

Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuaire  
le **07/03/2024**

Par : **Ikram RADI**

Titre : **Agrégation des marges vie et non-vie pour le calcul de la valeur client**

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

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

*Membres présents du jury de la filière*

*Nom : Olivier LOPEZ*

*Membres présents du jury de l'Institut  
des Actuaire*

*Entreprise : Prim'Act*

*Signature : MBenbrik*

*Directeur du mémoire en entreprise :*

*Nom : Frédéric PLANCHET et Maxime  
BENBRIK*

*Signature : MBenbrik*

**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

MBenbrik

Bibliothèque :

Signature du candidat

[Signature]



## Résumé

---

Un assureur IARD a construit un modèle de projection des marges de son portefeuille. Ce modèle est basé sur une évaluation de l'espérance des marges futures à partir d'un modèle multi-états markovien alimenté par des lois de comportement (équipement, résiliation, changement de garanties, etc.) avec une segmentation tarifaire fine. La somme des marges espérées futures actualisées fournit une mesure de la « valeur client » de chaque segment.

Parallèlement, il existe dans le groupe une société d'assurance-vie commercialisant des contrats d'épargne et de prévoyance, des clients pouvant être communs aux deux entités. Cette structure souhaite développer de son côté un modèle de calcul basé sur la *Value of In-Force* permettant de disposer d'une valeur de son portefeuille qu'il est possible de « découper » par segment le cas échéant.

L'assureur souhaite disposer d'une vision globale de la valeur client en additionnant les deux valeurs. Cette somme ne pose globalement pas de difficulté mais devient plus délicate :

- pour l'affecter par segment de clients, le calcul de la *Value of In-Force* n'étant pas effectué dans une logique « tête par tête » en assurance-vie ;
- pour décomposer la valeur actuelle par année, afin par exemple d'identifier le point mort d'un segment.

Le but de ce mémoire est de proposer dans un premier temps une méthodologie pour l'estimation de la marge vie et non vie par segment, puis de traiter ces deux problématiques.

---

*Mots-clés : valeur client, rentabilité, marge vie, marge non-vie, VIF, MCEV, Shapley, Modèle de Cox, Modèle de Cox étendu.*

## Abstract

---

A *P&C* insurer has built a projection model for its portfolio's margins. This model is based on an evaluation of the expected future margins by a Markovian multi-state model fed by behavioral distributions (equipment, cancellation, change of guarantees, etc.) and a segmentation of the tariff classes. The sum of the discounted future expected margins provides a measure of the customer value of each segment.

At the same time, there is a life insurance company in the group with customers that may be common to both entities. This company is looking to develop a calculation model based on the value of in-force (VIF) allowing it to have a value of its portfolio that is possible to « cut » by segment if necessary.

The insurer wishes to have an overall vision of the customer value by adding the two values. This sum is not of any particular difficulty globally, but it becomes more delicate:

- to allocate it by customer segment, as the calculation of the VIF is not carried out in a « head by head » logic in life insurance;
- to break down the current value by year, in order for example to identify the break-even point of a segment.

The purpose of this study is to firstly estimate the life and non-life profitability segment margins, and then to propose a methodology to deal with the two points above.

---

*Keywords : customer value; profitability; life margin; non-life margin; VIF; MCEV; Shapley; Cox Model; Extended Cox Model.*

# Note de Synthèse

## Contexte

Un groupe d'assurance, qui détient une filiale vie et une autre non-vie, souhaite avoir une idée sur la valeur des clients communs aux deux entités. En effet, l'étude de la rentabilité agrégée l'aidera à identifier les profils d'assurés les plus rentables afin de mieux allouer ses ressources.

D'un côté, la rentabilité en non-vie est calculée à un niveau individuel ou au niveau des *model points*, alors que la rentabilité en assurance vie se détermine généralement au niveau portefeuille. Cette disparité de niveaux de granularité entraîne une difficulté lorsqu'on cherche à agréger les montants pour un segment d'assurés donné.

Comment l'assureur peut-il alors convertir la rentabilité en vie calculée à une large maille en une quantité estimée à un niveau de granularité plus fin, similaire à celui de l'assurance non-vie, pour permettre l'agrégation des deux quantités ?

## Méthodologie

Il est d'abord important de présenter les indicateurs de rentabilité utilisés par chaque entité pour déterminer la valeur client de ses assurés.

D'une part, la filiale non-vie utilise un modèle markovien pour estimer la marge de rentabilité des *model points* de cette branche. Cette marge se constitue des éléments suivants :

- une marge technique qui représente le gain réalisé après l'acquisition des primes et le règlement des sinistres ;
- une marge non technique qui englobe le gain provenant des frais généraux, auxquels on retire les frais d'acquisition ;
- le coût de l'immobilisation de la marge de solvabilité réglementaire (CoC) .

Le modèle markovien se base donc sur l'usage d'un grand nombre d'états, chaque état correspondant à une combinaison de caractéristiques. Les probabilités de transition ainsi que les montants de marge estimés sont calibrés sur l'expérience passée et sont ajustés en permanence. En sortie, le modèle fournit en plus de la marge de l'année en cours, des marges projetées sur 30 ans. Cependant, ces résultats ne sont pas facilement exploitables par les équipes dans leurs stratégies commerciales, parce que l'effet des variables utilisées n'est pas explicitement mis en lumière. Cela justifie donc l'intérêt d'utiliser un modèle statistique plus simple qui vise à reconstruire les marges de rentabilité, estimées par le modèle markovien, et qui se base sur les caractéristiques des assurés.

D'une autre part, la rentabilité pour l'assureur vie sera modélisée en utilisant la *value of in-force*

(VIF). Cet indicateur encapsule la valeur actuelle des polices d'assurance vie en cours, incluant les primes futures anticipées, les prestations à prévoir et les dépenses futures. Une VIF élevée signale que la compagnie possède un portefeuille robuste de polices en cours qui génèrent des flux de trésorerie futurs. Elle se calcule comme suivant :

$$VIF = PVFP - TVOG - FCRC - CNHR$$

avec :

- PVFP (*Present Value of Future Profits*) : la valeur actualisée des flux de trésorerie générés par un segment de clients durant la durée de vie de leurs contrats ;
- TVOG (*Time Value of Options and Guarantees*) : la valeur temps des options et garanties qui se calcule comme la différence entre les cashflows stochastiques et déterministes ;
- FCRC (*Frictional Cost of Required Capital*) : le coût d'immobilisation du capital requis ;
- CNHR (*Cost of Non Hedgeable Risk*) : le coût des risques non répliquables qui n'ont pas été pris en compte dans le calcul.

Pour parvenir à une estimation de la rentabilité au niveau des MP, diverses méthodes d'allocation de capital peuvent être employées. Parmi ces méthodes, les plus répandues sont la méthode proportionnelle et la méthode marginale. La méthode proportionnelle implique l'utilisation des provisions mathématiques comme base de répartition de la rentabilité totale. En revanche, la méthode marginale repose sur le calcul de l'effet marginal de chaque *model point*, en recalculant la VIF après avoir exclu ce dernier de la base de données.

L'inconvénient de ces méthodes est qu'elles ne reflètent pas la réalité des risques intrinsèques à cette base de données. Les résultats obtenus ne retracent pas les différentes interactions possibles entre les *model points* du portefeuille.

La théorie des jeux coopératifs offre une solution qui peut être adaptée à cette problématique : la méthode de Shapley. Aujourd'hui, les valeurs de Shapley sont très utilisées en machine learning pour expliquer les prédictions des modèles : chaque valeur de Shapley quantifie l'impact d'une variable d'entrée spécifique sur la prédiction du modèle pour une observation donnée. Mais la valeur de Shapley a été initialement utilisée pour répondre à la question de la répartition équitable des gains d'une coalition entre les différents joueurs d'un jeu coopératif. En adaptant cette méthode à la problématique de rentabilité d'un assureur, les joueurs deviennent les *model points* et le capital à allouer est le montant global de la VIF. La valeur de Shapley pour le segment  $i$  est définie par la formule suivante :

$$SHAP_i = \sum_{\substack{S \in \mathcal{P}(N) \\ i \in S}} \frac{(s-1)! \times (n-s)!}{n!} [\nu(S) - \nu(S \setminus \{i\})]$$

où  $N$  est un ensemble de  $n$  joueurs,  $S$  une coalition de ce jeu de cardinal  $s$  et  $\nu(S)$  le gain de cette coalition.

En effet, la valeur de Shapley du joueur  $i$  peut être exprimée comme étant la moyenne des contributions marginales aux coalitions auxquelles il est susceptible de participer :

$$SHAP_i = \frac{1}{n} \sum_{\substack{S \in \mathcal{P}(N) \\ i \in S}} \frac{\nu(S) - \nu(S \setminus \{i\})}{C_{n-1}^{s-1}}$$

Ensuite, chaque segment se verra attribuer le poids suivant :

$$w_i = \frac{SHAP_i}{\sum_j SHAP_j}$$

Contrairement aux méthodes citées précédemment, cette approche permet de considérer l'effet des sous-ensembles de joueurs. Cependant, l'effort nécessaire pour calculer les valeurs de Shapley évolue exponentiellement avec le nombre de joueurs, ce qui peut être chronophage lorsque ce dernier est élevé.

Dans ce cas, les valeurs de Shapley peuvent être approchées à l'aide des simulations Monte-Carlo avec échantillonnage aléatoire. Pour chaque MP, on calculera la valeur de Shapley en se basant sur  $n$  échantillons choisis aléatoirement. Ce process sera réitéré  $m$  fois, pour enfin prendre la valeur moyenne qui sera utilisée pour estimer la VIF de chaque *model point*.

$$SHAP_i = \frac{1}{m} \sum_{k=1}^m \frac{1}{n} \sum_{\substack{S \in \mathcal{P}_k(N) \\ i \in S}} \frac{\nu(S) - \nu(S \setminus \{i\})}{C_{n-1}^{s-1}}$$

Où  $m$  est le nombre des scénarios et  $\mathcal{P}_k(N)$  est l'ensemble des combinaisons déterminé par le scénario  $k$ .

Dès lors la valeur client agrégée des *model points* communs aux deux filiales peut être déterminée comme la somme de la marge non-vie calculée par le modèle markovien et la marge vie estimée à la même maille via les poids de Shapley.

## Mise en pratique

Dans le cadre continu de l'analyse de rentabilité et pour approfondir l'étude actuelle, il a été envisagé d'intégrer la marge agrégée dans la détermination du point mort des assurés communs aux deux filiales. En assurance, le point mort d'un segment est défini comme étant l'année où la marge de rentabilité cumulée passe d'une valeur négative à une valeur positive. Le point mort est alors défini sur la base de la projection des marges agrégées sur une période donnée, en l'occurrence 31 ans pour cette étude.

Pour les cinq premières années, nous utiliserons le même process explicité dans la partie précédente afin de déterminer les marges vie par MP en vieillissant le portefeuille d'une année à chaque fois. Pour les années restantes, nous allons utiliser une régression Beta pour estimer les poids de Shapley des MP à partir des données trouvées sur les 5 premières années en utilisant les caractéristiques des assurés comme l'âge, l'ancienneté et les provisions mathématiques.

La base de données utilisée pour cette partie n'est composée que des polices ayant une ancienneté maximale d'une année. Cela est dû au fait que les calculs de marge sont établis à jour, c'est-à-dire qu'ils ne tiennent pas compte des flux réalisés dans le passé, mais uniquement des flux futurs sur la base des données actuelles. Les anciennes polices du portefeuille créent alors un biais de sous-estimation du point mort, vu que la profitabilité des produits en question a tendance à s'améliorer globalement avec le temps. Et pour assurer une homogénéité avec la définition du point mort présentée précédemment, les observations qui changent de signe plus d'une fois, ainsi que celles qui gardent un signe négatif tout au long de la période de projection seront écartées du périmètre de l'analyse.

Enfin, seuls les assurés de la filiale IARD qui ont souscrit un contrat d'épargne seront considérés pour la détection du point mort des clients communs entre les deux assureurs. Les résultats montrent que ces polices deviendraient, en moyenne, rentables après 7 ans. Cette valeur est égale à 10 ans quand elle n'est calculée que sur la base des marges non-vie.

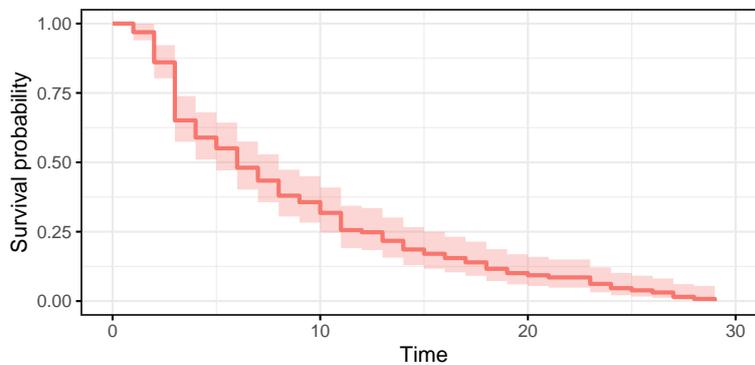
Dans le but d'isoler l'effet de la détention d'un contrat d'assurance vie sur la marge des assurés, nous allons étudier la marge non-vie et analyser l'effet de la variable `detention_epargne` sur cette dernière.

Le graphique met en évidence que l'adhésion à une police d'assurance vie influence favorablement la rentabilité des assurés : la souscription à un produit d'épargne fait avancer le point mort. Cette observation est plus marquée chez les assurés qui détiennent à la fois des polices automobile (AUTO) et multirisques habitation (MRH). Ces polices combinées avec un produit d'épargne deviennent rentables 25 ans avant celles des mêmes assurés qui n'ont aucun produit d'assurance vie.



Répartition des points morts par équipement

En outre, en considérant le point mort comme la durée avant la survenance de l'événement « passage en positif de la marge totale cumulée », il devient apparent que les modèles de survie peuvent être exploités dans cette étude. Les observations qui gardaient un signe négatif pendant toute la durée de la projection seront réinjectées dans la base de données comme données censurées. L'estimateur Kaplan-Meier de la fonction de survie donne la courbe suivante :



Fonction de survie estimée par Kaplan-Meier avec IC=95%

La probabilité de survie atteint le seuil de 50% au bout de la sixième année. Date à partir de laquelle, les assurés ont plus de chance de rentabiliser leurs polices, précédemment non profitables, que de rester déficitaires.

Le test du log-rank montre une différence significative entre les courbes de survie des détenteurs d'un produit d'épargne et des non-détenteurs. Ce test évoque également la différence entre la survie des groupes répartis selon les variables « canal de souscription » et « âge des assurés ». En vue de tenir compte de chacune de ces variables  $x_i$ , le modèle de Cox propose une estimation de la fonction du

hasard  $h(t)$  selon la formule suivante :

$$h(t) = h_0(t) \times \exp(b_1x_1 + \dots + b_px_p)$$

Mais comme les données ne vérifient pas l'hypothèse fondamentale de proportionnalité, nous allons utiliser un modèle stratifié tout en incluant des variables temporelles, comme l'âge, sur la base d'une version étendue du modèle :

$$h(t) = h_0(t) \times \exp\left(\sum_{i=1}^p b_ix_i + \sum_{i=1}^p \delta_ix_i g_i(t)\right)$$

où  $g_i(t)$  est la fonction temporelle pour la covariable  $i$ .

## Conclusion

En conclusion, l'utilisation de la valeur de Shapley permet à l'assureur de répartir la rentabilité vie calculée au niveau portefeuille sur les *model points* tout en reflétant les risques inhérents aux interactions entre les assurés de ce portefeuille.

Le groupe assureur a donc dorénavant une idée sur la rentabilité agrégée des assurés IARD qui détiennent une police d'assurance vie. Cette dernière aide à identifier les polices ou les lignes de produits qui pourraient nécessiter des ajustements dans le but d'améliorer la pérennité de l'activité de l'assureur. Le groupe assureur peut donc adapter son plan de pilotage stratégique en fonction des résultats de cette analyse. De plus, il peut tirer profit de l'aspect temporel de la rentabilité soulevée par l'étude du point mort, comme l'effet des caractéristiques des assurés sur la survenance du point mort



# Executive summary

## Context

An insurance group, which owns a life and a non-life subsidiary, would like to have an idea of the customer value of policyholders common to both entities. A study of aggregate profitability could be used to help the insurer identify the most profitable policyholder profiles, so as to better allocate its resources.

On the one hand, non-life profitability is calculated at individual or model point level, while life insurance profitability is generally determined at portfolio level. This disparity in granularity levels makes it difficult to aggregate the profitability amounts for a specific segment.

In other words, how can the insurer convert the life profitability calculated at a large mesh size into a quantity estimated at a finer level of granularity, similar to that of non-life insurance, to enable the two quantities to be aggregated?

## Methodology

It is important to start by presenting the profitability indicators used by each entity to determine the customer value of its policyholders.

The non-life subsidiary uses a Markov model to estimate the profitability margin of the model points. This margin is made up of the following elements:

- a technical margin: the gain realized after the acquisition of premiums and the settlement of claims;
- a non technical margin: the gain from overhead costs, from which the costs of acquisition are removed;
- a capital cost of the regulatory solvency margin (CoC).

The Markov model is based on the use of a large number of states, each state corresponding to a combination of characteristics. Transition probabilities and estimated margin amounts are calibrated to experience and are continuously adjusted. The model outputs projected margins over 30 years, in addition to the current year's margin. However, these results are not easily exploited for sales strategies, because the effect of the variables used is not explicit. This justifies the use of a simpler statistical model, based on policyholder characteristics, to reconstruct the profitability margins previously estimated by the Markov model.

For the life insurer, profitability will be modeled using the value of in-force (VIF). This indicator encapsulates the present value of life insurance policies in force, including anticipated future premiums,

expected benefits and future expenses. A high VIF indicates that the company has a robust portfolio of in-force policies that generate future cash flows. It is calculated as follows :

$$VIF = PVFP - TVOG - FCRC - CNHR$$

where :

- PVFP : the present value of future profits generated by a client segment over the life of their contracts;
- TVOG : the time value of options and guarantees that is calculated as the difference between stochastic and deterministic cashflows;
- FCRC : the frictional cost of the required capital;
- CRNH : the cost of non hedgeable risks, as in the non-replicable risks that were not included in the elements above.

At this point, to get an estimate of profitability at the MP level, various capital allocation methods can be employed. Among these, the most widespread are the proportional and marginal methods. The proportional method involves using mathematical reserves as the basis for allocating total profitability. The marginal method, on the other hand, is based on calculating the marginal effect of each model point, by recalculating the VIF after excluding it from the database.

The disadvantage of these methods is that they do not reflect the risks intrinsic to this database. The results obtained do not embody the various possible interactions between the portfolio's model points.

Cooperative game theory offers a solution that can be adapted to this problem: the Shapley method. Nowadays, Shapley values are used in machine learning to explain model predictions: each Shapley value quantifies the impact of a specific input variable on the model prediction for a given observation. But Shapley values were initially used to find a fair distribution of the gains of a coalition amongst the different players of a cooperative game. By adapting this method to the problem of the insurer's profitability, the players become the model points, and the capital to be allocated is the overall amount of the VIF. The Shapley value for segment  $i$  is then defined by the following formula:

$$SHAP_i = \sum_{\substack{S \in \mathcal{P}(N) \\ i \in S}} \frac{(s-1)! \times (n-s)!}{n!} [\nu(S) - \nu(S \setminus \{i\})]$$

where  $N$  is the set of  $n$  players,  $S$  is a coalition of this game of cardinal  $s$  and  $\nu(S)$  the gain of this coalition.

Shapley value can also be viewed as the average of the marginal contributions of player  $i$  to coalitions in which he is likely to participate.

$$SHAP_i = \frac{1}{n} \sum_{\substack{S \in \mathcal{P}(N) \\ i \in S}} \frac{\nu(S) - \nu(S \setminus \{i\})}{C_{n-1}^{s-1}}$$

Each segment will be assigned the following weight, according to its Shapley value :

$$w_i = \frac{SHAP_i}{\sum_j SHAP_j}$$

Unlike the marginal method, this approach allows to consider the effect of subsets. However, the effort required to calculate Shapley values is exponential in the number of players, which can be prohibitive when the number of players is high.

In this case, Shapley values can be approximated using Monte Carlo simulations with random sampling. For each MP, the Shapley value is calculated on the basis of  $n$  randomly selected samples. This process will be repeated  $m$  times, to finally take the mean value which will be used to estimate the VIF of each model point :

$$SHAP_i = \frac{1}{m} \sum_{k=1}^m \frac{1}{n} \sum_{\substack{S \in \mathcal{P}_k(N) \\ i \in S}} \frac{\nu(S) - \nu(S \setminus \{i\})}{C_{n-1}^{s-1}}$$

where  $m$  represents the number of scenarios and  $\mathcal{P}_k(N)$  is the set of combinations determined by scenario  $k$ .

The final aggregate customer value of the model points common to both subsidiaries can then be determined as the sum of the non-life margin calculated by the Markov model and the life margin estimated at the same level using Shapley weights.

## Break-even analysis

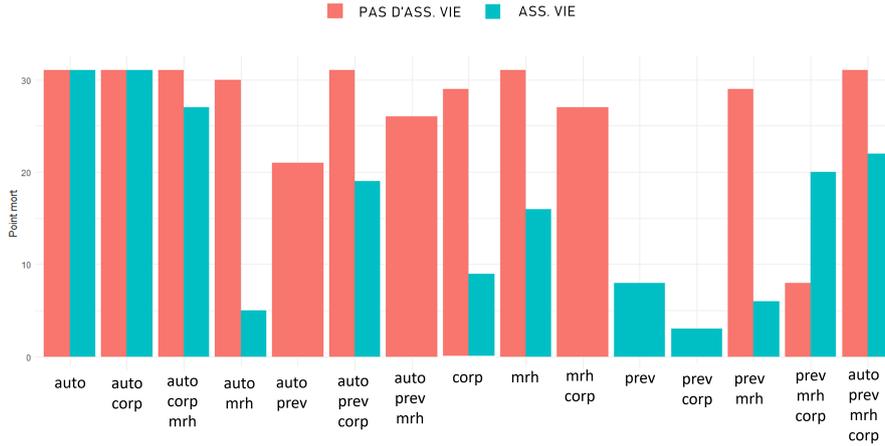
In order to further dissect the profitability, we will focus on examining the break-even point of common policyholders between the two subsidiaries. The break-even point of an insurance segment is defined as the year in which the accumulated profit margin, which at the beginning of the year was negative, becomes positive. The total margin of the insured is calculated as the sum of the life and non-life margins previously estimated. It will be defined on the basis of the projection of the total margins over 31 years.

For the first five years, we'll use the same process explained in the previous section to determine the life margins per MP by aging the portfolio one year at a time. For the remaining years, we'll use Beta regression to estimate the Shapley weights based on the first 5 years' results, using policyholder characteristics such as age, seniority and reserves.

The database used for this part is composed only of policies with a maximum seniority of one year. Since margin calculations are up-to-date, they do not take into account flows that have occurred in the past, but only future flows based on current data. The old policies in the portfolio then create a bias of underestimating the break-even point, since the profitability of the products in question tends to overall improve over time. And to ensure homogeneity with the definition suggested earlier for a break-even point, observations that change sign more than once, as well as those that keep a negative sign throughout the projection period will be excluded from the scope of the analysis.

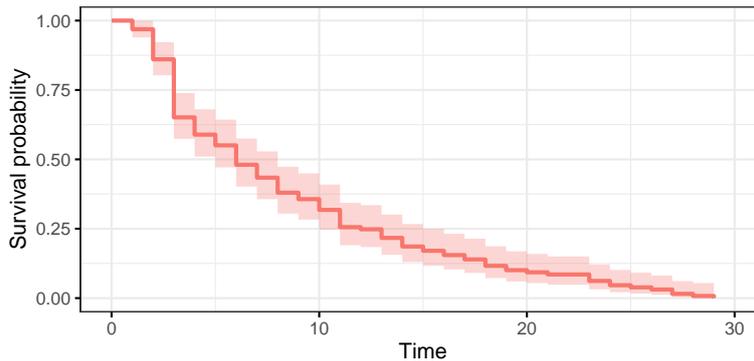
Finally, only the non-life policyholders who also own a savings product will be considered for the detection of the break-even point of joint customers between the two insurers. The results show an average value of 7 years. This value is lower than the one for non-life policyholders, calculated on only the p&c margin which is equal to 10 years.

Consequently, in order to isolate the effect of holding a savings policy on profitability, we will study the non-life margin and analyze the effect of the variable `detention_epargne` on the latter. The graph shows that taking out a life insurance policy favorably influences the profitability of policyholders: underwriting a savings product advances the break-even point. This observation is particularly evident among policyholders who hold both car insurance (AUTO) and multi-risk home (MRH) policies : the break-even point of the latter with a savings product occurs 25 years before those without one.



Distribution of break-even points among products

Furthermore, if we consider the break-even point as the time to occurrence of the event « sign change of the negative cumulative aggregated margin », it becomes apparent that survival models can be used for the purposes of this study. Observations that kept a negative sign throughout the projection will be fed back into the database as censored data. The Kaplan-Meier estimator for the survival function gives the following curve:



Kaplan-Meier Survival curve with CI=95%

It turns out that after the 6th year, the probability of survival reaches the threshold of 50%. From this year onwards, policyholders are more likely to start making profits on their previously loss-making policies than to remain in deficit.

The log-rank test shows the significant difference between the survival curves of holders of a savings product and non holders. This test also evokes the difference in survival within groups of the variables: subscription channel and age. To account for each of these variables  $x_i$ , the Cox model proposes an estimate of the hazard function  $h(t)$  according to the following formula:

$$h(t) = h_0(t) \times \exp(b_1x_1 + \dots + b_px_p)$$

But since the data does not verify the fundamental proportionality hypothesis, we will rely on stratification and including time variables, such as age, based on the following extended version:

$$h(t) = h_0(t) \times \exp\left(\sum_{i=1}^p b_ix_i + \sum_{i=1}^p \delta_ix_i g_i(t)\right)$$

where  $g_i(t)$  is the time function for the covariable  $i$ .

## Conclusion

In conclusion, the use of the Shapley value enables the insurer to spread the life profitability calculated at portfolio level over the model points, while reflecting the risks inherent in the interactions between the policyholders in this portfolio.

The insurer group can now have an idea of the aggregate profitability of P&C policyholders who hold a life insurance policy. This will help the company identify policies or product lines that may need to be adjusted in order to improve the sustainability of the insurer's business.

The insurer group can adjust its strategic steering plan based on the findings of this analysis. Furthermore, the insurer can leverage the temporal dimension of profitability highlighted by this study, including how policyholder characteristics impact the occurrence of the break-even point.



# Remerciements

Je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont contribué, directement ou indirectement, à la réalisation de ce mémoire.

Tout d'abord, je tiens à remercier toute l'équipe Prim'Act, pour leur accueil et leur bienveillance.

Je tiens tout particulièrement à exprimer ma reconnaissance envers Frédéric planchet, mon directeur de mémoire, pour sa guidance tout au long de ce projet. Votre expertise a été une source d'inspiration, et j'ai énormément appris sous votre direction.

Mes sincères remerciements vont également à Maxime Ben-brik pour sa relecture et ses conseils judicieux qui ont grandement enrichi ce travail.

Je tiens à exprimer ma gratitude envers Olivier Lopez pour ses remarques constructives et sa disponibilité. Mais également envers l'ensemble du corps enseignant de la filière d'actuariat et du personnel administratif de l'ENSAE pour leur dévouement et leur disponibilité.

Enfin, je tiens à exprimer ma reconnaissance à ma famille pour leur soutien inconditionnel.



# Acronymes

**CRNHR** *Cost of Residual Non Hedgeable Risks*

**FCRC** *Financial Cost of Required Capital*

**IARD** *Incendie, Accidents et Risques Divers*

**MCEV** *Market Consistent Embedded Value*

**NBV** *New Business Value*

**PVFP** *Present Value of Future Profits*

**TVOG** *Time Value of Options and Guarantees*

**VIF** *Value of In-Force*

**EV** *Embedded Value*

**BE** *Best Estimate*

**RM** *Risk Margin*

**GLM** *General Linear Models*



# Table des matières

<b>Note de Synthèse</b>	<b>5</b>
<b>Executive summary</b>	<b>11</b>
<b>Remerciements</b>	<b>17</b>
<b>Acronymes</b>	<b>19</b>
<b>Table des matières</b>	<b>21</b>
<b>Introduction</b>	<b>23</b>
<b>1 Cadre théorique</b>	<b>25</b>
1.1 Assurance vie . . . . .	25
1.1.1 Caractéristiques de l'assurance vie . . . . .	25
1.1.2 Indicateur de rentabilité . . . . .	27
1.1.3 Calcul de l'indicateur de rentabilité . . . . .	28
1.2 Assurance non-vie . . . . .	32
1.2.1 Caractéristiques de l'assurance non-vie . . . . .	32
1.2.2 Indicateur de rentabilité . . . . .	35
1.2.3 Calcul de l'indicateur de rentabilité . . . . .	36
1.2.4 Estimation avec un modèle markovien multiétats . . . . .	38
<b>2 Mise en pratique</b>	<b>41</b>
2.1 Calcul de la marge vie . . . . .	41
2.1.1 Données et hypothèses . . . . .	41
2.1.2 Outil interne . . . . .	44

2.1.3	Allocation de la VIF . . . . .	47
2.1.4	Résultats . . . . .	50
2.2	Calcul de la marge non-vie . . . . .	51
2.2.1	Données et hypothèses . . . . .	51
2.2.2	GLM . . . . .	57
2.2.3	LASSO . . . . .	58
2.2.4	Résultats . . . . .	59
<b>3</b>	<b>Agregation des marges</b>	<b>63</b>
3.1	Détermination du point mort . . . . .	63
3.1.1	Calcul du point mort . . . . .	63
3.1.2	Analyse univariée . . . . .	66
3.2	Modèle de survie . . . . .	69
3.2.1	Estimateur de Kaplan-Meier . . . . .	70
3.2.2	Estimation de la variance . . . . .	72
3.2.3	Comparaison des courbes . . . . .	73
3.3	Impact des variables . . . . .	76
3.3.1	Modèle de Cox . . . . .	76
3.3.2	Vérification des hypothèses . . . . .	78
3.3.3	Modèle de Cox étendu . . . . .	80
	<b>Conclusion</b>	<b>85</b>
	<b>Bibliographie</b>	<b>87</b>
	<b>Annexes</b>	<b>93</b>
<b>A</b>	<b>Une annexe</b>	<b>95</b>
A.1	Avec des tableaux . . . . .	95
A.2	Avec des graphiques . . . . .	95

# Introduction

L'importance de la rentabilité dans le secteur de l'assurance est indéniable, car elle joue un rôle prépondérant dans l'évaluation de la stabilité financière globale de la compagnie. Par conséquent, le régulateur, comme les investisseurs, incorporent dans leurs analyses les marges de rentabilité des assureurs pour juger la durabilité de leur profitabilité.

La marge de rentabilité d'une entreprise d'assurance est tributaire de divers facteurs. De fait que les compagnies d'assurance tirent leurs revenus des polices qu'elles souscrivent et des primes qu'elles perçoivent. Et qu'elles ont, en contrepartie, divers coûts ; notamment les frais généraux et les sinistres. Par ailleurs, quelle que soit la taille de la compagnie, la marge de rentabilité nette diffère d'une branche à l'autre, en fonction de la manière dont l'assureur exerce son activité et gère ses dépenses.

En effet, l'activité en assurance-vie se caractérise par l'offre des produits destinés à fournir une protection financière en cas de sinistre de l'assuré ou à constituer un capital pour des besoins futurs, tels que la retraite. Les primes d'assurance vie sont généralement payées sur une base régulière, et les indemnités sont versés aux bénéficiaires en cas de sinistre. Et donc la rentabilité des assureurs vie dépend principalement des primes collectées et des rendements générés par les investissements.

En revanche en assurance non-vie, l'assureur couvre une gamme plus large de risques, notamment l'assurance automobile, l'assurance habitation, l'assurance santé et d'autres biens et responsabilités. Les polices d'assurance non-vie sont souvent contractées pour des périodes plus courtes, généralement d'un an, et couvrent les pertes ou les dommages subis par l'assuré pendant cette période. Dans ce cas, la dimension financière de l'assurance vie n'est plus présente. Les bénéfices proviennent de façon générale des primes diminuées des coûts des sinistres et des frais administratifs.

C'est dans ce contexte qu'un groupe d'assurance, qui restera anonyme, souhaite obtenir une perspective consolidée et harmonisée de sa rentabilité. Cet assureur détient des filiales dans les deux branches vie et non-vie et s'attend à avoir une estimation des marges de rentabilité agrégée des assurés communs aux deux entités. Dans le cadre de cette démarche, nous procéderons au calcul des marges de rentabilité de chacune des deux branches de manière distincte. La problématique qui se pose par la suite réside dans la recherche d'une méthodologie appropriée pour agréger ces deux valeurs.

Le premier chapitre discute alors le choix des indicateurs de rentabilité à utiliser pour chacune des branches. Pour la partie vie, la première section se base sur la détermination d'un bon indicateur pour calculer la profitabilité des assurés de la branche. Alors que pour la partie non-vie, la filiale en question dispose d'un modèle markovien pour calculer la rentabilité de ses assurés. Cependant le calcul s'avère très lourd et peu interprétable. Les équipes ne peuvent pas déterminer explicitement l'effet des variables employées par le modèle sur la valeur de la marge. Alors pour enrichir les résultats de profitabilité de cette branche, l'objectif de la deuxième section du premier chapitre est de proposer un modèle statistique qui reproduira la construction des marges calculées par le modèle markovien mais en se basant sur les caractéristiques des assurés.

Cependant, l'intégration de ces deux modèles pour obtenir une vision globale de la valeur client pose des défis méthodologiques non négligeables. La complexité réside dans la méthode d'affectation de la somme totale aux segments des clients. Les indicateurs de rentabilité de l'assurance vie sont généralement déterminés sur une maille plus large que celle de la non-vie, souvent au niveau portefeuille ; contrairement à la non-vie où le calcul est au niveau des assurés.

L'apport de cette étude est de proposer dans le deuxième chapitre une méthode pour comparer les deux marges au niveau des *model points*, comme l'assureur veut s'en servir pour étudier la profitabilité des assurés communs entre les deux filiales.

De plus, la décomposition de la marge actuelle par année s'avère nécessaire pour des analyses fines telles que l'identification du point mort d'un groupe d'assurés. La marge obtenue est donc utilisée dans le troisième chapitre pour calculer ce dernier ainsi que pour étudier l'effet de la détention simultanée des deux types de produits sur la durée avant la réalisation du point mort.

# Chapitre 1

## Calcul de la marge en assurance

La rentabilité est un indicateur de santé financière d'une compagnie d'assurance. Elle permet d'évaluer si l'entreprise génère suffisamment de revenus et de bénéfices durables pour couvrir ses obligations.

L'analyse de la rentabilité aide alors les compagnies d'assurance à prendre des décisions éclairées sur leurs stratégies commerciales. En comprenant quelles sont les lignes de métier ou les produits les plus rentables, les assureurs peuvent ainsi se concentrer sur ces derniers et apporter des ajustements stratégiques qui optimisent leurs opérations.

C'est dans cette optique qu'un groupe d'assurance peut s'intéresser à identifier les profils des clients les plus rentables, afin de pouvoir allouer efficacement leurs ressources et de prioriser les initiatives qui visent à améliorer la rentabilité.

Ce mémoire porte sur un groupe d'assurance qui détient une filiale vie et une autre non-vie. Nous chercherons donc à développer des modèles de calcul qui estiment les marges de rentabilité et ce pour des segments de clients propres à chaque filiale.

L'un des principaux enjeux de ce projet consiste en la mesure et le pilotage de la rentabilité globale à moyen et long terme des clients qui possèdent à la fois des produits non-vie et des produits d'épargne.

### 1.1 Assurance vie

#### 1.1.1 Caractéristiques de l'assurance vie

L'assurance touche à des secteurs hétérogènes et répond à des besoins variés. Pour que la mutualisation des risques prenne tout son sens, l'assurance est divisée à minima en deux branches : l'assurance vie et l'assurance non-vie.

La notion d'assurance vie « recouvre l'ensemble des opérations d'assurance qui se rapportent directement ou indirectement à la couverture des risques liés à la durée de la vie humaine. Dans sa plus large acception, elle recouvre le périmètre des assurances en cas de vie, des assurances en cas de décès, ou des assurances mixtes » ([ACPR, 2016](#)).

- Les contrats d'assurance-vie prévoient le versement d'un capital accumulé ou de rentes si l'assuré est toujours en vie à la fin du contrat. Ce type de contrat bénéficie d'une fiscalité avantageuse.

- Les contrats décès permettent aux souscripteurs de constituer une épargne au profit d'un tiers. Le contrat stipule qu'au décès du souscripteur, un certain capital sera versé à la personne de son choix, le bénéficiaire.
- Et les contrats mixtes représentent une garantie de versement du capital ou de la rente à l'échéance, au souscripteur, s'il est vivant, ou au bénéficiaire, si le souscripteur est décédé.

En assurance-vie, des garanties type « capitalisation » sont aussi de mise. Elles ne reposent pas sur la vie humaine mais l'aléa est plutôt lié à des remboursements anticipés sous certaines conditions.

Les garanties en vie sont généralement de type forfaitaire dont le montant est déterminé ou déterminable à la souscription. Les conditions particulières du contrat individuel ou collectif doivent indiquer les montants de cotisations et les capitaux à régler par l'assureur. (Bocquaire & Camus-Paradis, 2022)

Au-delà de la différence au niveau de l'élément déclencheur de la garantie, une seconde différence caractérise les contrats en assurance-vie : certains contrats sont investis en un seul support, d'autres sur plusieurs. Nous parlons dès lors des supports :

- en fonds en euros quand on investit sur les marchés monétaires peu risqués ;
- ou des fonds investis en unités de compte (actions, obligations, parts de sociétés d'investissement à capital variable ou de fonds communs de placements, etc).

D'après les chiffres de France Assureurs, pour la première moitié de l'année 2023 : « les cotisations en assurance vie s'établissent à 81,8 milliards d'euros, en hausse de +6% par rapport à la même période de 2022. Cette augmentation est quasi-identique pour les fonds en euros et les supports en UC : 48,6 milliards d'euros pour les premiers, en hausse de +6%, et 33,2 milliards d'euros, en hausse de +7%, pour les seconds. La part des cotisations en UC est de 41% depuis le début de l'année, à comparer à 40% pour l'ensemble de l'année 2022. » (France Assureurs, 2023)

Le groupe assureur en question, anonyme dans le cadre de cette étude, possède une filiale spécialisée dans la production des produits en épargne, prévoyance et gestion de patrimoine. Elle commercialise des produits aux particuliers comme aux entreprises. Cependant, notre étude sur la partie vie se focalisera sur des produits retraite pour les particuliers en phase de constitution, assimilables à des produits d'épargne euro.

En effet, cet assureur explore les subtilités qui sous-tendent sa rentabilité pour plus d'une raison, les principales étant :

- la fidélisation de sa clientèle ;
- l'amélioration du développement des nouveaux produits ;
- l'évaluation de ses risques ;
- une meilleure gestion de ses investissements.

Pour cela, il cherche à modéliser la rentabilité de ses clients en modélisant une marge de rentabilité appelée « valeur client ».

En assurance vie, plusieurs facteurs font que le calcul de rentabilité revêt une certaine complexité, notamment :

- les engagements de l'assureur se caractérisent par un horizon long terme ;
- les options et garanties financières accordées peuvent représenter des montants très importants qui ont des comportements asymétriques ;
- la valeur des polices ne peut pas se limiter aux éléments qui figurent dans les états financiers.

C'est pour cette raison que l'assureur de cette branche doit accorder une attention particulière au choix des indicateurs de rentabilité.

### 1.1.2 Indicateur de rentabilité

Les compagnies d'assurance-vie ont une place importante dans le secteur financier. Elles occupent une position unique en mettant l'accent sur la gestion des risques et la protection des assurés. Cependant, ces compagnies recherchent constamment un équilibre délicat qui assure leur viabilité économique et leur croissance à long terme, surtout dans le contexte actuel de multiplication des normes réglementaires.

En effet, la rentabilité des compagnies d'assurance-vie repose sur une approche qui combine des pratiques de souscription astucieuses, des stratégies d'investissement prudentes, des offres de produits innovantes et un engagement sans faille en faveur de la satisfaction des clients. A mesure que le cadre économique évolue et que les préférences des clients changent, ces compagnies doivent s'adapter rapidement pour rester compétitives tout en respectant leur promesse de préserver le bien-être financier des assurés et de leurs bénéficiaires.

Les compagnies d'assurance vie analysent donc régulièrement plusieurs indicateurs afin d'évaluer leurs performances et d'identifier les domaines à améliorer pour accroître non seulement leur rentabilité, mais aussi leur compétitivité. Depuis les années 90, un nouvel outil d'évaluation s'est développé pour mieux appréhender la valeur des activités long terme de l'assurance vie : l'*Embedded Value* (EV). Il a été introduit afin d'évaluer au plus juste la richesse et le potentiel des compagnies d'assurance et de réassurance. Au cours des dernières années, de très nombreuses évolutions ont été apportées à cette notion, en réponse avec les conjonctures économiques rencontrées.

Aujourd'hui, l'un des ICP (Indicateur Clé de Performance) les plus connus de l'industrie de l'assurance vie est la *Market Consistent Embedded Value* (MCEV). Elle mesure la valeur de l'entreprise du point de vue actionnaire et a fait suite à d'autres versions de l'EV. C'est la méthode de calcul préconisée par le CFO Forum ([CFO forum, 2016](#)), puisqu'elle répond à des principes de calcul bien spécifiques dont des hypothèses financières qui respectent l'approche market consistent et des paramètres de projection déterminés de façon *best estimate*.

Toutefois, afin d'analyser la rentabilité des différents segments d'un assureur vie, et non sa profitabilité globale, nous pouvons nous focaliser sur un élément de la MCEV, à savoir la *Value of In-Force* (VIF). La VIF fait partie des principaux indicateurs fréquemment employés pour évaluer la rentabilité dans le domaine de l'assurance vie. Cette mesure encapsule la valeur actuelle des polices d'assurance vie en cours, qui inclue les primes futures anticipées, les prestations à prévoir, et les dépenses futures.

La VIF sert aussi à évaluer la stabilité financière d'une compagnie d'assurance. Une VIF élevée signifie que la compagnie possède un portefeuille robuste de polices en cours qui génèrent des flux de trésorerie futurs. Cette situation peut positivement impacter la solvabilité et la santé financière de la compagnie. Les autorités de réglementation des assurances peuvent donc l'utiliser comme indicateur de la capacité d'un assureur à remplir ses obligations. Dans ce cas, le maintien d'un niveau adéquat de VIF peut démontrer la conformité aux exigences réglementaires. Cette mesure joue aussi un rôle crucial dans l'évaluation des compagnies d'assurance lors des fusions et acquisitions. Les investisseurs potentiels évaluent la VIF comme un élément essentiel de leur compréhension de la valeur des affaires qu'ils acquièrent.

En complément de l'évaluation de la rentabilité et de la fiabilité financière, la VIF assiste les assureurs dans l'appréhension de la valeur des assurés. Cet indicateur est étroitement lié à la contribution des assurés à la valeur globale de l'entreprise. Une VIF plus basse reflète une rentabilité moindre. Par conséquent, les assureurs s'efforcent d'augmenter la VIF afin d'amplifier la valeur apportée par les assurés. Par conséquent, elle aide les assureurs dans l'orientation de leurs décisions commerciales.

La VIF aide ainsi à identifier les polices ou les lignes de produits sous performantes qui pourraient nécessiter des ajustements pour améliorer la rentabilité.

Globalement, la *value of in-force* est un indicateur essentiel qui permet aux assureurs de mesurer leurs performances financières, de gérer les risques et de prendre des décisions stratégiques. Elle fournit des informations clés sur les flux de trésorerie futurs de l'entreprise et sur la santé globale de ses activités. Par ailleurs, en décembre 2021, S&P Global Ratings a publié un appel aux commentaires à l'égard des changements proposés à leur méthodologie d'évaluation de l'adéquation des fonds propres des assureurs. L'un de ces éléments de changement est l'inclusion potentielle d'un crédit de 100% pour la VIF, qui augmenterait le TAC (*Total Adjusted Capital*) et améliorerait le ratio d'adéquation des fonds propres. Les assureurs qui calculent donc déjà la VIF bénéficient d'une meilleure compréhension de la gestion de leurs activités. Si cette composante est adoptée, ces mêmes assureurs bénéficieront d'un avantage supplémentaire dans leurs notations S&P (Menmeymer, 2022).

Naturellement, les assureurs vie peuvent utiliser d'autres indicateurs pour discuter de la rentabilité inhérente à leurs activités. C'est dans ce sens que se présente la New Business Value (NBV). Cette métrique représente la même idée fondamentale que la VIF et suit la même logique de calcul, mais elle s'applique sur un périmètre différent : au lieu de considérer le stock global du portefeuille, la NBV se concentre uniquement sur les nouvelles souscriptions. De là découle un autre indicateur de profitabilité ; la NBM (*New Business Margin*). Il s'agit cette fois-ci d'un ratio de profitabilité, calculé comme le rapport entre la NBV et le montant des primes actualisées.

### 1.1.3 Calcul de l'indicateur de rentabilité

Plusieurs principes ont été établis dans le secteur de l'assurance pour formaliser le cadre de la VIF ; toutefois ces principes comportent un certain degré de flexibilité, et les assureurs qui mettent en œuvre la VIF doivent prendre un nombre de décisions clés, notamment en ce qui concerne :

- le taux d'actualisation à utiliser ;
- la mesure à employer pour définir le capital cible ;
- la méthode d'allocation des dépenses.

Parmi ces principes, nous retrouvons celui de la *market consistency*. Ce dernier énonce que la valeur d'un actif ou d'un passif conforme au marché est sa valeur de marché, s'il est facilement négociable sur un marché au moment où l'évaluation est effectuée. Sinon pour tout autre actif ou passif, la meilleure estimation de sa valeur est ce qu'aurait été sa valeur de marché s'il avait été facilement négociable au moment de l'évaluation.

Cette notion se base sur le concept d'absence d'opportunité d'arbitrage et est devenue un standard dans la valorisation des portefeuilles des assureurs. Elle permet d'obtenir une juste valeur cohérente avec les valeurs et les risques du marché, raison pour laquelle elle constitue une référence dans le cadre prudentiel Solvabilité II.

C'est dans ce contexte qu'intervient l'idée du calcul de la *value of in-force* en valeur économique. Cette notion est semblable à celle du *best estimate* dans le référentiel Solvabilité II. La valeur économique tient compte d'un côté de l'espérance mathématique des *cashflows* futurs actualisés ( dont les primes, les indemnités de décès, les commissions ou les frais généraux) mais prend également en considération le coût des options et garanties implicites. Ces derniers peuvent consister des éléments suivants :

- Option de rachat : elle permet au titulaire du contrat de retirer une partie (rachat partiel) ou la totalité (rachat total) des fonds accumulés avant la fin de la période initialement convenue. Nous

faisons la distinction entre le rachat structurel, qui dépend des caractéristiques de la police et du comportement de l'assuré et le rachat conjoncturel, qui est corrélé à la conjoncture économique.

- Option d'arbitrage : elle fait référence à la possibilité pour le titulaire du contrat de réallouer les fonds investis entre différentes options d'investissement proposées par la compagnie.
- Option d'avance : elle donne la possibilité au titulaire du contrat de demander un prêt auprès de la compagnie d'assurance. Elle assure une certaine flexibilité financière qui permet au client d'accéder à une partie des fonds de la police sans effectuer de rachat.
- Taux minimum garanti : c'est le taux d'intérêt minimal que la compagnie d'assurance s'engage à verser sur le capital investi.
- Participation aux bénéfices : Il s'agit d'une garantie propre aux contrats d'assurance vie en euro participatifs. Cette disposition permet aux assurés de partager les bénéfices réalisés par l'assureur.
- Garantie plancher : elle garantit à l'assuré que la valeur de son contrat ne pourra pas descendre d'un certain niveau prédéterminé, même en cas de baisse significative des performances des investissements sous-jacents.

Comme expliqué ci-dessus, la *value of in-force* désigne la valeur générée dans le futur par le stock des contrats dans le portefeuille. Elle correspond à la valeur actuelle des profits probables actualisés et distribuables aux actionnaires sur les contrats présents à la date de calcul. Par conséquent, sa valeur économique se calcule comme la valeur actuelle des profits futurs (*Present Value of Future Profits* (PVFP)) diminuée de la valeur temps des options et garanties (*Time Value of Options and Guarantees* (TVOG)), des coûts frictionnels (*Financial Cost of Required Capital* (FCRC)) et des risques résiduels non couvrables (*Cost of Residual Non Hedgeable Risks* (CRNHR)).

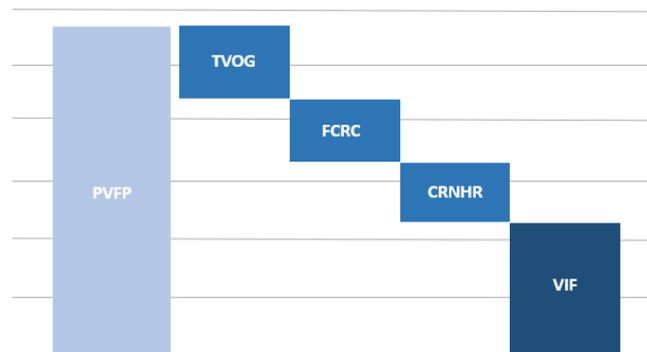


FIGURE 1.1 – Décomposition de la VIF

La détermination de la VIF implique les étapes suivantes :

— **Etape 1 : calcul de la PVFP déterministe**

Il s'agit de la valeur actuelle des profits futurs générés par le portefeuille des contrats, calculée suivant un scénario déterministe sans prime de risque. Les projections sont effectuées en run off. Pour un horizon de projection  $T$ , la PVFP est donnée par :

$$PVFP_{deterministe} = \sum_{t=1}^T \frac{CF_t}{(1+r_t)^t}$$

avec :

- $CF_t$  : Cashflows ou résultat de l'année t.
- $r_t$  : taux d'actualisation de l'année t.

Les résultats d'une année se calculent en sommant les gains réalisés par l'assureur :

- primes ;
- frais ;
- produits financiers.

Et en soustrayant les charges telles les :

- prestations payées en cas de sinistres ;
- coûts de gestion ;
- impôts.

Ce calcul est obtenu suite à la mise en place par l'assureur de nombreuses hypothèses techniques et financières *best estimate*. Ces hypothèses doivent refléter l'expérience passée spécifique à l'assureur, ainsi que toute autre donnée pertinente.

La PVFP inclut la valeur intrinsèque des options et garanties, mais pas leur valeur temps qui est incorporée dans le calcul de la TVOG.

## — Etape 2 : calcul de la TVOG

La TVOG révèle le caractère de l'asymétrie généré par les dispositions des polices d'épargne en assurance vie. Cette asymétrie provient principalement du risque et de l'incertitude intrinsèques aux marchés financiers.

Le calcul de la TVOG peut se faire dans un premier temps selon la formule suivante :

$$TVOG = PVFP_{stochastique} - PVFP_{deterministe}$$

La valeur temps des options et garanties se calcule dès lors comme la différence entre les cashflows stochastiques et déterministes en lien avec les garanties en question. Il faut donc déterminer la valeur stochastique PVFP en univers risque-neutre pour respecter la particularité *market consistent*.

Cette quantité s'établit comme l'espérance des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle attendue des flux de trésorerie futurs), estimée sur la base de la courbe des taux sans risque appropriée. Les modèles stochastiques et leurs paramètres associés doivent être adéquats et basés sur les données des marchés financiers les plus récentes. Les hypothèses qui concernent la volatilité doivent être fondées sur la volatilité historique observée des instruments sous-jacents, si possible.

$$PVFP(t)_{stochastique} = \mathbb{E}_{\mathbb{P} \otimes \mathbb{Q}} \left[ \sum_{u>t} \delta_u CF_u | \mathcal{F}_t \right]$$

avec :

- $\delta_u = e^{-\int_t^u r_s ds}$  : déflateur stochastique en date u ;
- $CF_u$  : flux de trésorerie à la date u.

Généralement, cette valeur est calculée en pratique avec des simulations type Monte-Carlo :

$$PVFP(t)_{stochastique} \approx \frac{1}{N} \sum_{n=1}^N \sum_{u>t} \delta_{u,n} CF_{u,n}$$

### — Etape 3 : calcul du FCRC

C'est le coût d'immobilisation du capital requis. Il s'agit du coût de friction pour les actionnaires, lié à l'immobilisation du capital qui aurait pu être placé. En effet, l'assureur détient un capital requis pour assurer sa solvabilité. Le CFO Forum a défini ce capital requis au niveau du principe 5 dans leur publication sur les principes de calcul de la MCEV :

*«Le capital requis est la valeur de marché des actifs attribués à l'entreprise couverte en sus de celui requis pour couvrir les engagements de la compagnie et dont la distribution aux actionnaires est restreinte.» (CFO forum, 2008)*

En immobilisant ces fonds requis, la compagnie d'assurance perd alors la possibilité de les investir pour réaliser des bénéfices, ce qui entraîne un manque de gain pour les actionnaires.

Il existe plusieurs méthodes pour calculer le coût de capital requis. Une de ces méthodes consiste à le considérer comme la somme actualisée des charges liées au capital requis à savoir les impôts et les coûts de gestion :

$$FCRC = \sum_{t=1}^T \frac{(CR_t \times \theta_t^{impot} \times \theta_t^{rend}) + CR_t \times (1 - \theta_t^{impot}) \times \theta_t^{gestion}}{(1 + \theta_t^{ref})}$$

avec :

- $CR_t$  : capital requis à la date t.
- $\theta_t^{impot}$  : taux d'imposition à la date t.
- $\theta_t^{rend}$  : taux de rendement à la date t.
- $\theta_t^{gestion}$  : taux de gestion à la date t.
- $\theta_t^{ref}$  : taux d'actualisation à la date t.

Dans un cadre *market consistent*, le taux de rendement à la date t peut être assimilé au taux *forward* à 1 an  $fdw_t$ , et le taux d'actualisation peut se calculer de façon à avoir l'égalité suivante :

$$\left( \frac{1}{1 + \theta_t^{ref}} \right)^t = \prod_{i=0}^{t-1} \frac{1}{1 + fdw_i}$$

Quant à l'évaluation du capital requis, il doit nécessairement être au moins égal au capital de solvabilité SCR imposé par le régulateur en Solvabilité II. Il peut être calculé suivant l'approche standard ou selon un modèle interne.

### — Etape 4 : calcul du CRNHR

C'est le coût des risques résiduels non diversifiables. Les risques non répliquables qui n'ont pas été pris en compte dans le calcul de la TVOG et de la PVFP doivent être estimés. Selon [The American Academy of Actuaries \(March 2011\)](#), quelques exemples de ces risques seraient :

- le risque opérationnel ;
- le risque de niveau dans les tables de mortalité ou de rachat ;
- le risque de volatilité pour la longévité.

Le CRNHR représente donc le montant qui comprend l'impact des risques non répliquables financiers et non financiers. Il est souvent déterminé suivant une approche « coût de capital ». Et souvent par analogie au concept de la marge de risque en Solvabilité II, cette quantité est calculée comme étant :

$$CRNHR = \sum_{t>0} \frac{6\% \times SCR_t}{(1 + \theta_t^{ref})^t}$$

Il faut souligner que la méthode d'actualisation de la *value of in-force* doit être cohérente avec les taux des marchés financiers.

Il est tout à fait possible de concevoir d'autres visions de décomposition de la VIF ; par exemple plutôt que de dissocier les éléments FCRC et CRNHR, la formule peut se baser sur une présentation alternative qui combine les deux en une seule composante «coût de capital».

Dans l'ensemble, le calcul de cet indicateur en assurance vie est généralement réalisé à l'échelle du portefeuille. Le calcul doit intégrer un ensemble de polices pour plusieurs raisons, dont la politique de la participation aux bénéfices. Par conséquent, le résultat obtenu constitue un agrégat basé sur une maille beaucoup plus large que celle requise dans le contexte du calcul de rentabilité individuelle ou par *model point*.

## 1.2 Assurance non-vie

### 1.2.1 Caractéristiques de l'assurance non-vie

Le groupe d'assurance de cette étude détient, en plus de sa filiale assurance vie, une filiale qui commercialise les produits non-vie, c'est-à-dire des contrats dont les garanties ne dépendent pas d'événements aléatoires liés à la vie humaine. L'assurance non-vie se répartit en deux catégories principales :

- assurances pour personnes ou dommages corporels, qui ont vocation de protéger des personnes, comme l'assurance maladie par exemple ;
- assurances de biens et de responsabilité.

L'assureur en question commercialise des produits Incendie, Accidents et Risques Divers (IARD). Ce type de contrats est axé uniquement sur la première catégorie de l'assurance non-vie.

Selon l'article R. 321-1 du code des assurances, les opérations de l'assurance non-vie comprennent les branches suivantes, sachant qu'en pratique la majorité des contrats comportent des garanties qui relèvent de plusieurs branches :

1. **Accidents** : y compris les accidents de travail et les maladies professionnelles ;
2. **Maladie** ;
3. **Corps de véhicules terrestres (autres que ferroviaires)** : tout dommage subi par les véhicules terrestres à moteur ou non automoteurs ;
4. **Corps de véhicules ferroviaires** ;
5. **Corps de véhicules maritimes lacustres et fluviaux** ;
6. **Marchandises transportées** : tout dommage subi par les marchandises quelque soit le moyen de transport ;
7. **Incendie et éléments naturels** : tout dommage subi par par les biens (autres que 3, 4, 5,6 et 7) ;

8. **Autres dommages aux biens** : tout dommage subi par les biens (autres que 3, 4, 5,6 et 7) et causé par la grêle ou la gelée, ainsi que tout par tout événement, tel le vol, autre que ceux compris dans la branche 8.10 ;
9. **RC véhicules terrestres** : toute responsabilité résultant de l'emploi des véhicules terrestres automoteurs
10. **RC véhicules aériens** : Toute responsabilité résultant de l'emploi des véhicules aériens ;
11. **RC véhicules maritimes lacustres et fluviaux** : toute responsabilité résultant de l'emploi des véhicules fluviaux, lacustres et maritimes ;
12. **RC générale** : toute autre responsabilité autre que celle mentionnées précédemment ;
13. **Caution et crédit** : caution directe ou indirecte, insolvabilité, crédit hypothécaire, . . . ;
14. **Pertes pécuniaires diverses** : risque d'emploi, insuffisance de recettes, pertes de bénéfices, . . . ;
15. **Protection juridique** ;
16. **Assistance** : assistance aux personnes en difficulté.

En assurance non-vie, les garanties sont généralement de type indemnitaire et le montant du sinistre est déterminé *a posteriori*. Le principe indemnitaire est un principe fondamental selon lequel la prestation de l'assureur ne peut en aucun cas dépasser le préjudice réel, subi ou causé par l'assuré (Bocquaire & Camus-Paradis, 2022).

La répartition du marché français des assurances entre la branche vie et non-vie varie d'une année à l'autre en fonction de plusieurs critères comme les conditions économiques ou les préférences des consommateurs. Cependant, de manière générale, l'assurance vie représente une part significative du marché assurantiel français, étant donné la popularité des produits d'assurance vie en tant que produit d'épargne et de protection.

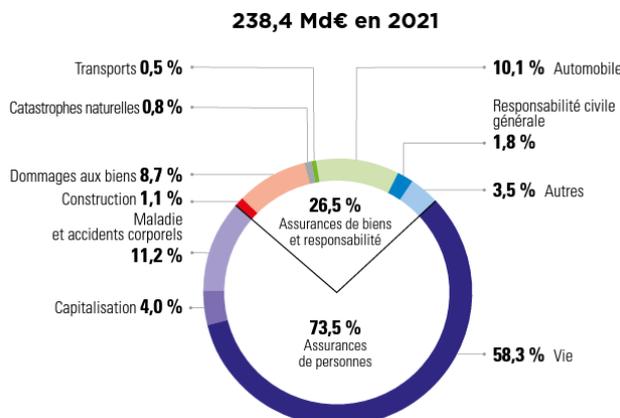


FIGURE 1.2 – Répartition des cotisations sur le marché français en 2021. Source : France Assureurs

D'après le rapport annuel de France Assureurs pour l'année 2021, les cotisations des assurances de bien et de responsabilité s'établissent à 63,2 Md€ en plus des 26,6 Md€ pour l'assurance maladie et accidents corporels :



d'une source supplémentaire qui réside dans sa composante investissement. L'inflation peut également peser sur la sinistralité de certaines activités couvertes par l'assurance dommage, tout comme les montants exorbitants des sinistres liés aux catastrophes naturelles. En se basant sur le résultat net rapporté aux fonds propres, le graphique ci-dessous affirme ce constat :

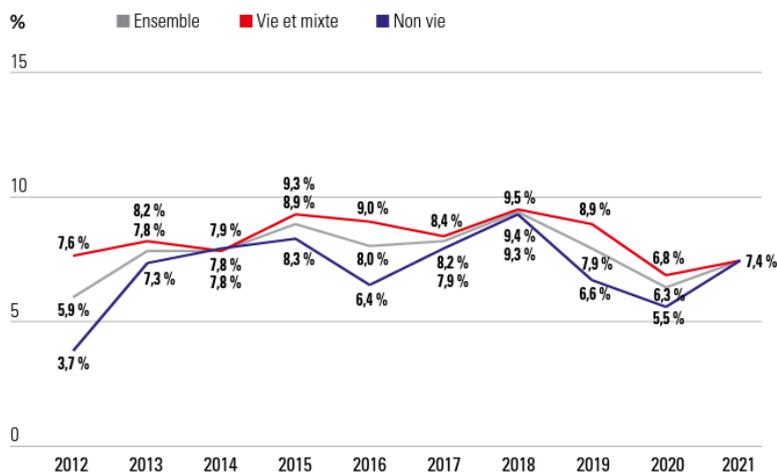


FIGURE 1.4 – Rentabilité des assureurs français. Source : France Assureurs

### 1.2.2 Indicateur de rentabilité

Dans la même logique que la première partie, la filiale IARD cherche à déterminer les marges de rentabilité de ses clients, dite « la valeur client ». C'est une marge individuelle qui renseigne le gain réalisé par la compagnie d'assurance sur chaque assuré.

Cet indicateur sera l'élément clé dans l'analyse de la profitabilité du portefeuille non-vie. Il permettra à l'assureur de déterminer d'éventuelles déformations de portefeuille dans le temps sur des segments déficitaires. Il peut également être utilisé pour d'autres raisons que le pilotage de la rentabilité, comme la défense du portefeuille : l'assureur pourra dès lors détecter les clients à bonne valeur et leur proposer des garanties supplémentaires pour améliorer leur satisfaction et optimiser leur fidélité.

Cette valeur client est surtout primordiale en assurance dommage où la concurrence se resserre sur le marché. La plupart des acteurs de l'assurance IARD font face à un marché qui stagne. De plus, la concurrence s'intensifie non seulement entre les acteurs traditionnels, mais aussi avec l'émergence de nouveaux assureurs digitaux : les Assurtech.

Ainsi, avec une croissance organique minimale, la majorité des assureurs IARD se retrouvent dans l'obligation de gagner et de conserver les clients les plus attractifs, ceux avec les plus grandes valeurs client, pour réaliser une croissance rentable et se démarquer dans ce marché compétitif.

Pour discuter de la santé financière en secteur de l'assurance non-vie, les acteurs ont tendance à communiquer le ratio combiné (CR) qui se calcule comme suit :

$$CR = \frac{\text{Coûts sinistres réglés} + \text{commissions} + \text{frais généraux}}{\text{Primes versées par l'assuré}}$$

Le ratio combiné fournit une vision plus complète que le *loss ratio*. Toutefois, c'est un ratio qui est habituellement calculé au niveau de la branche pour évaluer sa performance globale, alors que le but du projet « valeur client » est de calculer une marge individuelle par assuré. Il est tout à fait possible

de le calculer individuellement selon la disponibilité et l'exactitude des données, mais le fait que cette mesure soit un pourcentage rend sa manipulation difficile et les résultats seront plus volatiles si analysés au niveau individuel.

Un autre indicateur qui pourrait être utile dans le cadre de cette étude est la *Customer Lifetime Value* (CLV). Cette notion se rapproche du concept de la PVFP. Elle correspond à la valeur actualisée des flux de trésorerie générés par un client ou une cohorte de clients tout au long de la durée de vie de leurs contrats.

Pour rester homogène avec ce qui a été fait pour l'assureur vie, l'évaluation de la marge de rentabilité en assurance non-vie, détaillée dans la partie ci-dessous, se rapprochera de la logique de calcul en vie, mais en considérant les différences du fonctionnement des deux types de produits.

### 1.2.3 Calcul de l'indicateur de rentabilité

La marge individuelle en assurance non-vie se base sur les caractéristiques des assurés, supplémentées des informations des contrats souscrits par le passé ainsi que des primes, des sinistres et des frais. Globalement, elle se calcule comme la somme d'une marge technique qui concerne le gain réalisé par l'acquisition des primes et le règlement des sinistres, d'une marge non technique qui provient des frais généraux, auxquels on retire les coûts des frais d'acquisition et d'immobilisation de la marge de solvabilité règlementaire :

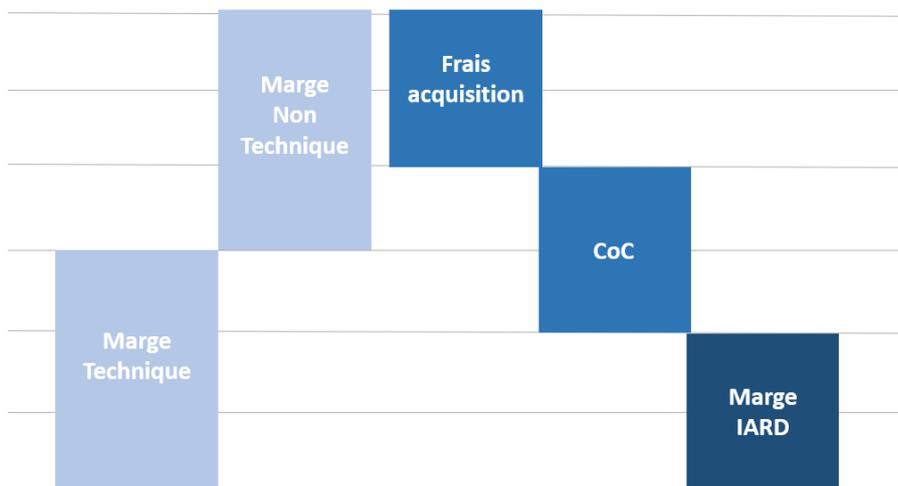


FIGURE 1.5 – Décomposition de la marge IARD

Pour un portefeuille d'assurés, la procédure de calcul de la marge non-vie se déroule selon les étapes suivantes :

— **Etape 1 : table de référence et initialisation des *model points***

Cette étape consiste à créer une table détaillée à partir du portefeuille existant avec les variables retenues et leurs modalités choisies. Des *model points* sont ensuite initialisés avec les valeurs du portefeuille stock.

— **Etape 2 : bases techniques**

Le calcul de la marge se base sur plusieurs bases techniques comme le montant des primes, la charge des sinistres ou le nombre de sinistres. L'objectif de cette étape est d'obtenir les montants techniques et non techniques de l'année de projection afin de calculer la marge technique unitaire.

— **Etape 3 : initialisation des frais généraux et projection**

Cette étape spécifie les différents types de frais qui rentrent en jeu : frais d'acquisition, frais de gestion, ...

— **Etape 4 : autres paramètres**

D'autres paramètres sont à considérer dans le modèle qui permettent le calcul de la marge comme le taux d'inflation, le taux d'actualisation, le taux d'imposition ou le coût de portage du capital (CoC).

— **Etape 5 : table de marge unitaire de MP**

Nous pouvons désormais calculer la marge unitaire pour chaque *model point* pour l'année en cours. Il s'agit de la valeur actuelle probable des résultats futurs de l'assuré pendant toute la durée résiduelle de ses contrats. Pour un assuré  $i$  qui détient un produit  $j$ , le résultat de l'année  $t$  s'écrit :

$$Resultat_{t,j}^i = (MT_{t,j}^i - ACQ_{t,j}^i + MNT_{t,j}^i) \times (1 - TX_{imp}) - CoC_{t,j}^i$$

Avec

- $MT_{t,j}^i$  : marge technique ;
- $MNT_{t,j}^i$  : marge non technique ;
- $ACQ_{t,j}^i$  : frais d'acquisition ;
- $TX_{imp}$  : taux d'imposition ;
- $CoC_{t,j}$  : coût du capital immobilisation .

La marge technique est la somme de la prime acquise moyenne annuelle diminuée de la charge des sinistres moyenne annuelle en vision ultime :

$$MT_{t,j}^i = \sum_{t=0}^n PA_{t,j}^i - CTF_{t,j}^i$$

La marge non technique, quant à elle, se calcule comme le placement des charges de sinistres diminué par les frais de gestion.

$$MNT_{t,j}^i = -(TFG_j \times PA_j^i) + TPF_{t,j}^i \times CTB_j$$

Avec

- $TFG_j$  : taux de frais de gestion ;
- $PA_j^i$  : assiette des primes acquises ;
- $TPF_{t,j}^i$  : taux de produit des placements net de frais de gestion ;
- $CTB_j$  : charge de sinistre moyenne annuelle en vision dossier/dossier.

Ainsi, la valeur de la marge d'un assuré  $i$  se détermine comme étant la somme de ses résultats pour tous les produits souscrits et ce sur sa durée de maintien dans le portefeuille de l'assureur.

$$MargeIARD_i = \sum_{t=0}^n \frac{1}{(1+r)^t} \sum_{j=1}^J Resultat_{t,j}^i$$

— **Etape 6 : loi de comportement**

Les lois de comportement permettent de faire évoluer les équipements des *model points* pour chaque année de projection. Elles prennent la forme de matrices qui contiennent les probabilités de passage d'une combinaison d'équipements à une autre.

— **Etape 7 : table des expositions MP projetées**

A l'aide des tables de survies et de comportement, le nombre d'assurés pour chaque année est estimé.

— **Etape 8 : table de marges projetées**

Cette étape est obtenue à l'aide d'un calcul matriciel : le produit de la matrice des marges unitaires et des expositions MP projetées.

— **Etape 9 : sensibilité et *stress test***

Pour vérifier la robustesse des hypothèses utilisées pour la modélisation de la marge, des travaux de sensibilité et de *stress test* ont été menés.

Après avoir calculé les marges des assurés à l'instant  $t=0$ , l'assureur s'intéresse à l'estimation de la marge future sur les 31 prochaines années. La projection de cette marge est particulièrement judicieuse car elle permet à l'assureur d'identifier les profils les plus rentables dans le futur, ce qui lui permet ainsi de mieux orienter son plan stratégique.

Cependant, et vu le nombre des paramètres en question, la projection de cette variable sur un horizon lointain est une tâche compliquée qui nécessite un temps de mise en place important, d'où le choix d'adopter un modèle markovien multi états pour estimer la marge non-vie.

#### 1.2.4 Estimation avec un modèle markovien multiétats

Les modèles multi-états sont des modèles de processus, décrivant par exemple le cycle de vie d'un individu, qui occupe à tout moment l'un des états possibles. Ils sont définis par des états (représentés par les cases de la figure 1.6) et des transitions entre eux (représentées par des flèches). Ces modèles sont souvent utilisés en épidémiologie et en biostatistique pour modéliser la progression d'une maladie en décrivant la probabilité pour un individu de passer d'un état pathologique à un autre.

Le modèle le plus simple contient deux états, par exemple vivant et mort ou sain et malade, et est communément appelé le modèle de mortalité. Ce type de modèle peut être analysé à l'aide de méthodes d'analyse de survie. Quant à la nature des états, ils peuvent être :

- transitoires, c'est-à-dire que les individus peuvent y entrer et en sortir ;
- ou absorbants, c'est-à-dire que les individus n'en sortent jamais une fois qu'ils y sont entrés (par exemple, morts dans un modèle de mortalité).

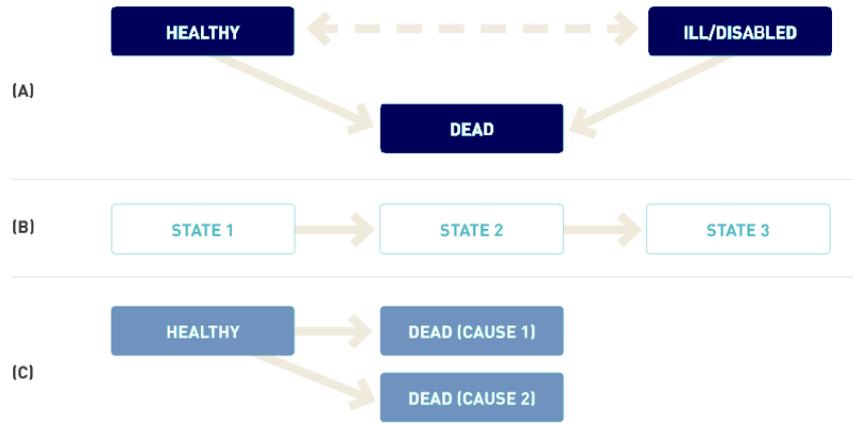


FIGURE 1.6 – Modèle markovien à 3 états

Les modèles de Markov sont le plus souvent utilisés en raison de leur simplicité. Un modèle à plusieurs états est considéré comme markovien s'il suppose que la probabilité de transition vers un nouvel état ne dépend que de la valeur actuelle du modèle.

En reprenant les notations de [Pierre \(2018\)](#), soit  $\{X(t), t \in [0, +\infty[ \}$  un processus de Markov à temps continu et à espace d'états fini  $S = 1, \dots, s$ . On définit les probabilités de transition entre l'instant  $s$  et  $t$  :

$$P(s, t) = p_{hj}(s, t)$$

Où  $p_{hj}(s, t)$  est la probabilité de passer de l'état  $h$  à l'instant  $s$  vers l'état  $j$  à l'instant  $t$  :

$$p_{hj}(s, t) = \mathbb{P}(X(t) = j | X(s) = h)$$

On définit les intensités de transition entre les états  $Q(t) = \alpha_{hj}(t)$  de façon que :

$$\begin{cases} \alpha_{hj}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{hj}(t, t+\Delta t) - p_{hj}(t, t)}{\Delta t} & \text{si } h \neq j \\ \alpha_{hh}(t) = -\sum_{h \neq j} \alpha_{hj}(t) \end{cases}$$

Il existe plusieurs approches pour déterminer les intensités de transition selon s'il s'agit d'un modèle de Markov homogène ou non homogène. Un modèle markovien est dit homogène si l'intensité de transition est indépendante de la durée de suivi, c'est-à-dire :

$$\alpha_{hj}(t) = \alpha_{hj}$$

Cette classe de modèles permet une approche extrêmement flexible qui peut modéliser presque n'importe quel type de données longitudinales. En effet, les modèles de Markov sont beaucoup plus simples que les autres modèles du point de vue des probabilités, ce qui simplifie l'évaluation de la vraisemblance. Les modèles markoviens multi-états peuvent alors être utilisés pour estimer les intensités et probabilités de transition ainsi que le temps de séjour moyen dans un état donné.

Pour notre cas d'étude, à savoir l'estimation de la marge de rentabilité non-vie future, le modèle considère que chaque état correspond à une combinaison des caractéristiques de l'assuré. Par exemple, chaque état peut correspondre au croisement des variables suivantes : **age** et **équipement**. Soit le cas

d'un assureur qui ne propose que deux produits IARD : PROD1 et PROD2. La figure ci-dessous illustre un schéma simplifié du fonctionnement du modèle de l'assureur avec tous les états possibles, les probabilités de transition et les marges calculées pour chaque état.

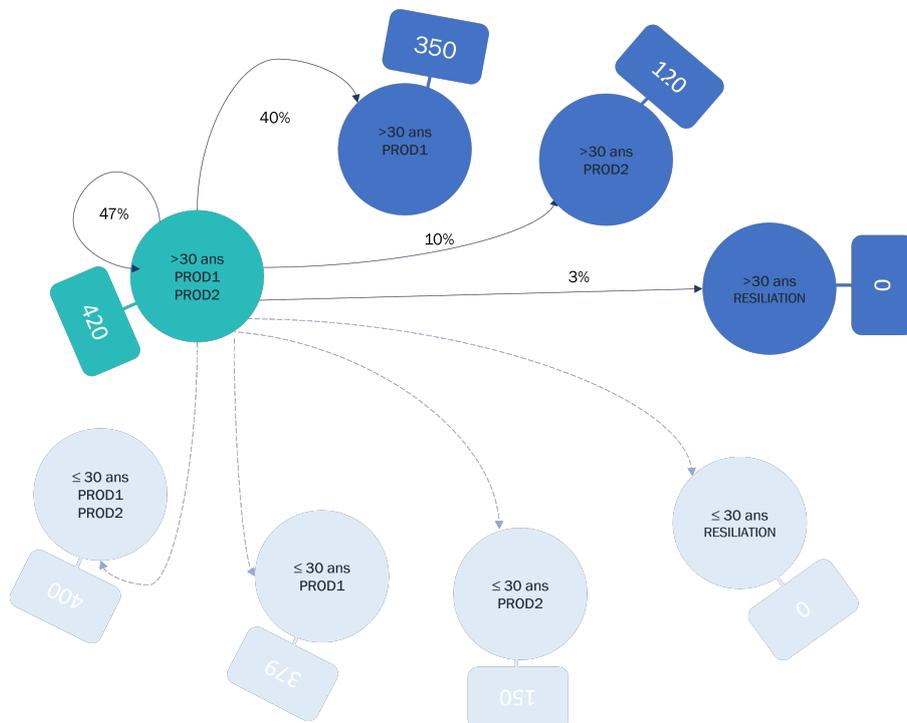


FIGURE 1.7 – Schéma simplifié du modèle markovien de l'assureur non-vie

Les probabilités de transition ainsi que les montants de marge estimée sont calibrés sur l'expérience passée de l'assureur et sont ajustés en permanence. Les marges qui correspondent à tous les états possibles sont préalablement calculées en utilisant la formule de calcul précédemment présentée. Cela dit, si les hypothèses de base changent, les résultats obtenus par le modèle markovien ne seront plus adaptés. C'est la raison pour laquelle, une revue des hypothèses et un recalcul des marges sont effectués tous les deux ans afin de calibrer en continu le calcul.

Ce procédé est appliqué séparément pour chaque produit, et fournit alors une valeur par produit. Un produit regroupe un ou plusieurs contrats jugés suffisamment similaires pour les analyser ensemble. Par conséquent, nous obtenons une valeur par individu, par échéance et par produit.

Pourtant, l'utilisateur du modèle markovien aura du mal à expliquer les marges de rentabilité, puisque ce modèle ne met pas en lumière les déterminants de la valeur finale de la marge. Mais comme il est évident qu'il existe une forte corrélation entre les caractéristiques des assurés et la marge fournie par l'assureur, des études se sont alors basées sur cette corrélation pour essayer de cerner les déterminants de la marge non-vie et mieux comprendre les résultats obtenus. C'est le but de la partie non-vie du deuxième chapitre, où nous allons chercher un modèle plus interprétable pour reconstituer la marge future d'un portefeuille d'assurés non-vie. L'intégration de ce nouveau modèle statistique a pour objectif de fournir davantage d'informations à l'utilisateur qui souhaite mieux interpréter les marges des assurés non-vie.

# Chapitre 2

## Mise en pratique

Dans ce chapitre, nous chercherons à mettre en pratique les concepts précédemment présentés. Sa structure se répartit alors en deux sous parties : une concernant l'assureur vie, où nous allons chercher à calculer la marge de rentabilité individuelle basée sur la VIF. Et une deuxième partie pour l'assureur IARD, qui estime la marge de rentabilité individuelle, actuellement calculée par le modèle markovien, avec des modèles de régression statistique.

### 2.1 Calcul de la marge vie

Cette première partie a pour but de calculer les marges de rentabilité pour la filiale vie du groupe d'assurance. D'après le premier chapitre, nous avons choisi d'assimiler la marge de rentabilité vie à la VIF. Le calcul de cette quantité s'effectuera par un outil interne de Prim'Act : SimBEL. Tout d'abord, nous veillerons à présenter les données exploitées dans cette analyse, pour ensuite expliquer le fonctionnement du calcul derrière l'outil en précisant les hypothèses et input utilisés. La valeur finale de la VIF doit être répartie sur les *model points* ; nous discuterons alors des méthodes d'allocation suivantes : proportionnelle, marginale et Shapley. Les résultats obtenus seront exposés et discutés en dernier lieu.

#### 2.1.1 Données et hypothèses

La filiale propose des produits vie qui mettent en place des dispositifs d'épargne retraite. C'est ainsi que l'assureur commercialise des plans d'épargne destinés aux particuliers mais aussi aux entreprises. Nous nous focaliserons uniquement sur le premier produit dans cette étude. Le portefeuille de ce dernier se partage entre deux types de segments :

- les contrats en phase de constitution : pendant cette phase, l'adhérent verse des primes sur son contrat pour constituer son épargne ;
- les contrats en phase de rente : à la date du départ à la retraite, l'adhérent commence à recevoir des rentes. Les assurés peuvent également choisir l'option de sortie en capital.

L'assureur a communiqué une base de données pour chacun des segments respectifs. La première base de données « Epargne » se compose de 125 *model points*, alors que la deuxième « Rente » en compte 212.

Le contexte d'étude suivant portera sur le segment des clients en phase de constitution, puisque cela se compare à un produit d'épargne classique. La base de données « Epargne » contient les caractéristiques suivantes de chaque *model point* :

- identifiant ;
- sexe ;
- age ;
- ancienneté ;
- nombre de polices ;
- cotisations ;
- provisions mathématiques ;
- taux de sortie en rente.

Cette base de données comporte des *model points* ayant des cotisations nulles mais qui ont pourtant une provision mathématique non nulle. Il s'agit des assurés âgés qui n'ont toujours pas liquidé leur retraite. Ils sont gérés ainsi dans le modèle interne de la compagnie :

- de 62 à 70 ans : sortie en fonction de la cadence de liquidation ;
- à partir de 70 ans : sortie dès la première année.

Le graphique suivant illustre le nombre de polices par *model point*. On peut observer la présence de deux courbes ayant des formes similaires ; en effet le premier groupe représente les *model points* pour les hommes et le second celui des femmes. La distribution des nombres de polices suit donc la même structure chez les hommes que chez les femmes.

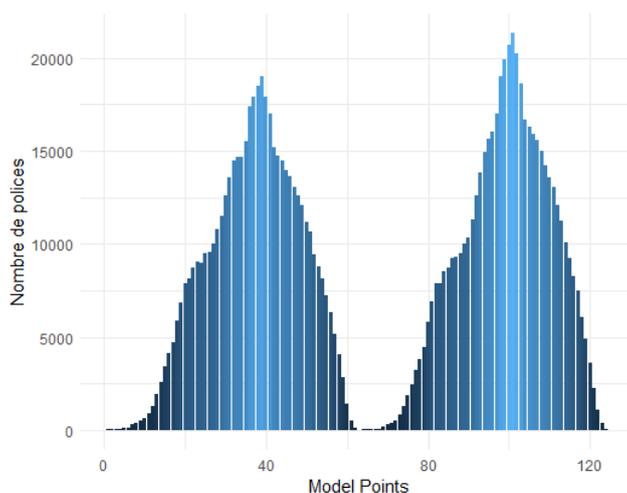


FIGURE 2.1 – Répartition des polices par *model point*

En effet, chaque *model point* se caractérise par le croisement des variables **age**, **sexe** et **anciennete**. Cependant la variable **anciennete** a été fixée à 1 an pour tous les individus de cette base de données : il s'agit d'une hypothèse de projection liée à la frontière des contrats sur SimBEL. Ce qui fait que seules les deux autres variables déterminent un *model point* dans ce cas : Les premiers 62 *model points* sont ceux des assurés hommes classés par ordre décroissant de l'âge, les 63 restant sont ceux des femmes classés selon le même critère. Les *model points* les plus représentés dans ce portefeuille sont les individus avec un âge entre 40-44 ans pour les hommes, et entre 39-44 ans pour les femmes :

Sexe	Ancienneté	Age	Nombre de Polices
F	1	41	21 318
F	1	42	20 715
F	1	40	20 266
F	1	43	19 936
F	1	44	19 029
M	1	41	18 988
F	1	39	18 626
M	1	42	18 464
M	1	43	17 903
M	1	40	17 894
M	1	44	17 377

TABLE 2.1 – Les 10 *model points* les plus fréquents de l'assureur vie

La répartition des âges par le nombre des contrats est la même chez les deux sexes : la majorité des assurés ont un âge entre 29 et 51 environ : 69% des femmes et 66% des hommes. Avec un minimum de 18 ans et un âge maximum de 79 ans :

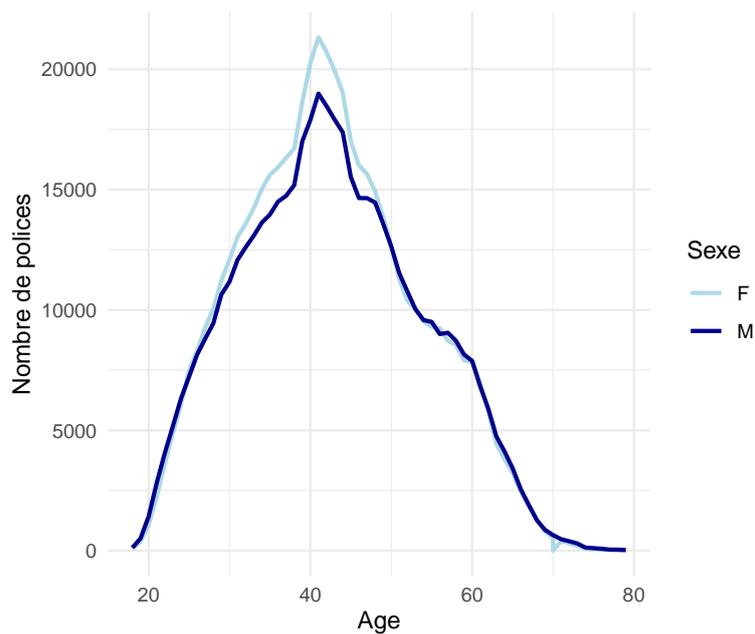


FIGURE 2.2 – Répartition des polices par âge

Quant aux provisions mathématiques, elles augmentent jusqu'à un certain âge (58 ans), et ensuite connaissent une tendance baissière, que ce soit chez la population masculine ou féminine. Le graphique suivant met en contraste le nombre de police par âge (histogramme) et le montant des encours (courbe orange) chez les deux populations. On remarque alors que jusqu'à l'âge de 41 ans, le nombre des polices et le montant des provisions augmentent. Entre 41 et 60 ans, le nombre de polices diminue mais le montant des provisions continue à augmenter jusqu'à 60 ans, où les deux variables diminuent.

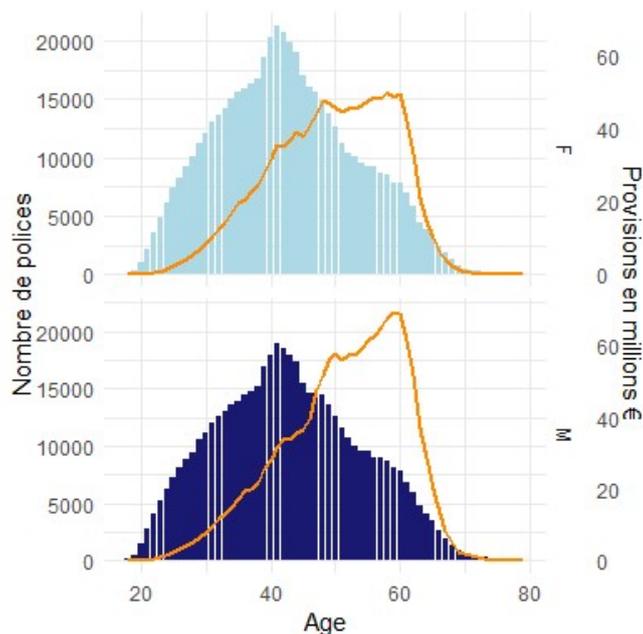


FIGURE 2.3 – Répartition des provisions mathématiques et du nombre de polices par âge

La compagnie a également fourni :

- un fichier des **flux du passif** qui contient : le flux économique avant et après plusieurs chocs appliqués à la longévité, la mortalité et le rachat. Ainsi que les charges IT avant et après les mêmes chocs.
- une **base d’actifs transparisés** où chaque fond est détaillé ligne par ligne. Elle contient les montants, les caractéristiques financières, les émetteurs et les notations des différents types d’actifs. Le portefeuille d’actifs de cet assureur est bien diversifié : actions, immobilier, monétaire, OAT (obligations souveraines), obligations d’entreprises, Actions OPCVM, OTF et de la private equity ;
- les **rappports réguliers** réglementaires pour l’année 2020 et 2021 des résultats S2 communiqués au régulateur.

### 2.1.2 Outil interne

Prim’Act dispose d’un outil interne pour le calcul des provisions *Best Estimate* (BE) : SimBEL. Il s’agit d’un modèle de simulation Monte Carlo s’appuyant sur une projection d’un canton (actif et passif) permettant l’évaluation des provisions BE d’un contrat d’épargne et retraite français en euros. L’outil prend la forme d’un package développé sous le logiciel open source R. A l’addition du BE central de l’assureur, il permet de produire les BE qui correspondent aux principaux scénarios de chocs marché et souscription, au sens de la formule standard de la réglementation Solvabilité II.

Pour assurer le bon fonctionnement de l’outil, il est impératif que le paramétrage soit effectué selon l’architecture suivante :

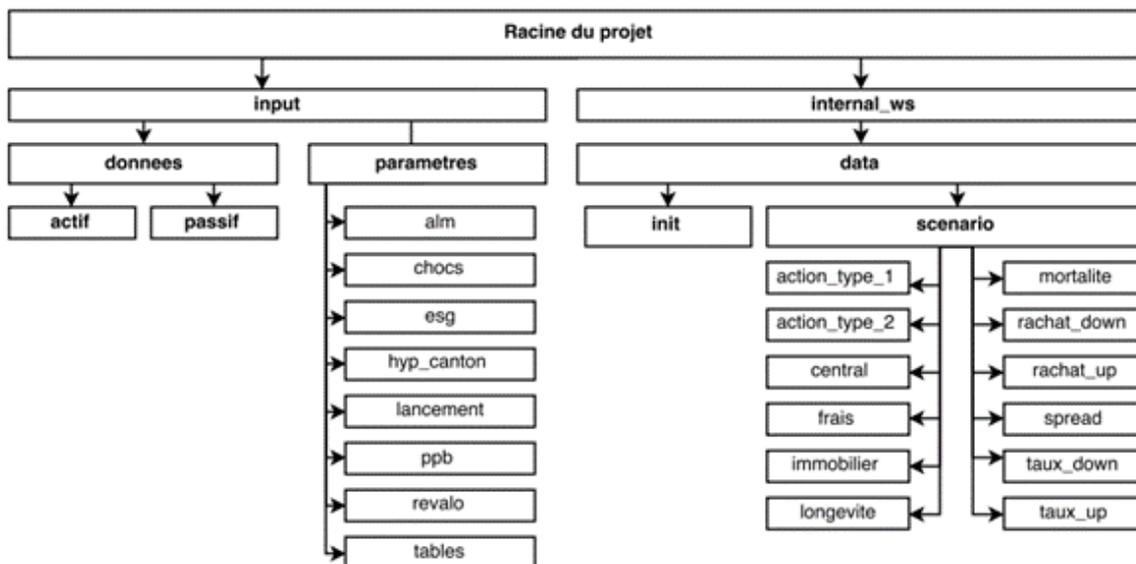


FIGURE 2.4 – Architecture de SimBEL

## — Données :

## 1. actif :

il faut préciser le portefeuille action, immobilier et obligation ainsi qu'un fichier de trésorerie et un autre pour les frais financiers. On se base sur la base d'actifs transparisée communiquée par l'assureur vie pour renseigner cet input.

## 2. passif :

cet onglet contient la table présentée ci-dessus des données *model points*, ainsi que les frais sur passif : frais sur primes, frais sur prestations et les frais sur encours. Cette information est tirée d'une matrice de frais partagée par la compagnie.

## — Paramètres :

cette partie de l'arborescence concerne les hypothèses techniques et financières qui se cachent derrière le calcul dont les plus importants :

## 1. alm :

une table pour renseigner la stratégie d'allocation cible de l'actif : pourcentage de répartition entre actions, immobilier et obligations.

## 2. chocs :

une liste de fichiers pour les chocs à appliquer pour chacun des risques suivants : devise, spread, souscription, mortalité, longévité, rachat et inflation.

## 3. esg :

le générateur des scénarios économiques GSE (ou en anglais *Economic Scenario Generator* ESG) est un ensemble de modèles (taux, actions, spreads de crédit, inflation, etc.) d'occurrences aléatoires des marchés financiers, qui ont un impact sur la valeur et la performance des actifs de l'assureur. Pour être en cohérence avec la réglementation, cet outil s'appuie sur la probabilité risque neutre pour générer ses résultats.

Dans cette branche de l'outil, il faut donc mettre à jour les fichiers générés par le générateur des scénarios économiques. Il s'agit des fichiers contenant les indices des actions, de l'immobilier et de l'inflation, ainsi que la courbe des taux et les déflateurs utilisés par l'outil.

## 4. lancement :

on précise les paramètres de lancement de l'outil, à savoir le nombre de simulations et le nombre d'années de projection (ici 31 ans).

## 5. tables :

on précise les tables de mortalité utilisées, notamment les tables TGF05-TGH05 ici. Il s'agit d'un ensemble de tables par génération. Elles comprennent en fait autant de tables de mortalité qu'il y a d'années de naissance comprises entre 1886 et 2005. On présente également les tables de lois de rachat. Dans ce cas de figure et du fait des dispositions contractuelles du produit, les rachats sont faibles (de l'ordre de 0.3%) et seront donc négligés dans les projections.

Chaque *model point* est projeté jusqu'au départ en retraite comme un contrat d'épargne en euro. À ce moment, l'épargne acquise fournit, pour chaque scénario économique, un capital constitutif de rente. Pour le calcul de la rente à servir à partir de cette date, on utilise la moyenne empirique, sur l'ensemble des scénarios économiques, de ces capitaux constitutifs.

Finalement, après avoir renseigné les inputs nécessaires, l'outil est en mesure de produire les flux de trésorerie requis pour le calcul du BE. En revanche, l'étude de rentabilité cherche à calculer la VIF et non pas le BE. En réalité, les mêmes flux générés par SimBEL peuvent être employés dans le calcul de plusieurs autres métriques en assurance vie dont la VIF. C'est la raison pour laquelle, nous avons tiré profit de l'outil existant pour valoriser la rentabilité.

La PVFP déterministe est alors calculée suivant un scénario central. L'outil calcule ensuite la PVFP stochastique sur la base des scénarios économiques fournis par l'assureur et mis en input. On conclut dès lors facilement la valeur de la TVOG, qui représente la différence entre la PVFP stochastique et la PVFP déterministe. Pour le calcul des deux composantes CRNHR et FCRC, nous allons procéder par l'approximation suivante (où *Risk Margin* (RM) représente la marge pour risque sous la directive S2) :

$$CRNHR + FCRC = RM$$

Cette estimation semble logique si on rapproche le bilan S2 et le bilan économique MCEV (Planchet, 2023) :

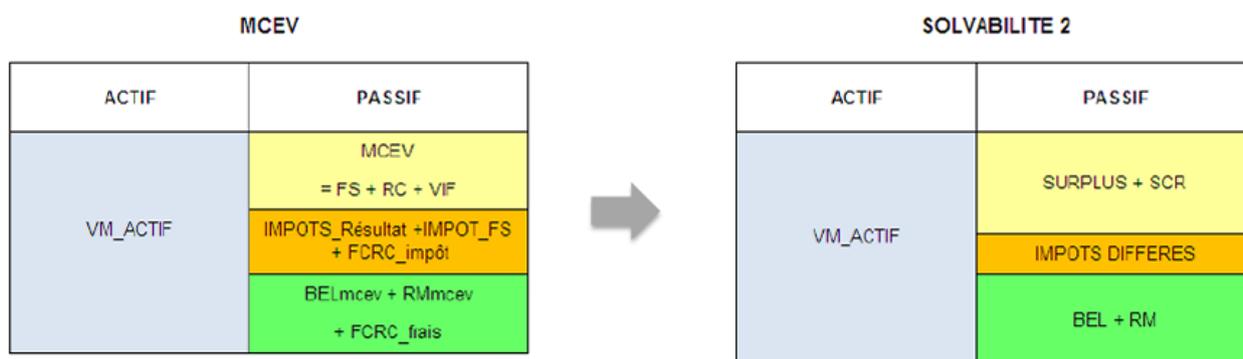


FIGURE 2.5 – Passage du bilan MCEV au bilan S2

Pour le calcul de RM, nous allons utiliser les rapports prudentiels réglementaires communiqués par l'assureur et tirer un pourcentage moyen du RM par rapport au BE, que nous allons appliquer par la suite au BE calculé par l'outil pour avoir le montant de la marge pour risque.

Le tableau suivant récapitule les résultats obtenus en million d'euros :

VIF	10,5
PVFP	125,9
TVOG	67,2
FCRC + CRNHR	48,2

TABLE 2.2 – Résultats de la VIF en million d'euros

Le résultat de la VIF obtenue est alors de 10,5 M€. Ce montant représente une mesure agrégée de la rentabilité du portefeuille de cet assureur vie. Pour répondre à la problématique du mémoire, et pour des raisons de comparaison avec les marges de rentabilité en non-vie, il va falloir chercher à estimer une marge de profitabilité pour chaque segment. En d'autres termes, chercher une méthode d'allocation pour répartir la valeur totale de la VIF sur les *model points*.

### 2.1.3 Allocation de la VIF

L'allocation équitable d'un capital présente un enjeu de taille pour différents acteurs. Par exemple :

- les assureurs qui souhaitent déterminer le capital de solvabilité à allouer à chaque segment en Solvabilité 2 ;
- les investisseurs qui cherchent à répartir leurs ressources financières entre différentes opportunités d'investissement.

L'approche suivante peut donc être réadaptée dans plusieurs situations. Dans ce cas de figure, le problème d'allocation peut être décrit comme suivant :

Soit un portefeuille d'assurance composé de  $n$  membres indexés par  $i = 1, \dots, n$ . Ici, le capital à allouer est la VIF. On nommera segment, la maille sur laquelle est alloué le capital, dans ce cas un segment  $i$  est assimilé à un model point déterminé par l'intersection des variables : **age**  $\times$  **sexe**  $\times$  **anciennete**. Nous essaierons de trouver le poids  $w_i$  à attribuer à chaque segment tel que la marge de rentabilité de chaque *model point* vérifie  $VIF_i = w_i \times VIF$ . L'allocation du capital sera réalisée selon l'une des méthodes suivantes :

- **Méthode proportionnelle :**

Le processus d'allocation nécessite de calculer la contribution de chaque segment au montant de la rentabilité globale. La méthode proportionnelle vise à allouer au segment  $i$  une proportion d'une métrique à définir.

La vision la plus simple, dans le cas d'un produit d'épargne en euro, est de répartir la valeur totale de la VIF selon l'encours de chaque *model point*.

$$w_i = \frac{PM_i}{PM}$$

où  $PM_i$  est le montant de la provision mathématique du *model point*  $i$ , et  $PM$  le montant total des provisions mathématiques du portefeuille.

L'avantage de cette méthode est qu'elle est intuitive et simple à mettre en œuvre. Cependant, elle ne tient pas compte de la dépendance entre les risques. La police qui possède la plus grande provision mathématique ne contribue pas nécessairement le plus à la VIF, ceci dépend d'autres facteurs comme la probabilité de sortie de ce segment par exemple.

— **Méthode marginale :**

Cette méthode repose sur l'évaluation de l'impact marginal de chaque segment. Elle considère comme poids une proportion qui se réfère à la contribution marginale. Cependant, cette méthode ne considère que la contribution marginale de chaque segment à la population, et ignore l'effet des sous-ensembles des joueurs. En pratique, l'impact marginal de chaque *model point* est la variation de la VIF après élimination de ce segment de la base de données :

$$w_i = \frac{VIF - VIF_{\setminus\{i\}}}{\sum_j VIF - VIF_{\setminus\{j\}}}$$

où  $VIF_{\setminus\{i\}}$  représente la valeur de la VIF sans tenir compte du *model point*  $i$ .

— **Méthode de SHAPLEY :**

La méthode de SHAPLEY est une méthode issue de la théorie des jeux coopératifs. Nous commencerons par reprendre la définition figurant dans « A course in game theory » ([Osborne & Rubinstein, 1994](#)) pour préciser ce que représente un jeu coopératif. Il s'agit de la donnée d' :

- un ensemble fini de joueurs,  $N = \{1, \dots, n\}$  ;
- une application  $\nu : \mathcal{P}(N) \rightarrow \mathbb{R}$ , avec  $\nu(\emptyset) = 0$  et  $\nu$  sur-additive, *i.e.*

$$\forall S, T \in \mathcal{P}(N), S \cap T = \emptyset \implies \nu(S) + \nu(T) \leq \nu(S \cup T)$$

Un élément  $S$  de  $\mathcal{P}(N)$  est appelé une coalition du jeu, et le nombre  $\nu(S)$  représente le gain de la coalition dans le jeu. La fonction  $\nu$  est aussi appelée fonction caractéristique du jeu, et le jeu noté  $(N, \nu)$ . Le jeu dans ce cas est le calcul de la rentabilité pour  $n$  *model points* où le gain à partager est le montant de la VIF.

Le caractère sur-additif de  $\nu$  est assuré dans le cas de l'assurance par le principe de mutualisation.

Une imputation est une fonction  $\bar{x}$  qui à un jeu  $(N, \nu)$  associe un vecteur gain  $(x_1, \dots, x_n) \in \mathbb{R}^n$  tel que :

$$\begin{cases} \sum_{i=1}^n x_i = \nu(N) \\ x_i \geq \nu(i) \end{cases}$$

Le nombre  $x_i$  représente le gain pour le joueur  $i$ . On note  $I(\nu)$  l'ensemble des imputations pour ce jeu. La notion d'imputation est comparable à la notion de fonction de coût. La valeur de Shapley représente donc l'unique imputation qui vérifie les axiomes suivants :

- **Symétrie :** Les joueurs qui contribuent de façon égale, recevront un gain égal. Ce qui veut dire que pour des joueurs  $i$  et  $j$  interchangeable, nous avons la relation suivante :

$$\forall S \subset N \setminus \{i, j\}, \Delta(i, S) = \Delta(j, S) \implies SHAP_i = SHAP_j$$

- **Dummy player** : Un joueur qui n'apporte aucun gain à une collaboration, ne recevra aucun gain en contrepartie. Si la contribution marginale d'un joueur à n'importe quelle coalition est constante égale à  $\nu(i)$ . Alors :

$$SHAP_i = \nu(i)$$

- **Additivité par rapport aux jeux** : Pour tout jeu  $(N, \nu_1)$  et  $(N, \nu_2)$  on doit avoir :

$$SHAP(N, \nu_1 + \nu_2) = SHAP(N, \nu_1) + SHAP(N, \nu_2)$$

où le jeu  $N(\nu_1 + \nu_2)$  est défini par  $(\nu_1 + \nu_2)(S) = \nu_1(S) + \nu_2(S)$  quel que soit  $S \subset N$ .

En raison de ces propriétés souhaitables exprimées par ces axiomes, la valeur de Shapley a été initialement utilisée pour répondre à la question de la répartition équitable des gains d'une coalition entre les différents joueurs d'un jeu coopératif. Aujourd'hui, elle est employée pour résoudre les problématiques d'allocation de capital. De plus, cette méthode a connu un grand nombre d'extensions et de variations : récemment, les valeurs de Shapley ont été utilisées pour interpréter les résultats des modèles de prédiction du *Machine Learning*, tels que les forêts aléatoires et les réseaux de neurones.

La valeur de Shapley pour le segment  $i$  est définie par la formule suivante :

$$SHAP_i = \sum_{\substack{S \in \mathcal{P}(N) \\ i \in S}} \frac{(s-1)! \times (n-s)!}{n!} [\nu(S) - \nu(S \setminus \{i\})]$$

En effet la valeur de Shapley peut être exprimée comme étant la moyenne des contributions marginales du joueur  $i$  aux coalitions auxquelles il est susceptible de participer :

$$SHAP_i = \frac{1}{n} \sum_{\substack{S \in \mathcal{P}(N) \\ i \in S}} \frac{\nu(S) - \nu(S \setminus \{i\})}{C_{n-1}^{s-1}}$$

Ensuite, chaque segment  $i$  se verra attribuer le poids ci-après :

$$w_i = \frac{SHAP_i}{\sum_j SHAP_j}$$

Contrairement à la méthode marginale, cette approche permet de considérer l'effet des sous-ensembles. Cependant, son inconvénient principal réside dans sa complexité lorsque le nombre de joueurs devient trop grand. En général, l'effort nécessaire pour calculer les valeurs de Shapley est exponentiel en nombre de joueurs, ce qui peut rendre le processus chronophage lorsque ce dernier est élevé. Dans ce cas, les valeurs de Shapley peuvent être approchées à l'aide des simulations Monte-Carlo.

Afin d'alléger les calculs, nous ferons abstraction de la variable **sexe** des *model points* : deux modèles du même âge mais de sexe différent seront assimilés à un seul *model point* pour cette partie de l'étude. Nous allons utiliser la méthode basée sur l'échantillonnage aléatoire :

$$SHAP_i = \frac{1}{m} \sum_{k=1}^m \frac{1}{n} \sum_{\substack{S \in \mathcal{P}_k(N) \\ i \in S}} \frac{\nu(S) - \nu(S \setminus \{i\})}{C_{n-1}^{s-1}}$$

Où  $m$  représente le nombre des scénarios et  $\mathcal{P}_k(N)$  est l'ensemble des combinaisons déterminé par un scénario  $k$ .

Sur le plan pratique, les étapes ci-après ont été suivies :

### 1. Génération d'échantillons aléatoires :

Pour chaque *model point*, on définit l'ensemble des combinaisons possibles dont il pourra faire partie. Le choix des combinaisons est fait de façon aléatoire.

### 2. Calcul des valeurs de Shapley pour chaque joueur :

La VIF est déterminée selon la méthodologie explicitée dans la partie théorique : l'outil interne SimBEL est utilisé pour calculer la nouvelle valeur de la VIF de chaque combinaison. Pour obtenir par la suite la valeur de Shapley pour chaque joueur selon la formule présentée en haut.

### 3. Calcul des valeurs moyennes :

Cette procédure est répétée pour plusieurs simulations (choix de combinaisons) afin d'obtenir une distribution d'estimations des valeurs de Shapley pour chaque joueur. Ainsi la valeur finale est la moyenne des valeurs calculées pour chaque simulation.

## 2.1.4 Résultats

Pour refléter la différence entre les résultats des trois méthodes, le graphique suivant illustre les 15 premiers *model points* en termes de poids d'allocation pour chaque méthode :

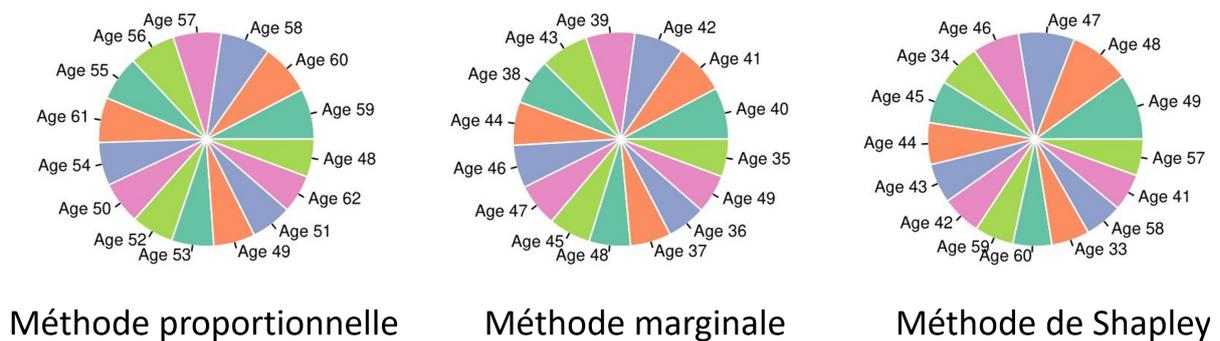


FIGURE 2.6 – 15 premiers MP par les trois méthodes d'allocation

C'est ainsi qu'il apparaît que les trois méthodes mettent en avant des *model points* différents et avec des poids différents. Les profils les plus prépondérants selon la première méthode sont les plus âgés, puisqu'ils détiennent la plus grande part de l'encours du portefeuille. Ces profils ont un âge qui varie entre 47 et 61 ans. La méthode marginale donne plus de poids aux *model points* avec un âge entre 35 et 48 ans. Quant à Shapley, l'âge des *model points* se situe entre 32 et 60 ans.

Si on observe que les 15 premiers *model points*, seulement trois sont communs aux trois méthodes. La méthode marginale et celle de Shapley se partagent cependant 67% de profils mais avec des poids différents comme le montre le graphique ci-après :

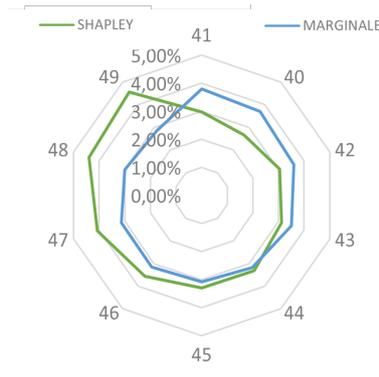


FIGURE 2.7 – Caption

Les poids de Shapley sont retenus dans le reste de l'étude pour estimer la marge de rentabilité vie des *model points* :

ID	Sexe	Age	Nombre Police	Cotisations	PM	VIF
1	M	26	8144	900 064	316 655	36 852
2	M	27	8806	1 030 775	413 481	48 134
3	M	28	9438	1 046 328	5 054 716	58 862
4	M	29	10639	1 205 805	6 596 335	76 814
5	M	30	11169	1 200 395	7 911 590	92 130
6	M	31	12081	1 379 883	9 801 018	114 133

TABLE 2.3 – Marge de rentabilité finale pour quelques *model points*

## 2.2 Calcul de la marge non-vie

Comme expliqué dans le premier chapitre, la compagnie d'assurance IARD utilise un modèle interne markovien à plusieurs états pour estimer la marge non-vie de ses clients. Mais en utilisant ce type de modèle, les équipes concernées ne peuvent pas interpréter les résultats et comprendre les déterminants de la valeur client obtenue. Dans cette partie, nous allons donc essayer de reconstruire la marge non-vie avec un modèle plus facilement interprétable. Pour ce faire, nous allons utiliser un modèle de régression GLM, suivi par un modèle de régression pénalisée LASSO. Nous commencerons alors par présenter les données utilisées dans le présent mémoire. Pour ensuite expliciter les modèles statistiques employés dans l'estimation de la marge, tout en expliquant la théorie derrière. Enfin, nous présenterons les résultats après les avoir validés. Nous tiendrons à faire le lien avec la problématique de l'agrégation des deux marges en soulignant l'impact de la souscription des assurés non-vie à un produit d'épargne.

### 2.2.1 Données et hypothèses

Comme la marge à reconstruire est une quantité elle-même estimée par la compagnie, il est indispensable de comprendre les données et leur provenance pour mieux interpréter les résultats. En effet, la filiale de l'assurance IARD commercialise un catalogue diversifié de produits :

- AUTO : un produit automobile qui propose de prendre en charge les dommages subis par le véhicule, en plus d'une garantie responsabilité civile et des services d'assistance ;
- MRH : un produit multirisques habitation qui offre des garanties pour protéger l'essentiel du logement de l'assuré comme l'incendie et les dégâts des eaux ;
- PREV : un produit de prévoyance qui couvre les accidents de la vie courante. Il protège ainsi les bénéficiaires des incidents tels les accidents domestiques, sportifs, de bricolage ou de loisirs ;
- CORP : un produit d'assurance corporelle en cas d'accident automobile. Il offre également des garanties d'incapacité et de perte d'autonomie, qui ne sont pas proposées par le produit AUTO. C'est pourquoi ce produit est majoritairement souscrit avec ce dernier.

Mais comme l'assureur propose plusieurs variantes du même produit : soit pour des risques différents, des garanties différentes ou des conditions de souscription différentes, les assurés peuvent dès lors détenir plusieurs contrats d'assurance. Les quatre produits précédemment cités ont été catégorisés selon les contrats proposés :

- Par exemple, le produit MRH est subdivisé en deux variables : MRH 1 et MRH 2. Ces contrats sont très similaires mais exclusifs : l'assuré n'en peut détenir qu'un. MRH 1 était l'ancienne offre du produit, qui n'est plus commercialisée aujourd'hui mais qui est toujours dans le portefeuille actuel. Cependant, ces deux variantes ont été scindée en une seule pour le calcul de la marge par le modèle markovien multi-états. Cette logique sera répliquée dans notre estimation de la marge.
- Le produit AUTO est également réparti en deux contrats : AUTO1 et AUTO2 qui couvrent deux risques distincts. Cependant pour le calcul interne, le modèle a considéré que la variante AUTO1 va être utilisée pour désigner le contrat du véhicule le plus fréquemment utilisé, si le client assure plus d'un véhicule ; sinon AUTO1 représentera son contrat auto.
- De même pour le produit PREV, deux contrats coexistent : PREV 1 et PREV 2. Nous allons appliquer la même logique que celle employée pour les contrats MRH ; puisqu'il s'agit de deux contrats exclusifs mais qui se rapprochent tout de même dans leur fonctionnement.
- Le produit CORP quant à lui se compose uniquement d'un seul contrat.

Les contrats eux-mêmes regroupent plusieurs formules. Par exemple AUTO1 se partage entre la formule initiale, la formule essentielle et la formule proportionnelle.

A l'addition des produits IARD, ce jeu de données explicite la détention de l'assuré d'un produit épargne ou décès.

Dans le cadre de cette étude, l'assureur a communiqué les données qui concernent le stock de son portefeuille. Elles peuvent se répartir en deux types de bases :

- Une base de données « Assurés » : qui décrit les caractéristiques, aussi bien qualitatives que quantitatives, des assurés présents dans le portefeuille actuel :
  - `age` ;
  - `sexe` ;
  - `ancienneté` ;
  - `situation familiale` ;

- catégorie socioprofessionnelle.

A l'addition de ces colonnes, la base contient l'information sur le type d'équipement de l'assuré ; les produits auxquels il a souscrit. Ces données se résument dans les variables binaires suivantes :

- détention AUTO ;
  - détention MRH ;
  - détention CORP ;
  - détention PREV ;
  - détention Epargne ;
  - détention CAPDECES.
- Des bases de données « Contrats » : elles contiennent des informations sur les contrats souscrits. Elles contiennent des variables générales et communes entre tous les produits, mais également des variables propres à chacun. Par exemple, pour les produits :
    - AUTO : des variables type CRM qui représente le coefficient de réduction-majoration , **categorie** pour désigner si le véhicule a 2 roues ou 4 roues, **energie**, **puissance** ;
    - MRH : des variables comme le nombre de propriétés, le nombre de pièces, la tranche mobilière ou le lieu utilitaire.

Cependant, nous ne possédons que les caractéristiques spécifiques liées aux contrats non-vie. En ce qui concerne les produits vie : épargne et décès, la seule information disponible dans cette base de données est la détention ou non de ces polices.

Toutes les variables présentées ci-dessus sont des variables catégorielles : les valeurs ont été codées sous un nombre limité de modalités.

La compagnie a également fourni une table par produit des marges calculées par le modèle markovien. Chaque *model point* a 31 valeurs correspondantes à la projection des marges de rentabilité futures. Ces marges ont été répliquées pour passer du calcul au niveau *model point* à l'estimation au niveau assuré.

La base de données finale se présente comme une matrice de 2 769 583 individus avec 45 variables : les caractéristiques des assurés, les variables détention de produit ainsi que les attributs liés aux contrats de chaque produit et enfin les marges de rentabilité pour les 31 prochaines années. Elle se présente comme suivant :

ID	AGE	...	AUTO1	AUTO1 ENERGIE	...	MRH	MRH Uti- litaire	...	Marge 1	...	Marge 31

La construction des bases de données fait de telle sorte à ce que deux types de données « manquantes » apparaissent : d'un côté, si l'assuré ne détient pas un contrat en particulier, les valeurs afférentes à ce produit seront manquantes, et d'un autre côté si la compagnie d'assurance ne possède pas l'information sur l'une des variables par rapport à l'assuré ; dans ce cas la variable aurait la valeur « indéterminé » dans les bases de données initiales. Pour différencier entre ces deux types de données, si l'individu ne possède pas un contrat en particulier, la valeur « Absence » sera affectée aux variables concernées. S'il s'agit d'une variable ordinale, où les valeurs suivent une échelle d'ordre, ce label sera classé en premier afin de mieux assimiler les résultats du modèle linéaire. Si la valeur d'une variable est cependant indéterminée parce que l'assureur n'en dispose pas, la valeur NA de R lui sera affectée.

Le tableau suivant récapitule le nombre des observations manquantes pour chaque variable. Les colonnes associées aux produits automobiles présentent les plus grandes valeurs de données manquantes, avec les variables `anc_permis` et `agecond` en tête de liste.

Nom de la variable	Descripton de la variable	Nombre des NA
auto2_ancpermis	Ancienneté du permis	75 004
auto2_agecond	Age du conducteur	56 345
auto1_ancpermis	Ancienneté du permis	33 126
auto1_agecond	Age du conducteur	15 887
mrh2_lieuutil	Présence d'un lieu utilitaire	11 182
mrh2_naturelri	Nature du bien assuré	7 829
auto1_energie	Type énergie du moteur du véhicule assuré	18
auto2_energie	Type énergie du moteur du véhicule assuré	15

TABLE 2.4 – Distribution des valeurs manquantes

Un total de 138 479 individus sur 2 769 583 ont au moins une valeur manquante. Pour y remédier, nous allons adopter trois approches. Premièrement, comme il s'avère que la proportion des valeurs manquantes est comparable à celles des autres classes pour chacune des quatre premières variables selon le graphe 2.8, nous allons implémenter une nouvelle classe NA pour chacune de ces dernières.

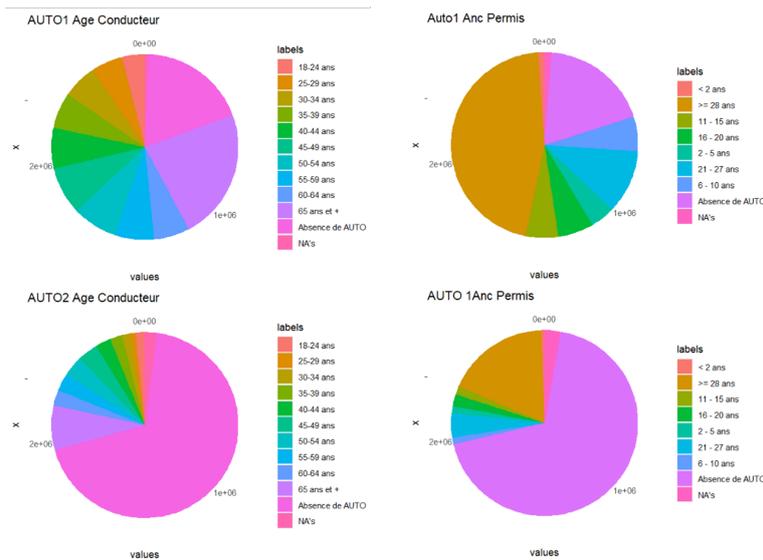


FIGURE 2.8 – Part des valeurs manquantes pour les variables : auto1\_age, auto1\_ancpermis, auto2\_age, auto2\_ancpermis

Il faudra donc changer l'ordre des modalités des variables puisqu'il s'agit de variables catégorielles ordinales. Nous allons faire l'hypothèse que la classe NA est au-dessous de la classe « Absence » et au-dessus de la première modalité. Ensuite, en ce qui concerne les variables `mrh2_lieuutil` et `mrh2_naturelri`, nous allons utiliser le package R `missRanger` pour remplacer les valeurs manquantes. Ce package nous permettra d'effectuer une imputation en utilisant l'algorithme Random Forest. Finalement, nous allons omettre les lignes qui restent (33 observations).

La base finale se compose alors de 2 769 550 individus. Vu la grande taille de la base, elle sera décomposée en 4 sous-bases : une pour chaque produit. Le tableau suivant donne la proportion des détenteurs de chaque produit dans le portefeuille :

Produit	Nombre d'observations	Proportion
AUTO	2 239 492	81%
MRH	1 889 060	68%
PREV	1 269 772	46%
CORP	2 055 992	74%

TABLE 2.5 – Proportion des détenteurs d'un produit d'épargne par produit IARD

Pour avoir une idée sur la composition du portefeuille de cet assureur IARD, le graphique ci-dessous montre la corrélation entre la souscription des différents produits en illustrant le nombre d'assurés pour chaque combinaison d'équipement.

Ainsi, il apparaît que Les combinaisons de produits les plus populaires sont représentées par : AUTO  $\times$  CORP , AUTO  $\times$  CORP  $\times$  PREV et AUTO  $\times$  CORP  $\times$  MRH. Elles comportent toutes l'association des produits AUTO et CORP : puisque le produit CORP vient compléter le produit AUTO en offrant des garanties corporelles liées aux accidents automobiles non couvertes par ce dernier.

Dans une logique d'analyse bivariée pour le produit AUTO par exemple, l'effectif le plus élevé des assurés qui ont souscrit un autre produit est évidemment celui des clients CORP. Ensuite en deuxième position, nous retrouvons les assurés AUTO et MRH. Enfin le nombre d'assurés AUTO le plus bas est celui de la population qui détient un produit PREV.

Pour le produit PREV, les assurés se répartissent à peu près de la même manière entre les trois autres produits. La majorité des assurés CORP et MRH détiennent un produit AUTO.

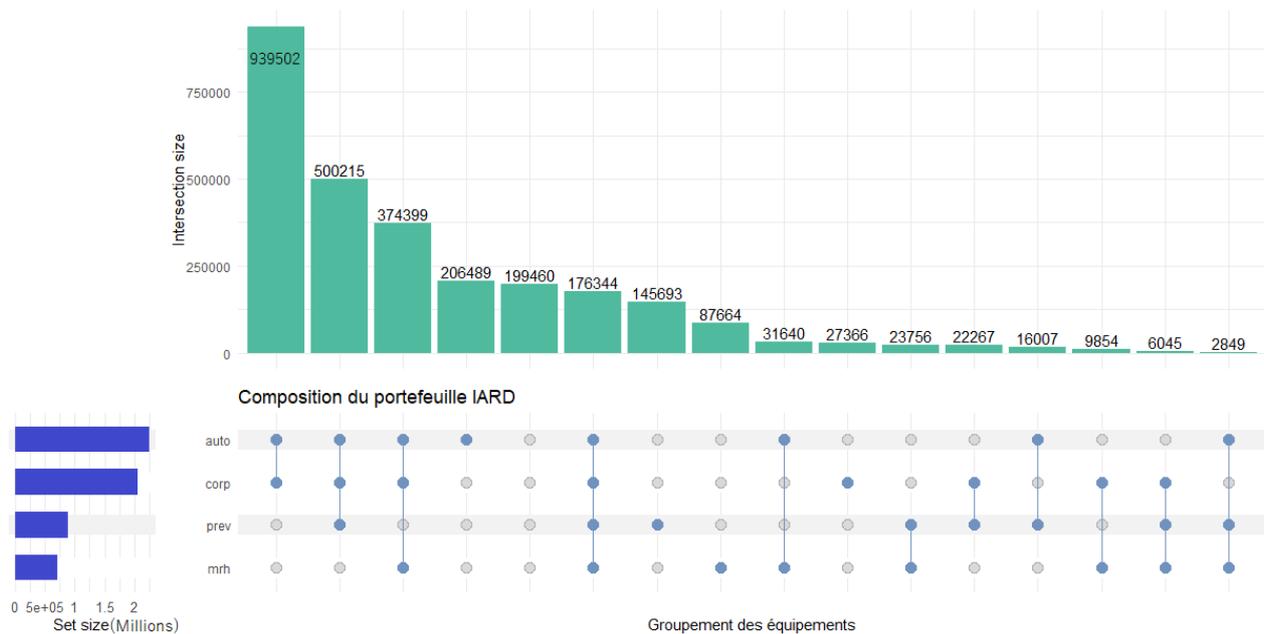


FIGURE 2.9 – Composition du portefeuille IARD

Par ailleurs, l'étude accorde une attention particulière aux individus ayant souscrit à la fois à une assurance non-vie et à un produit d'épargne. Ces individus se partagent entre les quatre produits

non-vie comme le montre le tableau suivant :

	Nombre Epargne	Nombre Non Epargne	% Epargne	% Non Epargne
AUTO	198 414	2 041 079	9%	91%
MRH	194 664	1 682 957	10%	90%
PREV	160 787	1 103 456	13%	87%
CORP	194 138	1 854 548	9%	91%

TABLE 2.6 – Répartition des assurés par produit IARD selon la détention d'un produit d'épargne

En termes de nombre de souscription, ce sont les personnes qui détiennent un produit AUTO qui souscrivent le plus à un produit d'épargne en plus. En revanche, en termes de pourcentage, les clients PREV connaissent le taux le plus élevé des souscriptions épargne.

L'âge et l'ancienneté sont considérés parmi les facteurs habituels qui impactent le plus la souscription d'un contrat d'assurance. Par conséquent, il serait pertinent d'analyser le lien entre ces variables et la détention d'un produit épargne. En effet, d'après les graphiques ci-dessous la proportion de assurés IARD qui ont acquis un contrat d'épargne tend à augmenter avec l'âge et les années passées dans le portefeuille.

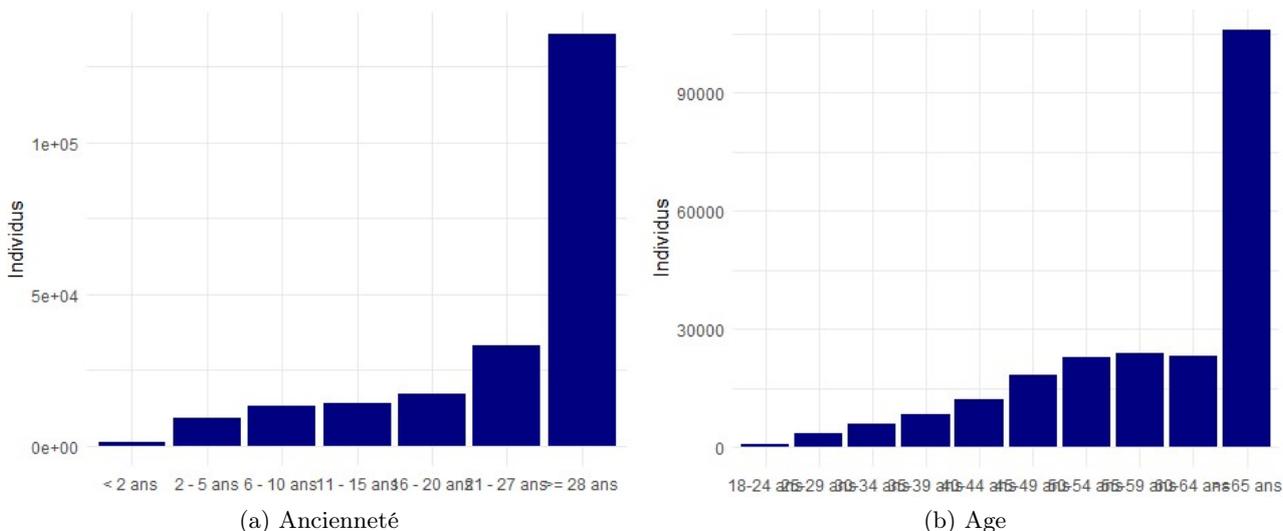


FIGURE 2.10 – Distribution de l'âge et de l'ancienneté des détenteurs d'un produit épargne

Il est également important de s'intéresser à la façon dont évoluent les marges estimées. Pour cela, nous allons considérer une seule marge par échéance à savoir la somme des marges des quatre produits. La représentation graphique ci-après illustre les trajectoires des marges au cours des années projetées, pour 10 *model points* choisis de manière aléatoire. Nous constatons que cette grandeur tend progressivement vers 0, ce phénomène étant cohérent avec la diminution progressive de la probabilité de maintien d'un assuré dans le portefeuille au fil des années. En effet, elle décroît également vers 0, puisque les individus ont davantage de chances de quitter le portefeuille de l'assureur avec le temps, que ce soit en résiliant leur contrat, en subissant un sinistre ou simplement par échéance du contrat. De plus, il convient de noter que la plupart des marges présentent des valeurs positives.

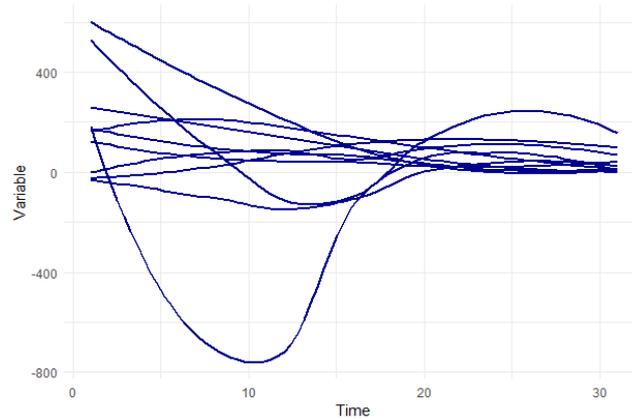


FIGURE 2.11 – Trajectoires des marges IARD

Afin de généraliser ce dernier constat, nous allons considérer l'évolution de la moyenne des marges totales dans le temps. Le graphique suivant montre que cette quantité demeure positive au fil des années avant de s'annuler vers la fin de la période de projection.

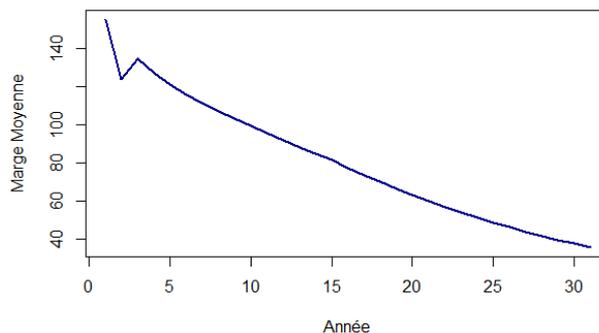


FIGURE 2.12 – Évolution de la marge totale moyenne IARD

### 2.2.2 GLM

Pour rappel, l'objectif de cette section est d'évaluer la marge IARD des assurés à l'aide d'un modèle facilement compréhensible, en utilisant les caractéristiques disponibles dans les bases de données.

C'est dans cette optique, que le mémoire d'Oleksandr Sorochynskyi ([Sorochynskyi, 2020](#)) a exploré la reconstruction des marges par des modèles du *Machine Learning* à partir des attributs des assurés. Comme les bases de données des marges se décomposent en 31 valeurs par échéance, et afin de simplifier le travail, il s'est basé sur l'estimation d'une seule marge. Il utilise la somme des valeurs actualisées des marges projetées. Dans la suite de cette étude, nous allons modéliser la même quantité agrégée. Le mémoire de Paul De Arce ([Arce, 2022](#)) a traité le même sujet mais avec des arbres CART et des modèles de partitionnement récursif (*model based recursive partitioning*).

De la même manière, la première solution qui se présente devant cette problématique de modélisa-

tion est l'utilisation d'un modèle de régression linéaire. Nous allons procéder dans un premier lieu par un modèle linéaire généralisé (*General Linear Models* (GLM)) :

En général, les modèles de régression linéaire estiment l'espérance d'une variable dépendante  $Y$  à partir d'une combinaison linéaire d'un ensemble de covariables observées  $X_1, \dots, X_n$ , comme dans l'exemple suivant :

$$E(Y) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_n X_n$$

Les modèles GLM offrent une certaine flexibilité pour tenir compte d'un plus large éventail de distributions de résultats en permettant à une fonction de lien de varier linéairement par rapport aux covariables, comme dans :

$$f(E(Y)) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_n X_n$$

Les coefficients  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n$  sont typiquement estimés par calcul du maximum de vraisemblance.

L'intérêt de cette approche est qu'elle permet d'estimer l'impact de chaque variable. En effet, Les coefficients associés à chaque prédicteur mesurent l'effet moyen sur la variable de réponse lorsque la variable explicative « augmente d'une unité », en maintenant toutes les autres variables constantes. L'ampleur et la direction de l'effet dépendent du coefficient lui-même et de la fonction de lien. Ceci est particulièrement utile dans le cadre d'une étude de rentabilité, puisque ce modèle permet de mieux comprendre les déterminants des marges clients et saisir l'effet marginal de chaque variable.

### 2.2.3 LASSO

Un autre modèle qui peut s'avérer utile dans le cas de cette étude, vu le nombre de variables pour certains produits est la régression LASSO (*Least Absolute Shrinkage and Selection Operator*). Comme son nom l'indique, LASSO utilise une technique de « réduction » (shrinkage) pour trouver des modèles simples qui possèdent moins de paramètres. Nous obtenons ainsi une meilleure interprétation des modèles.

L'objectif principal de la régression LASSO est de trouver un équilibre entre la simplicité et la précision du modèle. Il y parvient en ajoutant un terme de pénalité au modèle de régression linéaire traditionnel, ce qui encourage les solutions peu nombreuses où certains coefficients sont forcés d'être exactement nuls. Cette caractéristique rend LASSO particulièrement utile pour la sélection des caractéristiques, car il peut automatiquement identifier et éliminer les variables non pertinentes ou redondantes.

En pratique, la régression LASSO commence par un modèle linéaire simple :  $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$  où  $\epsilon$  représente le terme erreur du modèle.

Ensuite, la procédure LASSO introduit un terme de pénalité additionnel basé sur la valeur des coefficients. Le terme de régularisation  $L_1$  s'écrit donc comme suit :

$$L_1 = \lambda \times (|\beta_1| + |\beta_2| + \dots + |\beta_n|)$$

Où  $\lambda$  est le paramètre de régularisation ou de réglage qui contrôle la force de la pénalité  $L_1$ .

L'objectif de la régression LASSO est donc de trouver les valeur des coefficients qui minimisent la fonction objective constituée de la somme des carrés des résidus à laquelle se rajoute le terme de régularisation  $L_1$  :

$$L_{lasso} = \operatorname{argmin}_{\hat{\beta}} \{ \|Y - \beta X\|^2 + L_1 \}$$

En rajoutant ce terme  $L_1$ , et quand  $\lambda$  est suffisamment élevée, la régression LASSO peut réduire les coefficients vers une valeur nulle.

Le choix de  $\lambda$  représente alors une étape cruciale dans ce modèle. Une grande valeur de  $\lambda$  accentue la force de pénalité, poussant le modèle à rejeter plus de variables. Et inversement, une faible valeur de  $\lambda$  réduit l'effet de pénalité, retenant ainsi plus de variables.

### 2.2.4 Résultats

Comme explicité ci-dessus, le modèle de régression cherche à estimer une valeur de marge totale par produit et par assuré. Le travail sera alors réparti en cinq modèles, un pour chaque produit : AUTO1, AUTO2, MRH, PREV et CORP. La méthodologie étant la même, seuls les résultats d'un seul produit seront détaillés dans cette partie. De plus, la base de données d'un produit X par exemple, ne contiendra que les individus qui ont déjà souscrit à ce produit. Cette approche s'oppose à celle qui chercherait à estimer la marge de rentabilité pour un nouveau client.

Nous allons décomposer la base de données en deux bases : l'une pour l'apprentissage, la seconde pour l'entraînement du modèle. L'estimation du modèle sera faite sur la première base, et la validation des résultats sur la seconde.

Pour le modèle GLM, et puisque la marge prend des valeurs tout autant négatives que positives, la famille de loi sera fixée comme gaussienne avec l'identité comme fonction de lien.

Tous les calculs seront réalisés avec R, on utilisera la fonction *glm* et le package *glmnet*.

#### Valeurs extrêmes

Étant donné que les valeurs aberrantes peuvent influencer les résultats du modèle linéaire généralisé, une étape préliminaire consiste à procéder par une analyse des valeurs des marges en vue de détecter la présence de ces valeurs atypiques. À cette fin, le diagramme *boxplot* est utilisé comme un outil graphique qui résume la dispersion des valeurs des marges au sein de la base de données. Il apparaît alors que la problématique liée aux valeurs extrêmes constitue une préoccupation substantielle pour ce jeu de données :

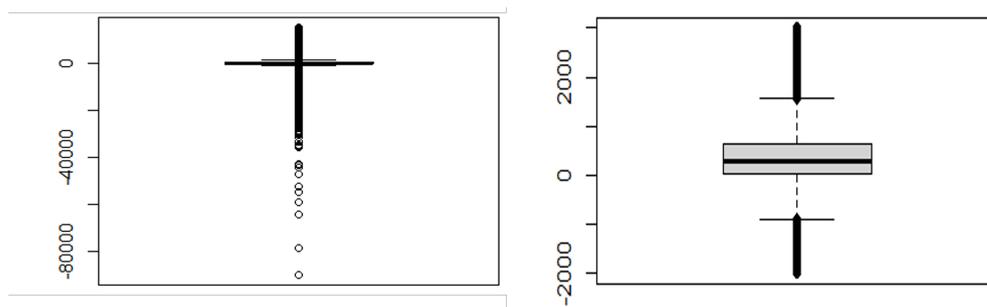


FIGURE 2.13 – *Boxplots* des marges pour le produit AUTO1

Le graphe à gauche représente le *boxplot* des marges pour la base initiale du produit AUTO1, alors que celui de droite illustre la répartition des valeurs après avoir éliminé une proportion des données extrêmes (environ 3% de la base de données) et qui permet une meilleure visualisation du problème.

Plusieurs solutions se présentent devant cette problématique :

- **transformation des variables :**

il s'agit d'une approche souvent utilisée pour traiter les valeurs extrêmes, notamment la trans-

formation logistique. Cependant, le choix de la fonction de transformation dépend de la nature des données et affecte considérablement l'interprétation des résultats.

- **méthodes robustes du GLM :**

R offre la possibilité d'utiliser des modèles GLM qui ont été conçues pour faire des statistiques robustes et qui prennent en considération la présence des valeurs extrêmes avec la fonction *glmRob* du package *Robustbase*. Cependant, vu la taille de la base de données actuelle, ce calcul est extrêmement lourd à exécuter.

- **élimination des valeurs extrêmes de la base de données :**

dans la suite, cette méthode a été appliquée et nous avons procédé par suppression des observations aberrantes.

## Corrélation des variables

Il est crucial d'étudier la relation de corrélation entre les variables explicatives du modèle. Vu la nature catégorielle de ces dernières, nous nous baserons sur la statistique V de Cramer pour estimer la dépendance entre les prédicteurs. Cette mesure correspond à une transformation du test de  $\chi^2$ . La fonction *cramerV* du package R *rcompanion* propose une version non biaisée telle que :

$$\tilde{V} = \sqrt{\frac{\tilde{\phi}^2}{(\tilde{k} - 1)(\tilde{r} - 1)}} ;$$

où  $\phi^2 = \max\left(\frac{\chi^2}{n} - \frac{(k-1)(r-1)}{n-1}, 0\right)$ ,  $\tilde{k} = k - \frac{(k-1)^2}{n-1}$  et  $\tilde{r} = r - \frac{(r-1)^2}{n-1}$

$\chi^2$  est la statistique du test du même nom de Pearson,  $n$  le nombre d'observations,  $k$  le nombre de colonnes et  $r$  le nombre de lignes. On remarque que problème de multi colinéarité est surtout présent pour les bases de données automobile comme l'explique le graphique ci-après :

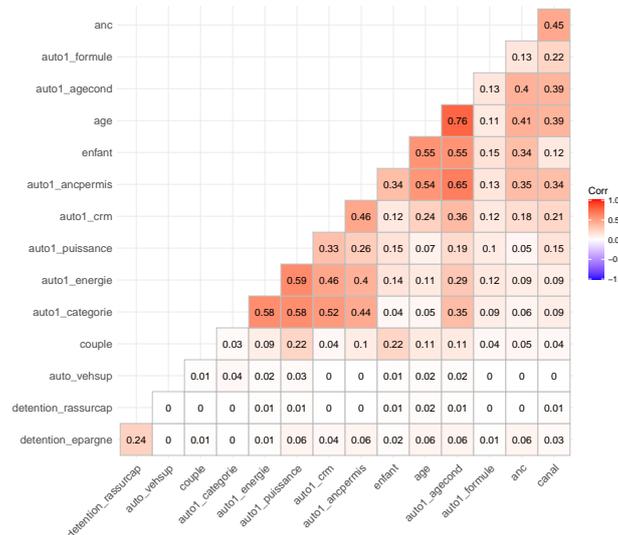


FIGURE 2.14 – Corrélation des variables pour le produit AUTO1

Les variables les plus corrélées sont l'âge de l'assuré et l'âge du conducteur, suivies par l'ancienneté du permis et l'âge du conducteur, et enfin on remarque une corrélation entre les variables liées au véhicule telle l'énergie, la catégorie et la puissance. La variable âge du conducteur sera retirée pour la

modélisation GLM, vu qu'elle contient quasi les mêmes valeurs que l'âge de l'assuré pour ce stock de polices.

## GLM

La variable `detention_epargne` est une variable significative dans l'estimation des marges de tous les produits, à part le produit PREV. L'estimation du coefficient de cette variable est en annexe A.1. Les résultats du  $R^2$ , calculé sur les bases d'entraînement, sont récapitulés dans le tableau suivant :

PRODUIT	AUTO1	AUTO2	PREV	CORP	MRH
$R^2$	49%	45%	42%	31%	57%

TABLE 2.7 –  $R^2$  du GLM par produit

## Overfitting

La prochaine étape consiste à examiner le surajustement du modèle estimé. L'une des techniques les plus utilisées pour cette fin est la validation croisée. Cette méthode permet de diviser l'ensemble de données en  $k$  sous-ensembles ou plis, généralement de manière aléatoire. Le modèle est entraîné sur  $k-1$  plis (le sous-ensemble d'entraînement) et évalué sur le pli restant (le sous-ensemble de validation). Ce processus est répété  $k$  fois, chaque pli étant utilisé comme ensemble de validation exactement une fois. Dans notre cas, pour évaluer la performance du modèle, nous utiliserons 5 blocs. Le tableau suivant récapitule les résultats obtenus sous R :

Bloc	RMSE	$R^2$	MAE
<b>Bloc 1</b>	132,088	49%	69,528
<b>Bloc 2</b>	132,146	49%	69,382
<b>Bloc 3</b>	132,102	49%	69,318
<b>Bloc 4</b>	131,884	49%	69,324
<b>Bloc 5</b>	131,577	49%	69,138

TABLE 2.8 – Résultats de la validation croisée

Les résultats de l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE) sont plutôt stables d'un bloc à un autre ce qui signifie que le modèle ne montre pas de signe de surajustement.

## Analyse des résidus

Enfin, nous nous intéresserons à l'analyse des résidus de Pearson pour étudier la surdispersion du modèle. Les résidus de Pearson sont des résidus standardisés par l'écart-type théorique des valeurs prédites.

$$r_i^p = \frac{Y_i - \hat{Y}_i}{\sqrt{\text{Var}(\hat{Y}_i)}}$$

Le graphique suivant montre la distribution des résidus de Pearson :

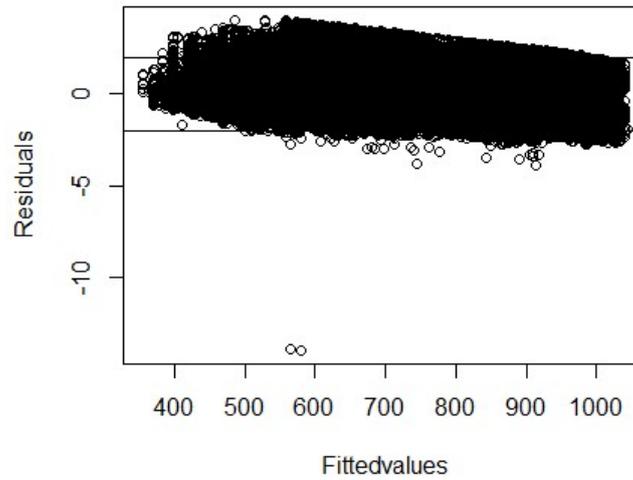


FIGURE 2.15 – Distribution des résidus de Pearson pour l’estimation de la marge AUTO1

Généralement, les résidus de Pearson se situent entre -2 et 2 pour un bon modèle. Vu le grand nombre des observations, le graphique n’est pas facilement interprétable. En effet 95% des résidus se situent dans cette plage pour le produit AUTO1. Cependant, la proportion des résidus restants peut indiquer une surdispersion due à plusieurs causes comme la présence de valeurs extrêmes ou la non-linéarité des variables.

## Régression LASSO

En tenant compte du nombre important de variables pour certains produits, et de la présence potentielle de multicollinéarité entre celles-ci, il est plutôt judicieux d’utiliser une régression LASSO afin de sélectionner les variables pertinentes. Les valeurs de  $R^2$  sont plus faibles par rapport aux résultats du GLM, ce qui peut être expliqué par l’omission de certaines variables pour résoudre le problème de corrélation entre les prédicteurs :

PRODUIT	AUTO1	AUTO2	PREV	CORP	MRH
$R^2$	35%	21%	43%	17%	43%

TABLE 2.9 –  $R^2$  du LASSO par produit

Ce niveau de performance, quoique faible, est comparable avec ce qui a été obtenu avec les méthodes du *Machine Learning* par [Sorochnykyi \(2020\)](#) qui affichaient en moyenne un  $R^2$  de 39%.

Toutefois, il faut souligner que la régression LASSO associe un facteur positif pour la variable `detention_epargne`, sauf pour le produit PREV où la régression lui attribue une valeur négative. Ce qui implique que la marge de rentabilité non-vie, pour les produits AUTO, CORP et MRH, est plus élevée chez les assurés ayant souscrit un produit d’épargne.

# Chapitre 3

## Agregation des marges

### 3.1 Détermination du point mort

Selon la [Banque de France \(Juillet 2019\)](#), le point mort d'un segment est le moment où une affaire devient rentable en couvrant ses charges. En assurance, il est défini comme l'année où la marge de rentabilité cumulée, qui en début d'année était négative, devient positive pour un segment donné. Cette notion permet à l'assureur de déterminer quand est-ce qu'un profil de client commencera à être rentable. L'assureur réalise donc des pertes pendant toute la durée antérieure au point mort. Le groupe d'assurance, en question, détenant les deux filiales vie et non-vie, peut s'intéresser à déterminer cette mesure pour ses clients, en particulier les assurés communs aux deux filiales.

#### 3.1.1 Calcul du point mort

La marge totale d'un client ayant souscrit un produit vie et non vie peut s'évaluer comme la somme des marges calculées au préalable. D'un côté, nous avons estimé dans le chapitre précédent les marges de rentabilité vie pour les différents *model points*. Et d'un autre côté, l'assureur IARD a déjà fourni les valeurs clients de ses assurés. La proportion des assurés non-vie ayant souscrit un produit d'épargne se présente comme suivant :

Nombre total des assurés non-vie	2 769 550
Proportion avec produit d'épargne	8.13%
Proportion sans produit d'épargne	91.87%

TABLE 3.1 – Répartition des assurés non-vie selon la détention d'un produit d'épargne

Pour lier les résultats des deux tables, et comme les identifiants ne sont pas les mêmes entre les deux bases de données, et que nous ne détenons pas le détail des polices pour chaque *model point* pour la partie vie, la jointure se fera par le biais du croisement des deux variables : **sexe** et **age**. La variable de l'ancienneté n'a pas été prise en compte pour deux raisons majeures :

- la première étant que l'ancienneté figurant dans chacune des bases de données ne concerne que les produits de la base ; ce qui veut dire qu'un individu ayant une police épargne depuis 5 ans et

qui vient de souscrire un produit MRH il y a un an, aura deux valeurs différentes d'ancienneté dans chacune des bases respectives.

- la deuxième étant que dans ce cas de figure, l'étude du point mort n'est pertinente que pour les nouvelles souscriptions : les marges sont calculées à jour, ce qui fait que les valeurs actuelles des marges ne tiennent pas compte des pertes réalisées dans le passé par les polices présentes dans le portefeuille depuis des années. Si les anciennes polices sont donc considérées dans l'étude, elles induiront à un biais de sous-estimation du point mort. Ce constat se justifie davantage par le graphique suivant :

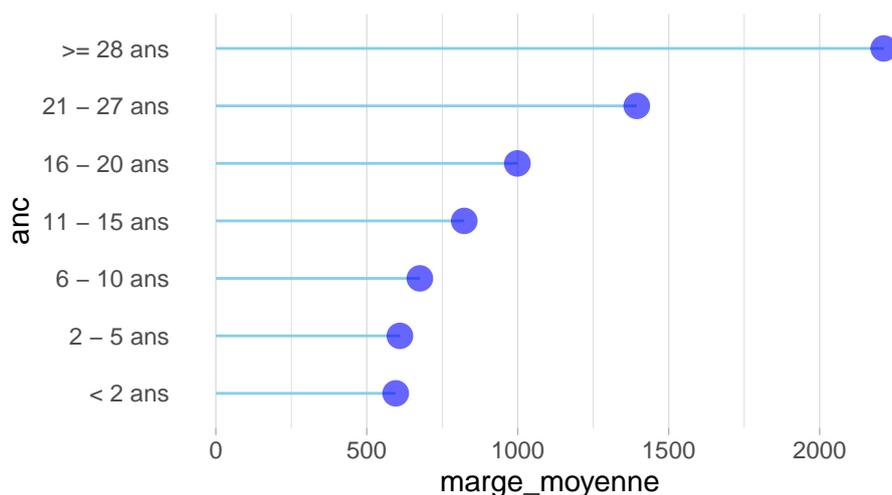


FIGURE 3.1 – Evolution de la marge cumulée moyenne par ancienneté

La marge moyenne des polices qui appartiennent à la classe d'ancienneté supérieure est considérablement plus grande que la marge moyenne des autres classes.

La base de données utilisée se limitera donc aux polices qui ont au maximum un an d'ancienneté. Pour les clients communs entre les deux filiales, le tableau ci-après récapitule la répartition de cette population selon l'ancienneté :

Nombre total des assurés non-vie ayant un contrat d'épargne	225 144
Proportion ancienneté $\leq 1$ an	0.65%
Proportion ancienneté $> 1$ an	99.35%

TABLE 3.2 – Répartition des assurés communs aux deux filiales selon l'ancienneté IARD

Naturellement, pour aborder la notion du point mort, il est impératif d'analyser l'évolution temporelle des marges de rentabilité. Pour rappel, une des bases de données de l'assureur non-vie comportait la projection des marges de rentabilité par produit sur 31 ans. Nous utilisons alors la marge non-vie totale cumulée par assuré.

Pour la partie assurance vie, nous allons réaliser la projection des marges vie sur 31 ans. Pour ce faire, nous procédons par la même approche explicitée dans le chapitre précédent. L'outil interne SimBEL sera utilisé pour déterminer les deux premières composantes de la VIF, à savoir la PVFP et la TVOG. Les éléments restants sont estimés suivant la même approximation expliquée dans la section 2.1.2.

Pour la répartition de cette valeur sur les *model points*, et comme la méthode de Shapley est ardue à mettre en œuvre pour un horizon aussi lointain ; nous allons estimer les poids de Shapley par les variables explicatives : `age`, `provisions_mathematiques` et `nombre_polices` par *model point*. Afin d’avoir assez de données, nous allons faire le calcul pour les 5 premières années. Sur la base des résultats obtenus, une régression Bêta est effectuée pour estimer les nouveaux poids.

La classe des modèles de régression Bêta est couramment utilisée pour modéliser des variables dont les valeurs sont comprises dans l’intervalle  $[0, 1]$ . Elle repose sur l’hypothèse que la variable dépendante est distribuée selon une loi bêta et que sa moyenne est liée à un ensemble de régresseurs par l’intermédiaire d’un vecteur de prédicteurs aléatoires à coefficients inconnus et d’une fonction de lien (Francisco Cribari-Neto, s. d.). La fonction `betareg()` de R donne les coefficients qui sont résumés dans le tableau suivant :

Variable	Coefficient
Age	$2.75 \times 10^{-5}$
Nombre de polices	$6.63 \times 10^{-7}$
Provisions mathématiques	$1.26 \times 10^{-10}$

TABLE 3.3 – Coefficients de la régression Bêta pour l’estimation des poids de Shapley

Pour respecter la contrainte que la somme des poids soit égale à 1, nous trierons les *model points* par ordre décroissant des poids de Shapley et nous annulerons les poids des *model points* au-delà du poids cumulé de 100%.

Afin de réaliser le croisement avec la base de données non-vie, les *model points* de la base de données vie seront regroupés selon les modalités d’âge de la base IARD. Enfin, la VIF de chaque *model point* sera divisée par le nombre de police pour avoir un montant individuel de la rentabilité vie à attribuer aux assurés non-vie avec un contrat d’épargne. Nous obtenons finalement une base de données où chaque assuré se caractérise par une marge totale, égale à la somme des deux marges, projetée sur un horizon de 31 ans, ainsi que les caractéristiques de leurs polices non-vie décrites dans les chapitres précédents.

Il serait alors nécessaire d’utiliser différents sous-ensembles de ce jeu de données en fonction des exigences de l’analyse. On pourrait ainsi distinguer deux bases différentes :

- **base 1** : la table des assurés non-vie, d’ancienneté inférieure ou égale à 1 an et avec la marge IARD uniquement. Le nombre d’observations de cette base est de 94 310 ;
- **base 2** : la table des assurés non-vie qui détiennent un contrat d’épargne, d’ancienneté inférieure ou égale à 1 an et avec la marge globale. Le nombre d’observations est de 1 473.

Ensuite, afin de déterminer la composition définitive de la base de données à utiliser pour l’analyse du point mort, il convient de définir de manière précise ce concept. Conformément à la définition énoncée au début de ce chapitre, il est envisageable qu’un assuré ne présente pas de point mort, signifiant ainsi que la marge bénéficiaire demeure constamment négative tout au long de la période de projection. Par ailleurs, il existe des contrats d’assurance pour lesquels la valeur de la marge bénéficiaire totale change de signe à plusieurs reprises ; ces situations seront exclues de l’étude ultérieure dans le but d’identifier le point mort.

Le diagramme suivant présente la structure de la base de données initiale : 20% des assurés présentent une valeur négative de rentabilité au cours de la première année. Parmi ces individus, 60% présentent un point mort où ils ne changent de signe qu’une seule fois durant la durée de projection ;

passant ainsi d'une rentabilité négative à une rentabilité positive. En revanche, 37% des assurés, ayant une marge négative, maintiennent des pertes tout au long de la période de projection, tandis que les 3% restants connaissent une rentabilité fluctuante qui changerait de signe plus d'une fois.

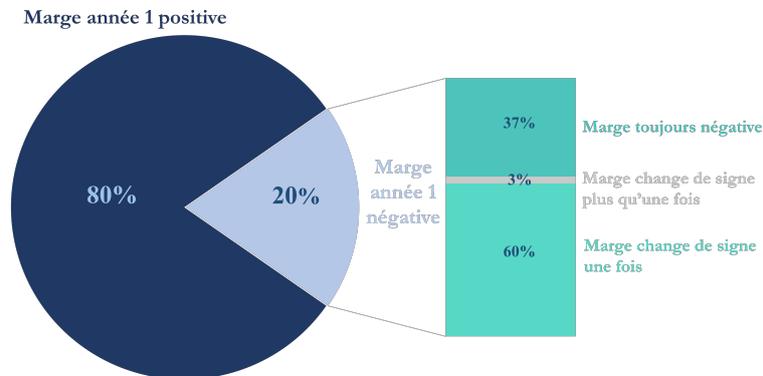


FIGURE 3.2 – Répartition des observations selon le signe de la rentabilité

### 3.1.2 Analyse univariée

La première partie de l'étude repose sur la détection de l'année où la marge cumulée négative devient positive. La base de données contient les assurés non-vie, d'ancienneté maximale inférieure à 1 an, qui ont souscrit à un produit d'épargne et qui ont commencé par une valeur négative mais finisse l'étude avec une mesure de rentabilité positive. Les résultats suivants concernent l'analyse du vecteur des segments morts obtenus sur ce jeu de données.

Dans le graphique qui suit, il est possible de visualiser le point mort moyen pour chaque tranche d'âge (point rouge) mais également de le comparer avec celui qui a été calculé avec la marge IARD uniquement (point bleu). La taille de l'effet de la marge vie est dès lors évaluée par la longueur des segments du graphe :

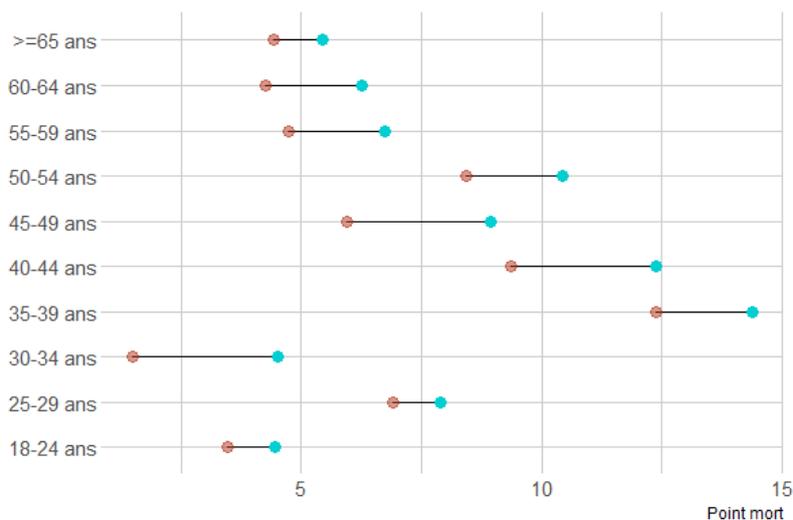


FIGURE 3.3 – Point mort moyen par tranche d'âge

En premier lieu, il est normal que le point mort calculé sur la base des marges IARD uniquement survienne après celui estimé avec la marge globale, puisque la marge vie est une quantité positive, dans ce cas de figure, qui se rajoute à la marge non-vie dans la construction de la marge globale.

Ensuite, il est apparent que les profils d'âge qui nécessitent, en moyenne plus de temps, pour passer d'une valeur de rentabilité négative à une valeur positive sont ceux des assurés âgés entre 35 et 44 ans. Ceci s'explique par une marge IARD très négative qui n'a pas été rattrapée par l'agrégation des résultats de rentabilité.

Cependant, les assurés d'un âge entre 30-34 ans nécessitent le moins de nombre d'années pour générer des bénéfices : en moyenne seulement 1 année. Ceci est dû à un effet combiné : ce segment est l'un des plus rentables dans les deux branches. Le point mort non-vie est parmi les plus avancés et le segment de cette classe d'âge est l'un des plus longs dans la figure.

L'agrégation de la rentabilité a également changé l'ordre de classement des points morts de certains segments. Par exemple les assurés «  $\geq 65$  ans » qui rentabilisaient leurs polices en moyenne avant les assurés « 60-64 ans » chez l'assureur non-vie, se voient attribuer un point mort postérieur à ces derniers dans la vision consolidée du groupe d'assurance. Ce qui veut dire que le poids du segment « 60-64 ans » dans la marge vie est plus grand que celui des «  $\geq 65$  ans ».

Au bout de la dixième année, 75% du portefeuille actuel du groupe d'assurance aurait commencé à générer une rentabilité positive selon la fonction de répartition empirique présentée ci-dessous.

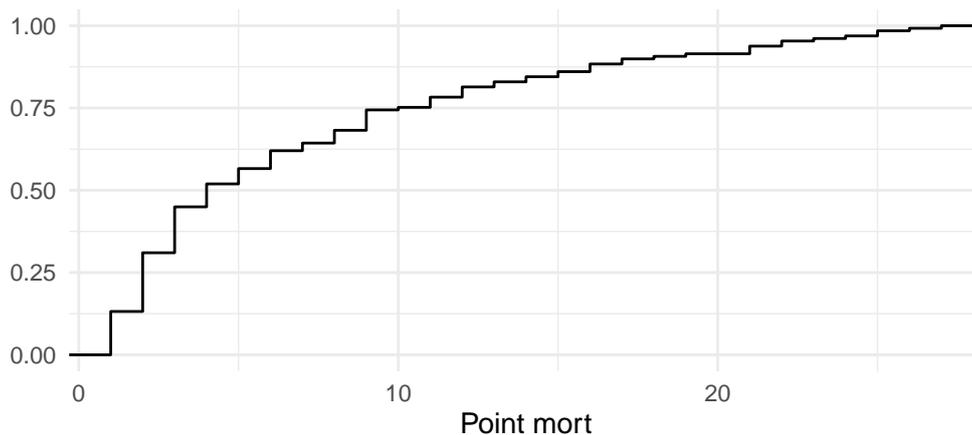


FIGURE 3.4 – Fonction de répartition empirique du point mort

La valeur du point mort prend des valeurs allant jusqu'à la 27<sup>ème</sup> année. Il apparaît que la valeur moyenne du point mort pour cette base de données est de 7 ans. C'est-à-dire, qu'en moyenne le groupe d'assurance a besoin de 7 années pour réaliser du profit sur ce stock de polices.

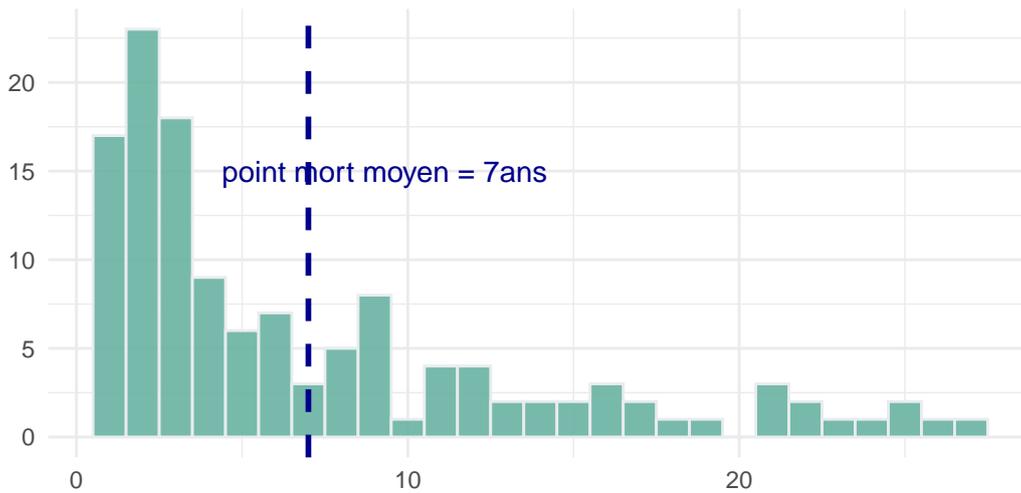


FIGURE 3.5 – Histogramme des points morts

En contraste, le point mort moyen des assurés non-vie d'ancienneté inférieure ou égale à 1 an est de 10 ans selon le mémoire de [Pham \(2022\)](#), qui a étudié le point mort des assurés IARD. Le passage de la marge non-vie à la marge globale a fait avancer le point mort de 3 ans.

Dans le contexte de la planification stratégique, l'attention de l'assureur se concentre spécifiquement sur la rentabilité vis-à-vis son portefeuille de produits. Par conséquent, il est essentiel d'examiner la répartition des segments morts au sein des différentes configurations d'équipement envisageables. Pour isoler l'effet de la détention d'un produit d'épargne sur le point mort, nous allons utiliser par la suite la base 1 qui tourne autour de la marge non-vie uniquement.

La figure ci-après représente les disparités des points morts observées entre les divers produits IARD, qu'ils soient souscrits de manière indépendante ou en association avec une police d'assurance vie pour les assurés en commun entre les deux filiales.

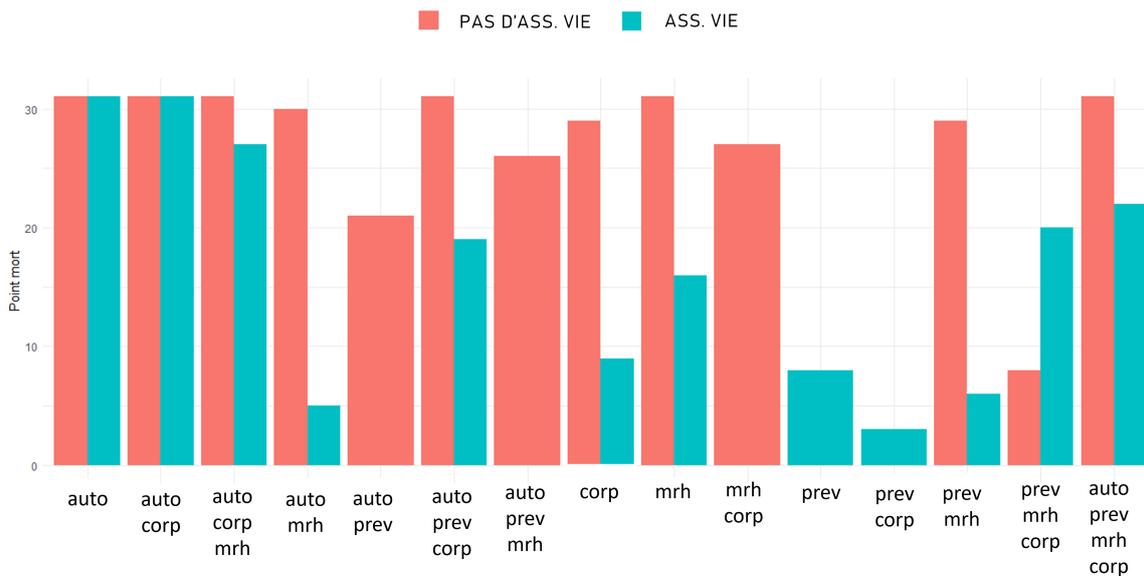


FIGURE 3.6 – Répartition des segments morts par équipement

En résumé, le graphique met en évidence que l'adhésion à une police d'assurance vie influence favorablement la survenance du point mort. Cette observation est particulièrement manifeste chez les assurés qui détiennent à la fois des polices automobile (AUTO) et multirisques habitation (MRH) : le point mort de ces derniers avec un produit d'épargne intervient 25 ans avant celui des assurés qui n'ont aucun produit d'assurance vie.

## 3.2 Modèle de survie

Si on considère le point mort comme étant la durée avant la survenance de l'événement « changement du signe négatif de la marge totale cumulée », il devient apparent que le cadre des modèles de survie est pertinent et approprié pour les fins de cette étude. Les modèles de durée ont pour but d'expliquer les valeurs prises par des durées ou par, de façon générale, des variables aléatoires positives, en fonction des valeurs de certaines caractéristiques. En outre, ces modèles se caractérisent par deux particularités vis-à-vis les données incomplètes.

La première concernant les données tronquées : il s'agit des données manquantes dues au fait que la variable aléatoire n'est observable que sur une sous partie de  $[0, +\infty[$  (Planchet, 2022). Les données tronquées se produisent lorsque certaines observations de l'échantillon ne sont pas incluses dans l'analyse en raison de certaines restrictions ou limitations de l'étude. Par exemple, si on étudie la durée de vie d'une certaine population de patients, mais que seuls les patients qui ont survécu au moins un certain temps sont retenus dans l'étude, alors les données sont tronquées. Cela signifie qu'il n'est pas possible d'observer les événements (dans ce cas, les décès) qui surviennent avant le temps de troncature.

La seconde particularité aborde le point des données censurées : on parle de censure, lorsque le résultat de l'expérience n'est observé que partiellement pour certains individus. Les données censurées se produisent lorsqu'on ne dispose pas d'informations complètes sur le moment où l'événement étudié a eu lieu pour ces observations. Dans la suite pour un individu  $i$ , en reprenant les notations de Pierre (2021), considérons :

- son temps de survie  $X_i$  ;
- son temps de censure  $C_i$  ;
- la durée réellement observée  $T_i$ .

De façon générale, deux types de censure existent :

- censure à droite : dans ce cas, les observations sont suivies pendant un certain temps, mais l'événement d'intérêt n'est pas observé pour toutes les observations pendant la période d'étude. Par exemple, dans une étude de survie, si un patient est encore en vie à la fin de l'étude, la durée de vie est censurée à droite car nous ne savons pas quand il décédera après la fin de l'étude. On utilise la notation suivante :

$$T_i = X_i \wedge C = \min(X_i, C)$$

Ce type de censure à droite est souvent connu sous le nom de censure de type I. Quand  $C$  est aléatoire, on parle de censure de type III.

- censure à gauche : cela se produit lorsque les données sont censurées avant un certain point dans le temps. Par exemple, si l'étude porte sur le temps qu'il faut pour qu'une machine tombe en panne, mais que la durée de vie de la machine avant son achat n'est pas connue, alors les données sont censurées à gauche. Pour chaque observation, on peut associer un couple de variables aléatoires  $(T_i, \delta_i)$  :

$$T_i = X_i \wedge C = \max(X_i, C)$$

$$\delta_i = \mathbb{1}_{X_i \geq C}$$

Le problème de troncature diffère de celui de la censure au sens où la troncature concerne l'échantillonnage lui-même. Ainsi, une variable  $X$  est tronquée par un sous-ensemble  $A$  de  $\mathbb{R}^+$  si au lieu de  $X$ , on observe  $X$  uniquement si  $X$  appartient à  $A$ .

Par exemple, dans le cas de notre étude, les données sont tronquées du fait qu'on n'observe que le stock des polices actives dans le portefeuille actuel : les polices résiliées, annulées ou à terme ne sont pas prises en compte.

Les données sont également censurées à droite, parce que l'horizon de projection limite le cadre de l'étude. On ignore l'existence d'un point mort pour les individus ayant une marge cumulée négative au bout de la dernière année. La variable  $C$  présentée précédemment prend ici la valeur 31 ans.

Le problème de censure à gauche, quant à lui, est résolu pour cet échantillon. Car, comme précisé au début du chapitre, l'étude se concentre exclusivement sur les nouvelles souscriptions, vu le biais induit par l'ancienneté. L'échantillon est constitué de polices d'assurance ayant une ancienneté maximale d'une année.

De plus, la loi de probabilité de la durée étudiée peut être définie par l'une des fonctions suivantes :

- **fonction de densité** : puisque le point mort est une quantité exprimée en nombre d'années, alors  $T$  est une variable positive discrète. Sa distribution est décrite par les  $p_k = P(T = k)$ .

- **fonction de survie** : elle représente la probabilité d'avoir survécu au-delà d'un instant  $k$ , c'est-à-dire que l'observation garde une marge cumulée négative jusqu'à cet instant. Elle s'écrit simplement comme :

$$S(k) = \sum_{m \geq k+1} p_m$$

- **fonction de hasard** : ou fonction de risque, évalue le risque que l'évènement survienne à un instant précis, sachant qu'il n'a jamais été observé auparavant.

Elle se définit comme :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \ln S(t)$$

Dans le cas discret, cette expression devient :

$$h(k) = P(T = k | T > k - 1) = \frac{p_k}{S(k - 1)}$$

Elle permet d'étudier si et quand un individu a changé d'état depuis le début de l'étude et se retrouve donc au cœur de l'analyse des durées de survie. Ses variations permettent d'identifier les instants où le risque est élevé mais aussi d'étudier comment il évolue au fur et à mesure que le temps s'écoule.

### 3.2.1 Estimateur de Kaplan-Meier

En pratique, on estime la fonction de survie à partir des données où l'estimateur le plus fréquemment utilisé est celui Kaplan-Meier. L'idée derrière cet estimateur dans le cas de l'étude de la mortalité est la suivante : survivre après un temps  $t$  c'est être en vie (plus globalement, c'est que l'évènement étudié n'a pas eu lieu) juste avant  $t$  et ne pas mourir au temps  $t$  (l'évènement ne se réalise pas à l'instant  $t$ ),

c'est-à-dire, si  $t'' < t' < t$  :

$$\begin{aligned} P(X > t) &= P(X > t', X > t) \\ &= P(X > t | X > t') \times P(X > t') \\ &= P(X > t | X > t') \times P(X > t' | X > t'') \times P(X > t'') \end{aligned}$$

En considérant les temps d'événements distincts  $T_{(i)}$  ( $i = 1, \dots, n$ ) rangés par ordre croissant, on obtient :

$$P(X > T_{(j)}) = \prod_{k=1}^j P(X > T_{(k)} | X > T_{(k-1)})$$

avec  $T_{(0)} = 0$ . Considérons les notations suivantes :

- $Y_i$  le nombre des individus à risque de subir l'événement juste avant le temps  $T_{(i)}$  ;
- $d_i$  le nombre des individus qui ont subi l'événement en  $T_{(i)}$  .

Alors la probabilité  $p_i$  que l'événement se réalise dans l'intervalle  $]T_{(i-1)}, T_{(i)}]$  sachant que l'événement n'a pas eu lieu en  $T_{(i-1)}$ , i.e.  $p_i = P(X \leq T_{(i)} | X > T_{(i-1)})$ , peut être estimé par :

$$\hat{p}_i = \frac{d_i}{Y_i}$$

Comme les temps d'événements sont supposés distincts, on a :

- $d_i = 0$  en cas de censure en  $T_{(i)}$ , i.e. quand  $\delta_i = 0$  ;
- $d_i = 1$  en cas de décès en  $T_{(i)}$ , i.e. quand  $\delta_i = 1$ .

On obtient alors l'estimateur de Kaplan-Meier, pour des données censurées :

$$\hat{S}_i = \prod_{\substack{(i=1, \dots, n) \\ T_{(i)} \leq t}} \left(1 - \frac{\delta_i}{Y_i}\right) = \prod_{i: T_{(i)} \leq t} \left(1 - \frac{\delta_i}{n - (i - 1)}\right) = \prod_{i: T_{(i)} \leq t} \left(1 - \frac{n - i}{n - (i - 1)}\right)^{\delta_i}$$

où  $\hat{S}(t)$  est une fonction en escalier décroissante, continue à droite.

Sur le plan pratique, la même base de données manipulée pour déterminer le point mort des assurés communs est utilisée. Mais cette fois-ci, les observations qui avaient une marge de rentabilité négative au bout de l'horizon de projection sont réinjectées dans la base de données pour tenir compte de la censure.

Pour différencier les observations censurées des observations non-censurées, une colonne *statut* a été rajoutée. Elle prend la valeur *FALSE* lorsque l'observation n'est pas observée pleinement, c'est le cas des individus qui ont une marge négative pour la 31<sup>ème</sup> année. On utilisera les packages *survival* et *survminer* pour l'estimation de la courbe par Kaplan-Meier sur R.

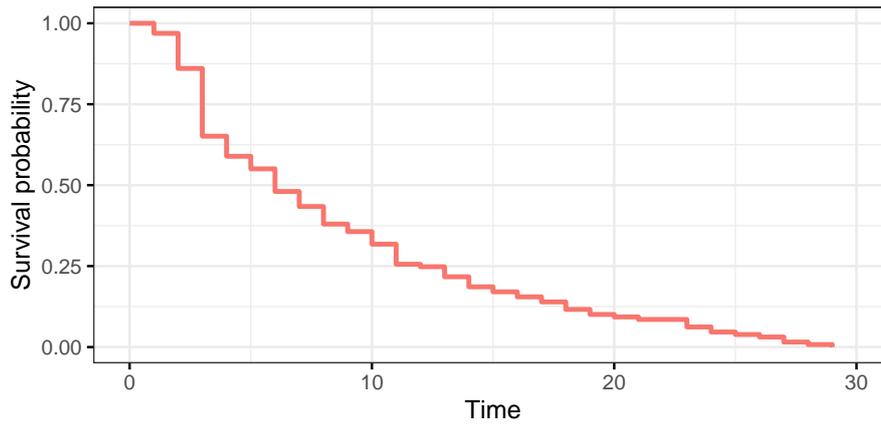


FIGURE 3.7 – Fonction de survie des assurés des deux filiales estimée par Kaplan-Meier

Il s'avère alors qu'au bout de la sixième année, la probabilité de survie atteint le seuil de 50%. A partir de cette année, les assurés communs entre les deux filiales ont plus de chance de rentabiliser l'ensemble de leurs polices, précédemment non profitable, que de rester déficitaires. En outre, nous observons que le saut le plus significatif se produit en transition de la deuxième année vers la troisième année, où la probabilité de survie passe de 86% à 65%.

On peut également estimer la fonction du hasard cumulé  $H(t) = -\ln(S(t))$ . Elle peut être interprétée comme la force cumulative du risque étudié. En d'autres termes, elle correspond au nombre d'événements auxquels on s'attendrait pour chaque individu au moment  $t$  si l'événement était un processus répétable.

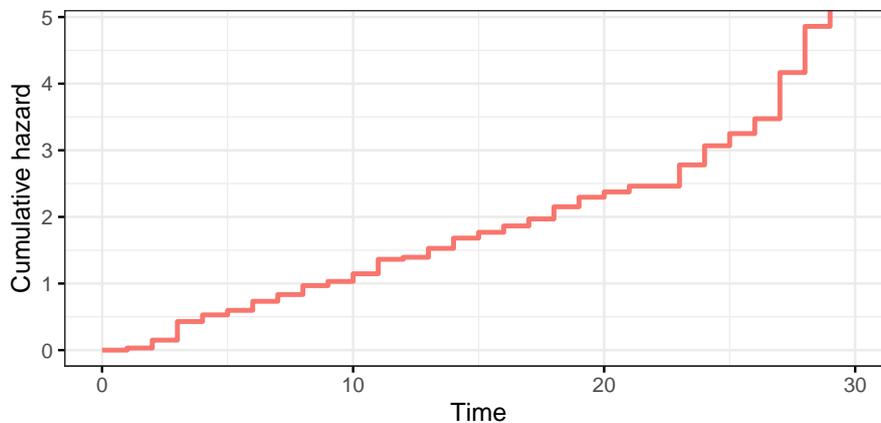


FIGURE 3.8 – Fonction de hasard cumulé des assurés des deux filiales

### 3.2.2 Estimation de la variance

La formule de Greenwood propose une estimation de la variance de la fonction de survie  $S$  :

$$\hat{\sigma}^2(\hat{S}(t_j)^2) = \hat{S}(t_j)^2 \sum_{i=1}^j \frac{d_i}{n_i(n_i - d_i)}$$

avec  $n_i$  le nombre d'observations restantes non censurées juste avant  $T_{(i)}$ .

Sous R, la sortie *std.error* donne le terme de la somme mentionnée dans la formule ci-dessus. On la multiplie par le carré des estimateurs de survie pour avoir la valeur de l'estimation de la variance.

Dans chaque intervalle de temps, l'estimation de la survie est une proportion. On peut donc, sous certaines conditions, faire une approximation par la loi normale et ainsi obtenir un intervalle de confiance :

$$IC_\alpha = [\hat{S}(t) \pm z \frac{\alpha}{2} \sqrt{\hat{\sigma}(\hat{S}(t))}]$$

avec  $z$  : fractile d'ordre  $\alpha$  de la loi normale.

Cette illustration présente l'estimation de la survie accompagnée de son intervalle de confiance à un niveau de confiance de 95%.

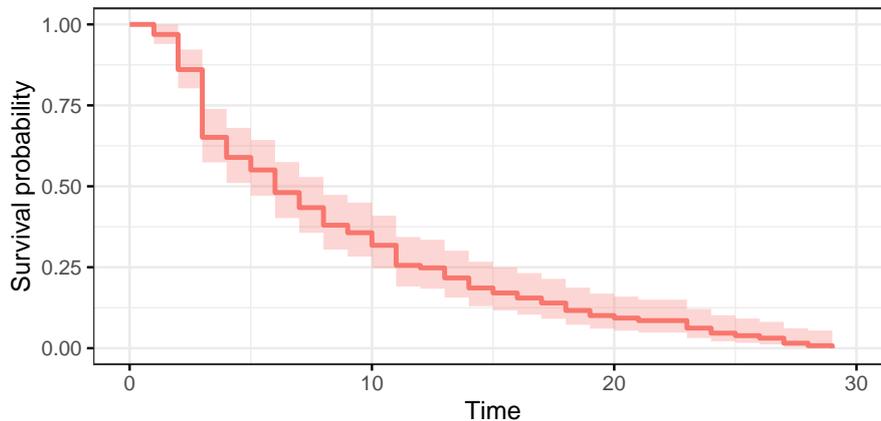


FIGURE 3.9 – Fonction de survie avec IC 95% des assurés des deux filiales

### 3.2.3 Comparaison des courbes

Il est possible de comparer les survies de deux groupes à un instant  $t$  donné. En effet, la survie au temps  $t$  est une proportion, ainsi, en utilisant l'approximation de la loi binomiale par la loi normale on montre que la statistique suivante suit asymptotiquement une loi  $\mathcal{N}(0, 1)$  sous l'hypothèse  $H_0 : S_A(t) = S_B(t)$ .

$$\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\sqrt{\hat{v}ar(\hat{S}_A(t)) - \hat{v}ar(\hat{S}_B(t))}}$$

Néanmoins, cela ne permet pas de tester globalement l'égalité des distributions de survie ce qui limite l'intérêt de la méthode. La faiblesse de cette approche est qu'elle ne fournit pas une comparaison de l'expérience de survie totale des deux groupes, mais plutôt une comparaison à un (des) point(s) de temps arbitraire(s).

Nous décrivons ici le test du log-rank, la méthode la plus populaire pour comparer la survie des groupes et qui prend en compte l'ensemble de la période de suivi. Elle présente l'avantage considérable de ne pas considérer la forme de la courbe de survie ni celle de la distribution des temps de survie.

Le test du log-rank est utilisé pour tester, à tout moment, l'hypothèse nulle selon laquelle il n'y a pas de différence entre les populations en ce qui concerne la probabilité d'un événement. L'analyse est basée sur les temps des événements. Pour chacun de ces moments, nous calculons le nombre observé des décès dans chaque groupe et le nombre attendu s'il n'y avait en réalité aucune différence entre les groupes.

Le principe du test consiste à ordonner par ordre croissant les individus par les temps observés dans les deux groupes A et B réunis. Et on note :

- $d_{B,i}$  : nombre de décès observés au temps  $T_{(i)}$  dans le groupe B.
- $R_{B,i}$  : nbre de sujets exposés au risque de décès juste avant  $T_{(i)}$ , dans le groupe B.
- $e_{B,i}$  : nombre de décès attendus (i.e sous  $(H_0)$ ) au temps  $T_{(i)}$  dans le groupe B tel que :

$$e_{B,i} = \frac{d_{A,i} + d_{B,i}}{R_{A,i} + R_{B,i}} \times R_{B,i}$$

- $w_i$  : poids associé au temps  $T_{(i)}$ . Pour le test du log-rank,  $w_i = 1 \forall i = 1, \dots, n$ . En changeant les poids, on obtient les statistiques de différents tests ( Gehan-Wilcoxon, Peto/Prentice, ...).

La statistique de test compare les décès observés dans le groupe B aux décès attendus sous  $(H_0)$  dans le groupe B :

$$U = \sum_{i=1}^l w_i \times (d_{B,i} - e_{B,i})$$

Le test du log-rank a toutes les chances de détecter une différence entre les groupes lorsque le risque d'un événement est systématiquement plus élevé dans un groupe que dans l'autre. Il est peu probable qu'il détecte une différence lorsque les courbes de survie se croisent. Lors de l'analyse des données de survie, les courbes de survie doivent toujours être tracées.

Les figures ci-dessous représentent les courbes de survies estimés pour les différents groupes, ainsi que la p-value du test log-rank relative à chaque variable. Nous avons analysé les disparités au sein des classes d'âge (A.1) et des segments du canal de souscription.

Le graphique suivant montre les résultats pour la variable `canal` ; où il existe deux types de souscription :

- interne : qui cible une clientèle particulière et offre des tarifs privilégiés ;
- externe : qui reste accessible au grand public.

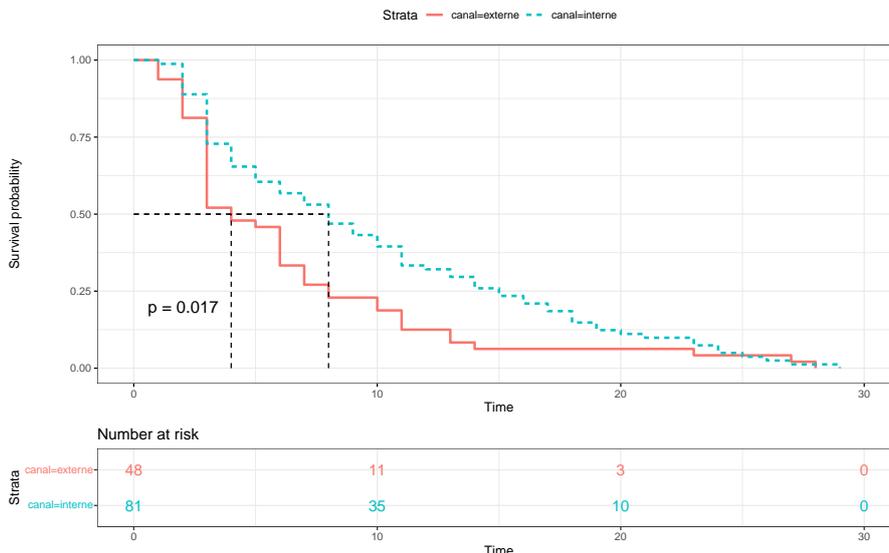


FIGURE 3.10 – Comparaison des courbes de survie selon le canal

La p-value du test du log-rank est égale à 0.017 ; d'où la différence significative entre les courbes de survie des deux segments d'assurés. Il apparaît que les contrats souscrits en interne sont moins

favorables en termes de rentabilité : la courbe de survie des *externes* se trouvent au-dessous de celle des *internes*, et donc pour la même année les polices souscrites en externe ont moins de chance de survivre et donc plus de chance de changer de signe que les polices distribuées en interne.

Le graphique pour la variable âge est présent dans l'annexe A. Il dévoile une p-value significative inférieure à 5%.

Une fois de plus, pour évaluer l'influence de la possession d'un contrat d'épargne, nous allons procéder à une modification de la base de données afin d'y intégrer tous les assurés non-vie qui ont une ancienneté maximale d'un an. La mesure de rentabilité prise en considération est basée exclusivement sur la marge non-vie. Le graphe ci-après montre les résultats du test log-rank pour la variable `detention_epargne` :

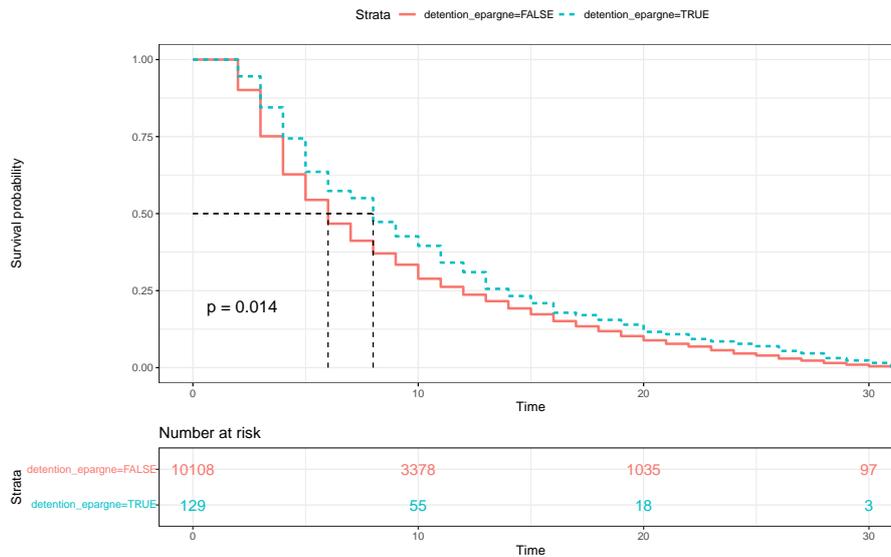


FIGURE 3.11 – Comparaison des courbes de survie selon la détention d'un produit d'épargne

On remarque que la courbe de survie des détenteurs d'un produit d'épargne se trouve au-dessus de celle des non-détenteurs. Ceci signifie que la probabilité de garder une valeur négative de rentabilité est supérieure pour les détenteurs puisqu'ils ont plus de chance de survivre à l'évènement changement de signe. En d'autres termes, les assurés n'ayant pas souscrit à un produit d'assurance vie ont une plus grande probabilité pour devenir rentables plus tôt. Ce résultat est plutôt contre-intuitif, puisque d'après les résultats précédent, l'effet de la détention d'un produit d'épargne était favorable sur la rentabilité des assurés non-vie.

Mais comme ce test est purement un test de signification, il ne peut fournir une estimation de la taille de la différence entre les groupes ou un intervalle de confiance. Pour cela, nous devons faire des hypothèses sur les données. Les méthodes courantes utilisent le rapport des risques, y compris le modèle des risques proportionnels de Cox, que nous décrirons dans la partie suivante et qui permettra une exploration plus approfondie des implications de la détention d'un produit d'épargne.

### 3.3 Impact des variables

Les méthodes mentionnées ci-dessus - courbes de Kaplan-Meier et tests du Log-rank - sont des exemples d'analyse univariée. Elles décrivent la survie en fonction d'un facteur étudié, mais ignorent l'impact des autres facteurs. Afin d'élargir notre compréhension des dynamiques des profils de rentabilité, il serait pertinent d'intégrer les autres variables de la base de données et d'élucider la manière dont elles exercent leur influence sur le point mort.

#### 3.3.1 Modèle de Cox

C'est dans ce sens, que nous adopterons le modèle de régression proportionnelle de Cox. Il s'agit d'une des méthodes les plus largement utilisées pour modéliser les données d'analyse de survie. La section suivante présente les bases du modèle de régression de Cox.

L'objectif du modèle est d'évaluer simultanément l'effet de plusieurs facteurs sur la survie. Il permet d'examiner comment certains facteurs influencent la fonction de hasard. Les variables prédictives sont généralement appelées covariables dans la littérature de l'analyse de survie.

Le modèle de Cox est exprimé par la fonction de hasard estimée comme suit :

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 + \dots + \beta_p x_p)$$

où :

- $(x_1, x_2, \dots, x_p)$  représente le vecteur de p covariables ;
- $(\beta_1, \beta_2, \dots, \beta_p)$  est le vecteur des coefficients qui mesurent l'effet des variables explicatives ;
- $h_0(t)$  appelé la fonction de hasard de base et correspond au taux de risque lorsque toutes les variables sont nulles ou correspondent aux modalités de référence.

Le modèle de Cox peut être écrit comme une régression linéaire multiple du logarithme du risque sur les variables  $x_i$ , avec le risque de base étant un terme « *intercept* » qui varie avec le temps.

Les quantités  $\exp(\beta_i)$  sont appelées risque proportionnel (HR pour *hasard ratio* en anglais). Une valeur de  $\beta_i$  supérieure à zéro, ou de manière équivalente, un risque proportionnel supérieur à 1, indique que plus la valeur de la covariable  $i$  augmente, le risque d'événement augmente et donc la durée de survie diminue. En d'autres termes, un risque proportionnel supérieur à 1 indique que la covariable est positivement associée à la probabilité de réalisation de l'événement, et donc négativement associée à la durée de survie.

Ce modèle repose sur la notion de risque proportionnel : considérons deux individus k et k'. La fonction hasard s'écrit respectivement pour chacun des deux, comme suit :

$$h_k(t) = h_0(t) e^{\sum_{i=1}^p (\beta_i x_i^k)} \quad \text{et} \quad h_{k'}(t) = h_0(t) e^{\sum_{i=1}^p (\beta_i x_i^{k'})}$$

Le taux de risque, ou de hasard, pour ces deux individus se calcule comme le rapport entre leurs fonctions de hasard. Il représente une quantité indépendante du temps :

$$\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t) e^{\sum_{i=1}^p (\beta_i x_i^k)}}{h_0(t) e^{\sum_{i=1}^p (\beta_i x_i^{k'})}} = e^{\sum_{i=1}^p \beta_i (x_i^k - x_i^{k'})}$$

Par conséquent, le modèle de Cox est un modèle à risques proportionnels ; le risque de l'événement dans un groupe quelconque est un multiple constant du risque dans tous les autres groupes. Cette

hypothèse implique que, comme mentionné ci-dessus, les courbes de risque pour les groupes doivent être proportionnelles et ne peuvent pas se croiser.

Dans cette étude, la notion de proportionnalité se traduit par l'exemple suivant : si le « risque » de devenir rentable pour un segment d'assurés à un moment donné est deux fois plus élevé que celui d'un autre segment, ce risque restera deux fois plus élevé à tous les moments ultérieurs. C'est une conséquence du modèle mais c'est aussi une hypothèse qu'il faudra vérifier.

Pour la mise en pratique sur R, et en se basant sur les résultats du test log-rank, le modèle choisi est le suivant :

$$h(t) = h_0(t) \times \exp(\beta_1 \times canal + \beta_2 \times detention\_epargne + \beta_3 \times nombre\_polices)$$

La variable `nombre_police` représente le nombre total des produits détenus par l'assuré. L'âge n'a pas été inclus dans le modèle puisqu'il s'agit d'une variable qui dépend du temps. Les résultats de cette régression se présentent dans le tableau qui suit :

	coef	exp(coef)	se(coef)	z	Pr(> z )
canal INTERNE	-0.66569	0.51392	0.02101	-31.684	< 2e - 16
detention epargne TRUE	-0.49990	0.60659	0.09062	-5.517	3.46e - 08
nombre polices	0.22329	1.25019	0.01379	16.189	< 2e - 16

TABLE 3.4 – Résultats du modèle de Cox

D'après les résultats ci-dessus, nous pouvons conclure que les trois variables ont des coefficients statistiquement significatifs.

La deuxième caractéristique à noter dans les résultats du modèle de Cox est le signe des coefficients de régression. Un signe positif signifie que le hasard (risque de changer de signe) est plus élevé, et donc que le pronostic est meilleur dans ce cas, pour les sujets ayant des valeurs plus élevées de cette variable :

- Le coefficient bêta pour le canal = -0,67 indique que les polices *internes* ont un « risque » de rentabilisation plus faible (taux de survie plus élevé) que les polices *externes*. Les polices souscrites en interne connaissent un point mort plus tardif que les polices distribuées en externe. Ce qui rejoint les résultats obtenus par le test du log-rank.
- La même analyse s'applique à la détention d'une police d'assurance vie qui se caractérise par un coefficient négatif. L'absence de cette variable a un effet bénéfique sur le point mort.
- En revanche, le nombre de polices a un coefficient positif, ce qui signifie que le risque de changer de signe, et donc que l'assuré devienne rentable, s'accroît avec le nombre de polices IARD souscrites.

Les coefficients exponentiels donnent l'ampleur de l'effet des covariables. C'est ainsi que le nombre des produits des assurés impacte le plus le point mort, suivi par l'effet de la détention d'un produit d'épargne qui réduit le risque d'un facteur de 0,60, soit 40 % par rapport aux non-détenteurs. Et en dernière position, la variable canal où les internes ont 49% moins de chance que les externes de devenir rentables pour une année donnée.

Enfin, la sortie ci-dessous donne les p-values pour trois tests alternatifs de signification globale du modèle : le test du rapport de vraisemblance, le test de Wald et la statistique du score log-rank. Ces trois méthodes sont asymptotiquement équivalentes. Lorsque N est suffisamment grand, elles donnent des résultats similaires. Dans ce cas, les trois tests confirment que le test est globalement statistiquement significatif.

```

Concordance= 0.617 (se = 0.003 )
Likelihood ratio test= 1074 on 3 df, p=<2e-16
Wald test              = 1084 on 3 df, p=<2e-16
Score (logrank) test = 1108 on 3 df, p=<2e-16

```

FIGURE 3.12 – Résultats des tests de signification du modèle de Cox

### 3.3.2 Vérification des hypothèses

#### La proportionnalité

Pour tester cette hypothèse sur R, on fait appel à la fonction `cox.zph()`.

```

                chisq df      p
canal           38.04  1 6.9e-10
detention_epargne  1.18  1  0.28
nb_po           77.20  1 < 2e-16
GLOBAL          154.06  3 < 2e-16

```

FIGURE 3.13 – Résultats du test de proportionnalité pour le modèle de Cox

D'après les résultats ci-dessus, le test est statistiquement significatif pour chacune des covariables hormis la variable `detention_epargne`. Ce qui veut dire que les variables `canal` et `nombre_polices` n'ont pas un effet constant dans le temps. De plus, le test global est statistiquement significatif. Par conséquent, nous ne pouvons pas supposer que les risques sont proportionnels.

Il est également possible d'effectuer un diagnostic graphique à l'aide de la fonction `ggcoxzph()` qui produit pour chaque covariable, des graphiques des résidus de Schoenfeld mis à l'échelle en fonction du temps transformé. Ces résidus concernent les covariables et non la fonction de hasard, et ne sont estimés que pour les cas non-censurés. Le résidu lié à une covariable représente l'écart entre la valeur prise par cette variable pour un individu au moment de la survenance de l'événement et la moyenne de cette covariable parmi tous les individus exposés au risque à ce moment. La formule suivante représente la formule de ces résidus dans le cas d'une covariable  $x$  pour le  $i$ -ème événement :

$$x_i - \bar{x}_i(\hat{\beta})$$

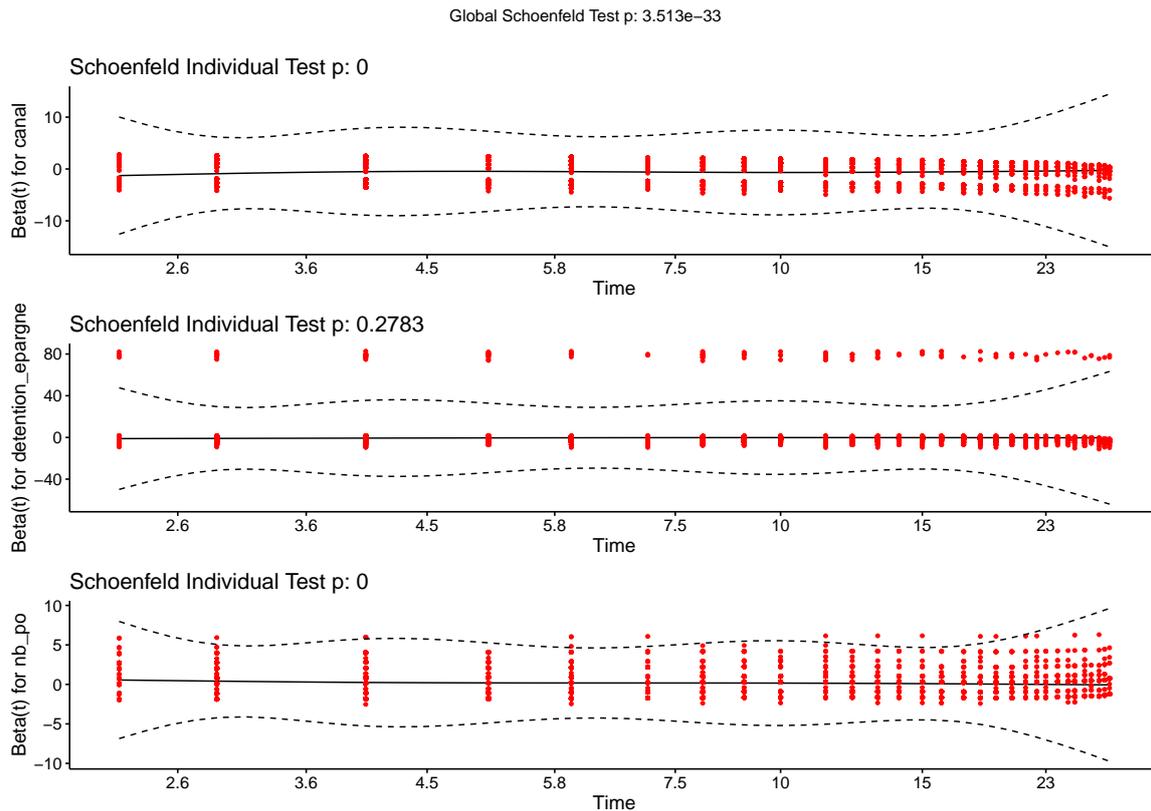
Dans le cas de plusieurs covariables  $x_k$ , chacune fournira ses résidus :

$$x_{ki} - \bar{x}_{ki}(\hat{\beta})$$

Les résidus de Schoenfeld servent à tester l'hypothèse des risques proportionnels :

- si l'hypothèse des risques proportionnels est vérifiée, ces résidus ont en théorie un aspect totalement aléatoire et l'évolution temporelle moyenne est une droite horizontale ;
- sinon, par exemple si le facteur de risque est important au début du suivi du patient mais pas à la fin, alors les résidus seront, sur le schéma, négatifs puis positifs, et l'évolution temporelle moyenne sera une courbe croissante (ou le contraire selon le codage du facteur risque).

R permet de générer les sorties suivantes pour analyser la distribution des résidus de Schoenfeld liés à la régression de Cox proposée dans la partie précédente. Dans les figures ci-dessous, la ligne continue est une spline de lissage ajustée au tracé, les lignes en pointillé représentant une bande d'erreur de autour de l'ajustement.



Il convient de noter que les écarts systématiques par rapport à une ligne horizontale indiquent des risques non proportionnels, puisque les risques proportionnels supposent que les estimations  $\beta_1, \beta_2, \beta_3$  ne varient pas beaucoup dans le temps.

Il est difficile de détecter visuellement un pattern avec le temps pour les variables `canal` et `nombre_polices`. Les trois plots cependant montrent une évolution plutôt horizontale avec le temps. Toutefois, nous nous baserons sur les résultats des *p-value* pour décider de la violation de l'hypothèse de proportionnalité : la seule variable qui vérifie cette hypothèse dans ce contexte est `detention_epargne`.

### La log-linéarité

Le modèle de Cox repose sur une deuxième hypothèse fondamentale qui est la log-linéarité des variables : nous supposons souvent que les covariables continues ont une forme linéaire.

Le tracé des résidus de Martingale en fonction des covariables continues est une approche communément utilisée pour détecter la non-linéarité. Pour une covariable continue donnée, les motifs du graphique peuvent suggérer si la variable n'est pas correctement ajustée.

Néanmoins, la non-linéarité ne constitue pas un problème pour les variables catégorielles. Et étant donné que toutes les variables de la base de données sont de nature catégorielle, il n'est pas nécessaire d'examiner cette hypothèse dans le cadre de cette étude.

## Stratification

Une solution pour remédier à la non-conformité à l'hypothèse de proportionnalité est la stratification. Cette méthode consiste à calculer un modèle de Cox en attribuant une valeur différente du risque de base  $h_0(t)$  à chaque catégorie de la variable de stratification. En revanche, l'influence des variables explicatives, et donc les valeurs estimées des paramètres  $\beta$ , est commune à toutes les catégories. Il est aussi possible de stratifier en fonction de plusieurs variables simultanément. Si on considère la variable `canal` par exemple, la stratification permet de générer les deux modèles de Cox suivants :

$$h_{interne}(t) = h_{0,interne}(t) \exp(\beta_1 \times \text{detention\_epargne} + \beta_2 \times \text{nombre\_polices})$$

$$h_{externe}(t) = h_{0,externe}(t) \exp(\beta_1 \times \text{detention\_epargne} + \beta_2 \times \text{nombre\_polices})$$

Le modèle stratifié sur R donne le tableau suivant :

	coeff	exp(coef)	se(coef)	z	Pr(> z )
detention epargne TRUE	-0.48588	0.61515	0.09074	-5.355	8.58e-08
nombre polices	0.22586	1.25340	0.01381	16.355	< 2e - 16

TABLE 3.5 – Résultats du modèle de Cox stratifié par rapport au canal

Ensuite, il convient d'évaluer si des modèles séparés sont plus performants que le modèle stratifié. Le modèle stratifié de Cox permet aux deux groupes de la variable de stratification d'avoir des risques de base différents, mais calcule le même coefficient pour les autres covariables. On peut alors comparer ce modèle aux modèles de Cox élaborés pour chaque groupe de la variable de stratification et qui calculent des coefficients différents pour les autres covariables sur chacun des groupes :

$$h_{interne}(t) = h_{0,interne}(t) \exp(\beta_{1,interne} \times \text{detention\_epargne} + \beta_{2,interne} \times \text{nombre\_polices})$$

$$h_{externe}(t) = h_{0,externe}(t) \exp(\beta_{1,externe} \times \text{detention\_epargne} + \beta_{2,externe} \times \text{nombre\_polices})$$

Pour ce faire, le test du rapport de vraisemblance permet de comparer les deux modèles. La *p-value* du test utilisé correspond à :

$$P(\chi_{(k-1)(q)}^2) > \ell_{\text{modèle stratifié}} - \sum \ell_{\text{modèle séparé}}$$

avec  $k$  le nombre des groupes de stratification,  $q$  le nombre des covariables et  $\ell$  la log-vraisemblance du modèle. Le test affiche une *p-value* = 0.000852, d'où le rejet de l'égalité des modèles stratifiés et séparés. Les modèles séparés apportent alors une amélioration significative en termes de performance statistique.

Une autre méthode qui permet d'inclure dans un modèle de Cox, une variable ne satisfaisant pas à l'hypothèse de proportionnalité des risques est de tenir compte de l'effet temporel.

### 3.3.3 Modèle de Cox étendu

Dans le cas où les variables explicatives ne sont pas indépendantes du temps, ce qui est le cas pour la variable `age` dans la base de données par exemple, le modèle de Cox étendu substitue le cadre classique.

Le modèle de Cox étendu fait référence à une version plus complexe et flexible du modèle de Cox classique. Il permet de relâcher l'hypothèse du hasard proportionnel et offre plus de flexibilité dans la modélisation des données de survie.

Les variables explicatives s'écrivent sous la forme suivante :  $x_i(t) = x_i \times g_i(t)$  où  $x_i$  est une covariable indépendante du temps. Cependant, comme elle ne répond pas à l'hypothèse de risque proportionnel, elle doit être mise en interaction avec la fonction temporelle. Si toutes les  $p$  covariables ne satisfont pas l'hypothèse de risque proportionnel, alors le nombre de  $p$  covariables doit être mis en interaction avec le temps. La fonction temporelle pour la covariable- $i$  est définie comme  $g_i(t)$ , obtenue par la forme de Cox étendue comme suit :

$$h(t) = h_0(t) \times \exp\left(\sum_{i=1}^p \beta_i x_i + \sum_{i=1}^p \delta_i x_i g_i(t)\right)$$

Si le nombre des variables qui respectent l'hypothèse de proportionnalité est de  $p_1$ , et  $p_2$  le nombre des facteurs qui ne la respectent pas avec  $p = p_1 + p_2$ , le modèle devient :

$$h(t) = h_0(t) \times \exp\left(\sum_{i=1}^{p_1} \beta_i x_i + \sum_{i=p_1+1}^{p_1+p_2} \beta_i x_i + \sum_{i=p_1+1}^{p_1+p_2} \delta_i x_i g_i(t)\right)$$

où la fonction temporelle  $g_i(t)$  peut prendre les formes suivantes de manière générale :

$$g_i(t) = 0 \quad , \quad g_i(t) = t \quad , \quad g_i(t) = \ln(t).$$

Pour intégrer l'évolution des variables dans le temps dans le cadre du modèle de Cox, l'idée consiste à réévaluer la valeur de ces dernières variant dans le temps pour chaque terme (chaque valeur  $t_i$ ). Dans ce cas, un nouvel encodage de la base de données est nécessaire. Selon la vignette R ([Terry Therneau, 2023](#)), la manière la plus simple consiste en l'usage des intervalles de temps ou ce qu'on appelle sous-épisodes. Un nouveau sous-épisode débute à chaque changement de la valeur de la co-variable. En effet, nous allons essayer de créer des variables par période. Considérons la variable âge du modèle par exemple : cette variable est catégorielle et se compose de 10 classes où les neuf premières classes sont des plages de 5 ans, et la dernière concerne les âges supérieurs à 65 ans. Et notons que dans ce cas, l'horizon de projection est de 31 ans. Si on suppose une distribution uniforme des âges, au bout de tous les 5 ans, les observations changeront de classe d'âge jusqu'à atteindre la dernière. Une façon de coder ces données pour la machine est de diviser le période de projection en 10 intervalles de temps, avec une fenêtre de 5 ans pour chaque intervalle correspondant aux classes d'âge. La base de données ressemble à ce qui suit :

ID	$T_{dbut}$	$T_{fin}$	Age	Changement de signe	Detention epargne	Canal	Point mort
1	0	5	55-59 ans	1	TRUE	INTERNE	3
2	0	5	25-29 ans	0	FALSE	EXTERNE	13
2	5	10	30-34 ans	0	FALSE	EXTERNE	13
2	10	15	35-39 ans	1	FALSE	EXTERNE	13
3	0	5	$\geq 65$ ans	0	TRUE	INTERNE	7
3	5	10	$\geq 65$ ans	1	TRUE	INTERNE	7

TABLE 3.6 – Extrait de la base personne-période

Cela signifie que sur l'intervalle de 0 à 5 ans pour le deuxième individu par exemple, l'âge du sujet était dans la classe 25-29 ans et que cet intervalle ne s'est pas terminé par un changement de signe. Ce

n'est que 10 ans après, que la rentabilité de cette observation passe en positif. Le code R sous-jacent traite les intervalles comme ouverts à gauche et fermés à droite.

La nouvelle régression de Cox, en utilisant la nouvelle base de données personnes-périodes et en stratifiant selon la variable *nombre\_police* donne les résultats suivants pour le canal interne :

	coef	exp(coef)	se(coef)	z	Pr(> z )						
detention_epargneTRUE	0.18805	1.20689	0.08298	2.266	0.023445	*					
t_age18-24 ans	0.20261	1.22460	0.04884	4.149	3.35e-05	***					
t_age25-29 ans	-0.11145	0.89454	0.05122	-2.176	0.029573	*					
t_age30-34 ans	-0.01081	0.98924	0.05251	-0.206	0.836828						
t_age35-39 ans	-0.20482	0.81480	0.05455	-3.755	0.000174	***					
t_age40-44 ans	-0.11093	0.89500	0.05206	-2.131	0.033104	*					
t_age45-49 ans	-0.42842	0.65154	0.05149	-8.321	< 2e-16	***					
t_age50-54 ans	-0.98908	0.37192	0.05586	-17.706	< 2e-16	***					
t_age55-59 ans	-1.10897	0.32990	0.05679	-19.527	< 2e-16	***					
t_age60-64 ans	-0.36506	0.69416	0.04810	-7.590	3.20e-14	***					
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

FIGURE 3.15 – Régression de Cox pour le canal interne

Et nous obtenons le tableau ci-après pour le canal externe, :

	coef	exp(coef)	se(coef)	z	Pr(> z )						
detention_epargneTRUE	-0.24852	0.77996	0.14133	-1.758	0.0787	.					
t_age18-24 ans	0.39714	1.48757	0.06185	6.421	1.36e-10	***					
t_age25-29 ans	0.14314	1.15389	0.06211	2.304	0.0212	*					
t_age30-34 ans	-0.59657	0.55070	0.07550	-7.901	2.76e-15	***					
t_age35-39 ans	-0.31869	0.72710	0.07164	-4.448	8.65e-06	***					
t_age40-44 ans	-0.00771	0.99232	0.06906	-0.112	0.9111						
t_age45-49 ans	-0.12124	0.88582	0.06872	-1.764	0.0777	.					
t_age50-54 ans	-0.11661	0.88994	0.06825	-1.708	0.0876	.					
t_age55-59 ans	-0.17371	0.84054	0.07228	-2.403	0.0162	*					
t_age60-64 ans	0.12921	1.13793	0.07474	1.729	0.0839	.					
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

FIGURE 3.16 – Régression de Cox pour le canal externe

Le test du ratio de vraisemblance, ainsi que celui de Wald et de Score génèrent des *p-value* <  $2e-16$  pour les deux catégories du canal de souscription.

On remarque que la classe d'âge la plus jeune a un effet positif sur le point mort des assurés non-vie pour les deux modèles de régression. Pour une année *t*, les assurés entre 18 et 24 ans ont plus de chance de rentabiliser leurs polices IARD que toutes les autres tranches d'âge. En ce qui concerne la détention d'un produit d'épargne, elle impacte favorablement les polices souscrites en interne mais défavorablement celles distribuées selon le mode externe.

Dans le but d'approfondir l'analyse sur les polices externes qui affichent un résultat contre-intuitif pour la variable détention d'une police d'assurance vie, nous allons comparer la marge moyenne des assurés non-vie d'un 1 an d'ancienneté détenant un contrat d'épargne et des non-détenteurs :

Avec épargne	Sans épargne
41.80	222.53

TABLE 3.7 – Marge IARD moyenne pour les assurés d'un 1 an d'ancienneté

Il s'avère que le premier groupe est largement plus déficitaire que le second, ce qui expliquerait le coefficient négatif attribué à la variable `detention_epargne`. Ce constat n'est pourtant pas vrai pour le reste des polices supérieures à 5 ans d'ancienneté où on retrouve :

Ancienneté	Avec épargne	Sans épargne
$\leq 1$ an	41.8026	222.5313
2 - 5 ans	163.7810	183.0415
6 - 10 ans	250.8515	198.2654
11 - 15 ans	315.3590	255.1208
16 - 20 ans	349.5168	304.1636
21 - 27 ans	410.2271	325.4330
$\geq 28$ ans	415.3470	387.5068

TABLE 3.8 – Marge IARD moyenne pour les assurés IARD par ancienneté

Une analyse de la composition est réalisée sur le portefeuille des polices externes qui sont en vigueur depuis au plus un an pour comprendre les causes sous-jacentes du déficit. On constate alors, dans la figure ci-dessous, que le portefeuille des polices externes qui ne détiennent aucun produit d'épargne présente une diversité plus large en matière d'équipements. Cette diversification pourrait être à l'origine des marges plus rentables en faveur de ce groupe d'assurés. Par contre, les polices externes avec un contrat d'assurance vie se concentrent principalement sur quatre combinaisons de produit.

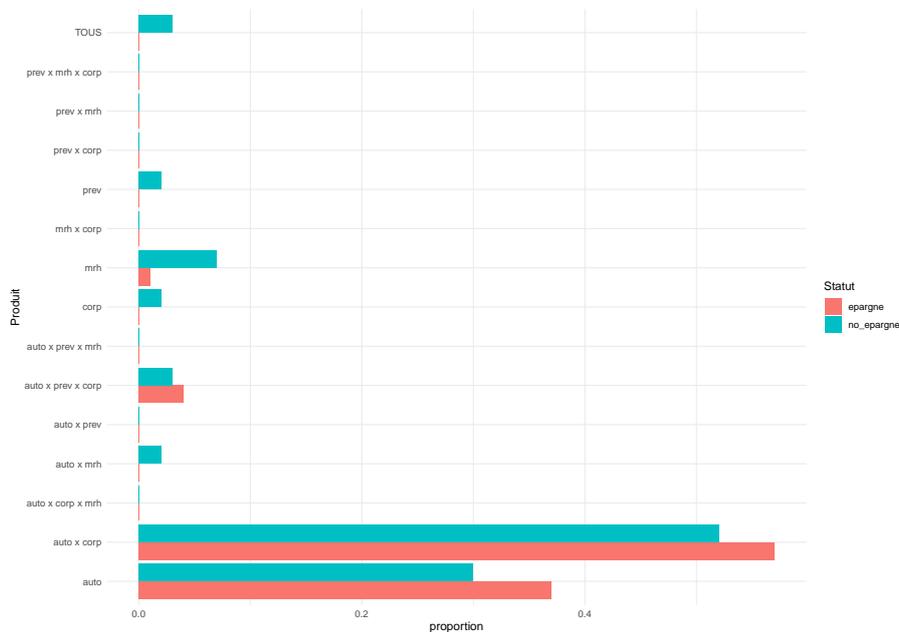


FIGURE 3.17 – Décomposition du portefeuille externe avec 1 an d'ancienneté



# Conclusion

La rentabilité occupe une place centrale dans la gestion de l'activité des compagnies d'assurance. Les indicateurs de rentabilité permettent ainsi à un assureur de mettre en œuvre efficacement ses stratégies commerciales, ainsi que de prendre des décisions éclairées concernant les nouveaux produits à proposer ou les segments de clients à fidéliser. Il est alors primordial qu'il accorde une attention particulière au choix des métriques utilisées pour évaluer la profitabilité de ces segments.

En assurance-vie, il existe plusieurs mesures qui communiquent la valeur d'un portefeuille d'assurés. Dans le cadre de cette étude, la *value of in-force* a été retenue comme indicateur parce qu'elle tient en compte la vision économique de la rentabilité. Elle estime la valeur du stock des polices de l'assureur tout en tenant compte de la valeur temps des options et garanties. Le calcul de la VIF a été assuré par *SimBEL*, un package R qui réalise des projections de cantons d'actifs et de passifs, utilisé en interne chez Prim'Act.

En assurance non-vie, l'assureur dispose d'un modèle markovien multi-états qui calcule la marge de rentabilité par *model point*. Toutefois, ce modèle nécessite des inputs de plusieurs départements : historique des sinistres, table de résiliation des contrats, etc. Il emploie également un grand nombre d'états qui correspondent aux caractéristiques des assurés. L'interprétabilité de ses résultats se révèle complexe, étant donné que l'effet de chaque variable n'est pas explicite. C'est la raison pour laquelle, ces marges ont été reconstruites avec des modèles plus simples, à savoir un modèle GLM et puis un LASSO. Il faut noter que ces modèles attribuent une valeur positive à l'effet de la détention d'un produit d'épargne sur la rentabilité des assurés. Mais les valeurs de leurs  $R^2$  sont proches de 0,5. Ce niveau de performance peut être expliqué par la dépendance entre plusieurs variables.

Au-delà du calcul indépendant des indicateurs de rentabilité vie et non-vie, un groupe d'assurance qui détient des filiales dans les deux branches peut souhaiter avoir une marge de rentabilité consolidée. Cette agrégation est délicate car les deux métriques sont calculées sur des mailles différentes : le calcul en non-vie est réalisé au niveau *model point* ou assuré, alors que les métriques en vie sont souvent calculées suivant une logique portefeuille.

Le deuxième chapitre de ce mémoire s'articule alors autour des méthodes d'allocation de capital appropriées à utiliser pour partager le montant global de la VIF sur les segments de l'assureur. La première méthode repose sur la répartition de la VIF selon le montant des provisions mathématiques de chaque segment. Cette approche est assez simpliste et ne tient pas compte des dépendances marginales. La seconde méthode propose de tenir compte de la contribution marginale de chaque segment à la valeur globale de la VIF. Cependant, l'effet des sous-ensembles est toujours ignoré. La méthode de Shapley est donc la méthode qui semble répondre à cette problématique. Il s'agit d'une méthode initialement introduite par la théorie des jeux coopératifs allouant à chaque segment un poids qui tient compte des dépendances de risques entre les différents segments. L'inconvénient de cette méthode est qu'elle est lourde à mettre en place en présence d'un grand nombre de *model points*. C'est pourquoi une approximation Monte Carlo avec échantillonnage aléatoire a été retenue pour estimer les marges vie

segmentaires.

Le troisième chapitre approfondit l'examen de la rentabilité en explorant les points morts des segments de l'assureur. Le point mort est défini comme l'année où la marge de rentabilité cumulée, qui était négative la première année, devient positive. Cette métrique fournit à l'assureur une indication temporelle de la rentabilité, puisque la police est considérée déficitaire durant toute la période antérieure à son point mort. En se basant sur la théorie des modèles de durée, nous avons soulevé une différence significative entre les courbes de survie et donc la réalisation du point mort des assurés avec et sans produit d'épargne. De plus, le canal de souscription ainsi que l'âge sont des facteurs qui s'avèrent influencer la durée avant la réalisation du point mort.

# Bibliographie

- ACPR. (2016). *Analyses et synthèses : Les différentes composantes de l'assurance vie et leur évolution* (Rapport N° 65). ACPR.
- Arce, P. D. (2022). *Analyse des déterminants de la marge client d'un portefeuille d'assurance*.
- Banque de France. (Juillet 2019). *Le « point mort » : un outil d'aide à la décision*. Consulté sur <https://entreprises.banque-france.fr/le-point-mort-un-outil-daide-la-decision>
- Bocquaire, E., & Camus-Paradis, M. L. (2022). *Les grands principes de l'actuariat*. L'Argus de l'assurance.
- CFO forum. (2008). *Market Consistent Embedded Value Principles*. Consulté sur <https://cfoforum.eu/downloads/CFO-Forum-MCEV-Principles-and-Guidance.pdf>
- CFO forum. (2016). *Market Consistent Embedded Value Basis for Conclusions*. Consulté sur <https://cfoforum.eu/downloads/CFO-Forum-MCEV-Basis-for-Conclusions-April-2016.pdf>
- France Assureurs. (2023). *Communiqué de presse : Assurance vie : net rebond des cotisations au mois de juin, l'encours total dépasse la barre des 1 900 milliards d'euros pour la première fois*.
- Francisco Cribari-Neto, A. Z. (s. d.). *Beta regression in r*. Consulté sur <https://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf>
- Menneymer, D. L. M. (2022). *Life insurers should adopt value of in-force to improve capital position*. Consulté sur <https://www.wtwco.com/en-bm/insights/2022/07/life-insurers-should-adopt-value-of-in-force-to-improve-capital-position>
- Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, USA : The MIT Press. (electronic edition)
- Pham, M. (2022). *Analyse des profils de rentabilité dans un portefeuille d'assurance non-vie*.
- Pierre, P. S. (2018). *Cours sur les modèles multi-états markoviens et semi-markoviens*.
- Pierre, P. S. (2021). *Introduction à l'analyse des durées de survie*. Consulté sur [https://perso.math.univ-toulouse.fr/psaintpi/files/2021/04/Cours\\_Survie\\_1.pdf](https://perso.math.univ-toulouse.fr/psaintpi/files/2021/04/Cours_Survie_1.pdf)
- Planchet, F. (2022). *Support de cours : Modèles de durée*. Consulté sur [http://www.ressources-actuarielles.net/C1256F13006585B2/0/1430AD6748CE3AFFC1256F130067B88E/\\$FILE/Seance1.pdf](http://www.ressources-actuarielles.net/C1256F13006585B2/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance1.pdf)
- Planchet, F. (2023, Feb). *Calcul du SCR : méthode LSMC*.
- Sorochynskyi, O. (2020). *Analyse des déterminants de la marge dans des contrats iard. mém. de mast. strasbourg : Université de strasbourg*.
- Terry Therneau, E. A., Cynthia Crowson. (2023). *Using time dependent covariates and time dependent coefficients in the cox model*. Consulté sur <https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf>
- The American Academy of Actuaries. (March 2011). *A public policy practice note : Market Consistent Embedded Values*. Consulté sur <https://www.actuary.org/content/market-consistent-embedded-values>



# Table des figures

1.1	Décomposition de la VIF . . . . .	29
1.2	Répartition des cotisations sur le marché français en 2021. Source : France Assureurs .	33
1.3	Répartition des prestations sur le marché financier en 2021. Source : France Assureurs	34
1.4	Rentabilité des assureurs français. Source : France Assureurs . . . . .	35
1.5	Décomposition de la marge IARD . . . . .	36
1.6	Modèle markovien à 3 états . . . . .	39
1.7	Schéma simplifié du modèle markovien de l'assureur non-vie . . . . .	40
2.1	Répartition des polices par <i>model point</i> . . . . .	42
2.2	Répartition des polices par âge . . . . .	43
2.3	Répartition des provisions mathématiques et du nombre de polices par âge . . . . .	44
2.4	Architecture de SimBEL . . . . .	45
2.5	Passage du bilan MCEV au bilan S2 . . . . .	46
2.6	15 premiers MP par les trois méthodes d'allocation . . . . .	50
2.7	Caption . . . . .	51
2.8	Part des valeurs manquantes pour les variables : <code>auto1_age</code> , <code>auto1_ancpermis</code> , <code>auto2_age</code> , <code>auto2_ancpermis</code> . . . . .	54
2.9	Composition du portefeuille IARD . . . . .	55
2.10	Distribution de l'âge et de l'ancienneté des détenteurs d'un produit épargne . . . . .	56
2.11	Trajectoires des marges IARD . . . . .	57
2.12	Évolution de la marge totale moyenne IARD . . . . .	57
2.13	<i>Boxplots</i> des marges pour le produit AUTO1 . . . . .	59
2.14	Corrélation des variables pour le produit AUTO1 . . . . .	60
2.15	Distribution des résidus de Pearson pour l'estimation de la marge AUTO1 . . . . .	62
3.1	Evolution de la marge cumulée moyenne par ancienneté . . . . .	64
3.2	Répartition des observations selon le signe de la rentabilité . . . . .	66
3.3	Point mort moyen par tranche d'âge . . . . .	66
3.4	Fonction de répartition empirique du point mort . . . . .	67

3.5	Histogramme des points morts . . . . .	68
3.6	Répartition des segments morts par équipement . . . . .	68
3.7	Fonction de survie des assurés des deux filiales estimée par Kaplan-Meier . . . . .	72
3.8	Fonction de hasard cumulé des assurés des deux filiales . . . . .	72
3.9	Fonction de survie avec IC 95% des assurés des deux filiales . . . . .	73
3.10	Comparaison des courbes de survie selon le canal . . . . .	74
3.11	Comparaison des courbes de survie selon la détention d'un produit d'épargne . . . . .	75
3.12	Résultats des tests de signification du modèle de Cox . . . . .	78
3.13	Résultats du test de proportionnalité pour le modèle de Cox . . . . .	78
3.14	Résidus de Schoenfeld . . . . .	79
3.15	Régression de Cox pour le canal interne . . . . .	82
3.16	Régression de Cox pour le canal externe . . . . .	82
3.17	Décomposition du portefeuille externe avec 1 an d'ancienneté . . . . .	83
A.1	Comparaison des courbes de survie pour les assurés communs aux deux filiales selon les classes d'âge . . . . .	95

# Liste des tableaux

1.1	Répartition des cotisations IARD de l'année 2021 pour le marché français . . . . .	34
2.1	Les 10 <i>model points</i> les plus fréquents de l'assureur vie . . . . .	43
2.2	Résultats de la VIF en million d'euros . . . . .	47
2.3	Marge de rentabilité finale pour quelques <i>model points</i> . . . . .	51
2.4	Distribution des valeurs manquantes . . . . .	54
2.5	Proportion des détenteurs d'un produit d'épargne par produit IARD . . . . .	55
2.6	Répartition des assurés par produit IARD selon la détention d'un produit d'épargne .	56
2.7	$R^2$ du GLM par produit . . . . .	61
2.8	Résultats de la validation croisée . . . . .	61
2.9	$R^2$ du LASSO par produit . . . . .	62
3.1	Répartition des assurés non-vie selon la détention d'un produit d'épargne . . . . .	63
3.2	Répartition des assurés communs aux deux filiales selon l'ancienneté IARD . . . . .	64
3.3	Coefficients de la régression Bêta pour l'estimation des poids de Shapley . . . . .	65
3.4	Résultats du modèle de Cox . . . . .	77
3.5	Résultats du modèle de Cox stratifié par rapport au canal . . . . .	80
3.6	Extrait de la base personne-période . . . . .	81
3.7	Marge IARD moyenne pour les assurés d'un 1 an d'ancienneté . . . . .	83
3.8	Marge IARD moyenne pour les assurés IARD par ancienneté . . . . .	83
A.1	Résultats GLM pour la variable <code>detention_epargne</code> . . . . .	95



# Annexes



# Annexe A

## Une annexe

### A.1 Avec des tableaux

Produit	Estimate	Std. Error	t value	Pr( t )
PREV	- 15,6904	0,6050	25,9360	<2e-16
MRH	61,8000	1,9100	32,4000	1,89e-230
CORP	8,5836	0,1498	57,2814	<2e-16
AUTO2	19,4000	2,0300	9,5438	1,38e-21
AUTO1	21,7000	1,5900	13,6134	3,35e-42

TABLE A.1 – Résultats GLM pour la variable detention\_epargne

### A.2 Avec des graphiques

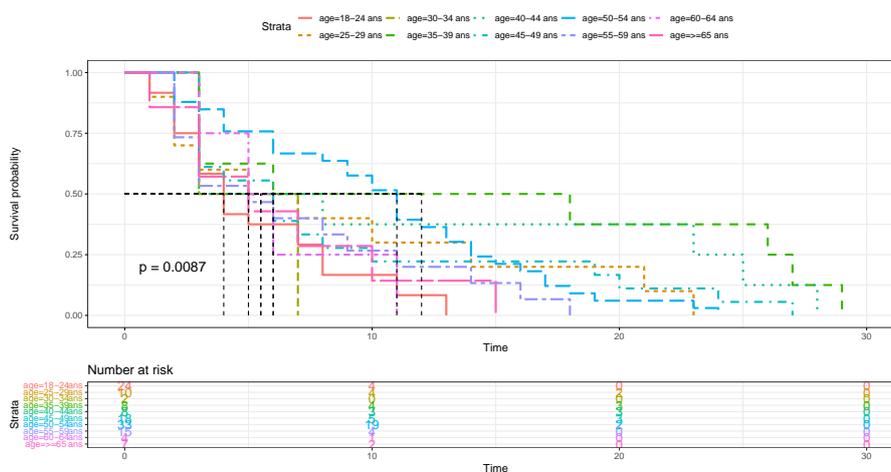


FIGURE A.1 – Comparaison des courbes de survie pour les assurés communs aux deux filiales selon les classes d'âge