon_{I_n}).$$

Alors la probabilité peut être exprimée en fonction de ϵ_{ni} et V_{ni} comme suit :

$$P\left(i \mid \mathcal{C}_n\right) = \int_{\epsilon_i=-\infty}^{+\infty} \int_{\epsilon_1=-\infty}^{V_{n,i}-V_{n,1}+\epsilon_i} \cdots \int_{\epsilon_{i-1}=-\infty}^{V_{n,i}-V_{n,i-1}+\epsilon_i} \int_{\epsilon_{i+1}=-\infty}^{V_{n,i}-V_{n,i+1}+\epsilon_i} \cdots \int_{\epsilon_{I_n}=-\infty}^{V_{n,i}-V_{n,I_n}+\epsilon_i} f_{\boldsymbol{\epsilon}_\mathbf{n}}\left(\epsilon_1, \epsilon_2, \ldots, \epsilon_{I_n}\right) d\boldsymbol{\epsilon}_\mathbf{n}.$$

et également en fonction des dérivées partielles de la CDF comme suit (Démonstration en Bierlaire (sans date)) :

$$P_n\left(i \mid \mathcal{C}_n\right) = \int_{\epsilon=-\infty}^{+\infty} \frac{\partial F_{\boldsymbol{\epsilon}_\mathbf{n}}}{\partial \epsilon_i}\left(\ldots, V_{n,i} - V_{n,i-1} + \epsilon, \epsilon, V_{n,i} - V_{n,i+1} + \epsilon, \ldots\right) d\epsilon.$$

Par conséquent, lorsque la CDF est exprimée en une forme fermée, le modèle de choix peut être déterminé par la résolution d'une intégrale unidimensionnelle. Cette résolution peut être effectuée de manière analytique pour des modèles simples, et de façon numérique pour des modèles plus complexes.

Hypothèse Indépendance des alternatives non pertinentes (IANP)

Lors de l'introduction de ce modèle, deux axiomes jouent un rôle central pour décrire les choix discrets. Le premier est l'axiome de positivité, qui stipule que pour toute option, il existe une probabilité positive d'être choisie (voir McFadden (1973)). Cela garantit que les probabilités sont positives, ce qui facilite les estimations.

Hypothèse : $a \in A$, avec $A \subseteq C$, nous avons $P[a|A] > 0$.

Le deuxième axiome est l'axiome de l'indépendance des alternatives non pertinentes (IANP, ou IIA - Independent and Irrelevant Alternative en anglais), qui affirme que l'ajout ou la suppression d'une option dans le panier affecte tous les autres choix proportionnellement. Ce dernier axiome assure que le rapport des probabilités de choix entre deux options ne dépend pas des autres options dans le menu.

Hypothèse : Pour tout $A \subset C$ et $B \subset C$ tels que $\{a, b\} \subset A \cap B$, alors $\frac{P[a|A]}{P[b|A]} = \frac{P[a|B]}{P[b|B]}$.

La positivité et l'axiome IANP impliquent que la préférence stochastique $\succ$ est transitive (voir Luce (1959)), ce qui signifie qu'il existe un ordre de préférence assurant la cohérence et la rationalité d'un individu. Cet axiome est restrictif, mais il est largement utilisé dans les travaux empiriques pour des raisons de faisabilité et de simplicité. Nous montrons dans les sections suivantes qu'il existe en réalité des cas violant cette hypothèse d'IANP et comment la relâcher afin de capturer plus largement les comportements des agents.

Points spécifiques du modèle d'utilité aléatoire

D'après Train (2009), lors de l'identification et de la spécification des modèles de choix discrets, deux principes majeurs émergent : "Seules les différences d'utilité comptent" et "L'échelle d'utilité est arbitraire". Ces principes ont des implications profondes pour les comportements des décideurs et la modélisation des chercheurs. En essence, seul l'écart entre les utilités a un impact, indépendamment de leur niveau absolu. Cela influence la spécification des modèles, avec des constantes spécifiques aux alternatives qui capturent les différences entre elles. Cependant, seules les différences relatives de ces constantes sont pertinentes, les niveaux absolus étant normalisés pour refléter cette réalité.

4.2.5 Modèle Logit Multinomial (MNL)

Formulation du modèle

McFadden (1973) a introduit le premier modèle de la famille des modèles de choix discrets : le Modèle Logit Multinomial. Si l'on suppose que les utilités inobservables de l'individu n, notées $\epsilon_{ni, i\in[1,..,I_n]}$, suivent une loi de valeur extrême de type I, c'est-à-dire la loi de Gumbel II avec une

fonction de répartition $F(\epsilon_{ni}) = e^{-e^{-\epsilon_{ni}}}$, alors la probabilité de choisir l'option i parmi l'ensemble $\mathcal{C}_n$ est donnée par :

$$P(i|\mathcal{C}_n) = P_n(i|\mathcal{C}_n) = \frac{e^{V_{ni}}}{\sum_{j=1}^{I_n} e^{V_{nj}}}.$$

Dans la suite du mémoire, nous noterons $P_n(i|\mathcal{C}_n)$ pour la probabilité que l'individu n choisisse l'alternative $i \in \mathcal{C}_n$, afin d'éviter l'ambiguïté en présentant différents individus.

Pour simplifier les calculs et se concentrer sur l'application du modèle, la démonstration de cette formule est intrinsèquement liée à la formule de probabilité dérivée du modèle d'utilité aléatoire, avec deux éléments supplémentaires : la différence de deux variables aléatoires suivant la distribution logistique et le fait que les $\epsilon_{ni,i\in[1,..,I_n]}$ sont indépendantes et identiquement distribuées (i.i.d.). Cette démonstration ne sera donc pas abordée dans ce mémoire. Les lecteurs peuvent se référer à Train (2009) pour une meilleure compréhension.

Maximum de vraisemblance

Dans une configuration simple, supposons que notre utilité soit représentée par une combinaison linéaire des caractéristiques observables de l'individu n et de l'alternative i (où $i \in \mathcal{C} = \{1, .., I_n\} \subseteq \mathcal{C} = \{1, .., I\}$) :

$$V_{ni} = \beta_{i0} + \sum_{k=1}^{K_i} \alpha_{ik} S_{nik} = \beta_{i0} + S_{ni}^T \alpha_i,$$

avec :

- $\beta_{10}, \ldots, \beta_{I_n 0}$ sont les constantes spécifiques à chaque alternative, on note $\beta = (\beta_{10}, \ldots, \beta_{I_n 0})$,
- K_i représente le nombre total de variables caractéristiques observables de l'individu et de l'alternative i, on note ce vecteur $\mathbf{S_{ni}} = (S_{ni1}, \ldots, S_{niK_i})$. Chaque option pourrait avoir des caractéristiques propres à elle que les autres possibilités n'ont pas, mais le nombre de caractéristiques observables des individus reste le même pour tous les individus,
- Le vecteur $\alpha_\mathbf{i} = (\alpha_{i1}, \ldots, \alpha_{iK_i})$ représente les paramètres associés au vecteur $\mathbf{S_{ni}}$, qui indiquent les poids attribués par les décideurs à chaque attribut qu'ils perçoivent.

Comme la formule de probabilité est bien une formule fermée, l'estimation du modèle (estimation des paramètres β et α_i) peut se faire à travers la maximisation de la fonction de vraisemblance. Plus précisément, on considère un échantillon de N individus auquel notre individu n appartient, donc on connaît tous les vecteurs $\mathbf{S_{ni}}, i \in [1, ..I_n]$ et sur tous les individus $n \in [1, .., N]$. De plus, on dispose des choix réellement faits par tous ces individus $\mathbf{Y_n} = (y_{n1}, .., y_{nI_n}), n \in \{1, .., N\}$ avec $y_{ni} = 1$ si l'individu n a choisi l'alternative i et 0 sinon. La vraisemblance étant égale au produit des probabilités, conditionné aux paramètres β et α_i, pour les N personnes de choisir l'alternative qu'elles ont effectivement été observées à choisir, sous réserve que les individus sont supposés indépendants. Elle peut être écrite comme :

$$\mathcal{L}(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}) = \prod_{n=1}^{N} \prod_{i=1}^{I_n} P_n(i|\mathcal{C}_n)^{y_{ni}} = \prod_{n=1}^{N} \prod_{i=1}^{I_n} \left(\frac{e^{V_{ni}}}{\sum_{j=1}^{I_n} e^{V_{nj}}}\right)^{y_{ni}} = \prod_{n=1}^{N} \prod_{i=1}^{I_n} \left(\frac{e^{\beta_{i0} + \sum_{k=1}^{K_i} \alpha_{ik} S_{nik}}}{\sum_{j=1}^{I_n} e^{\beta_{j0} + \sum_{k=1}^{K_j} \alpha_{jk} S_{njk}}}\right)^{y_{ni}}.$$

Si on prend le logarithme de la vraisemblance, on obtiendra :

$$l(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}) = ln(\mathcal{L}(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]})) = ln\left(\prod_{n=1}^{N} \prod_{i=1}^{I_n} P_n(i|\mathcal{C}_n)^{y_{ni}}\right) = \sum_{n=1}^{N} \sum_{i=1}^{I_n} y_{ni} ln(P_n(i|\mathcal{C}_n))$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{I_n} y_{ni} ln(\frac{e^{V_{ni}}}{\sum_{j=1}^{I_n} e^{V_{nj}}}) = \sum_{n=1}^{N} \sum_{i=1}^{I_n} y_{ni} ln(\frac{1}{1 + \sum_{j=1, j\neq i}^{I_n} e^{\beta_{j0} + \sum_{k=1}^{K_j} \alpha_{jk} S_{njk}}}).$$

McFadden (1973) a démontré que la log-vraisemblance est globalement concave pour les formes d'utilité linéaires en paramètres, garantissant ainsi l'existence d'une solution unique au problème de maximisation de la log-vraisemblance (équivalence au problème de maximisation de la fonction de vraisemblance). Cependant, l'introduction de formes d'utilité non linéaires en paramètres peut entraîner la perte de la concavité de la fonction de log-vraisemblance, conduisant à des optima locaux. Les algorithmes d'optimisation numérique introduits par Train (2009) qui permettent de résoudre ce problème, sont généralement mis en œuvre dans les bibliothèques existantes de langages de programmation tels que le Newton-Raphson, BHHH, BHHH-2, Steepest Ascent, DFP, etc.

Après avoir appliqué la procédure d'optimisation, on obtient ainsi une approximation du véritable vecteur de paramètres du modèle, noté $\hat{\beta}$ et $\hat{\alpha}_{\mathbf{i}}$ (pour $i \in \{1, .., I\}$).

Constraint de normalisation des paramètres

Comme abordé dans la section sur les points particuliers du modèle, seule la différence relative d'utilité est importante. Pour réduire la dimension des paramètres, il convient de réinitialiser un des coefficients d'une variable dont les coefficients sont spécifiés pour chaque alternative. Il s'agit, par exemple, du coefficient $\beta_{i0} = 0$ où $i \in C$ quelconque et $\alpha_{ik} = 0$ avec n'importe quel $i \in C$ pour chaque k tel que $S_{nmk} \neq S_{nlk} \forall m, l \in C, m \neq l$.

Problème de bus rouge/bleu

D'après la formulation du modèle Logit multinomial, on observe que pour deux choix distincts i et j dans l'ensemble $\mathcal{C}_n$, le rapport des probabilités de choisir i par rapport à j est donné par :

$$\frac{P_n(i|\mathcal{C}_n)}{P_n(j|\mathcal{C}_n)} = \frac{\frac{e^{V_{ni}}}{\sum_{k=1}^{I_n} e^{V_{nk}}}}{\frac{e^{V_{nj}}}{\sum_{k=1}^{I_n} e^{V_{nk}}}} = \frac{e^{V_{ni}}}{e^{V_{nj}}} = e^{V_{ni} - V_{nj}}.$$

Comme ce rapport de probabilité dépend uniquement de l'utilité déterministe des alternatives i et j, il met clairement en évidence le caractère IANP. L'hypothèse d'indépendance des alternatives non pertinentes (IANP) du modèle MNL stipule que les chances de choisir entre différentes alternatives sont indépendantes les unes des autres. Cela signifie que les rapports de probabilité entre deux choix spécifiques restent inchangés, même lorsque d'autres options sont introduites dans le modèle. Cette propriété peut sembler contre-intuitive, car elle impliquerait que des alternatives sans lien apparent influencent les choix de manière uniforme, quelle que soit la présence d'autres alternatives similaires. Il existe un paradoxe illustrant la violation de cette hypothèse.

Ce paradoxe est parfois illustré par l'exemple du "bus bleu/rouge" : imaginons un modèle avec trois choix de transport - voiture, bus rouge et train. L'IANP signifie que les probabilités relatives entre la voiture et le bus rouge resteront les mêmes, que le troisième choix soit un bus bleu ou un train. Cependant, dans le modèle logit multinomial, on observe que l'introduction du bus rouge fait augmenter la probabilité que les gens choisissent le bus (qu'il soit bleu ou rouge) en diminuant la probabilité de choisir la voiture. Cette observation crée une contradiction apparente avec l'IANP.

L'hypothèse d'IANP est pratique pour les estimations, mais elle peut ne pas toujours être réaliste. Hausman et McFadden (1984) souligne que l'IANP est théoriquement peu probable dans de nombreuses situations, mais empiriquement, le modèle MNL reste robuste dans de nombreux cas où l'IANP est peu plausible. Des tests peuvent être effectués pour évaluer la validité de l'IANP, et certains critères suggèrent que si des alternatives sont véritablement non pertinentes, leur exclusion du modèle n'aura pas un impact significatif sur les estimations des paramètres. Cependant, si les probabilités de choix

sont affectées par d'autres alternatives (c'est-à-dire si l'IANP n'est pas vérifiée), alors les paramètres estimés peuvent ne pas être significatifs.

4.3 Capture des hétérogénéités comportementales

La violation flagrante de la caractéristique IANP dans un contexte de choix discret conduit à l'inefficacité des modèles Logit Multinomial, voire même des modèles Logit Multinomial Emboîtés (voir l'annexe B.3). Cette inefficacité découle en grande partie de l'incompatibilité entre notre hypothèse selon laquelle les termes aléatoires sont identiquement et indépendamment distribués selon la même loi de valeur extrême et la réalité. En effet, d'autres sources d'aléas altèrent cette distribution inobservable pour le modélisateur. L'utilisation du modèle Probit Multinomial (présenté en annexe B.4) peut alléger cette contrainte en permettant les corrélations entre les termes aléatoires, mais il présente un inconvénient considérable en raison de la complexité de modélisation des matrices de corrélation, ainsi que de la demande importante de ressources pour l'entraînement lorsque la dimension augmente. Il faudra donc revenir aux autres modèles.

Les modèles présentés dans cette section proposent une manière de spécifier l'utilité en tenant compte des effets aléatoires provenant spécifiquement de l'individu ou d'une classe d'individus, ce qui permet de mieux refléter la réalité et d'améliorer la pertinence des résultats.

4.3.1 Modèle Logit Multinômial Hétérogène (HMLM)

Ce modèle repose sur le fait que les vecteurs aléatoires ϵ_n, pour $n \in \{1, .., N\}$, sont indépendants mais distribués selon une loi de valeur extrême, avec différentes variances variant en fonction des caractéristiques de l'individu n considéré. Mathématiquement, l'utilité de l'individu n face à un choix $i \in \mathcal{C}_n$ est exprimée comme suit :

$$U_{ni} = V_{ni} + \epsilon'_{ni} = V_{ni} + \frac{1}{\sigma_n}\epsilon_{ni}, \quad i \in \{1, .., I_n\}.$$

Ici, les $\epsilon_{ni, i\in\{1,..,I_n\}}$ sont indépendants et suivent la loi de Gumbel avec une variance unitaire, tandis que σ_n dépend de l'individu n.

Afin de pouvoir estimer le modèle, les termes aléatoires sont normalisés par σ_n, ce qui les ramène à une variance unitaire (consistant avec l'hypothèse d'indépendance et d'identité des lois des termes aléatoires du modèle Logit Multinomial). Cette normalisation n'affecte pas les probabilités de choix (en vertu du principe de l'échelle arbitraire d'utilité) :

$$U'_{ni} = U_{ni}\sigma_n = V_{ni}\sigma_n + \epsilon_{ni}.$$

En pratique (voir Tutz (2021)), le paramètre σ_n est souvent modélisé par l'exponentielle de la combinaison linéaire des caractéristiques observables de l'individu n (qui peuvent être distinctes ou incluses dans $\mathbf{X_n}$). On note $\mathbf{W_n} = (W_1, .., W_L) \in \mathbb{R}^L$ et $\gamma = (\gamma_1, .., \gamma_L) \in \mathbb{R}^L$ respectivement comme le vecteur des variables caractéristiques prises en compte et le vecteur de paramètres correspondants, avec L représentant le nombre total de variables considérées. La probabilité de choisir l'option i est définie par :

$$P_n(i|\mathcal{C}_n) = \frac{e^{V_{ni}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}{\sum_{j=1}^{I_n} e^{V_{nj}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}.$$

La différence en probabilité entre deux choix i et j s'écrit donc :

$$\frac{P_n(i|\mathcal{C}_n)}{P_n(j|\mathcal{C}_n)} = \frac{e^{V_{ni}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}{e^{V_{nj}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}} = e^{e^{\gamma^{\mathbf{T}}\mathbf{W_n}}(V_{ni}-V_{nj})}.$$

Même si la formule de probabilité de ce modèle diffère de celle du modèle logit habituel, les paramètres sont fixes et donc peuvent être estimés par la méthode du maximum de la vraisemblance. On peut écrire la vraisemblance et le log-vraisemblance comme suit :

$$\mathcal{L}(\beta,\gamma,\alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}) = \prod_{n=1}^{N}\prod_{i=1}^{I_n} P_n(i|\mathcal{C}_n)^{y_{ni}} = \prod_{n=1}^{N}\prod_{i=1}^{I_n}(\frac{e^{V_{ni}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}{\sum_{j=1}^{I_n} e^{V_{nj}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}})^{y_{ni}}$$
$$= \prod_{n=1}^{N}\prod_{i=1}^{I_n}(\frac{e^{(\beta_{i0}+\sum_{k=1}^{K_i}\alpha_{ik}S_{nik})e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}{\sum_{j=1}^{I_n} e^{(\beta_{j0}+\sum_{k=1}^{K_j}\alpha_{jk}S_{njk})e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}})^{y_{ni}}.$$

La log-vraisemblance s'écrit donc :

$$l(\beta,\gamma,\alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}) = ln(\mathcal{L}(\beta,\gamma,\alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]})) = ln(\prod_{n=1}^{N}\prod_{i=1}^{I_n} P_n(i|\mathcal{C}_n)^{y_{ni}}) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni}ln(P_n(i|\mathcal{C}_n))$$
$$= \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni}ln(\frac{e^{V_{ni}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}{\sum_{j=1}^{I_n} e^{V_{nj}e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni}ln(\frac{e^{(\beta_{i0}+\sum_{k=1}^{K_i}\alpha_{ik}S_{nik})e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}{\sum_{j=1}^{I_n} e^{(\beta_{j0}+\sum_{k=1}^{K_j}\alpha_{jk}S_{njk})e^{\gamma^{\mathbf{T}}\mathbf{W_n}}}}).$$

Ce modèle prend en compte l'hétérogénéité déterministe explicite dans la perception individuelle. En effet, le paramètre σ_n représente l'échelle d'utilité considérée, et le fait que chaque individu ou classe d'individus admet un σ_n différent implique que leurs différences d'utilité entre deux options varient en termes d'échelle. Cela résulte en ce que, pour certaines personnes, la différence entre deux choix semble beaucoup plus importante que pour d'autres personnes pour les deux mêmes options, car leurs échelles sont plus grandes.

Les paramètres γ étant constants à travers la population, on peut dire que c'est un effet hétérogène moyen. Malgré la simplicité du modèle encore en formule fermée, ce type de spécification d'hétérogénéité basée sur les caractéristiques de l'individu ne serait robuste que si la vraie source d'aléatoire comporte de la même façon que celui supposé par le modèle. Il est très important de tester l'adaptation du modèle aux données.

4.3.2 Modèle mixed logit (MXL) : Cas de variables continues

En ce qui concerne l'hypothèse IANP, qui rend le modèle MNL, le plus basique et facile à appliquer, peu efficace dans de nombreuses études réelles, le modèle Probit tend à capturer davantage d'inférences statistiques dans les comportements, mais il reste peu utilisé en raison de sa complexité d'estimation. En parallèle de ce modèle d'erreur normale, il existe un modèle doté d'une plus grande capacité, comme celui du probit, mais qui n'est devenu populaire que récemment grâce aux progrès de la capacité de simulation numérique des ordinateurs : le modèle logistique à loi mélangée (Mixed Logit Model). Ce modèle repose sur l'hypothèse du modélisateur qu'il existe une autre source aléatoire que celle de la loi de Gumbel, ou simplement que le terme aléatoire est une somme de lois différentes (et capable de rendre les termes d'utilité dépendants à travers les alternatives).

Dans le guide de Train (2009), on peut formuler cet effet aléatoire par un vecteur aléatoire ξ_n que l'on connaît à priori (par hypothèse posée pour la modélisation) sa distribution, et en conditionnant par ce vecteur, les termes aléatoires restants sont indépendants et identiquement distribués selon une loi de Gumbel (autrement dit, le modèle devient un MNL avec ξ fixé). Dans ce cas, la probabilité de choix de l'alternative i par l'individu n est donnée par :

$$P(i|\mathcal{C}_n) = \int_{\xi} P(i|\mathcal{C}_n,\xi_n=\xi)f(\xi)d\xi. \qquad (**)$$

Où $f()$ est la distribution de ξ et $P(i|\mathcal{C}_n, \xi_n = \xi)$ est la probabilité du modèle logit conditionné au paramètre ξ :

$$P(i|\mathcal{C}_n, \xi_n = \xi) = \frac{e^{V_{ni}(\xi)}}{\sum_{j=1}^{I_n} e^{V_{nj}(\xi)}}.$$

La probabilité associée au modèle logit mixte résulte d'une moyenne pondérée de la formule logit évaluée à diverses valeurs de ξ, et ces pondérations sont déterminées par la densité $f(\xi)$. Dans la terminologie statistique, cette moyenne pondérée de plusieurs fonctions est appelée une fonction mixte, et la densité qui définit les pondérations est désignée comme la distribution de mélange. Le modèle logit mixte peut être considéré comme une combinaison de la fonction logit calculée pour différentes valeurs de ξ, avec ξ agissant comme la distribution de mélange.

Le modèle logit standard, quant à lui, représente un cas particulier où la distribution de mélange $f(\xi)$ est dégénérée avec des paramètres fixes b : $f(\xi) = 1$ lorsque $\xi = b$ et égale à zéro pour $\xi \neq b$. En conséquence, la probabilité de choix ci-dessus se réduit à la formule logit classique.

McFadden et Train (2000b) ont montré un résultat important de ce modèle. Selon ses conclusions, sous les bonnes spécifications du modèle Logit à loi de mélange continue (ou Mixed Multinomial Logit - MMNL), on peut approximer un modèle d'utilité aléatoire avec n'importe quelle erreur donnée. Ce résultat est l'une des motivations de l'utilisation largement répandue dans la littérature et dans la pratique.

Méthode de simulation Monte Carlo

La progression significative de ces modèles découle de l'augmentation considérable de la puissance de calcul des ordinateurs au fil des décennies. Aujourd'hui, les ordinateurs disposent d'une mémoire considérable et d'une vitesse de lecture très rapide, ce qui réduit considérablement le temps nécessaire pour effectuer des calculs d'intégrales (ou des approximations d'intégrales) que les machines du passé ne pouvaient pas réaliser. Pour évaluer les intégrales définies dans l'équation (**), il est nécessaire de recourir à des méthodes de simulation Monte Carlo et des techniques de réduction de la variance afin d'obtenir une approximation de l'intégrale.

La loi faible des grands nombres nous indique que pour une suite de variables aléatoires $(X_m)_{0 \leq m}$ indépendantes et identiquement distribuées selon la même loi de variable X, dont l'espérance mathématique $E(X)$ existe, la moyenne empirique $\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ tend vers $E(X)$ en probabilité. Ainsi, l'estimateur $\overline{X}_m$ est à la fois convergent et non biaisé pour l'espérance $E(X)$. De plus, si $g(.)$ est une fonction mesurable, $\overline{g}_m(X)$ est également un estimateur convergent et non biaisé de $E(g(X))$:

$$\overline{g}_m = \frac{1}{m}\sum_{i=1}^{m} g(X_i) \xrightarrow[m \to +\infty]{\mathcal{P}} E[g(X)].$$

Par conséquent, si X suit une densité $f_X(.)$, nous pouvons estimer la quantité $\int_x g(x) f_X(x) dx$ en utilisant des échantillons générés à partir de $f_X(.)$. En appliquant cette approche à la formule *, en supposant que nous connaissions la famille de lois de ξ mais pas sa distribution exacte (nous présumons à priori que ξ suit une loi de cette famille), nous pouvons estimer à la fois les paramètres du modèle et les paramètres de la distribution de ξ par la méthode du maximum de vraisemblance. L'approximation de la probabilité de choix est ainsi formulée dans l'expression de la log-vraisemblance comme suit :

$$l(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}, \mu) = \ln(\mathcal{L}(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}, \mu)) = \ln(\prod_{n=1}^{N}\prod_{i=1}^{I_n} P_n(i|\mathcal{C}_n)^{y_{ni}}) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(P_n(i|\mathcal{C}_n))$$

$$\approx \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\frac{1}{M}\sum_{m=1}^{M} P_n(i|\mathcal{C}_n, \xi_{nm})) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\frac{1}{M}\sum_{m=1}^{M} \frac{e^{V_{ni}(\xi_{nm})}}{\sum_{j=1}^{I_n} e^{V_{nj}(\xi_{nm})}}).$$

Une fois que les paramètres sont estimés, on peut estimer la probabilité de choix i par simulation Monte Carlo, et plus le nombre de simulations M est élevé, plus la probabilité est bien approximée :

$$P(i|\mathcal{C}_n) = \int_{\xi} P(i|\mathcal{C}_n, \xi_n = \xi) f(\xi) d\xi \approx \frac{1}{M} \sum_{m=1}^{M} \frac{e^{V_{ni}(\xi_{nm})}}{\sum_{j=1}^{I_n} e^{V_{nj}(\xi_{nm})}}.$$

Malgré les capacités de calcul actuelles des supercalculateurs, l'estimation de ces modèles de lois de mélange continu sur un ordinateur classique reste relativement lourde, car cet estimateur n'est toujours pas très performant pour un petit nombre de simulations. En effet, sa vitesse de convergence est de l'ordre de $\mathcal{O}(\frac{1}{\sqrt{M}})$, ce qui signifie que pour réduire l'erreur de 10 fois, il faudrait augmenter le nombre de simulations de 100 fois, ce qui peut être très coûteux en termes de calculs. Dans le cas où $g(\xi)$ est intégrable, l'erreur quadratique moyenne de l'estimateur de Monte Carlo est donnée par :

$$\mathbb{E}\left[\{\bar{g}_M - \mathbb{E}[g(\xi)]\}^2\right] = \frac{\sigma^2}{M} =, \quad \text{où} \quad \sigma^2 = \text{Var}[g(\xi)].$$

Pour surmonter ce problème, diverses méthodes de réduction de la variance sont disponibles, notamment l'utilisation de variables antithétiques, l'échantillonnage préférentiel, etc. Une autre approche consiste à utiliser des méthodes quasi aléatoires telles que la suite de Halton, qui permettent d'améliorer la convergence de l'estimateur en répartissant de manière plus efficace les échantillons.

MXL : Modèle Factor Analysis

Ce modèle flexible, parfois appelé modèle Logit Kernel (Ben-Akiva et al. (2001)), peut être formulé comme une Analyse Factorielle. Cependant, ce modèle est relativement complexe à introduire dans un cadre simplifié du mémoire, notamment en raison des contraintes de mémoire et du caractère parcimonieux des informations disponibles dans le domaine de l'assurance santé, en grande partie dues aux lois de protection des assurés. Par conséquent, nous nous référerons à l'utilisation de ce modèle plus général telle que décrite dans l'article de référence Walker et al. (2004). Par la suite, nous examinerons deux cas simplifiés de ce modèle dans le contexte des choix discrets : le modèle de variance spécifique à l'alternative et le modèle à coefficient aléatoire.

MXL : Modèle de variance spécifique à l'alternative

Considérons l'hypothèse selon laquelle certains choix ou produits dans le panier peuvent susciter une réflexion plus variable chez les agents, par exemple, un produit totalement nouveau sur le marché. Les agents confrontés à un nouveau choix pourraient donc hésiter entre tester les caractéristiques ou revenir à leur choix habituel, ce qui entraînerait une grande variation d'utilité accordée au nouveau produit à travers la population, car chacun a sa propre référence en ce qui concerne la nouveauté par rapport à l'habitude. On peut donc formuler cette variation dans le terme aléatoire $\epsilon_{ni, i\in\{1,..,I_n\}}$ pour une configuration linéaire de la partie déterministe de l'utilité :

$$\forall i \in \{1, .., I_n\}, U_{ni} = V_{ni} + \eta_{ni} + \epsilon_{ni} = \beta_{i0} + \sum_{k=1}^{K_i} \alpha_{ik} S_{nik} + \eta_{ni} + \epsilon_{ni},$$

où :

- ϵ_{ni} sont iid de loi Gumbel standard $\forall n, i$.
- $\eta_{ni, i\in\{1,..,I_n\}}$ sont indépendants et peuvent être de lois différentes (dans la pratique souvent de même loi mais de variance différente), les vecteurs aléatoires $\eta_{\mathbf{n}} = (\eta_{n1}, .., \eta_{nI_n})$ sont iid de la même distribution $\forall n \in \{1, .., N\}$.

On a donc, par la formulation du paragraphe précédent, noté $V_{ni}(\eta_{\mathbf{n}}) = \beta_{i0} + \sum_{k=1}^{K_i} \alpha_{ik} S_{nik} + \eta_{ni}, \forall i \in \{1, .., I_n\}$.

Pour illustrer le modèle avec une perturbation normale, supposons que $\eta_{ni} \sim \mathcal{N}(0, \sigma_i), \forall i \in \{1, .., I_n\}$ (les paramètres de la distribution ne varient pas parmi les individus) et que les η_{ni} sont indépendants, on obtient la densité du vecteur $\eta_{\mathbf{n}} = (\eta_{n1}, .., \eta_{nI_n})$ comme suit :

$$f_{\eta_{\mathbf{n}}}(x_1, .., x_{I_n}) = \prod_{i=1}^{I_n} f_{\eta_{ni}}(x_i) = \prod_{i=1}^{I_n} \phi_{(0,\sigma_i)}(x_i) = \prod_{i=1}^{I_n} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2\sigma_i^2}\right).$$

On voit donc que simuler les vecteurs $\eta_{\mathbf{n}}$ revient à simuler indépendamment chaque composant normal centré de ce vecteur.

Afin de pouvoir estimer tous les paramètres du modèle avec perturbation normale par maximisation de la vraisemblance, il convient d'écrire $\eta_{ni} = \sqrt{\sigma_i}\zeta_{ni}$ avec $\zeta_{ni} \sim \mathcal{N}(0, 1)$. Désormais, la log-vraisemblance peut être approchée par un estimateur de Monte Carlo comme suit :

$$l(\beta, \alpha_{\mathbf{i},\mathbf{i}\in\{\mathbf{1},..,\mathbf{I}\}}, \sigma_{i,i\in\{1,..,I\}}) = \ln(\mathcal{L}(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}, \sigma_{i,i\in\{1,..,I\}})$$

$$\approx \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\sum_{m=1}^{M} P_n(i|\mathcal{C}_n, \eta_{nm})) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\sum_{m=1}^{M} \frac{e^{V_{ni}(\eta_{nm})}}{\sum_{j=1}^{I_n} e^{V_{nj}(\eta_{nm})}})$$

$$= \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\sum_{m=1}^{M} \frac{e^{\beta_{i0}+\sum_{k=1}^{K_i} \alpha_{ik} S_{nik}+\eta_{nmi}}}{\sum_{j=1}^{I_n} e^{\beta_{i0}+\sum_{k=1}^{K_i} \alpha_{ik} S_{nik}+\eta_{nmi}}}) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\sum_{m=1}^{M} \frac{e^{\beta_{i0}+\sum_{k=1}^{K_i} \alpha_{ik} S_{nik}+\sqrt{\sigma_i}\zeta_{nmi}}}{\sum_{j=1}^{I_n} e^{\beta_{i0}+\sum_{k=1}^{K_i} \alpha_{ik} S_{nik}+\sqrt{\sigma_i}\zeta_{nmi}}}),$$

avec ζ_{nmi} iid $\sim \mathcal{N}(0, 1)(\forall n, m, i)$.

La procédure d'estimation peut donc être réalisée par une étape de maximisation de vraisemblance. Cependant, un tel modèle présente une limitation, à savoir qu'on ne peut pas estimer tous les paramètres spécifiques de l'alternative, comme mentionné dans Ben-Akiva et al. (2001). Selon eux, il existe une condition appelée "condition d'ordre" qui limite le nombre de paramètres que l'on peut estimer à $\frac{I(I-1)}{2} - 1$ spécifiquement liés à chaque alternative. Par exemple, dans le cas de deux alternatives, il n'est pas possible d'estimer les paramètres liés à la distribution σ_i. Dans un cas de trois alternatives, il est possible d'identifier jusqu'à 2 paramètres des distributions, de sorte qu'un σ_j doit être fixé à 0 mais ce n'est pas n'importe quelle variance d'alternative qui peut être restreinte. Le processus de sélection des paramètres à contraindre est expliqué dans Walker et al. (2004), et selon eux, la tendance à restreindre à zéro la variance de l'alternative avec la plus petite variance semble valide pour permettre une bonne estimation du modèle. Pour la suite, nous effectuerons l'estimation de ce modèle pour 3 choix d'alternative en deux étapes :

1. L'estimation du modèle avec tous les paramètres de variance de chaque alternative sans avoir restreint aucun paramètre à zéro.

2. La sélection du paramètre de la variance le plus bas parmi tous les paramètres de variance estimés du modèle précédent, puis réestimer le modèle avec le paramètre de la variance choisi fixé à zéro. Nous obtiendrons le modèle identifiable.

MXL : Modèle de coefficient aléatoire

Le modélisateur peut, dans plusieurs cas, dire qu'il y a une hétérogénéité dans la façon dont les individus accordent de l'importance à certains attributs. En effet, la taille du produit peut être importante pour certaines personnes et, en même temps, ne pas peser beaucoup sur le choix d'autres individus. Une telle hétérogénéité pourrait être capturée en disant que le coefficient associé à l'attribut considéré est aléatoire et n'est pas observable. On pourrait très bien supposer que la distribution de

ce paramètre est connue et donc calculer la probabilité de choix comme l'espérance de la probabilité conditionnelle à ces paramètres. Par exemple, soit $S'_{ni,i\in\{1,..,I_n\}}$ l'attribut pour lequel on définit que son coefficient η_{ni} est aléatoire suivant une distribution $\mathcal{L}(\mu_i)$ et associé linéairement à cet attribut dans l'utilité $V_{ni,i\in\{1,..,I_n\}}$, on peut le réécrire sous la forme linéaire d'utilité comme pour le modèle Logit :

$$\forall i \in \{1,..,I_n\}, U_{ni} = V_{ni}(\eta) + \epsilon_{ni} = \beta_{i0} + \sum_{k=1}^{K_i} \alpha_{ik} S_{nik} + \eta_{ni} S'_{ni} + \epsilon_{ni}.$$

Les coefficients η_{ni} peuvent être variés par l'individu seulement ($\eta_{ni} = \eta_n, \forall i$) ou par alternative aussi ($\eta_{ni,i\in\{1,..,I_n\}}$) mais il y a une légère différente lors de procédure d'estimation avec modèle de variance spécifique à alternative :

- Si le coefficient aléatoire est associé à des dummies variables attributs spécifiquement de l'alternative, on se retrouve dans le cas similaire de modèle de variance spécifique à alternative. La condition d'ordre devrait appliquer sur les paramètres de distribution $\mu = (\mu_1, .., \mu_I)$, il faut donc normaliser certain paramètre de μ afin de pouvoir identifier le modèle.

- Si le coefficient aléatoire est associé à des variables liées au individu, on peut identifier autant de paramètre de distribution que les données permettent (mais il est conseiller de tenter pour des modèles parcimonieuses).

La différence fondamentale lors de l'application de ces deux modèles découle essentiellement de la connaissance du problème par le modélisateur. Si un attribut commun de la plupart de l'option dans le panier est susceptible d'influencer les comportements de manière différente pour chaque assuré en raison de l'appréciation propre à chaque individu - par exemple, la prime d'assurance ou le délai de remboursement, ainsi que la qualité des services administratifs dans le cas de l'assurance - alors il est plus incitatif de capturer cette hétérogénéité par le biais d'un coefficient aléatoire. En revanche, si un trait de l'option dans le panier a pour objectif de fournir des incitations différentes à chaque individu en raison de son caractère innovant ou de sa nature différente par rapport aux autres options du panier, il peut être préférable de spécifier plutôt la variance spécifique de l'alternative.

4.3.3 Modèle de Classe Latente : Cas discret de MXL

Les modèles de type lois mélangées continues associent des effets aléatoires à chaque individu en supposant que l'hétérogénéité partielle d'utilité individuelle soit distribuée selon une distribution connue. Même si leur grande capacité à capturer les comportements des assurés est reconnue, ils peuvent sembler trop coûteux en termes de calcul en raison des nombreuses simulations Monte Carlo nécessaires pour approcher au mieux les intégrations dans la log-vraisemblance. Une solution moins coûteuse, tout en permettant d'approcher l'hétérogénéité, consiste à supposer qu'il existe des classes de personnes avec le même degré d'hétérogénéité lorsque l'on dispose des informations liées à leur référence. Une telle information peut être déduite des connaissances sur la population étudiée à travers des méthodes de détection de classe latente existantes. Les méthodes de recherche des classes latentes consistent à rechercher et à expliquer les classes d'homogénéité cachées dans la population à partir des caractéristiques observées, telles que les indicateurs continus comme le poids, la taille, ou les indicateurs ordinaux comme le type de sang, le statut de fumeur, etc. Il y a deux types d'approches généralement utilisées en fonction de la nature des variables considérées :

- Méthode de classification non supervisée (K-means, classification hiérarchique) : Ces méthodes de classification non supervisée cherchent à regrouper des individus jugés similaires en utilisant des mesures de distance spécifiques pour lier des attributs continus. Leur principal avantage est leur très bonne performance dans le domaine continu, ainsi que leur flexibilité en matière de

métrique choisie, ce qui rend les clusters plus interprétables. Le seul inconvénient est qu'elles ne s'appliquent pas aux variables discrètes.

- Analyse en classes latentes : L'Analyse en Classes Latentes (LCA) est une méthode statistique qui consiste à construire des classes latentes, c'est-à-dire des groupes non observés, à partir de données observées sur un ensemble de variables indicatrices. Ces classes sont créées en regroupant des cas similaires en fonction de leurs réponses aux variables observées. Chaque classe est représentée par une catégorie distincte d'une variable latente. Contrairement aux méthodes de clustering, la LCA opère selon un modèle statistique qui associe chaque individu à un groupe de manière probabiliste. L'avantage de la LCA par rapport à ces méthodes réside dans la possibilité de sélectionner des modèles en utilisant des critères de sélection, d'estimer les probabilités d'appartenance à chaque classe, et de prendre en compte des variables de différentes échelles (continues, ordinales ou nominales) au sein du même modèle.

Une fois que l'on postule l'existence de classes latentes dans la population, par exemple deux classes avec deux types d'utilité différents U_{classe_1} et U_{classe_2}, on peut calculer la probabilité de choix a priori lorsqu'un individu appartient à l'une des deux classes. Pour tout $i \in C_n$, soit $U_{ni,classe_1} = V_{ni,classe_1} + \epsilon_{ni,classe_1}$ et $U_{ni,classe_2} = V_{ni,classe_2} + \epsilon_{ni,classe_1}$ respectivement, l'utilité que la personne n'accorde à l'option i si elle appartient au sous-groupe 1 (respectivement 2). Dans le cas le plus simple, on suppose qu'au sein d'une classe latente, tout le monde est homogène en préférence, ce qui donne que chaque $V_{ni,classe_l} (i \in C_n, l \in \{1;2\})$ prend la forme linéaire avec les coefficients fixes, comme la partie déterministe d'un modèle logit multinomial, sauf que les coefficients changent en fonction de la classe latente :

$$\forall n \in \{1,..,N\}, \forall i \in \mathcal{C}_n, \forall l \in \{1,2\} : V_{nil} = \beta_{i0l} + \sum_{k=1}^{K_i} \alpha_{ikl} S_{nik} = \beta_{i0l} + S_{ni}^T \alpha_{il}.$$

Notons l_n la classe latente à laquelle appartient l'individu n. On fait l'hypothèse que toutes les classes latentes fournissent le même ensemble de choix $\mathcal{C}_\backslash$. Les probabilités de choix sachant que l'individu appartient à une classe sont données par la formule du modèle logit :

$$P_n(i|\mathcal{C}_n, l_n = l) = \frac{e^{V_{nil}}}{\sum_{j=1}^{I_n} e^{V_{njl}}}.$$

Les classes latentes étant inobservables individuellement, grâce aux processus de détection de classe latente, on peut supposer qu'il y a une probabilité discrète d'affectation des classes latentes $\pi_{\mathbf{n}} = (\pi_{nl})_{l \in \{1,2\}}$. On peut déduire la probabilité de choix moyen sans savoir à quelle classe (2 classes) sont associés les individus :

$$P_n(i|\mathcal{C}_n) = \pi_{n1} P_n(i|\mathcal{C}_n, l_n = 1) + \pi_{n2} P_n(i|\mathcal{C}_n, l_n = 2) = \sum_{l=1}^{2} \pi_{nl} \frac{e^{V_{nil}}}{\sum_{j=1}^{I_n} e^{V_{njl}}}.$$

L'essentiel du modèle réside dans la maximisation de l'information apportée par la vraisemblance. Même si la forme exacte de la distribution de l_n n'est pas connue explicitement, il est tout à fait possible de l'approcher par un modèle de type logistique multinomial ou probit multinomial. Dans notre cas, qui se limite à deux classes, nous supposons que les classes latentes peuvent être capturées linéairement à travers les caractéristiques sociodémographiques observables. De plus, nous supposons que les probabilités d'appartenance à chaque classe peuvent être approchées par une régression logistique. Ainsi, la probabilité d'être dans la classe 1 selon le modèle logistique peut être formulée comme suit :

$$\pi_{n1} = P(l_n = 1|X_n) = \frac{1}{1 + e^{\mathbf{X'_n}^{\mathbf{T}}\tau}} = \frac{1}{1 + e^{\sum_{q=1}^{Q} X'_{nq}\tau_q}},$$

$$\pi_{n2} = P(l_n = 2|X_n) = 1 - P(l_n = 1|Xn) = 1 - \frac{1}{1 + e^{\sum_{q=1}^{Q} X'_{nq}\tau_q}}.$$

Avec :

- $\mathbf{X'_n} = (X'_{nq})_{q\in\{1,..,Q\}}$ le vecteur d'attributs observables de l'individu pris en compte pour la prédiction de classe latente (on suppose que $\mathbf{X_n}$ comporte toutes les informations observables de l'individu n, donc $\mathbf{X'_n} \subset \mathbf{X_n}$).
- $\tau = (\tau_q)_{q\in\{1,..,Q\}}$ le vecteur de paramètres à estimer du modèle par algorithme de maximum de vraisemblance.

De même façon, on peut utiliser le modèle de régression probit (un cas spécial du modèle probit multinomial décrit en B.4) avec $\Phi(.)$, la fonction de répartition d'une loi normale centrée réduite, et on obtient :

$$\pi_{n1} = P(l_n = 1|X_n) = \Phi(\mathbf{X'_n}^{\mathbf{T}}\tau) = \Phi(\sum_{q=1}^{Q} X'_{nq}\tau_q) \text{ et } \pi_{n2} = P(l_n = 2|X_n) = 1-P(l_n = 1|X_n) = 1-\Phi(\sum_{q=1}^{Q} X'_{nq}\tau_q).$$

Une fois que tous les paramètres sont spécifiés clairement dans la composante de l'utilité ainsi que dans la composition de classe latente, nous pouvons donc estimer en une étape de maximisation de vraisemblance totale sur l'ensemble d'observation. En effet, on écrit la log-vraisemblance comme :

$$l(\beta_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \alpha_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \tau) = \ln(\mathcal{L}(\beta_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \alpha_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \tau)$$
$$= \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(P_n(i|\mathcal{C}_n)) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\sum_{l=1}^{2} \pi_{nl} \frac{e^{V_{nil}}}{\sum_{j=1}^{I_n} e^{V_{njl}}}).$$

Selon les sous-modèles de classification de classe latente, on distingue les deux cas :

Sous-modèle Logit : $l(\beta_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \alpha_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \tau) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\frac{1}{1+e^{\sum_{q=1}^{Q} X'_{nq}\tau_q}} \times \frac{e^{V_{ni1}}}{\sum_{j=1}^{I_n} e^{V_{nj1}}} + (1 - \frac{1}{1+e^{\sum_{q=1}^{Q} X'_{nq}\tau_q}}) \times \frac{e^{V_{ni2}}}{\sum_{j=1}^{I_n} e^{V_{nj2}}})$.

Sous-modèle Probit : $l(\beta_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \alpha_{\mathbf{l,l}\in\{\mathbf{1,2}\}}, \tau) = \sum_{n=1}^{N}\sum_{i=1}^{I_n} y_{ni} \ln(\Phi(\sum_{q=1}^{Q} X'_{nq}\tau_q) \times \frac{e^{V_{ni1}}}{\sum_{j=1}^{I_n} e^{V_{nj1}}} + (1 - \Phi(\sum_{q=1}^{Q} X'_{nq}\tau_q)) \times \frac{e^{V_{ni2}}}{\sum_{j=1}^{I_n} e^{V_{nj2}}})$.

Il faut faire attention ici que la log-vraisemblance n'est plus concave en global, donc il peut y avoir des solutions locales du problème d'optimisation. Comme le fait d'ajouter une classe latente augmente significativement le nombre de paramètres à identifier et donc le temps d'estimation du modèle, il est conseillé de considérer que deux classes latentes dans la population lorsque l'étude préalable sur les données ne détecte pas plus de deux grands clusters.

L'étude menée par Zhou (2017) a réalisé une comparaison entre les modèles de classe latente et les modèles à loi de mélange continue en ce qui concerne les préférences en matière de santé. Elle a tiré une conclusion favorable pour l'utilisation des modèles de classe latente dans le domaine du traitement médical. Dans la même optique, l'étude menée par Shen (2009) compare les modèles de classe latente et les modèles à loi de mélange continue en se basant sur les données de choix de transport. Les résultats ont montré que le modèle de classe latente performe statistiquement mieux que le modèle de mélange

de loi continue en termes de probabilités prédites. Malgré certaines limitations, telles que l'hypothèse d'une distribution discrète des préférences, le modèle de classe latente a montré sa capacité à mieux représenter les groupes de préférences au sein de la population étudiée et offre une approche plus facilement interprétable que d'autres modèles. En effet, par comparaison des paramètres de différentes classes, nous pourrions voir les différences de perception sur certaines caractéristiques de l'alternative.

4.4 Test d'hypothèse - Validation du modèle

La validation d'un modèle constitue une étape cruciale dans le processus de modélisation statistique. Cette démarche suit généralement un ordre bien défini visant à garantir la fiabilité des résultats. Tout d'abord, nous effectuons un test des coefficients du modèle (les paramètres de la fonction d'utilité déterministe) afin de déterminer s'ils sont significativement différents de zéro. Cela nous permet d'identifier les variables explicatives pertinentes. Ensuite, nous pouvons explorer les tests de modèles emboîtés en incorporant progressivement des variables ou des termes pour évaluer leur impact sur la qualité globale du modèle. Parmi les modèles de différents types de spécifications présentés ci-dessus, le test de modèle non emboîté peut révéler si, par exemple, un type de spécification est plus adéquat que l'autre. Enfin, pour conclure la procédure de validation, il est crucial de comparer la performance des modèles considérés. Cette comparaison peut s'effectuer en utilisant diverses mesures d'ajustement du modèle, telles que le critère d'information d'Akaike (AIC) ou le critère d'information Bayésien (BIC).

4.4.1 Test de l'importance des coefficients

Lorsque nous cherchons à évaluer l'importance d'un attribut pour le modèle, nous réalisons un test d'hypothèse visant à déterminer si le coefficient estimé de cet attribut, noté $\bar{\beta}$ dans les formules d'utilité, est significativement différent de zéro ($\mathcal{H}_1$). Pour ce faire, nous utilisons le test de Student avec la statistique de test $t = \sqrt{n}\frac{\bar{\beta}-0}{Var(\bar{\beta})}$, où $Var(\bar{\beta})$ représente l'estimateur non biaisé de la variance de $\bar{\beta}$ et on suppose que $\bar{\beta}$ suit une loi normale. Selon les résultats du test statistique, nous savons que lorsque le nombre d'observations dans l'échantillon est suffisamment grand, la statistique de test suit approximativement une distribution de Student ($t \sim \text{Student}(N-1)$). Si l'échantillon est de taille importante, la distribution de Student tend à se rapprocher d'une distribution normale $\mathcal{N}(0,1)$, pour laquelle nous pouvons calculer la p-valeur sous l'hypothèse nulle comme suit : $p = 2(1-\Phi(t))$. Par conséquent, nous pouvons rejeter l'hypothèse nulle à un niveau de confiance de 95 % lorsque $p < 0,05$, ce qui signifie que la variable joue un rôle significatif dans le modèle. Ce test peut également être appliqué pour tester si un paramètre est vraiment égal à une valeur $b \neq 0$, en utilisant la statistique $t = \sqrt{n}\frac{\bar{\beta}-b}{Var(\bar{\beta})}$.

4.4.2 Test du modèle emboîté

Le test du modèle emboîté - souvent appelé test de rapport de vraisemblance, est employé lorsque nous souhaitons comparer deux modèles, $\mathcal{M}_0$ et $\mathcal{M}_1$, où $\mathcal{M}_0$ est une version avec des restrictions linéaires imposées sur les paramètres de $\mathcal{M}_1$. Dans ce scénario, nous parlons de modèles emboîtés, et l'hypothèse nulle notée $\mathcal{H}_0 : \mathcal{M} = \mathcal{M}_0$ (c'est-à-dire que $\mathcal{M}_0$ est le vrai modèle), est testée contre l'hypothèse alternative $\mathcal{H}_1 : \mathcal{M} = \mathcal{M}_1$ (c'est-à-dire que $\mathcal{M}_0$ n'est pas le vrai modèle). Selon les travaux de Wilks (1938), sous l'hypothèse nulle, nous avons la statistique suivante :

$$-2\left(l\left(\hat{\beta}_R\right) - l\left(\hat{\beta}_U\right)\right) \sim \chi^2_{(K_U-K_R)} \text{ ou } -2\frac{\mathcal{L}(\hat{\beta}_R)}{\mathcal{L}(\hat{\beta}_U)} \sim \chi^2_{(K_U-K_R)},$$

où

- $l(\hat{\beta}_R) = ln(\mathcal{L}(\hat{\beta}_R))$ est le log-vraisemblance du modèle $\mathcal{M}_0$,
- $l(\hat{\beta}_U) = ln(\mathcal{L}(\hat{\beta}_U))$ est le log-vraisemblance du modèle $\mathcal{M}_1$,
- K_R est le nombre de paramètres du modèle $\mathcal{M}_0$,
- K_U est le nombre de paramètres du modèle $\mathcal{M}_1$.

Nous pouvons rejeter l'hypothèse nulle à un niveau de confiance de 95% lorsque $-2\left(l(\hat{\beta}_R) - l(\hat{\beta}_U)\right) > q_{\chi_{(K_U-K_R)^2,95\%}}$, où $q_{\chi_{(K_U-K_R)^2,95\%}}$ représente le quantile à 95% de la distribution $\chi_{(K_U-K_R)^2}$.

Ce test revêt une importance particulière lors de l'évaluation de l'hétérogénéité de préférence, comme dans le cas des paramètres de loi de distribution dans les modèles à loi de mélange continue. Ces paramètres peuvent être contraints linéairement, ce qui équivaut à exclure l'effet aléatoire. Par exemple, si nous forçons $\sigma = 0$ dans un modèle avec une distribution normale sous-jacente $\mathcal{N}(0, \sigma)$, le modèle résultant est dépourvu d'effet aléatoire, ce qui signifie qu'il n'y a pas d'hétérogénéité individuelle. Le test de log-vraisemblance nous permet donc de déterminer si le modèle sans effet aléatoire est préférable, indiquant ainsi l'absence d'hétérogénéité individuelle, ou s'il faut envisager une forme différente d'hétérogénéité.

En partant de ce test, les tests des deux modèles non emboîtés sont ainsi construits et peuvent être utilisés dans la pratique (présents dans l'annexe B.5), mais nous n'étendons pas notre application à leur utilisation dans le cadre de ce mémoire.

4.4.3 Test d'hypothèse IANP

Il existe bien sûr dans la littérature des tests pour l'hypothèse IANP, avec lesquels on peut déterminer s'il y a une nécessité de développer d'autres modèles plus compliqués ou si l'on peut se contenter des modèles logit multinomiaux. D'après Small et Hsiao (1985), il existe deux types de test de violation de l'hypothèse IANP :

- Tests par partitionnement de l'ensemble des choix.
- Tests par la spécification du modèle.

Le premier type de test est couramment utilisé dans la pratique, car il permet d'utiliser uniquement les modèles logit multinomiaux et d'éviter en avance l'estimation coûteuse de modèles plus compliqués s'ils ne sont pas nécessaires. Il comprend des tests comme le test de Hausman-McFadden (Hausman et McFadden (1984)) et le test de Small-Hsiao (Small et Hsiao (1985)). Malgré leur large application sur divers ensembles de données, les tests de Hausman-McFadden et Small-Hsiao ont été jugés peu performants, même sur des données de grande taille, selon des études de simulations réalisées par Fry et Harris (1996), Fry et Harris (1998) et Cheng et Long (2007). C'est la raison pour laquelle nous ne présentons pas ces tests dans ce mémoire, mais nous recourons au deuxième type de tests.

Le deuxième type de test repose davantage sur la spécification des modèles, car son idée est de tester si une spécification de l'hétérogénéité est correcte. En effet, l'hypothèse IANP se conserve lorsque le vrai modèle est un modèle logit multinomial. En estimant les modèles avec de l'hétérogénéité inobservable plus générale, dont le modèle logit multinomial est un cas particulier, puis en effectuant le test de modèle emboîté en imposant des contraintes sur les paramètres pour rendre le modèle logit multinomial, on peut simplement voir s'il est possible de rejeter l'hypothèse nulle. Dans le cas où le modèle logit multinomial est rejeté, on peut conclure que l'hypothèse IANP est violée. Cette approche nous semble plus logique dans un contexte d'exploration de préférences individuelles, où les spécifications seront testées pour leur adaptation aux données. Cependant, cette méthode présente un inconvénient majeur, à savoir que les modèles sont très paramétriques et que seules certaines bonnes spécifications peuvent mener au rejet de l'hypothèse IANP. Il faut donc estimer de nombreux modèles pour parvenir à une conclusion sur l'IANP.

4.4.4 Coefficient de ρ^2

L'indicateur ρ^2 (indicateur de rapport de vraisemblance) tel que présenté par Train (2009) est un coefficient calculé pour évaluer l'amélioration du modèle estimé (avec une log-vraisemblance $l(\hat{\beta})$) par rapport au modèle nul où toutes les utilités déterministes $V_{ni,\forall n,i}$ sont nulles, et la log-vraisemblance $l(0)$ vaut $-Nln(I)$ pour un modèle avec I options disponibles pour tous les individus. Ce coefficient est calculé comme suit :

$$\rho^2 = 1 - \frac{l(\hat{\beta})}{l(0)}.$$

Voici quelques points importants à noter à propos de ρ^2 :

- ρ^2 est toujours compris entre 0, lorsque le modèle tend vers le modèle nul sans paramètres à estimer, et 1, lorsque le modèle tend vers un modèle parfait avec une vraisemblance de 1.
- ρ^2 permet de comparer la performance relative de deux modèles différents, mais estimés sur les mêmes données, en choisissant le modèle avec le ρ^2 le plus élevé.
-

Cependant, il est important de noter que l'ajout de variables non pertinentes peut également augmenter ρ^2, ce qui peut conduire à un modèle surparamétré. C'est pourquoi le coefficient $\bar{\rho}^2$ (indicateur de rapport de vraisemblance ajusté) a été introduit pour prendre en compte la dimension du modèle dans l'évaluation. En notant K comme le nombre de paramètres du modèle estimé et $l(\hat{\beta})$ comme la log-vraisemblance du modèle estimé, $\bar{\rho}^2$ est calculé comme suit :

$$\bar{\rho}^2 = 1 - \frac{l(\hat{\beta} - K)}{l(0)}.$$

Lors de la comparaison de modèles sur la même base d'estimation, il est préférable de privilégier le modèle qui présente le coefficient $\bar{\rho}^2$ le plus élevé, car il tient compte de la parcimonie du modèle tout en évaluant sa performance.

Le log-vraisemblance, AIC et BIC comme critère de performance

Similaire au coefficient ρ^2, le log-vraisemblance sera utilisé comme un critère pour comparer les modèles sur la même base d'estimation, car il présente l'adaptation du modèle aux données.

On peut associer deux mesures souvent regardées lors de la procédure de sélection du modèle en apprentissage automatique : les critères d'information d'Akaike (AIC) et de Bayes (BIC). Les critères d'information d'Akaike (AIC) et de Bayes (BIC) sont des outils essentiels pour choisir entre différents modèles statistiques en évaluant leur ajustement aux données tout en prenant en compte la complexité du modèle. Les formules mathématiques sont les suivantes :

$$AIC = -2\ln(\mathcal{L}(\hat{\beta})) + 2K = -2l(\hat{\beta}) + 2K,$$

$$BIC = -2\ln(\mathcal{L}(\hat{\beta})) + K\ln(N) = -2l(\hat{\beta}) + K\ln(N),$$

où :

- $\mathcal{L}(\hat{\beta})$ est la vraisemblance finale du modèle avec tous les paramètres estimés (que l'on note tous dans $\hat{\beta}$).
- K est le nombre de paramètres du modèle.
- N est le nombre d'observations utilisées pour la procédure d'estimation.

On peut interpréter ces deux coefficients comme suit :

- L'idée derrière l'AIC est de trouver un équilibre entre la précision de l'ajustement (déviance plus faible) et la parcimonie (moins de paramètres). Ainsi, un modèle avec un AIC plus faible est préféré, car il offre un bon ajustement aux données avec un nombre minimal de paramètres. Cependant, l'AIC ne tient pas compte de la taille de l'échantillon.

- Le critère d'information bayésien (BIC), quant à lui, est également utilisé pour choisir entre différents modèles, mais il pénalise davantage les modèles avec un grand nombre de variables. Le BIC favorise la parcimonie en pénalisant fortement les modèles complexes, ce qui en fait un choix approprié lorsque la taille de l'échantillon est petite par rapport au nombre de paramètres ($\frac{N}{K} < 40$).

Si vous vous trouvez dans une situation où le nombre de paramètres est grand par rapport au nombre d'observations (c'est-à-dire lorsque $\frac{N}{K} < 40$), comme c'est souvent le cas dans les expériences de choix discrets, il peut être préférable d'utiliser le critère d'information d'Akaike corrigé. Ce critère est présenté par Hurvich et Tsai (1995) comme suit :

$$AIC_c = AIC + \frac{2K(K+1)}{N-K-1}.$$

Pour récapituler les méthodes appliquées à la suite, nous cherchons à évaluer quatre critères sur les modèles entraînés sur la même base afin de sélectionner le modèle le plus adapté au problème tout en étant parcimonieux. Ces critères sont les suivants :

- $\bar{\rho}^2$: nous préférons le modèle maximisant cette quantité.

- $l(\hat{\beta}) = \ln(\mathcal{L}(\hat{\beta}))$: nous préférons le modèle maximisant cette quantité.

- AIC : nous préférons le modèle minimisant cette quantité.

- BIC : nous préférons le modèle minimisant cette quantité.

4.5 Impact du changement d'utilité - Probabilité de transition des choix

L'application des modèles de choix discret reste très utile en raison de sa formulation probabiliste des utilités. Dans cette section, nous introduisons les probabilités de transition de choix comme une estimation projetée du modèle d'une période à l'autre sous des changements de caractéristiques des alternatives du panier de choix, en particulier pour le modèle Logit Multinomial.

4.5.1 Hypothèse d'indépendance entre utilités observables - inobservables

Restons dans nos notations, soit un individu n et le panier de choix $\mathcal{C} = \{1, .., I\}$ qui dérive les utilités systémiques et les utilités inobservables V_{ni},$\epsilon_{ni}(\forall i \in \mathcal{C})$. Nous pourrions donc estimer le modèle sur la base des choix observés dans les données, ce qui nous donne les coefficients du modèle.

Supposons que nous voulons changer les prix, les qualités des alternatives ou encore les niveaux de vie des individus (ceux qui sont observables) pour la prochaine période, ce qui entraînerait un changement dans les utilités systémiques. Ces changements mènent à de nouvelles utilités systémiques et inobservables V'_{ni},$\epsilon'_{ni}(\forall i \in \mathcal{C})$. On suppose que $\mathcal{C}$ englobe ainsi de nouveaux changements pour la simplification. Afin de pouvoir calculer les probabilités de transition des choix causées par ce changement, nous devons faire l'hypothèse suivante :

Hypothèse : Nous supposons que tous les facteurs observables ou partiellement observables sont tous présents dans l'utilité systémique et que la partie aléatoire de l'utilité ne contient que des facteurs inobservables. En d'autres termes, pour tout (n, i), V_{ni} et ϵ_{ni} sont indépendants ; les variables présentes dans V_{ni} ne sont pas endogènes pour ϵ_{ni}, et le composant aléatoire reste inchangé après des changements des variables dans le composant observable.

Avec cette hypothèse, nous pourrions dire que les nouveaux changements ne concernent que l'utilité systémique et donc que la partie aléatoire reste constante pendant le changement, avec $\epsilon'_{ni} = \epsilon_{ni} \forall i \in \mathcal{C}$ (Hypothèse du terme aléatoire inchangé, Delle Site et Salucci (2013)).

4.5.2 Probabilité de transition

Ce qui est intéressant dans cette application des modèles de choix discrets réside dans les probabilités de transition de choix, c'est-à-dire les pourcentages de personnes qui restent dans la même alternative après le changement d'utilité. Ces quantités permettent de quantifier les flux d'entrée-sortie des alternatives lors des changements de caractéristiques, comme par exemple les prix des alternatives. Nous notons arbitrairement que t est l'instance avant le changement et t' est l'instance après le changement. Ce que nous cherchons à calculer est donc la probabilité que la personne n, ayant choisi l'alternative i à l'instance t, la quitte pour choisir l'alternative j à l'instance t' $(\forall i, j \in \mathcal{C})$, en d'autres termes $P_n(Choix_t = i \cap Choix_{t'} = j|\mathcal{C})$. Nous simplifions les instances t et t', et nous retenons les notations suivantes en général pour tout individu :

- $P(Choix_t = i \cap Choix_{t'} = j|\mathcal{C}) = P_{i\to j}(V_1, \ldots, V_I, V'_1, \ldots, V'_I)$.
- $P(Choix_t = i|\mathcal{C}) = P_i(V_1, \ldots, V_I)$.
- $P(Choix_{t'} = i|\mathcal{C}) = P_i(V'_1, \ldots, V'_I)$.

Après Delle Site et al. (2022), considérons un changement tel que les termes aléatoires restent inchangés. Le changement d'utilité systémique est donné par :

$$\mathbf{V} = (V_1, \ldots, V_I) \to \mathbf{V}' = \left(V'_1, \ldots, V'_I\right).$$

Nous notons $\delta_i = V'_i - V_i, i \in [1 \ldots I]$ et supposons, sans perte de généralité, que $\delta_1 \leq \ldots \leq \delta_I$. La probabilité de transition du choix i au choix $j \in [1 \ldots I]$, $P_{i\to j}$, est donc donnée par le théorème 1 de de Palma et Kilani (2011) comme suit :

$$(\mathbf{V}, \mathbf{V}') \mapsto P_{i\to j}\left(\mathbf{V}, \mathbf{V}'\right) = \begin{cases} P_i\left(V_1 + (\delta_1 - \delta_i)^+, \ldots, V_I + (\delta_I - \delta_i)^+\right) & \text{si} \quad j = i \\ \int_{\delta_i}^{\delta_j} \Pi_i^j\left(V_1 + (\delta_1 - z)^+, \ldots, V_I + (\delta_I - z)^+\right) dz & \text{si} \quad j > i \\ 0 \quad \text{si} \quad j < i \end{cases}$$

où $x \mapsto x^+ = \max(x, 0)$ et $(x_1, \ldots, x_I) \mapsto \Pi_i^j(x_1, \ldots, x_I) = -\partial P_i(x_1, \ldots, x_I)/\partial x_j$.

Dans le cas particulier du modèle logit multinomial, de Palma et Kilani (2011) dérive la formule suivante :

$$(\mathbf{V}, \mathbf{V}') \mapsto P_{i\to j}\left(\mathbf{V}, \mathbf{V}'\right) = \begin{cases} \frac{e^{V_i}}{\Omega_i} \quad \text{si} \quad j = i \\ \sum_{r=i}^{j-1}\left(\frac{e^{V_i}}{\Omega_{r+1}} - \frac{e^{V_i}}{\Omega_r}\right)\frac{e^{V'_j}}{\sigma_r} \quad \text{si} \quad j > i \\ 0 \quad \text{si} \quad j < i \end{cases}$$

où :

$$\sigma_r = \sigma_0 - \sum_{k \leq r} e^{V'_k}, \quad r \in \{1, \ldots, I\}.$$

$$\sigma_0 = \sum_{k=1}^{I} e^{V'_k}.$$

$$\Omega_r = s_r + \sigma_r e^{-\delta_r}, \quad r \in \{1, \ldots, I\}.$$

$$s_r = \sum_{k \leq r} e^{V_k}, \quad r \in \{1, \ldots, I\}.$$

Les probabilités de transition peuvent être facilement vérifiées à l'aide des formules élémentaires de somme des probabilités :

$$\sum_{j=1}^{I} P_{i \to j}\left(\mathbf{V}, \mathbf{V}'\right) = P_i(V_1, \ldots, V_I) = P_i(\mathbf{V}),$$

$$\sum_{i=1}^{I} P_{i \to j}\left(\mathbf{V}, \mathbf{V}'\right) = P_j(V'_1, \ldots, V'_I) = P_j(\mathbf{V}').$$

En outre, on peut calculer la probabilité conditionnelle de changer pour l'alternative j sachant que l'individu a choisi l'alternative i auparavant comme suit :

$$P(Choix_{t'} = j | Choix_t = i, \mathcal{C}) = \frac{P_{i \to j}\left(\mathbf{V}, \mathbf{V}'\right)}{P_i(\mathbf{V})}.$$

Cette quantité est relativement importante dans notre contexte d'assurance car elle représente la probabilité qu'un assuré quitte le contrat actuel pour souscrire à un autre contrat après un changement de caractéristiques du contrat de la part de l'assureur. Cela permet à l'assureur d'estimer la perte associée à chaque type de contrat. Plus tard, au chapitre 5, nous évaluons la probabilité moyenne de transition entre les contrats pour les groupes homogènes d'assurés, ce qui peut être interprété comme la proportion de la population restante ou partant de chaque niveau de couverture.

Chapitre 5

Application à la modélisation de l'anti-sélection

5.1 Application des modèles dans chaque cas de portefeuille

Après avoir préparé des bases de données d'assurés de plus en plus sophistiquées, il est temps de tester les modèles décrits au chapitre 4 avec une procédure de post-validation. Les modèles complexes seront employés lorsque le modèle le plus simple de type Logit Multinomial n'est pas adapté et est rejeté par le test de rapport de vraisemblance pour les modèles emboîtés. Les modèles entraînés sur la même base seront ensuite comparés par la relation de dominance de type optimum Pareto décrite au chapitre 3. Pour l'entraînement de tous les modèles utilisés, nous utilisons la bibliothèque Pandas Biogeme en Python (voir Bierlaire (2016)), développée par le professeur Michel Bierlaire (Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Switzerland) et spécialisée pour le problème de maximum de vraisemblance, qui sera abordé en annexe (Annexe B.2).

Dans cette application, puisque les deux bases d'assurés sont simulées à partir de nos statistiques présentées au chapitre 3, nous avons décidé de retenir deux bases de données d'assurés de 50 000 lignes chacune.

5.1.1 Application à la première base : Distinction de l'effet salaire et plafond

Commençons par la première base avec uniquement les influences des variables de tranche d'âge, de sexe et de région d'habitation. On voit clairement qu'une telle base ne tient pas compte des relations liées aux attributs des contrats ou au niveau de vie des assurés. Or, cela fait un exemple parfait pour commencer la procédure de modélisation parce qu'on dispose, dans un premier temps, de ces informations de segmentation. Nous procédons aux étapes générales présentées dans la table 5.1.

Etape 1	Concevoir les segmentations par l'analyse de dépendence des variables de segmentation sur la variable choix.
Etape 2	Intégration des nouvelles variables avec les tests de modèles emboîtés.
Etape 3	Test de l'hétérogénéité inobservable et modèle de loi de mélange.
Etape 4	Choix du meilleur modèle au sens de la relation de dominance de type optimum de Pareto.

Table 5.1 : Plan de modélisation pour le modèle de choix discret

Concevoir les variables socio-démographiques

Avant d'utiliser les trois variables de segmentation de la base Damir dans la partie d'utilité systémique, il faut s'assurer du fait qu'il y a une différence significative de proportion de choix pour chaque variable. Pour cette question, on utilise le test du Chi-deux sur chaque variable parmi les trois variables suivantes : Âge x Sexe x Région, afin de tester si les proportions de choix de niveau de couverture sont homogènes pour chaque modalité de ces variables.

Soit le nombre total d'observations $N = 50000$, C la variable de choix telle que $C \in \{1, 2, 3\}$, X la variable socio-démographique de M modalités. Nous voulons tester deux hypothèses :

$\mathcal{H}_0$: Les choix sont distribués de manière homogène pour toutes les modalités de la variable X

Contre $\mathcal{H}_1$: Les choix ne sont pas distribués de manière homogène et il y a de l'hétérogénéité selon X

C ‖ X	1	...	m	...	M	Total
1	n_{11}	...	n_{1m}	...	n_{1M}	$n_{1.}$
2	n_{21}	...	n_{2m}	...	n_{1M}	$n_{2.}$
3	n_{31}	...	n_{3m}	...	n_{1M}	$n_{3.}$
Total	$n_{.1}$	...	$n_{.m}$	...	$n_{.M}$	N

TABLE 5.2 : Table de contingence issue de la base de donnée 1

On regarde la table de contingence 5.2 après avoir regroupé les variables binaires issues des trois variables socio-démographiques. Sous l'hypothèse $\mathcal{H}_0$, la statistique $\chi = \sum_{j=1}^{3}\sum_{m=1}^{M} \frac{(n_{jm}-T_{jm})^2}{T_{jm}}$, où $T_{jm} = \frac{n_{j.}n_{.m}}{N}$ suit la loi du Chi-deux de degré de liberté $(3-1)(M-1)$. On rejette l'hypothèse nulle au niveau de confiance de 5% lorsque $\chi > \alpha_{\chi^2_{(3-1)(M-1)},0.05}$, où $\alpha_{\chi^2_{(3-1)(M-1)},0.05}$ est le quantile d'ordre 95% de la loi du Chi-deux de degré de liberté $(3-1)(M-1)$.

Variable	Tranche d'âge	Sexe	Région
χ	3961.449	86.462	122.886
Degré de liberté	12	2	24
$\alpha_{\chi^2_{(3-1)(M-1)},0.05}$	21.026	5.991	36.415
Résultat	Rejette $\mathcal{H}_0$	Rejette $\mathcal{H}_0$	Rejette $\mathcal{H}_0$

TABLE 5.3 : Test d'hétérogénéité systémique des trois variables socio-démographiques sur la base 1

La table 5.3 nous donne les résultats des tests sur les trois variables : Âge (7 modalités), Sexe (2 modalités), Région (13 modalités). Comme les trois tests rejettent l'hypothèse que les proportions de choix soient homogènes pour les modalités, on peut conclure que les trois variables affectent les probabilités de choix et donc affectent l'utilité des assurés. Sachant que ces trois variables sont observables, nous pourrions les inclure dans la partie d'utilité déterministe sous forme de combinaisons linéaires des variables binaires représentant les modalités comme suivant.

Soit $i \in \{1, 2, 3\}$ et $n \in \{1, \ldots, N = 50000\}$. On note V_{ni} l'utilité déterministe que l'assuré n accorde au niveau de couverture i. On suppose que l'utilité déterministe est additive et égale à la somme des utilités composantes :
$V_{ni} = \beta_{i,const} + \beta_{i,Age\_20}\mathbb{1}_{[Age\_20_n=1]} + \beta_{i,Age\_30}\mathbb{1}_{[Age\_30_n=1]} + \ldots + \beta_{i,R5}\mathbb{1}_{[R5_n=1]} + \beta_{i,R11}\mathbb{1}_{[R11_n=1]} + \ldots + \beta_{i,Homme}\mathbb{1}_{[Homme_n=1]} + \ldots$

Comme présenté dans le chapitre précédent, on ne peut pas estimer tous les coefficients, donc il faut restreindre à zéro l'un des trois coefficients d'une même variable présente sur les trois utilités. Avec trois variables socio-démographiques, on a 22 variables binaires et donc 23 coefficients mis à zéro, en prenant en compte la constante spécifique de chaque niveau de couverture.

Afin de raccourcir l'écriture, nous notons $Age = \{Age\_20, \ldots, Age\_80\}$ l'ensemble des modalités de la tranche d'âge de l'assuré, $Sexe = \{Femme, Homme\}$ celui du sexe de l'assuré, et $Region = \{R5, \ldots, R93\}$ l'ensemble des modalités de la région d'habitation de l'assuré. Pour un assuré n, on dispose de ses informations telles que $age_n \in Age$, $sexe_n \in Sexe$, et $region_n \in Region$.

On estime donc le premier modèle logit multinomial avec cette utilité déterministe, et on appelle ce premier modèle Logit_ASR. La spécification est présentée en 5.4, le modèle logit multinomial avec trois variables distinctes : Âge x Sexe x Région. On regarde les principaux indicateurs tels que la log-vraisemblance ainsi que les deux coefficients AIC et BIC. Les résultats sont présentés dans la table C.1.

$$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n = j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n = k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n = l]}$$

TABLE 5.4 : La spécification du modèle Logit_ASR : Logit Multinômial avec variables Âge x Sexe x Région

Intégration de l'effet de plafond de remboursement et du revenu de l'assuré

Une fois que la base de segmentation est établie, on peut penser à intégrer les nouvelles variables liées aux attributs des choix ou aux caractéristiques individuelles des assurés sans que ce soit des variables définissant la nouvelle segmentation. Dans notre cas, le portefeuille comporte des informations comme le salaire par segment de la population défini en haut et la prime technique pour chaque segment. À partir de ces informations, nous cherchons à construire des modèles incorporant l'effet de revenu abordé avant ainsi que l'attribut lié au niveau de remboursement du contrat complémentaire santé.

L'effet de revenu est modélisé par la variable budget restant après la souscription du contrat complémentaire santé, notée Finance (FIN). Elle est égale à la transformation Box-Cox de la différence de niveau de vie et de la prime correspondante au niveau de couverture considéré, car il est nécessaire de présenter l'hypothèse de diminution de la sensibilité marginale par rapport au niveau de vie (budget) initial de chaque assuré :

$$\forall i \in \{1,2,3\}, FIN_{ni} = Box\_Cox(Niv\_de\_vie_n - Prime_{ni}, \lambda)$$

Nous introduisons cet effet dans les utilités déterministes en mettant les conditions sur les coefficients à estimer : l'effet revenu corrèle positivement aux utilités, $0 \leq \beta_{FIN}$, et la transformation Box-Cox est concave, $\lambda \leq 1$. Le modèle s'appelle donc Logit_ASR_S (S pour salaire/revenu/niveau de vie) et est décrit par la table 5.5.

$$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n = j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n = k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n = l]} + \beta_{FIN} FIN_{ni}$$

TABLE 5.5 : La spécification du modèle Logit_ASR_S : Logit Multinômial avec variables Âge x Sexe x Région x Finance

Pour prendre en compte les attributs du niveau de couverture, car on sait que le plafond de remboursement ou la qualité objective du contrat perçue par les assurés influence fortement la demande d'assurance. En vue de la modélisation globale du contrat, il est plausible de chercher en premier lieu à utiliser les proxies de qualité relative du contrat, par exemple les rapports du prime technique entre deux niveaux de couverture différents. En effet, le modèle de choix discret ne cherche qu'à différencier les utilités de façon relative, alors la représentation de l'ordre de qualité relative à n'importe quelle

option dans le panier est bonne pour normaliser cette quantité. Nous appelons le modèle intégrant uniquement ces variables proxies de qualité du contrat *Logit_ASR_P* (P pour plafond/limite de remboursement), et nous écrivons les utilités déterministes comme dans la table 5.6 en calculant les proxies relatifs par rapport au contrat de couverture moyen :

$$\forall i \in \{1,2,3\}, PLAF_{ni} = \frac{Prime_{ni}}{Prime_{n2}}$$

$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n=l]} + \beta_{PLAF} PLAF_{ni}$

TABLE 5.6 : La spécification du modèle Logit_ASR_P : Logit Multinômial avec variables Âge x Sexe x Région x Plafond

Afin de pouvoir comparer deux modèles avec deux effets distincts, nous estimons ainsi un troisième modèle intégrant les deux notions, Logit_ASR_PS, englobant les deux modèles Logit_ASR_P et Logit_ASR_S, avec l'utilité décrit dans la table 5.7

$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n=l]} + \beta_{PLAF} PLAF_{ni} + \beta_{FIN} FIN_{ni}$

TABLE 5.7 : La spécification du modèle Logit_ASR_PS : Logit Multinômial avec variables Âge x Sexe x Région x Plafond x Finance

Grâce à la table C.1, nous pouvons en premier lieu effectuer les tests de modèle emboîté par rapport à la vraisemblance entre les modèles estimés, et les résultats sont récapitulés dans la table C.2. On constate très clairement que dans ce cas où le choix réel n'est pas corrélé au niveau de vie ni à l'attribut du contrat, les résultats obtenus par les modèles nous confirment que la modélisation par trois variables de segments est déjà suffisante, et il n'est pas nécessaire d'incorporer ni l'effet de revenu ni l'effet de qualité du contrat. Nous retenons le modèle Logit_ASR comme modèle suffisant grâce au test de confirmation, ainsi que ses meilleurs scores des critères AIC et BIC, même si sa log-vraisemblance est moins élevée que les trois autres modèles.

L'hétérogénéité inobservée de préférence

Suite au paragraphe précédent, nous avons testé la possibilité d'ajouter de nouvelles variables au modèle en question. Il reste maintenant à évaluer l'effet de l'hétérogénéité pour déterminer si le modèle logit multinomial simple est approprié. Nous estimons un modèle de type logit de loi mélange avec les variances spécifiques d'alternative issues du modèle Logit_ASR, car il est le meilleur modèle retenu. L'idée est d'essayer de capturer s'il y a de l'hétérogénéité dans la perception des assurés causée par les alternatives. Sa spécification d'utilité est présentée dans la table 5.8, et nous l'appelons modèle ML_VS_ASR.

$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n=l]} + \sigma_i \epsilon_{ni}$

TABLE 5.8 : La spécification du modèle MLogit_VS_ASR : Modèle Logit Mixte Multinomial avec Variances Spécifiques d'Alternative, comprenant les variables Âge x Sexe x Région et un terme d'erreur gaussien.

Où $\epsilon_{i,i\in\{1,2,3\}}$ sont des variables aléatoires iid de loi normale centrée réduite et $|\sigma_{i,i\in\{1,2,3\}}|$ sont les écarts types de la loi normale supposés comme bruit aléatoire. Le défaut de ce modèle est qu'il n'est

pas possible théoriquement d'estimer les trois variances du modèle, et il est nécessaire de restreindre l'un de ces trois coefficients à zéro. C'est également la raison pour laquelle nous faisons le choix de modéliser des bruits aléatoires indépendants, où la matrice de covariance est diagonale. Nous estimons le modèle appelé MLogit_VS_ASR (pour variance spécifique d'alternative) en procédant en deux étapes suivantes en utilisant 10 simulations Monte Carlo pour l'estimation :

1. Estimer le modèle avec les trois coefficients $\sigma_i,\ i \in \{1, 2, 3\}$. Sur la base du résultat d'estimation, nous mettons à zéro le coefficient $\sigma_{i_{argmin}}$ le plus petit en terme de valeur absolue.
2. Ré-estimer le modèle avec le coefficient $\sigma_{i_{argmin}} = 0$.

Les résultats du modèle sont présentés dans la table C.1, après le test de modèle emboîté avec le modèle Logit_ASR (table C.2), nous retenons le modèle Logit simple, car il n'est pas possible de rejeter l'hypothèse $\mathcal{H}_0$ à 5%.

En résumé, la première base ne nécessite pas l'entraînement de modèles complexes avec la possibilité de capturer l'hétérogénéité inobservable. Nous nous contentons du modèle Logit_ASR comme le meilleur modèle parcimonieux, dont les coefficients sont présentés en annexe C.3.

5.1.2 Deuxième base avec l'effet réel de revenu et de la perception de qualité du contrat

Méthode de construction de la deuxième base

En raison du manque d'avis d'experts et de données sur la manière dont les deux effets, plafond et revenu, affectent les probabilités de souscription, nous avons eu recours à la réutilisation du modèle Logit_ASR_PS, entraîné sur la première base tenant compte de ces deux effets, afin de construire la deuxième base de données présentant une corrélation assez forte entre les probabilités de choix et les deux effets abordés. Cependant, comme les coefficients liés aux deux effets dans le modèle Logit_ASR_PS sont incohérents en raison de la nature de l'indépendance entre les probabilités de choix et les deux variables FIN - PLAF, nous avons modifié les coefficients $\beta_{FIN}, \beta_{PLAF}, \lambda$ de manière que les coefficients soient plus importants et que les probabilités de souscription prédites soient presque inchangées. Les probabilités de choix d'un segment de la population défini par les trois variables Âge x Sexe x Région sont estimées par le nouveau modèle, puis utilisées pour la génération de données. La figure 5.1 présente plus clairement le principe grâce auquel on parvient à déformer les probabilités de la base 2. Les coefficients utilisés pour générer la deuxième base seront présentés en annexe C.6.

Nous remarquons que le niveau de vie et la prime pour les couvertures sont des données agrégées moyennes selon les segments (Âge x Sexe x Région), donc il est normal de prendre en compte seulement ces trois variables de segment lors de la simulation de la deuxième base. Une fois que la deuxième base est simulée, nous procédons à la modélisation de la même manière que pour la première base.

Vérification de la spécification du modèle logit multinomial

Pour cette deuxième base, où nous avons effectivement incorporé les effets du revenu et de la qualité du contrat, les étapes de modélisation sont les mêmes que pour la première base. Nous commençons par examiner la base pour les hétérogénéités suivant les trois variables de segment, puis nous intégrons des variables supplémentaires, pour finir par les modèles qui capturent les hétérogénéités inobservables.

Nous effectuons les tests du chi-deux pour nous assurer qu'il y a toujours de l'hétérogénéité suivant trois variables : âge, sexe et région. La table 5.9 montre que nous pouvons modéliser, dans un premier temps, par l'utilisation de ces trois variables. Par conséquent, nous réutilisons la spécification du modèle Logit_ASR (spécification d'utilité dans la table 5.4) de la section 5.1.1.

De plus, nous construisons trois modèles pour l'intégration des effets du revenu et de la qualité du contrat, en intégrant chaque effet séparément, puis les deux effets ensemble : Logit_ASR_S (spécification

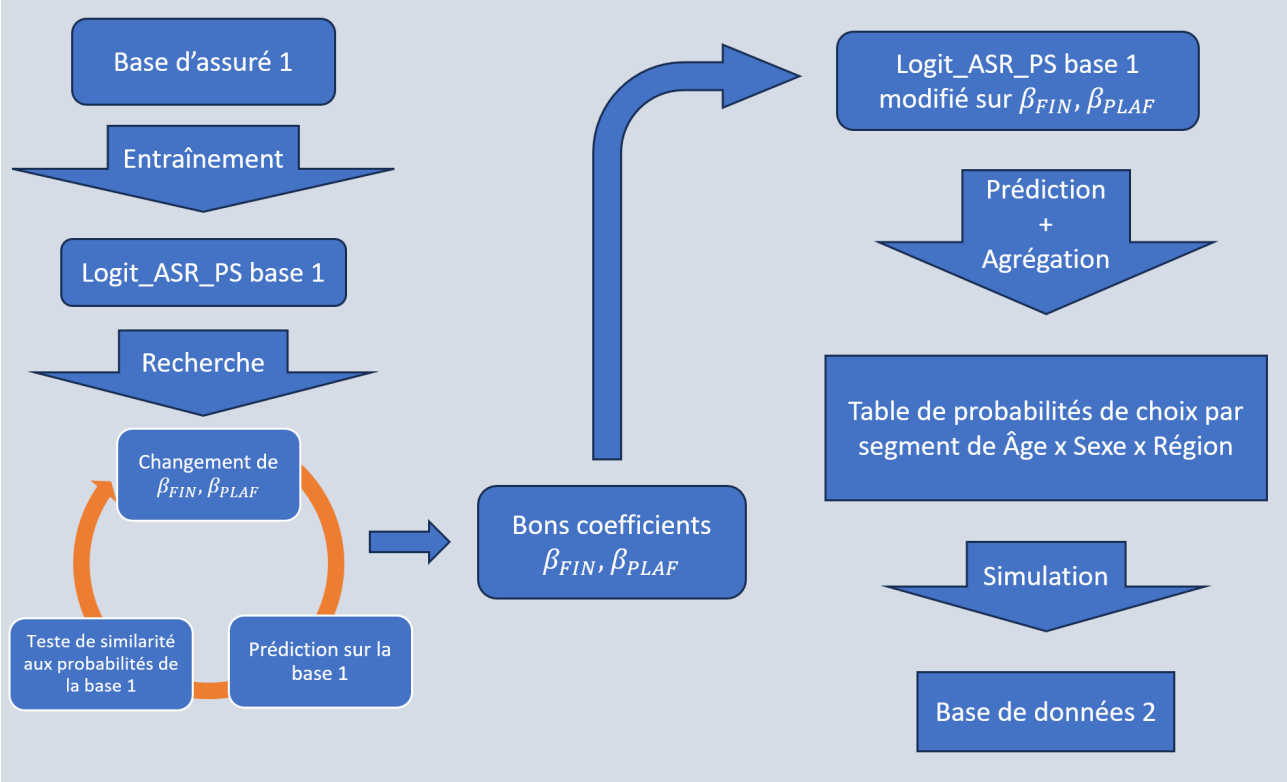


FIGURE 5.1 : Méthode de construction et de simulation da la deuxième base d'assuré (base 2)

Variable	Tranche d'âge	Sexe	Région
χ	2802.142	25.518	366.727
Degré de liberté	12	2	24
$\alpha_{\chi^2_{(3-1)(M-1)},0.05}$	21.026	5.991	36.415
Résultat	Rejette $\mathcal{H}_0$	Rejette $\mathcal{H}_0$	Rejette $\mathcal{H}_0$

TABLE 5.9 : Test d'hétérogénéité systémique des trois variables socio-démographiques sur la base 2

d'utilité dans la table 5.5), Logit_ASR_P (spécification d'utilité dans la table 5.6) et Logit_ASR_PS (spécification d'utilité dans la table 5.7). Les évaluations des quatre modèles sont présentées dans la table C.4, et d'après les résultats des tests de modèles emboîtés de la table C.5, nous retenons le modèle Logit_ASR_PS comme le meilleur modèle en raison de ses trois meilleurs scores d'AIC, BIC et de log vraisemblance. Les coefficients du modèle sont présentés en annexe C.6. On peut donc conclure que l'effet de revenu est clairement présent dans la base 2, tandis que la perception de la qualité du contrat est beaucoup moins claire dans les données.

Modèle de type mélange de loi continue

Comme le modèle Logit_ASR_PS est le mieux adapté et que nous avons inclus toutes les variables possibles de manière linéaire dans la formule d'utilité, il reste à essayer de capturer l'hétérogénéité non observée au sein de la population, s'il existe. Nous poursuivons avec deux modèles logit à loi mélangée, modifiés à partir du modèle Logit_ASR_PS. Le premier modèle suppose un effet aléatoire

spécifique à chaque alternative. Cette spécification est similaire au modèle MLogit_VS_ASR de la section précédente, avec les ϵ_i iid de même loi normale centrée réduite. Nous précisons que le modèle MLogit_VS_ASR_PS est estimé deux fois en supprimant l'écart type le plus petit, et dont l'utilité est décrite comme dans la table 5.10.

$$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j\in Age} \beta_{i,j}\mathbb{1}_{[age_n=j]} + \sum_{k\in Sexe} \beta_{i,k}\mathbb{1}_{[sexe_n=k]} + \sum_{l\in Region} \beta_{i,l}\mathbb{1}_{[region_n=l]} + \beta_{PLAF}PLAF_{ni} + \beta_{FIN}FIN_{ni} + \sigma_i\epsilon_i$$

TABLE 5.10 : La spécification du modèle MLogit_VS_ASR_PS : Modèle Logit Mixte Multinomial avec Variances Spécifiques d'Alternative, comprenant les variables Âge x Sexe x Région x Plafond x Finance et un terme d'erreur gaussien.

Ce modèle suppose donc que l'hétérogénéité est liée aux aléas créés par la perception de l'assuré sur les contrats de l'assurance, par exemple s'il a un doute de façon générale sur un niveau de couverture par rapport aux autres. Il se peut que les aléas proviennent seulement d'évaluations individuellement différentes par rapport à une ou des caractéristiques du contrat. On peut utiliser dans ce cas le modèle Logit à coefficient aléatoire afin de supposer l'aléatoire de certains coefficients de variable dans la fonction de l'utilité.

Notre deuxième modèle suppose que le coefficient de la variable Finance est aléatoire, présentant ainsi l'hétérogénéité inobservable des assurés pour l'effet revenu. Ce modèle est donc un modèle à coefficients aléatoires, inspiré par la spécification des modèles de choix discrets pour différencier les produits, comme décrit dans BIRCHALL et VERBOVEN (2022). Par hypothèse, les assurés devraient systématiquement être mieux moralement lorsque leur somme restante augmente, ce qui indique que le coefficient doit être positif. Par conséquent, il est plausible de supposer que la distribution de ce coefficient suit une loi lognormale ou d'autres lois à support positif. Il n'est pas approprié de choisir la loi normale car la densité de la loi normale (figure 5.2) prend en charge $\mathbb{R}$ et peut donc être négative. Pour la suite, on utilise à titre d'exemple la loi lognormale pour sa simplicité d'implémentation et de simulation.

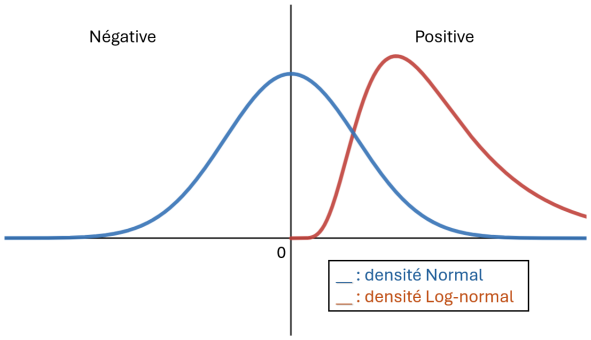


FIGURE 5.2 : Densités de la loi normale et log-normale.

Nous notons ce modèle MLogit_RC_ASR_PC (pour Random Coefficient Logit) et écrivons sa partie utilité V_{ij} dans la table 5.11.

Nous avons séparé la moyenne et l'écart type dans l'expression ci-dessus, car $e^{\mu+\sigma\epsilon} \sim \mathcal{LN}(\mu, \sigma^2)$ lorsque ϵ suit la loi normale centrée réduite.

Les estimations de ces deux modèles sont fournies dans la table C.9 pour 10 simulations de Monte-Carlo. Nous nous concentrons sur les tests de modèle emboîté de la table C.9, qui indiquent que le modèle Logit_ASR_PS ne peut pas être rejeté. Cependant, ce résultat n'est pas très informatif sur l'hypothèse IANP, car il conduit seulement à la conclusion que les deux modèles logit à loi mélange continue estimés ne conviennent pas au jeu de données. D'autres modèles de l'espace de modèle à loi

$$\forall i \in \{1,2,3\}, \quad V_{ni} = \beta_{i,const} + \sum_{j \in Age} \beta_{i,j} \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \beta_{i,k} \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \beta_{i,l} \mathbb{1}_{[region_n=l]} + \beta_{PLAF} PLAF_{ni} + e^{\mu+\sigma\epsilon} FIN_{ni}$$

TABLE 5.11 : La spécification du modèle MLogit_RC_ASR_PC : Modèle Logit Mixte Multinomial avec Coefficient Aléatoire de la variable Finance, comprenant les variables Âge x Sexe x Région x Plafond x Finance et un terme d'erreur gaussien.

mélange plus complexes pourraient être mieux adaptés que Logit_ASR_PS. Selon le résultat important de McFADDEN et TRAIN (2000a), les modèles à loi mélange continue peuvent bien approximer n'importe quel système d'utilité aléatoire sous l'hypothèse d'une bonne spécification. Cependant, en raison du manque de capacités de simulation et de données disponibles, nous nous contentons du modèle Logit_ASR_PS comme le meilleur modèle retenu.

En outre, les deux modèles entraînés avec seulement 10 simulations de type pseudo-aléatoire ne sont clairement pas suffisants pour approcher les vrais paramètres du composant aléatoire. Il n'existe pas un nombre de simulations universel pour la convergence des modèles, mais dans notre cas, avec 3 alternatives comportant au plus deux paramètres du composant aléatoire, il convient de prendre au moins 25 simulations de Halton (voir HENSHER et GREENE (2003)) pour que le modèle soit stabilisé. Nous notons que les simulations de Halton sont des simulations intelligentes, c'est-à-dire qu'elles réduisent considérablement la variance par rapport aux méthodes de simulation pseudo-aléatoire. Par conséquent, il peut être plus prudent de prendre au moins 100 simulations pseudo-aléatoires dans ce cas. Cependant, la vitesse de l'optimisation de la vraisemblance est fortement affectée par le nombre de simulations, et 100 simulations semblent impossibles compte tenu de notre capacité informatique. Par conséquent, nous avons choisi d'entraîner ces modèles de mélange continu pour illustrer la théorie, mais les résultats non robustes ne peuvent pas être pris en compte. Nous préférons donc utiliser le modèle sûr Logit_ASR_PS.

Modèle de type mélange de loi discret

Après avoir exploré les deux modèles à loi de mélange continue, nous examinons des modèles de type classe latente dérivés du modèle Logit_ASR_PS pour déterminer s'il existe un effet d'hétérogénéité inter-classes d'individus. Nous supposons qu'il existe deux classes $m \in \{1,2\}$ d'individus dont tous les coefficients dans l'utilité diffèrent entre les deux classes. Ce modèle reflète le fait que deux assurés de même classe sociale peuvent avoir deux perceptions différentes du risque monétaire (moins sensible et très sensible à la perte monétaire), liées au coefficient de FINANCE, λ, et deux niveaux d'anticipation de dépenses (apprécier les bons remboursements ou être indifférent), liés au coefficient de PLAFOND, en raison des informations endogènes sur les préférences non observables. Nous soulignons que le modèle devrait avoir au moins une classe latente avec les coefficients β_{PLAF}, β_{FIN} positifs et $\lambda < 1$ pour être cohérent avec l'hypothèse de l'existence de l'effet revenu, de la diminution de l'utilité marginale du revenu et de l'anticipation de soins. Néanmoins, le fait qu'il existe un coefficient négatif ou que $\lambda \geq 1$ ne signifie pas que le modèle est faux, mais indique seulement que nous capturons certains comportements cachés par le fait de modéliser le comportement en général sur toutes les classes latentes.

La spécification du modèle de classe latente nécessite des probabilités d'appartenance des classes d'un individu. Nous supposons dans cette application que les probabilités d'appartenance à la classe 1 ou 2 peuvent être évaluées par une régression logistique ou probit binaire sur trois variables socio-démographiques : la tranche d'âge, le sexe et la région. Le modèle avec les probabilités de classe logistique, appelé LCLogit_ASR_PS_L (pour modèle logit de classe latente à probabilité de membre logit), exprime les probabilités d'appartenance de l'individu n à l'une ou l'autre classe par le système suivant :

$$\forall i \in \{1,\dots,N\}, \omega_n = \gamma_{const} + \sum_{j \in Age} \gamma_j \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \gamma_k \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \gamma_l \mathbb{1}_{[region_n=l]}$$

D'où $P(n \in Classe_1) = \frac{1}{1+e^{\omega_n}}$ et $P(n \in Classe_2) = 1 - P(n \in Classe_1) = 1 - \frac{1}{1+e^{\omega_n}}$.

Si l'on change le sous-modèle de probabilité du logit au probit, nous obtenons une autre spécification appelée LCLogit_ASR_PSV_P (pour modèle logit de classe latente à probabilité de membre probit). Les probabilités d'appartenance à chaque classe sont définies comme suit, où $\Phi()$ représente la fonction de répartition de la loi normale centrée réduite. De même, nous obtenons le système suivant :

$$\forall i \in \{1,\dots,N\}, \omega_n = \gamma_{const} + \sum_{j \in Age} \gamma_j \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \gamma_k \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \gamma_l \mathbb{1}_{[region_n=l]}.$$

D'où $P(n \in Classe_1) = \Phi(\omega_n)$ et $P(n \in Classe_2) = 1 - P(n \in Classe_1) = 1 - \Phi(\omega_n)$.

Les utilités déterministes des deux classes sont communes pour les deux modèles, comme indiqué dans le tableau 5.12. Les évaluations des modèles sont présentées dans le tableau C.8, et les coefficients sont disponibles dans le tableau C.11.

$\forall i \in \{1,2,3\}, \forall m \in \{1,2\}, FIN_{ni,Classe_m} = Box\_Cox(Niv\_de\_vie_n - Prime_{ni}, \lambda_{Classe_m})$
$\forall i \in \{1,2,3\}, V_{ni,Classe_1} = \beta_{i,const,Classe_1} + \sum_{j \in Age} \beta_{i,j,Classe_1} \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \beta_{i,k,Classe_1} \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \beta_{i,l,Classe_1} \mathbb{1}_{[region_n=l]} + \beta_{PLAF,Classe_1} PLAF_{ni} + \beta_{FIN,Classe_1} FIN_{ni,Classe_1}$
$\forall i \in \{1,2,3\}, V_{ni,Classe_2} = \beta_{i,const,Classe_2} + \sum_{j \in Age} \beta_{i,j,Classe_2} \mathbb{1}_{[age_n=j]} + \sum_{k \in Sexe} \beta_{i,k,Classe_2} \mathbb{1}_{[sexe_n=k]} + \sum_{l \in Region} \beta_{i,l,Classe_2} \mathbb{1}_{[region_n=l]} + \beta_{PLAF,Classe_2} PLAF_{ni} + \beta_{FIN,Classe_2} FIN_{ni,Classe_2}$

TABLE 5.12 : La spécification d'utilité de la première et de la deuxième classe des modèles LCLogit_ASR_PS_L et LCLogit_ASR_PS_P, avec tous les coefficients variant entre les deux classes

Le tableau C.9 présente les résultats des tests de modèle emboîté par rapport à Logit_ASR_PS et Logit_ASR. Cette comparaison est réalisable car le modèle Logit_ASR_PS est un cas spécial de ces deux modèles avec les contraintes linéaires imposées, où les coefficients des deux classes latentes sont exactement les mêmes. Ces tests concluent qu'il n'existe pas d'hétérogénéité latente, car il n'est pas possible de rejeter le modèle Logit_ASR_PS avec un niveau de confiance de 5%. Sinon, il devrait être possible de rejeter le modèle simple lorsqu'il y a des hétérogénéités interclasses. Le fait que le modèle LCLogit_ASR_PS_P puisse rejeter le modèle Logit_ASR peut simplement signifier que le modèle Logit_ASR manque de l'effet de revenu et de la qualité de l'alternative.

Nous pourrons vérifier notre hypothèse de deux classes latentes qui diffèrent en termes de préférence pour le risque et d'anticipation des dépenses en analysant les deux modèles LCLogit_ASR_PS_L et LCLogit_ASR_PS_P. Nous observons d'abord le modèle LCLogit_ASR_PS_L, où :

$$\beta_{FIN\_1} = 0.624; \beta_{PLAF\_1} = 0.936; \alpha_1 = 1.42.$$

$$\beta_{FIN\_2} = 0.334; \beta_{PLAF\_2} = 1.01; \alpha_2 = 0.783.$$

Comme les coefficients estimés affectent l'utilité relativement, il ne faut pas les comparer directement, mais les comparer normalisés par rapport à un coefficient. Nous choisissons les coefficients β_{PLAF_1} et β_{PLAF_2} comme dénominateurs car ils sont les coefficients d'un quotient de deux grandeurs de même unité et n'ont donc pas d'unité. En divisant tous les coefficients présentés dans l'utilité déterministe par ces deux dénominateurs, sauf pour α_1 et α_2 (comme il s'agit d'un paramètre exact de la transformation Box-Cox), nous obtenons :

$$\bar{\beta}_{FIN\_1} = \frac{0.624}{0.936} \approx 0.6667; \bar{\beta}_{FIN\_2} = \frac{0.334}{1.01} \approx 0.3306; \bar{\beta}_{PLAF\_1} = \bar{\beta}_{PLAF\_2} = 1.$$

En constatant que $\bar{\beta}_{FIN\_2} < \bar{\beta}_{FIN\_1}$ pour le même niveau de $\bar{\beta}_{PLAF\_2} = \bar{\beta}_{PLAF\_1} = 1$, nous pouvons conclure que la classe 1 accorde plus de poids à ses contraintes budgétaires et que sa sensibilité au changement de prix est différente de celle de la classe 2. Cela s'explique par la convexité de la transformation de Box-Cox avec $\alpha_1 = 1.42$, indiquant une augmentation de la sensibilité marginale de la prime d'assurance lorsque le revenu augmente. Ce modèle distingue deux classes d'assurés différentes dans leur perspective : la classe 1 avec une plus grande aversion au petit risque monétaire et la classe 2 qui accorde plus d'importance à la qualité du contrat. Cette conclusion semble plausible par rapport à notre hypothèse du modèle.

Pour le modèle LCLogit_ASR_PS_P, les résultats de l'estimation sont les suivants :

$$\beta_{FIN\_1} = 0.123; \beta_{PLAF\_1} = 2.03; \alpha_1 = 0.872.$$

$$\beta_{FIN\_2} = 0.542; \beta_{PLAF\_2} = 0.967; \alpha_2 = 0.694.$$

Après avoir normalisé les coefficients en utilisant les coefficients de la variable Plafond, de la même manière que pour le modèle LCLogit_ASR_PS_L, nous obtenons :

$$\bar{\beta}_{FIN\_1} \approx 0.0606; \bar{\beta}_{FIN\_2} \approx 0.5605; \bar{\beta}_{PLAF\_1} = \bar{\beta}_{PLAF\_2} = 1.$$

On constate que la perception des deux classes diffère très peu en ce qui concerne les effets ; peu importe la classe latente, les individus auront une aversion à la perte monétaire et une diminution de l'utilité marginale avec l'augmentation du revenu. En comparant les deux modèles, il semble que le modèle LCLogit_ASR_PS_P soit préférable, car il obtient de meilleures performances sur les cinq scores de performance, y compris la log-vraisemblance, l'AIC, le BIC, ρ^2, et $\bar{\rho}^2$. De plus, il permet de conclure qu'il n'y a pas d'hétérogénéité remarquable sur la perception du risque dans les données générées, ce qui était bien notre cas. En somme, une explication simplifiée pourrait être la suivante : le modèle de classe latente basé sur la régression probit se comporte mieux que le modèle basé sur la régression logistique, car il permet une meilleure capture des classes latentes et présente des avantages pour relâcher l'hypothèse IANP en autorisant une corrélation entre les deux classes latentes.

Modèle Logit de l'échelle hétérogène

En arrivant à la fin des méthodes couramment utilisées, nous allons à présent explorer les modèles qui capturent l'hétérogénéité d'échelle. L'idée sous-jacente à ce modèle est que certaines variables influencent plus profondément la structure des données que d'autres, et qu'elles interagissent avec les coefficients des autres variables de manière à modifier de manière homogène l'échelle de l'utilité déterministe pour toutes les alternatives. Plus précisément, nous supposons qu'il existe une variable d'échelle ω fonction des variables observées, de telle sorte que les nouvelles utilités déterministes peuvent être calculées comme suit, où V_{ni} est l'utilité déterministe spécifiée normalement pour le modèle Logit Multinomial Logit_ASR_PS.

$$\forall i \in \{1, 2, 3\}, \quad V'_{ni} = e^{\omega_n} V_{ni}.$$

Par la suite, nous modélisons la variable ω selon la spécification en fonction des variables de segmentation telles que l'âge, le sexe et la région, tandis que les variables contenant des attributs spécifiques des trois types de contrat ne sont pas utilisées. En effet, nous supposons que seuls ces trois variables catégorielles affecte la variance des utilités inobservables de façon significative pour prendre en compte. Nous appelons le modèle HLogit_ASR_PS_ASR et présente sa formule d'utilité par la table 5.13.

$\forall n \in \{1,\ldots,N\}, \omega_n = \sum_{j\in Age}\gamma_j \mathbb{1}_{[age_n=j]} + \sum_{k\in Sexe}\gamma_k \mathbb{1}_{[sexe_n=k]} + \sum_{l\in Region}\gamma_l \mathbb{1}_{[region_n=l]}$
$\forall i \in \{1,2,3\}, \quad V'_{ni} = e^{\omega}(\beta_{i,const} + \sum_{j\in Age}\beta_{i,j}\mathbb{1}_{[age_n=j]} + \sum_{k\in Sexe}\beta_{i,k}\mathbb{1}_{[sexe_n=k]} + \sum_{l\in Region}\beta_{i,l}\mathbb{1}_{[region_n=l]} + \beta_{PLAF}PLAF_{ni} + \beta_{FIN}FIN_{ni} = e^{\omega_n}V_{ni}$

TABLE 5.13 : La spécification du modèle Logit Hétérogène HLogit_ASR_PS_ASR avec une variable d'échelle dépend des variables Tranche d'âge, Région et Sexe.

Son évaluation après l'estimation est présentée dans la table C.8, et son acceptation par le test de modèle emboîté (contre le modèle Logit_ASR_PS) est présentée dans la même table. Nous remarquons que le modèle Logit_ASR_PS reste préféré sans avoir besoin de spécifier des hétérogénéités supplémentaires. Même en incluant davantage de facteurs explicatifs dans la variable d'échelle, le modèle n'améliore pas significativement les coefficients ρ^2 et $\bar{\rho}^2$. Ceci est renforcé par le fait que le modèle HLogit_ASR_PS_ASR n'est pas capable de rejeter le modèle simple avec un niveau de confiance de 5%. De plus, les coefficients liés au sexe dans ω estimés du modèle HLogit_ASR_PS_ASR, présentés en annexe C.12, sont assez proches (0.851 pour les hommes et 0.85 pour les femmes) qu'ils nous conduisent à conclure que pour la variable sexe, il n'y a pas suffisamment d'impact sur l'échelle des hétérogénéités inobservables pour qu'elle puisse être utilisée comme prédicteur pour la variable ω. Nous ne retenons donc pas ce modèle pour la suite en raison de son incapacité à rejeter le modèle Logit_ASR_PS (table C.9).

5.1.3 Récapitulatif des modèles appliqués et pistes de modélisation

Tout d'abord, pour la première base où nous n'avons pas intégré l'effet du revenu et la perception de la qualité du contrat aux données, le fait que les facteurs observables spécifiques à l'âge, au sexe et à la région expliquent parfaitement les relations des données sans avoir besoin d'explications supplémentaires par des facteurs aléatoires montre que le modèle Logit_ASR est le meilleur modèle. Cependant, cette base de données ne nous permet pas de modéliser l'effet du revenu et du tarif du contrat, qui sont des effets importants pour la demande d'assurance, comme montré par GRIGNON, KAMBIA-CHOPIN et al. (2009) et CAUSSAT et GLAUDE (1993a).

Sans les statistiques liant l'effet revenu et la qualité du contrat aux probabilités de choix utilisées pour la simulation de la base d'assurés, nous avons choisi de modifier le modèle Logit_ASR_PS entraîné sur la base 1 afin d'obtenir les probabilités de choix utilisées pour la simulation de la deuxième base d'assurés présentée au chapitre 3. Grâce à la méthode de simulation du deuxième scénario, les choses sont plus intéressantes car les données sous-jacentes sont réellement corrélées avec les variables FINANCE et PLAFOND. Nous sommes ainsi capables de rejeter le modèle simple Logit_ASR (entraîné sur la base 2) proprement pour adopter le modèle Logit_ASR_PS (entraîné sur la base 2), soulignant la présence d'un effet de revenu et de caractéristiques des contrats.

Afin de tester la validité de l'hypothèse IANP dans la deuxième base d'assurés, nous avons donc testé les différents modèles englobant Logit_ASR_PS présentant l'hétérogénéité non observable, s'ils peuvent rejeter le modèle simple Logit_ASR_PS. Nous avons essayé les variantes suivantes en accompagnant les hypothèses pour l'hétérogénéité de préférence :

- Modèle de type variance spécifique avec l'hypothèse que les trois niveaux de couverture pourraient créer une sorte de corrélation avec les informations personnelles de la personne, présente donc par les quantités aléatoires ajoutées à chaque utilité déterministe.
- Modèle de type coefficient aléatoire où notre assuré est supposé associer un poids individuel à l'effet de revenu, d'où le coefficient β_{FIN} est considéré aléatoire mais positif.
- La littérature sur l'anti-sélection a effectivement inspiré l'application de nos modèles de classe latente sur l'hypothèse qu'il existe deux classes d'assurés avec deux attitudes différentes par

rapport à la préférence pour l'assurance, dont certains sont vraiment adverses au risque monétaire et d'autres moins. Les modèles sont étendus pour voir si les profils de risque (approximés par les variables socio-économiques) affectent ainsi les préférences de l'assurance.

- Nous supposons que l'hétérogénéité des préférences vient de la différence de la variance des utilités inobservables et testons les modèles de type logit multinomial hétérogène.

Grâce aux résultats des tests de modèle emboîtés, montrant un manque de capacité à rejeter le modèle simple, nous nous sommes permis de conclure que les données n'ont pratiquement pas d'hétérogénéité en préférences inobservables, l'hypothèse d'IANP est préservée. Ce résultat paraît plausible dans le cadre de cette modélisation, car les données à notre disposition sont assez simples et génériques (celles qui sont générées sans l'hétérogénéité inobservable).

Au final, le modèle à retenir est Logit_ASR_PS avec un nombre de coefficients parcimonieux et des scores AIC, BIC acceptables. Son score de ρ^2 et $\bar{\rho}^2$ est moins bon que les autres modèles, du fait que les modèles plus complexes s'adaptent davantage aux données, mais risquent de surajuster le vrai comportement des assurés. Ce modèle que nous décidons de garder serait utilisé pour prédire les probabilités de choix agrégées, c'est-à-dire la structure du portefeuille après avoir introduit des changements au prime d'assurance.

Il est important de noter ici que, contrairement à l'approche de l'apprentissage automatique, où les modèles peuvent être soumis à une procédure de sélection des variables pour choisir les variables explicatives les plus informatives selon les critères d'information, et où les interprétations reposent sur le modèle le plus optimal post-estimation, la modélisation des choix discrets repose sur l'incrémentation expérimentale des variables issues de la connaissance du modélisateur. On ne peut pas inclure n'importe quelle variable dans n'importe quelle forme dans l'utilité déterministe, car l'interprétation du modèle repose sur les hypothèses selon lesquelles les variables sont incorporées. Après avoir choisi le meilleur modèle pour chaque scénario, il faut effectuer la procédure de post-validation du modèle, souvent par comparaison de la prédiction du modèle avec les proportions réelles de données, ou mieux par les procédures dites de calibration du modèle.

5.1.4 Analyse de la qualité de calibration du modèle

Dans notre application, nous nous concentrons sur la capture des comportements des assurés présents dans le portefeuille, et donc sur la calibration du modèle au maximum des données, sans nous pencher sur la généralisation du modèle aux nouveaux profils d'assurés. En effet, nous cherchons à évaluer les préférences de ceux qui sont déjà présents, et non des nouveaux entrants. Par conséquent, l'étape d'évaluation de la capacité de généralisation du modèle n'est pas nécessaire.

Nous vérifions dans deux scénarios la capacité réelle du modèle à approximer les données. Pour ce faire, nous cherchons à comparer les proportions de choix réels et les probabilités de choix agrégées selon les tranches de segment de chaque variable de tranche d'âge, de sexe et de région. En effet, l'évaluation de la base d'assurés par le modèle choisi donne les probabilités de choix de couverture pour chaque individu dans la base. Il est donc nécessaire de calculer la moyenne des probabilités de choix de niveau de couverture pour un segment a de la base d'assurés comme suit :

$$\forall i \in \{1,2,3\}, P'(C^a = i) = \frac{1}{n_a} \sum_{m \in seg(a)} P_{\text{prédit}}(C_m = i),$$

où :

- $\{1,2,3\}$ correspond à {couverture minimale, couverture moyenne, couverture maximale}.
- n_a : nombre d'assurés du segment a,

- C_m : choix de niveau de couverture de l'individu m,
- C^a : choix de niveau de couverture agrégé des assurés du segment a.

Les proportions de différents contrats dans un segment a sont calculées par :

$$\forall i \in \{1,2,3\}, P(C^a = i) = \frac{1}{n_a} \sum_{m \in \text{seg}(a)} \mathbb{1}_{[C_m = i]}.$$

La table 5.14 nous indique que le modèle Logit_ASR est très adapté à la base 1, avec peu de différence pour n'importe quel segment. Pour la base 2, le modèle choisi de la section précédente est Logit_ASR_PS, et ce modèle est très adapté aux données de la base 2 (table 5.15). On pourrait sans doute considérer que les probabilités agrégées des modèles choisis sont en réalité les vraies proportions des données de la base 1 et 2.

Segment	$P(C^a=1)$	$P(C^a=2)$	$P(C^a=3)$	$P'(C^a=1)$	$P'(C^a=2)$	$P'(C^a=3)$
Sexe = Homme	0.2901	0.5545	0.1554	0.2901	0.5545	0.1554
Sexe = Femme	0.2560	0.5690	0.1749	0.2560	0.5690	0.1749
Tranche d'âge = 20	0.4089	0.5033	0.0878	0.4089	0.5033	0.0878
Tranche d'âge = 30	0.3639	0.5244	0.1117	0.3639	0.5244	0.1117
Tranche d'âge = 40	0.3255	0.5573	0.1173	0.3255	0.5573	0.1173
Tranche d'âge = 50	0.2580	0.6064	0.1356	0.2580	0.6064	0.1356
Tranche d'âge = 60	0.1859	0.6250	0.1892	0.1859	0.6250	0.1892
Tranche d'âge = 70	0.1441	0.5675	0.2884	0.1441	0.5675	0.2884
Tranche d'âge = 80	0.1124	0.5318	0.3558	0.1124	0.5318	0.3558
Code région = 5	0.3062	0.5438	0.1500	0.3062	0.5438	0.1500
Code région = 11	0.2828	0.5649	0.1522	0.2829	0.5649	0.1522
Code région = 24	0.2760	0.5626	0.1614	0.2760	0.5626	0.1614
Code région = 27	0.2560	0.5808	0.1631	0.2560	0.5808	0.1631
Code région = 28	0.2646	0.5726	0.1628	0.2646	0.5726	0.1628
Code région = 32	0.2984	0.5532	0.1484	0.2984	0.5532	0.1484
Code région = 44	0.2778	0.5634	0.1588	0.2778	0.5634	0.1588
Code région = 52	0.2822	0.5716	0.1461	0.2822	0.5716	0.1461
Code région = 53	0.2824	0.5569	0.1607	0.2824	0.5569	0.1607
Code région = 75	0.2661	0.5564	0.1775	0.2661	0.5564	0.1775
Code région = 76	0.2496	0.5652	0.1852	0.2496	0.5652	0.1852
Code région = 84	0.2567	0.5679	0.1754	0.2567	0.5679	0.1754
Code région = 93	0.2557	0.5471	0.1972	0.2557	0.5471	0.1972
Globale	0.2722	0.5621	0.1657	0.2722	0.5621	0.1657

TABLE 5.14 : Validation du modèle Logit_ASR sur la base 1 par la comparaison des probabilités de prédiction et des proportions réelles des contrats dans le portefeuille

5.2 Mesure de la présence d'anti-sélection avant et après changement tarifaire.

5.2.1 Mesure de l'anti-sélection sur le portefeuille d'assurance

Cette section consiste à appliquer les formules de l'anti-sélection au portefeuille de l'assureur une fois que les modèles de choix discrets sont bien estimés et calibrés sur les bases stratifiées du por-

Segment	$P(C^a = 1)$	$P(C^a = 2)$	$P(C^a = 3)$	$P'(C^a = 1)$	$P'(C^a = 2)$	$P'(C^a = 3)$
Sexe = Homme	0.3196	0.5378	0.1427	0.3196	0.5378	0.1427
Sexe = Femme	0.3163	0.5249	0.1587	0.3163	0.5249	0.1587
Tranche d'âge = 20	0.4233	0.5046	0.0722	0.4233	0.5046	0.0722
Tranche d'âge = 30	0.3787	0.5157	0.1055	0.3787	0.5157	0.1055
Tranche d'âge = 40	0.3531	0.5258	0.1211	0.3531	0.5258	0.1211
Tranche d'âge = 50	0.3237	0.5585	0.1178	0.3237	0.5585	0.1178
Tranche d'âge = 60	0.2622	0.5713	0.1665	0.2622	0.5713	0.1665
Tranche d'âge = 70	0.2171	0.5166	0.2663	0.2171	0.5166	0.2663
Tranche d'âge = 80	0.1728	0.5054	0.3219	0.1728	0.5054	0.3219
Code région = 5	0.3205	0.5445	0.1349	0.3205	0.5445	0.1349
Code région = 11	0.3674	0.4915	0.1411	0.3674	0.4915	0.1412
Code région = 24	0.3301	0.5419	0.1279	0.3301	0.5419	0.1279
Code région = 27	0.2663	0.5754	0.1582	0.2663	0.5754	0.1582
Code région = 28	0.2826	0.5714	0.1460	0.2826	0.5714	0.1460
Code région = 32	0.3515	0.5144	0.1341	0.3515	0.5144	0.1341
Code région = 44	0.3077	0.5374	0.1549	0.3077	0.5374	0.1549
Code région = 52	0.3138	0.5713	0.1150	0.3137	0.5713	0.1150
Code région = 53	0.3078	0.5415	0.1507	0.3078	0.5415	0.1507
Code région = 75	0.2732	0.5707	0.1560	0.2732	0.5708	0.1560
Code région = 76	0.2898	0.5432	0.1670	0.2898	0.5432	0.1670
Code région = 84	0.3290	0.5179	0.1531	0.3290	0.5179	0.1531
Code région = 93	0.2973	0.5025	0.2003	0.2973	0.5025	0.2003
Globale	0.3179	0.5310	0.1511	0.3179	0.5310	0.1511

TABLE 5.15 : Validation du modèle Logit_ASR_PS sur la base 2 par la comparaison des probabilités de prédiction et des proportions réelles des contrats dans le portefeuille

tefeuille. Le graphe 5.3 synthétise la méthode explicite pour l'application, où la sortie du modèle peut être utilisée directement pour l'évaluation des coefficients d'anti-sélection. Dans le contexte de l'exemple utilisé dans ce mémoire, notre base assurée stratifiée est la base d'assurés simulée selon les deux scénarios, et notre base de sinistres est celle de la base Open Damir accordée à la population française en 2021 traité au chapitre 2. Comme notre portefeuille d'assurance ne dispose pas de l'information liée à la sinistralité individuelle au niveau d'agrégation de la base Open Damir, les individus sont estimés avoir une sinistralité égale à la sinistralité moyenne ou la prime technique par segment auquel ils appartiennent. Nous ne sommes donc pas capables d'utiliser le coefficient d'anti-sélection relatif introduit précédemment, et nous devrons faire appel au coefficient d'anti-sélection global du portefeuille. Dans le but d'illustrer comment calculer ce coefficient, nous notons les quantités suivantes :

- N, le nombre total d'assurés du portefeuille : c'est la population française totale en 2021 segmentée en plusieurs segments correspondant au maillage de la base Open Damir et des modèles entraînés.
- N_a, le nombre total d'assurés du segment a du portefeuille.
- $Sin\_moy\_seg_j^a$, la prime technique du niveau de couverture j que l'assureur doit faire payer à un assuré du segment a si tous les assurés du segment a ont effectivement choisi la couverture j.
- $Sin\_total\_seg_j^a = Sin\_moy\_seg_j^a \times N_a$, le sinistre total du segment a si tout le monde du segment a a choisi le niveau de couverture j.

- $Sin\_portef\_total_j$, le sinistre total de l'ensemble du portefeuille si tous les assurés ont choisi le niveau de couverture j.
- $\theta_a = 1$ p.s $\forall a \in \{1, .., A\}$, comme hypothèse de simplification du calcul, on a donc l'exposition totale égale au nombre de contrats existants.
- $P'(C^a = j)$, la probabilité agrégée des probabilités que chaque assuré de segment a choisisse le niveau de couverture j, prédite par le modèle Logit_ASR sur la base d'assurés 1 et Logit_ASR_PS sur la base d'assurés 2.

Nous remarquons que les sinistres ici peuvent être considérés comme des sinistres par poste de garantie du contrat ou sur l'ensemble des postes de garantie, car le sinistre sur l'ensemble de poste est simplement la somme des sinistres de tous les postes. Le coefficient d'anti-sélection du niveau de couverture $j \in \{1, 2, 3\}$ pour un poste de garantie arbitraire s'écrit donc :

$$Coef\_global_j = \frac{Sin\_moy_j}{Sin\_portef\_moy_j} = \frac{\frac{\sum_{a=1}^{A} Sin\_moy\_seg_j^a \times N_a \times P'(C^a=j)}{\sum_{a=1}^{A} N_a \times P'(C^a=j)}}{\frac{Sin\_portef\_total_j}{N}}.$$

Dans un premier temps, il peut être intéressant de regarder comment l'anti-sélection se comporte sur certains postes jugés principaux causes de l'anti-sélection et certains postes considérés comme beaucoup moins anticipés par la littérature.

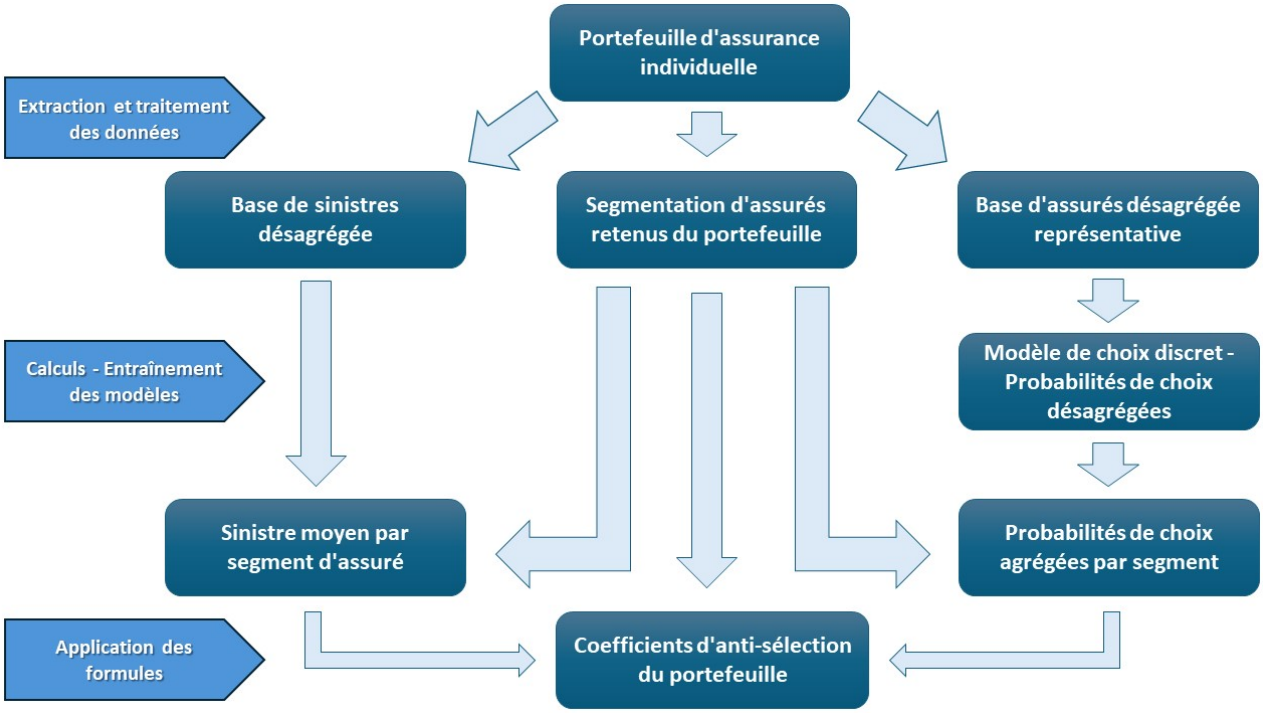


FIGURE 5.3 : Stratégie de la modélisation de l'anti-sélection grâce au modèle de choix discrets

Contraste de l'anti-sélection sur différentes postes de garantie

La littérature sur l'assurance santé en France a cité que des postes de type dentaire, optique ou auditif peuvent conduire à des choix stratégiques, tandis que des postes comme les consultations chez un médecin généraliste ne le font pas. Afin d'illustrer cette hypothèse, nous calculons les coefficients d'anti-sélection sur deux sous-familles de garanties, à savoir :

- Consultation chez un médecin généraliste : On suppose que ce poste de garantie ne présente pas beaucoup d'anti-sélection, et nous cherchons à vérifier si le vrai sinistre moyen ne peut pas dépasser de plus de 10% le sinistre moyen calculé pour l'ensemble du portefeuille.

- Prothèse auditive : Ce poste de garantie expose une forte incitation à l'anti-sélection, car la base de remboursement de la Sécurité Sociale est souvent trop faible par rapport au coût de ces appareils. Il convient de noter que la réforme du 100% Santé vise à inciter les personnes déjà couvertes par une assurance complémentaire santé à s'équiper gratuitement, ce qui relève davantage de la problématique de l'aléa moral que de l'anti-sélection. Nous excluons le poste du 100% Santé de cette évaluation.

Les coefficients d'anti-sélection sur l'ensemble du contrat sont ainsi évalués, car les assureurs peuvent s'intéresser aux résultats globaux de leur portefeuille pour évaluer l'impact de l'anti-sélection sur l'ensemble de leurs activités.

Base d'assuré 1	Entrée de gamme	Moyenne gamme	Haut de gamme
Prothèse auditive	57.393%	100.755%	167.603%
Consultation généraliste	98.312%	100.872%	101.072%
Ensemble de contrat	84.511%	101.610%	117.838%
Base d'assuré 2	Entrée de gamme	Moyenne gamme	Haut de gamme
Prothèse auditive	70.710%	98.705%	166.339%
Consultation généraliste	99.974%	98.652%	101.370%
Ensemble de contrat	90.057%	99.833%	117.811%

TABLE 5.16 : Comparaison des coefficients d'anti-sélection sur les postes de garantie

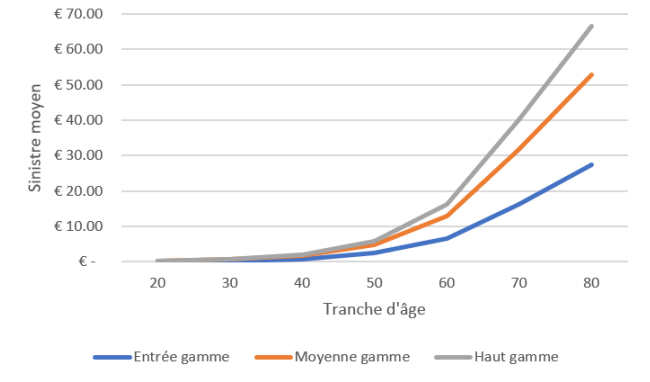


FIGURE 5.4 : Sinistre moyen pour le poste appareil auditif hors de 100% Santé

La table 5.16 nous montre des résultats similaires pour les deux bases, ce qui est normal, car la base d'assurés 2 n'est qu'une modification de la structure d'indépendance des variables, et les probabilités de choix ont très peu changé. En comparant les coefficients des deux sous-familles de garanties, nous pouvons dire que le poste "Consultation généraliste" n'a effectivement pas d'anti-sélection, car les coefficients ne varient pas de manière significative d'une couverture à l'autre pour un seuil de 10%. Cependant, selon plusieurs études différentes, telles que LEGAL (2008), VALDIGUIE (2017), et WEISS (2017), le poste de médecin généraliste présente essentiellement de l'asymétrie d'information en raison de l'existence de niveaux de remboursement différents pour les médecins adhérant à OPTAM/OPTAM-CO, ce qui crée en réalité un risque moral. Le fait que nous n'observions pas cette anti-sélection réside dans le fait que l'asymétrie de l'information se produit à un niveau plus fin que ce que nous

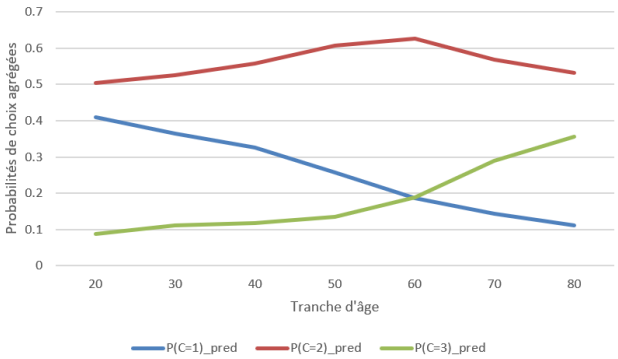


FIGURE 5.5 : Probabilités de choix de niveau de couverture agrégées en fonction de tranche d'âge

pouvons observer, car les données ont été simplifiées pour ne pas tenir compte de la distinction entre OPTAM/OPTAM-CO. Par conséquent, nous n'observons pas d'anti-sélection de manière plus générale sur ce poste.

En revanche, le poste de l'appareil auditif présente une forte anti-sélection liée à la stratégie anticipée des assurés. La couverture minimale n'est pas assez utilisée, tandis que la couverture maximale présente une sinistralité de plus de 50% par rapport au sinistre moyen sur l'ensemble du portefeuille. Une explication simple peut être déduite de la formule des coefficients d'anti-sélection, selon laquelle les proportions d'assurés ayant choisi la couverture minimale ne sont pas réparties de façon homogène dans la population. En effet, une grande proportion de jeunes se concentre sur la couverture de base, tandis qu'une faible proportion de personnes âgées opte pour ce niveau de couverture, ce qui conduit à une pondération de la sinistralité faible, étant donné que la sinistralité des personnes âgées est nettement plus élevée que celle des jeunes (comme le montrent les tables 5.4 et 5.5). De la même manière, la couverture haut de gamme attire un grand nombre d'assurés plus âgés, qui sont les principaux facteurs de consommation médicale, entraînant ainsi une forte sinistralité pour cette couverture.

Cette comparaison des coefficients d'anti-sélection nous permet de confirmer l'hypothèse tirée de la littérature, ainsi que l'existence du phénomène d'anti-sélection dans notre portefeuille, selon notre configuration. Sur l'ensemble du contrat, nous observons moins d'anti-sélection, en raison de la compensation avec les autres postes présentant moins de ce phénomène. Cependant, nous pouvons quand même confirmer qu'il y a de l'anti-sélection dans l'ensemble du portefeuille d'assurance pour des variations de plus de 10%.

5.2.2 L'effet du changement de la stratégie de tarification sur l'anti-sélection

Les avantages des modèles de choix discret ne se limitent pas à la modélisation de la structure du portefeuille, mais s'étendent également à leur capacité à prédire les changements potentiels du portefeuille lors de l'introduction de modifications liées au prix ou à la stratégie de tarification. Pour illustrer cette section, nous utilisons la base d'assurés 2 ainsi que le modèle Logit_ASR_PS que nous avons conservé pour prédire les modifications possibles des coefficients d'anti-sélection. Nous introduisons deux scénarios de modification de la stratégie de tarification en conservant la même segmentation des tarifs des complémentaires santé :

- Augmentation homogène des tarifs des couvertures de moyenne et haut de gamme afin de dissuader ceux qui souhaitent anticiper stratégiquement les contrats de haute qualité et d'encourager les personnes à souscrire la couverture minimale.
- Augmentation hétérogène des tarifs des couvertures de moyenne et haut de gamme.

Après avoir estimé l'impact de l'anti-sélection, les assureurs pourraient modifier leur stratégie de

tarification pour réduire ce phénomène. En pratique, on pourrait dire que l'augmentation des tarifs demandés aux assurés pour les couvertures de moyenne gamme et de haut de gamme peut dissuader l'anticipation stratégique du choix de couverture. De plus, le fait que, dans notre cas, la couverture minimale soit sous-consommée alors que la couverture maximale soit surconsommée peut perturber le résultat du portefeuille. Nous mettons en œuvre une telle stratégie et prédisons à nouveau les probabilités agrégées de choix de couverture, puis calculons les nouveaux coefficients d'anti-sélection. Les nouveaux coefficients prédits par le modèle pourraient être utilisés comme une estimation du niveau d'anti-sélection du portefeuille après l'introduction des modifications, sous l'hypothèse que la consommation moyenne des assurés de chaque segment ayant choisi une couverture n'est que légèrement affectée par le changement de proportion des niveaux de couverture au sein de ce segment. Cette hypothèse stricte est très peu probable d'être le cas, car les assurés ajustent leur niveau de risque et influencent essentiellement la consommation médicale moyenne lorsqu'ils changent de niveau de couverture, tandis que les assureurs n'ont pas accès aux nouvelles données de sinistres de l'année suivante. C'est la raison pour laquelle notre étude de l'impact des stratégies de tarification basée sur une approximation est valable pour une période très courte, soit pour la prédiction de l'anti-sélection du portefeuille dans un an.

Soit ϕ le paramètre de changement des primes demandées, nous recalculons les primes demandées aux assurés de segment a comme suit :

- $Prime_1^{a\prime} = Prime_1^a$.
- $Prime_2^{a\prime} = Prime_2^a \times (1 + \phi)$.
- $Prime_3^{a\prime} = Prime_3^a \times (1 + \phi)$.

Ce changement de tarification n'affecte que la variable Finance et non la variable Plafond, car cette dernière est une mesure de la qualité du contrat telle que perçue par les assurés avant la souscription. En effet, les assurés sont censés percevoir les contrats d'une certaine manière, et dans notre application, nous avons construit cette variable comme le rapport de la prime technique du niveau de couverture considéré à la couverture de moyenne gamme, sous l'hypothèse que cela approxime la qualité relative entre les contrats. Un changement de prix sans modification des attributs de garantie ne peut pas modifier la qualité perçue, sauf dans le cas où ce changement de prix résulte d'une modification des garanties, auquel cas la variable Plafond doit être modifiée.

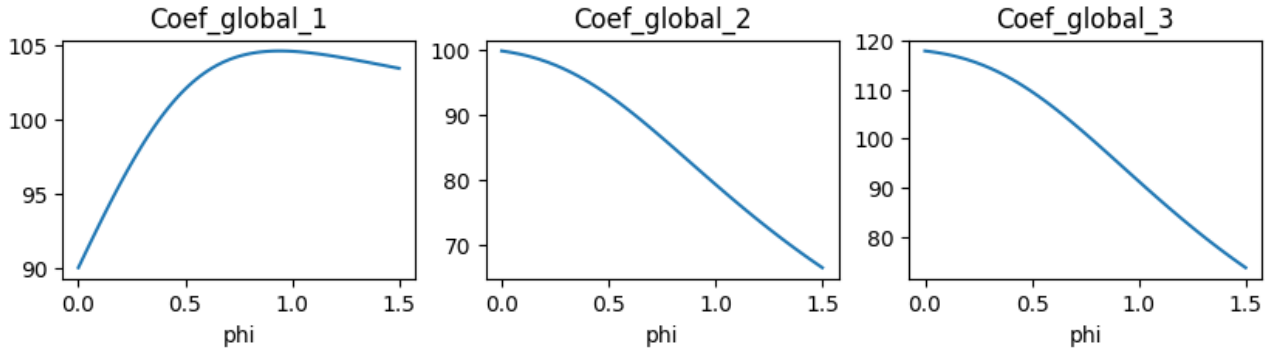


FIGURE 5.6 : Scénario avec l'augmentation homogène : Changement des coefficients d'anti-sélection selon ϕ

Les résultats de la Figure 5.6 montrent une augmentation de $Coef\_global_1$ en raison du fait que les assurés à haut risque des couvertures moyenne et maximale passent à la couverture minimale, ce qui entraîne une baisse des deux autres coefficients d'anti-sélection. Sur la base des résultats, nous

pourrions choisir $\phi \approx 0.75$ comme paramètre de changement, car à cette valeur de ϕ, la couverture maximale ne présente pas d'anti-sélection ($Coef\_global_3 \approx 100\%$). De plus, l'augmentation du sinistre liée à la concentration sur la couverture minimale peut être compensée par une sous-consommation de la couverture de moyenne gamme. Cependant, il ne faut pas exagérer en fixant le paramètre ϕ trop élevé, car cela pourrait dépasser la limite de ce que les assurés sont prêts à payer pour les couvertures de moyenne et de haut de gamme, ce qui risquerait de compromettre la stabilité du portefeuille.

Nous pourrions également envisager des cas où les modifications des primes ne sont pas homogènes entre les différents niveaux de couverture, ce qui se traduirait du scénario 2 par une augmentation des tarifs des contrats de moyenne gamme moins importante que celle des contrats de haut de gamme, soit la moitié du changement de prime de la couverture maximale :

- $Prime_1^{a\prime} = Prime_1^a$.
- $Prime_2^{a\prime} = Prime_2^a \times (1 + 0.5 \times \phi)$.
- $Prime_3^{a\prime} = Prime_3^a \times (1 + \phi)$.

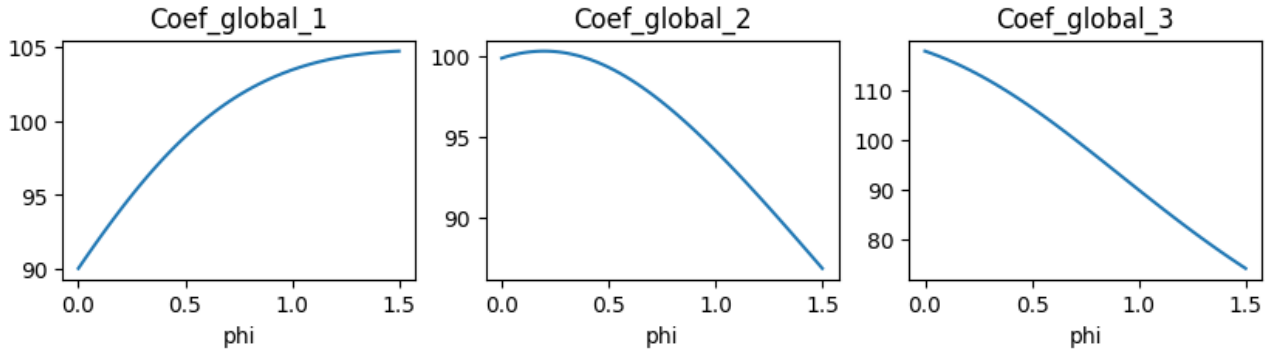


FIGURE 5.7 : Scénario avec l'augmentation homogène : Changement des coefficients anti-sélection selon ϕ

Après avoir examiné les résultats de la Figure 5.7, nous pourrions préférer le deuxième scénario, car pour la même valeur de ϕ permettant d'obtenir $Coef\_global_3 \approx 100\%$, $Coef\_global_1$ augmente moins rapidement dans le scénario 1.

5.3 L'impact d'une agrégation tarifaire sur l'anti-sélection

La section précédente nous aide à comprendre l'impact total de la distorsion des distributions de profils de risque à travers les trois niveaux de couverture considérés. Avec les nouvelles stratégies de tarification évoquées précédemment, on observe une diminution de l'anti-sélection au niveau de la couverture moyenne et maximale, ce qui résulte de la redistribution des profils de risque sur les trois niveaux de couverture. En général, les profils à haut risque ont tendance à opter pour une couverture moins généreuse car l'augmentation très forte de la couverture maximale et un peu moins forte de la couverture moyenne les dissuadent de souscrire ou de maintenir la souscription à ces contrats. Ces changements entraînent également une augmentation des assurés à haut risque (c'est-à-dire à dépenses élevées et plus coûteuses) en couverture minimale et une dilution des profils à haut risque en couverture moyenne et maximale, d'où le résultat constaté.

Cependant, une telle solution peut nuire à l'ensemble de la couverture car elle peut réduire drastiquement le nombre d'assurés dans les deux couvertures généreuses et donc réduire la mutualisation nécessaire pour la stabilisation des contrats complémentaires santé. Il faut donc chercher un moyen

de comprendre les changements de comportement en termes de souscription des profils d'assurés sur chaque niveau de couverture face à un changement de tarification des assureurs. Pour cela, nous utilisons les probabilités de transition abordées à la fin du chapitre 4 pour réaliser cette analyse de distribution des assurés.

Cette section sert donc à illustrer l'analyse des mouvements de distribution de profil de risque lorsqu'on souhaite induire un changement de tarif et expliquer les changements dans les composants de l'anti-sélection. Nous choisissons le changement de tarif actuel (totalement segmenté) à un seul tarif pour tout le monde et analysons à chaque étape de réduction de segmentation, comment le comportement stratégique des assurés change. C'est aussi pour cette raison qu'au début, nous avons retenu la tarification par sexe afin de tenir compte du changement de distribution homme-femme dans chaque couverture. Nous avons choisi l'ordre de segmentation comme tranche d'âge - région - sexe puisque le sexe étant un élément sensible pour la tarification donc il convient de l'enlever en premier, puis les régions car les profiles de risque étant très dépendant aux tranches d'âge alors la variable liée à l'âge serait le dernier à être enlevé. Nous procédons à chaque étape du graphique 5.8 avec le tarif modifié par la réduction de segments de population et interprétons les coefficients anti-sélection correspondants.

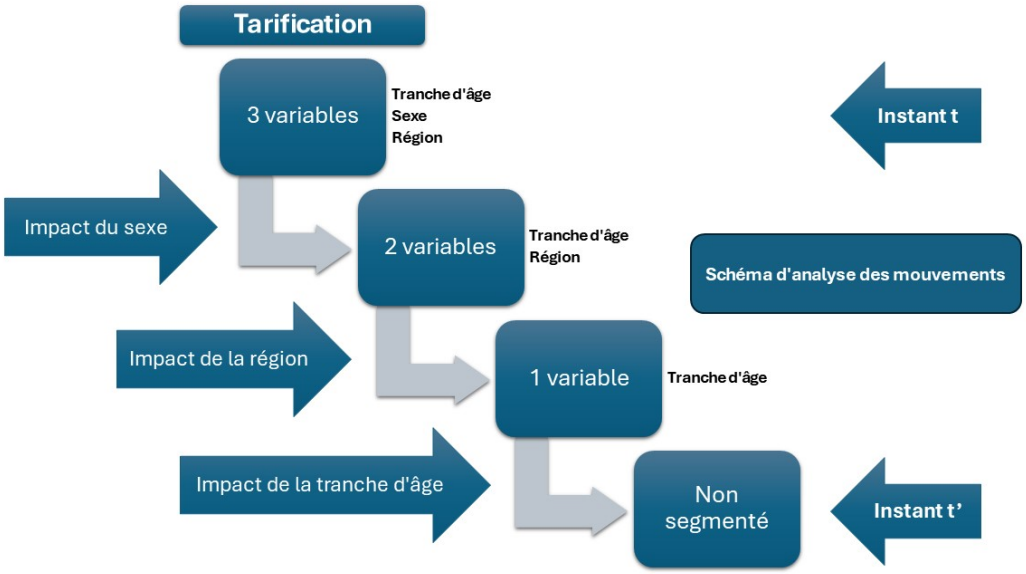


FIGURE 5.8 : Analyse des mouvements des distributions des assurés lors du changement du tarif segmenté à un tarif sans segmentation

5.3.1 Changement de tarification segmentée à un seul tarif

Afin de pouvoir estimer les changements, nous devrons calculer les primes techniques associées à chaque étape de l'analyse. Grâce à la méthode de tarification par expérience, nous avons seulement besoin de connaître le coût total et la population de chaque segment afin de pouvoir donner la sinistralité moyenne par personne de chaque segment. En commençant par les tarifs pleinement segmentés, nous avons 3 variables de segmentation correspondant aux tranches d'âge (7 modalités), au sexe (2 modalités) et aux régions (13 modalités). Pour passer au tarif moins segmenté, comme le tarif segmenté par 2 variables tranche d'âge et région, nous agrégeons les sinistralités et la population des profils de sexe différents afin d'obtenir les sinistralités totales et la population de chaque

segment segmenté par tranche d'âge et région. De même pour les tarifs segmentés par tranche d'âge ou le tarif non segmenté, où nous faisons la somme de tous les sinistres et de la population, puis nous divisons ces deux quantités en supposant que toute l'exposition au risque reste constante et vaut 1. Notons T(Tranche d'âge, Sexe, Région) la tarification totale par les trois variables de segmentation et les autres tarifs respectivement par T(Tranche d'âge, Région), T(Tranche d'âge), T(Unique). Nous présentons des exemples de comparaison des tarifs à chaque étape sur la figure 5.8 pour le niveau de couverture moyen :

1. La figure 5.9 montre que la suppression de la segmentation par Sexe dans les tarifs rend les tarifs des personnes du même profil de tranche d'âge et de région plus élevés pour les hommes et plus bas pour les femmes. Ceci est logique car il prend en compte les fortes dépenses des femmes par rapport aux hommes du même profil. Cet écart s'atténue avec l'augmentation de la tranche d'âge.

2. La figure 5.10 présente les différences de tarif entre les profils de régions différentes avec le tarif lissé commun pour toute région. Nous remarquons que les régions d'Île-de-France, de Provence-Alpes-Côte d'Azur et de Corse, ainsi que l'Auvergne-Rhône-Alpes sont des régions dépensant le plus en santé, notamment l'Île-de-France.

3. La figure 5.11 donne une idée sur les tranches d'âge bénéficiant de ce lissage moyen des tarifs. Pour le niveau de couverture moyen, nous voyons que les personnes à partir de 40 ans peuvent bénéficier du prix baissé, contrairement aux profils jeunes de moins de 40 ans.

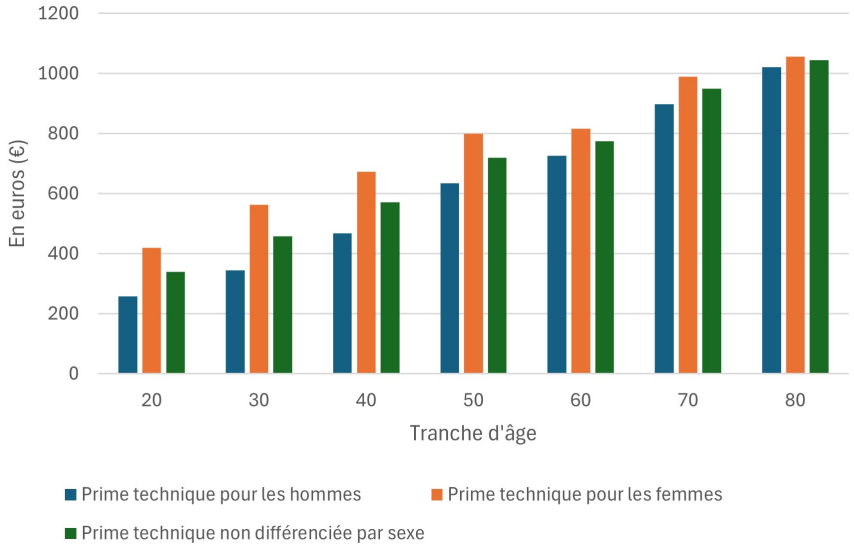


FIGURE 5.9 : Comparaison des primes de niveau de couverture moyen pour le changement de tarif T(Tranche d'âge, Sexe, Région) à T(Tranche d'âge, Région), pour les profils de région d'Île-de-France.

5.3.2 Mouvement des coefficients anti-sélection

Comme l'application de majoration des tarifs, nous appliquons les nouveaux tarifs au modèle Logit_ASR_PS entraîné sur la base 2. Nous obtenons donc des coefficients anti-sélection associés à chaque étape grâce à des nouvelles probabilités de choix. Nous regardons sur les graphes 5.12, 5.13, 5.14 sur l'évolution des coefficients d'anti-sélection pour les postes de consultation généraliste, appareil auditif hors de programme 100% Santé puis sur l'ensemble des postes de garantie.

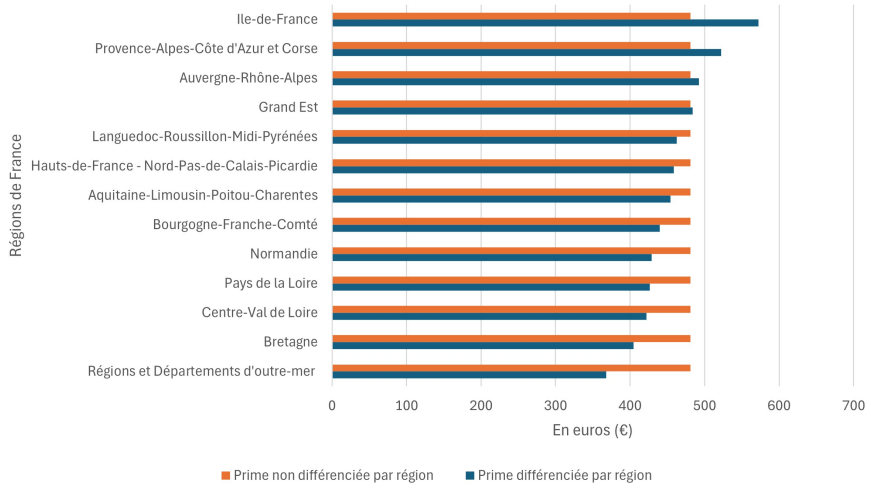


FIGURE 5.10 : Comparaison des primes de niveau de couverture moyen pour le changement de tarif T(Tranche d'âge, Région) à T(Tranche d'âge), pour les profils de tranche d'âge 40-49 ans.

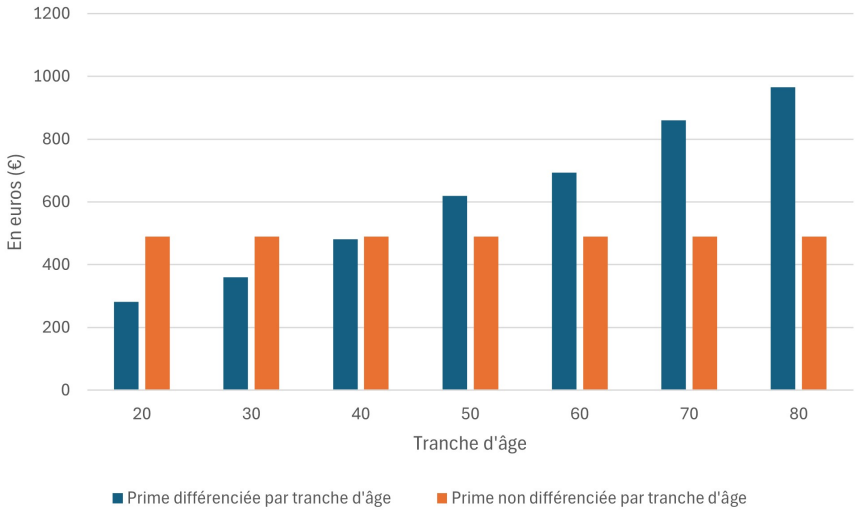


FIGURE 5.11 : Comparaison des primes de niveau de couverture moyen pour le changement de tarif T(Tranche d'âge) à T(Unique).

Nous remarquons qu'en général, les coefficients d'anti-sélection pour les niveaux de couverture moyen et maximum ont tendance à augmenter à chaque étape, lorsque les tarifs deviennent de moins en moins segmentés. Ce phénomène est généralement causé par la migration des profils à haut risque vers les contrats plus généreux, où ils perçoivent une réduction de la prime à payer, tandis que les profils à bas risque migrent vers les contrats moins généreux, où ils doivent payer une prime plus élevée qu'auparavant.

Nous commençons par les deux postes de garanties : l'auditif hors 100% Santé et la consultation généraliste. L'observation quantitative sur les deux figures 5.13 et 5.12 nous montre que le sexe semble être le facteur déterminant, prévalant sur la région et la tranche d'âge lorsqu'il s'agit de consultations médicales générales, alors qu'il l'est beaucoup moins sur le poste auditif. En revanche, la tranche d'âge est le seul facteur à augmenter fortement l'anti-sélection lorsqu'on agrège le tarif par ce facteur, suggérant que dans notre cas, la segmentation par tranche d'âge est essentielle pour maintenir un

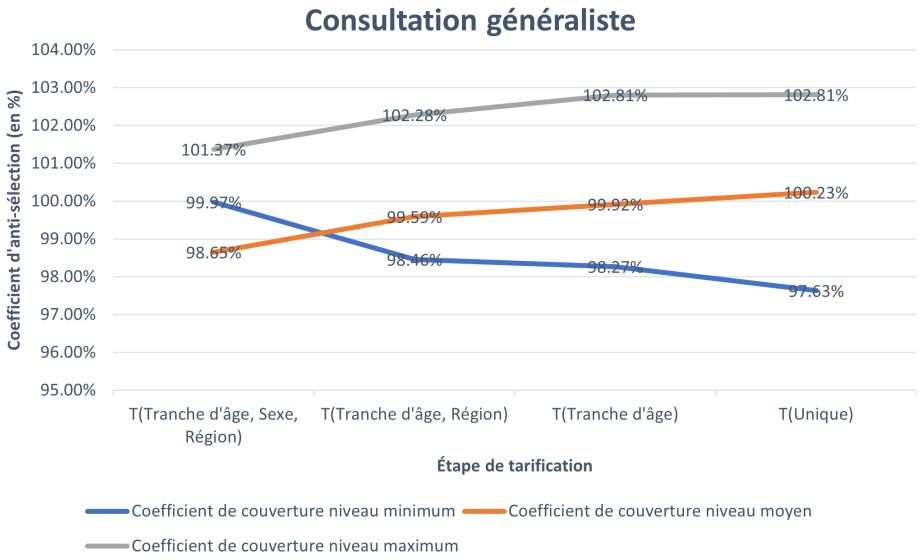


FIGURE 5.12 : L'évolution du coefficient d'anti-sélection du poste de consultation généraliste par étape de changement de tarification

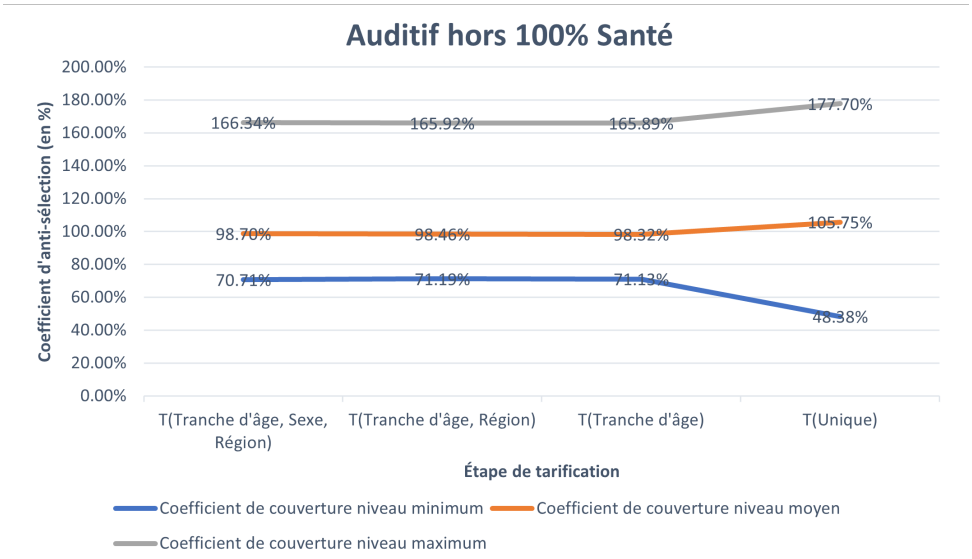


FIGURE 5.13 : L'évolution du coefficient d'anti-sélection du poste d'auditif hors 100% Santé par étape de changement de tarification

niveau d'anti-sélection plus faible.

En ce qui concerne l'ensemble des postes de garantie, nous constatons que la variable "sexe de l'assuré" semble avoir un impact plus significatif sur l'anti-sélection à travers les tarifs que la variable "région de l'assuré". Cependant, ni la variable "sexe" ni la variable "région" ne semblent avoir un impact majeur sur le poste auditif ou sur le total des postes de garantie. Ainsi, nous pouvons conclure que les tranches d'âge de l'assuré demeurent la variable déterminante de l'anti-sélection dans notre portefeuille. Cette analyse par cumul permet de séparer les effets marginaux de chaque variable tout en tenant compte de l'évolution globale du problème.

Cependant, bien que nous puissions expliquer ce phénomène de manière générale et rapide, nous ne comprenons pas entièrement le mécanisme par lequel les distributions des assurés changent après

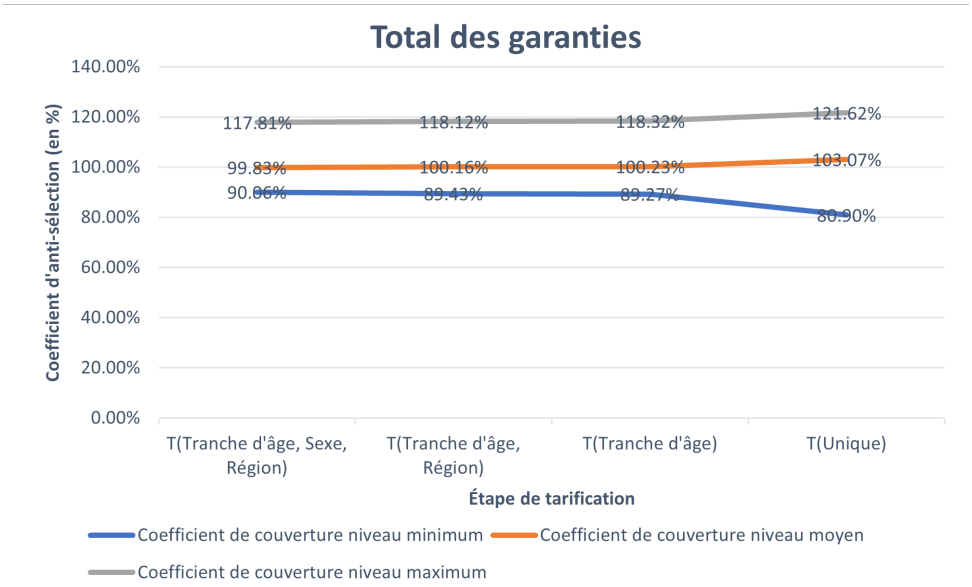


FIGURE 5.14 : L'évolution du coefficient d'anti-sélection du total des postes de garantie par étape de changement de tarification

chaque étape, ce qui rendrait difficile la résolution de l'anti-sélection. En effet, les assureurs ont besoin d'identifier les profils à haut risque et très adverses au risque, principale cause de l'anti-sélection, ainsi que leur comportement, afin de pouvoir ajuster leur stratégie de tarification de manière appropriée. Pour pallier cette difficulté, nous utilisons les matrices de probabilités de transition présentées au chapitre 4 à chaque étape de changement de tarification.

Nous calculons pour chaque assuré de la base 2, face à deux tarifs, l'un avant changement et l'autre après changement, les probabilités conditionnelles de changement $P(Choix_{t'} = j | Choix_t = i, \mathcal{C})$ pour $i, j \in \{1, 2, 3\}$. À travers des résultats agrégés sur la base 2, nous estimons les probabilités de transition moyennes pour les segments de population afin d'expliquer le mouvement des coefficients d'anti-sélection obtenus. Nous cherchons d'abord à expliquer la transition totale de l'anti-sélection d'un changement de tarification de tarif T(Tranche d'âge, Sexe, Région) à T(Unique).

Appliquons le tarif unique sans segmentation à chaque assuré dans la base 2 comme changement de tarif, nous obtenons les probabilités conditionnelles de transition moyennes comme présentées dans les figures C.1 et C.2.

Regardons rapidement les résultats, et nous constatons les effets causés par les primes sur la décision de changement de niveau de couverture des assurés suivants :

- Il y a une tendance de changement vers les couvertures plus généreuses chez les personnes âgées qui ont déjà souscrit aux couvertures minimum, et cette tendance augmente avec la tranche d'âge.
- À l'inverse, les jeunes assurés vont considérablement abandonner les couvertures moyennes et maximales car elles coûtent désormais plus cher qu'auparavant, avec un taux de rabais de niveau de couverture maximale plus élevé qu'au niveau de couverture moyen.
- Entre les hommes et les femmes, les femmes ont plus tendance à changer vers les couvertures plus généreuses ou à rester à ces niveaux de couverture que les hommes.
- Les trois régions qui bénéficient d'une baisse de prime, dont l'Île-de-France, ont tendance à améliorer ou à conserver leur niveau de couverture.

Avec ce raisonnement, plus les niveaux de dépenses de chaque segment déjà connus, nous identifions donc la déformation des distributions de profils à haut risque (notamment les personnes âgées, les femmes, les assurés de la région Île-de-France, etc.) comme la cause de l'augmentation de l'anti-sélection lors d'un changement de tarification à un tarif unique.

Afin d'évaluer l'importance de chaque variable de segmentation dans le tarif sur le mouvement de l'anti-sélection, nous examinons dans les sections suivantes les transitions de l'anti-sélection lors du changement marginal de chaque variable de segmentation considérée.

5.3.3 Analyse de l'impact marginal des variables de segmentation

Étape 1 : Impact de la segmentation par le sexe

Comme décrit dans le schéma 5.8, nous décomposons le changement total de tarif en trois sous-changements. La première étape consiste à analyser le changement de la variable "Sexe de l'assuré" lorsque nous appliquons le tarif à priori T(Tranche d'âge, Sexe, Région) et le tarif à postériori T(Tranche d'âge, Région) aux assurés de la base 2, toujours pour le modèle Logit_ASR_PS.

Les résultats présentés par les graphiques C.3 indiquent que l'égalisation des tarifs entre hommes et femmes forcera une partie des hommes à opter pour des couvertures moins élevées et attirera les femmes vers des couvertures plus élevées. Ce phénomène va probablement générer de l'anti-sélection car les femmes ont en moyenne des dépenses médicales plus élevées que les hommes, ce qui signifie que les couvertures plus élevées auront plus de personnes à haut risque que de personnes à faible risque. Cependant, le fait que les coefficients d'anti-sélection totale (figure 5.14) et du poste auditif hors 100% (figure 5.13) de la figure n'augmentent pas beaucoup à cette étape peut être dû au fait que les entrées-sorties des plans d'assurance ne concernent que les profils jeunes. En effet, plus les tranches d'âge augmentent, moins l'écart de dépenses entre hommes et femmes est constaté, ce qui signifie que les vrais profils à haut risque restent immobiles dans le choix du niveau de couverture. Cet effet marginal du sexe est donc assez faible en présence de la segmentation par tranche d'âge. D'autre part, l'augmentation relativement considérable des coefficients d'anti-sélection pour le poste de consultation généraliste (figure 5.12) peut indiquer un risque d'anti-sélection lié à la fréquence des visites chez les hommes et les femmes, car les femmes prévoient une plus grande fréquentation chez les médecins. C'est donc l'une des raisons pour lesquelles nous avons choisi de conserver la tarification par sexe de l'assuré dans ce mémoire, exposant ainsi le risque d'anti-sélection lié au sexe.

Étape 2 : Impact de la segmentation par région

La deuxième étape consiste à retirer la variable "Région" des assurés après avoir retiré la variable "Sexe". Nous constatons à travers les figures C.4 que, en moyenne, les assurés de trois régions - notamment l'Île-de-France, Provence-Alpes-Côte d'Azur et Corse, ainsi que Auvergne-Rhône-Alpes - sont plus susceptibles d'augmenter leur niveau de couverture lorsque leur niveau initial est minimum, ou de rester sur les couvertures les plus élevées si tel est leur choix initial. Il est très important de remarquer que la segmentation par région n'a d'impact marginal que sur ces trois régions, et que la tendance à l'augmentation du niveau de couverture pour tous ceux qui choisissent initialement le niveau minimum sur les figures C.2 résulte de l'agrégation par tranche d'âge. De plus, même s'il y a des sorties de couverture moyenne et maximale des autres régions qui contribuent à la concentration des hauts risques dans ces deux couvertures, l'augmentation très limitée des coefficients anti-sélection dans la figure 5.14, encore moins que l'augmentation lors de la suppression de la variable sexe, nous suggère que la tarification par région a réellement un impact moins important que par le sexe.

Étape 3 : Impact de la segmentation par tranche d'âge

Pour la troisième et dernière étape, nous remplaçons le tarif segmenté par tranche par un tarif unique pour tous les assurés dans le portefeuille. À travers les figures C.5, nous constatons que les profils jeunes de moins de 39 ans ont tendance à réduire leur niveau de couverture initial, car il est trop coûteux pour eux de mutualiser avec les profils plus âgés. À partir de 40 ans, nous observons une tendance à choisir d'améliorer le niveau de couverture initiale, augmentant avec la tranche d'âge. L'augmentation considérable des coefficients d'anti-sélection lors de la dernière étape pour l'ensemble des garanties (figure 5.14) et le poste d'auditif hors 100% Santé (figure 5.13) montre que cette concentration des profils plus âgés aux niveaux de couverture moyen et maximum est le principal facteur de la forte distorsion de la distribution des profils à haut risque ou d'anti-sélection. En revanche, le coefficient d'anti-sélection du niveau de garantie maximum pour le poste de consultation généraliste ne change pas, seuls les coefficients des couvertures minimum et moyen sont impactés en raison de la tendance de changement de couverture minimale au moyen chez les personnes âgées (figure C.5).

La agrégation par tranche d'âge est également la cause de l'effet systémique du changement de niveau de couverture dans toute la région (figures C.6), car les profils plus âgés, y compris le profil de tranche d'âge représentatif de la France 40-49 ans, sont incités à se couvrir davantage avec une prime réduite pour eux. La sortie de couvertures plus généreuses dans toute la région est donc causée par les profils jeunes. En combinant l'effet de base de la tranche d'âge (figures C.6) et l'effet marginal de la région (figures C.4), nous obtenons les résultats du changement de tarif présentés aux figures C.2, où les probabilités sont plus claires pour les sortants de la couverture minimum.

5.4 Les limites théoriques et pratiques du mémoire

Cette partie aborde les difficultés rencontrées par ce mémoire ainsi que les limites prises en compte. Nous abordons tout d'abord le problème du manque de données des assureurs. En effet, l'utilisation de la base Open Damir peut être assimilée à un portefeuille d'assureur, mais elle présente des inconvénients importants à noter.

Premièrement, cela ne permet pas d'adopter une approche de fréquence-sévérité, car les lignes sont agrégées, ce qui nous force à calculer les sinistres moyens qui ne sont pas très stables pour les postes de garantie moins présents dans la base Damir.

Deuxièmement, le fait que la base manque d'informations liées aux remboursements des médecins OPTAM/OPTAM-CO et aux lignes de régulation de la base risque de biaiser nos résultats et empêche l'utilisation de la base pour évaluer correctement les niveaux de remboursement des complémentaires santé.

Par rapport à notre décision de construire la base d'assurés, bien que l'utilisation de la variable sexe pour la tarification des complémentaires santé soit interdite par la loi, elle a été distinguée volontairement pour ajouter une variable supplémentaire sur laquelle peut être observée l'anti-sélection. Nos résultats de l'analyse de l'impact marginal des variables de segmentation sur l'anti-sélection ont même indiqué que le sexe affecte l'anti-sélection plus que la variable de région, suggérant que les portefeuilles des contrats individuels des assureurs sont actuellement exposés au risque d'anti-sélection créé par le sexe.

Par ailleurs, la généralisation de la méthode d'analyse des changements de comportement suite aux changements de tarification à des portefeuilles réels pourrait être difficile au niveau de la variable de revenu. Cela vient du fait que la souscription à des contrats d'assurance santé individuels ne nécessite pas de revenu de l'assuré, ce qui est différent pour des portefeuilles collectifs. Malgré ce manque de données, l'incorporation des tranches de revenu par segment extraites des portefeuilles collectifs ou des données publiques pourrait éventuellement pallier ce défi. Ce pourrait même être une amélioration des modèles, car les assureurs pourraient potentiellement relever des informations inobservées dans les

données initiales.

Les modèles de choix discrets sont très utiles pour leur interprétation et leur compréhension simples, mais des difficultés surviennent dès que l'on aborde des modèles plus compliqués que le modèle logit multinomial. Les tests de l'hypothèse IANP présentés dans la littérature ne sont pas très forts et donnent parfois de mauvais résultats. Cela conduit au fait que pour tester si l'hypothèse IANP est rejetée, il faut entraîner plusieurs modèles d'hétérogénéité de préférence capables de rejeter le modèle logit simple.

En particulier, pour les modèles de lois mélangées continues, notre approximation peut révéler un manque de considération pour la corrélation possible entre les facteurs aléatoires (ou entre les alternatives) s'il en existe. En effet, le fait de supposer que la matrice de covariance est diagonale pour le modèle de variance spécifique à l'alternative représente une limite importante de la modélisation si la matrice de covariance ne l'est pas en réalité. Cette hypothèse d'indépendance entre les alternatives sur les composants aléatoires n'est pas testable, donc il s'agit bien d'une limite à considérer lors de la spécification du modèle. De plus, l'utilisation des modèles de lois mélangées nécessite une grande capacité de calcul de l'ordinateur pour un grand nombre de simulations. Nos modèles avec seulement 10 simulations ne constituent clairement pas une bonne approximation des coefficients optimaux et représentent donc une limite pratique importante.

Nous remarquons que l'entraînement des modèles de choix discret nécessiterait une procédure de validation croisée, consistant à répartir les données en une base d'entraînement et une base de validation où la vraisemblance du modèle est évaluée sur la base de validation, pour la comparaison des modèles. En effet, l'entraînement des modèles sur l'ensemble des données tend à surajuster les données et à biaiser les scores du modèle. Or, il faut également remarquer que le modèle final doit être entraîné sur toutes les données car le but est de calibrer le modèle aux préférences actuelles des assurés et non de généraliser la prédiction aux nouveaux assurés. La répartition des données entre l'entraînement et le test peut améliorer l'évaluation de la capacité de généralisation du modèle, mais également biaiser les vraies préférences ou les coefficients du modèle en raison du manque de préférences à ajuster. Dans notre cas, la structure mathématique des utilités définit le comportement sous-jacent des assurés. Il est donc normal d'entraîner le modèle sur toute la base de données et de le surajuster aux données, car le but est de prédire le nouveau comportement des assurés en cas de changement des options, en supposant que le modèle reflète la structure de comportement existante, et non de prédire le comportement des nouveaux assurés.

La dernière limite vient de notre choix de modélisation, car nous avons supposé que les assurés devraient choisir un niveau de couverture à souscrire. Ce choix de modélisation simplifie la réalité selon laquelle les assurés peuvent choisir de ne pas souscrire à notre produit et chercher une autre alternative ou un autre assureur, ce qui peut changer les résultats de l'analyse des changements de comportement présentés auparavant.

Conclusion

Ce mémoire a pour l'objectif d'introduire une approche économique de préférence de choix afin de résoudre le problème de quantification et d'explication du phénomène d'anti-sélection en assurance santé. Les principales étapes que nous avons suivies sont :

1. Analyse du portefeuille de l'assureur
2. Traitement des données
3. Entraînement et la sélection des modèles
4. Prédiction ainsi que l'analyse des résultats

En premier lieu, l'analyse de la base de données Open Damir nous a fourni des informations sur les habitudes de consommation des assurés, ainsi que les disparités significatives liées à l'âge des assurés. La tarification par méthode d'expérience nous a donné un *benchmark* des primes techniques sur le marché qui sont comparables à des tarifs d'un vrai assureur. Cependant, nous avons fait des compromis liés au manque d'informations des actes des médecins OPTAM/OPTAM-CO, aux régulations de la Sécurité Sociale et aux nombres des assurés de portefeuille. À la fin de cette étape, nous avons identifié les profils à haute dépense dans le portefeuille sur la base de trois variables : tranche d'âge, sexe et région, qui sont les suivants :

- Les femmes consomment en moyenne plus que les hommes pour n'importe quelle tranche d'âge ;
- Les personnes âgées à partir de 50 ans ont une dépense moyenne qui augmente rapidement avec la tranche d'âge ;
- Les trois régions Île-de-France, Provence-Alpes-Côte d'Azur (et Corse) et Auvergne-Rhône-Alpes, consomment en moyenne plus que les autres régions, en particulier l'Île-de-France.

Dans un deuxième temps, nous avons effectué une revue bibliographique sur le sujet de l'anti-sélection en assurance santé individuelle, dont nous avons déduit l'importance du choix de couverture anticipé des assurés. Nous avons fait l'hypothèse de séparabilité de l'anti-sélection et l'aléa moral pour les produits individuels à adhésion facultative, grâce au niveau de remboursement minimum (panier de soin) des complémentaires santé posé par la Sécurité Sociale. L'étude de la littérature a révélé que la demande d'assurance en santé ne dépend pas seulement de l'aversion au risque monétaire, mais aussi du risque de préférence en santé. Reformulé de manière mathématique, il s'agit d'une probabilité de choix de couverture, nous introduisons ainsi une mesure de distorsion de distribution de profil de risque dans le portefeuille qui prend en compte à la fois les composants de segmentation de population dans chaque niveau de couverture et leur sinistralité. Nous avons ainsi généré des portefeuilles avec un certain niveau d'anti-sélection à partir de données publiques qui sera la source de l'anti-sélection dans notre étude. Nous avons trouvé un niveau d'anti-sélection élevé sur toutes les postes de garantie, mais cette dernière est particulièrement importante pour l'auditive d'auditive hors 100%.

Ensuite, pour répondre à la question de la modélisation de référence de choix, nous avons eu recours aux modèles de choix discrets, tout en mettant en évidence les deux effets importants dans le contexte de l'anti-sélection en assurance santé : l'effet de salaire (proxy de risque monétaire) et l'effet de qualité du contrat d'assurance (proxy de référence pour le risque de préférence en santé). Ces modèles, grâce à l'explicabilité en matière d'utilité pour l'assuré, nous permettent de comprendre facilement l'impact de chaque facteur sur le choix d'assurance, d'où le point fort des modèles d'utilité aléatoire pour notre problématique de recherche sur l'anti-sélection. La prédictivité du modèle économétrique est ensuite abordée et nous présentons les probabilités de transition d'alternative lors d'un changement d'hypothèses.

Dans un dernier temps, notre application des modèles sur deux bases de données simulées a confirmé leur pertinence et leur capacité à capturer les préférences des assurés. Les résultats obtenus correspondent à nos attentes, démontrant que même des modèles simples sont en mesure de représenter efficacement des données complexes. En effet, nous avons retenu les modèles *Logit Multinomial* en acceptant qu'il n'existe pas d'hétérogénéité inobservable dans nos données générées.

Nous démontrons ainsi qu'il est possible de réduire les coefficients d'anti-sélection, c'est-à-dire de diluer les profils à haut risque dans les couvertures moyenne et maximum à travers la stratégie de majoration des tarifs. Dans notre exemple, la majoration des primes pour les niveaux de couverture moyen et maximum peut réduire l'anti-sélection. Cependant, elle reste irréalisable dans la pratique, car le niveau de majoration est de plus de 50% de la prime actuelle, ce qui est très cher pour l'assuré et peut inciter les assurés à se retirer. Avec pour objectif de trouver des solutions efficaces à l'anti-sélection à partir des stratégies de tarification, nous pouvons essayer de construire un système de coefficients de majoration pour chaque segment de population, basé sur leur sinistralité et leur concentration dans chaque niveau de couverture, au lieu d'un coefficient de majoration unique.

La fin du mémoire porte sur l'analyse des mouvements d'anti-sélection lors d'une déségmentation des tarifs, nous éclairant sur les effets marginaux des variables de segmentation du tarif sur les préférences des assurés. À travers les résultats, nous constatons que les deux variables "Sexe de l'assuré" et "Région de l'assuré" ont des effets marginaux plus faibles que l'effet de base causé par "Tranche d'âge de l'assuré". À savoir que par la tarification avec le sexe et la construction de la base d'entraînement sur ces tarifs, nous sommes capables de séparer cet effet marginal alors qu'en pratique les assureurs n'auront pas cette possibilité car la tarification par sexe n'est pas autorisée par la Gender Directive. Il s'agit d'une différence entre ce mémoire et la réalité.

En parallèle de notre application du modèle sur la prédiction du niveau de l'anti-sélection lors d'un changement de tarification, l'élasticité de la prime d'assurance et du salaire apparaît comme des instruments pour analyser l'anti-sélection. En effet, Ringel et al. (2002) parle de l'anti-sélection comme du fait que les assurés de haut risque tendent à souscrire les couvertures généreuses et ceux de faible risque tendent à éviter ces couvertures. Dans cette situation de présence de l'anti-sélection, l'élasticité des primes d'assurance va mesurer les besoins des assurés et donc distinguer les profils facteurs de l'anti-sélection.

Pendant la réalisation de ce mémoire, nous avons été confrontés à une limite importante de manque de données réelles pour laquelle nous avons dû recourir à la méthode de tarification par expérience. Dans le cas où nous aurions des données de sinistralité individuelle sur une période de temps continue et des choix individuels observables, il serait intéressant d'appliquer les méthodes de tarification de type *Modèle Linéaire Généralisé* ou *Machine Learning* et de calibrer les distributions de sinistres conditionnellement au choix de niveau de couverture. Ainsi, les modèles de choix discrets pourraient être améliorés par l'ajout de données panel. Dans ce cas, nous pourrions tenter de calculer la limite asymptotique des coefficients d'anti-sélection et d'étudier la relation de causalité entre le sinistre et le choix de couverture.

Pour les modèles de choix discrets, Feldman et al. (1989) ont également indiqué que l'utilisation des modèles *Logit Emboîté* (ou *Nested Logit Model*) dans un contexte de choix d'assurance santé aux

États-Unis est très recommandée, sous prétexte qu'il existe des alternatives semblables (les mêmes fournisseurs de soins, les mêmes hôpitaux, le même mécanisme de remboursement, etc.). En France, les contrats complémentaires santé se différencient uniquement par leur niveau de couverture, ce qui les rend très semblables, et l'utilisation des modèles Logit Emboîté peut s'avérer plus adaptée que notre modèle *Logit Multinomial* retenu.

Une autre limite est apparue lors de l'entraînement des modèles, à savoir le temps de calcul et les ressources nécessaires pour entraîner les modèles à l'hétérogénéité inobservable avec la méthode de Monte Carlo. Afin d'exécuter les modèles sous contraintes matérielles, nous avons seulement réalisé cinq simulations de calcul de vraisemblance. Certes insuffisante, mais nous permet d'effectuer les premières analyses. Une solution possible pour pallier cette difficulté serait de réduire le nombre de modalités de chaque variable qualitative pour ne retenir que les modalités réellement importantes.

Enfin, ce mémoire propose une réflexion sur les possibilités d'amélioration de la tarification pour réduire l'anti-sélection en assurance santé. Il suggère également des pistes pour de futures recherches, notamment en termes de modélisation des comportements des assurés et d'exploitation des données disponibles. En définitive, ce travail a contribué à la compréhension et à la gestion de l'anti-sélection en assurance santé, offrant des perspectives prometteuses pour l'industrie de l'assurance et ouvrant la voie à des développements futurs dans ce domaine.

Bibliographie

Abdullah, S., Markandya, A. et Nunes, P. (2011). Introduction to economic valuation methods. *Research tools in natural resource and environmental economics* 5, p. 143-187.

Alan (11 août 2020). Pourquoi on fait une pause sur les contrats individuels. Visité le 29/04/2024. url : https://alan.com/fr-fr/blog/etre-partenaire-sante/a/pourquoi-on-fait-une-pause-sur-les-contrats-individuels.

Albouy, V et Crepon, B (2007). Aléa moral en santé: une évaluation dans le cadre du modèle causal de Rubin. *Institut National de la Statistique et des Etudes Economiques.(INSEE). Direction des Etudes et Synthèses Economiques*.

Alessie, R. J. M., Angelini, V., Mierau, J. O. et Viluma, L. (juill. 2020). Moral hazard and selection for voluntary deductibles. *Health Economics* 29.10, 1251–1269.

Assurance Maladie (3 nov. 2023a). Consultations en métropole : vos remboursements. Visité le 15/11/2023. url : https://www.ameli.fr/assure/remboursements/rembourse/consultations-telemedecine/metropole.

Assurance Maladie (3 juill. 2023b). Le tiers payant. Visité le 16/08/2023. url : https://www.ameli.fr/hauts-de-seine/assure/remboursements/etre-bien-rembourse/tiers-payant#:~:text=D%C3%A9finition%20du%20tiers%20payant,de%20sant%C3%A9%20exer%C3%A7ant%20en%20ville..

Assurance Maladie (16 avr. 2023c). Notre environnement : la Sécurité sociale. Visité le 16/04/2023. url : https://assurance-maladie.ameli.fr/qui-sommes-nous/organisation/securite-sociale.

Assurance Maladie (5 mai 2023d). Notre histoire. Visité le 03/05/2023. url : https://assurance-maladie.ameli.fr/qui-sommes-nous/histoire.

Assurance Maladie (10 oct. 2023e). Présentation du système national des données de santé (SNDS). Visité le 23/10/2023. url : https://assurance-maladie.ameli.fr/etudes-et-donnees/en-savoir-plus-snds/presentation-systeme-national-donnees-sante-snds.

Assurance Maladie (19 avr. 2024). Open Damir : base complète sur les dépenses d'assurance maladie - 2009 à 2023. Visité le 16/04/2023. url : https://assurance-maladie.ameli.fr/etudes-et-donnees/open-damir-depenses-sante-interregimes.

Bakker, F. et van Vliet, R. (1993). The Effect of Deductibles on Premiums in Health Insurance: A Case Study on Prescription Drugs. *2nd European Workshop on Econometrics and Health Economics*. Center for Health Economics, University of York.

Barlet, M., Gaini, M, Gonzalez, L et Legal, R (2019). La complémentaire santé : acteurs, bénéficiaires, garanties - édition 2019. *Panoramas de la DREES*.

Becker, K. et Zweifel, P. (jan. 2008). Age and Choice in Health Insurance: Evidence from a Discrete Choice Experiment. *The Patient: Patient-Centered Outcomes Research* 1.1, 27–40.

Ben-Akiva, M., Bolduc, D. et Walker, J. (2001). Specification, Identification, and Estimation of the Logit Kernel (or Continuous Mixed Logit) Model. Rapp. tech. Working paper. Massachusetts Institute of Technology.

Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H. et Rao, V. (1994). Combining revealed and stated preferences data. *Marketing Letters* 5, p. 335-349.

Bien, F. (2001). Essais en Économie de la Santé et Assurance. Thèse de doct. Paris 10.

Bierlaire, M. (2016). PythonBiogeme: a short introduction. Rapp. tech. TRANSP-OR 160706. Transport, Mobility Laboratory, School of Architecture, Civil et Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Bierlaire, M. (2020). A short introduction to PandasBiogeme. *A short introduction to PandasBiogeme*.

Bierlaire, M. (sans date). EPFLx: Introduction to Discrete Choice Models. Visité le 21/10/2023. URL : https://www.edx.org/learn/discrete-mathematics/ecole-polytechnique-federale-de-lausanne-introduction-to-discrete-choice-models.

Billot, A. et Thisse, J.-F. (1995). Modèles de choix individuels discrets: théorie et applications à la micro-économie. *Revue économique*, p. 921-931.

Birchall, C. et Verboven, F. (2022). Estimating Substitution Patterns and Demand Curvature in Discrete-Choice Models of Product Differentiation. *CEPR Discussion Paper* DP16981.

Caillol, H. (2015). Ouverture des données de santé: l'expérience de l'Assurance maladie. *Informations sociales* 5, p. 60-67.

Cameron, A. C., Trivedi, P. K., Milne, F. et Piggott, J. (jan. 1988). A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia. *The Review of Economic Studies* 55.1, p. 85.

Carson, R. T., Flores, N. E. et Meade, N. F. (2001). Contingent valuation: controversies and evidence. *Environmental and resource economics* 19, p. 173-210.

Caussat, L. et Glaude, M. (1993a). Dépenses médicales et couverture sociale. *Economie et statistique* 265.1, p. 31-43.

Caussat, L. et Glaude, M. (1993b). Dépenses médicales et couverture sociale. *Economie et statistique* 265.1, 31–43.

Chen, Q. (1995). A comparison between revealed preference (RP) and stated preference (SP) based on results of simulations. Thèse de doct. New Jersey Institute of Technology.

Cheng, S. et Long, J. S. (2007). Testing for IIA in the multinomial logit model. *Sociological methods & research* 35.4, p. 583-600.

Cherchi, E., Polak, J. et Hyman, G. (2004). The impact of income, tastes and substitution effects on the assessment of user benefits using discrete choice models. *European Transport Conference, Strasbourg*. T. 10.

Cutler, D. M. et Zeckhauser, R. J. (jan. 1998). Adverse Selection in Health Insurance. *Forum for Health Economics amp; Policy* 1.1.

Davidson, R. et MacKinnon, J. (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. *Econometrica* 49.3, p. 781-93.

De Palma, A. et Kilani, K. (2011). Transition choice probabilities and welfare analysis in additive random utility models. *Economic Theory* 46, p. 427-454.

De Williencourt, C. (2022). Rapport 2022 Sur la situation financière des organismes complémentaires assurant une couverture santé. Rapport. DRESS. URL : https://drees.solidarites-sante.gouv.fr/publications-communique-de-presse-documents-de-reference/rapports/rapport-2022-sur-la-situation.

Delle Site, P., de Palma, A. et Kilani, K. (2022). Consumers' welfare and compensating variation: survey and mode choice application.

Delle Site, P. et Salucci, M. V. (2013). Transition choice probabilities and welfare analysis in random utility models with imperfect before–after correlation. *Transportation Research Part B: Methodological* 58, p. 215-242.

Denuit, M. et Charpentier, A. (2005). Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement.

Dgccrf (Direction Générale De La Concurrence, De La Consommation Et De La Répression Des Fraudes) (sans date). Assurance complémentaire santé. Visité le 04/08/2023.

URL : https://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Assurance-complementaire-sante.

Dutang, C. (2021). Cours d'Actuariat 1. Support de cours. Université Paris-Dauphine, p. 88-89.

Einav, L. et Finkelstein, A. (2011). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic perspectives* 25.1, p. 115-138.

Ericson, K. M. et Sydnor, J. (2017). The questionable value of having a choice of levels of health insurance coverage. *Journal of Economic Perspectives* 31.4, p. 51-72.

Feldman, R., Finch, M., Dowd, B. et Cassou, S. (1989). The demand for employment-based health insurance plans. *Journal of Human Resources*, p. 115-142.

Fotsing, A. C. M. (2018). L'open DAMIR : apport à la maîtrise des dépenses de santé. Mémoire d'actuariat. ISFA, Univ. Clauder Bernard Lyon 1.

Franc, C., Perronnin, M et Pierre, A (2010a). Qui a souscrit une sur-complémentaire ? Une analyse dynamique de l'auto-sélection. *Questions d'économie de la santé* 150.

Franc, C., Perronnin, M. et Pierre Aurélie in collaboration with Cases, C. (2010b). Who Took out Additional Supplementary Health Insurance? A dynamic Analysis of Adverse-Selection.

Fry, T. R. et Harris, M. N. (1996). A Monte Carlo study of tests for the independence of irrelevant alternatives property. *Transportation Research Part B: Methodological* 30.1, p. 19-30.

Fry, T. R. et Harris, M. N. (1998). Testing for independence of irrelevant alternatives: some empirical results. *Sociological Methods & Research* 26.3, p. 401-423.

Gardiol, L., Geoffard, P.-Y. et Grandchamp, C. (oct. 2005). Separating selection and incentive effects in health insurance. working paper or preprint. URL : https://shs.hal.science/halshs-00590713.

Gaudry, M. J., Jara-Diaz, S. R. et de Dios Ortúzar, J. (1989). Value of time sensitivity to model specification. *Transportation Research Part B: Methodological* 23.2, p. 151-158.

Genier, P. (1998). Assurance et recours aux soins. Une analyse microéconométrique à partir de l'enquête Santé 1991-1992 de l'Insee. *Revue économique*, p. 809-819.

Geruso, M., Layton, T. et Leive, A. (juill. 2023). The Incidence of Adverse Selection: Theory and Evidence from Health Insurance Choices.

Gorter, J. et Schilp, P. (2012). Risk Preferences Over Small Stakes: Evidence from Deductible Choice. *SSRN Electronic Journal*.

Grignon, M., Kambia-Chopin, B. et al. (2009). Income and the demand for complementary health insurance in France. Institut de recherche et documentation en économie de la santé.

Hausman, J. et McFadden, D. (1984). Specification Tests for the Multinomial Logit Model. *Econometrica* 52.5, p. 1219-40.

Hensher, D. A. et Greene, W. H. (2003). The mixed logit model: the state of practice. *Transportation* 30, p. 133-176.

Hurvich, C. M. et Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, p. 1077-1084.

Huynh, S. (2021). Open data et Assurance santé : l'union fait la force ? Mémoire d'actuariat. ISFA, Univ. Claude Bernard Lyon 1.

Insee (18 oct. 2021). Niveau de vie. Visité le 25/08/2023. URL : https://www.insee.fr/fr/metadonnees/definition/c1890.

Insee (24 jan. 2023a). Estimation de la population au 1er janvier 2023. Séries par région, département, sexe et âge de 1975 à 2023. Estimation de population par région, sexe et âge quinquennal - Années 1975 à 2023. Visité le 20/07/2023. URL : https://www.insee.fr/fr/statistiques/1893198.

Insee (14 nov. 2023b). Niveau de vie selon l'âge - Données annuelles de 1996 à 2021. Visité le 03/08/2023. URL : https://www.insee.fr/fr/statistiques/2416878.

Insee (25 juill. 2023c). Salaire net horaire moyen selon la catégorie socioprofessionnelle, le sexe et l'âge en 2021. Insee, Bases tous salariés, fichier salariés au lieu de résidence. Résultats pour les communes, arrondissements, régions, départements, zones d'emploi, EPCI de France hors Mayotte. Salaire net

horaire moyen selon la catégorie socioprofessionnelle, le sexe et l'âge en 2019 (en géographie au 01/01/2022). Visité le 25/08/2023. URL : https://www.insee.fr/fr/statistiques/2021266.

INSEE (sans date). Pyramide des âges au 1er janvier 2024. Visité le 03/05/2023. URL : https://www.insee.fr/fr/outil-interactif/5014911/pyramide.htm.

JOHNSON, F. R., LANCSAR, E., MARSHALL, D., KILAMBI, V., MÜHLBACHER, A., REGIER, D. A., BRESNAHAN, B. W., KANNINEN, B. et BRIDGES, J. F. (2013). Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in health* 16.1, p. 3-13.

KALOUGUINA, V. et WAGNER, J. (avr. 2020). How Do Health, Care Services Consumption and Lifestyle Factors Affect the Choice of Health Insurance Plans in Switzerland? *Risks* 8.2, p. 41.

KEANE, M. (2004). Modeling health insurance choice using the heterogeneous logit model.

KIRCHER, P., ERICSON, K. M. M., SPINNEWIJN, J. et STARC, A. (2015). Inferring Risk Perceptions and Preferences Using Choice from Insurance Menus: Theory and Evidence.

KRUEGER, R., BIERLAIRE, M., GASOS, T. et BANSAL, P. (2023). Robust discrete choice models with t-distributed kernel errors. *Statistics and Computing* 33.1, p. 2.

KRUEGER, R., RASHIDI, T. H. et VIJ, A. (2020). A Dirichlet process mixture model of discrete choice: Comparisons and a case study on preferences for shared automated vehicles. *Journal of choice modelling* 36, p. 100229.

LA SÉCURITÉ SOCIALE (22 avr. 2022). Quels sont les tarifs d'un médecin (conventionné ou non) ? Visité le 03/05/2023. URL : https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-branches.

LA SÉCURITÉ SOCIALE (2023). Les chiffres clés de la Sécurité Sociale 2022 - Edition 2023. Rapp. tech. URL : https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/DSS/2023/Chiffres%20cles%20de%20la%20DSS%202022.pdf.

LA SÉCURITÉ SOCIALE (sans date[a]). Les branches. Visité le 03/05/2023. URL : https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-branches.

LA SÉCURITÉ SOCIALE (sans date[b]). Les grandes dates. Visité le 03/05/2023. URL : https://www.securite-sociale.fr/la-secu-cest-quoi/histoire/les-grandes-dates.

LEGAL, R. (2008). Les déterminants de la demande individuelle de couverture complémentaire santé en France. Rapp. tech. Université Paris-Dauphine.

LEGAL, R. (mars 2009). L'influence de l'offre de soins et du niveau des primes sur la demande d'assurance complémentaire santé en France. *Revue économique* Vol. 60.2, 441–453.

LIU, Y. et BOES, S. (avr. 2022). On the Relative Importance of Different Factors Explaining Health Plan Choices: Evidence From Mandatory Health Insurance in Switzerland. *Frontiers in Health Services* 2.

LUCE, R. D. (1959). Individual choice behavior, 1959. *Publisher: John Wiley*.

MAJNONI D'INTIGNANO, B. (2013). Capital santé et demande de soins. *Santé et économie en Europe*. Sous la dir. de MAJNONI D'INTIGNANO, B. Que sais-je ? Paris cedex 14 : Presses Universitaires de France, p. 63-77.

MALAKOFF HUMANIS (30 nov. 2020). Frais de gestion et taux de redistribution des mutuelles : de quoi parle-t-on ? Visité le 03/05/2023. URL : https://www.malakoffhumanis.com/s-informer/sante/frais-gestion-et-taux-redistribution-de-quoi-parle-t-on/.

MANGHAM, L. J., HANSON, K. et MCPAKE, B. (2009). How to do (or not to do)... Designing a discrete choice experiment for application in a low-income country. *Health policy and planning* 24.2, p. 151-158.

MARQUIS, M. S. (1992). Adverse selection with a multiple choice among health insurance plans: a simulation analysis. *Journal of Health Economics* 11.2, p. 129-151.

MARQUIS, M. S. et HOLMER, M. R. (1986). Choice Under Uncertainty and the Demand for Health Insurance. Santa Monica, CA : RAND Corporation.

Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. *Econometrica: Journal of the Econometric Society*, p. 111-141.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behaviour. *Frontiers in Econometrics*. Sous la dir. de Zarembka, P. New York, NY, USA : Academic Press New York, p. 105-142.

McFadden, D. (2000). Disaggregate behavioral travel demand's RUM side. *Travel behaviour research*, p. 17-63.

McFadden, D. et Train, K. (2000a). Mixed MNL models for discrete response. *Journal of applied Econometrics* 15.5, p. 447-470.

McFadden, D. et Train, K. (2000b). Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15.5, p. 447-470. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/1099-1255%28200009/10%2915%3A5%3C447%3A%3AAID-JAE570%3E3.0.CO%3B2-1.

Newhouse, J. P., Price, M., Huang, J., McWilliams, J. M. et Hsu, J. (déc. 2012). Steps To Reduce Favorable Risk Selection In Medicare Advantage Largely Succeeded, Boding Well For Health Insurance Exchanges. *Health Affairs* 31.12, 2618–2628.

Pierre, A. et Rochereau, T. (mai 2022). L'absence de couverture par une complémentaire santé en France en 2019. Premiers résultats de l'Enquête santé européenne (EHIS). *QUESTIONS D'ECONOMIE DE LA SANTE (IRDES)* 268, p. 1-6.

Plantier, M. (2021). Essays on Behavioral Economics on Social Protection. Thèse de doct. Université Claude Bernard Lyon 1 (UCBL).

Powell, D. et Goldman, D. (2021). Disentangling moral hazard and adverse selection in private health insurance. *Journal of Econometrics* 222.1, Part A. Annals Issue: Structural Econometrics Honoring Daniel McFadden, p. 141-160.

Remmerswaal, M., Boone, J., Douven, R. C. et al. (2019). Selection and moral hazard effects in healthcare. Rapp. tech. CPB Netherlands Bureau for Economic Policy Analysis.

Ringel, J. S., Hosek, S. D., Vollaard, B. A. et Mahnovski, S. (2002). The elasticity of demand for health care a review of the literature and its application to the military health system. *RAND-PUBLICATIONS-MR-ALL SERIES-*.

Rothschild, M. et Stiglitz, J. (nov. 1976). Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information*. *The Quarterly Journal of Economics* 90.4, p. 629-649. eprint : https://academic.oup.com/qje/article-pdf/90/4/629/5331107/90-4-629.pdf.

Régime Local d'Assurance Maladie d'Alsace Moselle (4 juill. 2022). Rapport d'Activité 2021. Rapp. tech. url : https://regime-local.fr/2022/07/04/rapport-dactivite-2022/.

Saha, A. (1993). Expo-power utility: a 'flexible'form for absolute and relative risk aversion. *American Journal of Agricultural Economics* 75.4, p. 905-913.

Saliba, B. et Ventelou, B. (mai 2007). Complementary health insurance in France Who pays? Why? Who will suffer from public disengagement? *Health Policy* 81.2–3, 166–182.

Schneider, P. (nov. 2004). Why should the poor insure? Theories of decision-making in the context of health insurance. *Health Policy and Planning* 19.6, 349–355.

Sfeir, G., Rodrigues, F. et Abou-Zeid, M. (2022). Gaussian process latent class choice models. *Transportation Research Part C: Emerging Technologies* 136, p. 103552.

Shen, J. (oct. 2009). Latent class model or mixed logit model? A comparison by transport mode choice data. *Applied Economics* 41, p. 2915-2924.

Sifringer, B., Lurkin, V. et Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological* 140, p. 236-261.

Simplicia (2 sept. 2020). Norme NOEMIE OC Ce qu'il faut savoir sur la Norme NOEMIE OC. Visité le 03/05/2023. url : https://www.simplicia.co/blog/norme-noemie-oc/.

Small, K. A. et Hsiao, C. (1985). Multinomial Logit Specification Tests. *International Economic Review* 26.3, p. 619-627.

Snds (Système National Des Données De Santé) (sans date). Documentation technique. Visité le 03/05/2023. url : https://www.snds.gouv.fr/SNDS/Documentation-technique.

Soekhai, V., de Bekker-Grob, E. W., Ellis, A. R. et Vass, C. M. (2019). Discrete choice experiments in health economics: past, present and future. *Pharmacoeconomics* 37, p. 201-226.

Train, K. E. (2009). Discrete choice methods with simulation. Cambridge university press.

Tranthimy, L. (19 juin 2020). 17 291 adhérents aux OPTAM, bientôt des indicateurs avec warning. Le quotidien du medecin.

Tutz, G. (2021). Uncertain choices: the heterogeneous multinomial logit model. *Sociological Methodology* 51.1, p. 86-111.

Valdiguie, M. (2017). Mesure du risque d'anti-sélection en Assurance Santé Collective. Mémoire d'actuariat. Université de Paris Dauphine.

Van Vliet, R. C. J. A. (déc. 2004). Deductibles and Health Care Expenditures: Empirical Estimates of Price Sensitivity Based on Administrative Data. *International Journal of Health Care Finance and Economics* 4.4, 283–305.

Van de Ven, W. P. et Van Praag, B. M. (nov. 1981). The demand for deductibles in private health insurance. *Journal of Econometrics* 17.2, 229–252.

Walker, J., Ben-Akiva, M. et Bolduc, D. (2004). Identification of the logit kernel (or mixed logit) model. *10th International Conference on Travel Behavior Research, Lucerne, Switzerland, August*.

Wang, J. (2015). Tarification santé : Mesure des risques associés aux produits modulaires. Mémoire d'actuariat. EURIA.

Weiss, A. (2017). Mesure du phénomène d'antisélection sur les options facultatives en santé. Mémoire d'actuariat. ISFA, Univ. Claude Bernard Lyon 1.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics* 9.1, p. 60-62.

Yerle, L. (2020). Apport de l'Open Data à la modélisation de la consommation médicale. Mémoire d'actuariat. Université de Paris Dauphine.

Zhou, M. (2017). Exploring heterogeneity of stated preferences through latent class analysis. Thèse de doct. Johns Hopkins University.

Annexe A

Traitement des données de la base Open Damir

A.1 Tableau des variables de la base Open Damir

Période	
FLX_ANN_MOI	Année et Mois de Traitement
SOI_ANN	Année de Soins
SOI_MOI	Mois de Soins
Prestation	
PRS_NAT	Nature de Prestation
ASU_NAT	Nature d'Assurance
ATT_NAT	Nature de l'Accident du Travail
CPT_ENV_TYP	Type d'Enveloppe
CPL_COD	Complément d'Acte
EXO_MTF	Motif d'Exonération du Ticket Modérateur
MTM_NAT	Modulation du Ticket Modérateur
PRS_REM_TAU	Taux de Remboursement
PRS_PPU_SEC	Code Secteur Privé/Public
PRS_FJH_TYP	Type de Prise en Charge Forfait Journalier
ETE_IND_TAA	Indicateur TAA Privé/Public
PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)
DRG_AFF_NAT	Nature du Destinataire de Règlement affiné
PRS_REM_TYP	Type de Remboursement
TOP_PS5_TRG	Top Périmètre hors CMU-C et prestations pour information
Organisme de prise en charge	
ORG_CLE_ZEAT	ZEAT de l'Organisme de Liquidation jusqu'à 2014
ORG_CLE_REG	Région de l'Organisme de Liquidation à partir de 2015
Bénéficiaire des soins	
BEN_SEX_COD	Sexe du Bénéficiaire
AGE_BEN_SNDS	Tranche d'Age Bénéficiaire au moment des soins
BEN_QLT_COD	Qualité du Bénéficiaire
BEN_RES_ZEAT	ZEAT de Résidence du Bénéficiaire jusqu'à 2014
BEN_RES_REG	Région de Résidence du Bénéficiaire à partir de 2015
BEN_CMU_TOP	Top Bénéficiaire CMU-C (Bénéficiaire CSS depuis avril 2021)

Professionnel de santé exécutant	
PSE_ACT_CAT	Catégorie de l'Exécutant
PSE_SPE_SNDS	Spécialité Médicale PS Exécutant
PSE_ACT_SNDS	Nature d'Activité PS Exécutant
EXE_INS_ZEAT	ZEAT du PS Exécutant jusqu'à 2014
EXE_INS_REG	Région du PS Exécutant à partir de 2015
PSE_STJ_SNDS	Statut Juridique PS Exécutant
MFT_COD	Mode de Fixation des Tarifs Etb Exécutant
ETE_ZEAT_COD	ZEAT d'Implantation Etb Exécutant jusqu'à 2014
ETE_REG_COD	Région d'Implantation Etb Exécutant à partir de 2015
ETE_TYP_SNDS	Type Etb Exécutant
ETE_CAT_SNDS	Catégorie Etb Exécutant
DDP_SPE_COD	Discipline de Prestation Etb Exécutant
MDT_TYP_COD	Mode de Traitement Etb Exécutant
Professionnel de santé prescripteur	
PSP_ACT_CAT	Catégorie du Prescripteur
PSP_SPE_SNDS	Spécialité Médicale PS Prescripteur
PSP_ACT_SNDS	Nature d'Activité PS Prescripteur
PRE_INS_ZEAT	ZEAT du PS Prescripteur jusqu'à 2014
PRE_INS_REG	Région du PS Prescripteur à partir de 2015
PSP_STJ_SNDS	Statut Juridique PS Prescripteur
ETP_ZEAT_COD	ZEAT d'Implantation Etb Prescripteur jusqu'à 2014
ETP_REG_COD	Région d'Implantation Etb Prescripteur à partir de 2015
ETP_CAT_SNDS	Catégorie Etb Prescripteur
Indicateurs bruts	
PRS_ACT_COG	Coefficient Global
PRS_ACT_NBR	Dénombrement
PRS_ACT_QTE	Quantité
PRS_DEP_MNT	Montant du Dépassement
PRS_PAI_MNT	Montant de la Dépense
PRS_REM_MNT	Montant Versé/Remboursé
PRS_REM_BSE	Base de Remboursement
Indicateurs préfiltrés	
FLT_ACT_COG	Coefficient Global de la Prestation Préfiltré
FLT_ACT_NBR	Dénombrement de la Prestation Préfiltré
FLT_ACT_QTE	Quantité de la Prestation Préfiltrée
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltré
FLT_PAI_MNT	Montant de la Dépense de la Prestation Préfiltrée
FLT_REM_MNT	Montant Versé/Remboursé Préfiltré

TABLE A.1 : Tableau des variables de la base Open Damir en 2021

BEN_RES_REG	Libellé Région de Résidence du Bénéficiaire	BEN_RES_ZEAT	Libellé ZEAT de Résidence du Bénéficiaire
5	Régions et Départements d'outre-mer	0	Inconnu
11	Ile-de-France	1	Région Parisienne (11)
24	Centre-Val de Loire	2	Bassin Parisien (21,22,23,24,25,26)
27	Bourgogne-Franche-Comté	3	Nord (31)
28	Normandie	4	Est (41,42,43)
32	Hauts-de-France - Nord-Pas-de-Calais-Picardie	5	Ouest (52,53,54)
44	Grand Est	6	Sud-Ouest (72,73,74)
52	Pays de la Loire	7	Centre-Est (82,83)
53	Bretagne	8	Méditerrannée (91,93,94)
75	Aquitaine-Limousin-Poitou-Charentes	9	Régions et Départements d'outre-mer (01,02,03,04,06)
76	Languedoc-Roussillon-Midi-Pyrénées		
84	Auvergne-Rhône-Alpes		
93	Provence-Alpes-Côte d'Azur et Corse		
99	Inconnu		

TABLE A.2 : Différence dans la codification des régions avant et après 2015

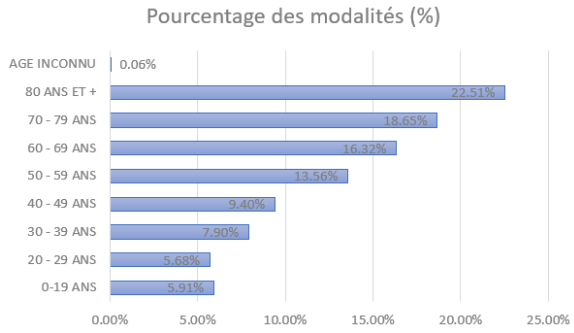
AGE_BEN_SNDS	Libellé Tranche d'Age Bénéficiaire au moment des soins
0	0-19 ANS
20	20 - 29 ANS
30	30 - 39 ANS
40	40 - 49 ANS
50	50 - 59 ANS
60	60 - 69 ANS
70	70 - 79 ANS
80	80 ANS ET +
99	AGE INCONNU
BEN_RES_REG	Libellé Région de Résidence du Bénéficiaire
5	Régions et Départements d'outre-mer
11	Ile-de-France
24	Centre-Val de Loire
27	Bourgogne-Franche-Comté
28	Normandie
32	Hauts-de-France - Nord-Pas-de-Calais-Picardie
44	Grand Est
52	Pays de la Loire
53	Bretagne
75	Aquitaine-Limousin-Poitou-Charentes
76	Languedoc-Roussillon-Midi-Pyrénées
84	Auvergne-Rhône-Alpes
93	Provence-Alpes-Côte d'Azur et Corse
99	Inconnu
BEN_SEX_COD	Libellé Sexe du Bénéficiaire
0	INCONNU OU PERSONNE MORALE SANS CIVILITE
1	MASCULIN
2	FEMININ
9	VALEUR INCONNUE

TABLE A.3 : Modalités des variables d'âge, de sexe et de région de bénéficiaire de la base Open Damir

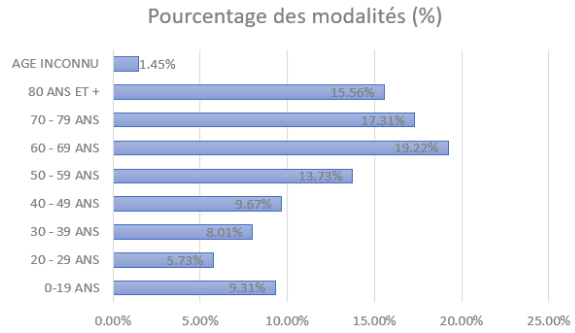
A.2 Traitement des données manquantes et analyse préliminaire

A.2.1 Traitement des valeurs manquantes

Lors de l'étude des variables principales de la base Damir, il apparaît que certaines lignes de la base correspondent à des prestations non identifiables, ce qui signifie que les informations liées au bénéficiaire ou au régime de remboursement sont inconnues. Une présentation en termes de pourcentage de dépenses totales et du nombre total d'actes occupés par ces prestations est utile pour décider de l'élimination des valeurs inconnues. Les variables indicatrices utilisées pour les statistiques globales de consommation (dépenses et nombre d'actes) seront des variables préfiltrées (FLT_PAI_MNT et FLT_ACT_QTE), car ces variables permettent de distinguer les lignes de prestations complémentaires pour un même soin et donc de compter le soin une seule fois. Nous présentons donc les graphiques des pourcentages du volume d'actes consommés et de dépenses prises par les différentes modalités de 4 variables : Tranche d'âge, Sexe, Région, Bénéficiaire CSS.

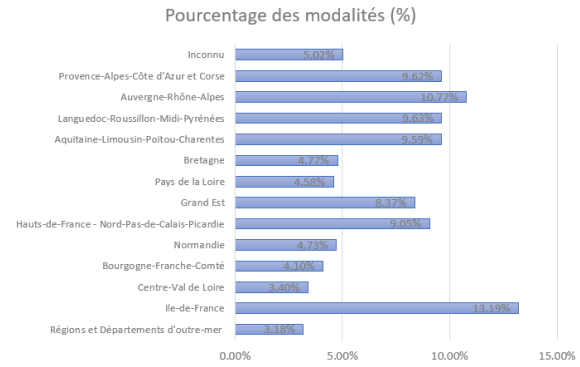


(a) Répartition des tranches d'âge de bénéficiaire en fonction des actes totaux

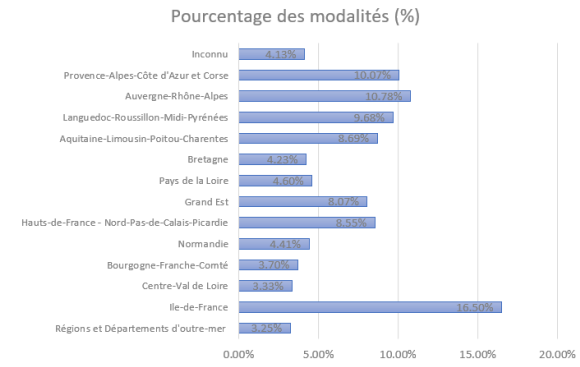


(b) Répartition des tranches d'âge de bénéficiaire en fonction des dépenses totaux

FIGURE A.1 : Proportion de consommation par âge de bénéficiaires du régime de la Sécurité Sociale



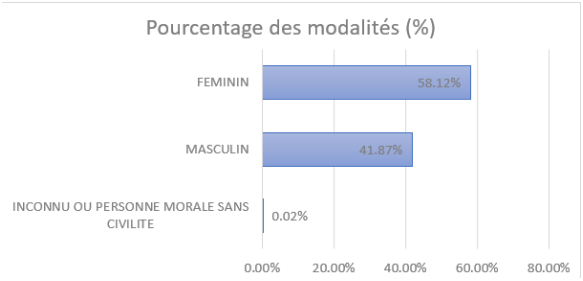
(a) Répartition des régions de bénéficiaire en fonction des actes totaux



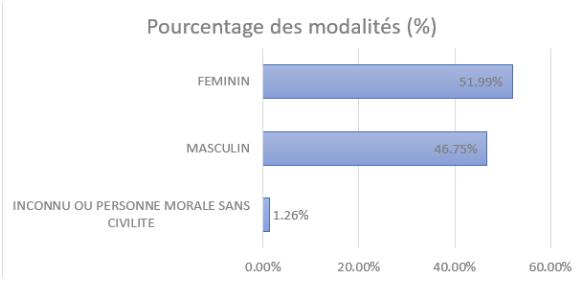
(b) Répartition des régions de bénéficiaire en fonction des dépenses totaux

FIGURE A.2 : Proportion de consommation par région de bénéficiaires du régime de la Sécurité Sociale

Par les trois figures A.1, A.2, A.3, les modalités inconnues des variables de tranche d'âge et de sexe représentent des parts très faibles, soit moins de 2% du volume d'actes consommés et de dépenses, tandis que pour la variable de région du bénéficiaire, ces proportions atteignent 5%. En ce qui concerne les variables de sexe et d'âge, il n'y a pas de difficulté à supprimer les valeurs inconnues. Cependant, la variable de région du bénéficiaire pose un problème, car le poids des valeurs inconnues est plus

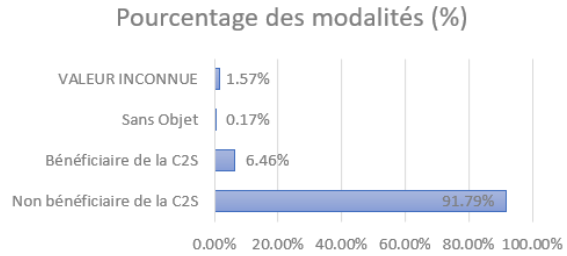


(a) Répartition des sexes de bénéficiaires en fonction des actes totaux

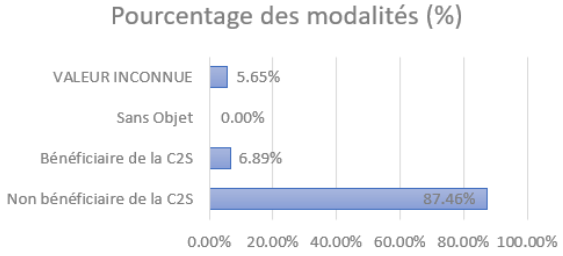


(b) Répartition des sexes de bénéficiaires en fonction des dépenses totaux

FIGURE A.3 : Proportion de consommation par sexe de bénéficiaires du régime de la Sécurité Sociale



(a) Répartition des bénéficiaires de la CSS en fonction des actes totaux



(b) Répartition des bénéficiaires de la CSS en fonction de dépenses totaux

FIGURE A.4 : Proportion de consommation de bénéficiaires CSS parmi tous les bénéficiaires de Sécurité Sociale

élevé ici. Malgré cela, il est préférable de les supprimer car elles ne sont pas exploitables en termes de consommation et de suivi des remboursements par région.

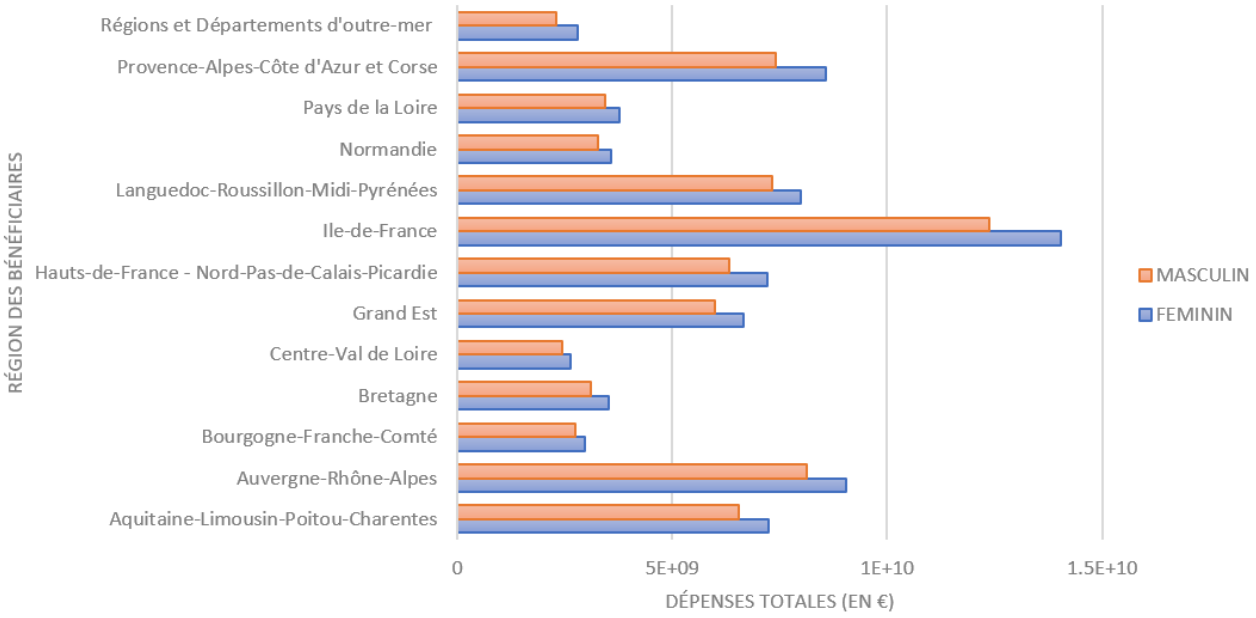
Pour la variable BEN_CMU_TOP (variable permettant de distinguer les bénéficiaires de la CSS ou du régime obligatoire seul), une partie importante de la dépense totale (5.65%) de modalités de valeur inconnue rend l'analyse de la consommation en modulant les flux de remboursement parmi les assurés de la complémentaire santé CSS ou privée plus difficile. En effet, toutes ces lignes aberrantes devraient être enlevées, ce qui entraîne une perte considérable de données dans la base.

Après cette étape, notre base ne comporte plus d'inconnus pour les trois variables importantes de segmentation de la population, ni pour les sous-catégories de la population dans les prestations payées. Cela s'applique aussi bien aux bénéficiaires de la couverture complémentaire santé CSS qu'à ceux relevant du régime obligatoire, y compris les personnes sans couverture CSS.

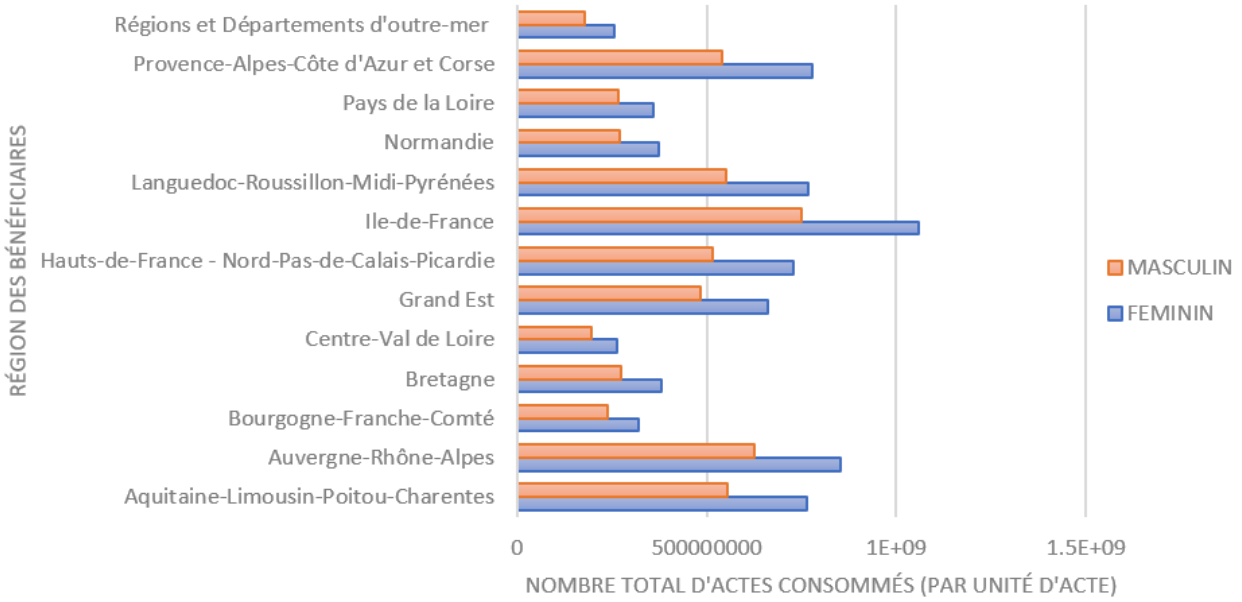
A.2.2 Analyse préliminaire de la consommation globale

Après avoir rejeté toutes les lignes de données manquantes, rendant ainsi les quatre variables ci-dessus non identifiables pour l'analyse, nous examinons désormais le volume d'actes médicaux consommés ainsi que les dépenses totales payées pour ces actes, pour l'ensemble des régimes, en fonction de deux variables différentes simultanément, à savoir la tranche d'âge par région et le sexe par région. La figure A.5 montre clairement une tendance à des dépenses plus élevées et à un plus grand nombre d'actes chez les femmes par rapport aux hommes dans la même région. De plus, la région Île-de-France est celle où les dépenses et le nombre d'actes consommés sont les plus élevés.

La figure A.6 montre clairement une corrélation de la consommation médicale avec l'âge des assurés : plus la tranche d'âge est grande, plus les assurés consomment.



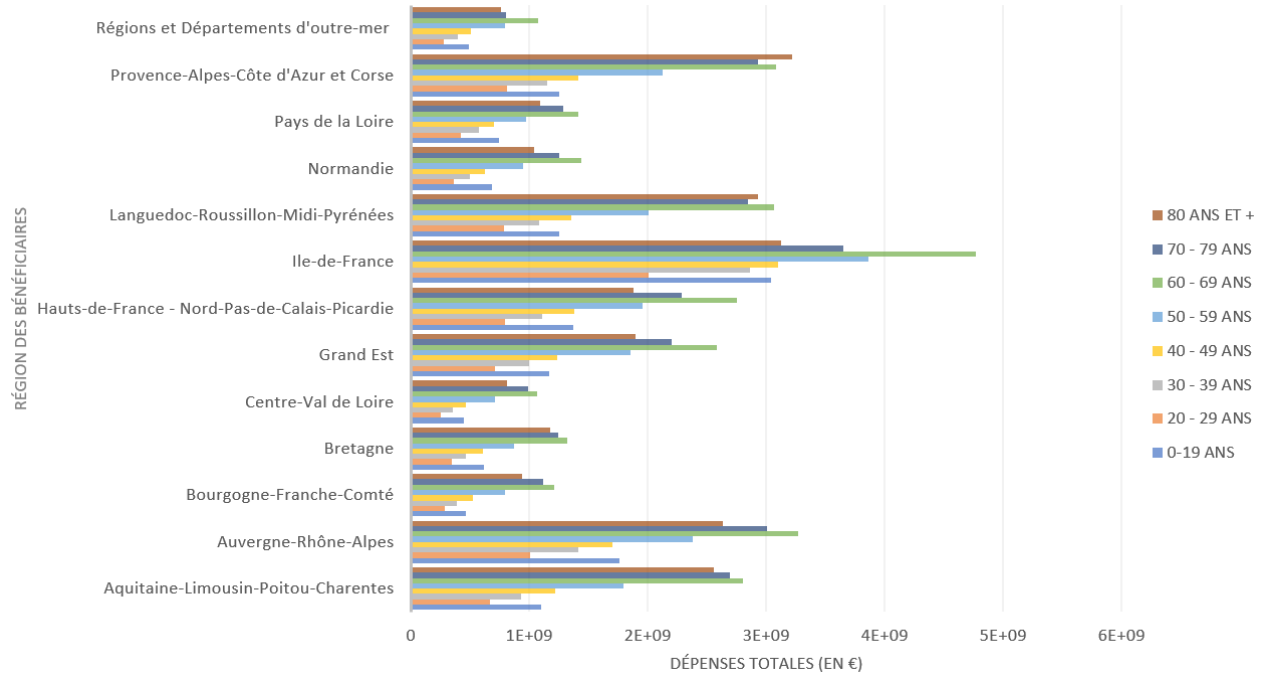
(a) Dépenses totales (en €)



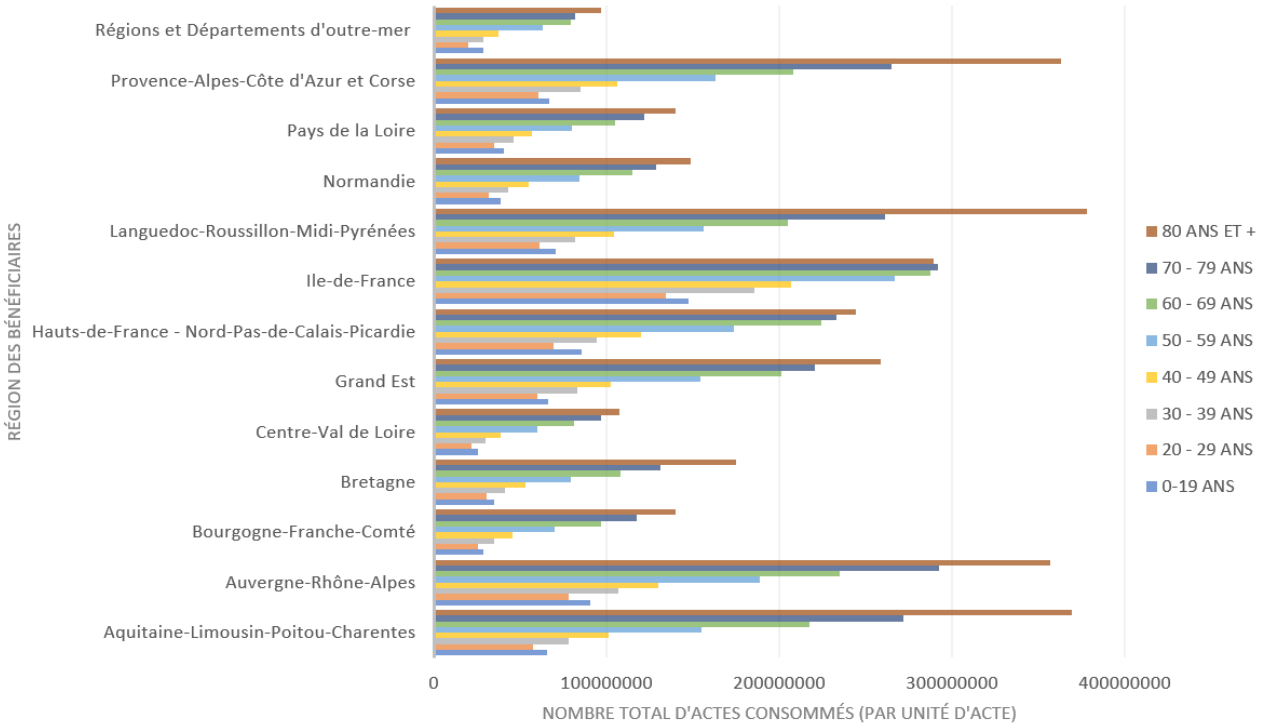
(b) Nombre total d'actes consommés (en unités d'acte)

FIGURE A.5 : Répartition de la consommation des bénéficiaires par région et par sexe

Pour mieux comprendre les bénéficiaires de soins, la variable BEN_QLT_COD devrait être regroupée en 4 modalités, car plusieurs modalités de cette variable sont minoritaires par rapport aux trois principales : Assuré, Enfant, Conjoint et assimilé. À partir de la figure A.7, la modalité "Conjoint et assimilé" peut être regroupée avec la modalité "Autre ayant-droit", ce qui laisse trois modalités : Assuré, Enfant, Autre ayant-droit. Il est remarquable que la consommation par les enfants représente une part importante de la consommation médicale des assurés de la sécurité sociale.



(a) Dépenses totales (en €)

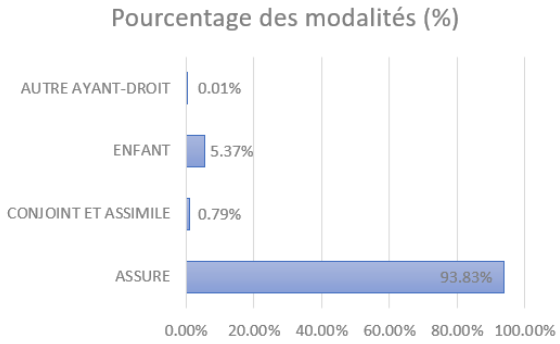


(b) Nombre total d'actes consommés (en unités d'acte)

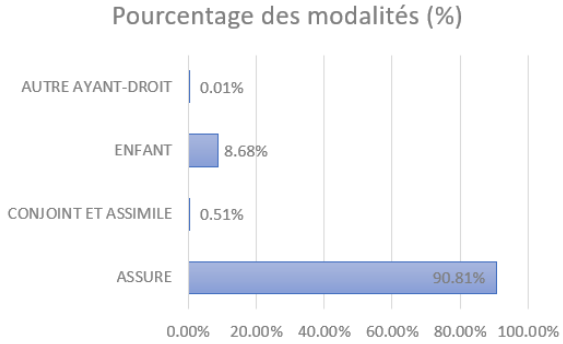
FIGURE A.6 : Répartition de la consommation des bénéficiaires par région et par tranche d'âge

A.2.3 Les lignes de prestation liées au régime local d'Alsace-Moselle

Il est nécessaire de comprendre le principe de séparation du remboursement par régime. Si une prestation a une partie complémentaire au régime général (CSS, le montant résultant de la différence



(a) Répartition des qualités des bénéficiaires en fonction des actes totaux



(b) Répartition des qualités des bénéficiaires en fonction des dépenses totaux

FIGURE A.7 : Proportion de consommation par qualité des bénéficiaires du régime de la Sécurité Sociale

de taux de remboursement entre le régime d'Alsace-Moselle et le régime général), alors les prestations complémentaires seront placées dans d'autres lignes. Elles auront la même dépense brute et la même quantité d'actes bruts que la ligne de prestation du régime général, mais le remboursement effectué et le taux de remboursement seront différents. De plus, tous les indicateurs préfiltrés seront nuls, servant ainsi de marqueur indiquant qu'il s'agit d'une ligne de prestation complémentaire.

Comme mentionné dans les parties précédentes, en plus des prestations du régime obligatoire enregistrées dans la base Damir, il existe également des lignes de prestations non classées en tant que complémentaire santé, provenant du régime local d'Alsace-Moselle, qui rembourse un supplément par rapport aux prestations habituelles du régime principal. Par exemple, pour une consultation chez un médecin généraliste de secteur 1 d'un montant de 25€, le régime obligatoire rembourse 17.5€, et ce remboursement est enregistré dans la base Damir comme un remboursement normal du régime obligatoire. Si l'assuré relève du régime local d'Alsace-Moselle, il recevra un remboursement supplémentaire de 20% de la base de remboursement (25€). Ce montant supplémentaire est également enregistré dans la base Damir, mais cette fois-ci dans une ligne distincte avec le même nom d'acte (table A.4). Cependant, par défaut, tous les indicateurs préfiltrés de cette ligne sont mis à zéro pour éviter un double comptage, tandis que les indicateurs bruts sont différents de zéro. On peut distinguer cette situation en particulier en se référant aux variables suivantes :

- Type de remboursement = "SUPPLEMENT ALSACE MOSELLE"
- Modulation du Ticket Modérateur = "REGIME LOCAL ALSACE MOSELLE"
- Type d'Enveloppe = "ALSACE MOSELLE"

Après avoir identifié la distinction entre cette prestation supplémentaire et les prestations du régime général, il convient de supprimer toutes ces lignes de prestations supplémentaires.

A.2.4 Les lignes de prestations liées au dispositif CSS

Le dispositif CSS, en remboursant les assurés dans la plupart des cas le ticket modérateur comme une complémentaire santé, est inscrit dans la base Damir sous forme de lignes de prestations complémentaires des actes de soins dont les indicateurs préfiltrés sont nuls. Comme il présente une forme très particulière dans la base Damir (le seul à générer des remboursements de la complémentaire santé dans la base Damir), il existe des variables utiles pour identifier ces prestations :

Variables de la base Damir	Exempe ligne 1	Exempe ligne 2
Année et mois de remboursement	2021-10	2021-10
Tranche d'âge du bénéficiare	30-39 ans	30-39 ans
Région de résidence du bénéficiaire	Grand Est	Grand Est
Sexe du bénéficiaire	Femme	Femme
Qualité du bénéficiaire	Assuré	Assuré
Année de soins	2021	2021
Mois de soins	9	9
Type d'enveloppe de prestation	ALSACE-MOSELLE	SOINS DE VILLE
Nature de prestation	CONSULTATION MEDECINE GENERALE	CONSULTATION MEDECINE GENERALE
Type de remboursement	SUPPLEMENT ALSACE-MOSELLE	PRESTATION DE REFERENCE
Base de remboursement	0 €	850 €
Taux de remboursement	20%	70%
Quantité d'acte brut	34	34
Montant de la dépense brut	850 €	850 €
Montant remboursé brut	170 €	595 €
Quantité d'acte préfiltré	0	34
Montant de la dépense préfiltré	0 €	850 €
Montant remboursé préfiltré	0 €	595 €

TABLE A.4 : Exemple de remboursement de régime local d'Alsace Moselle

- Type de remboursement = "TICKET MODERATEUR C2S","FORFAIT C2S"
- Top Bénéficiaire CMU-C = "Bénéficiaire de la C2S"
- Type d'Enveloppe = "C2S, AME ou ACS"
- Top Périmètre hors CMU-C et Prestations pour Information = "Type de Remboursement C2S/CMU C ou Prestations pour Information"

En rappelant le principe de remboursement d'une complémentaire santé, pour une dépense en consultation généraliste en secteur 1, le bénéficiaire de la CSS est remboursé de la totalité du ticket modérateur (7.5€), et cette prestation est enregistrée dans la modalité "TICKET MODERATEUR C2S" sur une ligne distincte de la ligne décrivant le remboursement de base du régime obligatoire (voir référence A.5). Afin de rendre le mémoire conforme aux critères définis précédemment, il est nécessaire de supprimer ces lignes de prestations complémentaires de la CSS.

A.2.5 Franchise et régulation dans la base Damir

Les participations forfaitaires et les régulations de remboursement (en cas d'erreur de remboursement de la Sécurité Sociale) sont ajoutées dans la base Damir après les lignes de prestations dites de référence. Pour une consultation généraliste, par exemple, la participation forfaitaire est déduite des remboursements aux assurés. Cependant, cette déduction ne s'effectue pas directement sur le montant remboursé, mais plutôt par l'ajout d'une ligne de montant remboursé négatif, traduisant le

Variables de la base Damir	Exempe ligne 1	Exempe ligne 2
Année et mois de remboursement	2021-08	2021-08
Tranche d'âge du bénéficiare	50-59 ans	50-59 ans
Région de résidence du bénéficiaire	Île de France	Île de France
Sexe du bénéficiaire	Homme	Homme
Qualité du bénéficiaire	Assuré	Assuré
Année de soins	2021	2021
Mois de soins	1	1
Bénéficiaire CSS	Oui	Oui
Nature de prestation	CONSULTATION MEDECINE GENERALE	CONSULTATION MEDECINE GENERALE
Type de remboursement	TICKET MODERATEUR CSS	PRESTATION DE REFERENCE
Base de remboursement	0 €	300 €
Taux de remboursement	30%	70%
Quantité d'acte brut	12	12
Montant de la dépense brut	300 €	300 €
Montant remboursé brut	90 €	210 €
Quantité d'acte préfiltré	0	12
Montant de la dépense préfiltré	0 €	300 €
Montant remboursé préfiltré	0 €	210 €

TABLE A.5 : Exemple de remboursement de Complémentaire Santé Solidaire

fait que la Sécurité Sociale n'a pas remboursé ces participations forfaitaires en réalité (illustré par la table A.6). Du point de vue des assureurs proposant des complémentaires santé, la franchise de 1€ n'est pas remboursée par les contrats responsables, il convient donc de ne conserver que les lignes de remboursement positif. Les régulations dans la base Damir se font par l'ajout de lignes de dépense brute/préfiltrée négatives ou de quantité d'actes brute/préfiltrée négatives. Toutefois, en raison de la difficulté à identifier l'ensemble de ces régulations, nous avons choisi de ne pas les prendre en compte.

A.2.6 Une tendance de tiers payant importante en 2021

Avec la base obtenue, nous avons réexaminé les destinations des remboursements. Après avoir calculé la somme des dépenses totales (total_dep) et du nombre d'actes totaux (total_act) pour chaque modalité de la variable DRG_AFF_NAT (Nature du Destinataire du Règlement Affiné), il est clair, comme le montre la figure A.8 que la prestation liée au tiers payant (modalité 35) représente le poids le plus important parmi toutes les modalités, tandis que le tiers payant intégral (modalité 36) constitue également une partie non négligeable de l'ensemble des prestations (de l'ordre de grandeur de la modalité "Assuré"). Il est important de noter que la modalité du tiers payant intégral (souvent interprétée comme le paiement du régime obligatoire plus la part des organismes complémentaires) est répertoriée dans la base Damir uniquement en tant qu'indicateur de destination du flux de remboursement, alors que le remboursement réel peut ne contenir que la partie du régime obligatoire. Par conséquent, aucune modification n'est nécessaire pour cette variable, car elle ne présente pas de difficultés pour les étapes ultérieures de l'analyse de la base Damir.

Variables de la base Damir	Exempe ligne 1	Exempe ligne 2
Année et mois de remboursement	2021-04	2021-04
Tranche d'âge du bénéficiare	50-59 ans	50-59 ans
Région de résidence du bénéficiaire	Auvergne-Rhône-Alpes	Auvergne-Rhône-Alpes
Sexe du bénéficiaire	Femme	Femme
Qualité du bénéficiaire	Assuré	Assuré
Année de soins	2021	2021
Mois de soins	2	2
Bénéficiaire CSS	Non	Non
Nature de prestation	CONSULTATION MEDECINE GENERALE	PARTICIPATION FORFAITAIRE HORS TIERS PAYANT
Type de remboursement	PRESTATION DE REFERENCE	PRESTATION DE REFERENCE
Base de remboursement	150 €	0 €
Taux de remboursement	70%	100%
Quantité d'acte brut	6	6
Montant de la dépense brut	150 €	0 €
Montant remboursé brut	105 €	-6 €
Quantité d'acte préfiltré	6	6
Montant de la dépense préfiltré	150 €	0 €
Montant remboursé préfiltré	105 €	-6 €

TABLE A.6 : Exemple de réduction des franchises sur les remboursements de la Sécurité Sociale

A.3 Primes pures de la tarification basée sur la base Open Damir pour l'année 2021

Tranche d'âge	**Région**	**Sexe**	**Prime de niveau de couverture minimum**	**Prime de niveau de couverture moyen**	**Prime de niveau de couverture maximum**
20-29 ans	Régions et Départements d'outre-mer	Homme	85.41 €	122.06 €	122.88 €
20-29 ans	Régions et Départements d'outre-mer	Femme	137.39 €	197.57 €	198.61 €
30-39 ans	Régions et Départements d'outre-mer	Homme	128.74 €	187.21 €	189.99 €
30-39 ans	Régions et Départements d'outre-mer	Femme	197.34 €	288.93 €	292.28 €

40-49 ans	Régions et Départements d'outre-mer	Homme	180.09 €	310.33 €	317.35 €
40-49 ans	Régions et Départements d'outre-mer	Femme	236.31 €	415.39 €	423.37 €
50-59 ans	Régions et Départements d'outre-mer	Homme	236.85 €	422.22 €	432.83 €
50-59 ans	Régions et Départements d'outre-mer	Femme	279.16 €	507.42 €	520.72 €
60-69 ans	Régions et Départements d'outre-mer	Homme	298.19 €	525.34 €	538.55 €
60-69 ans	Régions et Départements d'outre-mer	Femme	328.44 €	591.15 €	605.90 €
70-79ans	Régions et Départements d'outre-mer	Homme	389.93 €	680.31 €	692.95 €
70-79ans	Régions et Départements d'outre-mer	Femme	385.44 €	699.05 €	711.97 €
80+ ans	Régions et Départements d'outre-mer	Homme	418.03 €	849.35 €	858.64 €
80+ ans	Régions et Départements d'outre-mer	Femme	419.15 €	896.79 €	904.74 €
20-29 ans	Ile-de-France	Homme	124.99 €	258.37 €	264.93 €
20-29 ans	Ile-de-France	Femme	199.13 €	419.09 €	428.26 €
30-39 ans	Ile-de-France	Homme	165.94 €	344.76 €	358.31 €
30-39 ans	Ile-de-France	Femme	255.68 €	562.88 €	580.21 €
40-49 ans	Ile-de-France	Homme	209.02 €	468.04 €	488.20 €
40-49 ans	Ile-de-France	Femme	289.78 €	673.51 €	696.93 €
50-59 ans	Ile-de-France	Homme	285.76 €	634.62 €	663.73 €
50-59 ans	Ile-de-France	Femme	350.05 €	799.15 €	830.97 €
60-69 ans	Ile-de-France	Homme	357.38 €	727.06 €	760.82 €
60-69 ans	Ile-de-France	Femme	394.81 €	816.57 €	851.36 €
70-79ans	Ile-de-France	Homme	472.95 €	898.54 €	936.63 €
70-79ans	Ile-de-France	Femme	513.10 €	990.10 €	1,028.89 €
80+ ans	Ile-de-France	Homme	553.40 €	1,021.06 €	1,063.23 €
80+ ans	Ile-de-France	Femme	596.56 €	1,056.97 €	1,093.23 €
20-29 ans	Centre-Val de Loire	Homme	104.83 €	178.12 €	180.83 €
20-29 ans	Centre-Val de Loire	Femme	181.40 €	312.60 €	315.78 €
30-39 ans	Centre-Val de Loire	Homme	133.68 €	221.69 €	227.57 €

30-39 ans	Centre-Val de Loire	Femme	209.96 €	357.44 €	363.84 €
40-49 ans	Centre-Val de Loire	Homme	180.35 €	351.52 €	361.76 €
40-49 ans	Centre-Val de Loire	Femme	248.25 €	490.95 €	502.27 €
50-59 ans	Centre-Val de Loire	Homme	251.42 €	495.18 €	510.92 €
50-59 ans	Centre-Val de Loire	Femme	307.52 €	604.66 €	622.48 €
60-69 ans	Centre-Val de Loire	Homme	342.54 €	597.62 €	618.62 €
60-69 ans	Centre-Val de Loire	Femme	364.95 €	652.74 €	674.42 €
70-79ans	Centre-Val de Loire	Homme	473.13 €	760.23 €	786.84 €
70-79ans	Centre-Val de Loire	Femme	484.80 €	792.40 €	817.16 €
80+ ans	Centre-Val de Loire	Homme	519.59 €	795.55 €	822.97 €
80+ ans	Centre-Val de Loire	Femme	531.78 €	800.55 €	823.24 €
20-29 ans	Bourgogne-Franche-Comté	Homme	113.58 €	183.10 €	185.61 €
20-29 ans	Bourgogne-Franche-Comté	Femme	200.02 €	329.40 €	332.73 €
30-39 ans	Bourgogne-Franche-Comté	Homme	155.68 €	240.87 €	247.12 €
30-39 ans	Bourgogne-Franche-Comté	Femme	232.06 €	375.83 €	382.56 €
40-49 ans	Bourgogne-Franche-Comté	Homme	208.73 €	371.66 €	382.11 €
40-49 ans	Bourgogne-Franche-Comté	Femme	275.19 €	506.39 €	518.43 €
50-59 ans	Bourgogne-Franche-Comté	Homme	284.96 €	524.21 €	540.34 €
50-59 ans	Bourgogne-Franche-Comté	Femme	338.61 €	627.55 €	646.00 €
60-69 ans	Bourgogne-Franche-Comté	Homme	373.37 €	624.65 €	644.89 €
60-69 ans	Bourgogne-Franche-Comté	Femme	395.21 €	675.80 €	696.82 €
70-79ans	Bourgogne-Franche-Comté	Homme	500.40 €	784.18 €	810.27 €
70-79ans	Bourgogne-Franche-Comté	Femme	519.09 €	821.45 €	844.97 €
80+ ans	Bourgogne-Franche-Comté	Homme	551.38 €	834.07 €	860.40 €
80+ ans	Bourgogne-Franche-Comté	Femme	576.25 €	857.71 €	878.25 €
20-29 ans	Normandie	Homme	109.37 €	176.20 €	178.77 €
20-29 ans	Normandie	Femme	195.12 €	316.58 €	319.73 €
30-39 ans	Normandie	Homme	149.65 €	233.98 €	240.10 €
30-39 ans	Normandie	Femme	228.72 €	365.70 €	372.47 €
40-49 ans	Normandie	Homme	198.41 €	360.21 €	370.30 €
40-49 ans	Normandie	Femme	267.92 €	495.59 €	507.08 €
50-59 ans	Normandie	Homme	269.65 €	505.35 €	520.46 €
50-59 ans	Normandie	Femme	321.72 €	602.61 €	619.23 €

60-69 ans	Normandie	Homme	362.26 €	602.57 €	621.17 €
60-69 ans	Normandie	Femme	381.60 €	648.31 €	667.78 €
70-79ans	Normandie	Homme	495.74 €	761.75 €	785.10 €
70-79ans	Normandie	Femme	510.77 €	796.02 €	817.93 €
80+ ans	Normandie	Homme	540.56 €	806.54 €	830.04 €
80+ ans	Normandie	Femme	570.95 €	848.61 €	867.08 €
20-29 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	117.66 €	187.77 €	190.18 €
20-29 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	196.98 €	318.83 €	321.81 €
30-39 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	169.78 €	260.53 €	266.45 €
30-39 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	246.96 €	394.39 €	401.08 €
40-49 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	228.65 €	393.38 €	403.33 €
40-49 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	299.38 €	522.52 €	533.72 €
50-59 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	307.68 €	543.77 €	558.38 €
50-59 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	365.24 €	640.37 €	656.39 €
60-69 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	395.95 €	636.29 €	653.77 €
60-69 ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	425.39 €	685.50 €	702.57 €
70-79ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	546.18 €	823.16 €	845.51 €
70-79ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	574.62 €	869.43 €	889.00 €
80+ ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Homme	642.81 €	941.61 €	963.35 €
80+ ans	Hauts-de-France - Nord-Pas-de-Calais-Picardie	Femme	732.32 €	1,046.38 €	1,063.85 €

20-29 ans	Grand Est	Homme	119.72 €	195.93 €	198.84 €
20-29 ans	Grand Est	Femme	209.07 €	346.58 €	349.98 €
30-39 ans	Grand Est	Homme	168.75 €	271.64 €	279.49 €
30-39 ans	Grand Est	Femme	252.65 €	418.95 €	426.76 €
40-49 ans	Grand Est	Homme	226.09 €	410.37 €	424.14 €
40-49 ans	Grand Est	Femme	302.35 €	556.75 €	571.69 €
50-59 ans	Grand Est	Homme	311.56 €	586.29 €	607.20 €
50-59 ans	Grand Est	Femme	372.48 €	699.53 €	722.47 €
60-69 ans	Grand Est	Homme	413.88 €	711.44 €	736.89 €
60-69 ans	Grand Est	Femme	437.05 €	761.01 €	786.43 €
70-79ans	Grand Est	Homme	555.16 €	893.74 €	923.20 €
70-79ans	Grand Est	Femme	568.10 €	927.04 €	954.18 €
80+ ans	Grand Est	Homme	619.68 €	970.67 €	998.90 €
80+ ans	Grand Est	Femme	655.00 €	1,008.10 €	1,030.24 €
20-29 ans	Pays de la Loire	Homme	115.24 €	182.31 €	184.54 €
20-29 ans	Pays de la Loire	Femme	198.79 €	319.27 €	321.79 €
30-39 ans	Pays de la Loire	Homme	151.73 €	235.67 €	240.75 €
30-39 ans	Pays de la Loire	Femme	238.99 €	381.59 €	387.13 €
40-49 ans	Pays de la Loire	Homme	196.89 €	361.28 €	370.50 €
40-49 ans	Pays de la Loire	Femme	261.89 €	491.16 €	501.54 €
50-59 ans	Pays de la Loire	Homme	272.96 €	512.83 €	527.78 €
50-59 ans	Pays de la Loire	Femme	321.33 €	607.72 €	623.87 €
60-69 ans	Pays de la Loire	Homme	367.54 €	615.97 €	636.40 €
60-69 ans	Pays de la Loire	Femme	379.52 €	656.36 €	677.21 €
70-79ans	Pays de la Loire	Homme	507.89 €	782.77 €	809.70 €
70-79ans	Pays de la Loire	Femme	514.51 €	806.42 €	831.07 €
80+ ans	Pays de la Loire	Homme	559.35 €	814.96 €	842.37 €
80+ ans	Pays de la Loire	Femme	573.64 €	828.40 €	850.34 €
20-29 ans	Bretagne	Homme	111.56 €	173.97 €	176.41 €
20-29 ans	Bretagne	Femme	195.05 €	303.85 €	306.78 €
30-39 ans	Bretagne	Homme	147.53 €	224.25 €	230.04 €
30-39 ans	Bretagne	Femme	232.12 €	354.20 €	360.07 €
40-49 ans	Bretagne	Homme	192.36 €	342.14 €	352.23 €
40-49 ans	Bretagne	Femme	259.47 €	466.31 €	477.40 €
50-59 ans	Bretagne	Homme	256.94 €	472.43 €	487.68 €
50-59 ans	Bretagne	Femme	318.59 €	583.48 €	601.35 €
60-69 ans	Bretagne	Homme	342.94 €	570.24 €	590.61 €
60-69 ans	Bretagne	Femme	366.70 €	624.85 €	646.71 €
70-79ans	Bretagne	Homme	465.73 €	714.77 €	739.78 €
70-79ans	Bretagne	Femme	486.17 €	756.36 €	779.44 €
80+ ans	Bretagne	Homme	504.04 €	772.06 €	796.17 €
80+ ans	Bretagne	Femme	545.32 €	841.43 €	860.62 €
20-29 ans	Aquitaine-Limousin-Poitou-Charentes	Homme	129.40 €	202.56 €	205.25 €
20-29 ans	Aquitaine-Limousin-Poitou-Charentes	Femme	214.28 €	345.88 €	349.09 €

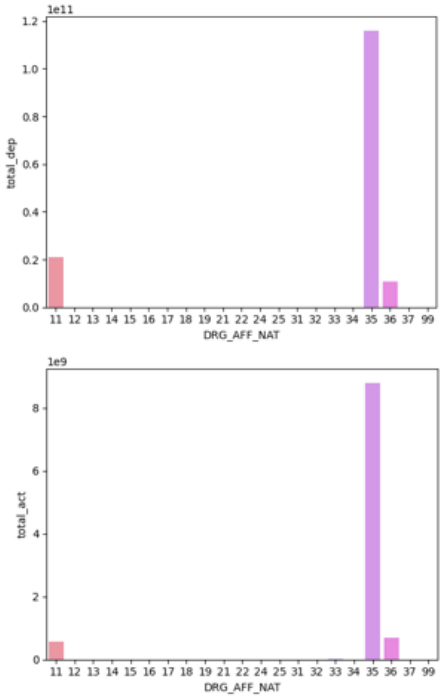
30-39 ans	Aquitaine-Limousin-Poitou-Charentes	Homme	163.94 €	254.54 €	260.59 €
30-39 ans	Aquitaine-Limousin-Poitou-Charentes	Femme	255.33 €	413.07 €	419.67 €
40-49 ans	Aquitaine-Limousin-Poitou-Charentes	Homme	213.01 €	382.10 €	393.19 €
40-49 ans	Aquitaine-Limousin-Poitou-Charentes	Femme	287.10 €	523.63 €	535.26 €
50-59 ans	Aquitaine-Limousin-Poitou-Charentes	Homme	287.56 €	533.78 €	550.42 €
50-59 ans	Aquitaine-Limousin-Poitou-Charentes	Femme	342.98 €	633.83 €	651.89 €
60-69 ans	Aquitaine-Limousin-Poitou-Charentes	Homme	387.83 €	654.71 €	677.52 €
60-69 ans	Aquitaine-Limousin-Poitou-Charentes	Femme	409.39 €	696.36 €	719.63 €
70-79ans	Aquitaine-Limousin-Poitou-Charentes	Homme	534.86 €	837.71 €	866.17 €
70-79ans	Aquitaine-Limousin-Poitou-Charentes	Femme	547.23 €	856.50 €	883.22 €
80+ ans	Aquitaine-Limousin-Poitou-Charentes	Homme	598.65 €	911.62 €	940.36 €
80+ ans	Aquitaine-Limousin-Poitou-Charentes	Femme	628.22 €	947.98 €	971.36 €
20-29 ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	136.76 €	205.21 €	208.28 €
20-29 ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	217.72 €	337.58 €	341.04 €
30-39 ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	178.37 €	267.53 €	274.49 €
30-39 ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	266.04 €	412.08 €	419.19 €
40-49 ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	235.17 €	396.60 €	409.44 €
40-49 ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	308.53 €	526.66 €	539.96 €
50-59 ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	323.07 €	557.60 €	576.99 €
50-59 ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	376.75 €	652.97 €	674.20 €

60-69 ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	424.02 €	670.08 €	694.57 €
60-69 ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	440.51 €	708.17 €	733.27 €
70-79ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	574.09 €	848.21 €	877.44 €
70-79ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	588.87 €	878.26 €	905.65 €
80+ ans	Languedoc-Roussillon-Midi-Pyrénées	Homme	662.91 €	989.48 €	1,017.76 €
80+ ans	Languedoc-Roussillon-Midi-Pyrénées	Femme	729.66 €	1,078.71 €	1,102.04 €
20-29 ans	Auvergne-Rhône-Alpes	Homme	125.89 €	222.58 €	226.66 €
20-29 ans	Auvergne-Rhône-Alpes	Femme	209.25 €	372.36 €	377.22 €
30-39 ans	Auvergne-Rhône-Alpes	Homme	161.66 €	280.60 €	289.31 €
30-39 ans	Auvergne-Rhône-Alpes	Femme	252.86 €	448.57 €	457.95 €
40-49 ans	Auvergne-Rhône-Alpes	Homme	209.49 €	416.61 €	431.55 €
40-49 ans	Auvergne-Rhône-Alpes	Femme	280.72 €	567.20 €	583.38 €
50-59 ans	Auvergne-Rhône-Alpes	Homme	290.79 €	594.11 €	617.27 €
50-59 ans	Auvergne-Rhône-Alpes	Femme	342.49 €	701.43 €	726.35 €
60-69 ans	Auvergne-Rhône-Alpes	Homme	384.81 €	707.55 €	736.34 €
60-69 ans	Auvergne-Rhône-Alpes	Femme	397.09 €	742.83 €	772.03 €
70-79ans	Auvergne-Rhône-Alpes	Homme	519.55 €	887.86 €	921.88 €
70-79ans	Auvergne-Rhône-Alpes	Femme	528.92 €	911.15 €	943.21 €
80+ ans	Auvergne-Rhône-Alpes	Homme	576.47 €	956.05 €	989.81 €
80+ ans	Auvergne-Rhône-Alpes	Femme	600.50 €	975.87 €	1,003.37 €

20-29 ans	Provence-Alpes-Côte d'Azur et Corse	Homme	159.17 €	246.27 €	249.97 €
20-29 ans	Provence-Alpes-Côte d'Azur et Corse	Femme	247.56 €	402.90 €	407.29 €
30-39 ans	Provence-Alpes-Côte d'Azur et Corse	Homme	203.41 €	314.98 €	323.22 €
30-39 ans	Provence-Alpes-Côte d'Azur et Corse	Femme	292.98 €	484.09 €	492.84 €
40-49 ans	Provence-Alpes-Côte d'Azur et Corse	Homme	256.96 €	442.05 €	456.68 €
40-49 ans	Provence-Alpes-Côte d'Azur et Corse	Femme	341.27 €	597.09 €	612.80 €
50-59 ans	Provence-Alpes-Côte d'Azur et Corse	Homme	350.09 €	613.71 €	636.62 €
50-59 ans	Provence-Alpes-Côte d'Azur et Corse	Femme	415.22 €	730.13 €	754.61 €
60-69 ans	Provence-Alpes-Côte d'Azur et Corse	Homme	437.83 €	722.13 €	749.53 €
60-69 ans	Provence-Alpes-Côte d'Azur et Corse	Femme	469.81 €	779.37 €	807.59 €
70-79ans	Provence-Alpes-Côte d'Azur et Corse	Homme	554.52 €	884.58 €	915.49 €
70-79ans	Provence-Alpes-Côte d'Azur et Corse	Femme	594.96 €	943.19 €	972.71 €
80+ ans	Provence-Alpes-Côte d'Azur et Corse	Homme	630.14 €	1,051.56 €	1,082.26 €
80+ ans	Provence-Alpes-Côte d'Azur et Corse	Femme	746.86 €	1,210.70 €	1,236.63 €

TABLE A.7 : Primes pures retenues de la tarification sur la base Open Damir pour l'année 2021

	DRG_AFF_NAT	total_dep	total_act
0	11	2.089565e+10	564564244
1	12	2.359684e+06	98831
2	13	4.825000e+02	14
3	14	5.275784e+04	1412
4	15	8.770010e+06	31514
5	16	7.206726e+07	763090
6	17	4.188444e+06	151188
7	18	2.176764e+05	6710
8	19	1.133345e+05	3617
9	21	2.867220e+03	82
10	22	1.393999e+04	325
11	24	1.579000e+02	6
12	25	3.423130e+03	75
13	31	1.800690e+05	2440
14	32	4.639122e+04	1173
15	33	0.000000e+00	7698973
16	34	4.946579e+04	915
17	35	1.160771e+11	8796736082
18	36	1.069547e+10	697911144
19	37	0.000000e+00	70
20	99	0.000000e+00	0



DRG_AFF_NAT	Libellé Nature du Destinataire du Règlement Affiné
0	SANS OBJET
11	Assuré
12	Adresse Provisoire
13	Conjoint divorcé
14	Conjoint séparé
15	Procuration
16	Tutelle
17	Veuvage
18	conjoint
19	Enfant
21	Autre ayant droit
22	concubin
23	Ascendant, descendant
24	Concubin en rupture
25	Bénéficaire Pacsé
26	Pacs fin
27	A-D- RESIDENCE UE
31	Correspondant d'entreprise
32	Mutuelle
33	Subrogation
34	Titre Médecin
35	Tiers Payant
36	Tiers Payant Intégral (RO+OC)
37	Subrogation multiple
38	TP RO + OC C25
99	VALEUR INCONNUE

FIGURE A.8 : Destination des remboursements comprenant la plupart des prestations en tiers payant (non intégral + intégral)

Annexe B

Théorie de l'utilité aléatoire - Modèles de choix discret

B.1 Expérience de choix discret

En synthétisant la littérature, notamment les travaux de Soekhai et al. (2019), Mangham et al. (2009), et Johnson et al. (2013), cette section décrit l'Expérience de Choix Discret (ECD) dans son ensemble. Une ECD est une méthode de recherche utilisée dans des domaines tels que l'économie de la santé et l'évaluation environnementale pour comprendre les processus de prise de décision des individus impliquant des choix complexes. Les chercheurs présentent aux participants des scénarios hypothétiques avec différentes alternatives caractérisées par des attributs et des niveaux spécifiques. Ces attributs sont les catégories distinctes de caractéristiques qui définissent chaque alternative, tandis que les niveaux représentent les différentes options au sein de ces attributs. De plus, les ECD intègrent souvent un attribut de prix avec des niveaux de prix variables pour évaluer la disposition des participants à payer pour certaines caractéristiques. La sélection appropriée des attributs et des niveaux est cruciale pour obtenir des résultats de qualité, car cela influence directement la pertinence des informations extraites des choix des participants.

B.2 Étude de type Préférences Révélées ou Préférences Déclarées

Cette section est basée sur la compréhension des travaux suivants : Carson et al. (2001), Abdullah et al. (2011), Ben-Akiva et al. (1994), Chen (1995). Dans l'application des modèles de choix discrets, on retrouve deux approches principales : la Méthode des Préférences Révélées et la Méthode des Préférences Déclarées.

Méthode des Préférences Révélées (PR) : Les méthodes des Préférences Révélées impliquent l'observation et l'analyse des comportements et choix réels des individus dans les marchés ou situations existants. Ces méthodes infèrent les préférences des individus en se basant sur leurs actions et décisions concrètes au sein de ces marchés. Les méthodes RP capturent les relations existantes entre les différentes caractéristiques des biens et services ainsi que les choix effectués par les individus en réponse aux changements de ces caractéristiques. Cela permet d'estimer les valeurs économiques sans demander directement aux individus leurs préférences.

Avantages des Méthodes PR :

- Réalisme du Marché : Les méthodes PR sont ancrées dans les transactions et comportements réels du marché, offrant des aperçus sur la manière dont les individus prennent réellement des décisions et font des compromis.

- Choix Réels : Les méthodes PR fournissent des données basées sur les décisions effectives des individus, établissant un lien plus direct avec leurs préférences.

- Consistance Comportementale : Les méthodes PR prennent naturellement en compte les contraintes et compromis auxquels sont confrontés les individus dans leurs choix, conduisant à des résultats cohérents et réalistes.

Méthode des Préférences Déclarées (PD) : Les méthodes des Préférences Déclarées impliquent de présenter aux individus des scénarios hypothétiques et de leur demander d'exprimer leurs préférences ou de faire des choix dans ces scénarios. Ces méthodes se basent sur les réponses déclarées par les individus plutôt que sur leurs comportements observés. Les méthodes PD sont particulièrement utiles lorsque les données de marché sont indisponibles ou pour étudier des scénarios qui ne se sont pas encore produits. Elles sont souvent utilisées pour valoriser des biens et services qui n'ont pas de transactions directes sur le marché, tels que les aménités environnementales.

Avantages des Méthodes PD :

- Flexibilité : Les méthodes PD permettent aux chercheurs de concevoir et de manipuler des scénarios pour explorer un large éventail de situations hypothétiques, facilitant la valorisation de biens sans marché existant.

- Valeurs Hors-Marché : Les méthodes PD sont essentielles pour valoriser les valeurs de non-utilisation (par exemple, valeur d'existence, valeur de legs) qui ne laissent pas de traces observables sur le marché.

- Analyse des Politiques : Les méthodes PD peuvent être utilisées pour estimer l'impact potentiel des changements de politique ou des nouvelles initiatives environnementales, fournissant des aperçus sur les futures prises de décision.

Comparaison et utilisation complémentaire : Les méthodes PR offrent des aperçus sur les comportements et choix réels au sein des marchés existants, ce qui les rend adaptées lorsque les données de marché sont disponibles. En revanche, les méthodes PD permettent aux chercheurs de capturer les valeurs associées à des scénarios hypothétiques ou hors-marché. Les chercheurs utilisent souvent à la fois les méthodes PR et PD conjointement pour obtenir une compréhension plus complète des préférences des individus et estimer la valeur économique de divers biens et services.

L'intégration de données provenant de sources multiples constitue une approche prometteuse pour enrichir et améliorer la validité des modèles de choix discret. Cette méthodologie trouve son origine dans l'article Ben-Akiva et al. (1994), où les auteurs explorent les avantages et les défis associés à la combinaison de données de préférences révélées et déclarées dans le contexte des modèles de choix. L'objectif principal de cette approche est d'exploiter les atouts spécifiques de chaque type de données tout en surmontant leurs limites respectives. L'article met en lumière plusieurs avantages, tels que l'efficacité accrue des estimations, la correction des biais inhérents aux données PD, l'identification de préférences pour de nouvelles alternatives et attributs, ainsi que la possibilité de fusionner des variables issues de différentes sources. En adoptant des modèles de choix discrets, les auteurs démontrent comment les paramètres de préférence peuvent être simultanément estimés en utilisant des données PR et PD.

B.3 Modèle Nested Logit et GEV

À partir du paradoxe de Bus rouge et bleu, la famille de modèles Nested Logit est née afin de relâcher cette caractéristique IANP du modèle logit tout en conservant la simplicité du modèle Logit

Multinomial. Cependant, il n'est pas présenté dans ce mémoire, car son application n'est pas pertinente pour notre étude. Son principe est décrit ci-dessous pour une compréhension générale :

Prenons le problème de trois choix de transports (Voiture, Bus rouge, Bus bleu). Le modèle Nested Logit suppose que les deux choix de Bus rouge ou Bus bleu sont maintenant regroupés dans le même panier de "Transport en commun", tandis que l'option Voiture est placée dans "Transport privé". Les individus sont ramenés au problème du choix entre Transport en commun ou privé, puis en fonction de leur choix, ils auront accès aux options contenues dans les paniers de transport (voir figure B.1).

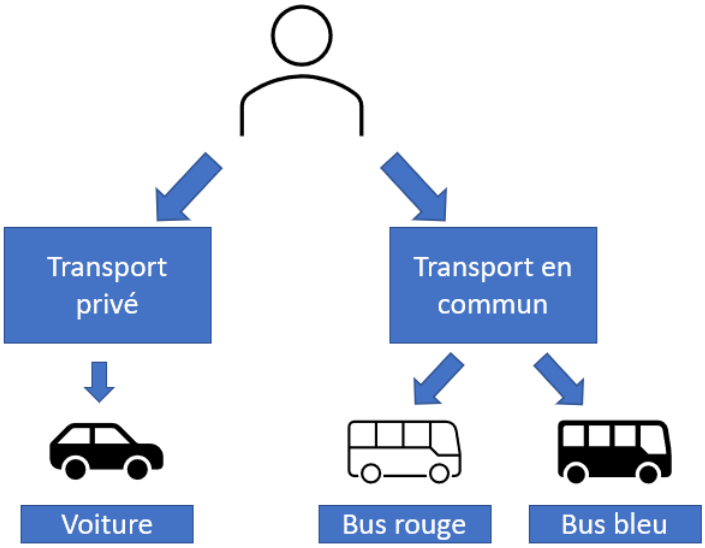


FIGURE B.1 : Modèle Nested Logit sur les choix de transport

Il est clairement visible que le modèle Logit est un cas particulier du Nested Logit lorsque le choix de panier se réduit à un panier. En dehors du modèle Nested Logit, il existe également un modèle plus général appelé GEV (Generalized Extreme Values), généré à partir d'une fonction génératrice G, qui permet au modélisateur de spécifier la structure spécifique de la distribution des valeurs extrêmes dans le terme aléatoire d'utilité considéré.

B.4 Modèle Probit Multinomial

Quant au modèle MNL, qui suppose que les termes aléatoires dans l'utilité sont i.i.d (indépendants et identiquement distribués) selon la loi de Gumbel, le fait de supposer que ces termes aléatoires suivent la loi normale conduit à un autre modèle appelé modèle probit multinomial. On suppose que le vecteur aléatoire $\epsilon_{\mathbf{n}}$ suit une loi normale multidimensionnelle avec un vecteur d'espérance nulle et une matrice de covariance $\mathbf{\Omega}$, et sa densité est définie par :

$$\phi\left(\epsilon_{\mathbf{n}}\right) = \frac{1}{(2\pi)^{\frac{I_n}{2}}\left|\mathbf{\Omega}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\epsilon_{\mathbf{n}}^{\mathbf{T}}\mathbf{\Omega}^{-\mathbf{1}}\epsilon_{\mathbf{n}}}.$$

Dans ce modèle, les termes aléatoires ne nécessitent pas d'être indépendants et identiquement distribués entre les alternatives, ce qui permet de relâcher l'hypothèse restrictive IANP du modèle MNL. De plus, il s'adapte de manière plus flexible aux données, car la structure de covariance Ω est estimée lors de la procédure d'estimation des autres paramètres dans V_n. Son seul compromis réside dans la difficulté d'approximation numérique de la fonction de répartition, étant donné que celle de la

loi normale n'est pas explicite. Un exemple de modèle à quatre choix est illustré par Train (2009), sur lequel les lecteurs pourraient s'appuyer. Nous illustrons ici le cas de de Probit bivarié.

Soit un modèle de choix avec $C = \{1, 2\}$, le vecteur d'erreur $\epsilon_{\mathbf{n}} = (\epsilon_{n1}, \epsilon_{n2})$ est i.i.d. pour tout n avec une moyenne nulle et une matrice variance-covariance $\mathbf{\Omega} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$.

Nous notons $\epsilon'_n = \epsilon_{n2} - \epsilon_{n1}$. En vertu de l'hypothèse du vecteur gaussien, nous pouvons déduire que ϵ'_n suit une loi normale avec une moyenne nulle et une variance $\sigma' = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$.

Nous pouvons donc calculer les probabilités de choix selon la définition de la théorie de l'utilité aléatoire, par exemple pour l'option 1 :

$$P_n(1|\mathcal{C}_n) = P(\epsilon_{n2} - \epsilon_{n1} \leq V_{n1} - V_{n2}) = P(\epsilon'_n \leq V_{n1} - V_{n2}).$$

Comme ϵ'_n suit une loi normale et que les utilités déterministes V_{n1} et V_{n2} sont constantes, on peut en déduire la formule de probabilité de choix comme suit :

$$P_n(1|\mathcal{C}_n) \int_{\epsilon=-\infty}^{V_{n1}-V_{n2}} \frac{1}{\sigma'\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\epsilon}{\sigma'})^2} d\epsilon.$$

Par un changement de variable $u = \frac{\epsilon}{\sigma'}$, nous obtenons :

$$P_n(1|\mathcal{C}_n) \int_{u=-\infty}^{\frac{V_{n1}-V_{n2}}{\sigma'}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \Phi(\frac{V_{n1} - V_{n2}}{\sigma'}),$$

où $\Phi()$ est la fonction de répartition d'une loi normale centrée réduite.

Si nous spécialisons les utilités déterministes $V_{ni,i\in C_n}$ comme pour le modèle Logit Multinomial, on peut écrire la vraisemblance comme :

$$\mathcal{L}(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},\mathbf{2}]}) = \prod_{n=1}^{N} P_n(1|\mathcal{C}_n)^{y_{n1}} P_n(2|\mathcal{C}_n)^{y_{n2}} = \prod_{n=1}^{N} \Phi(\frac{V_{n1} - V_{n2}}{\sigma'})^{y_{n1}} \Phi(\frac{V_{n2} - V_{n1}}{\sigma'})^{y_{n2}}$$

$$= \prod_{n=1}^{N} \Phi(\frac{\beta_{10} + \sum_{k=1}^{K_1} \alpha_{1k} S_{n1k} - (\beta_{20} + \sum_{k=1}^{K_2} \alpha_{2k} S_{n2k})}{\sigma'})^{y_{n1}} \Phi(\frac{\beta_{20} + \sum_{k=1}^{K_2} \alpha_{2k} S_{n2k} - (\beta_{10} + \sum_{k=1}^{K_1} \alpha_{1k} S_{n1k})}{\sigma'})^{y_{n2}}.$$

La log-vraisemblance s'écrit :

$$l(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]}) = ln(\mathcal{L}(\beta, \alpha_{\mathbf{i},\mathbf{i}\in[\mathbf{1},..,\mathbf{I}]})) = \sum_{n=1}^{N} (y_{n1} ln(\Phi(\frac{V_{n1} - V_{n2}}{\sigma'})) + y_{n2} ln(\Phi(\frac{V_{n2} - V_{n1}}{\sigma'})))$$

$$= \sum_{n=1}^{N} (y_{n1} ln(\Phi(\frac{\beta_{10} + \sum_{k=1}^{K_1} \alpha_{1k} S_{n1k} - (\beta_{20} + \sum_{k=1}^{K_2} \alpha_{2k} S_{n2k})}{\sigma'}))$$

$$+ y_{n2} ln(\Phi(\frac{\beta_{20} + \sum_{k=1}^{K_2} \alpha_{2k} S_{n2k} - (\beta_{10} + \sum_{k=1}^{K_1} \alpha_{1k} S_{n1k})}{\sigma'}))).$$

B.5 Test de deux modèles non emboîtés

Si nous disposons de deux modèles, $\mathcal{M}_1$ et $\mathcal{M}_2$, ayant des spécifications différentes, nous pouvons utiliser le test de Cox. Pour mieux comprendre ce test, considérons deux modèles dont l'utilité diffère partiellement, mais partage une partie déterministe commune, notée $V\prime_{ni}$ (fonction des mêmes variables et du même type de paramètres), comme suit :

$$\forall i \in \{1,..,I_n\}, U^1_{ni} = V^1_{ni} + \epsilon^1_{ni} = V\prime_{ni} + \sum_{k^1=1}^{K^1_i} \alpha^1_{ik^1} S^1_{nik^1} + \epsilon^1_{ni},$$

$$U^2_{ni} = V^2_{ni} + \epsilon^2_{ni} = V\prime_{ni} + \sum_{k^2=1}^{K^2_i} \alpha^2_{ik^2} S^2_{nik^2} + \epsilon^2_{ni}.$$

Le test de Cox est basé sur le test de modèle emboîté, plus précisément, nous devons tester les deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ par rapport au modèle combiné $\mathcal{M}_C$, où chaque modèle testé est une version avec des contraintes linéaires sur les paramètres du modèle mère $\mathcal{M}_C$.

$$\mathcal{M}_C : \forall i \in \{1,..,I_n\}, U^C_{ni} = V^C_{ni} + \epsilon^C_{ni} = V\prime_{ni} + \sum_{k^1=1}^{K^1_i} \alpha^1_{ik^1} S^1_{nik^1} + \sum_{k^2=1}^{K^2_i} \alpha^2_{ik^2} S^2_{nik^2} + \epsilon^C_{ni},$$

où :

- $\mathcal{M}_1$ est la version contrainte de $\mathcal{M}_C$ avec $\alpha^1_{ik^1} = 0, \forall i, k^1$.
- $\mathcal{M}_2$ est la version contrainte de $\mathcal{M}_C$ avec $\alpha^2_{ik^2} = 0, \forall i, k^2$.

Le tableau suivant présente le test de Cox pour deux modèles non emboîtés, impliquant deux tests de rapport de vraisemblance, ainsi que les conclusions correspondantes dans la table B.1.

$\mathcal{H}^1_0 : \mathcal{M} = \mathcal{M}_1$ contre $\mathcal{H}^1_1 : \mathcal{M} = \mathcal{M}_C$	$\mathcal{H}^2_0 : \mathcal{M} = \mathcal{M}_2$ contre $\mathcal{H}^2_1 : \mathcal{M} = \mathcal{M}_C$	Conclusion
$\mathcal{H}^1_0$ n'est pas rejetée	$\mathcal{H}^2_0$ est rejetée	Préférence pour le modèle 1
$\mathcal{H}^1_0$ est rejetée	$\mathcal{H}^2_0$ n'est pas rejetée	Préférence pour le modèle 2
$\mathcal{H}^1_0$ est rejetée	$\mathcal{H}^2_0$ est rejetée	Aucun des modèles n'est préféré, nécessité de développer un nouveau modèle
$\mathcal{H}^1_0$ n'est pas rejetée	$\mathcal{H}^2_0$ n'est pas rejetée	Aucune conclusion, recours à un autre test nécessaire

TABLE B.1 : Test de Cox pour deux modèles non emboîtés : deux tests de rapport de vraisemblance sont effectués

Lorsque le test de Cox ne permet pas de conclure quel modèle est préférable, il peut être utile d'utiliser le test de Davidson et MacKinnon J. comme décrit dans DAVIDSON et MACKINNON (1981)

B.6 Pandas Biogeme sur Python

D'après BIERLAIRE (2020), le package Biogeme est conçu pour estimer les paramètres de différents modèles en utilisant la méthode de maximisation de la vraisemblance. Il est particulièrement adapté aux modèles de choix discrets, offrant de nombreuses possibilités de spécification du modèle, y compris les modèles logit à loi mélangée et les modèles de classes latentes. PandasBiogeme est un véritable package Python écrit en Python et en C++, qui repose sur la bibliothèque Pandas pour la gestion des données. L'installation et la documentation sont disponibles sur le site : Site du package Biogeme.

Annexe C

Résultats des modèles entraînés - Désegmentation tarifaire

Soit {minimum, moyen, maximum} l'ensemble des niveaux de couverture de notre portefeuille, nous supposons que les modalités minimum, moyen et maximum correspondent respectivement aux niveaux 1, 2 et 3.

C.1 Résultat d'estimation sur le base 1

Nom	Logit_ASR	Logit_ASR_S	Logit_ASR_P	Logit_ASR_PS	MLogit_VS_ASR
Nombre de paramètres	46	48	47	49	48
Log-vraissemblance	-46804.4975	-46804.2686	-46804.4978	-46804.2710	-46804.3155
AIC	93700.9949	93702.5373	93704.9957	93706.5419	93704.6310
BIC	94106.7047	94117.0668	94128.3450	94138.7110	94127.9804
ρ^2	0.1479	0.1481	0.1605	0.1492	0.1784
$\bar{\rho}^2$	0.1471	0.1473	0.1596	0.1483	0.1776

TABLE C.1 : Evaluation des modèles sur la base 1

Modèle $\mathcal{H}_0$	Modèle $\mathcal{H}_1$	Modèle retenu
Logit_ASR	Logit_ASR_S	Logit_ASR
Logit_ASR	Logit_ASR_P	Logit_ASR
Logit_ASR	Logit_ASR_PS	Logit_ASR
Logit_ASR_S	Logit_ASR_PS	Logit_ASR_S
Logit_ASR_P	Logit_ASR_PS	Logit_ASR_P
Logit_ASR	MLogit_VS_ASR	Logit_ASR

TABLE C.2 : Test de rapport de vraissemblance des modèles estimés sur la base 1 au niveau de 5%

Nom du coefficient	Logit_ASR	Logit_ASR_P	Logit_ASR_S	Logit_ASR_PS	MLogit_ASR
Constant de niveau de couverture 2	0.000	0.000	0.000	0.000	0.000

Constant de niveau de couverture 3	-0.559	-0.586	-0.559	-0.586	-0.559
Constant de niveau de couverture 1	-0.598	-0.485	-0.598	-0.485	-0.598
Tranche d'âge 20-29 ans niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 20-29 ans niveau de couverture 3	-0.660	-0.649	-0.660	-0.649	-0.660
Tranche d'âge 20-29 ans niveau de couverture 1	0.787	0.799	0.787	0.799	0.787
Tranche d'âge 30-39 ans niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 30-39 ans niveau de couverture 3	-0.459	-0.453	-0.459	-0.453	-0.459
Tranche d'âge 30-39 ans niveau de couverture 1	0.631	0.642	0.631	0.642	0.632
Tranche d'âge 40-49 ans niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 40-49 ans niveau de couverture 3	-0.472	-0.468	-0.472	-0.468	-0.472
Tranche d'âge 40-49 ans niveau de couverture 1	0.458	0.497	0.458	0.497	0.458
Tranche d'âge 50-59 ans niveau de couverture 3	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 50-59 ans niveau de couverture 1	0.554	0.594	0.554	0.594	0.554
Tranche d'âge 50-59 ans niveau de couverture 2	0.414	0.412	0.414	0.412	0.414
Tranche d'âge 60-69 ans niveau de couverture 3	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 60-69 ans niveau de couverture 1	-0.109	-0.089	-0.109	-0.089	-0.110

Tranche d'âge 60-69 ans niveau de couverture	0.110	0.108	0.110	0.108	0.110
Tranche d'âge 70-79 ans niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 70-79 ans niveau de couverture 3	0.777	0.776	0.777	0.776	0.777
Tranche d'âge 70-79 ans niveau de couverture 2	0.372	0.368	0.372	0.369	0.372
Tranche d'âge 80+ niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
Tranche d'âge 80+ niveau de couverture 3	1.220	1.230	1.220	1.230	1.220
Tranche d'âge 80+ niveau de couverture 2	0.545	0.551	0.545	0.550	0.545
Sexe femme niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
Sexe femme niveau de couverture 3	-0.065	-0.099	-0.065	-0.099	-0.065
Sexe femme niveau de couverture 2	0.393	0.351	0.393	0.351	0.393
Sexe homme niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
Sexe homme niveau de couverture 3	-0.493	-0.487	-0.493	-0.487	-0.493
Sexe homme niveau de couverture 1	-0.270	-0.233	-0.270	-0.233	-0.270
R11 niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
R11 niveau de couverture 3	0.028	-0.031	0.028	-0.031	0.028
R11 niveau de couverture 2	0.101	0.043	0.101	0.043	0.101
R24 niveau de couverture 3	0.000	0.000	0.000	0.000	0.000
R24 niveau de couverture 1	0.080	0.093	0.080	0.093	0.080
R24 niveau de couverture 2	0.107	0.103	0.107	0.103	0.107
R27 niveau de couverture 2	0.000	0.000	0.000	0.000	0.000

R27 niveau de couverture 3	-0.146	-0.142	-0.146	-0.142	-0.146
R27 niveau de couverture 1	-0.116	-0.113	-0.116	-0.113	-0.116
R28 niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
R28 niveau de couverture 3	-0.033	-0.031	-0.033	-0.031	-0.033
R28 niveau de couverture 2	0.084	0.081	0.084	0.081	0.084
R32 niveau de couverture 3	0.000	0.000	0.000	0.000	0.000
R32 niveau de couverture 1	0.156	0.144	0.156	0.144	0.156
R32 niveau de couverture 2	0.125	0.119	0.125	0.119	0.125
R44 niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
R44 niveau de couverture 3	-0.102	-0.099	-0.102	-0.099	-0.102
R44 niveau de couverture 1	-0.039	-0.031	-0.039	-0.031	-0.039
R52 niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
R52 niveau de couverture 3	-0.200	-0.198	-0.200	-0.198	-0.200
R52 niveau de couverture 2	0.026	0.023	0.026	0.023	0.026
R53 niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
R53 niveau de couverture 3	-0.102	-0.099	-0.102	-0.099	-0.102
R53 niveau de couverture 1	0.008	0.006	0.008	0.006	0.008
R5 niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
R5 niveau de couverture 3	-0.033	-0.027	-0.033	-0.027	-0.033
R5 niveau de couverture 1	0.050	0.043	0.050	0.043	0.050
R75 niveau de couverture 3	0.000	0.000	0.000	0.000	0.000
R75 niveau de couverture 1	-0.009	-0.011	-0.009	-0.011	-0.009
R75 niveau de couverture 2	0.023	0.019	0.023	0.019	0.023
R76 niveau de couverture 1	0.000	0.000	0.000	0.000	0.000

R76 niveau de couverture 3	0.177	0.192	0.177	0.192	0.177
R76 niveau de couverture 2	0.137	0.149	0.137	0.149	0.137
R84 niveau de couverture 3	0.000	0.000	0.000	0.000	0.000
R84 niveau de couverture 1	-0.137	-0.112	-0.137	-0.112	-0.137
R84 niveau de couverture 2	-0.001	-0.003	-0.001	-0.003	-0.001
R93 niveau de couverture 1	0.000	0.000	0.000	0.000	0.000
R93 niveau de couverture 3	0.196	0.200	0.196	0.200	0.196
R93 niveau de couverture 2	0.074	0.076	0.074	0.076	0.074
Plafond	x	0.437	x	0.436	x
Finance	x	x	0.000	0.000	x
Lambda	x	x	0.565	0.772	x
Variance du niveau de couverture 3	x	x	x	x	0.000
Variance du niveau de couverture 1	x	x	x	x	0.009
Variance du niveau de couverture 2	x	x	x	x	0.019

TABLE C.3 : Tableau des coefficents des modèles entraînés sur la base 1

C.2 Résultat d'estimation sur la base 2

C.2.1 Modèle Logit Multinômial

Nom du modèle	Logit_ASR	Logit_ASR_S	Logit_ASR_P	Logit_ASR_PS
Nombre de paramètres	46	48	47	49
Log vraisemblance	-47821.8661	-47820.6305	-47802.5165	-47794.7224
AIC	95735.7322	95735.2611	95701.0330	95687.4449
BIC	96141.4420	96149.7907	96124.3823	96119.6140
ρ^2	0.1294	0.1461	0.1354	0.1405
$\bar{\rho}^2$	0.1286	0.1452	0.1345	0.1396

TABLE C.4 : Évaluation des modèles logit multinomial sur la base 2

Nom du coefficient	Logit_ASR	Logit_ASR_P	Logit_ASR_S	Logit_ASR_PS	Vrai modèle utilisé pour simuler la base 2
Constant de niveau de couverture 2	0.000	0.000	0.000	0.000	0.000
Constant de niveau de couverture 3	-0.696	-0.755	-0.532	-0.652	-0.568
Constant de niveau de couverture 1	-0.414	-0.171	-0.884	-0.297	0.019
Tranche d'âge 20-29 ans niveau de couverture 2	0.000	0.000	0.000	0.000	0.046
Tranche d'âge 20-29 ans niveau de couverture 3	-0.767	-0.743	-0.887	-0.855	-0.789
Tranche d'âge 20-29 ans niveau de couverture 1	0.531	0.558	0.894	1.120	0.742
Tranche d'âge 30-39 ans niveau de couverture 2	0.000	0.000	0.000	0.000	0.037
Tranche d'âge 30-39 ans niveau de couverture 3	-0.404	-0.391	-0.510	-0.504	-0.498
Tranche d'âge 30-39 ans niveau de couverture 1	0.401	0.425	0.671	0.867	0.461
Tranche d'âge 40-49 ans niveau de couverture 2	0.000	0.000	0.000	0.000	0.031
Tranche d'âge 40-49 ans niveau de couverture 3	-0.287	-0.277	-0.377	-0.379	-0.345
Tranche d'âge 40-49 ans niveau de couverture 1	0.316	0.402	0.366	0.666	0.314
Tranche d'âge 50-59 ans niveau de couverture 3	0.000	0.000	0.000	0.000	-0.304
Tranche d'âge 50-59 ans niveau de couverture 1	0.550	0.638	0.475	0.769	0.150
Tranche d'âge 50-59 ans niveau de couverture 2	0.375	0.370	0.446	0.455	0.155
Tranche d'âge 60-69 ans niveau de couverture 3	0.000	0.000	0.000	0.000	0.043

Tranche d'âge 60-69 ans niveau de couverture 1	-0.004	0.041	-0.097	0.037	-0.203
Tranche d'âge 60-69 ans niveau de couverture 2	0.048	0.045	0.107	0.118	0.161
Tranche d'âge 70-79 ans niveau de couverture 1	0.000	0.000	0.000	0.000	-0.609
Tranche d'âge 70-79 ans niveau de couverture 3	0.656	0.653	0.857	0.942	0.556
Tranche d'âge 70-79 ans niveau de couverture 2	0.137	0.128	0.385	0.467	0.053
Tranche d'âge 80+ niveau de couverture 1	0.000	0.000	0.000	0.000	-0.836
Tranche d'âge 80+ niveau de couverture 3	1.070	1.100	1.320	1.530	0.769
Tranche d'âge 80+ niveau de couverture 2	0.342	0.355	0.635	0.826	0.066
Sexe femme niveau de couverture 1	0.000	0.000	0.000	0.000	-0.046
Sexe femme niveau de couverture 3	-0.176	-0.248	0.013	-0.127	-0.238
Sexe femme niveau de couverture 2	0.277	0.186	0.495	0.315	0.284
Sexe homme niveau de couverture 2	0.000	0.000	0.000	0.000	0.265
Sexe homme niveau de couverture 3	-0.521	-0.507	-0.545	-0.524	-0.330
Sexe homme niveau de couverture 1	-0.313	-0.232	-0.376	-0.110	0.065
R11 niveau de couverture 1	0.000	0.000	0.000	0.000	0.113
R11 niveau de couverture 3	-0.099	-0.224	0.249	-0.131	-0.024
R11 niveau de couverture 2	-0.174	-0.298	0.152	-0.220	-0.089
R24 niveau de couverture 3	0.000	0.000	0.000	0.000	-0.192
R24 niveau de couverture 1	0.249	0.276	0.281	0.366	0.138
R24 niveau de couverture 2	0.222	0.214	0.245	0.225	0.054

R27 niveau de couverture 2	0.000	0.000	0.000	0.000	0.059
R27 niveau de couverture 3	-0.078	-0.069	-0.099	-0.076	0.006
R27 niveau de couverture 1	-0.241	-0.234	-0.219	-0.214	-0.065
R28 niveau de couverture 1	0.000	0.000	0.000	0.000	-0.002
R28 niveau de couverture 3	0.046	0.050	-0.021	-0.019	-0.128
R28 niveau de couverture 2	0.184	0.178	0.144	0.121	0.130
R32 niveau de couverture 3	0.000	0.000	0.000	0.000	-0.056
R32 niveau de couverture 1	0.181	0.154	0.237	0.158	0.047
R32 niveau de couverture 2	0.066	0.052	0.093	0.057	0.010
R44 niveau de couverture 2	0.000	0.000	0.000	0.000	0.040
R44 niveau de couverture 3	0.010	0.017	0.003	0.026	0.000
R44 niveau de couverture 1	-0.051	-0.033	-0.116	-0.108	-0.040
R52 niveau de couverture 1	0.000	0.000	0.000	0.000	0.083
R52 niveau de couverture 3	-0.300	-0.296	-0.359	-0.353	-0.209
R52 niveau de couverture 2	0.082	0.076	0.047	0.029	0.126
R53 niveau de couverture 2	0.000	0.000	0.000	0.000	0.073
R53 niveau de couverture 3	-0.049	-0.042	-0.072	-0.057	-0.074
R53 niveau de couverture 1	-0.047	-0.051	0.027	0.028	0.001
R5 niveau de couverture 2	0.000	0.000	0.000	0.000	0.008
R5 niveau de couverture 3	-0.046	-0.035	-0.070	-0.047	-0.123
R5 niveau de couverture 1	-0.055	-0.066	0.054	0.052	0.115
R75 niveau de couverture 3	0.000	0.000	0.000	0.000	-0.038
R75 niveau de couverture 1	-0.115	-0.119	-0.103	-0.131	-0.107
R75 niveau de couverture 2	0.091	0.083	0.106	0.084	0.145

R76 niveau de couverture 1	0.000	0.000	0.000	0.000	-0.079
R76 niveau de couverture 3	0.171	0.203	0.131	0.236	0.031
R76 niveau de couverture 2	0.112	0.138	0.084	0.169	0.047
R84 niveau de couverture 3	0.000	0.000	0.000	0.000	0.006
R84 niveau de couverture 1	0.014	0.068	-0.124	-0.006	0.011
R84 niveau de couverture 2	-0.037	-0.040	-0.038	-0.051	-0.017
R93 niveau de couverture 1	0.000	0.000	0.000	0.000	-0.195
R93 niveau de couverture 3	0.320	0.327	0.403	0.470	0.232
R93 niveau de couverture 2	0.008	0.011	0.090	0.139	-0.037
Plafond	x	0.945	x	3.130	2.300
Finance	x	x	0.028	1.030	0.086
Lambda	x	x	1.000	0.680	0.890

TABLE C.6 : Tableau des coefficients des modèles entraînés Logit Multinomial sur la base 2 et le vrai modèle pour la simulation de la base 2

C.2.2 Les autres modèles plus avancés

Les modèles plus complexes abordés dans le chapitre 5 sont décrits ici.

Nom du coefficient	MLogit_RC_ASR_PS	MLogit_VS_ASR_PS
Constant de niveau de couverture 2	0.000	0.000
Constant de niveau de couverture 3	-0.652	-0.641
Constant de niveau de couverture 1	-0.295	-0.397
Tranche d'âge 20-29 ans niveau de couverture 2	0.000	0.000
Tranche d'âge 20-29 ans niveau de couverture 3	-0.855	-0.851
Tranche d'âge 20-29 ans niveau de couverture 1	1.130	1.010
Tranche d'âge 30-39 ans niveau de couverture 2	0.000	0.000
Tranche d'âge 30-39 ans niveau de couverture 3	-0.504	-0.497
Tranche d'âge 30-39 ans niveau de couverture 1	0.869	0.771

Tranche d'âge 40-49 ans niveau de couverture 2	0.000	0.000
Tranche d'âge 40-49 ans niveau de couverture 3	-0.379	-0.369
Tranche d'âge 40-49 ans niveau de couverture 1	0.669	0.571
Tranche d'âge 50-59 ans niveau de couverture 3	0.000	0.000
Tranche d'âge 50-59 ans niveau de couverture 1	0.772	0.672
Tranche d'âge 50-59 ans niveau de couverture 2	0.455	0.444
Tranche d'âge 60-69 ans niveau de couverture 3	0.000	0.000
Tranche d'âge 60-69 ans niveau de couverture 1	0.039	-0.004
Tranche d'âge 60-69 ans niveau de couverture 2	0.119	0.109
Tranche d'âge 70-79 ans niveau de couverture 1	0.000	0.000
Tranche d'âge 70-79 ans niveau de couverture 3	0.943	0.885
Tranche d'âge 70-79 ans niveau de couverture 2	0.468	0.408
Tranche d'âge 80+ niveau de couverture 1	0.000	0.000
Tranche d'âge 80+ niveau de couverture 3	1.530	1.410
Tranche d'âge 80+ niveau de couverture 2	0.829	0.713
Sexe femme niveau de couverture 1	0.000	0.000
Sexe femme niveau de couverture 3	-0.128	-0.121
Sexe femme niveau de couverture 2	0.315	0.326
Sexe homme niveau de couverture 2	0.000	0.000
Sexe homme niveau de couverture 3	-0.524	-0.521
Sexe homme niveau de couverture 1	-0.108	-0.191
R11 niveau de couverture 1	0.000	0.000
R11 niveau de couverture 3	-0.135	-0.009
R11 niveau de couverture 2	-0.223	-0.104
R24 niveau de couverture 3	0.000	0.000
R24 niveau de couverture 1	0.367	0.343
R24 niveau de couverture 2	0.225	0.227
R27 niveau de couverture 2	0.000	0.000
R27 niveau de couverture 3	-0.076	-0.080
R27 niveau de couverture 1	-0.215	-0.207
R28 niveau de couverture 1	0.000	0.000
R28 niveau de couverture 3	-0.019	-0.020
R28 niveau de couverture 2	0.121	0.126
R32 niveau de couverture 3	0.000	0.000
R32 niveau de couverture 1	0.157	0.181

R32 niveau de couverture 2	0.057	0.064
R44 niveau de couverture 2	0.000	0.000
R44 niveau de couverture 3	0.026	0.018
R44 niveau de couverture 1	-0.108	-0.090
R52 niveau de couverture 1	0.000	0.000
R52 niveau de couverture 3	-0.353	-0.358
R52 niveau de couverture 2	0.029	0.031
R53 niveau de couverture 2	0.000	0.000
R53 niveau de couverture 3	-0.057	-0.059
R53 niveau de couverture 1	0.028	0.025
R5 niveau de couverture 2	0.000	0.000
R5 niveau de couverture 3	-0.047	-0.051
R5 niveau de couverture 1	0.052	0.044
R75 niveau de couverture 3	0.000	0.000
R75 niveau de couverture 1	-0.131	-0.113
R75 niveau de couverture 2	0.083	0.089
R76 niveau de couverture 1	0.000	0.000
R76 niveau de couverture 3	0.237	0.205
R76 niveau de couverture 2	0.170	0.142
R84 niveau de couverture 3	0.000	0.000
R84 niveau de couverture 1	-0.005	-0.021
R84 niveau de couverture 2	-0.051	-0.047
R93 niveau de couverture 1	0.000	0.000
R93 niveau de couverture 3	0.471	0.436
R93 niveau de couverture 2	0.140	0.110
Plafond	3.150	2.210
Finance	x	0.073
Lambda	0.674	0.921
Paramètre moyen de la loi du coefficient aléatoire	0.095	x
Paramètre variance de la loi du coefficient aléatoire	-0.012	x
Variance du niveau de couverture 1	x	-0.019
Variance du niveau de couverture 2	x	0.006
Variance du niveau de couverture 3	x	0.000

TABLE C.10 : Tableau des coefficients des modèles Logit Multinomial à loi mélange continue entraînés sur la base 2

Nom du coefficient détaillé	LCLogit_ASR_PS_L (Probabilité de membre logit)	LCLogit_ASR_PS_P (Probabilité de membre Probit)
Constant de niveau de couverture 3 - classe latente 1	-0.504	-0.832

Constant de niveau de couverture 3 - classe latente 2	0.135	0.145
Constant de niveau de couverture 1 - classe latente 2	0.000	0.000
Constant de niveau de couverture 1 - classe latente 1	-0.138	0.035
Constant de niveau de couverture 2 - classe latente 1	0.000	0.000
Constant de niveau de couverture 2 - classe latente 2	1.220	1.010
Finance - classe latente 1	0.624	0.123
Finance - classe latente 2	0.334	0.542
Plafond - classe latente 1	0.936	2.030
Plafond - classe latente 2	1.010	0.967
Tranche d'âge 20-29 ans - coefficient de probabilité de membre	-0.179	0.323
Tranche d'âge 20-29 ans niveau de couverture 3 - classe latente 1	-0.501	-0.114
Tranche d'âge 20-29 ans niveau de couverture 2 - classe latente 1	0.000	0.000
Tranche d'âge 20-29 ans niveau de couverture 1 - classe latente 1	0.421	1.080
Tranche d'âge 20-29 ans niveau de couverture 1 - classe latente 2	1.680	1.590
Tranche d'âge 20-29 ans niveau de couverture 3 - classe latente 2	0.000	0.000
Tranche d'âge 20-29 ans niveau de couverture 2 - classe latente 2	0.489	1.380
Tranche d'âge 30-39 ans - coefficient de probabilité de membre	0.082	0.144
Tranche d'âge 30-39 ans niveau de couverture 3 - classe latente 1	-0.478	-0.960
Tranche d'âge 30-39 ans niveau de couverture 2 - classe latente 1	0.000	0.000
Tranche d'âge 30-39 ans niveau de couverture 1 - classe latente 1	0.388	0.987
Tranche d'âge 30-39 ans niveau de couverture 1 - classe latente 2	1.170	-0.732
Tranche d'âge 30-39 ans niveau de couverture 2 - classe latente 2	0.120	-0.084
Tranche d'âge 30-39 ans niveau de couverture 3 - classe latente 2	0.000	0.000
Tranche d'âge 40-49 ans - coefficient de probabilité de membre	-0.033	0.081
Tranche d'âge 40-49 ans niveau de couverture 3 - classe latente 1	-0.277	-0.106
Tranche d'âge 40-49 ans niveau de couverture 3 - classe latente 2	0.018	-0.113

Tranche d'âge 40-49 ans niveau de couverture 2 - classe latente 1	0.000	0.000
Tranche d'âge 40-49 ans niveau de couverture 2 - classe latente 2	0.000	0.000
Tranche d'âge 40-49 ans niveau de couverture 1 - classe latente 1	0.301	0.603
Tranche d'âge 40-49 ans niveau de couverture 1 - classe latente 2	0.576	0.113
Tranche d'âge 50-59 ans - coefficient de probabilité de membre	0.069	0.055
Tranche d'âge 50-59 ans niveau de couverture 3 - classe latente 2	-0.050	0.155
Tranche d'âge 50-59 ans niveau de couverture 1 - classe latente 1	0.091	0.776
Tranche d'âge 50-59 ans niveau de couverture 1 - classe latente 2	0.218	0.154
Tranche d'âge 50-59 ans niveau de couverture 2 - classe latente 2	0.000	0.000
Tranche d'âge 50-59 ans niveau de couverture 3 - classe latente 1	0.000	0.000
Tranche d'âge 50-59 ans niveau de couverture 2 - classe latente 1	0.191	0.691
Tranche d'âge 60-69 ans - coefficient de probabilité de membre	0.179	-0.204
Tranche d'âge 60-69 ans niveau de couverture 3 - classe latente 2	0.512	-0.210
Tranche d'âge 60-69 ans niveau de couverture 1 - classe latente 1	-0.290	-0.479
Tranche d'âge 60-69 ans niveau de couverture 1 - classe latente 2	0.000	0.000
Tranche d'âge 60-69 ans niveau de couverture 3 - classe latente 1	0.000	0.000
Tranche d'âge 60-69 ans niveau de couverture 2 - classe latente 1	0.234	0.114
Tranche d'âge 60-69 ans niveau de couverture 2 - classe latente 2	0.219	-0.553
Tranche d'âge 70-79 ans - coefficient de probabilité de membre	0.449	-0.303
Tranche d'âge 70-79 ans niveau de couverture 3 - classe latente 1	0.483	-0.695
Tranche d'âge 70-79 ans niveau de couverture 3 - classe latente 2	1.310	1.370
Tranche d'âge 70-79 ans niveau de couverture 2 - classe latente 1	0.041	0.711
Tranche d'âge 70-79 ans niveau de couverture 2 - classe latente 2	0.436	-0.023
Tranche d'âge 70-79 ans niveau de couverture 1 - classe latente 1	0.000	0.000

Tranche d'âge 70-79 ans niveau de couverture 1 - classe latente 2	0.000	0.000
Tranche d'âge 80+ - coefficient de probabilité de membre	0.734	0.002
Tranche d'âge 80+ niveau de couverture 3 - classe latente 1	0.495	2.140
Tranche d'âge 80+ niveau de couverture 3 - classe latente 2	1.800	1.100
Tranche d'âge 80+ niveau de couverture 2 - classe latente 1	0.029	0.770
Tranche d'âge 80+ niveau de couverture 2 - classe latente 2	0.714	0.591
Tranche d'âge 80+ niveau de couverture 1 - classe latente 1	0.000	0.000
Tranche d'âge 80+ niveau de couverture 1 - classe latente 2	0.000	0.000
Sexe femme - coefficient de probabilité de membre	0.584	-0.115
Sexe femme niveau de couverture 3 - classe latente 1	0.121	-0.360
Sexe femme niveau de couverture 3 - classe latente 2	0.453	0.280
Sexe femme niveau de couverture 1 - classe latente 1	-0.319	-0.049
Sexe femme niveau de couverture 2 - classe latente 1	0.000	0.000
Sexe femme niveau de couverture 1 - classe latente 2	0.000	0.000
Sexe femme niveau de couverture 2 - classe latente 2	0.799	0.844
Sexe homme - coefficient de probabilité de membre	0.717	0.213
Sexe homme niveau de couverture 3 - classe latente 1	-0.625	-0.472
Sexe homme niveau de couverture 1 - classe latente 2	-0.101	-0.028
Sexe homme niveau de couverture 2 - classe latente 1	0.444	0.388
Sexe homme niveau de couverture 3 - classe latente 2	0.000	0.000
Sexe homme niveau de couverture 1 - classe latente 1	0.000	0.000
Sexe homme niveau de couverture 2 - classe latente 2	0.419	0.162
Coefficient constant de probabilité de membre	1.300	0.098
R11 - coefficient de probabilité de membre	-1.100	0.000
R11 niveau de couverture 3 - classe latente 1	-0.106	0.289

R11 niveau de couverture 3 - classe latente 2	1.070	-0.234
R11 niveau de couverture 1 - classe latente 2	0.000	0.000
R11 niveau de couverture 1 - classe latente 1	0.426	0.035
R11 niveau de couverture 2 - classe latente 1	0.000	0.000
R11 niveau de couverture 2 - classe latente 2	0.965	-0.170
R24 - coefficient de probabilité de membre	-0.257	0.222
R24 niveau de couverture 3 - classe latente 1	-0.056	-0.069
R24 niveau de couverture 3 - classe latente 2	-0.334	-0.363
R24 niveau de couverture 2 - classe latente 1	0.049	0.106
R24 niveau de couverture 1 - classe latente 1	0.000	0.000
R24 niveau de couverture 1 - classe latente 2	0.000	0.000
R24 niveau de couverture 2 - classe latente 2	-0.096	-0.144
R27 - coefficient de probabilité de membre	0.119	0.214
R27 niveau de couverture 3 - classe latente 1	-0.019	-0.155
R27 niveau de couverture 3 - classe latente 2	0.015	0.330
R27 niveau de couverture 1 - classe latente 1	-0.073	-0.412
R27 niveau de couverture 2 - classe latente 2	0.107	0.292
R27 niveau de couverture 2 - classe latente 1	0.000	0.000
R27 niveau de couverture 1 - classe latente 2	0.000	0.000
R28 - coefficient de probabilité de membre	1.770	-0.095
R28 niveau de couverture 3 - classe latente 1	-0.039	0.320
R28 niveau de couverture 1 - classe latente 2	0.539	-0.781
R28 niveau de couverture 2 - classe latente 1	0.117	-0.463
R28 niveau de couverture 2 - classe latente 2	0.163	0.873
R28 niveau de couverture 1 - classe latente 1	0.000	0.000
R28 niveau de couverture 3 - classe latente 2	0.000	0.000
R32 - coefficient de probabilité de membre	-0.500	0.180
R32 niveau de couverture 3 - classe latente 1	-0.085	-0.513
R32 niveau de couverture 3 - classe latente 2	-0.095	0.078
R32 niveau de couverture 1 - classe latente 1	0.050	-0.095
R32 niveau de couverture 2 - classe latente 2	-0.014	-0.181
R32 niveau de couverture 2 - classe latente 1	0.000	0.000
R32 niveau de couverture 1 - classe latente 2	0.000	0.000
R44 - coefficient de probabilité de membre	-0.929	-0.132
R44 niveau de couverture 3 - classe latente 1	-0.047	0.391
R44 niveau de couverture 1 - classe latente 2	-1.140	0.340
R44 niveau de couverture 2 - classe latente 1	0.057	0.292
R44 niveau de couverture 2 - classe latente 2	-0.021	0.076
R44 niveau de couverture 1 - classe latente 1	0.000	0.000
R44 niveau de couverture 3 - classe latente 2	0.000	0.000
R52 - coefficient de probabilité de membre	0.709	0.030
R52 niveau de couverture 3 - classe latente 1	-0.091	-0.552
R52 niveau de couverture 3 - classe latente 2	-0.741	-0.117
R52 niveau de couverture 1 - classe latente 1	-0.036	-0.060
R52 niveau de couverture 2 - classe latente 2	-0.336	0.237
R52 niveau de couverture 2 - classe latente 1	0.000	0.000
R52 niveau de couverture 1 - classe latente 2	0.000	0.000

R53 - coefficient de probabilité de membre	0.279	-0.024
R53 niveau de couverture 3 - classe latente 1	-0.035	-0.050
R53 niveau de couverture 1 - classe latente 2	0.355	-0.119
R53 niveau de couverture 2 - classe latente 1	0.084	-0.122
R53 niveau de couverture 2 - classe latente 2	0.071	0.200
R53 niveau de couverture 3 - classe latente 2	0.000	0.000
R53 niveau de couverture 1 - classe latente 1	0.000	0.000
R5 - coefficient de probabilité de membre	1.840	-0.312
R5 niveau de couverture 3 - classe latente 1	-0.035	-0.812
R5 niveau de couverture 3 - classe latente 2	-0.069	-0.173
R5 niveau de couverture 1 - classe latente 2	0.584	-0.647
R5 niveau de couverture 2 - classe latente 1	0.072	-0.631
R5 niveau de couverture 1 - classe latente 1	0.000	0.000
R5 niveau de couverture 2 - classe latente 2	0.000	0.000
R75 - coefficient de probabilité de membre	0.602	-0.027
R75 niveau de couverture 3 - classe latente 2	-0.102	-0.142
R75 niveau de couverture 1 - classe latente 1	-0.187	0.034
R75 niveau de couverture 1 - classe latente 2	0.101	-0.728
R75 niveau de couverture 2 - classe latente 1	0.210	0.139
R75 niveau de couverture 3 - classe latente 1	0.000	0.000
R75 niveau de couverture 2 - classe latente 2	0.000	0.000
R76 - coefficient de probabilité de membre	0.296	-0.191
R76 niveau de couverture 3 - classe latente 1	0.003	0.121
R76 niveau de couverture 1 - classe latente 2	0.066	-0.492
R76 niveau de couverture 2 - classe latente 1	0.130	-0.232
R76 niveau de couverture 2 - classe latente 2	-0.054	0.190
R76 niveau de couverture 3 - classe latente 2	0.000	0.000
R76 niveau de couverture 1 - classe latente 1	0.000	0.000
R84 - coefficient de probabilité de membre	-0.852	0.201
R84 niveau de couverture 3 - classe latente 2	0.048	0.569
R84 niveau de couverture 1 - classe latente 1	0.063	0.021
R84 niveau de couverture 1 - classe latente 2	-0.687	0.633
R84 niveau de couverture 2 - classe latente 1	-0.006	0.453
R84 niveau de couverture 3 - classe latente 1	0.000	0.000
R84 niveau de couverture 2 - classe latente 2	0.000	0.000
R93 - coefficient de probabilité de membre	-0.673	0.031
R93 niveau de couverture 3 - classe latente 2	0.325	0.484
R93 niveau de couverture 1 - classe latente 1	-0.079	-0.637
R93 niveau de couverture 1 - classe latente 2	-0.634	0.371
R93 niveau de couverture 2 - classe latente 1	-0.007	-0.208
R93 niveau de couverture 2 - classe latente 2	0.000	0.000
R93 niveau de couverture 3 - classe latente 1	0.000	0.000
Lambda - classe latente 1	1.420	0.872
Lambda - classe latente 2	0.783	0.694

TABLE C.11 : Tableau des coefficients des modèles Logit Multinomial à classe latente totale entraînés sur la base 2

Modèle $\mathcal{H}_0$	Modèle $\mathcal{H}_1$	Modèle retenu
Logit_ASR	Logit_ASR_S	Logit_ASR_S
Logit_ASR	Logit_ASR_P	Logit_ASR
Logit_ASR	Logit_ASR_PS	Logit_ASR_PS
Logit_ASR_S	Logit_ASR_PS	Logit_ASR_PS
Logit_ASR_P	Logit_ASR_PS	Logit_ASR_PS

TABLE C.5 : Test de rapport de vraisemblance des modèles logit multinomial estimés sur la base 1 au niveau de 5%

Nom du modèle	MLogit_RC_ASR_PS	MLogit_VS_ASR_PS
Nombre de paramètres	50	51
Log-vraisemblance	-47794.5653	-47795.0834
AIC	95689.1305	95692.1668
BIC	96130.1194	96141.9755
ρ^2	0.1382	0.1469
$\bar{\rho}^2$	0.1373	0.1460

TABLE C.7 : Évaluation des modèles à l'effet aléatoire sur la base 2

Nom du modèle	LCLogit_ASR_PS_L	LCLogit_ASR_PS_P	HLogit_ASR_PS_ASR
Nombre de paramètres	121	121	49
Log-vraisemblance	-47783.1444	-47752.0571	-47783.7666
AIC	95808.2888	95746.1142	95709.5332
BIC	96875.4820	96813.3074	96335.7375
ρ^2	0.1432	0.1437	0.1407
$\bar{\rho}^2$	0.1410	0.1416	0.1394

TABLE C.8 : Évaluation des modèles à classes latentes et du modèle logit d'échelle sur la base 2

Modèle $\mathcal{H}_0$	Modèle $\mathcal{H}_1$	Modèle retenu
Logit_ASR	MLogit_RC_ASR_PS	MLogit_RC_ASR_PS
Logit_ASR	MLogit_VS_ASR_PS	MLogit_VS_ASR_PS
Logit_ASR_PS	MLogit_RC_ASR_PS	Logit_ASR_PS
Logit_ASR_PS	Logit_VS_ASR_PS	Logit_ASR_PS
Logit_ASR_PS	LCLogit_ASR_PS_L	Logit_ASR_PS
Logit_ASR_PS	LCLogit_ASR_PS_P	Logit_ASR_PS
Logit_ASR	LCLogit_ASR_PS_L	Logit_ASR
Logit_ASR	LCLogit_ASR_PS_P	LCLogit_ASR_PS_P
Logit_ASR_PS	HLogit_ASR_PS_ASR	Logit_ASR_PS

TABLE C.9 : Test de rapport de vraisemblance des modèles estimés sur la base 2 au niveau de 5%

Nom du coefficient	HLogit_ASR_PS_ASR
Constant de niveau de couverture 3	-0.226
Constant de niveau de couverture 2	0.000

Constant de niveau de couverture 1	-0.002
Finance	0.592
Plafond	1.260
Tranche d'âge 20-29 ans - coefficient de la variable d'échelle	0.402
Tranche d'âge 20-29 ans niveau de couverture 2	0.000
Tranche d'âge 20-29 ans niveau de couverture 3	-0.136
Tranche d'âge 20-29 ans niveau de couverture 1	0.341
Tranche d'âge 30-39 ans - coefficient de la variable d'échelle	0.204
Tranche d'âge 30-39 ans niveau de couverture 2	0.000
Tranche d'âge 30-39 ans niveau de couverture 3	-0.138
Tranche d'âge 30-39 ans niveau de couverture 1	0.254
Tranche d'âge 40-49 ans - coefficient de la variable d'échelle	0.019
Tranche d'âge 40-49 ans niveau de couverture 3	-0.191
Tranche d'âge 40-49 ans niveau de couverture 2	0.000
Tranche d'âge 40-49 ans niveau de couverture 1	0.189
Tranche d'âge 50-59 ans - coefficient de la variable d'échelle	0.045
Tranche d'âge 50-59 ans niveau de couverture 3	0.000
Tranche d'âge 50-59 ans niveau de couverture 1	0.290
Tranche d'âge 50-59 ans niveau de couverture 2	0.208
Tranche d'âge 60-69 ans - coefficient de la variable d'échelle	0.211
Tranche d'âge 60-69 ans niveau de couverture 3	0.000
Tranche d'âge 60-69 ans niveau de couverture 1	-0.002
Tranche d'âge 60-69 ans niveau de couverture 2	0.021
Tranche d'âge 70-79 ans - coefficient de la variable d'échelle	0.276
Tranche d'âge 70-79 ans niveau de couverture 1	0.000

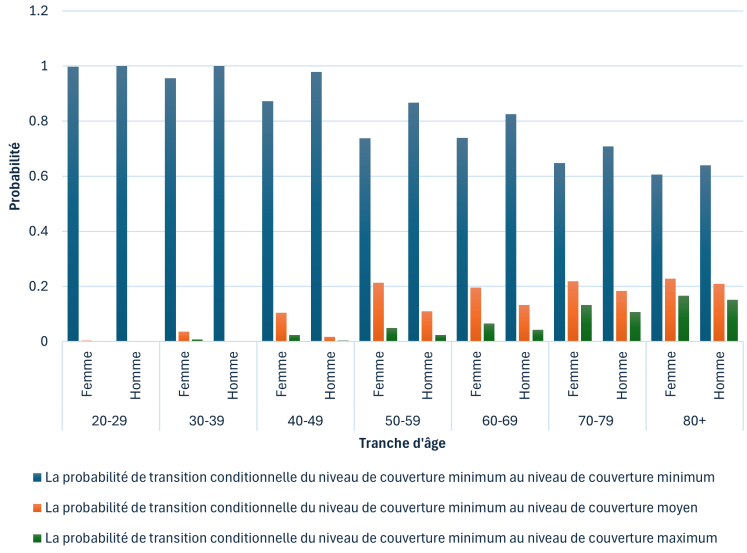
Tranche d'âge 70-79 ans niveau de couverture 3	0.310
Tranche d'âge 70-79 ans niveau de couverture 2	0.139
Tranche d'âge 80+ - coefficient de la variable d'échelle	0.544
Tranche d'âge 80+ niveau de couverture 1	0.000
Tranche d'âge 80+ niveau de couverture 3	0.446
Tranche d'âge 80+ niveau de couverture 2	0.180
Sexe femme - coefficient de la variable d'échelle	0.850
Sexe femme niveau de couverture 1	0.000
Sexe femme niveau de couverture 3	-0.069
Sexe femme niveau de couverture 2	0.065
Sexe homme - coefficient de la variable d'échelle	0.851
Sexe homme niveau de couverture 2	0.000
Sexe homme niveau de couverture 3	-0.158
Sexe homme niveau de couverture 1	-0.006
R11 - coefficient de la variable d'échelle	0.139
R11 niveau de couverture 1	0.000
R11 niveau de couverture 3	-0.094
R11 niveau de couverture 2	-0.124
R24 - coefficient de la variable d'échelle	0.281
R24 niveau de couverture 3	0.000
R24 niveau de couverture 1	0.090
R24 niveau de couverture 2	0.008
R27 - coefficient de la variable d'échelle	0.264
R27 niveau de couverture 2	0.000
R27 niveau de couverture 3	0.025
R27 niveau de couverture 1	-0.027
R28 - coefficient de la variable d'échelle	0.146
R28 niveau de couverture 1	0.000
R28 niveau de couverture 3	-0.007
R28 niveau de couverture 2	0.025
R32 - coefficient de la variable d'échelle	0.203
R32 niveau de couverture 3	0.000
R32 niveau de couverture 1	0.029
R32 niveau de couverture 2	-0.011
R44 - coefficient de la variable d'échelle	0.024
R44 niveau de couverture 2	0.000
R44 niveau de couverture 3	-0.028
R44 niveau de couverture 1	-0.038
R52 - coefficient de la variable d'échelle	0.188
R52 niveau de couverture 1	0.000
R52 niveau de couverture 3	-0.092
R52 niveau de couverture 2	-0.005

R53 - coefficient de la variable d'échelle	0.053
R53 niveau de couverture 2	0.000
R53 niveau de couverture 3	-0.044
R53 niveau de couverture 1	0.002
R5 - coefficient de la variable d'échelle	-0.090
R5 niveau de couverture 2	0.000
R5 niveau de couverture 3	-0.114
R5 niveau de couverture 1	-0.023
R75 - coefficient de la variable d'échelle	0.190
R75 niveau de couverture 3	0.000
R75 niveau de couverture 1	-0.043
R75 niveau de couverture 2	0.003
R76 - coefficient de la variable d'échelle	0.096
R76 niveau de couverture 1	0.000
R76 niveau de couverture 3	0.069
R76 niveau de couverture 2	0.060
R84 - coefficient de la variable d'échelle	0.162
R84 niveau de couverture 3	0.000
R84 niveau de couverture 1	0.019
R84 niveau de couverture 2	-0.025
R93 - coefficient de la variable d'échelle	0.046
R93 niveau de couverture 1	0.000
R93 niveau de couverture 3	0.127
R93 niveau de couverture 2	0.050
Lambda	0.617

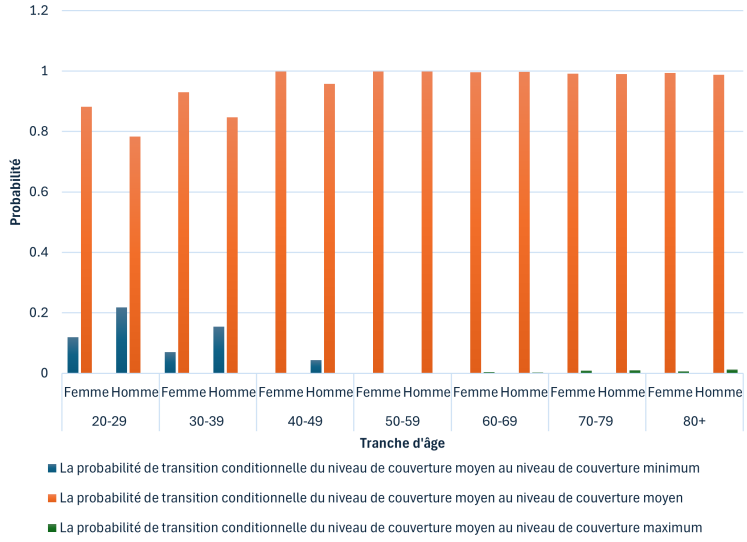
TABLE C.12 : Tableau des coefficients du modèle HLogit_ASR_PS_ASR

C.3 Résultats de l'agrégation des segments de tarifs

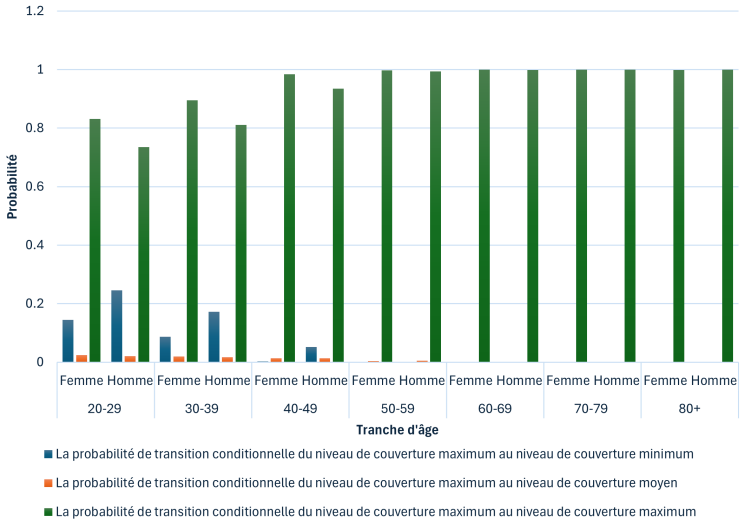
Nous présentons d'abord les probabilités de transition agrégées lorsque les tarifs sont totalement segmentés et deviennent uniques directement, puis les probabilités de transition lorsque les segments de tarifs sont agrégés étape par étape. Les probabilités sont agrégées sur les probabilités individuelles prédites par le modèle, soit par sexe et tranche d'âge, soit par région, afin de présenter les effets des variables de manière plus claire.



(a) Contrat de niveau de couverture minimum

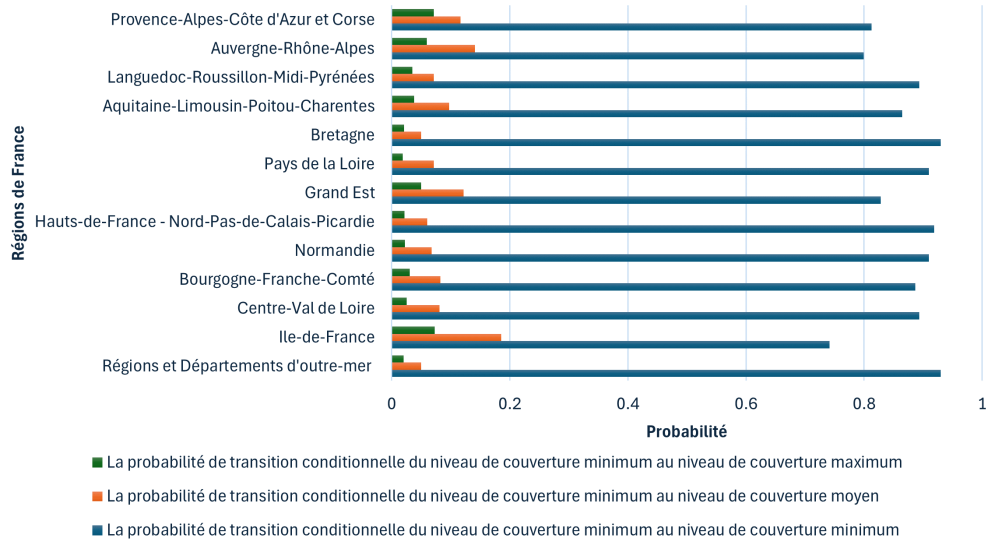


(b) Contrat de niveau de couverture moyen

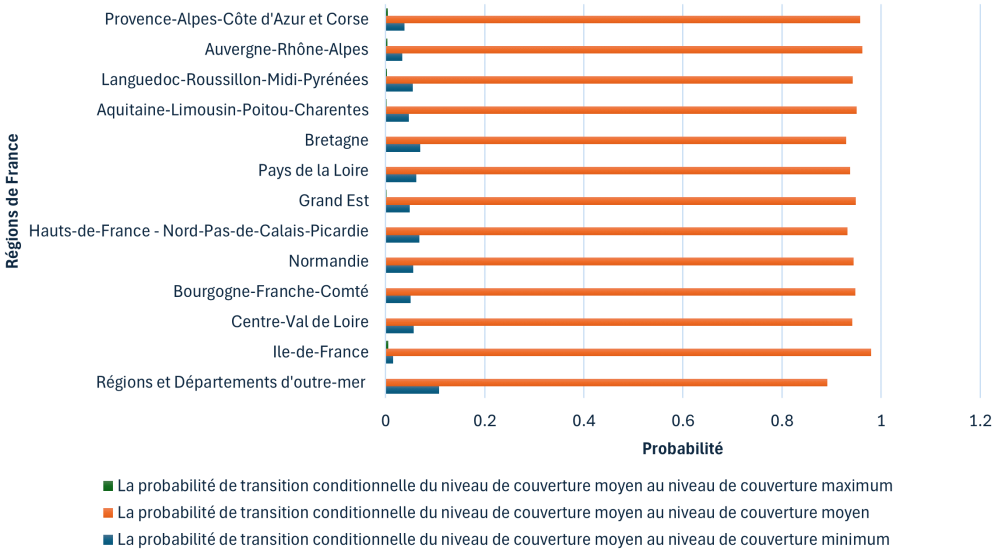


(c) Contrat de niveau de couverture maximum

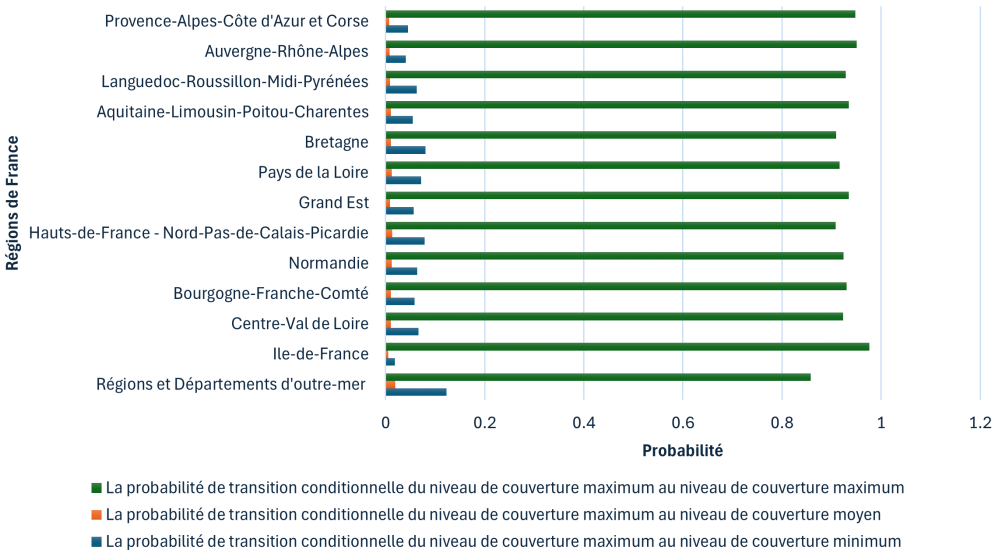
FIGURE C.1 : Probabilités de transition conditionnelles moyennes regroupées par sexe et tranche d'âge pour le changement T(Tranche d'âge, Sexe, Région) $\rightarrow T$(Unique).



(a) Contrat de niveau de couverture minimum

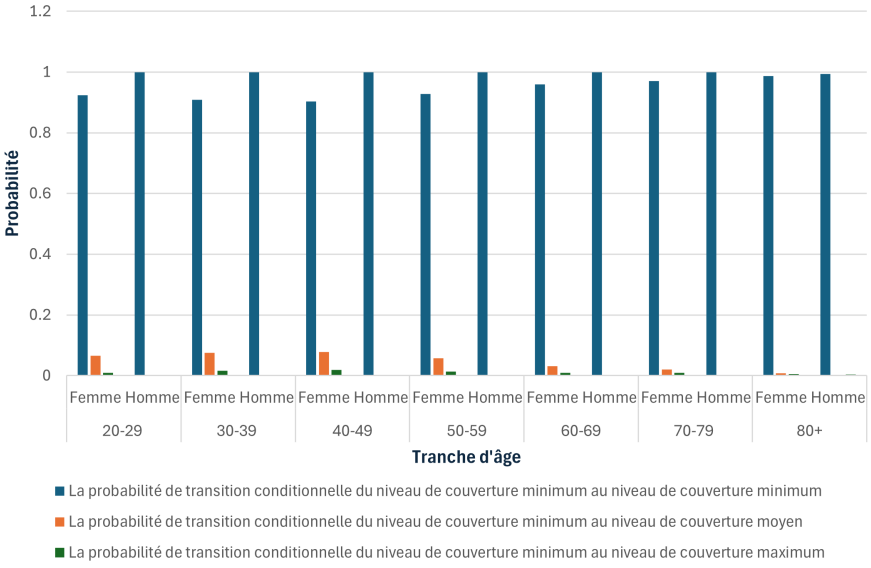


(b) Contrat de niveau de couverture moyen

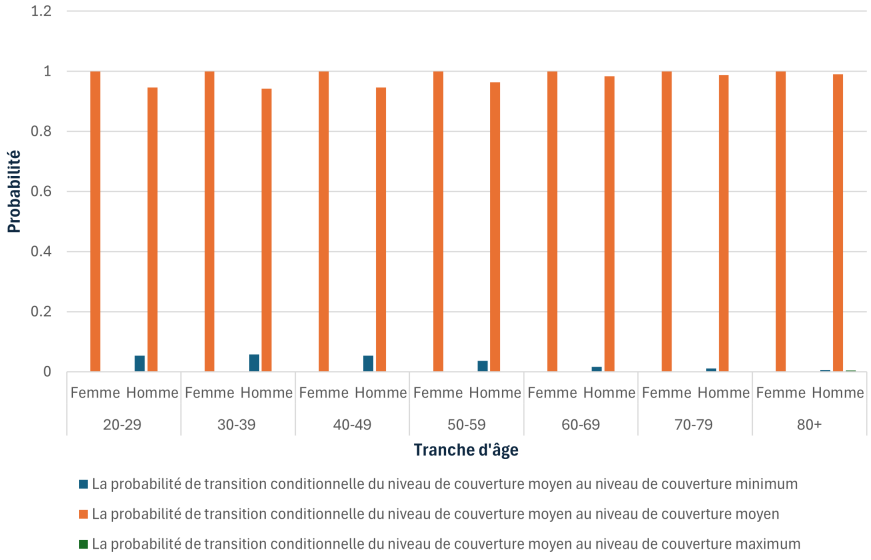


(c) Contrat de niveau de couverture maximum

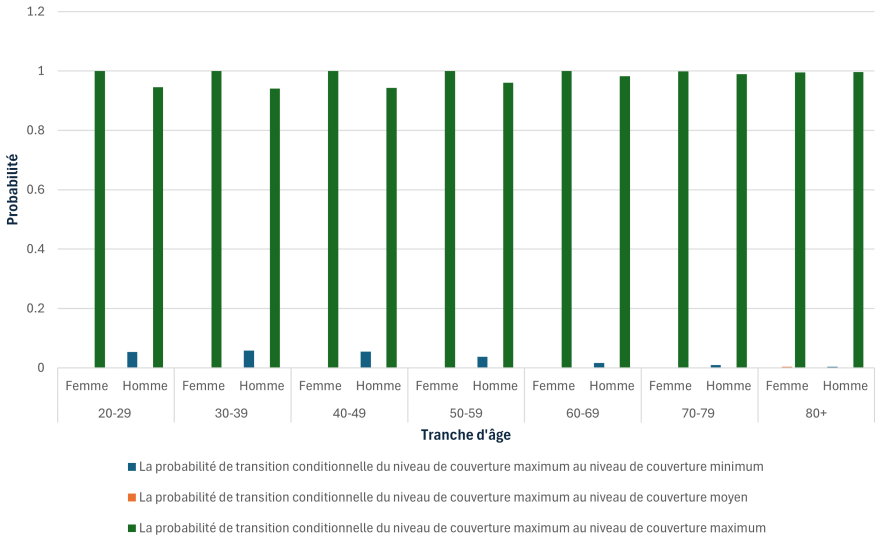
FIGURE C.2 : Probabilités de transition conditionnelles moyennes regroupées par région pour le changement $T(\text{Tranche d'âge, Sexe, Région}) \rightarrow T(\text{Unique})$.



(a) Contrat de niveau de couverture minimum

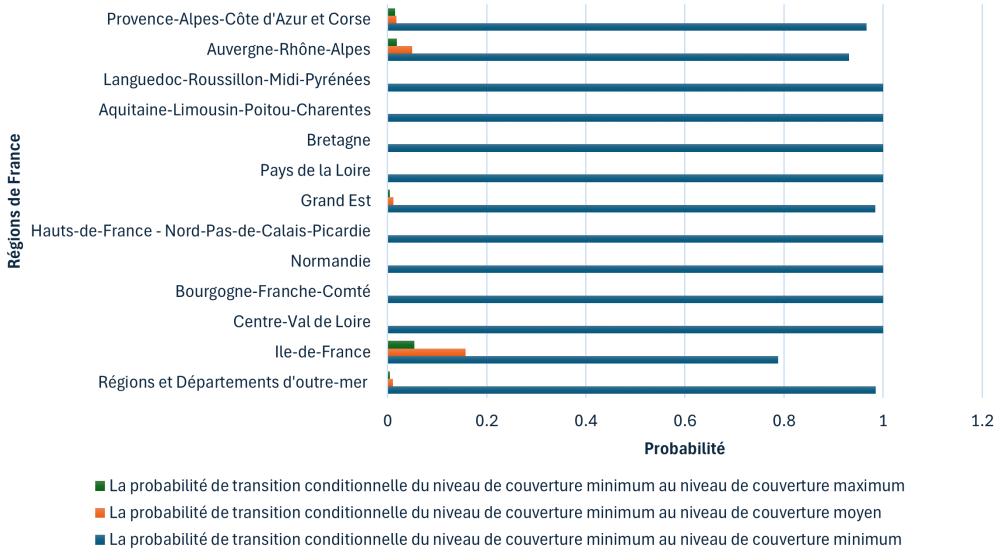


(b) Contrat de niveau de couverture moyen

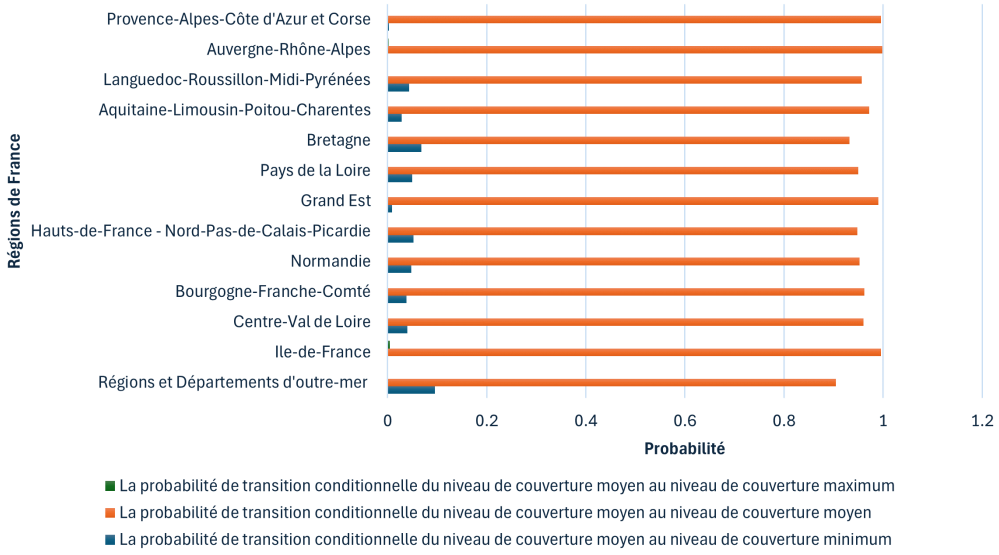


(c) Contrat de niveau de couverture maximum

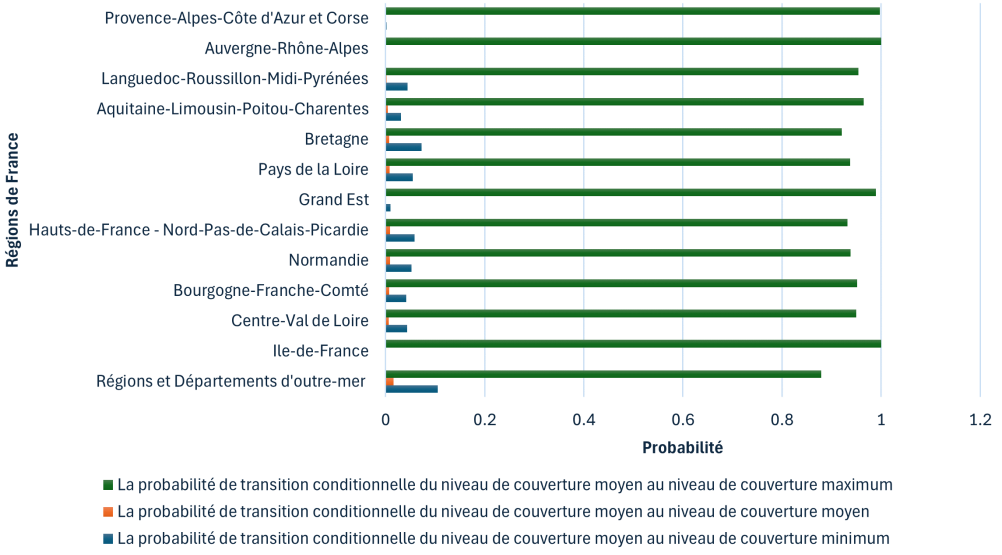
FIGURE C.3 : Probabilités de transition conditionnelles moyennes regroupées par sexe et tranche d'âge pour le changement T(Tranche d'âge, Sexe, Région) $\rightarrow$ T(Tranche d'âge, Région).



(a) Contrat de niveau de couverture minimum

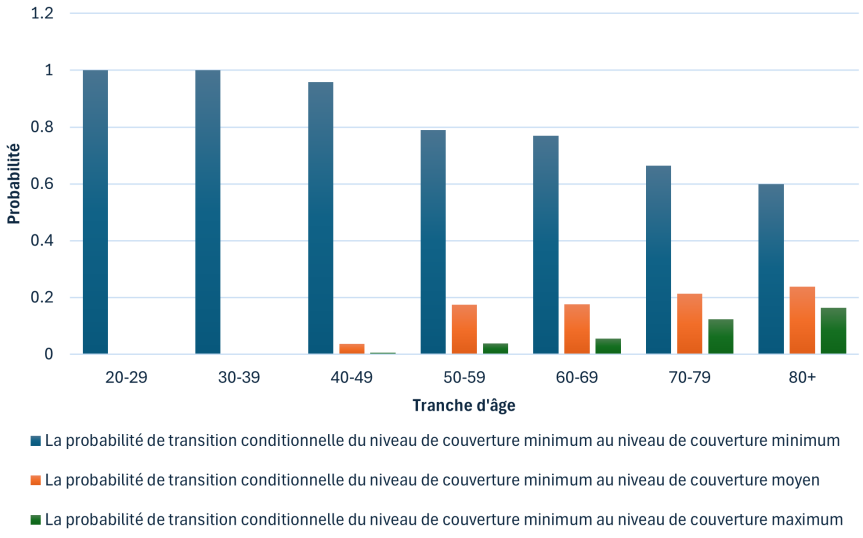


(b) Contrat de niveau de couverture moyen

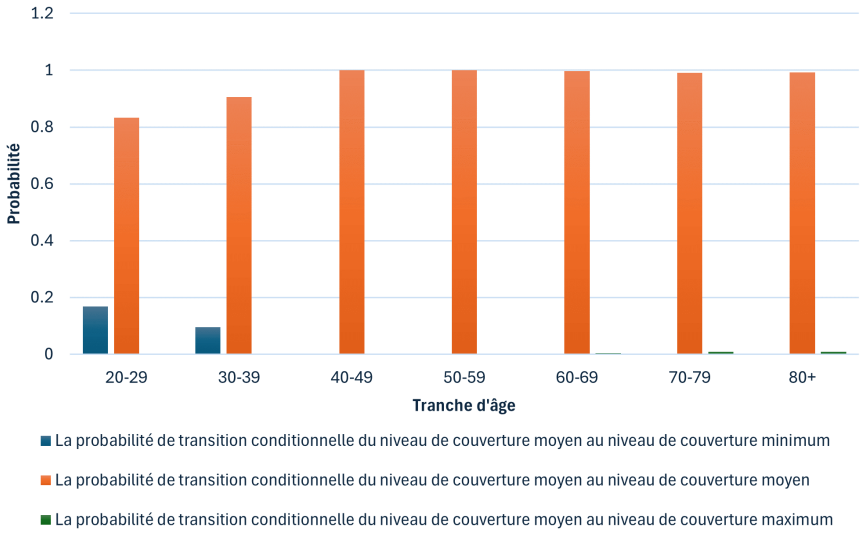


(c) Contrat de niveau de couverture maximum

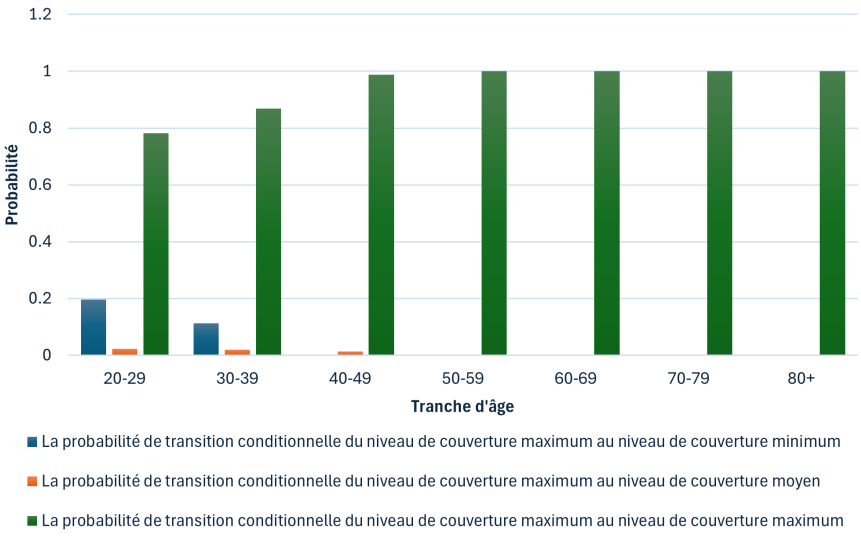
FIGURE C.4 : Probabilités de transition conditionnelles moyennes regroupées par région pour le changement $T(\text{Tranche d'âge, Région}) \rightarrow T(\text{Tranche d'âge})$.



(a) Contrat de niveau de couverture minimum

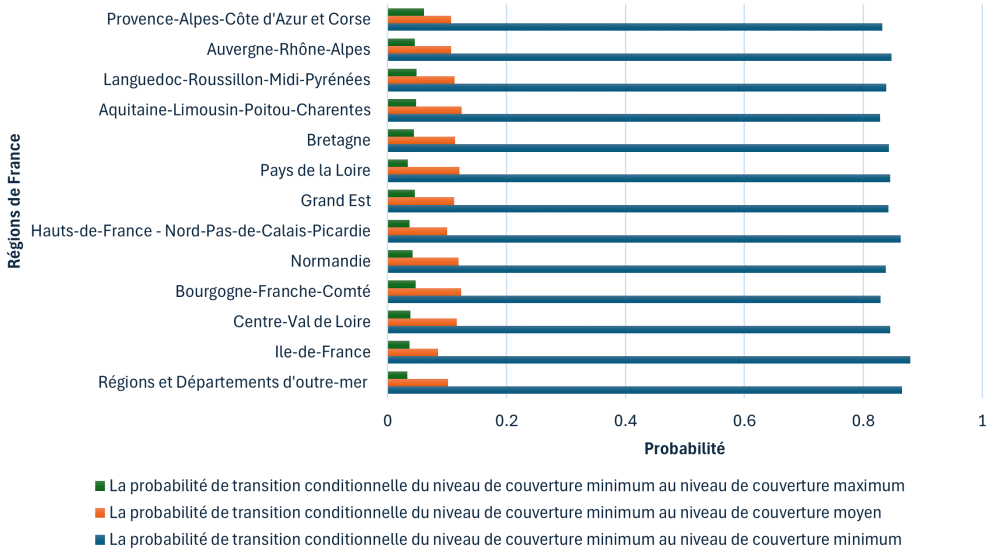


(b) Contrat de niveau de couverture moyen

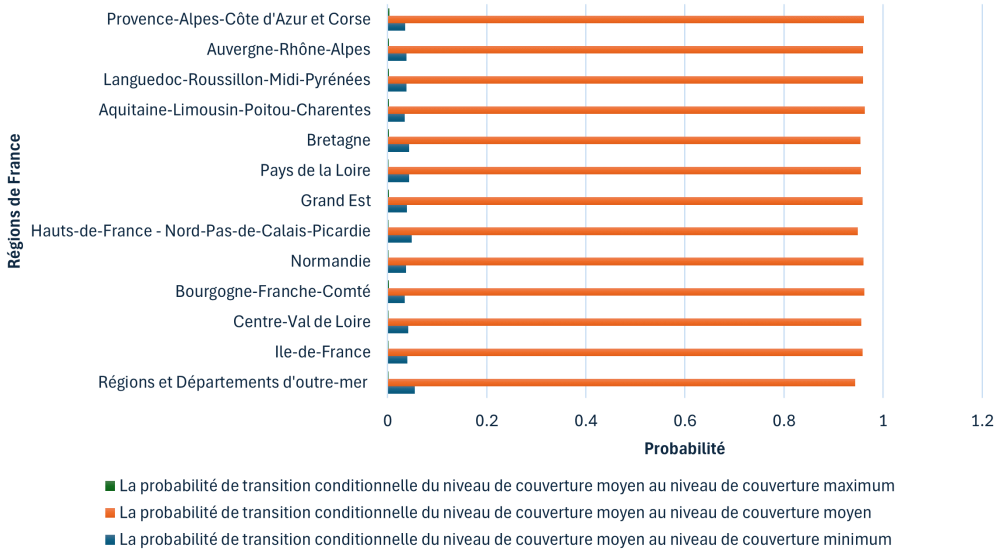


(c) Contrat de niveau de couverture maximum

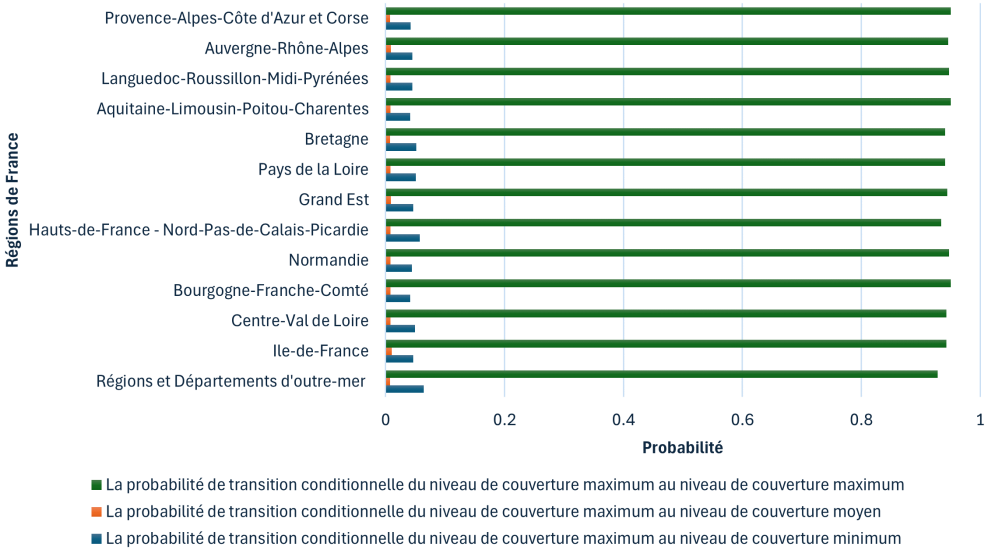
FIGURE C.5 : Probabilités de transition conditionnelles moyennes regroupées par tranche d'âge pour le changement $T(\text{Tranche d'âge}) \rightarrow T(\text{Unique})$.



(a) Contrat de niveau de couverture minimum



(b) Contrat de niveau de couverture moyen



(c) Contrat de niveau de couverture maximum

FIGURE C.6 : Probabilités de transition conditionnelles moyennes regroupées par région pour le changement $T(\text{Tranche d'âge}) \rightarrow T(\text{Unique})$.