

**Mémoire présenté le :**

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA  
et l'admission à l'Institut des Actuaires**

Par : Emilie SOIX

Titre Estimation du ratio de solvabilité à l'aide de méthodes d'apprentissage statistique supervisé

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

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

*Membre présents du jury de l'Institut  
des Actuaires*

signature

*Entreprise :*

Nom : ACTUARIS

Signature :

*Directeur de mémoire en entreprise :*

Nom : Janick JEMINET

Signature :

*Invité :*

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat



## Remerciements

Nombreux sont ceux et celles qui ont, de près ou de loin, participé à l'élaboration de ce mémoire. Les citer tous serait difficile et je risquerais d'en oublier. Quoi qu'il en soit je vous adresse à tous ma plus profonde gratitude.

Je tiens à remercier dans un premier temps Janick JEMINET pour ses remarques toujours pertinentes, sa disponibilité et ses idées sans lesquelles j'aurais parfois été bloquée. Je remercie également mes collègues pour leur soutien, leurs conseils et leur bonne humeur constante. Vous retrouver tous les jours est un plaisir et pour cela, merci ! Je voudrais aussi remercier Mathieu LE GOFF grâce à qui j'ai pu créer la base de données nécessaire à l'utilisation de méthodes d'apprentissage sans difficultés.

Je remercie également mes professeurs de l'INSA, Philippe BESSE, Béatrice LAURENT-BONNEAU et Hélène MILHEM pour la qualité des enseignements qu'ils m'ont prodigués et pour leur attrait pour les statistiques qu'ils m'ont transmis. Je remercie par ailleurs Cathy MAUGIS-RABUSSEAU pour la rigueur dont elle m'a appris à faire preuve et pour la clarté de ses explications qui ont toujours réussi à faire des statistiques quelque chose d'intuitif. Je remercie tout aussi chaleureusement mes professeurs de l'ISFA qui m'ont apporté de nombreuses connaissances complémentaires à celles acquises lors de ma formation d'ingénieur.

Je voulais également adresser tous mes remerciements à Anani OLYMPIO qui m'a donné envie de poursuivre mes études, a suscité mon goût pour l'actuariat et m'a toujours encouragée pour mon mémoire.

Je remercie évidemment mes amis et ma famille pour le soutien qu'ils m'ont tous apporté pendant ces sept longues années d'études. Je n'y serais jamais arrivée sans vous.

Enfin, je remercie la personne avec qui j'ai partagé les difficultés du mémoire, qui m'a apporté un soutien indéfectible et sur qui j'ai toujours pu compter. Tu as toujours su trouver les mots pour me rassurer dans les moments de doute et ce depuis bien longtemps. Ta bonne humeur et ton optimisme m'ont permis de venir à bout de mes études malgré les difficultés que j'ai pu rencontrer et je me devais de t'en remercier maintenant. Tout cela n'aurait certainement pas été possible sans toi, merci encore !

## Résumé

Le calcul de l'exigence en capital d'un organisme assurantiel nécessite la mise en place de calculs parfois lourds. Elle requiert en outre une bonne connaissance du portefeuille de l'assureur, parfois à une maille très fine.

Tous ces calculs induits dans le cadre de l'ORSA laissent encore aujourd'hui trop peu de temps à l'analyse et viennent ainsi réduire l'un des objectifs principaux de cet exercice : améliorer le pilotage de la compagnie.

Les méthodes d'apprentissage supervisé ont été largement développées ces dernières années grâce à l'amélioration de la puissance de calcul des ordinateurs et aux capacités de stockage dont ils disposent désormais.

Celles-ci peuvent donc être désormais utilisées pour fournir une première estimation du Capital de Solvabilité Requis (SCR). Les méthodes mises en place dans le présent mémoire (régressions linéaires pénalisées ou non, gradient boosting machine et *random forest*) n'ont pas pour but de remplacer les calculs proposés par Solvabilité II mais plutôt de permettre d'effectuer rapidement des sensibilités et ainsi permettre une analyse et un pilotage plus rapide.

Elles peuvent aussi être utilisées dans un but d'aide à la décision : proposer une nouvelle garantie ou en supprimer une ou encore mettre en place de nouvelles stratégies. Elles peuvent également aider à pallier le problème de manque de données à une maille fine.

Afin d'améliorer le calibrage des modèles proposés, nous avons créé une base de données de 10 000 observations, chacune représentant un organisme d'assurance. Afin de représenter au mieux le marché, différents types d'organismes ont été simulés. Nous avons calculé le SCR de ces différentes compagnies à l'aide du modèle prévoyance-santé standard d'ACTUARIS®. Nous considérons les fonds propres comme une donnée connue et nous ne les estimerons pas. L'étude sera donc focalisée uniquement sur la prévision du SCR et du ratio de solvabilité.

Nous conserverons l'approche bottom-up de calcul du SCR en estimant chacun des sous-modules, et en comparant les résultats obtenus par agrégation avec ceux obtenus par estimation directe. Nous avons également estimé des GLM avec des variables sélectionnées par avis d'expert afin d'obtenir des modèles facilement interprétables. Ceux-ci ne sont pas forcément les meilleures méthodes d'estimation des exigences en capital, mais ils permettent toutefois d'anticiper l'effet de certains indicateurs sur le SCR.

Après avoir calibré tous les modèles, nous arrivons à estimer le BSCR, le SCR et le ratio de solvabilité avec des erreurs respectives de 5,1 %, 5,1 % et 5,3 %. Nous avons également procédé à du backtesting sur des données d'entités réelles et ceux-ci se sont avérés concluants pour certains profils d'organismes.

La conclusion qui émerge du présent mémoire est que les méthodes proposées ne peuvent en aucun cas remplacer la formule standard, mais elles permettent toutefois d'estimer rapidement et parfois avec assez de précision l'exigence en capital, facilitant ainsi les sensibilités nécessaires au pilotage et à la prise de décision pour l'organisme. Nous avons également pu détecter les variables influentes pour l'estimation des différents SCR et nous pourrions ainsi proposer l'étude à des entités spécifiques, à condition de calibrer les paramètres des différents modèles pour refléter au mieux le profil de l'organisme.

Il serait intéressant de mener la même étude à partir de données réelles, d'autant plus que nous disposons aujourd'hui d'historiques acceptables. Cela permettrait en outre de contourner le problème de corrélations entre les variables qui est présent dans cet étude en raison de la méthode de simulation proposée.

**Mots-clés :** apprentissage statistique supervisé, *Gradient Boosting Machine*, *Random Forest*, Modèles Linéaires Généralisés, régression Ridge, régression LASSO, Solvabilité II.

## Abstract

Within an insurance company, the calculation of the capital requirement involves the implementation of heavy calculations processes. It also goes with a good knowledge of the insurer's portfolio, sometimes with a very fine mesh.

All these calculations induced by the ORSA still leave too little time for analysis and thus reduce one of the main objectives of this exercise: to improve the management of the company.

Supervised learning methods have been widely developed in recent years thanks to the improving computing power of computers and to the storage capacity they now have.

These methods can now be used to provide a first estimate of the Solvency Capital Requirement (SCR). The methods implemented in this paper (linear regression with or without penalization, gradient boosting machine and random forest) do not aim to replace the calculations proposed by Solvency II but rather allow to quickly make sensitivities and thus lead to faster analysis and control.

They can also be used for decision-making purposes such as offer a new guarantee or cancel one or to put in place new strategies. They can also help to overcome the problem of lack of data at a fine mesh.

To enhance the calibration of the proposed models, we created a database of 10,000 observations, each representing an insurance organization. In order to better reflect the market, different types of organizations have been simulated. We calculated the SCR for these different companies using ACTUARIS<sup>®</sup> standard health insurance model. We consider own funds as known and we will not estimate them. The study will therefore only focus on forecasting the SCR and the solvency ratio.

We will keep the bottom-up SCR calculation approach by estimating each of the sub-modules, and comparing the results obtained by aggregation with those obtained by direct estimation. We also estimated GLMs with variables selected by expert opinion in order to obtain easily interpretable models. These are not necessarily the best methods for estimating capital requirement, but they do allow anticipating the effect of some indicators on the SCR.

After calibrating all the models, we can estimate the BSCR, the SCR and the solvency ratio with respective errors of 5.1 %, 5.1 % and 5.3 %. We also performed backtesting on real entity data and the results of our models were quite satisfying for some kind of insurance companies.

The conclusion coming from this study is that the proposed methods can not in any case replace the standard formula. Nevertheless, it enables a quick estimation of the capital requirement, sometimes quite accurately and thus make easier the sensitivities needed for piloting and decision-making for the organization. We were also able to detect influential explanatory variables for the estimation of several SCRs and we could thus propose the study to specific entities, provided that we calibrate the parameters of the different models to better reflect the profile of the organization.

It would be interesting to conduct the same study from real data, especially since we now have acceptable historical data. This would also circumvent the problem of correlations between variables that is present in this study due to the proposed simulation method.

**Keywords :** supervised learning, Gradient Boosting Machine, Random Forest, Generalized Linear Models, Ridge regression, LASSO regression, Solvency II.

## Sommaire

Acronymes utilisés.....	1
Introduction.....	2
I. Solvabilité II et ratio de solvabilité .....	3
1. Ratio de solvabilité .....	3
2. Fonds propres.....	3
3. Capital de Solvabilité Requis (SCR) .....	4
II. Méthodes utilisées .....	6
1. Arbres boostés et forêts aléatoires.....	6
2. Modèles Linéaires Généralisés et régressions pénalisées .....	8
III. Présentation du modèle Prévoyance standard d'Actuaris® .....	12
IV. Estimation des différents SCR et du ratio de solvabilité .....	13
1. Création de la base de données .....	13
2. Modèles d'estimation des SCR .....	17
3. SCR de défaut .....	20
4. SCR Vie (Life).....	28
5. SCR Santé (Health).....	46
1. SCR Marché .....	72
2. BSCR.....	86
3. SCR.....	90
4. Ratio de solvabilité .....	92
V. Backtesting .....	96
Conclusion .....	97
Bibliographie.....	98
Liste des figures.....	99
Liste des tableaux.....	100
ANNEXES.....	102

## Acronymes utilisés

**AIC** : Akaike Information Criterion

**ACP** : Analyse en Composantes Principales

**ACPR** : Autorité de Contrôle Prudentiel et de Résolution

**AMCR** : Absolute Minimum Capital Requirement (Capital Minimum Requis Absolu)

**AT** : Arrêt de travail

**BEL** : Best Estimate of Liabilities (Meilleure estimation des engagements)

**BIC** : Bayesian Information Criterion

**CAT** : risque catastrophe

**FDB** : Future Discretionary Benefits (Bénéfices Discrétionnaires Futurs)

**GBM** : Gradient Boosting Machine

**GLM** : Generalized Linear Model (Modèle Linéaire Généralisé)

**HNSLT** : Health Non Similar to Life Techniques (Santé non similaire à la vie)

**HSLT** : Health Similar to Life Techniques (Santé similaire à la vie)

**LoB** : Line of Business (Ligne d'activité)

**MAE** : Mean Absolute Error (Erreur Absolue Moyenne)

**MCR** : Minimum Capital Requirement (Capital Requis Minimum)

**MGDC** : Maintien de la garantie décès

**RF** : Random Forest

**RM** : Risk Margin (Marge pour risque)

**SCR** : Solvency Capital Requirement (Capital de Solvabilité requis)

**S/P** : Sinistres/Primes

**UC** : Unités de Compte

**VM** : Valeur de marché

**PANE** : Primes Acquisées Non Emises

**PPNA** : Provisions pour Primes Non Acquisées

**RC** : Rente de conjoint

**RE** : Rente éducation

## Introduction

La mise en place de Solvabilité II nécessite le calcul de nombreuses valeurs relatives à la solvabilité des organismes d'assurance et de réassurance. Même si la formule standard permet de simplifier les calculs, elle requiert de nombreux paramètres qu'il peut parfois être difficile d'obtenir.

Les méthodes d'apprentissage se sont popularisées ces dernières années avec l'augmentation des capacités de calcul des ordinateurs et elles constituent un outil de plus à la disposition de l'actuaire, en complément des approches de régression traditionnelles.

Ces méthodes peuvent donc être utilisées dans le secteur de l'assurance afin d'estimer au mieux la solvabilité des entreprises, et ce de manière plus rapide qu'avec la formule standard. Elles permettent donc la mise en place de proxys qui peuvent être utilisés dans le cadre de l'ORSA ou comme outil d'aide à la décision. Ces approximations peuvent également être mises en œuvre en cas de manque de données à une maille fine grâce à l'utilisation de macro-indicateurs tels que les âges moyens des portefeuilles. Il ne s'agit pas de remplacer les calculs exigés par la réglementation mais d'approcher rapidement la solvabilité d'un organisme afin de laisser plus de temps à l'analyse des résultats et au pilotage de l'entreprise.

L'objectif du présent mémoire est de prédire le ratio de solvabilité d'un organisme assurantiel à l'aide de modèles linéaires généralisés et d'arbres de régression. L'approche *bottom-up* de calcul du SCR sera conservée, les modèles de prévision étant construits au niveau de chaque module de la pieuvre.

Afin de constituer une base de données nécessaire à la mise en place des méthodes d'apprentissage statistique, nous simulerons les hypothèses du modèle prévoyance-santé standard d'Actuaris®. Cela nous permettra, pour chacune des exécutions du modèle, d'obtenir un détail des différents modules de la pieuvre ainsi que des indicateurs sur le passif et l'actif de l'assureur. Pour chacun des SCR, nous disposerons donc de plusieurs variables explicatives (S/P, taux de réassurance, de PANE, chiffre d'affaires, montants de provisions, etc.) permettant l'estimation.

Pour chacun de ces modules, nous estimerons plusieurs modèles de prévision avant de déterminer quel est le plus précis et robuste pour l'estimation. Une analyse de la qualité des résultats obtenus et de la pertinence des indicateurs sélectionnés selon les différentes méthodes d'estimation sera effectuée.

Enfin, une analyse critique de l'étude menée ainsi que des idées d'amélioration seront faites.

# I. Solvabilité II et ratio de solvabilité

Nous présentons brièvement dans cette partie le calcul du SCR et du ratio de couverture.

## 1. Ratio de solvabilité

Le ratio de couverture est évalué de la manière suivante :

$$R = \frac{\text{Fonds propres éligibles}}{\text{SCR}}$$

Le ratio de solvabilité constitue un indicateur essentiel pour un organisme assurantiel car il représente la capacité de ce-dernier à payer ses engagements.

Le calcul du ratio de solvabilité nécessite donc d'évaluer les fonds propres de bases de l'entreprise ainsi que le capital de solvabilité requis, comme détaillé ci-après.

## 2. Fonds propres

Le schéma suivant représente le bilan économique simplifié d'une compagnie d'assurance :

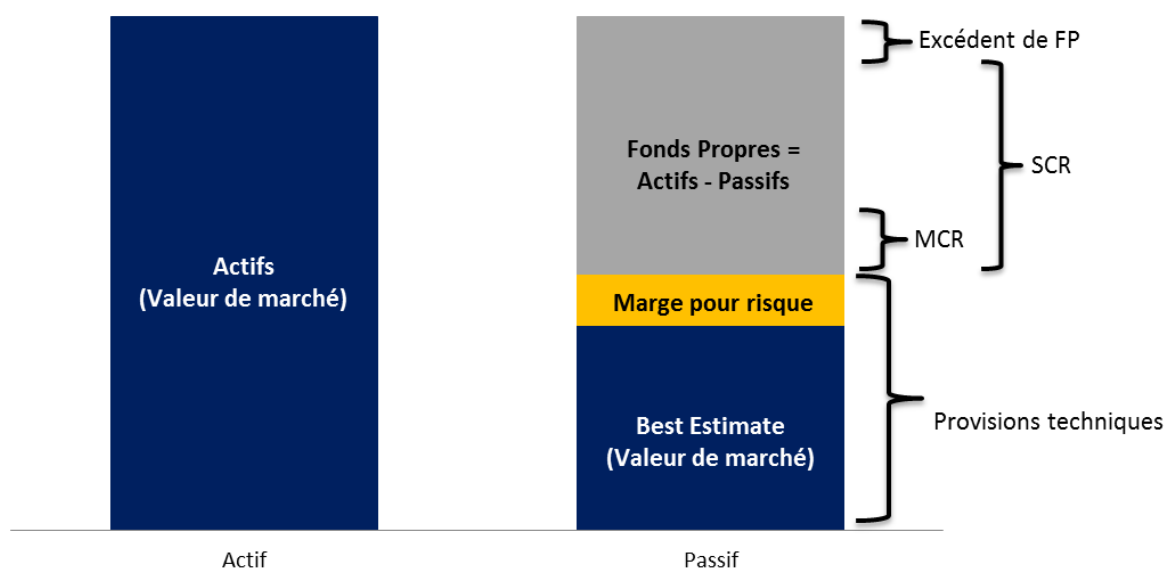


Figure 1 : Bilan économique simplifié

Dans ce contexte simplifié, le passif d'une entreprise d'assurance est constitué du *Best Estimate* des passifs (BEL) et de la Marge pour Risque.

Le BEL représente l'évaluation prospective des engagements de l'assureur (bruts de réassurance), tandis que la marge pour risque représente le coût que requerrait un assureur pour prendre à sa charge le portefeuille concerné. Elle est calculée dans une logique de coût d'immobilisation du capital nécessaire à la couverture des engagements.

Le BEL ainsi que les actifs nécessaires à la couverture des engagements sont valorisés en valeur de marché, c'est-à-dire le montant auquel ils pourraient être échangés entre deux parties informées et consentantes dans des conditions de concurrence normales.

Les fonds propres sont calculés comme la différence entre les actifs et les passifs de l'entreprise. Ils permettent de couvrir notamment :

- Le SCR,
- Le MCR

Les fonds propres en couverture du SCR constituent les fonds propres éligibles nécessaires au calcul du ratio de solvabilité.

### 3. Capital de Solvabilité Requis (SCR)

Le SCR est le niveau de capital permettant à l'organisme d'assurance d'absorber les sinistres imprévus significatifs et de continuer à assurer le paiement des engagements d'assurance à horizon d'un an. Le niveau du SCR doit être suffisant pour limiter la probabilité de ruine de l'entreprise à 1/200, sur une durée d'un an.

Le SCR est calculé par approche *bottom-up*, partant de modules spécifiques pour arriver ensuite, par agrégation, à un montant global pour l'entreprise, tout en tenant compte des spécificités de ses risques :

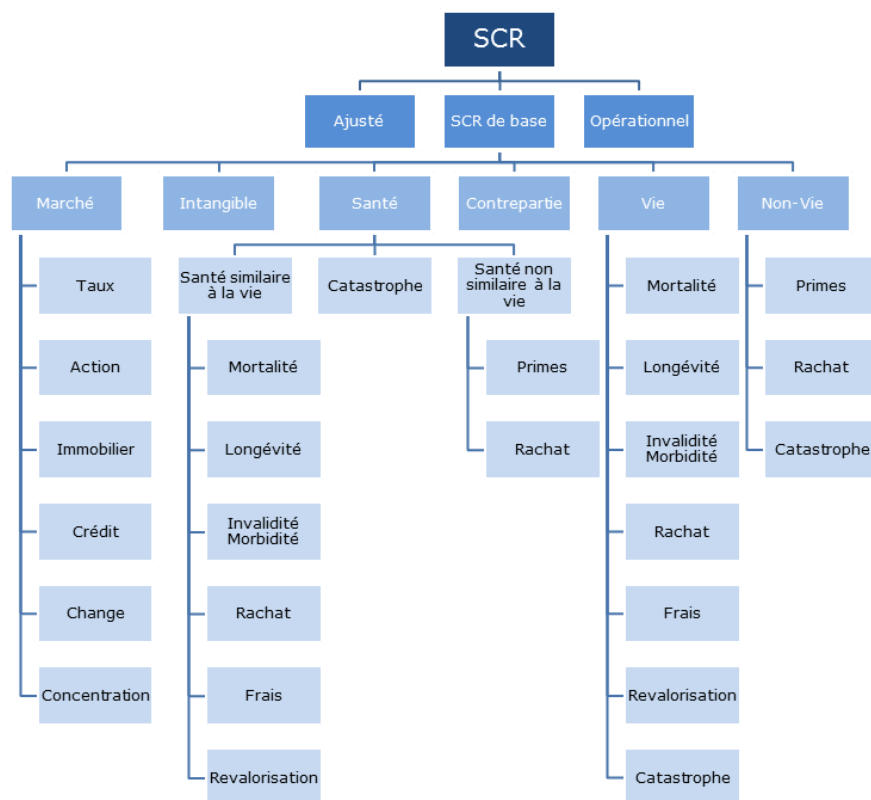


Figure 2: calcul du SCR (<https://www.financieredelacite.com/Solvency>)

Les éléments constitutifs du SCR et les chocs à appliquer pour leur calcul sont détaillés dans le règlement délégué. Pour tous les sous-modules, le SCR est calculé comme étant la différence entre les fonds propres de base en scénario central et en scénario choqué, chaque choc étant défini par la

règlementation. Les différentes corrélations utilisés pour le calcul des SCR sont elles aussi définies dans les spécifications techniques.

## II. Méthodes utilisées

Cette partie présente les méthodes mises en place dans le présent mémoire afin d'estimer les différents SCR de la pieuvre, d'une manière différente de celle proposée par la formule standard. Cette nouvelle approche peut s'avérer utile dans le cadre de l'ORSA, ou comme outil de pilotage et d'aide à la décision. Nous utiliserons les arbres de régressions et les méthodes d'agrégation de ceux-ci (*Random Forest* et *Gradient Boosting Machine*) ainsi que les modèles linéaires généralisés car ces méthodes permettent de gérer les données mixtes. Si les modèles linéaires offrent l'avantage d'être facilement interprétables, les méthodes fondées sur les arbres de régression nous permettront de déterminer quelles variables sont importantes.

### 1. Arbres boostés et forêts aléatoires

Avant de présenter les méthodes d'agrégation, nous présentons les arbres binaires de décision, qui constituent la brique de base de *Random Forest* et du *Gradient Boosting Machine*.

#### a. Arbre binaire de décision

Les arbres binaires de décision sont des méthodes itératives de discrimination. Ces méthodes ont l'avantage de pouvoir se présenter sous une forme graphique simple à interpréter et sont donc utiles à la prise de décision. Les arbres binaires de décision sont basés sur un découpage de l'espace engendré par les variables explicatives en plusieurs hyperplans.

La construction d'un arbre binaire de décision dépend du type de variable à expliquer : on parlera d'arbre de discrimination si la variable à modéliser est qualitative et d'arbre de régression si la variable à expliquer est quantitative.

Toutefois, quel que soit le type de variable à expliquer, le principe de construction d'un arbre binaire de décision est le même et repose sur différents aspects : un critère de division, une règle d'arrêt, un critère d'affectation ainsi qu'un critère d'homogénéité.

#### Critère de division :

Un arbre binaire est une succession de divisions, chaque nœud père donnant naissance à deux nœuds fils. Une division est admissible si aucun des deux nœuds fils n'est vide. Pour une variable explicative quantitative, la division peut découler de la mise en place d'un certain seuil. Pour des variables qualitatives, les divisions se font en fonction des modalités prises par la variable de réponse. Le critère de division repose sur la notion d'hétérogénéité dans les nœuds fils : les observations contenues dans un même nœud doivent être le plus homogènes possibles afin que l'arbre discrimine au mieux les observations, alors que les nœuds terminaux doivent présenter la plus forte hétérogénéité entre eux. La notion d'hétérogénéité est présentée en annexe.

#### Règle d'arrêt :

Afin d'éviter d'avoir un arbre trop « profond » et ainsi difficilement interprétable, nous devons nous donner une règle d'arrêt, c'est-à-dire un critère permettant de déterminer quand un nœud est terminal. On parle alors de feuille. Un nœud est terminal lorsqu'il est homogène ou lorsque le nombre d'observations qu'il contient est trop faible.

### Critère d'affectation :

Dans le cas d'une variable à expliquer quantitative, chaque feuille est représentée par la moyenne des observations qu'elle contient.

Un exemple d'arbre de régression est proposé à titre illustratif ci-dessous :

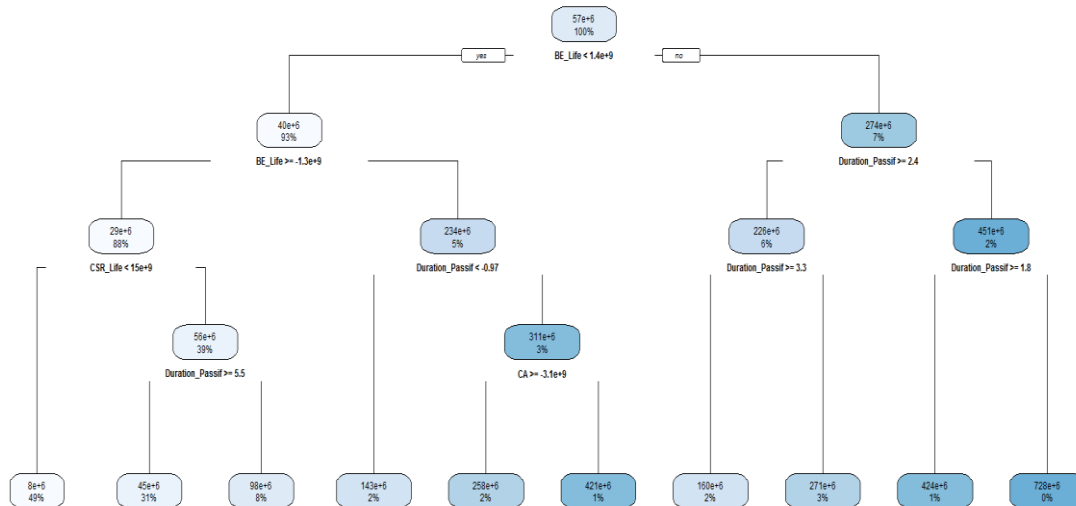


Figure 3: Arbre de régression pour l'estimation du SCR Life

Les arbres de régression ont l'avantage d'être facilement interprétables. Toutefois, afin de limiter le sur-apprentissage il est nécessaire de procéder à l'élagage. Cette partie est présentée en annexe.

### **b. Forêts aléatoires**

Cette méthode repose sur le principe du *bagging*, consistant à estimer un modèle sur plusieurs échantillons *bootstrap*. Le *bagging* permet, en moyennant les résultats de plusieurs modèles, de réduire la variance et par conséquent l'erreur de prévision.

Afin d'améliorer les résultats obtenus grâce au *bagging*, il est également possible de tirer aléatoirement le nombre de variables explicatives utilisées dans chacun des modèles. Cela permet de rendre les différents arbres de la forêt plus indépendants.

Dans l'algorithme que nous utiliserons, la taille de chaque arbre est dictée par le nombre minimal d'observations dans chaque feuille. Nous obtiendrons donc des arbres plutôt complets, de faible biais mais avec une variance importante. Nous utiliserons la validation croisée pour déterminer le nombre de variables qui interviennent dans chacun des modèles. Le nombre de prédicteurs choisi sera celui qui permet de réduire au maximum l'erreur quadratique moyenne calculée sur les

échantillons de validation croisée sans toutefois être trop élevé (observation d'un coude lors du tracé de l'erreur en fonction du nombre de variables).

Comme pour tout modèle construit par agrégation, l'interprétabilité des résultats est difficile. Il est toutefois possible d'obtenir en sortie de l'algorithme un critère d'importance de variables. Nous utiliserons celui fondé sur la décroissance d'entropie ou encore la décroissance de l'hétérogénéité définie à partir du critère de Gini (*Mean Decrease Gini*). L'importance d'une variable est alors une somme pondérée des décroissances d'hétérogénéité induites lorsqu'elle est utilisée pour définir la division associée à un nœud.

### c. Arbres boostés

#### Principe du *boosting* :

Le *boosting* repose sur l'agrégation de modèles à l'aide d'une moyenne pondérée dont les poids sont plus importants pour les observations mal prédites. Ce type d'algorithme se focalise donc sur les observations les plus difficiles à prédire. L'agrégation de modèles permet également de réduire l'erreur de prédiction, bien qu'elle rende l'interprétation difficile.

#### Gradient d'arbres boostés :

Cette méthode construit des modèles de régression en ajustant par moindres carrés, à chaque itération, une fonction appelée *base learner* à des pseudo-résidus. Ces derniers forment le gradient d'une fonction de perte qui doit être minimisée, ce gradient étant approché par un arbre de régression. À chaque étape, le modèle agrégé apparaît comme un pas vers la solution optimale, ce pas étant fait dans la direction du gradient de la fonction de perte. Des détails théoriques sur l'estimation sont présentés en annexe.

Le nombre d'arbres construits ainsi que le paramètre de rétrécissement et le niveau d'interactions seront choisis par validation croisée. Tout comme les forêts aléatoires, ces méthodes sont plus difficiles à interpréter qu'un arbre de régression. En revanche, le critère d'importance des variables peut s'avérer intéressant.

## 2. Modèles Linéaires Généralisés et régressions pénalisées

Nous présentons ici la seconde grande classe de modèles utilisée : les modèles linéaires généralisés. En plus de pouvoir gérer les données mixtes, ces méthodes ont l'avantage d'être intuitives et facilement interprétables.

### a. Le modèle linéaire général

#### Définitions :

Soit une variable  $Y$  à prédire et  $X_1, X_2, \dots, X_p$   $p$  variables explicatives. Le modèle linéaire général suppose :

$$Y_i = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j + \epsilon_i \quad \forall i = 1, \dots, n$$

Ou sous forme matricielle :

$$Y = X\beta + \epsilon$$

Où  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$  est un vecteur de paramètres à estimer. Nous supposons également que les erreurs (les  $\epsilon$ ) sont gaussiennes, centrées, homoscédastiques et non corrélées, c'est-à-dire :

$$\begin{cases} E[\epsilon_i] = 0 \quad \forall i = 1, \dots, n \\ \text{Var}[\epsilon_i] = \sigma^2 \quad \forall i = 1, \dots, n \\ \text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j \end{cases}$$

## b. Le modèle linéaire généralisé

Nous observons un vecteur  $Y$  de taille  $n$ , réalisation d'une variable aléatoire de moyenne  $\mu$  et dont les composantes sont indépendantes.

Le modèle linéaire généralisé possède trois caractéristiques principales :

1. Une composante aléatoire : le vecteur  $Y$  de moyenne  $\mu$
2. Une composante déterministe : les variables explicatives  $X^{(1)}, \dots, X^{(p)}$  qui définissent un prédicteur linéaire  $\eta = X\beta$
3. Une fonction de lien entre  $\mu$  et  $\eta$  :  $\eta = g(\mu)$

L'utilisation d'une fonction de lien permet d'élargir l'application des modèles linéaires vus précédemment : en effet, imposer un lien linéaire entre  $\mu$  et  $\eta$  peut paraître trop restrictif.

Caractérisation d'un modèle :

Nous supposerons par la suite que la variable à prédire  $Y$  suit une loi appartenant à la famille exponentielle, c'est-à-dire que sa densité peut s'écrire sous la forme :

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

Où  $a, b, c$  sont des fonctions réelles connues. Le paramètre  $\theta$  est le paramètre d'intérêt tandis que  $\phi$  est un paramètre associé à la dispersion, supposé connu dans un premier temps.

Pour les variables aléatoires dont la distribution appartient à la famille exponentielle, nous avons :

$$E[Y] = b'(\theta) \quad \text{et} \quad \text{Var}[Y] = b''(\theta)a(\phi)$$

En remarquant que  $\mu = E[Y] = b'(\theta) \Leftrightarrow \theta = (b')^{-1}(\mu)$ , nous voyons que le paramètre  $\theta$  est directement lié à la moyenne  $\mu$ . Ainsi, la variance des observations peut être réécrite sous la forme

$$\text{Var}[Y] = b''((b')^{-1}(\mu))a(\phi) = V(\mu)a(\phi)$$

Où  $V(\mu)$  est appelée « fonction variance ».

Chacune des distributions de la famille exponentielle possède une fonction de lien dit « canonique » qui relie directement le paramètre naturel  $\theta$  à  $\mu$  :

$$g_*(\mu) = \theta \Leftrightarrow \mu = g_*^{-1}(\theta)$$

Or, comme  $\mu = b'(\theta)$ , il vient  $g_*^{-1} = b'$ .

Cependant, d'autres fonctions de lien que la fonction canonique peuvent être utilisées. Ce choix est fait lors de l'étape de modélisation et dépend du problème étudié.

Le tableau suivant recense quelques exemples de fonctions de lien canoniques :

Distribution	Fonction de lien $g(\mu)$
Normale( $\mu, \sigma^2$ )	$\mu$
Binomiale( $n, \mu$ )	$\log\left(\frac{\mu}{1-\mu}\right)$
Poisson( $\mu$ )	$\log(\mu)$
Gamma	$1/\mu$

Tableau 1: Fonctions de liens canoniques

### c. Les régressions pénalisées

Le modèle linéaire généralisé peut poser problème lorsque :

- Le nombre de variables est important, rendant ainsi plus difficile l'interprétation du modèle
- Les variables présentent une structure de corrélation entre elles, gênant ainsi l'estimation des paramètres

Afin de pallier ces problèmes, des régressions pénalisées ont été introduites.

La régression ridge :

La régression ridge permet de contourner le problème de multicollinéarité. Soit le modèle de régression linéaire :

$$Y = X\beta + \epsilon$$

Où  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$  est un vecteur de paramètres à estimer.

L'estimateur ridge de  $\beta$ , noté  $\tilde{\beta}$  est solution du problème d'optimisation suivant :

$$\tilde{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^{p+1}} \left( \sum_{i=1}^n \left( Y_i - \sum_{j=0}^p X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

Avec  $\lambda$  un paramètre positif à choisir.

Plus le critère  $\lambda$  sélectionné est grand, plus la solution obtenue est régulière, diminuant ainsi la variance du modèle mais augmentant son biais. Le choix du critère de pénalisation est donc crucial et se fait généralement par validation croisée.

### La régression LASSO :

Contrairement à la régression Ridge où toutes les variables du modèle sont conservées, la régression LASSO permet de résoudre le problème de dimensions du modèle. L'estimateur LASSO de  $\beta$ , noté  $\check{\beta}$ , est solution du problème :

$$\check{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left( \sum_{i=1}^n \left( Y_i - \sum_{j=0}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$

Où  $\lambda$  est un paramètre positif à choisir.

Nous pouvons voir à travers cette expression que l'estimateur LASSO correspond à un seuillage doux de l'estimateur des moindres carrés.

De manière similaire à la régression *ridge*, la pénalisation LASSO est choisie par validation croisée.

### La régression *Elastic Net* :

Il s'agit d'une combinaison des régressions ridge et LASSO présentées ci-dessus. Le critère à minimiser peut s'écrire sous la forme :

$$\sum_{i=1}^n \left( Y_i - \sum_{j=0}^p X_{ij} \beta_j \right)^2 + \lambda \left( \alpha \times \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right)$$

Ainsi, le cas  $\alpha = 0$  correspond à une régression ridge et le cas  $\alpha = 1$  à un modèle LASSO.

### III. Présentation du modèle Prévoyance standard d'Actuaris®

Le modèle Prévoyance et Santé est un programme implémenté sous ADDACTIS® Modeling permettant de:

- Répondre aux calculs règlementaires du Pilier 1 selon la formule standard
- Projeter sur l'horizon souhaité les différents éléments qui composent l'activité d'assurance dans le cadre du Pilier 2
- Préparer les états de reporting quantitatifs du Pilier 3

Ce modèle est standard et sert d'architecture de base pour élaborer les solutions de modélisation spécifiques aux clients du secteur de la prévoyance et de la santé. Ainsi il prend en compte les garanties d'assurance suivantes :

- La santé
- L'arrêt de travail (incapacité, invalidité et maintien de la garantie décès)
- Les rentes d'éducation
- Les rentes de conjoint
- Le décès

Par ailleurs il est entièrement déterministe afin de réduire les temps de calculs et faciliter sa maintenance.

La présentation du modèle est proposée en annexe. Nous présentons toutefois ici les différents chocs appliqués :

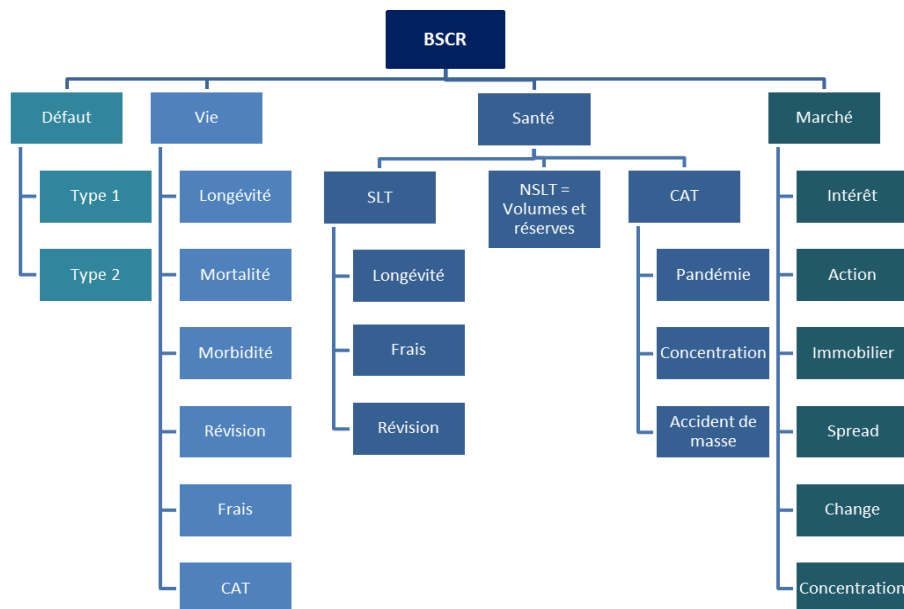


Figure 4: Détail des chocs appliqués dans le modèle prévoyance-santé

## IV. Estimation des différents SCR et du ratio de solvabilité

Cette partie constitue le cœur du mémoire et présente la construction et l'analyse des différents modèles utilisés pour la prévision du Ratio S2.

Dans un premier temps, nous présenterons la méthode de simulation des données que nous avons mise en place et les variables retenues pour l'estimation des différents SCR. Ensuite, une partie présentera la construction des différents modèles de type GLM/arbres boostés, ainsi que l'analyse de leur qualité à travers différents critères (analyse des résidus, MAE, distance de Cook, etc.). Enfin, un regard critique sur ces modèles ainsi que des pistes d'améliorations seront proposés.

### 1. Création de la base de données

Hypothèses du modèle Prévoyance-Santé :

Afin de projeter le business plan d'une compagnie d'assurance, le modèle prévoyance-santé d'ACTUARIS® repose sur différentes hypothèses :

- Comptables : bilans sociaux et compte de résultat détaillés par segment,
- Hypothèses relatives aux traités de réassurance passés et futurs,
- Fiscales,
- Hypothèses du business plan : nombre initial de cotisants, taux d'évolution des cotisations individuelles, hypothèses sur les taux renouvellements, de participation aux résultats, chroniques des taux techniques et des S/P,
- Hypothèses sur les frais,
- Hypothèses de sinistralité : chroniques de PSAP, cadences de règlement de la charge
- Hypothèses sur les portefeuilles en arrêt de travail, rente de conjoint et rente éducation, model point, tables de mortalité et de maintien en arrêt de travail,
- Hypothèses relatives à l'ALM et au générateur de scénarios économiques,
- Portefeuille de placements et d'actifs transparisés (il n'y aura pas de transparisation dans le cadre de ce mémoire),
- Hypothèses relatives à solvabilité 2 : méthode de calcul de la marge pour risque, taux d'intégration des FP futures, traités XS pour le risque catastrophe (Life), hypothèses sur les expositions pour le risque catastrophe (Health), capacité d'absorption des chocs par le FDB, classification des fonds propres par tier, détail des créances de type 1 (réassureurs et autres).

La segmentation du modèle utilisé comporte :

- 3 segments santé
- 2 segments arrêt de travail,
- 2 segments de rente éducation,
- 2 segments de rente de conjoint,
- 2 segments décès.

Simulation des hypothèses :

Les hypothèses du modèle étant trop nombreuses, seule une partie de celles-ci a été modifiée lors des simulations effectuées. Les hypothèses simulées, choisies grâce à un avis d'expert, sont les suivantes :

- Montant de cotisations,
- Chroniques des S/P,
- Taux d'évolution des affaires nouvelles,
- Taux d'évolution du montant de cotisations individuelles
- Portefeuilles en arrêt de travail, rente de conjoint et rente éducation
- Répartition des sexes pour les garanties rente éducation et rente de conjoint,
- Chroniques de frais de gestion, acquisition et administration,
- Montant des créances réassureurs,
- Taux de PANE et de PPNA,
- Portefeuille d'actifs,
- Traités de réassurance,
- Répartition des provisions mathématiques en arrêt de travail, rente de conjoint et rente éducation,
- Capitaux sous risques pour le risque catastrophe en vie,
- Expositions et montants sous risques pour les scénarios du risque catastrophe en santé,
- Coefficient d'absorption des chocs par le FDB.

Lors de la simulation des montants de cotisations, différents « profils » d'organismes assurantiels ont été simulés :

- Des organismes avec une activité exclusivement en santé,
- Des organismes avec des garanties santé et arrêt de travail,
- Des organismes proposant des garanties en santé, arrêt de travail, décès, rente de conjoint et rente éducation,
- Des organismes de type institution de prévoyance, proposant des garanties en arrêt de travail, décès, rente de conjoint et rente éducation.

Au niveau des portefeuilles de rentes, trois types de simulations sont possibles :

- Bootstrap simple,
- Bootstrap favorisant les individus « jeunes »,
- Bootstrap favorisant les individus « âgés ».

La modification des hypothèses sélectionnées a un impact sur les comptes de résultats, les bilans comptables et fiscaux. Ainsi, lors de la simulation la cohérence des hypothèses a été assurée. Le graphe ci-dessous résume la façon de simuler les paramètres du modèle :

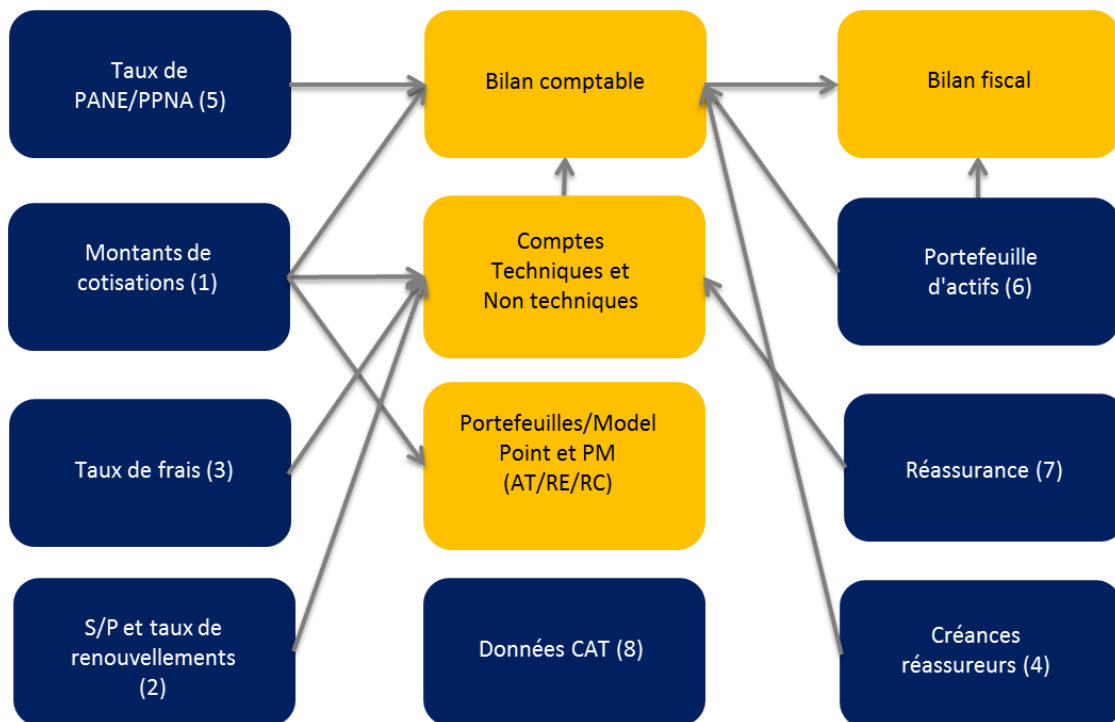


Figure 5: impact des simulations sur l'ensemble des hypothèses du modèle

En premier lieu, nous simulons les montants de cotisations des comptes techniques afin de créer divers profils d'organismes d'assurance :

- Des organismes 100% santé,
- Des organismes 80% santé et 20% arrêt de travail,
- Des organismes 20% santé et 80% arrêt de travail,
- Des organismes proposant des garanties arrêt de travail, décès, rente de conjoint et rente éducation,
- Des organismes ne proposant que des garanties vie (rente de conjoint, rente éducation, décès),
- Des organismes proposant toutes les garanties du modèle prévoyance-santé d'ACTUARIS®.

Chacun des cas susmentionnés est tiré aléatoirement. Les comptes de résultats, les bilans, les portefeuilles d'arrêt de travail, rente de conjoint et rente éducation et les éléments S2 relatifs aux risques catastrophe sont ensuite adaptés au type d'organisme simulé.

Ensuite, nous simulons les ratios S/P, les taux de renouvellements des cotisations et d'affaires nouvelles et les taux d'évolution des cotisations individuelles. Suite à la simulation des ratios S/P, nous adaptons les prestations des comptes techniques, et par conséquent les bilans. Nous procédons de même pour la simulation des taux de frais, des taux provisions sur primes (PPNA et PANE) et des traités de réassurance.

Nous simulons par la suite le portefeuille de placements de l'assureur, en faisant varier les valeurs de marché, les prix d'acquisition, les valeurs fiscales, les amortissements, les notations, les durations, les PDD, les nominaux, les surcotes-decotes et les durations. Nous effectuons également du bootstrap pour modifier la répartition des placements. Nous adaptons ensuite les bilans comptables et fiscaux pour que ceux-ci intègrent les valeurs des placements.

Cette méthode de simulation nous permet d'obtenir dans chacun des scénarii un fichier d'initialisation cohérent avec le type d'organisme créé, avec des bilans équilibrés. Toutefois la simulation des cotisations mène à des corrélations entre les proportions du chiffre d'affaires dédiées à chacune des garanties. De même, les portefeuilles d'arrêt de travail, de rente de conjoint et de rente éducation sont corrélés. Enfin, comme les capitaux sous risques pour le calcul du SCR catastrophe en vie ne sont simulés que si l'organisme créé propose des garanties de type vie, nous observons également une corrélation entre ces montants et la répartition du chiffre d'affaires. Ces corrélations posant problème pour la détermination des coefficients des modèles linéaires généralisés, nous proposerons des régressions pénalisées (*Ridge* et *Elastic Net*) afin de pouvoir intégrer la répartition du chiffre d'affaires dans les estimateurs car nous pouvons penser a priori que ces variables peuvent améliorer l'estimation de certains SCR. Dans un premier temps, la plupart des simulations de taux ont été effectuées selon des lois uniformes. Toutefois, le choix de cette loi a un impact non négligeable sur la répartition des SCR à prédire, alors que celle-ci doit appartenir à la famille exponentielle dans le cadre des modèles linéaires généralisés. Nous avons par la suite effectué la simulation des hypothèses à partir de lois normales dans le but d'obtenir une répartition des SCR adaptée à l'estimation par modèles linéaires généralisés.

#### Création d'indicateurs synthétiques :

Le but du présent mémoire étant de proposer des modèles d'estimation du ratio de solvabilité adapté à plusieurs organismes assurantiels, la création d'indicateurs synthétiques permet de généraliser l'application des modèles obtenus sans toutefois disposer d'informations trop précises. Les indicateurs créés sont les suivants :

- Taux d'évolution du chiffre d'affaires :  $Coef_{evolution\ CA} = (Tx_{renouvellement} + Tx_{AFN}) \times (1 + Tx_{evolution\ cotisations})$ ,
- Notation moyenne et durée moyenne des obligations,
- Répartition des actifs,
- Âge et ancienneté moyenne, dispersion des âges et des anciennetés du portefeuille d'arrêt de travail, ratio entre les provisions dédiées au maintien de la garantie décès et celles afférentes à l'arrêt de travail, durées moyennes des portefeuilles d'incapacité et d'invalidité,
- Âge moyen, dispersion des âges et proportion d'hommes pour les portefeuilles de rentes éducation et de rente de conjoint,
- Durée globale du passif et de l'actif,
- S/P moyens par type de garantie (santé, décès, arrêt de travail, rente éducation et rente de conjoint),
- Montants sous risques par type de garantie pour le risque catastrophe,
- Taux de frais totaux (en moyenne par type de garantie),
- Montant des créances réassureurs,
- Taux de réassurance par type de garantie,
- Montant des BEL par classification S2 (Life, SLT et NSLT),
- Coefficients d'absorption des chocs par le FDB,
- Répartition des cotisations entre les garanties santé, arrêt de travail, décès, rente de conjoint et rente éducation ainsi que le montant total de cotisations,
- Taux de PANE et de PPNA moyens par type de garantie.

La base finale utilisée est composée de 114 variables.

## **2. Modèles d'estimation des SCR**

### Méthodologie :

La méthodologie de modélisation de chaque SCR est la suivante :

- 1) Analyse en Composantes Principales (ACP) afin de déterminer les indicateurs liés à la variable à expliquer. L'analyse des variables se limite aux deux axes où la contribution de la variable de sortie est la plus importante,
- 2) Construction de modèles linéaires généralisés :
  - a. Modèle complet,
  - b. Sélection de variables selon le critère AIC,
  - c. Sélection de variables selon le critère BIC,
  - d. Modèle construit à l'aide des variables sélectionnées par l'ACP,
  - e. Modèle construit à l'aide des variables sélectionnées par l'ACP et sélection selon le critère AIC,
  - f. Modèle construit à l'aide des variables sélectionnées par l'ACP et sélection selon le critère BIC,
- 3) Estimation via un *Gradient Boosting Machine* (GBM),
- 4) Estimation à l'aide de forêts aléatoires,
- 5) Estimation à l'aide de régressions linéaires pénalisées :
  - a. Régression *RIDGE*,
  - b. Régression LASSO,
  - c. Régression *Elastic Net*,
- 6) Estimation à l'aide d'un modèle linéaire généralisé avec les variables retenues par le *GBM*,
- 7) Estimation à l'aide d'un modèle linéaire généralisé avec les variables retenues par *Random Forest*,
- 8) Estimation d'un modèle « combiné » (GLM ou régression pénalisée) avec les variables les plus sélectionnées par les méthodes précédentes.

### Sélection des indicateurs :

Les critères de sélection de variables sont les suivants :

- La variable est considérée comme significative au seuil 5% (test de Student pour les régressions linéaires),
- La variable a une forte importance (*GBM*, *Random Forest*),
- Le coefficient associé à la variable est non négligeable (*Ridge*, LASSO, *Elastic Net*).

### Critère de sélection de modèles :

Afin de sélectionner le meilleur modèle d'estimation des différents SCR, nous utiliserons l'erreur absolue moyenne (ou MAE : *Mean Absolute Error*). Celle-ci est définie par :

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

Nous l'exprimerons en pourcentage d'écart par rapport à la valeur à estimer :

$$MAE_{\%} = 100 \times \frac{1}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|$$

La base d'apprentissage sera constituée de 80% des observations. Les paramètres optimaux des modèles *GBM*, *Random Forest* et régressions pénalisées seront déterminés par validation croisée 10 plis.

### Graphes de validation :

Pour les modèles linéaires généralisés, les graphes des valeurs ajustées en fonction des observations, des résidus ou bien la distance de Cook seront également utilisés pour juger de la qualité du modèle et détecter les observations influentes qui devront être retirées du jeu de données car biaisant l'estimation des paramètres.

### Transformation des variables :

Afin de pouvoir estimer un modèle linéaire, nous devons nous intéresser à la distribution de la variable à prédire. Si cela est nécessaire, nous procéderons à des transformations sur celle-ci (translation, logarithme ou racine). Bien que ces transformations ne nous permettent pas par la suite d'estimer directement la valeur du SCR et qu'elles risquent de diminuer les taux erreurs que nous observerons, elles sont nécessaires à la modélisation par GLM. Dans ce cas, les modèles linéaires ne pourront pas être utilisés de façon opérationnelle pour estimer un SCR, mais pourront toutefois apporter une intuition et nous aider à juger de l'importance et de l'impact d'une variable sur celui-ci. De plus, comme nous utilisons des variables de grandeurs très différentes (pourcentages, millions d'euros, âges, etc.) nous réduirons systématiquement les variables afin de ramener celles-ci à une échelle commune. Cela nous permettra de voir dans quelle mesure l'impact d'une des variables sur le SCR étudié est important et facilitera également l'estimation des paramètres de la régression. Les méthodes d'estimation par arbres de régressions seront quant à elle effectuées sur les variables brutes (non transformées).

### Comparaison de l'estimation et de l'agrégation des meilleures méthodes d'estimation des sous-modules :

Nous effectuerons également des sensibilités sur les SCR Vie, Défaut, Marché, Santé et le BSCR afin de déterminer si l'estimation directe du module est plus performante que l'agrégation des meilleurs modèles d'estimation des sous-modules.

### Sensibilités :

Dans certains cas et selon les qualités des différents modèles, nous mènerons des études de sensibilités afin d'améliorer les qualités de prévision des GLM, en testant l'effet de l'ajout de certaines variables sur le SCR à estimer.

### Modèles construits à partir de variables retenues par avis d'expert :

Dans certains cas, nous estimerons également des modèles linéaires grâce à des variables sélectionnées par avis d'expert. Ces modèles n'ont pas nécessairement pour but d'être les meilleurs modèles d'estimation, mais d'être facilement interprétables et d'apporter une intuition quant à l'évolution de la variable à prédire.

### Gestion des corrélations :

Les montants de chiffre d'affaires en santé, arrêt de travail, rente de conjoint, rente éducation et décès étant le point de départ des simulations, ceux-ci sont fortement corrélés. En conséquence, ils sont exclus de la base servant à construire les modèles d'estimation, excepté dans le cas des régressions *ridge* et *Elastic Net* puisque ces dernières permettent de surpasser le problème de multicolinéarité. En revanche, le chiffre d'affaires global sera conservé car le lien entre celui-ci et le SCR d'un organisme assurantiel est direct.

### **3. SCR de défaut**

#### **a. SCR de défaut de type 1**

##### Analyse en composantes principales :

Les variables retenues grâce à l'ACP sont les suivantes :

- Montant du chiffre d'affaires,
- Dispersion des âges pour les garanties rente de conjoint et rente éducation,
- BE SLT et NSLT,
- Les taux de réassurance,
- Montants sous risques pour le risque catastrophe en vie,
- Âge moyen du portefeuille de rente de conjoint,
- Âge moyen du portefeuille d'arrêt de travail,
- Ratio entre les provisions dédiées au maintien de la garantie décès et celles afférentes à l'arrêt de travail,
- Dispersion des anciennetés en invalidité et en incapacité,
- Anciennetés moyennes en invalidité et en incapacité,
- Durations moyennes des portefeuilles d'incapacité et d'invalidité.

De par la méthode de simulation des données (en fonction du type d'organisme d'assurance), les indicateurs élaborés sur les portefeuilles d'arrêt de travail, rente de conjoint et rente éducation sont corrélés aux taux de réassurance pour ces mêmes garanties. De même, les montants sous risque pour le risque catastrophe en vie sont fortement corrélés aux chiffres d'affaires des différentes garanties. Ils sont donc également un indicateur du type d'activité de l'organisme assurantiel et de ses risques dans le cadre de notre étude.

Bien que la part d'information expliquée dans le plan factoriel où la variable d'intérêt présente les plus fortes contributions soit faible, l'ACP nous permet de présélectionner quelques variables pour l'estimation du SCR de défaut de type 1.

## Choix de la distribution pour l'estimation par GLM :

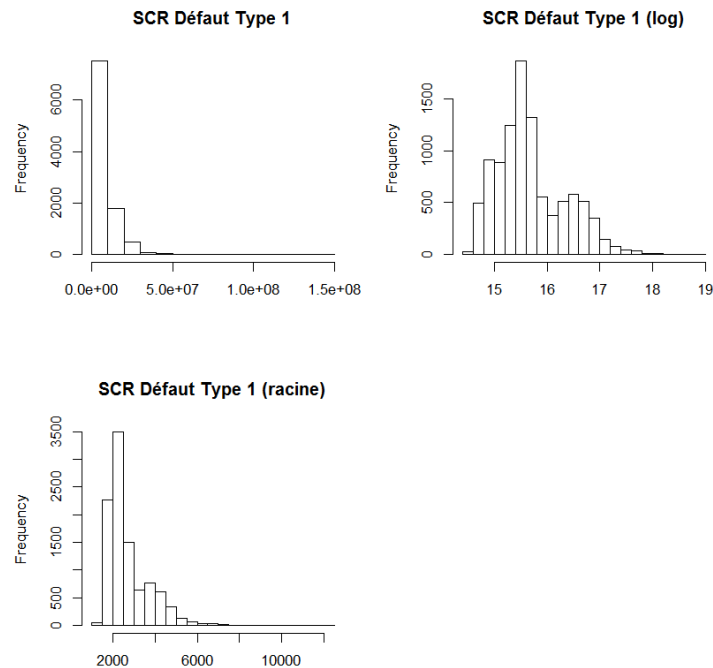


Figure 6: histogrammes du SCR de défaut de type 1

Nous avons premièrement essayé de construire un GLM en supposant une distribution Gamma car le QQ-plot renvoyé pour une telle distribution de la variable brute était assez satisfaisant. Toutefois l'algorithme n'a pas convergé, malgré la réduction des variables et le passage de coefficients initiaux en paramètres de l'algorithme. Nous utiliserons donc une transformation logarithmique. Le QQ-plot obtenu nous permet d'accepter l'hypothèse de normalité, bien qu'il existe des distorsions au niveau des queues de distribution. Les modèles linéaires construits intégreront donc cette hypothèse.

### Commentaires sur les variables sélectionnées et les modèles d'estimation :

Le tableau 1 (cf. Annexes) synthétise les variables sélectionnées par les différents modèles d'estimation.

Nous pouvons voir que les montants de provisions *Best Estimate*, le chiffre d'affaires, la répartition des provisions entre l'arrêt de travail et le maintien de la garantie décès, les coefficients d'évolution du chiffre d'affaires (en santé et décès) et les taux de réassurance sont souvent sélectionnés. Ces variables seront donc utilisées pour construire le dernier modèle d'estimation du SCR. De plus, comme les répartitions du chiffre d'affaires ont des coefficients non négligeables dans le cadre des régressions *Ridge* et *Elastic Net*, elles seront également incluses dans le dernier modèle.

Les montants de BE et les taux de réassurance nous indiquent quelle est la part des provisions cédées soumises au risque de défaut. Le chiffre d'affaire et son évolution selon les différentes garanties présentent un lien avec le niveau d'engagement de l'assureur, et donc l'engagement cédé s'il existe des traités de réassurance. La sélection de ces variables semble donc pertinente.

### Choix du modèle d'estimation :

Le tableau suivant résume les MAE sur les bases de test et d'apprentissage :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	21,5%	23,1%
Modèle complet + AIC	21,3%	22,4%
Modèle complet + BIC	21,5%	20,0%
Modèle ACP	22,1%	20,7%
Modèle ACP + AIC	22,1%	20,6%
Modèle ACP + BIC	22,1%	20,6%
Random Forest	11,3%	13,6%
<b>GBM</b>	<b>8,8%</b>	<b>10,2%</b>
Régression Ridge	19,7%	19,6%
Régression LASSO	21,0%	22,4%
Elastic Net	19,4%	19,5%
GLM avec variables issues de RF	21,4%	21,8%
GLM avec variables issues de GBM	26,5%	25,6%
Regression Ridge finale	20,9%	20,2%

*Tableau 2: Détail des résultats pour l'estimation du SCR de défaut de type 1*

Nous retenons le GBM comme le meilleur modèle, car il présente une bonne capacité de prévision.

Bien qu'ils soient moins complexes, les trois derniers modèles présentent une qualité prédiction satisfaisante.

### **b. SCR de défaut de type 2**

#### Analyse en composantes principales :

L'ACP nous indique que les variables liées aux SCR de défaut de type 2 sont les suivantes :

- La dispersion des âges dans les portefeuilles de rente de conjoint et rente éducation,
- Les durations moyennes en incapacité et invalidité,
- Les dispersions des anciennetés en incapacité et invalidité,
- Les créances des réassureurs,
- Les capitaux sous risques pour le risque catastrophe en vie,
- Le montant de *Best Estimate* en SLT,
- La répartition des provisions entre arrêt de travail et maintien de la garantie décès.

## Choix de la distribution pour l'estimation par GLM :

La distribution de la variable à expliquer est la suivante :

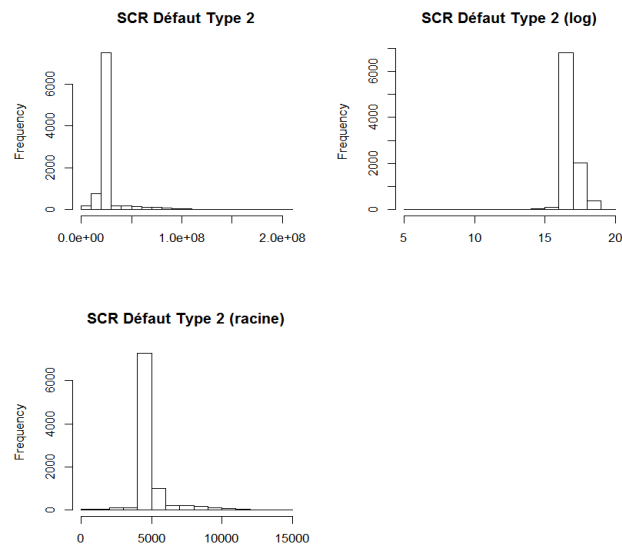


Figure 7: histogrammes du SCR de défaut de type 2

Il est difficile de déterminer quelle est la distribution du SCR de défaut de type 2, et aucun des QQ-plots construits à partir de lois de la famille exponentielle n'est satisfaisant. La transformation par la racine est la plus symétrique, c'est donc celle-ci que nous retiendrons par la suite. Nous acceptons l'hypothèse de normalité pour l'estimation de modèles linéaires, bien que cette dernière risque de poser problème quant à l'adéquation.

## Analyse des résultats et des modèles construits :

Lors de l'analyse des graphes des distances de Cook, nous remarquons que certaines observations sont influentes. Elles sont donc retirées de la base de données. Nous observons une structure dans le graphe des résidus en fonction des valeurs estimées et lors du tracé de la droite de régression pour le premier modèle linéaire, la forme du nuage de points nous incite à construire un modèle polynômial. Cependant les procédures de sélection de variables sont assez longues, donc nous limitons les modèles polynômiaux aux GLM construits à partir des variables sélectionnées selon *Random Forest* et GBM.

Le tableau 2 présenté en annexe résume les différentes sélections de variables. Nous pouvons voir que le chiffre d'affaires, les créances des réassureurs, les taux de PANE, les montants sous risques pour le risque catastrophe et les montants de *Best Estimate* en Vie et santé similaire à la vie sont souvent sélectionnés. Ces variables sont donc utilisées pour la construction du dernier GLM.

Les créances des réassureurs sont prises en compte pour le calcul du SCR de défaut de type 2, leur sélection est donc pertinente. Contrairement à l'ACP, les taux de PANE sont sélectionnés par les méthodes d'estimation du SCR de défaut de type 2. Ces taux représentent des créances et il est donc naturel de les retenir. Les capitaux sous risques sont souvent retenus car corrélés au montant et à la répartition du chiffre d'affaires, qui est impacté par les taux de PANE au travers de notre méthode de

simulation : après avoir simulé des taux de PANE, nous avons modifié le poste A6aa du bilan pour que ce dernier soit cohérent avec les taux et les montants de cotisations simulés.

Comme évoqué précédemment, nous avons construit des modèles linéaires généralisés quadratiques avec les variables sélectionnées selon *Random Forest* et le GBM. L'adéquation est nettement meilleure pour ces modèles par rapport à un modèle sans interaction, et les graphes des résidus sont plus convaincants.

#### Analyse de sensibilités :

Le dernier GLM construit est également quadratique. Lorsque nous traçons les valeurs ajustées en fonction des observations, nous notons une distorsion dans la forme du nuage de points, qui n'était pas présente dans les deux précédents modèles. Après avoir analysé les variables sélectionnées dans ces deux modèles, il semble que la seule qui n'ait pas été intégrée dans le modèle « combiné » soit la durée du passif. Après ajout de cette dernière, nous obtenons un nuage de points sans structure arrondie, nous indiquant ainsi que la durée du passif améliore l'estimation du SCR de défaut de type 2.

#### Choix du modèle d'estimation :

Les résultats obtenus par les différents modèles d'estimation sont synthétisés dans le tableau ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	5,9%	6,5%
Modèle complet + AIC	5,9%	6,1%
Modèle complet + BIC	5,9%	6,1%
Modèle ACP	6,4%	6,5%
Modèle ACP + AIC	6,4%	6,5%
Modèle ACP + BIC	6,4%	6,5%
Random Forest	2,3%	3,6%
<b>GBM</b>	<b>1,5%</b>	<b>3,4%</b>
Régression Ridge	6,8%	8,6%
Régression LASSO	6,0%	6,1%
Elastic Net	5,6%	9,2%
GLM avec variables issues de RF (quadratique)	3,4%	4,1%
GLM avec variables issues de GBM (quadratique)	3,7%	4,8%
GLM Final (quadratique)	4,5%	8,8%
GLM Final avec durée du passif (quadratique)	4,5%	4,9%

Tableau 3: Détail des résultats pour l'estimation du SCR de défaut de type 2

Comme pour le SCR de défaut de type 1, les meilleurs résultats sont obtenus pour les méthodes *Random Forest* et GBM. Nous pouvons voir que les résultats obtenus par le GLM élaboré avec les variables sélectionnées par *Random Forest* sont satisfaisants, mais les taux d'erreurs sont faibles en raison de la transformation effectuée sur la variable.

Nous retenons donc le GBM comme étant le meilleur modèle d'estimation du SCR de défaut de type 2.

### c. SCR de défaut

Analyse en composantes principales :

Les variables retenues grâce à l'ACP sont les suivantes :

- Le montant de *Best Estimate*,
- Le chiffre d'affaires et sa répartition,
- Les montants sous risques pour le risque catastrophe en vie,
- Les coefficients d'absorption par le FDB pour le de défaut,
- Les indicateurs (dispersion de l'âge, âge moyen, anciennetés moyennes et durations moyennes) construits sur les portefeuilles d'arrêt de travail, de rente de conjoint et de rente éducation,
- La durée du passif.

Les montants sous risque pour le calcul du SCR catastrophe en vie sont également révélateurs de l'engagement de l'assureur et de son exposition au risque de défaut. Cette variable est corrélée à la répartition du chiffre d'affaires et aux différents portefeuilles de passif en raison de la méthode de simulations proposée, expliquant la sélection de ces indicateurs.

Choix de la distribution pour l'estimation par GLM :

Le SCR de défaut étant l'agrégation du SCR de défaut de type 1 et de celui du type 2, nous rencontrons le même problème pour l'estimation de la loi sous-jacente que précédemment :

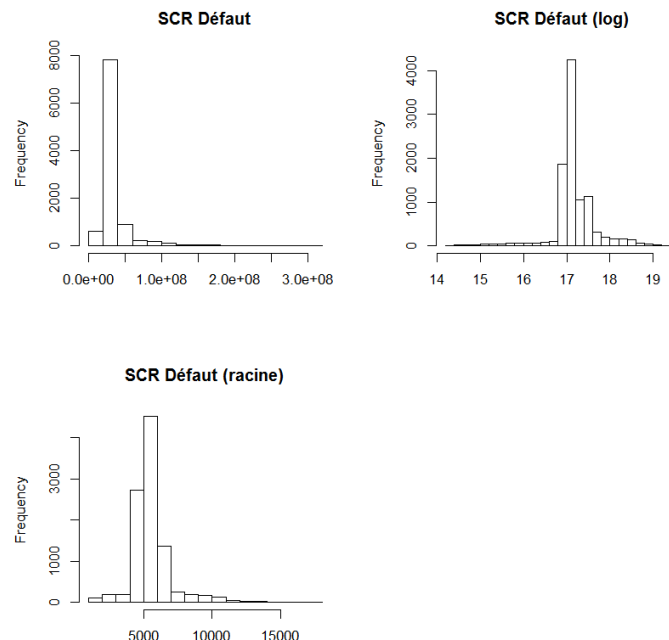


Figure 8: Histogrammes pour le SCR de défaut

Nous retenons la transformation logarithmique, car la distribution de la variable d'intérêt est plus symétrique dans ce cas. Toutefois, le QQ-plot obtenu avec les quantiles d'une loi normale n'est pas satisfaisant, réduisant grandement la fiabilité et la qualité des GLM construits par la suite. Nous

avons voulu ajuster un GLM avec une loi Gamma, mais l'estimation des coefficients de la régression n'a pas convergé.

Analyse des résultats :

Le tableau 3 présenté en annexe résume les différentes sélections de variables. Le chiffre d'affaires, les montants de *Best Estimate*, la durée moyenne en invalidité et le taux de PANE pour la garantie décès sont les variables le plus sélectionnées et sont donc utilisées dans la construction du dernier modèle. Les montants de provisions et le chiffre d'affaires sont des indicateurs de l'engagement de l'assureur et du risque de défaut auquel il est exposé en cas de cession. Les taux de PANE sont également représentatifs des montants de créances des assurés envers l'organisme d'assurance. La durée moyenne en invalidité présente des corrélations avec les différents taux de PANE, expliquant en partie pourquoi cette variable a été retenue dans plusieurs modèles.

Les graphes des résidus et des valeurs prédites en fonction des observations sont très peu satisfaisants. Les erreurs moyennes des GLM et les coefficients d'ajustement ne sont pas convaincants. En revanche, les résultats obtenus par *Random Forest* et le GBM sont bien meilleurs.

La structure du nuage de points des valeurs prédites en fonction des observations nous incite à construire un modèle polynômial (cubique). Cela est fait dans le cas des GLM construits à partir des variables sélectionnées par *Random Forest*, GBM et pour le modèle combiné, autrement la procédure de sélection de variables serait trop longue. L'ajout d'interactions améliore les graphes des résidus, mais il existe toujours une structure. De plus, les modèles ainsi construits sont assez complexes alors qu'ils n'améliorent pas – voire dégradent - l'ajustement ou la qualité de la prévision, comme le tableau des résultats en témoigne.

Choix de la méthode d'estimation :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	20,6%	22,4%
Modèle complet + AIC	20,2%	19,2%
Modèle complet + BIC	20,6%	17,8%
Modèle ACP	20,9%	18,8%
Modèle ACP + AIC	20,9%	18,4%
Modèle ACP + BIC	20,9%	18,2%
Random Forest	7,4%	9,1%
<b>GBM</b>	<b>5,6%</b>	<b>8,5%</b>
Régression Ridge	14,0%	18,8%
Régression LASSO	20,8%	28,0%
Elastic Net	17,5%	18,8%
GLM avec variables issues de RF (cubique)	18,9%	24,2%
GLM avec variables issues de GBM (cubique)	21,9%	25,0%
GLM Final (cubique)	24,5%	23,9%

Tableau 4: détail des résultats pour l'estimation du SCR de défaut

Nous retenons le GBM comme meilleur modèle pour l'estimation directe du SCR de défaut.

### Comparaison de l'estimation directe et de l'agrégation des sous-modules :

Le SCR de défaut étant l'agrégation du SCR de défaut de type 1 et du SCR de défaut de type 2, nous pouvons également agréger les résultats des estimations obtenues dans les deux parties précédentes et comparer avec les données réelles. Nous procédons ainsi :

- Estimation du SCR de défaut de type 1 et type 2 à l'aide d'un GBM,
- Comparaison avec les valeurs réelles du SCR de défaut

Le graphe des résidus en fonctions des valeurs ajustées ne présente pas de structure. Toutefois, les résultats par agrégation sont moins satisfaisants que ceux obtenus par estimation directe :

	Estimation directe	Agrégation des sous-modules
MAE (apprentissage)	5,6 %	8,1 %
MAE (test)	8,5 %	11 %

*Tableau 5: comparaison de l'estimation directe et de l'agrégation pour le SCR de défaut*

Nous retiendrons le GBM comme meilleure méthode pour estimer le SCR de défaut.

### Estimation d'un modèle linéaire avec des variables sélectionnées par avis d'expert :

Nous avons également voulu estimer un modèle linéaire à l'aide des variables suivantes :

- Les montants de provisions *Best Estimate*,
- Le chiffre d'affaires,
- Les taux de PANE,
- Les créances réassureurs,
- Les taux de réassurance.

Nous obtenons les coefficients suivants :

Variable	Coefficient
<b>BE Life</b>	0,52
<b>BE SLT</b>	0,19
<b>BE NSLT</b>	0,12
<b>Taux de PANE (Décès)</b>	0,05
<b>Taux de PANE (Santé)</b>	0,03

*Tableau 6: coefficients du GLM d'estimation du SCR de défaut*

Bien qu'extrêmement simple, ce modèle nous permet d'obtenir une MAE de l'ordre de 21%, ce qui est mieux que certaines régressions linéaires construites.

## 4. SCR Vie (Life)

### a. SCR Mortalité

Analyse en composantes principales :

L'analyse en composantes principales nous suggère que les variables le plus liées au SCR de mortalité sont les suivantes :

- Le chiffre d'affaires et sa répartition entre les différentes garanties,
- Les montants de provisions *Best Estimate*,
- Les dispersions des âges dans les portefeuilles de rente éducation et de rente de conjoint,
- La dispersion de l'ancienneté en incapacité,
- La durée moyenne en incapacité.

Analyse de la distribution pour modélisation par GLM :

L'analyse des histogrammes du SCR de mortalité nous suggère d'utiliser une transformation logarithmique :

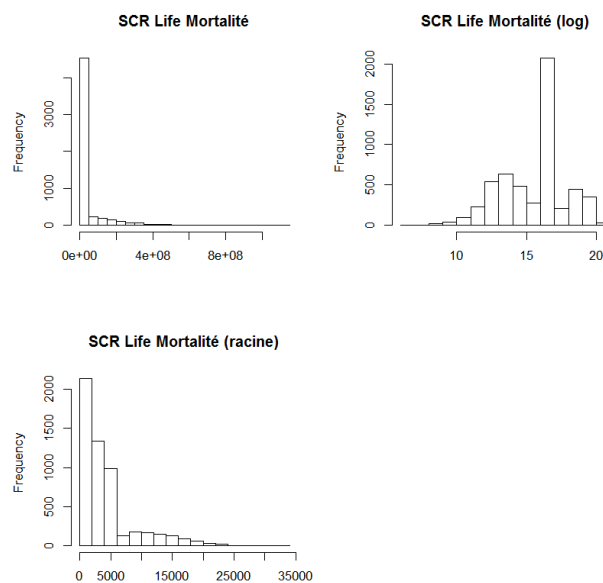


Figure 9: Histogrammes du SCR de mortalité (vie)

Le QQ-plot obtenu sur la variable transformée, nous permet de retenir l'hypothèse de normalité malgré quelques distorsions, notamment au niveau des queues de distribution.

Analyse des indicateurs sélectionnés et de la qualité des modèles :

Le tableau 4 (cf. Annexe) résume les variables sélectionnées par les différents modèles estimés.

Les différents modèles s'avèrent satisfaisants, au niveau des résidus et des graphes des valeurs estimées en fonction des observations. Toutefois, les trois modèles construits à partir des variables présélectionnées par l'ACP sont moins convaincants, notamment en raison de structures dans les graphes des résidus. Ce phénomène peut-être expliqué par l'absence de certaines variables utilisées

dans la plupart des autres modèles telles que le ratio entre les provisions dédiées au maintien de la garantie décès et celles afférentes à l'arrêt de travail ou la duration du passif.

Lors du tracé de la droite de régression pour le premier GLM, la forme légèrement incurvée du nuage de points nous incite à construire des modèles quadratiques. Cela est fait dans le cas des modèles construits à partir des variables sélectionnées par *random forest* ou le GBM. Une procédure de sélection de variables à l'aide du critère BIC est ensuite effectuée.

La sélection du chiffre d'affaires et de sa répartition et des montants de provisions *Best Estimate (vie)* est cohérente car ces variables sont représentatives de l'engagement de l'assureur et par conséquent du risque qu'il supporte. Les garanties rente de conjoint et rente éducation étant soumises au choc de mortalité, il semble naturel que des indicateurs sur les portefeuilles associés soient retenus. Le ratio entre le maintien de la garantie décès et l'arrêt de travail est construit comme la somme des provisions vie sur la somme des provisions non vie pour l'arrêt de travail. Ainsi, plus celui-ci est élevé, plus l'exposition de l'assureur au risque de mortalité est élevée. Cette variable semble donc assez naturellement importante pour l'estimation du SCR de mortalité. De manière similaire, la duration du passif de l'assureur est directement liée à l'engagement de celui-ci envers ses assurés surtout pour des garanties de type décès ou l'engagement est de long terme. Enfin, de par la méthode de simulation effectuée, les capitaux sous risques sont également impactés par les types de garanties proposées par l'organisme assurantiel car simulés uniquement si l'assureur propose des garanties de type vie.

#### Choix de la méthode d'estimation :

Nous obtenons les résultats suivants pour les différents modèles étudiés :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	13,5%	14,9%
Modèle complet + AIC	13,4%	13,4%
Modèle complet + BIC	13,5%	13,1%
Modèle ACP	37,7%	35,6%
Modèle ACP + AIC	37,7%	35,6%
Modèle ACP + BIC	37,8%	35,7%
Random Forest	9,2%	14,7%
GBM	5,4%	8,2%
Régression Ridge	16,7%	24,9%
Régression LASSO	13,7%	12,9%
Elastic Net	11,1%	12,0%
GLM avec variables issues de RF (quadratique)	10,1%	9,9%
GLM avec variables issues de GBM (quadratique)	12,6%	14,1%
GLM Final (quadratique)	10,6%	13,2%

Tableau 7: Détail des résultats des estimations du SCR mortalité (Vie)

Le GBM est le meilleur modèle d'estimation, au niveau des trois critères établis.

#### Estimation d'un modèle linéaire avec les variables retenues par avis d'expert :

Nous avons également voulu estimer un modèle à l'aide des variables sélectionnées par avis d'expert :

- Le chiffre d'affaires,
- Le montant de provisions *Best Estimate* (Life et Health SLT),
- Le ratio MGDC/AT,
- La durée du passif.

Les coefficients associés aux variables sont les suivants :

Variable	Coefficient
Chiffre d'affaires	-0,44
BE Life	1,16
BE SLT	-0,22
Ratio MGDC/AT	0,01
Duration passif	0,39

Tableau 8: Coefficients du GLM d'estimation du SCR mortalité (Ve)

Plus le montant de provisions en vie est important, plus l'impact du choc de mortalité sera grand pour l'assureur. De même, si la part dédiée au maintien de la garantie décès est conséquente, alors l'impact de la mortalité sur les fonds propres de l'organisme assurantiel le sera également. En revanche, le BE SLT représente la partie non-vie de l'engagement de l'assureur en arrêt de travail et est donc à l'opposé du ratio MGDC/AT dans le modèle estimé.

Les résultats obtenus avec ce modèle sont :

- MAE sur la base d'apprentissage = 15,8 %
- MAE sur l'échantillon de test = 16,5 %

Même si ce modèle présente l'avantage d'être facilement interprétable, il n'atteint pas les performances du GBM.

## b. SCR Longévité

### Analyse en composantes principales :

L'analyse en composante principale nous incite à retenir les variables suivantes :

- Les capitaux sous risques en vie,
- La dispersion de l'âge du portefeuille d'arrêt de travail,
- L'âge moyen des portefeuilles de rente de conjoint et d'arrêt de travail,
- L'ancienneté et la durée moyennes en incapacité,
- La proportion du chiffre d'affaires dédiée à la santé,
- Les dispersions des âges des portefeuilles de rente,
- Les montants de provisions *Best Estimate*,
- L'ancienneté moyenne en invalidité et la dispersion de l'ancienneté en invalidité.

La proportion du chiffre d'affaires dédiée à la garantie rente éducation, les indicateurs basés sur les portefeuilles de rente éducation et de rente de conjoint et les capitaux sous risques pour le risque catastrophe en vie sont retenus en raison des corrélations présentes dans notre base de données.

## Analyse de la distribution pour l'estimation par GLM :

Les histogrammes du SCR de longévité sont les suivants :

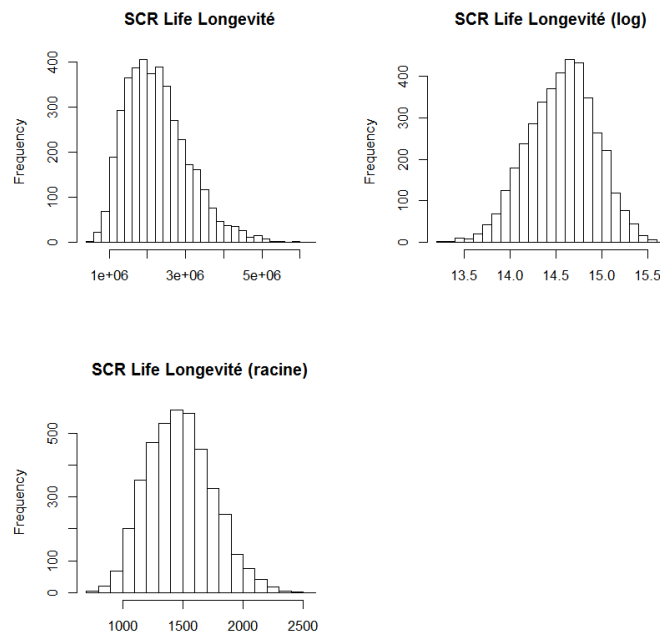


Figure 10: Histogrammes du SCR de longévité (Vie)

Le QQ-plot de la variable brute nous permet de retenir l'hypothèse de normalité malgré quelques distorsions observées au niveau de la queue supérieure de distribution. En revanche, le QQ-plot obtenu après transformation logarithmique est quasiment parfait. Nous retiendrons donc cette transformation.

## Analyse des résultats et des modèles :

Le tableau 5 présenté en annexe synthétise les différentes variables sélectionnées par les modèles estimés.

Outre les montants de *Best Estimate* dont le lien avec le SCR est direct, nous pouvons voir que les informations relatives à la garantie rente de conjoint apparaissent souvent comme significatives. Cela semble logique car il s'agit de la seule garantie donnant la possibilité d'une rente viagère, particulièrement sensible au risque de longévité.

La validation croisée pour le choix des paramètres à utiliser pour la régression *Elastic Net* nous suggère d'utiliser  $\alpha = 1$ , ce qui correspond à une régression LASSO. De plus, le critère de pénalisation retenu est le même que celui choisi pour cette dernière méthode.

L'analyse des résidus des modèles linéaires généralisés est très convaincante et il semble que toutes les hypothèses sous-jacentes soient respectées. En revanche, l'adéquation est beaucoup moins bonne que pour d'autres modèles précédemment construits. Nous observons en effet une forte dispersion du nuage de points qui semble limiter l'adéquation, malgré la réduction de la variable

d'intérêt. Cette dispersion peut être la conséquence d'une variable nécessaire à l'estimation dans notre base de données.

Choix de la méthode d'estimation :

Les résultats des différents modèles construits sont renseignés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	25,3%	26,7%
Modèle complet + AIC	24,8%	26,2%
Modèle complet + BIC	25,1%	27,1%
Modèle ACP	30,1%	31,2%
Modèle ACP + AIC	30,3%	31,2%
Modèle ACP + BIC	30,4%	31,0%
Random Forest	20,2%	23,7%
<b>GBM</b>	<b>10,6%</b>	<b>13,4%</b>
Régression Ridge	20,2%	21,3%
Régression LASSO	24,8%	26,8%
GLM avec variables issues de RF	25,6%	27,2%
GLM avec variables issues de GBM	28,3%	31,5%
GLM Final	25,8%	26,4%

*Tableau 9: Détail des résultats pour l'estimation du SCR de longévité (Vie)*

Nous pouvons voir que les erreurs absolues moyennes sont moins convaincantes que pour l'estimation des précédents SCR. Toutefois, le GBM apparaît encore comme étant la méthode la plus fiable.

La régression *Ridge* est, en comparaison avec les autres modèles, une des meilleures méthodes d'estimation par régression du SCR de longévité, nous confirmant encore une fois que la répartition du chiffre d'affaires d'un assureur est un indicateur pertinent des risques qu'il supporte et auxquels il est exposé.

Enfin, le dernier GLM construit dispose d'une capacité de prévision meilleure que la plupart des autres modèles, nous confortant dans l'idée que les indicateurs sur le portefeuille de rente de conjoint jouent un rôle important dans l'estimation du risque de longévité pour un organisme de prévoyance et santé.

### Estimation d'un modèle linéaire à l'aide de variables sélectionnées par avis d'expert :

Nous avons enfin testé d'estimer un modèle linéaire à l'aide de variables sélectionnées par avis d'expert. Nous obtenons alors les coefficients suivants :

Variable	Coefficient
Coefficient d'évolution du chiffre d'affaires (RC)	0,1
Âge moyen du portefeuille (RC)	0,56
BE Life	0,02

*Tableau 10 : coefficients du GLM d'estimation du SCR de longévité (vie)*

Les coefficients estimés sont cohérents : en effet, plus l'évolution du chiffre d'affaires pour la rente de conjoint est grande, plus l'assureur s'expose au versement de rentes viagères dans le futur. De même, un portefeuille comportant des individus âgés est synonyme de versement de rentes viagères soumises au risque de longévité. Enfin, le BE *Life* représente l'engagement de l'assureur et nous indique donc à quel point le passif de l'assureur sera impacté par un choc de longévité.

Ce modèle ne présente toutefois pas de meilleures qualités d'ajustement et de prévision que le GBM.

### **c. SCR morbidité**

#### Analyse en composantes principales :

L'ACP nous suggère de retenir les variables suivantes :

- Montants de provisions *Best Estimate*,
- La dispersion des âges dans les portefeuilles de rente éducation et de rente de conjoint,
- La répartition des provisions entre le maintien de la garantie décès et l'arrêt de travail,
- La dispersion de l'ancienneté en invalidité et l'ancienneté moyenne en invalidité.

## Étude de la distribution pour l'estimation par GLM :

Nous obtenons les histogrammes suivants pour la variable à prédire :

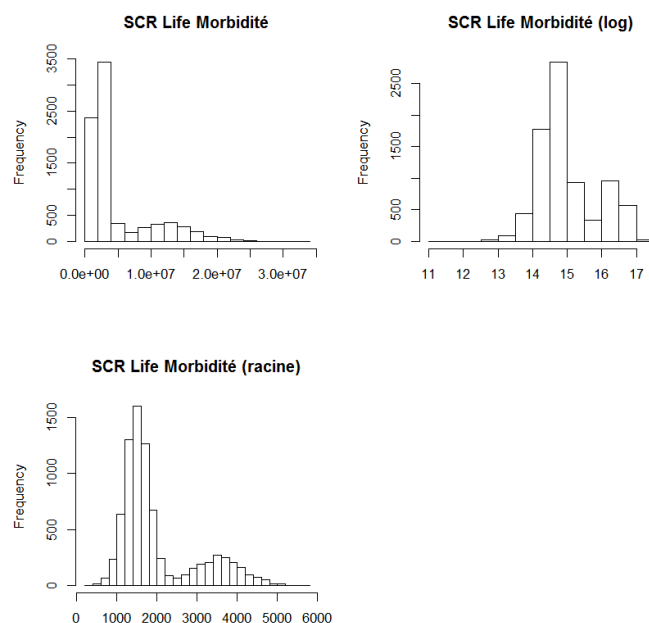


Figure 11: Histogrammes du SCR de morbidité (Vie)

La transformation logarithmique nous paraît la plus adéquate car donnant lieu à la répartition la plus symétrique. Le QQ-plot de la variable transformée nous permet d'accepter l'hypothèse de normalité, malgré quelques distorsions, notamment au niveau de la queue supérieure de la distribution.

## Analyse des résultats et des modèles estimés :

Le tableau 6 (cf. Annexe) résume les différentes sélections de variables selon les modèles estimés.

Les variables les plus sélectionnées sont :

- Le coefficient d'évolution du chiffre d'affaires en arrêt de travail,
- Les montants de provisions *Best Estimate* en santé similaire et non-similaire à la vie,
- Les montants sous risques pour le risque catastrophe en vie.

La répartition du chiffre d'affaire apparaît également comme significative dans les modèles *Ridge* et *Elastic Net*.

Le choc de morbidité concernant le maintien de la garantie décès, il paraît logique que les indicateurs sur le portefeuille d'arrêt de travail soient liés au SCR de morbidité, à même titre que le BE SLT et le ratio entre les provisions vie et non-vie en arrêt de travail.

Dans les modèles linéaires, le coefficient associé à l'évolution du chiffre d'affaires en arrêt de travail est positif. Cela est cohérent, car plus l'engagement de l'assureur en arrêt de travail est important, plus le coût du maintien de la garantie décès est grand. De façon analogue, les coefficients associés aux montants de BE SLT et NSLT sont positifs. Ces provisions sont constituées partiellement ou complètement par les provisions d'arrêt de travail, et plus celles-ci sont conséquentes plus le risque

de morbidité impacte l'assureur via le maintien de la garantie décès. Les capitaux sous risques pour le risque catastrophe en vie sont corrélés négativement à la proportion du chiffre d'affaire dédiée à l'arrêt de travail, et cela se retrouve dans le signe du coefficient associé à cette variable.

Ces variables seront utilisées pour l'estimation du SCR de morbidité à l'aide d'un GLM quadratique. En effet, lors des tracés des graphes des observations en fonction des valeurs ajustées, nous pouvons noter une forme incurvée du nuage de points. Cette observation nous incite à construire des modèles quadratiques. Cela est fait pour les modèles contenant les variables sélectionnées par *Random Forest* et le GBM. Avec l'ajout d'effets croisés, la droite estimée, bien que ne proposant pas une adéquation parfaite, semble toutefois plus proche de la moyenne du nuage de points. Cela est également visible à travers le coefficient d'ajustement et les erreurs moyennes. En revanche, l'analyse de la structure des résidus des différents modèles est globalement satisfaisante, excepté pour les modèles construits à partir des variables sélectionnées par l'ACP. Cela provient du fait que la sélection a été effectuée dans un plan factoriel n'expliquant qu'une faible part de l'information disponible dans la base de données.

#### Choix du meilleur modèle d'estimation :

Les résultats des différents modèles sont résumés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	32,0%	36,1%
Modèle complet + AIC	31,8%	33,8%
Modèle complet + BIC	32,4%	34,5%
Modèle ACP	33,8%	32,9%
Modèle ACP + AIC	33,8%	32,8%
Modèle ACP + BIC	33,8%	32,8%
Random Forest	13,3%	16,6%
<b>GBM</b>	<b>8,7%</b>	<b>11,7%</b>
Régression Ridge	25,0%	27,0%
Régression LASSO	32,6%	31,6%
Elastic Net	24,0%	23,7%
GLM avec variables issues de RF (quadratique)	27,9%	30,7%
GLM avec variables issues de GBM (quadratique)	26,6%	27,8%
GLM Final (quadratique)	27,9%	28,9%

Tableau 11: Détail des résultats pour l'estimation du SCR de morbidité (Vie)

Le GBM est la méthode présentant une erreur minimale sur l'échantillon de test, témoignant ainsi de la bonne capacité de généralisation de ce modèle.

Nous avons voulu estimer un modèle *Elastic Net* quadratique à partir des variables le plus sélectionnées, mais la détermination des paramètres par validation croisée n'a pas abouti à des résultats satisfaisants ( $\lambda = 0$ , soit aucune pénalisation).

Nous avons donc préféré déterminer un GLM quadratique.

### Analyse de sensibilités :

Le tableau ci-dessus nous indique cependant que l'adéquation et les erreurs absolues moyennes sont moins satisfaisantes que pour les GLM estimés avec les variables sélectionnées par *Random Forest* et *GBM*. Nous avons donc effectué des sensibilités en intégrant tour à tour la durée du passif, le chiffre d'affaires et le montant *Best Estimate* en vie, mais aucun de ces ajouts n'a permis d'améliorer significativement la qualité du modèle. L'amélioration de la qualité du modèle dépend donc des interactions entre ces trois variables.

### Estimation d'un modèle linéaire avec les variables retenues par avis d'expert :

Nous avons estimé un modèle linéaire à l'aide des variables suivantes :

- Le coefficient d'évolution du chiffre d'affaires en arrêt de travail,
- Le ratio entre provisions vie et non-vie en arrêt de travail,
- Le montant de provisions *Best Estimate* (Health NSLT),
- La proportion du chiffre d'affaires afférente à l'arrêt de travail,
- Les capitaux sous risques pour le SCR catastrophe en vie.

Les coefficients associés à ces variables sont :

Variable	Coefficient
<b>Coefficient d'évolution du chiffre d'affaires (AT)</b>	0,12
<b>Ratio MGDC/AT</b>	0,17
<b>BE NSLT</b>	0,65
<b>Proportion du chiffre d'affaires dédiée à l'AT</b>	0,03
<b>Capitaux sous risques</b>	-0,21

Tableau 12: coefficients du GLM d'estimation du SCR de morbidité (vie)

Le risque de morbidité se traduisant par un choc sur la table de passage en invalidité, les coefficients positifs associés à l'évolution du chiffre d'affaires en arrêt de travail et au BE NSLT sont pertinents. Le choc de morbidité affecte également le maintien de la garantie décès en arrêt de travail, le signe positif associé au ratio MGDC/AT nous indique donc que plus la part afférente à cette garantie au sein du portefeuille d'arrêt de travail est importante, plus le SCR de morbidité sera grand. De façon analogue le SCR de morbidité est d'autant plus élevé que la part du chiffre d'affaires dédié à l'arrêt de travail est importante. Les capitaux sous risques sont corrélés négativement avec la proportion du chiffre d'affaires afférente à l'arrêt de travail, impactant ainsi le coefficient qui lui est associé dans la régression.

#### d. SCR frais

##### Analyse en composantes principales :

L'Analyse en Composantes principales nous suggère de retenir les variables suivantes :

- Les montants de provision *Best Estimate*,
- La dispersion de l'âge en arrêt de travail,
- Les durations moyennes en incapacité et invalidité,
- L'âge moyen du portefeuille d'arrêt de travail,
- L'ancienneté moyenne en incapacité,
- Le rapport entre les provisions vie et non-vie en arrêt de travail,
- Les proportions du chiffre d'affaires dédiées à la santé et à la rente éducation.

La proportion du chiffre d'affaires dédiée à la santé est corrélée à -54,7 % à celle afférente à la garantie décès. La sélection de cette variable est donc cohérente eu égard à cette corrélation.

##### Analyse de la distribution pour l'estimation par GLM :

Les histogrammes du SCR de frais nous incitent à utiliser une transformation logarithmique :

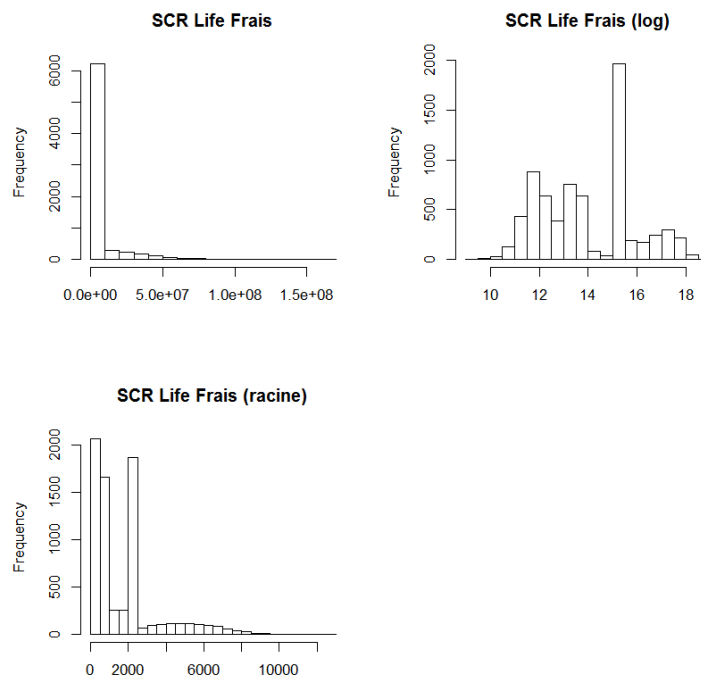


Figure 12: Histogrammes du SCR de Frais (vie)

Le QQ-Plot obtenu pour la variable transformée s'avère satisfaisant malgré quelques distorsions, et l'hypothèse de normalité de la variable est donc acceptée pour la suite de la modélisation par GLM.

##### Analyse des résultats et des modèles d'estimation :

Les variables sélectionnées par les différents modèles sont présentées dans le tableau 7 en annexe.

Les variables le plus retenues sont :

- Le chiffre d'affaires,
- Le taux de frais en arrêt de travail,
- Les montants de provisions *Best Estimate*,
- La durée du passif,
- Les capitaux pour le risque catastrophe en vie.

Le choc de dépenses en vie concerne les garanties de rente de conjoint, de rente éducation, la garantie décès et le maintien de la garantie décès en arrêt de travail. La sélection de variables relatives aux portefeuilles d'arrêt de travail est donc pertinente. Le chiffre d'affaires et les provisions représentent l'engagement de l'assureur, qui est impacté par le choc de frais. Le choix de ces variables est donc pertinent, au même titre que la sélection des taux de frais. Le risque de frais est un risque de long terme, il est ainsi logique de retenir la durée du passif de l'assureur. Enfin, les montants sous risque sont aussi révélateurs du risque supporté par l'organisme assurantiel et sont également sensibles au choc de frais.

Les graphes des résidus (résidus en fonction des valeurs ajustées, distance de Cook, QQ-plot) sont satisfaisants pour l'ensemble des modèles construits, à l'exception des GLM basés sur les variables présélectionnées par l'ACP. Cela s'explique encore une fois par le fait que ces variables ont été choisies à partir d'une représentation dans un plan factoriel n'expliquant que peu de l'information disponible dans le jeu de données.

Lors du tracé de la droite de régression du GLM complet, nous pouvons remarquer que le nuage de points présente une forme légèrement arrondie, notamment aux extrémités. Cette observation nous incite donc à construire un modèle polynômial à partir des variables sélectionnées par *Random Forest* et GBM. L'ajustement est en effet amélioré par l'ajout d'effets croisés.

Choix de la méthode d'estimation :

Les résultats des différents modèles sont résumés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	5,6%	5,1%
Modèle complet + AIC	5,6%	4,7%
Modèle complet + BIC	5,6%	4,5%
Modèle ACP	14,3%	13,7%
Modèle ACP + AIC/BIC	14,3%	13,7%
Random Forest	1,2%	24,4%
<b>GBM</b>	<b>1,0%</b>	<b>1,7%</b>
Régression Ridge	7,4%	6,7%
Régression LASSO	5,8%	4,6%
Régression Elastic Net	4,0%	4,2%
GLM avec variables issues de RF (quadratique)	2,9%	3,8%
GLM avec variables issues de GBM (quadratique)	6,5%	10,6%
GLM Final (quadratique)	3,7%	4,8%

Tableau 13: Détail des résultats des modèles d'estimation du SCR de Frais (vie)

Le GBM présente les meilleurs critères d’ajustement et de prévision, nous informant ainsi de l’importance des taux de frais pour les garanties décès et rente de conjoint. Le GLM estimé à partir des variables sélectionnées par *Random Forest* présente également de bons critères. Cela signifie qu’outre les taux de frais, les indicateurs sur les portefeuilles de rente jouent un rôle dans la prévision du SCR de frais. Cela est d’autant plus pertinent pour la rente de conjoint car il s’agit d’une garantie de long terme.

Analyse de sensibilités :

Nous avons effectué des sensibilités en intégrant les taux de frais pour les garanties décès et rente de conjoint au GLM combiné, mais cela n’a pas grandement amélioré l’ajustement ou diminué les erreurs absolues moyennes. Cela signifie donc que la méthode d’estimation par arbres de régression est également plus adaptée au SCR de frais que les modèles linéaires généralisés.

Estimation d’un modèle linéaire avec les indicateurs sélectionnés par avis d’expert :

Nous avons enfin estimé un modèle à l’aide des variables suivantes :

- Chiffre d’affaires,
- Taux de frais en arrêt de travail, décès, rente de conjoint et rente éducation,
- Duration du passif,
- Capitaux sous risques

Les coefficients estimés sont :

<b>Variable</b>	<b>Coefficient</b>
<b>Chiffre d’affaires</b>	0,38
<b>Taux de frais (AT)</b>	0,06
<b>Taux de frais (Décès)</b>	0,01
<b>Taux de frais (RE)</b>	0,004
<b>Taux de frais (RC)</b>	0,01
<b>Duration du passif</b>	0,04
<b>BE Life</b>	0,04
<b>Capitaux sous risques</b>	0,77

*Tableau 14: coefficients du GLM d’estimation du SCR de frais (Vie)*

Ce modèle, bien que simple et intuitif, ne permet pas d’obtenir de meilleurs résultats que le GBM car l’erreur absolue moyenne sur l’échantillon de test est de l’ordre 6 %.

## e. SCR révision

### Analyse en composantes principales :

L'analyse en composantes principales nous suggère que les variables liées au SCR révision en vie sont :

- Les montants de provision *Best Estimate*,
- Les dispersions des âges des portefeuilles de rente de conjoint et rente éducation,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- Les anciennetés moyennes, leurs dispersions et les durations moyennes en incapacité et invalidité,
- L'âge moyen du portefeuille d'arrêt de travail et la dispersion des âges du même portefeuille,
- La proportion du chiffre d'affaires afférente à la santé,
- Les capitaux pour le risque catastrophe en vie.

Nous pouvons voir qu'aucun indicateur relatif au décès n'a été directement retenu. En revanche, l'ACP nous suggère d'utiliser les capitaux pour le risque catastrophe en vie et la proportion du chiffre d'affaires dédiée à la santé, qui sont quant à eux corrélés aux chiffres d'affaires des garanties de type vie et arrêt de travail.

### Étude de la distribution pour l'estimation par GLM :

L'analyse des histogrammes de la variable d'intérêt nous incite à utiliser une transformation logarithmique :

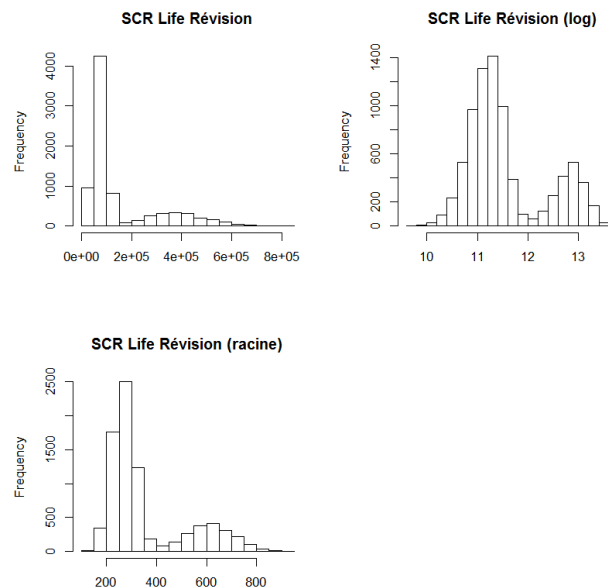


Figure 13: histogrammes du SCR révision (Life)

Aucun des QQ-plots construit à partir d'une distribution de la famille exponentielle n'est satisfaisant. Le QQ-plot du SCR révision transformé nous permet d'accepter l'hypothèse de normalité de la variable, même si des distorsions au niveau des queues de distribution sont visibles. En effet, nous pouvons constater que l'histogramme de la variable transformée présente une allure semblable à un

mélange de gaussiennes. Il pourrait être intéressant d'estimer chacune des composantes par modèle linéaire et de pondérer les résultats obtenus mais cela ne sera pas présenté ici.

Commentaire sur les variables sélectionnées par les différents modèles:

Le tableau 8 de l'annexe présente les sélections de variables effectuées selon les modèles d'estimation proposés.

Le risque de révision en vie affecte les garanties décès, rente de conjoint, rente éducation et le maintien de la garantie décès en arrêt de travail. La sélection de variables basées sur ces portefeuilles est donc cohérente, au même titre que le choix des provisions *best estimate* qui représentent le risque supporté par l'assureur dans ces différentes garanties. Nous pouvons voir que les variables relatives à l'arrêt de travail sont plus souvent sélectionnées que celles en rapport avec les garanties de type vie.

Lors du tracé des graphes des résidus et des observations en fonction des valeurs ajustées, nous avons remarqué une structure des nuages de points nous incitant à construire des modèles quadratiques avec les variables sélectionnées par l'ACP, *Random Forest* et le GBM. L'ajout d'interactions permet d'améliorer la qualité des modèles.

Choix de la méthode d'estimation :

Les résultats obtenus sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	21,6%	24,0%
Modèle complet + AIC	21,5%	22,1%
Modèle complet + BIC	21,6%	21,5%
Modèle ACP (quadratique)	16,9%	19,1%
Modèle ACP + AIC (quadratique)	16,8%	18,5%
Modèle ACP + BIC (quadratique)	16,9%	18,1%
Random Forest	12,5%	14,2%
<b>GBM</b>	<b>8,8%</b>	<b>11,5%</b>
Régression Ridge	18,3%	21,7%
Régression LASSO	21,6%	21,6%
Elastic Net	13,9%	17,8%
GLM avec variables issues de RF (quadratique)	15,1%	19,0%
GLM avec variables issues de GBM (quadratique)	16,8%	17,5%
GLM Final (quadratique)	16,7%	17,3%

Tableau 15: détail des résultats pour l'estimation du SCR révision (Vie)

Nous pouvons voir que les meilleurs résultats sont obtenus grâce au GBM. Nous pouvons aussi noter que la régression *Elastic Net* nous fournit des résultats assez convaincants, nous suggérant que la répartition du chiffre d'affaires selon les différentes garanties permet de mieux prédire le SCR révision en vie. Cela n'est pas surprenant car la répartition du chiffre d'affaires est un indicateur pertinent du type de risques supportés par l'assureur.

Analyse de sensibilités :

Comme le GBM nous indique que le chiffre d'affaires, la durée moyenne en invalidité, le BE Life, la durée du passif, le taux de PPNA de la garantie rente de conjoint et le taux de frais en arrêt de travail sont des variables importantes, nous avons testé d'ajouter tour à tour ces variables dans le dernier GLM estimé. Les meilleurs résultats sont obtenus après l'ajout du BE *Life* :

- MAE sur l'échantillon d'apprentissage = 16,7 %
- MAE sur la base de test = 17,1 %

Bien que l'amélioration du modèle ne soit pas significative, cela nous indique que –conformément à nos attentes – le montant de provisions pour les risques de type *Life* améliore l'estimation et la prévision du SCR révision en vie. La modélisation par arbres de régression semble quand même, au vu des résultats ci-dessus, plus adaptée à l'estimation du SCR révision que les régressions linéaires.

Estimation d'un modèle linéaire avec les variables retenues par avis d'expert :

Nous avons également construit un modèle à l'aide de variables sélectionnées par avis d'expert et nous avons obtenu les coefficients suivants :

Variable	Coefficient
Coefficient d'évolution du chiffre d'affaires (AT)	0,09
Ratio MGDC/AT	0,35
BE Life	-0,01
BE NSLT	0,83

Tableau 16: coefficients du GLM d'estimation du SCR révision (Vie)

Les résultats obtenus quant à l'ajustement et la capacité de prévision de ce modèle sont moins satisfaisants que pour d'autres modèles précédemment estimés car l'erreur est de l'ordre de 20%. Toutefois, nous pouvons constater que le rôle du BE NSLT est prépondérant car le coefficient associé est grand. Le choc de révision concerne les provisions mathématiques dédiées au maintien de la garantie décès pour les individus en incapacité, il est donc cohérent que le montant de provisions NSLT joue un rôle important dans l'estimation du SCR. De même, plus le ratio MGDC/AT est grand, plus la part afférente au maintien de la garantie décès dans les provisions d'arrêt de travail est importante, accentuant par conséquent l'impact du choc de révision.

#### f. SCR catastrophe

La formule de calcul du SCR Catastrophe en vie de Solvabilité II étant simple, nous n'utiliserons pas de méthode d'estimation de ce SCR.

#### g. SCR risque de souscription en vie

Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Les proportions de chiffre d'affaires afférentes à la rente de conjoint et à l'arrêt de travail,
- Le montant de provisions *Best Estimate* en vie,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- L'âge moyen et sa dispersion dans le portefeuille de rente de conjoint.
- La dispersion des âges dans le portefeuille de rente éducation,

- La duration moyenne en incapacité et en invalidité,
- La dispersion de l'ancienneté en invalidité,
- Les capitaux sous risques pour le risque catastrophe en vie.

Nous pouvons voir que les variables en rapport avec la garantie décès n'ont pas été retenues par l'ACP, sûrement en raison de corrélations au sein de notre base de données.

#### Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

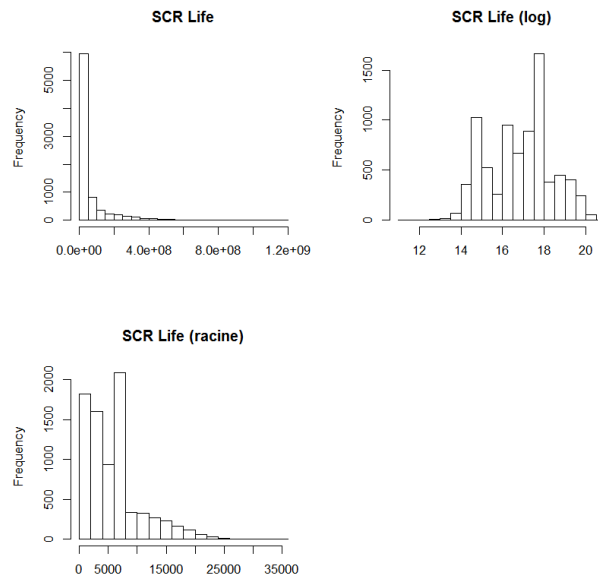


Figure 14: Histogrammes du SCR Vie

Bien que le QQ-plot obtenu pour la variable brute et les quantiles d'une loi Gamma soit très satisfaisant, l'estimation des coefficients de la régression n'a pas été possible car l'algorithme n'a pas convergé. Le QQ-plot de la variable transformée par le logarithme nous permet d'accepter l'hypothèse de normalité pour la suite de la modélisation par modèles linéaires.

#### Analyse des résultats :

Le tableau 9 présenté en annexe résume les variables sélectionnées par les différentes méthodes d'estimation du SCR vie.

Nous pouvons voir que les variables sélectionnées dans la majorité des modèles sont :

- Le chiffre d'affaires,
- Les coefficients d'évolution du chiffre d'affaires en arrêt de travail,
- Le ratio MGDC/AT,
- Les capitaux sous risques pour le risque catastrophe,
- Les montants de provisions *Best Estimate* en vie et en santé similaire à la vie.

Le chiffre d'affaires, les montants sous risques ainsi que le montant de provisions *Best Estimate* en vie étant révélateurs de l'engagement de l'assureur, leur sélection est pertinente. L'évolution du chiffre d'affaires en arrêt de travail, couplée au ratio MGDC/AT nous permet également de savoir dans quelle mesure le maintien de la garantie décès sera présent dans le portefeuille de l'assureur. Les indicateurs calculés sur le portefeuille de rente de conjoint sont également pertinents car il s'agit de la seule garantie du modèle touchée par le risque longévité (en vie).

Le montant de provisions en santé similaire à la vie n'a pas forcément de sens pour le calcul du SCR vie. Cette variable a certainement été retenue dans la plupart des modèles en raison de sa corrélation avec le BE Life et des liens existant entre les portefeuilles de rente et d'arrêt de travail.

#### Choix de la méthode d'estimation :

Les résultats des estimations sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	17,5%	19,8%
Modèle complet + AIC	17,4%	18,2%
Modèle complet + BIC	17,5%	18,0%
Modèle ACP	29,1%	29,5%
Modèle ACP + AIC/BIC	29,1%	29,4%
Random Forest	8,1%	11,3%
<b>GBM</b>	<b>3,5%</b>	<b>9,2%</b>
Régression Ridge	11,8%	13,2%
Régression LASSO	17,4%	18,2%
Elastic Net	9,7%	10,5%
GLM avec variables issues de RF	20,1%	20,8%
GLM avec variables issues de GBM	17,9%	18,6%
GLM Final	20,1%	20,7%

Tableau 17: Détail des résultats de l'estimation du SCR vie

Nous pouvons constater que les meilleurs résultats sont obtenus grâce au GBM. Toutefois, les bons résultats de la régression *Elastic Net* nous suggèrent que la répartition du chiffre d'affaires permet d'estimer de manière plus précise le SCR Life. Ce modèle nous permet également de retenir le chiffre d'affaires, les capitaux sous risques et le BE Life pour l'estimation du SCR vie.

#### Estimation d'un modèle linéaire avec les variables sélectionnées par avis d'expert :

Comme aucune variable relative à la garantie décès n'a été retenue, nous avons aussi calibré un modèle linéaire à l'aide de variables en rapport avec le risque souscription en vie, soit :

- Coefficient d'évolution du chiffre d'affaires pour les garanties décès et rente de conjoint,
- Les proportions de chiffre d'affaires dédiées au décès et à la rente de conjoint,
- Ratio entre les provisions vie et non-vie en arrêt de travail,
- Le BE Life,
- Les capitaux sous risques,
- La durée du passif.

Le modèle obtenu est le suivant :

Variable	Coefficient
Coefficient d'évolution du chiffre d'affaires (Décès)	0,01
Coefficient d'évolution du chiffre d'affaires (RC)	0,01
Proportion du chiffre d'affaires (Décès)	5,63
Proportion du chiffre d'affaires (RC)	5,60
Ratio MGDC/AT	0,03
BE Life	0,06
Capitaux sous risques	0,28
Duration du passif	0,03

Tableau 18: Coefficients du GLM d'estimation du SCR Vie

Les proportions du chiffre d'affaires en rente de conjoint et décès ont des coefficients significativement supérieurs aux autres variables de la régression, suggérant que ce type de garantie est prépondérant pour l'estimation du SCR vie.

Les résultats obtenus avec ce modèle sont :

- MAE sur la base d'apprentissage = 19,2 %
- MAE sur la base de test = 21,1 %

Les erreurs sont plus élevées que pour la régression *Ridge* ou *Elastic Net*. En effet, celles-ci tiennent compte de l'ensemble des variables du modèle ou permettent de tenir compte des corrélations entre celles-ci, améliorant ainsi l'estimation des coefficients et la prévision des résultats. Cependant, le précédent modèle possède de bonnes qualités prédiction et d'ajustement malgré sa simplicité. Il possède en outre la qualité d'être facilement interprétable.

Comparaison de l'estimation directe et de l'agrégation des sous-modules :

Nous avons enfin essayé d'agrèger les meilleurs modèles de prévision d'estimation des SCR longévité, mortalité, morbidité, frais, révision et catastrophe. Nous avons alors les résultats suivants :

	Estimation directe	Agrégation des sous-modules
MAE (apprentissage)	3,5 %	15,6 %
MAE (test)	9,2 %	17,5 %

Tableau 19: comparaison de l'agrégation et de l'estimation directe pour le SCR Vie

L'agrégation ne nous fournit donc pas de meilleurs résultats que l'estimation directe.

## 5. SCR Santé (Health)

### a. Santé similaire à la vie (Health SLT)

#### SCR Longévité

##### Analyse en composantes principales :

L'ACP dans le plan factoriel où les contributions du SCR de longévité sont les plus importantes nous permet de sélectionner les variables suivantes :

- Les provisions *Best Estimate* en santé,
- Les dispersions des âges dans les portefeuilles de rente,
- L'âge moyen du portefeuille d'arrêt de travail,
- L'ancienneté moyenne en invalidité,
- La dispersion de l'ancienneté en incapacité,
- Les montants sous risques pour le risque catastrophe en vie.

Les dispersions des âges pour les portefeuilles de rente de conjoint et rente éducation sont corrélées avec les indicateurs sur le portefeuille d'arrêt de travail. Cette corrélation peut expliquer la sélection de telles variables.

##### Étude de la distribution pour l'estimation par GLM :

Les histogrammes obtenus sont les suivants :

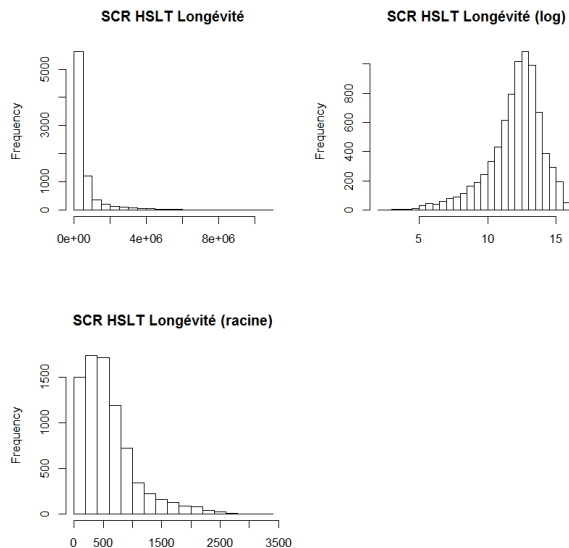


Figure 15: histogrammes du SCR longévité (HSLT)

La transformation par le logarithme est celle donnant lieu à la distribution la plus « gaussienne ». Le QQ-plot de la variable transformée est satisfaisant. Nous retiendrons donc l'hypothèse de normalité. Nous aurions voulu ajuster un GLM avec une distribution Gamma mais l'estimation des coefficients n'a pas été possible, malgré la réduction des variables et le passage de coefficients initiaux en paramètres.

### Analyse des résultats et des modèles d'estimation :

Les variables sélectionnées par les différentes méthodes d'estimation du SCR longévité sont résumées dans le tableau 10 (cf. Annexe).

Les variables le plus retenues sont :

- La répartition des provisions vie et non-vie en arrêt de travail,
- La duration moyenne en invalidité,
- L'ancienneté moyenne en invalidité,
- Les montants de BE en santé (similaire et non similaire à la vie).

Les coefficients négatifs associés au ratio entre provisions vie et non-vie, à l'ancienneté moyenne en invalidité et au BE NSLT semblent logiques. En effet, le ratio MGDC/AT permet de compenser les risques liés à l'invalidité. De plus, si l'ancienneté moyenne dans le portefeuille d'invalides est importante alors la durée restante jusqu'à l'âge limite pour le versement de telles rentes est faible, diminuant ainsi l'engagement de l'assureur.

Les coefficients associés à la duration moyenne en invalidité et au BE SLT ont également du sens : si la duration du portefeuille d'invalides est grande, alors l'intensité du choc de longévité sera d'autant plus importante. De même, le montant de provisions dédiées à la santé similaire à la vie témoigne de l'engagement de l'assureur pour la garantie d'invalidité. Ainsi, si le BE SLT est élevé alors le choc de longévité aura un impact non négligeable sur le passif de l'organisme assurantiel.

La variable d'intérêt présente une dispersion importante malgré la réduction, rendant l'estimation difficile. De plus, lors de la construction des premiers modèles linéaires généralisés, la forme incurvée du nuage de points semble préconiser des modèles polynômiaux.

Nous avons ainsi estimé des modèles quadratiques pour les régressions faites à partir des variables sélectionnées par le GBM et *Random Forest*. Comme le démontre le tableau ci-dessous, l'ajout d'effets croisés a permis de diminuer l'erreur absolue moyenne. De plus, les graphes des résidus des premiers GLM présentent une structure qui disparaît avec l'ajout d'interactions. Toutefois cette amélioration n'est pas significative. Nous avons également tenté d'augmenter le niveau d'interactions entre les variables pour le dernier modèle, mais cela n'améliore en rien la qualité des prévisions effectuées, et le modèle construit devient en outre difficilement interprétable.

### Choix de la méthode d'estimation :

Les résultats des estimations effectuées sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	26,9%	31,7%
Modèle complet + AIC	26,6%	29,0%
Modèle complet + BIC	27,0%	25,4%
Modèle ACP	34,5%	36,5%
Modèle ACP + AIC/BIC	34,5%	36,4%
Random Forest	11,8%	15,6%
<b>GBM</b>	<b>9,8%</b>	<b>12,1%</b>
Régression Ridge	59,4%	60,1%
Régression LASSO	26,7%	25,5%
Elastic Net	27,5%	28,3%
GLM avec variables issues de RF (quadratique)	10,8%	18,9%
GLM avec variables issues de GBM (quadratique)	23,0%	23,5%
GLM Final (cubique)	22,9%	23,9%

Tableau 20 : Détails des résultats pour l'estimation du SCR longévité (HSLT)

### Estimation d'un modèle linéaire avec des indicateurs retenus par avis d'expert :

Nous avons également estimé un GLM à l'aide de variables sélectionnées par avis d'expert. Le modèle obtenu est présenté ci-dessous :

Variable	Coefficient
<b>Ratio MGDC/AT</b>	-0,24
<b>Duration moyenne en invalidité</b>	0,38
<b>Ancienneté moyenne en invalidité</b>	-0,11
<b>BE SLT</b>	0,58

Tableau 21: coefficients du GLM d'estimation du SCR Longévité (HSLT)

Le BE NSLT représente l'assiette du choc. L'ancienneté moyenne en invalidité représente l'horizon du choc, la durée moyenne son intensité. Enfin, le ratio entre les provisions vie et non-vie en arrêt de travail permet de compenser le choc de longévité grâce au maintien de la garantie décès, qui est quant à elle soumise au risque de mortalité.

Ce modèle ne nous permet toutefois pas d'obtenir de meilleures prévisions que ceux présentés précédemment.

### **SCR frais**

#### Analyse en Composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Les montants de provisions *Best Estimate* en santé,
- Les dispersions des âges dans les portefeuilles de rente de conjoint et de rente éducation,
- Les dispersions, durations et anciennetés moyennes en incapacité et invalidité,
- L'âge moyen du portefeuille d'arrêt de travail,

- Le ratio entre les provisions vie et non-vie en arrêt de travail.

Comme nous l'avons vu précédemment, la méthode de simulation des données a contribué à créer une corrélation entre les portefeuilles d'arrêt de travail, de rente éducation et de rente de conjoint, expliquant pourquoi les dispersions des âges des portefeuilles de rente ont été retenues.

Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

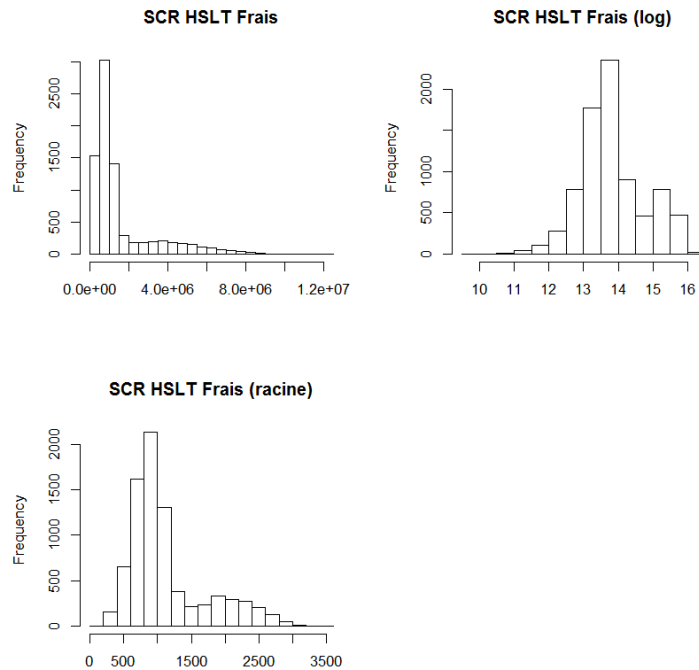


Figure 16: Histogrammes de SCR frais (HSLT)

Nous retiendrons la transformation logarithmique car cette dernière donne lieu à la répartition la plus gaussienne. L'analyse du QQ-plot de la variable transformée nous conforte dans l'idée de retenir l'hypothèse de normalité.

Analyse des résultats :

Les différentes variables sélectionnées par les modèles d'estimation du SCR de frais sont synthétisées dans le tableau 11 en annexe.

La seule garantie SLT du modèle prévoyance santé d'ACTUARIS® étant l'invalidité, la sélection des variables relatives à l'arrêt de travail est cohérente. Le risque de frais étant un risque de long terme, le choix d'indicateurs afférents à l'incapacité est moins pertinent que celui de facteurs relatifs à l'invalidité, mais nous pouvons imputer ces sélections aux corrélations existantes dans la base de données. De même, le BE NSLT et les capitaux sous-risques ont également pu être retenus en raison des corrélations qu'ils présentent avec d'autres variables de notre base.

### Choix de la méthode d'estimation :

Les résultats des différentes estimations sont résumés dans le tableau ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	19,2%	22,7%
Modèle complet + AIC	19,1%	21,2%
Modèle complet + BIC	19,2%	20,4%
Modèle ACP	21,5%	23,6%
Modèle ACP + AIC	21,5%	23,6%
Modèle ACP + BIC	21,5%	23,5%
Random Forest	7,1%	10,0%
<b>GBM</b>	<b>5,8%</b>	<b>7,9%</b>
Régression Ridge	19,3%	18,9%
Régression LASSO	19,2%	20,4%
Elastic Net	19,2%	18,3%
GLM avec variables issues de RF (quadratique)	10,0%	12,1%
GLM avec variables issues de GBM	19,8%	20,1%
GLM Final (quadratique)	11,2%	11,1%

Tableau 22: Résultats détaillés des estimations du SCR frais (HSLT)

Nous pouvons constater que les meilleures estimations sont obtenues avec le GBM et *Random Forest*.

### Sensibilités :

Toutefois, la capacité de prévision du GLM construit à partir des variables les plus sélectionnées est satisfaisante. À partir de ce dernier modèle, nous avons effectué deux sensibilités :

- Ajout de l'ancienneté moyenne en invalidité,
- Ajout de la dispersion de l'ancienneté en invalidité.

L'ajout de chacune de ces variables a le même impact sur le modèle, c'est-à-dire :

- Erreur absolue moyenne sur la base d'apprentissage = 10,3%
- Erreur absolue moyenne sur la base de test = 10,4%

### Estimation d'un modèle linéaire à l'aide d'indicateurs sélectionnés par avis d'expert :

Nous avons enfin estimé un GLM à l'aide des variables suivantes :

- Le ratio MGDC/AT,
- La duration moyenne en invalidité,
- Le taux de frais en arrêt de travail,
- Le BE SLT

Les coefficients estimés sont :

Variable	Coefficient
Ratio MGDC/AT	0,02
Duration moyenne en invalidité	0,10
Taux de frais (AT)	0,82
BE SLT	0,09

Tableau 23: coefficients du GLM d'estimation du SCR de frais (HSLT)

La duration nous indique quelle sera l'intensité du choc de frais. Ce choc sera d'autant plus important pour l'assureur que les provisions qu'il a constituées et que le taux de frais sont élevés. Bien qu'il présente l'avantage d'être simple, ce modèle n'atteint pas les performances du GBM.

## SCR révision

### Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Les montants de provisions *Best Estimate* en santé,
- Les dispersions des âges dans les portefeuilles de rente de conjoint et de rente éducation,
- Les dispersions des anciennetés, les anciennetés et durations moyennes en incapacité et invalidité,
- L'âge moyen du portefeuille d'arrêt de travail,
- La répartition des provisions entre vie et non-vie en arrêt de travail,
- Les capitaux sous risques pour le risque catastrophe en vie,
- La proportion du chiffre d'affaires dédiée à la santé.

La proportion du chiffre d'affaires dédiée à la santé et les capitaux sous risques sont corrélés à la part du chiffre d'affaires afférente à l'arrêt de travail et ont donc été sélectionnés.

## Étude de la distribution pour l'estimation par GLM :

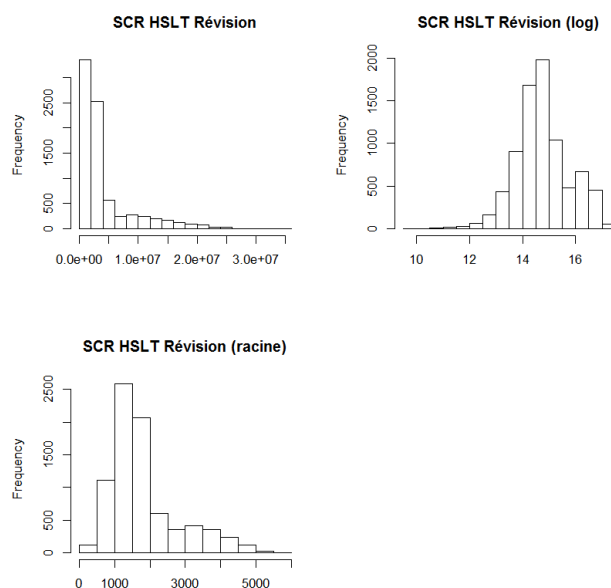


Figure 17: Histogrammes du SCR révision (HSLT)

Le QQ-plot de la variable transformée par le logarithme valide l'hypothèse de normalité. Nous avons dans un premier temps voulu estimer un GLM à l'aide d'une distribution Gamma mais encore une fois, l'algorithme d'estimation des coefficients n'a pas convergé.

### Commentaire sur les variables sélectionnées et analyse des modèles d'estimation :

Les variables sélectionnées par les différentes méthodes d'estimation sont résumées dans le tableau 12 (cf. Annexe). Le montant de provisions *Best Estimate* en santé similaire à la vie nous indique quel est l'engagement de l'assureur et à quel point le choc de révision impacte ses fonds propres. Le risque de révision en santé similaire à la vie affecte la garantie d'invalidité, il est donc cohérent que les variables relatives à l'arrêt de travail soient sélectionnées.

Lors de l'analyse des six premiers modèles linéaires généralisés, nous pouvons constater une forme arrondie du nuage de points et des résidus en fonction des valeurs ajustées. Pour cette raison, nous construisons des régressions quadratiques avec les variables sélectionnées par l'ACP, *Random Forest* et le GBM. Cela permet de faire disparaître les structures observées précédemment, et par conséquent d'améliorer l'ajustement et les erreurs absolues moyennes. Nous pouvons cependant noter une dispersion assez importante du nuage de points lors du tracé des valeurs ajustées en fonction des observations, et ce malgré la réduction de la variable d'intérêt. Cela signifie qu'il manque peut-être une variable explicative permettant d'améliorer l'estimation dans notre base d'étude.

### Choix de la méthode d'estimation :

Les résultats des modèles construits sont synthétisés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	18,2%	22,1%
Modèle complet + AIC	18,0%	19,4%
Modèle complet + BIC	18,2%	18,6%
Modèle ACP (quadratique)	18,3%	18,8%
Modèle ACP + AIC (quadratique)	10,0%	12,2%
Modèle ACP + BIC (quadratique)	10,2%	10,9%
Random Forest	7,2%	9,0%
<b>GBM</b>	<b>6,1%</b>	<b>7,2%</b>
Régression Ridge	17,6%	18,8%
Régression LASSO	18,3%	18,4%
Elastic Net	17,8%	18,7%
GLM avec variables issues de RF (quadratique)	10,2%	10,7%
GLM avec variables issues de GBM (quadratique)	18,7%	18,9%
GLM Final (quadratique)	13,5%	13,4%

Tableau 24: Détail des résultats de l'estimation du SCR révision (HSLT)

Nous retenons le GBM comme meilleur modèle.

### Analyse de sensibilités :

A partir du modèle final, nous avons effectué plusieurs sensibilités, en fonction des variables d'importance significative selon *Random Forest* et le GBM :

- Ajout de la durée du passif,
- Ajout du chiffre d'affaires,
- Ajout du BE Life,
- Ajout de la dispersion des âges en rente de conjoint,
- Ajout de la dispersion des âges en rente éducation.

L'ajout des trois premières variables n'améliore pas ou peu la qualité des prédictions. Les dispersions des âges en rente de conjoint et rente éducation ont le même effet, c'est-à-dire :

- Erreur absolue moyenne sur la base d'apprentissage = 12,4%
- Erreur absolue moyenne sur la base de test = 12,2%

Cette sensibilité nous indique que la dispersion des âges dans les portefeuilles de rente a un impact sur l'estimation du SCR révision en raison des corrélations existantes entre les portefeuilles d'arrêt de travail, de rente de conjoint et de rente éducation.

### Estimation d'un modèle à l'aide de variables sélectionnées par avis d'expert :

Nous avons estimé un GLM ne tenant compte que du BE SLT et du taux de frais en arrêt de travail. Les résultats obtenus sont les suivants :

Variable	Coefficient
Taux de frais (AT)	0,08
BE SLT	0,93

Tableau 25: coefficients du GLM d'estimation du SCR révision (HSLT)

Bien que très simple, ce modèle nous permet d'obtenir des erreurs absolues moyennes de l'ordre de 21%.

## SCR Santé similaire à la vie

### Analyse en composantes principales :

L'analyse en composantes principales nous indique que les variables liées au SCR Health SLT sont les suivantes :

- Les montants de provisions en santé (similaire et non similaire à la vie),
- Les durations moyennes, les anciennetés moyennes et leurs dispersions en incapacité et invalidité,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- L'âge moyen du portefeuille d'arrêt de travail,
- La proportion du chiffre d'affaires dédiée à la santé,
- Les dispersions des âges dans les portefeuilles de rente de conjoint et rente éducation,
- Les capitaux sous risques pour le risque catastrophe en vie.

Les dispersions des âges en rente de conjoint et rente éducation, la proportion du chiffre d'affaires afférente à la santé et les capitaux sous risques sont corrélés respectivement aux indicateurs du portefeuille et au chiffre d'affaires en arrêt de travail, expliquant pourquoi ces variables ont été retenues par l'ACP.

### Choix de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

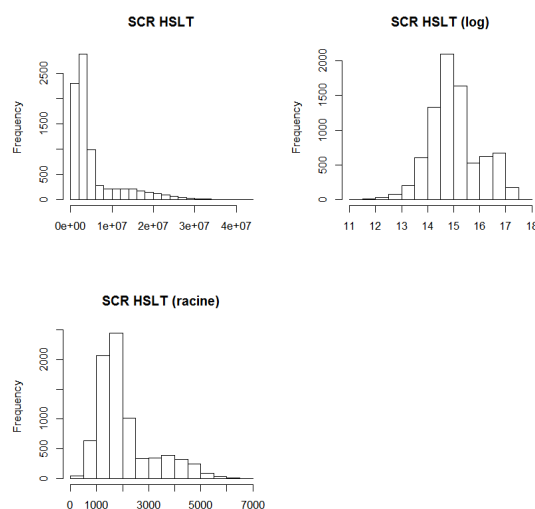


Figure 18: histogrammes du SCR santé similaire à la vie

L'histogramme de la variable après transformation par le logarithme est d'allure relativement gaussienne. Le QQ-plot de la variable transformée nous permet d'accepter l'hypothèse de normalité.

### Analyse des résultats et des modèles :

Le tableau 13 présenté en annexe résume les sélections de variables par les modèles d'estimation proposés.

Le risque de souscription en santé similaire à la vie ne concerne que la garantie d'invalidité. La sélection d'indicateurs relatifs à ce risque et plus globalement à l'arrêt de travail est ainsi cohérente. Nous pouvons également voir que le montant de provisions *Best Estimate* en santé similaire à la vie est systématiquement retenu. Cette variable traduisant le risque supporté par l'assureur quant à l'invalidité, sa sélection est logique tout comme celle du taux de frais qui a un impact sur les provisions de la garantie.

Lorsque nous traçons les graphes des résidus en fonction des valeurs ajustées pour les premiers modèles linéaires, nous observons une forme du nuage de points nous incitant à construire des modèles polynômiaux. Les procédures de sélection de variables étant longues, cela est fait pour les modèles construits grâce aux variables sélectionnées par GBM et *Random Forest*. De même, pour les modèles linéaires élaborés avec les variables présélectionnées par l'ACP, la structure du graphe des résidus nous suggère de construire un modèle quadratique. Les erreurs absolues détaillées ci-après nous prouvent que l'ajout d'effets croisés améliore l'ajustement du modèle.

### Choix de la méthode d'estimation :

Les détails des estimations sont résumés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	26,2%	29,7%
Modèle complet + AIC	26,0%	27,8%
Modèle complet + BIC	26,2%	27,3%
Modèle ACP (quadratique)	17,2%	19,6%
Modèle ACP + AIC (quadratique)	17,2%	19,2%
Modèle ACP + BIC (quadratique)	17,2%	18,6%
Random Forest	10,2%	11,8%
<b>GBM</b>	<b>8,0%</b>	<b>9,6%</b>
Régression Ridge	25,7%	27,2%
Régression LASSO	26,3%	27,3%
Elastic Net	25,6%	27,5%
GLM avec variables issues de RF (quadratique)	17,5%	16,9%
GLM avec variables issues de GBM (quadratique)	27,3%	25,0%
GLM Final (quadratique)	21,6%	19,1%

*Tableau 26: détail des résultats des estimations du SCR santé similaire à la vie*

Les meilleurs résultats sont obtenus grâce au GBM.

### Sensibilités :

L'analyse des variables retenues grâce à cette méthode nous indique que le ratio entre les provisions vie et non-vie en arrêt de travail et la durée du passif peuvent a priori améliorer l'adéquation des modèles aux données. Nous avons donc tenté de rajouter ces deux variables dans le dernier GLM construit, et dans chacun des cas nous obtenons :

- Erreur absolue moyenne sur la base d'apprentissage = 21 %
- Erreur absolue moyenne sur la base de test = 18,7 %

L'amélioration de la capacité de prévision du modèle confirme donc cette hypothèse. En revanche, les erreurs absolues moyennes sur la base de test obtenues ne sont pas plus satisfaisantes que celle du modèle construit avec les variables sélectionnées par *Random Forest*. Nous testons alors l'ajout de la dispersion de l'âge du portefeuille de rente éducation. Ainsi, nous avons :

- Erreur absolue moyenne sur l'échantillon d'apprentissage = 17,3 %
- Erreur absolue moyenne sur l'échantillon de test = 16,5 %

Les résultats nous indiquent que la dispersion de l'âge du portefeuille de rente éducation joue donc un rôle dans l'estimation du SCR santé similaire à la vie, certainement en raison de la corrélation observée entre cette variable et la durée moyenne en invalidité (corrélation de -32,4 %). Nous avons donc rajouté cette dernière variable au dernier GLM estimé. Nous avons aussi testé l'ajout de l'ancienneté moyenne en invalidité et de l'âge moyen du portefeuille d'arrêt de travail. Dans chacun des cas, nous avons obtenu des résultats similaires, à savoir :

- Erreur absolue moyenne sur la base d'apprentissage = 20,1 %
- Erreur absolue moyenne sur la base de test = 19 %

Comme cela pouvait être attendu, l'ajout d'indicateurs relatifs à l'arrêt de travail et à l'invalidité permet d'améliorer l'adéquation du modèle aux données. Cependant, l'amélioration n'est pas aussi importante qu'avec l'ajout de la dispersion des âges du portefeuille de rente éducation, certainement car celle-ci est corrélée à plusieurs variables afférentes à l'arrêt de travail telles que la dispersion des âges du portefeuille, le ratio entre les provisions vie et non-vie et les durées moyennes en incapacité et invalidité. Toutes ces corrélations sont négatives, expliquant donc pourquoi le coefficient associé à la dispersion des âges du portefeuille de rente éducation est négatif dans le modèle linéaire qui l'intègre.

Nous aurions pu tenter d'ajouter des effets croisés entre toutes les variables en rapport avec l'invalidité dans le dernier GLM construit, mais cela aurait sans doute complexifié le modèle sans pour autant atteindre les performances du GBM, que nous retiendrons comme meilleure méthode d'estimation du SCR santé similaire à la vie.

#### Estimation d'un modèle linéaire avec les variables sélectionnées par avis d'expert :

Nous avons également voulu construire un modèle facilement interprétable à l'aide de variables sélectionnées par avis d'expert. Nous obtenons ainsi :

Variable	Coefficient
Taux de frais (AT)	0,04
BE SLT	0,86
Duration moyenne en invalidité	0,03
Ancienneté moyenne en invalidité	0,06
Dispersion de l'ancienneté en invalidité	0,01
Ratio MGDC/AT	-0,01

Tableau 27: coefficients du GLM d'estimation du SCR HSLT

Le montant de provisions en santé similaire à la vie est prépondérant pour l'estimation du SCR Health SLT et cela se retrouve à travers le coefficient associé. Les indicateurs sur l'invalidité ont également tous des coefficients positifs. Cela est pertinent car l'invalidité est une garantie de long terme, et les risques du module Health SLT le sont également (longévité, révision et frais). Ces variables nous

fournissent des informations quant à l’horizon et à l’intensité des différents chocs. Comme nous l’avons vu précédemment, le ratio entre les provisions vie et non-vie en arrêt de travail permet de compenser les chocs au travers du maintien de la garantie décès.

Les taux d’erreurs obtenus sont de l’ordre de 27%. Cela est assez performant pour un modèle aussi simple en comparaison avec d’autres méthodes, mais ne permet pas de surpasser les performances du GBM.

Comparaison de l’estimation directe et de l’agrégation des meilleures méthodes d’estimation des sous-modules :

Enfin, nous avons voulu comparer l’estimation directe du SCR santé similaire à la vie par GBM et l’agrégation des prévisions des SCR sous-jacents. L’agrégation nous fournit des résultats satisfaisants, bien que ceux-ci soient moins convaincants que l’estimation directe :

	Estimation directe	Agrégation des sous-modules
MAE (apprentissage)	8 %	13 %
MAE (test)	9,6 %	14,5 %

*Tableau 28: comparaison de l’agrégation et de l’estimation directe pour le SCR santé similaire à la vie*

**b. Santé non similaire à la vie (Health NSLT)**

Dans le modèle utilisé pour construire notre base de données, aucun choc de rachat n’est appliqué en santé non similaire à la vie. Ainsi, le SCR HNSLT est composé uniquement du risque de primes et réserves.

Analyse en composantes principales :

L’analyse en composantes principales nous indique que les variables liées au SCR HNSLT sont les suivantes :

- Les montants de provisions *best estimate* en santé,
- Les durations moyennes en incapacité et en invalidité,
- Les dispersions des âges en rente de conjoint et rente éducation,
- La proportion du chiffre d’affaires dédiée à la santé,
- Les capitaux sous risques pour le risque catastrophe en vie.

Les dispersions des âges en rente de conjoint et rente éducation et les montants sous risques pour le risque catastrophe en vie présentent des corrélations respectives de 50,83%, 50,85% et -63,93% avec la proportion du chiffre d’affaires afférente à la santé. Ces corrélations peuvent en partie expliquer pourquoi ces variables relatives au SCR Life ont été retenues.

## Étude de la distribution pour l'estimation par GLM :

L'analyse des histogrammes de la variable d'intérêt nous suggère d'utiliser une transformation par le logarithme, car elle donne lieu à une distribution plus symétrique :

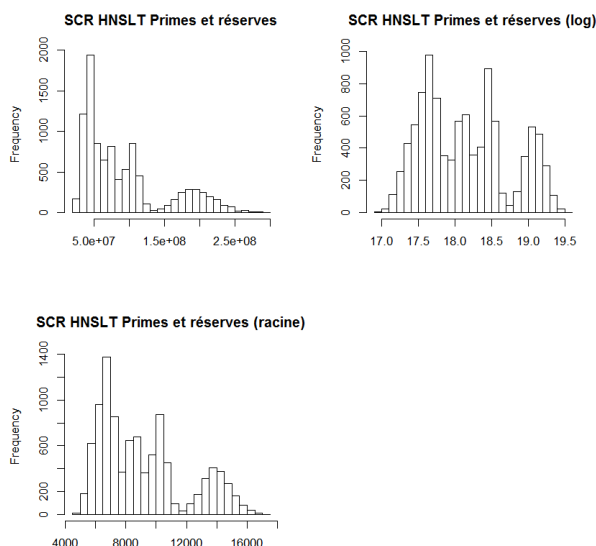


Figure 19: histogrammes du SCR HNSLT

Le QQ-plot de la variable ainsi transformée nous permet d'accepter l'hypothèse de normalité malgré des distorsions observées au niveau des queues de distribution.

### Analyse des résultats :

La liste des variables sélectionnées pour l'estimation du SCR HNSLT est présentée dans le tableau 14 en annexe.

Les coefficients d'évolution du chiffre d'affaires en santé et arrêt de travail et la répartition du chiffre d'affaires sont par construction directement liés au niveau de cotisations reçu par l'assureur, et par conséquent leur impact sur le SCR HNSLT est immédiat. De même, les taux de réassurance en santé et en arrêt de travail (pour les garanties non-vie) ont un effet sur le volume de primes, car les traités de réassurance incluent une cession des cotisations de l'organisme assureur vers la cédante. Ces traités vont donc venir diminuer le volume de primes, mais également le risque de l'assureur et les réserves qu'il doit constituer.

Le ratio entre les provisions afférentes aux garanties vie et celles dédiées à la non-vie en arrêt de travail a également du sens puisqu'il indique dans quelle mesure l'engagement de l'assureur pour les risques non SLT est important. La durée moyenne en incapacité témoigne du risque pris par l'assureur en santé non-similaire à la vie, mais également de la durée restante jusqu'à la fin de cette garantie.

Choix de la méthode d'estimation :

Les résultats des modèles d'estimation sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	6,7%	7,6%
Modèle complet + AIC	6,7%	7,3%
Modèle complet + BIC	6,7%	6,9%
Modèle ACP	7,9%	8,0%
Modèle ACP + AIC	7,9%	8,0%
Modèle ACP + BIC	7,9%	8,0%
Random Forest	2,8%	4,4%
<b>GBM</b>	<b>1,8%</b>	<b>3,6%</b>
Régression Ridge	5,8%	8,5%
Régression LASSO	6,9%	8,1%
Elastic Net	3,1%	8,1%
GLM avec variables issues de RF	8,8%	8,9%
GLM avec variables issues de GBM	7,3%	7,3%
GLM Final	6,9%	6,8%

Tableau 29 : Détail des résultats de l'estimation du SCR HNSLT

Nous retenons le GBM comme le meilleur modèle.

Estimation d'un modèle linéaire avec des indicateurs retenus par avis d'expert :

Nous avons estimé un GLM à l'aide des variables suivantes :

- Coefficients d'évolution du chiffre d'affaires en santé et arrêt de travail,
- Taux de réassurance en santé et en arrêt de travail (non-vie),
- BE NSLT,
- Duration moyenne en incapacité,
- Duration du passif,
- Proportions du chiffre d'affaires dédiées à la santé et à l'arrêt de travail.

Nous obtenons les coefficients suivants :

Variable	Coefficient
<b>Coefficient d'évolution du chiffre d'affaires (Santé)</b>	0,07
<b>Coefficient d'évolution du chiffre d'affaires (AT)</b>	0,08
<b>BE NSLT</b>	0,87
<b>Taux de réassurance (Santé)</b>	-0,02
<b>Taux de réassurance (AT – non vie)</b>	-0,02
<b>Duration moyenne en incapacité</b>	0,20
<b>Duration du passif</b>	0,07
<b>Proportion du chiffre d'affaires (Santé)</b>	0,74
<b>Proportion du chiffre d'affaires (AT)</b>	0,05

Tableau 30: Coefficients du GLM d'estimation du SCR HNSLT

Les coefficients d'évolution et les proportions du chiffre d'affaires sont révélateurs de l'exposition de l'assureur à des risques de type NSLT, mais également de son engagement et donc du niveau de primes qu'il reçoit et du niveau de réserves qu'il doit constituer. Les taux de réassurance permettent

quant à eux de céder le risque, mais également les cotisations. Ils agissent donc en sens inverse et cela se retrouve à travers les coefficients négatifs qui leur sont associés. Les deux durations nous permettent également de connaître l'horizon de l'engagement de l'assureur et donc la sensibilité des réserves de celui-ci au risque.

S'il a l'avantage d'être facilement interprétable, ce modèle ne permet pas d'atteindre les performances du GBM car les taux d'erreurs obtenus sont proches de 8%.

### **c. Risque catastrophe en santé (Health CAT)**

#### **SCR Concentration**

##### Analyse en composantes principales :

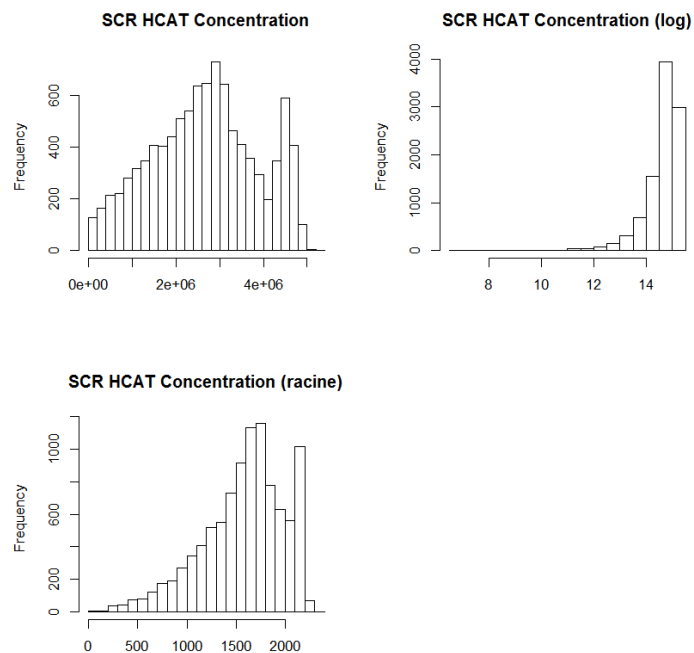
L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- L'exposition pour soins médicaux dans le cas du risque de concentration,
- L'exposition pour invalidité permanente dans le cas du risque de concentration,
- La dispersion de l'ancienneté et l'ancienneté moyenne en incapacité,
- L'âge moyen du portefeuille de rente de conjoint,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- Le taux de frais pour les garanties décès,
- La proportion du chiffre d'affaires dédiée à l'arrêt de travail,
- La durée de l'actif.

Les variables sélectionnées par l'ACP présentent toute une corrélation de l'ordre de 20% avec les variables relatives au scénario de concentration. Ces corrélations peuvent expliquer le résultat de l'ACP.

## Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :



*Figure 20: histogrammes du SCR concentration (risque catastrophe en santé)*

La variable brute présente une distribution globalement symétrique. L'analyse du QQ-plot nous permet d'accepter l'hypothèse de normalité.

### Analyse des indicateurs retenus et des modèles construits :

Les variables sélectionnées pour l'estimation du SCR concentration sont présentées dans le tableau 15 en annexe.

Les expositions pour invalidité permanente et incapacité d'au plus 12 mois sont les variables le plus retenues pour l'estimation du SCR. Le calcul du SCR concentration étant directement lié à ces expositions, leur sélection est logique.

En revanche, les expositions pour invalidité d'au plus 10 ans, pour décès accidentel et pour soins médicaux n'ont pas été retenues par la majorité des modèles alors qu'elles sont pourtant nécessaires au calcul du SCR concentration pour le risque catastrophe en santé.

L'estimation du SCR concentration est difficile, notamment en raison d'une forte dispersion du nuage de points que nous observons lorsque nous traçons les valeurs ajustées en fonction des observations, malgré la réduction de la variable d'intérêt. Cette dispersion ne disparaît pas avec l'ajout d'effets croisés, qui sont systématiquement évalués comme non-significatifs et supprimés par les procédures de sélection de variables AIC ou BIC. Les graphes des résidus en fonction des valeurs ajustées, les distances de Cook et QQ-plots sont pourtant satisfaisants pour chacun des GLM étudiés.

Les trois derniers modèles sélectionnent les mêmes variables et ont donc des résultats identiques.

### Choix de la méthode d'estimation :

Les résultats des différents modèles sont présentés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	99,8%	108,0%
Modèle complet + AIC	99,1%	101,3%
Modèle complet + BIC	99,5%	100,0%
Modèle ACP	99,7%	100,9%
Modèle ACP + AIC/BIC	99,6%	100,2%
Random Forest	99,7%	99,9%
<b>GBM</b>	<b>50,4%</b>	<b>71,4%</b>
Régression Ridge	99,5%	100,0%
Régression LASSO	99,0%	99,9%
Elastic Net	99,7%	100,4%
GLM avec variables issues de RF	99,6%	99,4%
GLM avec variables issues de GBM (quadratique)	99,6%	99,4%
GLM Final	99,6%	99,4%

*Tableau 31: détail des résultats de l'estimation du SCR concentration (catastrophe en santé)*

Nous retiendrons le GBM comme « meilleur » modèle.

### **SCR accident de masse**

#### Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir :

- Les expositions pour invalidité d'au plus 10 ans, incapacité d'au plus 12 mois et invalidité permanente (scénario accident de masse),
- Les expositions pour invalidité d'au plus 10 ans, incapacité d'au plus 12 mois et invalidité permanente (scénario concentration),
- Les expositions pour consultation, soins médicaux et hospitalisation (scénario pandémie),
- Le coefficient du chiffre d'affaires en santé.

La sélection de variables relatives au scénario accident de masse est logique. De plus, comme ces variables sont corrélées aux autres indicateurs relatifs au risque catastrophe en santé, il est cohérent que ces derniers soient également retenus. De même, la proportion du chiffre d'affaires afférente à la santé étant le point de départ des simulations, celle-ci est corrélée avec les variables permettant le calcul des SCR catastrophe en santé et est également retenue par l'ACP.

## Étude de la distribution pour l'analyse par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

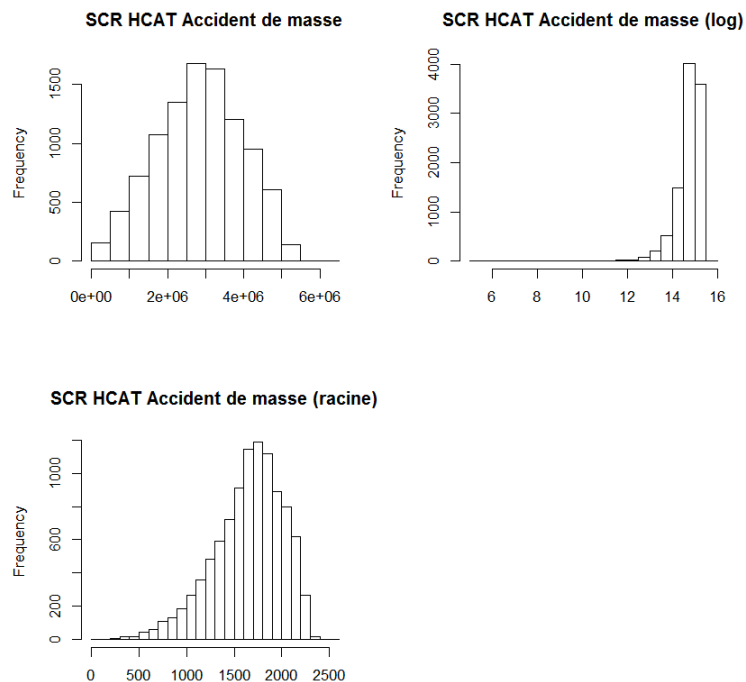


Figure 21: histogrammes du SCR Accident de masse (risque catastrophe en santé)

La variable brute présente une allure gaussienne. Cette hypothèse est confirmée par le QQ-plot et l'hypothèse de normalité peut donc être retenue.

### Analyse des prédicteurs sélectionnés et de la qualité des modèles :

Le tableau 16 de l'annexe résume les différentes variables sélectionnées pour l'estimation du SCR accident de masse.

Comme attendu, les expositions pour le scénario accident de masse sont les variables les plus sélectionnées. Toutefois, certains indicateurs nécessaires à l'évaluation du SCR accident de masse tels que les expositions pour décès accidentel et pour soins médicaux n'ont pas été retenus.

De manière tout à fait similaire au SCR concentration, la forte dispersion du nuage de points que nous observons lorsque nous traçons les valeurs ajustées en fonction des observations empêche l'ajustement. Cette dispersion ne disparaît pas avec l'ajout d'effets croisés, qui sont systématiquement évalués comme non-significatifs et supprimés par les procédures de sélection de variables AIC ou BIC. Les hypothèses sous-jacentes à l'utilisation de GLM sont pourtant respectées. Nous avons également testé l'ajout d'autres variables d'exposition nécessaires au calcul du SCR d'intérêt, testé l'adéquation à d'autres lois de la famille exponentielles ou d'autres transformations sur la variable de sortie, mais rien n'a permis d'améliorer les modèles construits.

### Choix de la méthode d'estimation :

Les résultats des estimations sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	99,1%	107,4%
Modèle complet + AIC	98,6%	101,5%
Modèle complet + BIC	99,2%	99,3%
Modèle ACP	99,3%	100,1%
Modèle ACP + AIC	99,2%	99,5%
Modèle ACP + BIC	99,3%	99,1%
Random Forest	99,2%	99,5%
<b>GBM</b>	<b>96,3%</b>	<b>97,3%</b>
Régression Ridge	98,9%	99,8%
Régression LASSO	99,0%	99,2%
Elastic Net	99,3%	99,6%
GLM avec variables issues de RF	99,3%	99,1%
GLM avec variables issues de GBM	99,3%	98,8%
GLM Final	99,3%	98,5%

*Tableau 32: détail des résultats de l'estimation du SCR accident de masse*

Même si les résultats ne sont pas satisfaisants, nous retiendrons le GBM comme meilleure méthode d'estimation du SCR accident de masse.

### **SCR Pandémie**

#### Analyse en composantes principales :

L'analyse en composantes principales nous indique que les variables liées au SCR pandémie sont les suivantes :

- Les expositions pour consultation, hospitalisation et soins médicaux (scénario pandémie),
- Le nombre de cotisants (scénario pandémie),
- La capacité d'absorption des chocs par la FDB (scénario pandémie),
- L'exposition pour soins médicaux (scénario concentration),
- L'exposition pour invalidité permanente (scénario accident de masse),
- Le nombre de bénéficiaires pour incapacité d'au plus 12 mois,
- Les taux de frais et de PANE en santé,
- La durée de l'actif.

La sélection de variables relatives au scénario pandémie est naturelle. De par les corrélations existantes entre les variables afférentes au risque catastrophe en santé, il est logique que certaines expositions pour le calcul des SCR concentration et accident de masse aient également été retenues. De manière analogue, le taux de frais et le taux de PANE en santé sont corrélés avec les variables liées au risque catastrophe en santé, d'où leur sélection par l'ACP.

## Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable à expliquer sont les suivants :

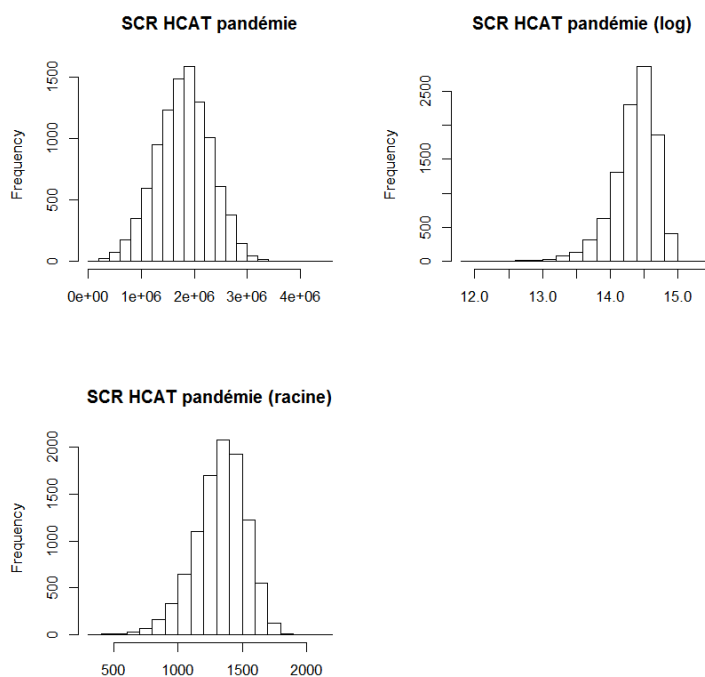


Figure 22: histogrammes du SCR pandémie (risque catastrophe en santé)

La variable brute présente un histogramme gaussien. Le QQ-plot nous permet d'accepter l'hypothèse de normalité.

### Analyse des résultats et choix de la méthode d'estimation :

Le tableau 17 (cf. Annexe) résume les différentes sélections de variables lors de l'estimation du SCR pandémie.

De manière cohérente, les variables les plus retenues sont celles relatives au scénario pandémie, à l'exception du nombre de cotisants. Nous pouvons également noter que des variables en rapport avec le portefeuille d'actifs de l'assureur ont été sélectionnées par plusieurs modèles.

Comme pour les SCR précédents, la forte dispersion des données limite grandement l'adéquation. L'ajout d'effets croisés, l'ajout de variables relatives au risque catastrophe en santé, les transformations sur la variable d'intérêt ou la modification de la loi sous-jacente n'ont pas d'effet sur l'ajustement et n'améliorent pas les modèles proposés. De plus, les graphes des résidus (en fonction des valeurs ajustées, distance de Cook ou QQ-plot) ne sont satisfaisants que lorsque nous considérons une loi normale.

Les résultats des différentes estimations sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	98,5%	108,7%
Modèle complet + AIC	98,0%	103,0%
Modèle complet + BIC	98,8%	99,2%
Modèle ACP	98,8%	99,8%
Modèle ACP + AIC	98,8%	99,2%
Modèle ACP + BIC	98,8%	99,2%
Random Forest	93,3%	99,3%
<b>GBM</b>	<b>88,1%</b>	<b>98,8%</b>
Régression Ridge	99,0%	99,5%
Régression LASSO	98,6%	99,0%
Elastic Net	99,1%	100,2%
GLM avec variables issues de RF	98,9%	99,4%
GLM avec variables issues de GBM	98,9%	99,4%
GLM Final	98,7%	99,4%

*Tableau 33: détail des résultats des estimations du SCR Pandémie*

Nous retiendrons le GBM comme meilleur modèle, bien que les résultats obtenus par cette méthode soient très peu convaincants.

### **SCR catastrophe en santé**

Analyse en composantes principales :

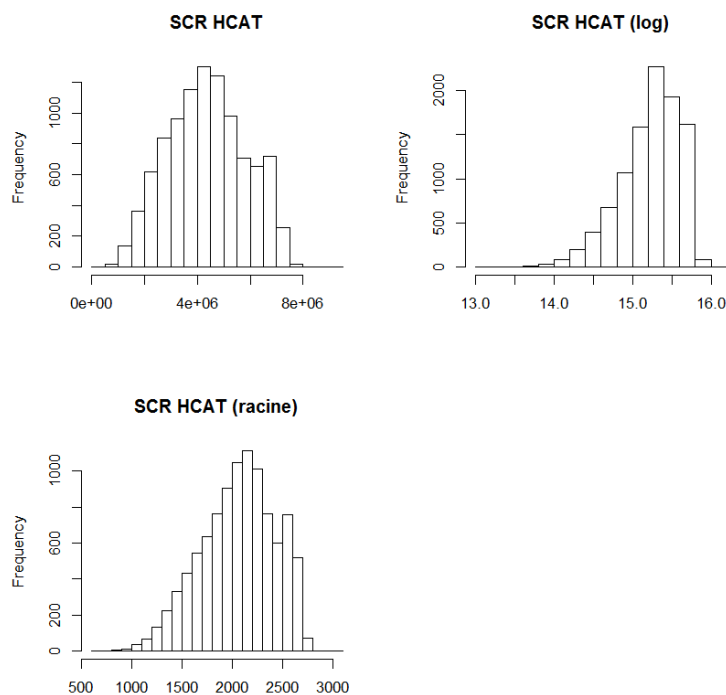
L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- L'exposition pour invalidité d'au plus 10 ans (scénario accident de masse),
- Les expositions pour incapacité d'au plus 12 mois, invalidité d'au plus 10 ans et invalidité permanente (scénario concentration),
- Le nombre de cotisants et l'exposition pour consultation (scénario pandémie),
- Le coefficient d'évolution du chiffre d'affaires et le taux de PANE en arrêt de travail.

Le SCR catastrophe en santé étant l'agrégation des trois précédents SCR, il est cohérent que les variables en rapport avec les scénarios concentration, accident de masse et pandémie soient sélectionnés par l'ACP. Les autres variables ont certainement été retenues en raison de corrélations avec les variables afférentes au risque catastrophe.

## Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable à prédire sont les suivants :



*Figure 23 : histogrammes du SCR catastrophe en santé*

L'histogramme de la variable brute et le QQ-plot de celle-ci nous permettent d'accepter l'hypothèse de normalité.

### Analyse des résultats et choix de la méthode d'estimation :

Les variables retenues selon les différentes méthodes d'estimation du SCR catastrophe en santé sont présentées en annexe (tableau 18).

La plupart des variables en rapport avec le risque catastrophe ont été sélectionnées au moins une fois. Toutefois, seules les expositions pour invalidité permanente et invalidité d'au plus 10 ans dans les scénarios concentration et accident de masse ont été retenues par la plupart des modèles.

Nous sommes encore confrontés à une forte dispersion du nuage de points lors des tracés de droite de régression, même si les données ont été réduites. De plus, les graphes des résidus et les QQ-plots sont satisfaisants, validant ainsi l'hypothèse de la loi sous-jacente.

Les transformations sur la variable à prédire, le changement de loi sous-jacente, l'ajout d'interactions et de variables relatives au risque catastrophe ne permettent pas d'améliorer les modèles construits. En outre, les procédures de sélection de variables par le critère BIC conduisent systématiquement à un modèle vide.

Les résultats des estimations sont présentés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	98,5%	107,5%
Modèle complet + AIC	99,3%	101,2%
Modèle ACP	99,8%	100,1%
Modèle ACP + AIC	99,8%	99,8%
Random Forest	99,8%	100,1%
<b>GBM</b>	<b>93,7%</b>	<b>95,1%</b>
Régression Ridge	99,6%	100,0%
Régression LASSO	99,7%	100,0%
Elastic Net	99,4%	102,2%
GLM Final	99,7%	100,4%

Tableau 34: détail des résultats de l'estimation du SCR catastrophe en santé

Nous retiendrons toutefois le GBM comme meilleur modèle d'estimation du SCR catastrophe en santé.

L'agrégation des SCR pandémie, accident de masse et concentration, au vu des erreurs d'estimation présentées dans les parties précédentes, ne donne pas de meilleur résultat.

Expression du SCR CAT en fonction des SCR Santé :

Après avoir analysé les données de clients, nous avons constaté que les valeurs des expositions, des montants sous risques sont très différents d'un organisme à l'autre, y compris pour des organismes de taille similaire. Nous avons donc estimé le SCR catastrophe en santé en fonction des SCR HSLT et HNSLT. Nous obtenons alors :

Variable	Coefficient
<b>SCR HSLT</b>	-0,04
<b>SCR HNSLT</b>	0,06

Tableau 35 : coefficients du GLM d'estimation du SCR catastrophe en santé

Cependant, ce modèle ne nous permet pas de mieux estimer le SCR catastrophe en santé.

## d. SCR Santé

Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Les montants de provisions *Best Estimate* en santé (similaire et non-similaire à la vie),
- Les durations moyennes en incapacité et en invalidité,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- Les dispersions des âges dans les portefeuilles de rente de conjoint et rente éducation,
- La répartition du chiffre d'affaires selon les différentes garanties.

Comme nous l'avons déjà vu, les indicateurs construits sur les portefeuilles de rente de conjoint et de rente éducation sont corrélés à ceux relatifs au portefeuille d'arrêt de travail. Pour cette raison, les dispersions des âges de ces portefeuilles ont été retenues par l'ACP.

Étude de la distribution pour l'estimation par GLM :

L'analyse des histogrammes de la variable d'intérêt nous suggère de retenir une transformation logarithmique :

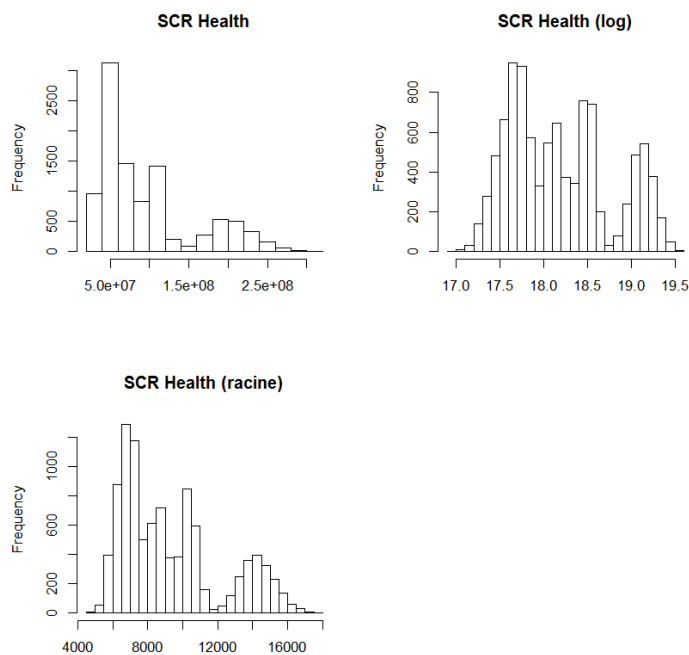


Figure 24: histogrammes du SCR santé

Le QQ-plot de la variable ainsi transformée nous permet d'accepter l'hypothèse de normalité malgré des distorsions observées au niveau des queues de distribution.

Commentaire sur les indicateurs retenus :

Les variables retenues selon les différents modèles d'estimation sont présentées dans le tableau 19 en annexe.

La sélection des montants de provisions en santé, traduisant la hauteur de l'engagement de l'assureur, est naturelle. L'invalidité étant le seul risque de type santé similaire à la vie du modèle, la sélection de variables en rapport avec celle-ci est logique. De façon analogue, l'incapacité entre en jeu dans le calcul des risques de type NSLT et doit donc être prise en compte dans le calcul du SCR santé. Enfin, les proportions du chiffre d'affaires dédiées à la santé et à l'arrêt de travail nous indiquent également le risque supporté par l'assureur. La dispersion de l'âge en rente éducation est corrélée négativement avec plusieurs indicateurs construits sur le portefeuille d'arrêt de travail, expliquant pourquoi cette variable est si souvent retenue. Cette corrélation se retrouve dans le coefficient négatif qui lui est associé. Enfin, les coefficients négatifs associés au montants de provisions Life et aux capitaux sous risques pour le risque catastrophe en vie sont pertinents. Nous avons déjà vu que les montants sous risques sont corrélés négativement à la proportion du chiffre d'affaires dédiée à la santé, qui joue un rôle dans l'estimation du présent SCR.

#### Choix de la méthode d'estimation :

Les résultats des estimations sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	6,6%	6,8%
Modèle complet + AIC	6,6%	6,5%
Modèle complet + BIC	6,6%	6,4%
Modèle ACP	22,8%	19,9%
Modèle ACP + AIC	22,8%	19,9%
Modèle ACP + BIC	22,8%	19,9%
Random Forest	2,9%	14,1%
<b>GBM</b>	<b>1,9%</b>	<b>5,8%</b>
Régression Ridge	5,5%	8,2%
Régression LASSO	6,8%	6,4%
Elastic Net	3,2%	7,8%
GLM avec variables issues de RF	6,7%	6,6%
GLM avec variables issues de GBM	7,6%	7,7%
GLM Final	6,7%	6,6%

*Tableau 36: détail des résultats de l'estimation du SCR santé*

Nous pouvons voir que la plupart des modèles d'estimation donnent des résultats assez convaincants, à l'exception des GLM construits avec les variables présélectionnées par l'ACP. La mauvaise adéquation du modèle peut être expliquée par une sélection de variables dans un plan factoriel n'expliquant qu'une faible partie de l'information disponible dans nos données.

Nous retiendrons le GBM comme meilleure méthode d'estimation du SCR santé, car il offre un bon compromis entre ajustement sur la base d'apprentissage et qualité de prévision sur l'échantillon de test. Nous pouvons aussi constater que la régression LASSO offre des résultats satisfaisants.

#### Comparaison de l'estimation directe et de l'agrégation des estimations des sous-modules :

Nous avons également tenté d'estimer le SCR Santé par agrégation des meilleurs modèles d'estimation des SCR santé NLST, santé SLT et santé CAT.

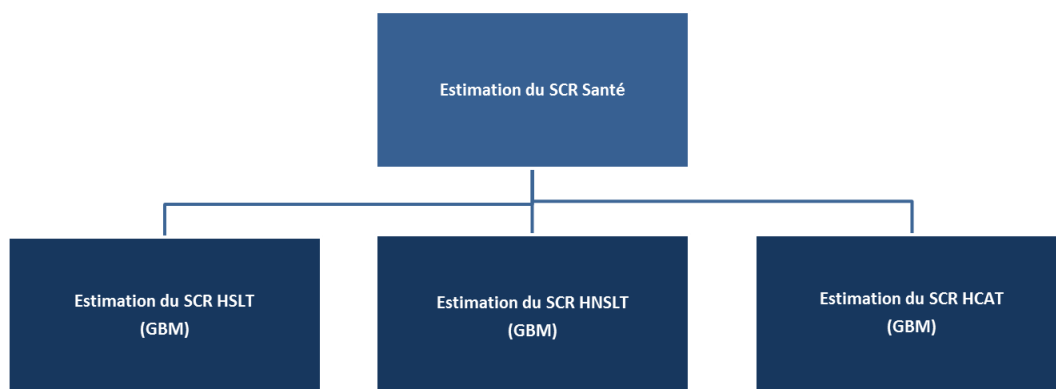


Figure 25: Agrégation des estimations des sous-modules du SCR santé

Les résultats de l'agrégation sont les suivants :

	Estimation directe	Agrégation des sous-modules
MAE (apprentissage)	1,9 %	69,7 %
MAE (test)	5,8 %	61,2 %

Tableau 37: comparaison de l'estimation directe et de l'agrégation des sous-modules pour le SCR santé

Les résultats obtenus par agrégation sont bien moins convaincants que l'ensemble des résultats pour l'estimation directe. L'erreur que nous observons pour l'estimation par agrégation est sûrement le résultat des mauvaises adéquations que nous avons observé sur le risque catastrophe en santé.

Estimation d'un modèle linéaire à l'aide d'indicateurs retenus par avis d'expert :

Nous obtenons le modèle suivant :

Variable	Coefficient
<b>Coefficient d'évolution du chiffre d'affaires (Santé)</b>	0,07
<b>Coefficient d'évolution du chiffre d'affaires (AT)</b>	0,07
<b>Chiffre d'affaires</b>	0,01
<b>Ratio MGDC/AT</b>	-0,13
<b>Duration moyenne en invalidité</b>	0,07
<b>BE SLT</b>	0,48
<b>BE NSLT</b>	0,57
<b>Duration du passif</b>	0,004
<b>Proportion du chiffre d'affaires (Santé)</b>	0,02
<b>Proportion du chiffre d'affaires (AT)</b>	0,03

Tableau 38 : coefficients du GLM d'estimation du SCR santé

Les durations nous indiquent l'intensité avec laquelle l'organisme assurantiel va subir les différents chocs du module et le ratio entre maintien de la garantie décès et arrêt de travail permet de compenser ces chocs à travers l'introduction d'un risque de catégorie *Life*.

Même s'il est simple, ce modèle présente des qualités d'ajustement et de prévisions satisfaisants car les taux d'erreurs obtenus sont de l'ordre de 9,5%.

## 1. SCR Marché

### a. SCR risque de taux d'intérêts

Analyse en composantes principales :

L'analyse en composantes principales nous suggère que les variables liées à la variable d'intérêt sont les suivantes :

- L'ancienneté moyenne en invalidité et sa dispersion,
- La dispersion des âges dans les portefeuilles d'arrêt de travail, rente éducation et rente de conjoint,
- La durée moyenne en incapacité,
- L'âge moyen en arrêt de travail,
- Les montants de provisions Best Estimate,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- La proportion du chiffre d'affaires afférente à la santé.

Étude de la distribution pour l'estimation par GLM :

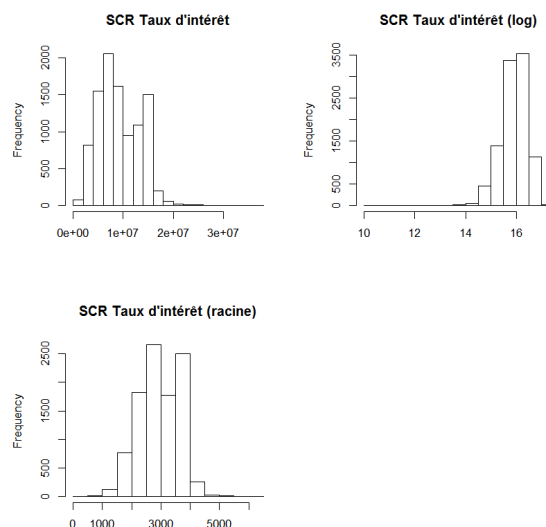


Figure 26: histogrammes du SCR risque de taux d'intérêts

L'analyse des histogrammes de la variable et le QQ-plot de celle-ci nous permettent de retenir l'hypothèse de normalité pour une transformation par la racine.

Analyse des indicateurs retenus et de la qualité des modèles :

Après avoir calibré les différents modèles d'estimation du SCR risque de taux d'intérêt (cf. Annexe, tableau 20), les variables les plus retenues sont :

- Les coefficients d'évolution du chiffre d'affaires en santé et rente de conjoint,

- Le ratio entre les provisions dédiées au maintien de la garantie décès et celles afférentes à l'arrêt de travail,
- Les durations moyennes en incapacité et invalidité,
- Les montants sous risques pour le SCR catastrophe en vie,
- Les montants de provisions *Best Estimate*,
- La durée moyenne des OATi.

Les montants de provisions représentent la valeur actualisée de l'engagement de l'assureur, elles sont donc sensibles à la courbe des taux utilisée pour l'actualisation et leur sélection est logique. L'invalidité et la rente de conjoint sont des garanties présentant une durée importante. Les provisions calculées sont donc très sensibles à la courbe des taux utilisée pour actualiser les flux futurs. Celle-ci impacte également les montants de provisions *Best Estimate*. Les OATi sont des instruments financiers sensibles au taux et leur sélection est cohérente.

Les erreurs de prévision sont importantes pour tous les modèles. Lors des tracés des graphes des résidus pour les GLM, nous pouvons cependant voir que les hypothèses sous-jacentes à l'utilisation de tels modèles sont respectées. Nous constatons également une structure dans la forme du nuage de points lorsque nous traçons le graphe des valeurs ajustées en fonction des observations. Pour cette raison, nous estimons des modèles polynômiaux à l'aide des variables retenues par GBM et *Random Forest*. Dans le cas du dernier modèle, l'ajout d'interactions améliore l'adéquation aux données.

#### Choix de la méthode d'estimation :

Les résultats des différentes estimations sont résumés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	32,6%	37,7%
Modèle complet + AIC	32,4%	35,1%
Modèle complet + BIC	32,8%	33,8%
Modèle ACP	34,9%	36,5%
Modèle ACP + AIC	34,9%	36,4%
Modèle ACP + BIC	34,9%	36,4%
Random Forest	14,8%	15,8%
<b>GBM</b>	<b>9,7%</b>	<b>11,5%</b>
Régression Ridge	46,2%	46,9%
Régression LASSO	32,7%	33,9%
Régression Elastic Net	30,1%	31,1%
GLM avec variables issues de RF (quadratique)	24,1%	26,3%
GLM avec variables issues de GBM (quadratique)	22,3%	25,2%
GLM Final (quadratique)	21,4%	24,6%

Tableau 39: détail des résultats de l'estimation du SCR risque de taux d'intérêts

Le GBM est la méthode d'estimation la plus performante.

#### Estimation d'un modèle linéaire avec les variables retenues par avis d'expert :

Enfin, nous avons voulu calibrer un GLM à l'aide des indicateurs suivants :

- Les montants de provisions Best Estimate,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- Les durations de l'actif et du passif,
- Le poids des obligations, des OPCVM et du monétaire dans le portefeuille de placements.

Nous obtenons les résultats suivants :

- MAE sur l'échantillon d'apprentissage = 34,6%
- MAE sur la base de test = 36,1%

Les coefficients associés à chacune des variables sont :

Variable	Coefficient associé
BE SLT	0,048
BE NSLT	0,035
BE Life	0,13
Duration de l'actif	-0,004
Duration du passif	-0,009
Poids du monétaire	0,023
Poids des OPCVM	0,035
Poids des obligations	0,026

Tableau 40 : coefficients du GLM d'estimation du SCR risque de taux d'intérêts

Les montants de provisions constituent les assiettes de passif utilisées pour le calcul du SCR. Plus celles-ci sont importantes, plus l'impact du choc de taux sera grand, accentuant ainsi le montant de SCR. Le monétaire, les OPCVM et les obligations sont sensibles aux taux d'intérêts et si la part qui leur est accordée au sein des placements de l'assureur est conséquente le SCR sera important.

## **b. SCR risque action**

Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Le poids des actions et des obligations convertibles dans le portefeuille de placements,
- Les durations moyennes des obligations,
- La durée de l'actif,
- La capacité d'absorption des chocs par le FDB pour le risque action.

La plupart des indicateurs retenus sont pertinents. En effet, l'estimation du SCR pour le risque action est calculé à partir d'un choc linéaire sur la valeur de marché des actions, leur sélection est par conséquent pertinente. De même, la capacité d'absorption par le FDB est logique car elle nous indique à quel point le choc sur les actions impacte les fonds propres de l'assureur. Les durations des obligations ont été retenues par l'ACP car elles présentent des corrélations de l'ordre de 15 % avec la part action.

Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

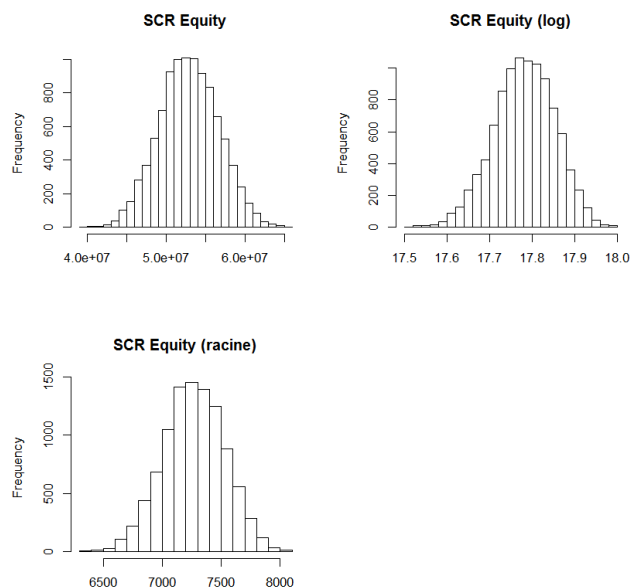


Figure 27: Histogrammes du SCR risque action

Tous les histogrammes présentent un histogramme d'allure gaussienne. Le QQ-plot de la variable brute nous permet d'accepter l'hypothèse de normalité pour la suite de la modélisation.

#### Analyse des résultats et des modèles :

Les variables sélectionnées par la différentes méthodes d'estimation du SCR risque action sont présentées dans le tableau 21 en annexe.

Les variables retenues par la majorité des modèles sont :

- La part des actions, des obligations, de l'immobilier et des OPCVM dans le portefeuille de placements de l'organisme assurantiel,
- La duration moyenne des OAT.

La sélection de la part action est cohérente. Les autres indicateurs sur la répartition du portefeuille de placements ont été sélectionnés en raison de corrélations (de l'ordre de 25%) avec la poche action.

Les premiers modèles linéaires estimés présentent des qualités d'ajustement peu convaincantes, notamment ceux construits à partir des variables présélectionnées par l'ACP. Comme ces variables ont été retenues à partir d'un plan factoriel n'expliquant qu'une faible part de l'information disponible dans notre jeu de données, la mauvaise qualité de ces modèles n'est pas surprenante. Nous constatons également une structure dans les graphes des résidus et des observations en fonction des valeurs ajustées, nous incitant à ajouter des effets croisés dans les modèles. Cela a été fait pour les modèles estimés à partir des variables indiquées comme importantes par le GBM et *random forest*. Nous pouvons voir sur le tableau ci-dessus que l'ajout d'interactions permet d'améliorer l'adéquation du modèle aux données et sa qualité de prévision. Comme nous ne disposons pas des détails des poches action de type 1 et 2, l'estimation par les régressions linéaires reste difficile.

### Choix de la méthode d'estimation :

Les résultats des estimations sont présentés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	40,9%	49,0%
Modèle complet + AIC	40,7%	42,7%
Modèle complet + BIC	41,0%	43,7%
Modèle ACP	99,5%	99,2%
Modèle ACP + AIC	99,5%	99,0%
Modèle ACP + BIC	99,5%	98,8%
Random Forest	20,4%	21,3%
<b>GBM</b>	<b>15,7%</b>	<b>17,7%</b>
Régression Ridge	47,8%	46,5%
Régression LASSO	46,3%	48,2%
Régression Elastic Net	47,2%	45,6%
GLM avec variables issues de RF (quadratique)	46,0%	46,7%
GLM avec variables issues de GBM (quadratique)	46,6%	47,8%
GLM Final (quadratique)	47,0%	48,2%

Tableau 41: détail des résultats de l'estimation du SCR risque action

Les meilleurs résultats sont obtenus grâce au GBM. Nous pouvons aussi constater que les résultats de la régression *Elastic Net* sont assez convaincants, nous indiquant que l'ajout d'effets croisés n'est pas nécessairement plus performant que la prise en compte des corrélations et la pénalisation.

### Sensibilités :

Le dernier GLM présente des erreurs absolues moyennes plus élevées que celui estimé avec les variables sélectionnées par *Random Forest*. Nous testons donc l'ajout du poids du monétaire dans la régression et nous obtenons alors :

- MAE sur la base d'apprentissage = 45,8 %
- MAE sur la base de test = 46,3 %

Cette variable, certainement en raison de sa corrélation avec le poids des actions, permet d'améliorer les capacités d'estimation et de prévision du modèle.

### **c. SCR risque immobilier**

#### Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Le poids de l'immobilier, des actions et des obligations dans le portefeuille de l'assureur,
- La durée moyenne des OTV.

## Étude de la distribution pour l'estimation par modèles linéaires :

L'analyse des histogrammes de la variable à prédire nous incite à choisir une transformation par le logarithme car l'histogramme obtenu est le plus symétrique :

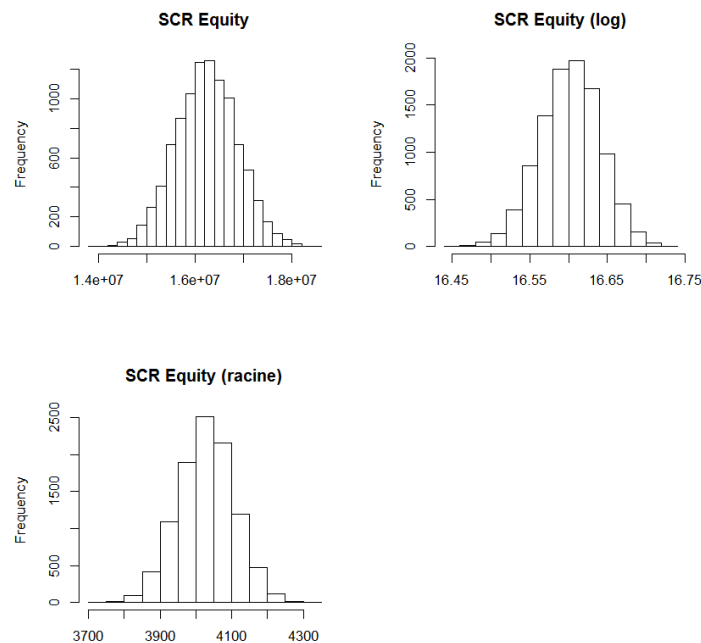


Figure 28 : histogrammes du SCR risque immobilier

Le QQ-plot de la variable brute nous permet d'accepter l'hypothèse de normalité de la variable pour la modélisation par modèles linéaires.

## Analyse des indicateurs retenus et qualité des modèles :

Les indicateurs sélectionnés par les différents modèles d'estimation du SCR risque immobilier sont présentés dans le tableau 22 (cf. annexe).

Les variables retenues par la majorité des méthodes sont :

- Le poids des actions,
- Le poids de l'immobilier,
- Le poids des obligations.

Le rôle de la part dédiée à l'immobilier dans le portefeuille de placements de l'organisme assurantiel est prépondérant pour l'estimation du SCR risque immobilier. La part des actions et des obligations ont été retenues car celles-ci présentent une corrélation avec la poche immobilier.

Les modèles linéaires construits ont du mal à ajuster les données. Cela n'est pas dû au non-respect des hypothèses sous-jacentes à l'utilisation de GLM car les graphes des résidus (en fonction des valeurs ajustées, QQ-plot, distance de Cook) sont satisfaisants. Nous avons toutefois voulu ajouter des interactions dans les modèles estimés à partir des variables retenues par GBM et *Random Forest*. Dans ces deux régressions, les effets croisés ont été systématiquement supprimés par la procédure

de sélection de variables. Ils ont cependant été conservés lors de la construction du dernier modèle linéaire, mais cela ne permet pas d'améliorer la qualité d'ajustement et de prévision.

Choix du meilleur modèle :

Les résultats des estimations sont présentés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	50,9%	57,8%
Modèle complet + AIC	50,4%	52,7%
Modèle complet + BIC	50,7%	51,1%
Modèle ACP	51,0%	49,4%
Modèle ACP + AIC	51,0%	49,3%
Modèle ACP + BIC	51,0%	49,3%
Random Forest	19,6%	21,0%
<b>GBM</b>	<b>13,8%</b>	<b>17,9%</b>
Régression Ridge	49,5%	51,9%
Régression LASSO	51,0%	49,4%
Régression Elastic Net	50,4%	52,2%
GLM avec variables issues de RF/GBM	50,6%	50,9%
GLM Final (quadratique)	50,6%	50,9%

*Tableau 42 : détail des résultats pour l'estimation du SCR risque immobilier*

Nous retiendrons le GBM comme le meilleur modèle d'estimation du SCR risque immobilier.

#### **d. SCR risque de change**

Analyse en composantes principales :

L'analyse en composantes principales nous incite à sélectionner les variables suivantes :

- Le poids des obligations, des actions et du monétaire dans le portefeuille de placements,
- Les durations moyennes des obligations.

Les durations moyennes des obligations sont retenues en raison des corrélations qu'elles présentent avec la répartition de l'actif.

## Étude de la distribution pour la modélisation par GLM :

Les histogrammes de la variable d'intérêt sont atypiques :

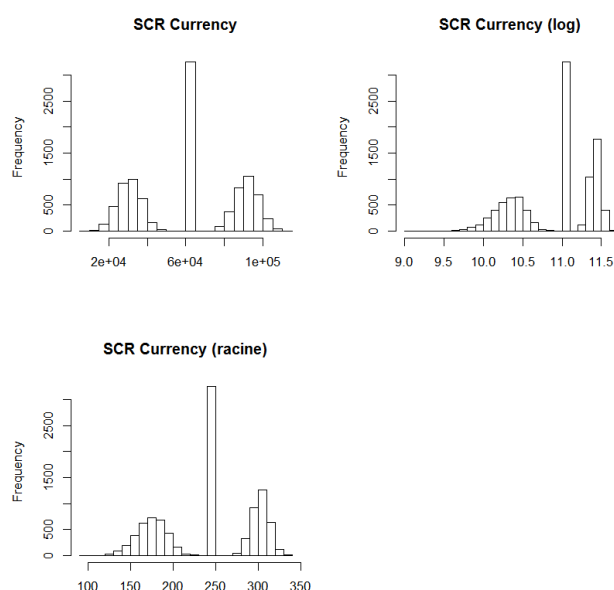


Figure 29: Histogrammes du SCR risque de change

Nous retiendrons la variable brute car la répartition ainsi obtenue est plus symétrique. Le QQ-plot ne nous permet pas d'accepter l'hypothèse de normalité de la variable, mais nous n'avons pas réussi à trouver d'adéquation convenable avec d'autres lois de la famille exponentielle. Nous utiliserons une loi gaussienne pour la suite de la modélisation, mais le non-respect de cette hypothèse risque de poser problème pour l'adéquation.

## Analyse des résultats et choix du meilleur modèle d'estimation :

Le tableau 23 en annexe présente les différentes sélections de variables effectuées. Le risque de change affecte tous les placements hors immobilier, il est donc cohérent que la répartition du portefeuille soit sélectionnée. La poche action reste toutefois la variable la plus retenue. La durée moyenne des OATi est également souvent sélectionnée, mais cela est la conséquence des corrélations présentes dans notre jeu de données. Les résultats des estimations sont présentés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	79,9%	89,5%
Modèle complet + AIC	79,4%	82,9%
Modèle ACP	80,0%	80,3%
Modèle ACP + AIC	80,0%	80,1%
Random Forest	30,0%	31,3%
<b>GBM</b>	<b>23,5%</b>	<b>25,0%</b>
Régression Elastic Net	80,0%	80,1%
GLM avec variables issues de RF (quadratique)	79,9%	80,4%
GLM avec variables issues de GBM (quadratique)	79,7%	81,2%
GLM Final (quadratique)	80,0%	80,1%

Tableau 43 : détail des résultats de l'estimation du SCR risque de change

Nous pouvons constater que l'ajustement des modèles est très peu satisfaisant et que les erreurs absolues moyennes sont très grandes. Cela provient certainement du fait que nous n'avons pas réussi à déterminer la distribution de la variable, et que les modèles proposés ne sont pas adaptés à la modélisation du SCR risque de change. De plus, les procédures de sélection de variables par critère BIC conduisent systématiquement à des modèles vides. Pour les régressions pénalisées *Ridge* et LASSO, la validation croisée nous incite à sélectionner des critères de pénalisation très grands conduisant ainsi à des modèles vides ou avec des coefficients très petits. Nous pouvons également voir que l'ajout d'effets croisés n'améliore en rien la qualité de prévision des modèles. Nous retiendrons toutefois le GBM comme « meilleur » modèle d'estimation du SCR risque de change.

## e. SCR risque de concentration

### Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Part de l'immobilier, des obligations et du monétaire dans le portefeuille de placements,
- Capacité d'absorption des chocs par le FDB pour le risque de concentration.

### Étude de la loi sous-jacente pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

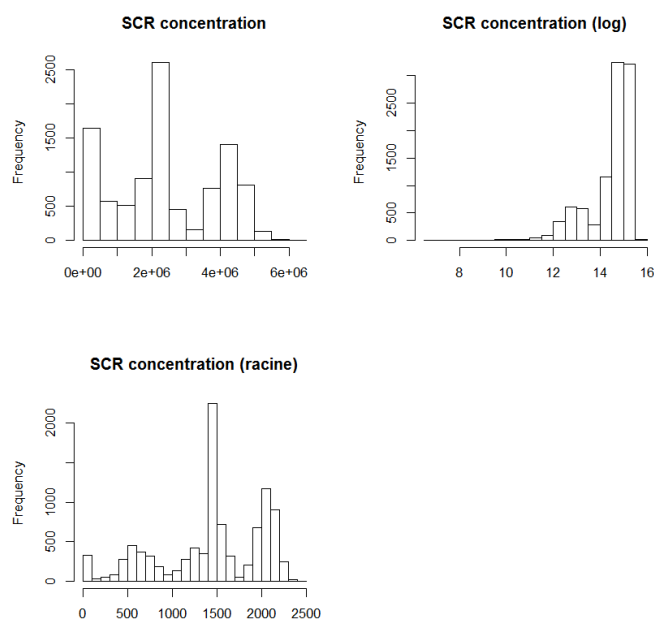


Figure 30 : histogrammes du SCR risque de concentration

Nous retiendrons la transformation par la racine, bien que la distribution obtenue ne soit pas vraiment gaussienne. Le QQ-plot de la variable transformée présente de trop fortes distorsions au niveau des queues de distribution pour nous permettre d'accepter l'hypothèse de normalité. Toutefois, nous n'avons pas réussi à ajuster une loi de la famille exponentielle et nous retiendrons

l'hypothèse de normalité. Ce problème d'adéquation risque cependant de nous poser problème pour l'ajustement des GLM.

#### Analyse des indicateurs sélectionnés et choix du meilleur modèle d'estimation :

Les différentes sélections de variables sont synthétisées dans le tableau 24 en annexe. Le risque de concentration concerne tous les types de placement, il est donc cohérent que les parts des différents types d'actifs du portefeuille soient retenues.

Les résultats des estimations sont détaillés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	29,0%	44,5%
Modèle complet + AIC	28,8%	41,1%
Modèle complet + BIC	29,1%	39,7%
Modèle ACP	29,7%	29,4%
Modèle ACP + AIC / BIC	29,7%	29,3%
Random Forest	21,7%	23,2%
<b>GBM</b>	<b>17,5%</b>	<b>21,7%</b>
Régression Ridge	29,5%	28,7%
Régression LASSO	29,4%	29,3%
Régression Elastic Net	29,1%	31,6%
GLM avec variables issues de RF (quadratique)	28,6%	29,8%
GLM avec variables issues de GBM (quadratique)	28,8%	29,4%
GLM Final (quadratique)	28,8%	29,6%

*Tableau 44 : détails des résultats de l'estimation du SCR risque de concentration*

Les critères d'ajustement et d'erreurs des GLM sont assez peu satisfaisants, certainement en raison du fait d'une mauvaise adéquation à la loi sous-jacente passée en paramètres des méthodes d'estimation. Toutefois, les graphes des résidus sont satisfaisants et les QQ-plot des résidus obtenus ne semblent pas rejeter l'hypothèse de normalité de ceux-ci. L'approche par modèles linéaires n'est pas nécessairement adéquate car le risque de concentration n'est pas calculé dans une approche linéaire. L'ajout d'interactions ne permet pas non plus d'améliorer la qualité du modèle. Nous retiendrons le GBM comme meilleure méthode de prévision du SCR risque de concentration.

#### **f. SCR risque de *spread***

##### Analyse en composantes principales :

L'analyse en composantes principales dans le plan factoriel où les contributions de la variable d'intérêt sont les plus importantes nous suggère de retenir les variables suivantes :

- La duration de l'actif et du passif,
- La duration moyenne des OAT et des OTF,
- Le poids des obligations perpétuelles et de l'immobilier dans le portefeuille de placements.

L'immobilier n'est pas soumis au risque de spread et ne devrait donc pas être retenu. Toutefois, le poids de l'immobilier présente une corrélation de l'ordre de -13 % avec les parts des obligations. Cette corrélation se retrouve dans les résultats de l'ACP.

## Étude de la distribution pour l'estimation par GLM :

L'analyse des histogrammes de la variable à prédire nous incite à utiliser la variable brute :

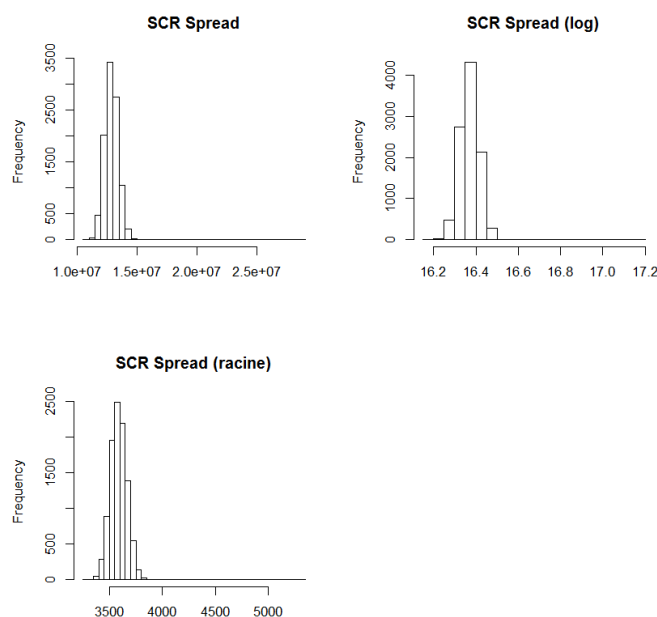


Figure 31: histogrammes du SCR risque de spread

Après suppression d'une valeur aberrante, le QQ-plot nous permet d'accepter l'hypothèse de normalité.

### Analyse des résultats et sélection du meilleur modèle :

Les sélections de variables sont présentées dans le tableau 25 en annexe. Nous pouvons constater que la répartition du portefeuille et la durée des obligations ainsi que celle de l'actif sont sélectionnées par la plupart des modèles d'estimation.

Le risque de *spread* concerne principalement les obligations et le monétaire. La sélection des obligations est donc cohérente. De plus, le choc à appliquer dépend de la notation et de la durée des obligations. Il est donc pertinent que celles-ci soient sélectionnées.

Les résultats des estimations sont présentés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	25,3%	30,9%
Modèle complet + AIC	25,1%	30,8%
Modèle complet + BIC	25,3%	30,5%
Modèle ACP	29,2%	33,1%
Modèle ACP + AIC	29,2%	33,0%
Modèle ACP + BIC	29,2%	33,0%
Random Forest	16,0%	16,1%
<b>GBM</b>	<b>10,1%</b>	<b>11,7%</b>
Régression Ridge	23,5%	27,5%
GLM avec variables issues de RF	27,4%	30,3%
GLM avec variables issues de GBM	27,9%	29,5%
GLM Final	27,5%	29,5%

Tableau 45 : détails des résultats de l'estimation du SCR risque de spread

Les résultats ci-dessus nous indiquent que le GBM présente les meilleurs critères d'ajustement et de prévision. Les taux d'erreurs ne sont toutefois pas convaincants et nous notons une forte dispersion du nuage de points. Cependant, cela ne semble pas être la conséquence du non-respect d'hypothèses sous-jacentes à la modélisation par modèles linéaires car les graphes des résidus (en fonction des valeurs ajustées, QQ-plot, distance de Cook) sont satisfaisants. De plus, l'estimation des critères de pénalisation par validation croisée pour les régressions LASSO et Elastic Net nous conduit à des modèles vides. Nous avons également essayé d'ajouter des interactions pour améliorer l'adéquation des modèles aux données mais les effets croisés ont systématiquement été supprimés par les procédures de sélection de variables.

### g. SCR marché

#### Analyse en composantes principales :

L'analyse en composantes principales nous indique que les variables explicatives présentant un lien avec le SCR marché sont :

- Le ratio AT/MGDC,
- L'âge moyen du portefeuille de rente éducation,
- Le poids des obligations perpétuelles, des obligations et des actions dans le portefeuille de placements,
- La dispersion des âges dans le portefeuille d'arrêt de travail,
- La durée moyenne en invalidité,
- Les proportions du chiffre d'affaires dédiées à l'arrêt de travail et au décès.

## Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable à expliquer sont les suivants :

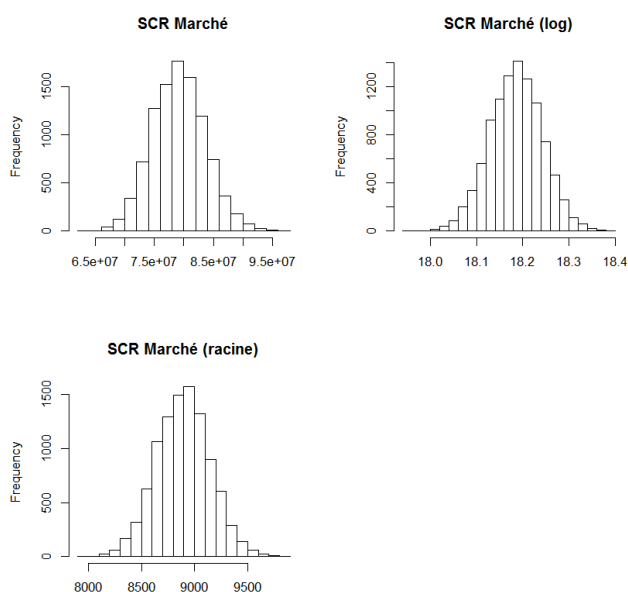


Figure 32 : histogrammes du SCR marché

Nous retiendrons la variable brute car le QQ-plot obtenu nous permet d'accepter l'hypothèse de normalité.

### Analyse des indicateurs sélectionnés et choix du meilleur modèle d'estimation :

Le tableau 26 (cf. Annexes) présente les variables retenues pour l'estimation du SCR marché par chacun des modèles proposés.

Le poids des différentes classes d'actifs et la durée de l'actif sont retenues de nombreuses fois, ainsi que le montant de provisions *Best Estimate* en santé similaire à la vie et les capitaux sous risques utilisés pour calculer le SCR catastrophe en vie. Nous retrouvons indirectement à travers ces sélections le rôle de l'ALM : si l'assureur souhaite honorer ses engagements, il doit se couvrir en adaptant son portefeuille d'actifs. Les capitaux sous risques sont corrélés avec de nombreuses autres variables de notre base de données et sont retenus pour cette raison. Ils sont également révélateurs du risque encouru par l'organisme assureur.

Les résultats des estimations sont détaillés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	26,6%	45,2%
Modèle complet + AIC	26,4%	45,3%
Modèle complet + BIC	26,7%	45,2%
Modèle ACP	33,2%	37,1%
Modèle ACP + AIC	33,2%	37,1%
Modèle ACP + BIC	33,2%	37,1%
Random Forest	16,7%	21,2%
<b>GBM</b>	<b>12,3%</b>	<b>15,4%</b>
Régression Ridge	21,4%	41,7%
Régression LASSO	23,3%	34,2%
Régression Elastic Net	21,2%	30,6%
GLM avec variables issues de RF	27,5%	29,5%
GLM avec variables issues de GBM	26,3%	27,7%
GLM Final	27,1%	29,3%

Tableau 46 : détail des résultats de l'estimation du SCR marché

Les meilleurs résultats sont obtenus grâce au GBM.

#### Sensibilités :

Le GBM nous indiquant que le chiffre d'affaires, les BE Life et NSLT et la durée du passif sont importantes, nous avons tenté de les ajouter dans le dernier modèle linéaire afin d'en améliorer l'adéquation aux données. Les meilleurs résultats sont obtenus grâce à l'ajout du BE NSLT :

- MAE sur la base d'apprentissage = 25,7 %
- MAE sur la base de test = 26,8 %

Bien que cela ne nous permette pas d'égaliser les performances du GBM, nous pouvons conclure que le BE NSLT permet d'estimer le SCR marché avec plus de précision.

#### Comparaison de l'estimation directe et de l'agrégation des estimations des sous-modules :

Nous avons également essayé d'agréger les meilleurs modèles d'estimation des sous-modules composant le SCR marché. Nous obtenons alors :

	Estimation directe	Agrégation des sous-modules
MAE (apprentissage)	12,3 %	21,6 %
MAE (test)	15,4 %	27,4 %

Tableau 47: comparaison de l'estimation directe et de l'agrégation des sous-modules pour le SCR marché

Ces résultats sont bien moins satisfaisants que ceux issus de l'estimation directe, notamment en raison des difficultés que nous avons rencontrées pour l'estimation des sous-modules risque de change et risque de concentration.

## 2. BSCR

### Analyse en composantes principales :

L'analyse en composantes principales nous incite à sélectionner les variables suivantes :

- Les montants de provisions *Best Estimate*,
- Les dispersions des âges dans les portefeuilles de rente de conjoint et de rente éducation,
- La dispersion des anciennetés et l'ancienneté moyenne en invalidité,
- Les capitaux sous risques pour le risque catastrophe en vie,
- L'âge moyen du portefeuille d'arrêt de travail,
- Le ratio entre provisions vie et non-vie en arrêt de travail,
- La proportion du chiffre d'affaires dédiée à la santé.

### Analyse de la distribution pour l'estimation par GLM :

Les histogrammes de la variable à prédire sont les suivants :

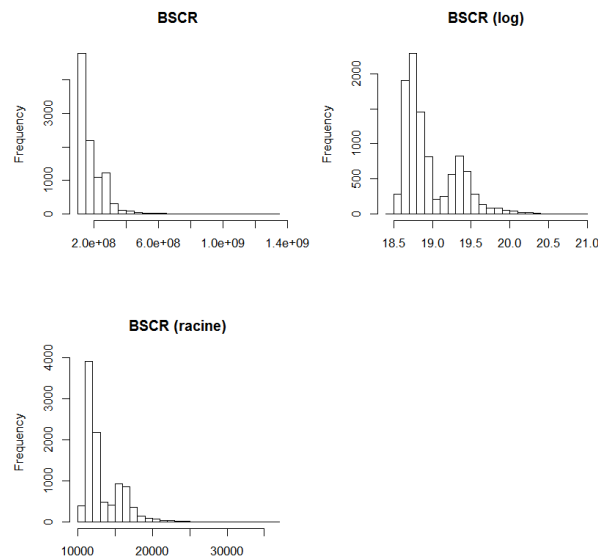


Figure 33: histogrammes du BSCR

Le QQ-plot obtenu entre les quantiles de la variable brute et une loi Gamma est le plus satisfaisant. Toutefois, nous n'arrivons pas à ajuster un GLM avec une telle distribution sur la variable brute. Nous translatoons donc la variable en lui retranchant sa valeur minimale, nous la réduisons et obtenons alors l'histogramme suivant :

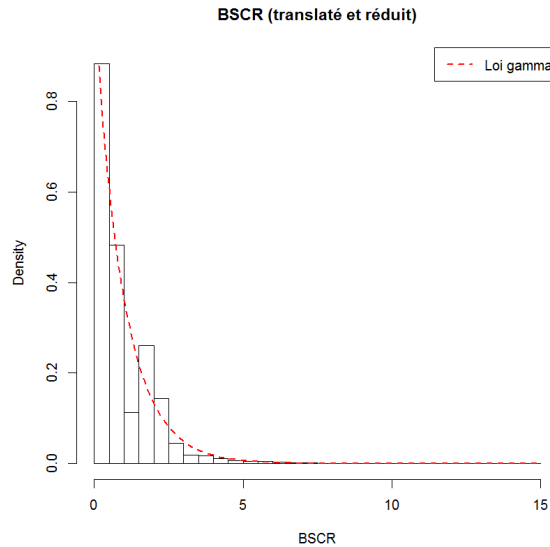


Figure 34: histogramme du BSCR après translation et réduction

Le QQ-plot nous permet d'accepter l'hypothèse d'adéquation à une loi Gamma, malgré quelques distorsions.

Analyse des indicateurs retenus :

Le tableau 27 présenté en annexe détaille les différentes sélections de variables. Les coefficients d'évolution du chiffre d'affaires en santé et arrêt de travail, le ratio entre les provisions dédiées au maintien de la garantie décès et celles afférentes à l'arrêt de travail, les capitaux sous risques, les montants de provisions *Best Estimate* et le poids des actions sont les indicateurs les plus sélectionnés.

Ces sélections sont pertinentes car les indicateurs sur le chiffre d'affaires, les capitaux sous risques et les provisions sont révélateurs de l'engagement de l'assureur et des risques auxquels il est exposé. Le ratio MGDC/AT permet d'expliquer quelle sera la compensation entre l'arrêt de travail (risque NSLT/SLT) et le maintien de la garantie décès (risque *Life*). Enfin, le poids de la poche action est un indicateur des placements de l'assureur, qui lui servent à honorer les engagements pris au passif.

### Choix du meilleur modèle d'estimation :

Les résultats des méthodes d'estimation sont détaillés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	14,4%	15,0%
Modèle complet + AIC	14,3%	14,7%
Modèle complet + BIC	14,4%	14,6%
Modèle ACP	15,8%	16,3%
Modèle ACP + AIC / BIC	15,8%	16,3%
Random Forest	7,9%	10,3%
<b>GBM</b>	<b>4,4%</b>	<b>5,1%</b>
Régression Ridge	14,4%	15,3%
Régression LASSO	14,4%	14,6%
Elastic Net	12,6%	13,8%
GLM avec variables issues de RF	14,0%	14,5%
GLM avec variables issues de GBM	14,1%	14,6%
GLM Final	14,5%	14,5%

Tableau 48 : détail des résultats de l'estimation du BSCR

Les résultats obtenus avec le GBM sont très satisfaisants. Nous pouvons également constater que les performances de *Random Forest* sont convaincantes, et que la régression *Elastic Net* présente de bonnes capacités d'adéquation et de prévision. La répartition du chiffre d'affaires semble donc être un indicateur pertinent pour l'estimation du BSCR. En effet, cette variable nous renseigne quels chocs vont être appliqués pour le calcul du BSCR.

### Sensibilités :

Puisque *Random Forest* et le GBM nous indiquent que le chiffre d'affaires et la durée du passif sont importants, nous les intégrons dans le dernier GLM construit. Ainsi :

- MAE sur la base d'apprentissage = 12,4 %
- MAE sur la base de test = 13,6 %

Nous constatons donc l'importance de ces variables pour la prévision du BSCR.

Comparaison de l'estimation directe et de l'agrégation des estimations des sous-modules :

Nous avons également essayé d'estimer le BSCR par agrégation des meilleures méthodes d'estimation des sous-modules.

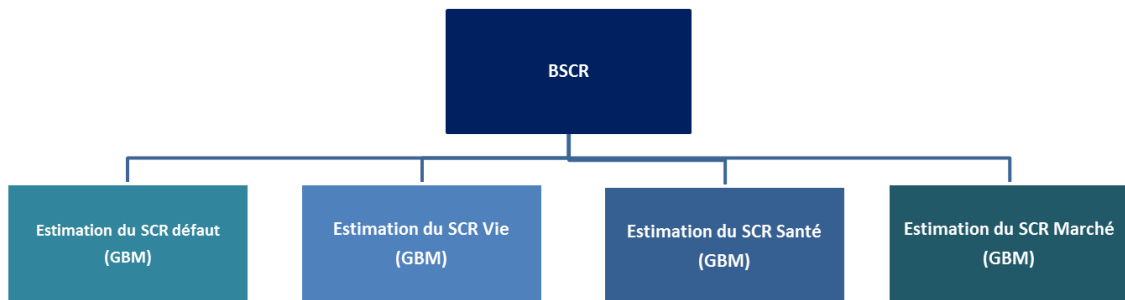


Figure 35: Estimation du BSCR par agrégation des meilleurs modèles d'estimation des sous-modules

Ainsi :

	Estimation directe	Agrégation des sous-modules
MAE (apprentissage)	4,4 %	12,5 %
MAE (test)	5,1 %	13,8 %

Tableau 49 : comparaison de l'estimation directe et de l'agrégation des sous-modules pour le BSCR

L'agrégation ne nous permet pas d'estimer le BSCR de façon plus précise que le GBM.

### 3. SCR

#### Analyse en composantes principales :

L'analyse en composantes principales nous suggère de retenir les variables suivantes :

- Les montants de provisions *Best Estimate*,
- La répartition du chiffre d'affaires,
- Le ratio entre les provisions vie et non-vie en arrêt de travail,
- Les durations moyennes en incapacité et invalidité,
- L'ancienneté moyenne et sa dispersion en invalidité,
- L'âge moyen du portefeuille d'arrêt de travail,
- Les capitaux sous risques (SCR catastrophe en vie),
- La dispersion de l'âge dans le portefeuille de rente de conjoint.

#### Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

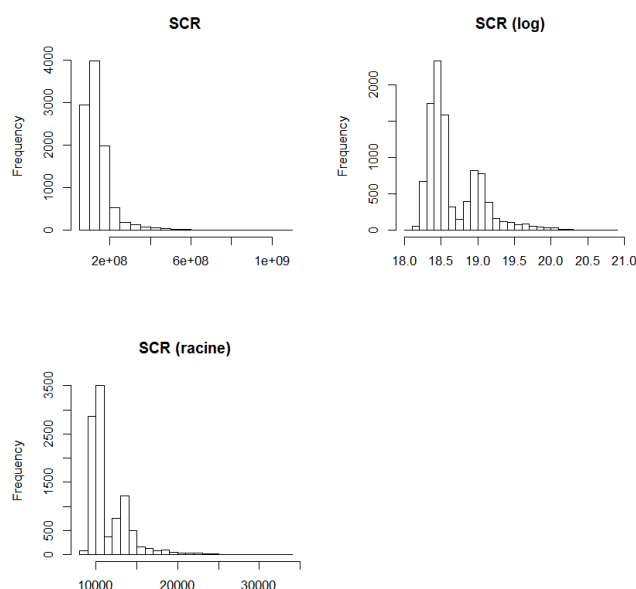


Figure 36 : histogrammes du SCR

Comme nous l'avons fait pour le BSCR, nous translatons et réduisons la variable. Ainsi :

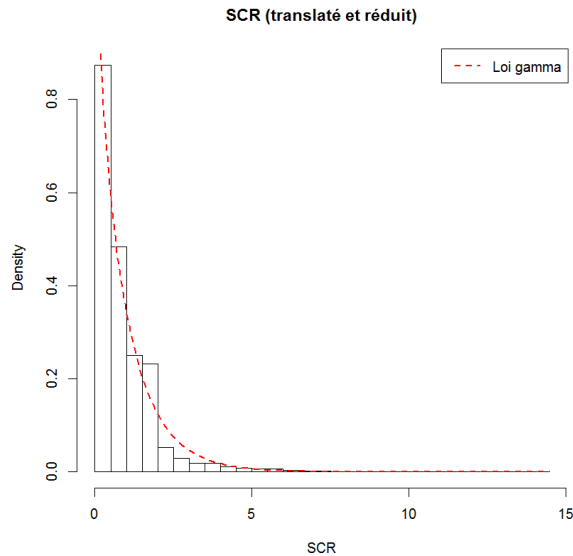


Figure 37 : comparaison de la distribution du SCR transformé avec une loi Gamma

Nous acceptons, grâce au QQ-plot, l'hypothèse d'une distribution Gamma pour les estimations par modèles linéaires généralisés.

Analyse des indicateurs sélectionnés par les différentes méthodes :

Le tableau 28 (cf. Annexes) présente les variables sélectionnées par les différentes méthodes d'estimation. Les coefficients d'évolution du chiffre d'affaires en santé et arrêt de travail, la durée moyenne en incapacité, le ratio MGDC/AT, les capitaux sous risques, les différents BE et le poids des obligations sont retenus dans la majeure partie des cas.

Ces indicateurs sont pour la plupart pertinents : les coefficients d'évolution du chiffre d'affaires, les capitaux sous risques et les montants de provisions nous informent quant à la taille de l'organisme, sa stratégie d'évolution, son type d'activité et son engagement. Le ratio entre les provisions vie et non-vie en arrêt de travail nous indique quelle sera la compensation entre les risques de type santé et ceux de type vie. La durée moyenne en incapacité ne nous semble pas être un indicateur pertinent, contrairement à la durée en invalidité qui est beaucoup plus longue. Cependant de par les corrélations que nous observons dans nos données, nous pouvons supposer que la durée en incapacité a été retenue en raison d'un fort lien avec la durée en invalidité. Nous pouvons toutefois constater que la plupart des indicateurs relatifs aux placements n'ont pas été retenus, alors qu'ils impactent pourtant le SCR.

### Choix de la méthode d'estimation :

Les résultats des estimations sont détaillés ci-après :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	16,2%	17,5%
Modèle complet + AIC	16,1%	17,0%
Modèle complet + BIC	16,2%	16,9%
Modèle ACP	16,6%	17,1%
Modèle ACP + AIC	16,6%	17,1%
Modèle ACP + BIC	16,6%	17,1%
Random Forest	6,6%	8,4%
<b>GBM</b>	<b>4,4%</b>	<b>5,1%</b>
Régression Ridge	16,3%	17,6%
Régression LASSO	16,2%	16,8%
Elastic Net	13,6%	14,1%
GLM avec variables issues de RF	16,3%	16,4%
GLM avec variables issues de GBM	16,3%	16,4%
GLM Final	16,3%	16,4%

Tableau 50 : détail des estimations du SCR

Comme dans le cas du BSCR, les méthodes d'estimation par arbres de régression nous donnent les meilleurs résultats. La régression pénalisée *Elastic Net* nous fournit également des résultats convaincants en comparaison avec les autres méthodes, témoignant ainsi de l'importance de la répartition du chiffre d'affaires pour l'estimation du SCR.

Comme précédemment, le dernier GLM que nous avons estimé n'intègre ni le chiffre d'affaires ni la durée du passif alors que ces indicateurs sont signalés comme importants par le GBM et *Random Forest*. Nous rajoutons donc ces variables dans à notre GLM mais l'amélioration n'est pas autant significative que ce qu'elle pouvait l'être pour le BSCR.

#### **4. Ratio de solvabilité**

##### Analyse en composantes principales :

L'analyse en composantes principales dans le plan factoriel où les contributions du ratio de solvabilité sont les plus fortes nous suggère d'utiliser les variables suivantes :

- Les capitaux sous risques (SCR catastrophe en vie),
- Les montants de provisions *Best Estimate*,
- La répartition du chiffre d'affaires,
- Les différents indicateurs sur l'invalidité : durée moyenne, ancienneté moyenne, dispersion de l'ancienneté,
- Les dispersions des âges dans les portefeuilles de rente de conjoint et de rente éducation,
- La dispersion de l'âge du portefeuille de rente de conjoint,
- Le ratio entre les provisions vie et non-vie en arrêt de travail.

## Étude de la distribution pour l'estimation par GLM :

Les histogrammes de la variable d'intérêt sont les suivants :

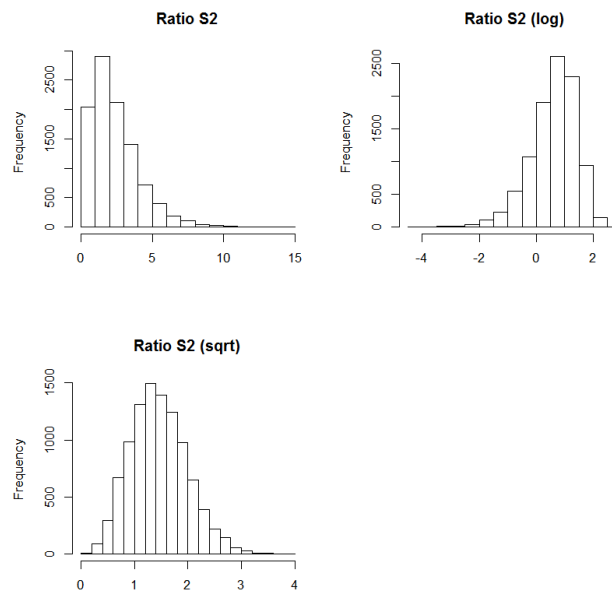


Figure 38 : histogrammes du ratio de solvabilité

La distribution du ratio de solvabilité nous fait penser à une loi gamma. En comparant les densités réelles et théoriques nous obtenons un histogramme convaincant :

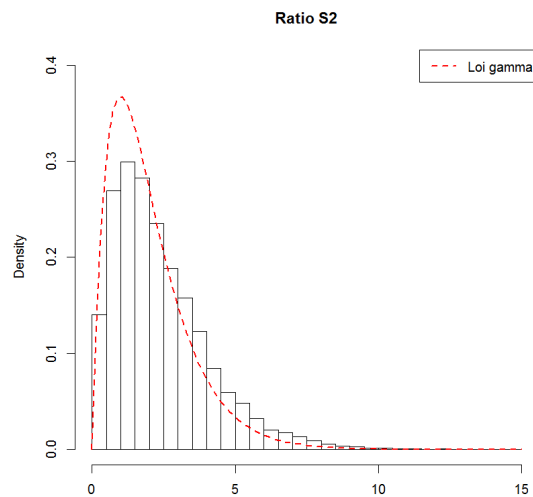


Figure 39 : comparaison de la distribution du SCR avec la loi Gamma

Le QQ-plot obtenu nous permet, malgré une légère distorsion observée au niveau de la queue supérieure de distribution, d'accepter l'hypothèse d'une telle distribution.

### Analyse des variables sélectionnées par les différents modèles :

Nous pouvons voir dans le tableau 29 en annexe que le chiffre d'affaires et son évolution, le ratio MGDC/AT, les capitaux sous risques, les montants de provisions, la durée du passif et le poids des obligations dans les placements sont les variables les plus sélectionnées. Le choix de ces indicateurs est pertinent. En effet, le ratio entre MGDC et arrêt de travail nous indique dans quelle mesure les risques de type « *Life* » vont compenser ceux liés à l'arrêt de travail. Les capitaux sous risques sont corrélés à la répartition du chiffre d'affaires qui est un indicateur des différents chocs subis par l'organisme dans le cadre de Solvabilité II. Les montants de provisions sont révélateurs de l'engagement pris et représentent les assiettes de choc. La durée du passif nous renseigne sur l'intensité avec laquelle seront subis les chocs. Le poids des obligations nous indique l'importance de certains chocs à l'actif. Certains modèles retiennent également d'autres indicateurs sur les placements (poids des actions, durée de l'actif) et cela semble pertinent. Toutefois, les répartitions des différentes classes d'actifs sont corrélées et la sélection d'un seul type de placement peut-être suffisant dans le cadre de nos modèles.

### Choix du meilleur modèle :

Les résultats des estimations sont présentés ci-dessous :

Modèle	MAE (Apprentissage)	MAE (Test)
Modèle complet	18,1%	20,3%
Modèle complet + AIC	17,8%	19,3%
Modèle complet + BIC	18,2%	21,0%
Modèle ACP	18,4%	20,6%
Modèle ACP + AIC	18,4%	20,6%
Modèle ACP + BIC	18,4%	20,4%
Random Forest	13,8%	15,6%
<b>GBM</b>	<b>10,9%</b>	<b>12,8%</b>
Régression Ridge	14,8%	16,5%
Régression LASSO	17,6%	19,5%
GLM avec variables issues de RF	18,6%	19,5%
GLM avec variables issues de GBM	17,1%	19,5%
GLM Final	17,3%	20,2%

*Tableau 51: détail des estimations du ratio de solvabilité*

Les méthodes par arbres de régression sont celles offrant les taux d'erreur les plus convaincants. En revanche, nous pouvons noter des résultats moins satisfaisants que pour les SCR ou le BSCR. En effet, la variable à prédire varie sur un intervalle beaucoup plus réduit, rendant ainsi sûrement plus difficile la discrimination.

### Sensibilité :

Nous avons tenté d'estimer le SCR grâce au GBM puis de calculer le ratio de couverture en considérant les fonds propres éligibles comme une donnée connue. Nous obtenons alors :

- MAE sur la base d'apprentissage = 4,3 %
- MAE sur la base de test = 5,3 %

Nous pouvons ainsi constater qu'il est plus robuste d'estimer le SCR directement et calculer le ratio de solvabilité à partir de cette estimation. Cela suppose de connaître la valeur exacte des fonds propres éligibles. Il aurait été intéressant d'estimer ceux-ci est de comparer l'agrégation des prévisions avec l'estimation directe du ratio de couverture.

## V. Backtesting

Afin de challenger les méthodes d'estimation construites, nous comparons les résultats prédits sur le ratio de solvabilité avec les ratios observés pour 3 entités :

- Une entité ne proposant que des garanties santé (1),
- Une entité proposant tous les types de garanties présentes dans le modèle prévoyance-santé standard d'ACTUARIS® (2),
- Une entité ne proposant que les garanties santé et arrêt de travail (3).

Nous obtenons les résultats suivants :

Entité	Ratio prédit	Ratio observé	Écart relatif (en %)
(1)	164,7%	220,4%	-25,3%
(2)	278,7%	276,5%	0,8%
(3)	172,3%	161,0%	7,0%

*Tableau 52: résultats du backtesting sur des entités réelles*

Nous pouvons voir que les résultats de l'estimation pour une entité présentant toutes les garanties du modèle prévoyance-santé d'ACTUARIS® sont très satisfaisants.

Nous pouvons constater que l'estimation du ratio de solvabilité pour l'entité (1) est particulièrement difficile et que le modèle sous-estime fortement le ratio de couverture. L'écart observé provient d'une difficulté à estimer le SCR de l'entité, car celui-ci présente une valeur bien moins élevée que les entités fictives présentes dans notre base de données.

En revanche, les deux autres entités présentent des caractéristiques plus proches des observations de la base sur laquelle nous avons calibré les modèles. Ainsi, les paramètres de ces-derniers sont adaptés à la prévision du ratio de couverture des entités (1) et (3).

Dans le cas du dernier organisme, le SCR marché représente une part très importante du SCR. Comme nous l'avons vu, l'estimation du SCR marché est moins satisfaisante que les prévisions proposées pour les SCR souscription. Nous pouvons dresser le même constat lorsque nous analysons les différents SCR estimés pour cette entité, expliquant ainsi l'erreur obtenue. Toutefois, les compensations entre les risques permettent d'obtenir une erreur relativement convaincante.

Au vu des précédents résultats, il aurait été intéressant de mener cette étude sur une base uniquement constitué de données réelles et peut-être plus représentatives du marché et des différences entre les organismes assurantiels.

## Conclusion

Nous avons réussi à mettre en place des méthodes rapides d'estimation du ratio de couverture, du SCR et de ses composantes. Celles-ci ont pu nous fournir des résultats relativement convaincants. Ces méthodes ne pourront en aucun cas permettre de contourner les calculs exigés par la réglementation mais peuvent être utilisés dans un but de pilotage, de sensibilités ou d'aide à la décision.

L'approche par modèles linéaires généralisés nous permet d'apporter un aspect intuitif, mais n'est pas utilisable de manière opérationnelle en raison des transformations que nous avons dû effectuer sur les variables et qui risquent d'« écraser » les erreurs. Nous avons pu en outre constater la difficulté d'estimation par modèles linéaires car les chocs appliqués ne supposent pas toujours une relation linéaire entre le SCR et les variables explicatives. Nous avons quand même pu apprécier les capacités des méthodes de régression pénalisées, qui permettent d'obtenir un bon ajustement avec des modèles relativement simples.

Les méthodes par arbres de régression telles que *Random Forest* et le *Gradient Boosting Machine* offrent des résultats plus performants et ne nécessitent quant à elle aucun traitement des données, ce qui leur confère un avantage non négligeable. L'estimation par réseaux de neurones ou les *Support Vector Machine* aurait pu s'avérer également pertinente car ces méthodes, bien que difficilement interprétables, offrent des performances satisfaisantes.

Nous avons également dû faire face aux corrélations induites par la simulation des données et celles-ci peuvent avoir faussé les estimations des coefficients des régressions linéaires ou les choix de variables et les successions des divisions dans les arbres de régression. Il serait intéressant de mener une étude similaire à la nôtre à partir de données d'entités réelles, bien que la constitution d'une base de données de taille suffisante à l'utilisation de méthodes d'apprentissage supervisée nécessite de nombreuses observations. Nous avons aussi piloté la distribution des variables grâce à notre méthode de simulation, qui ne reflète donc pas la réalité.

## Bibliographie

ACTUARIS [2017] « Pilier 1 et formule standard », support de formation

ACTUARIS [2018] « ADDACTIS MODELING®, Documentation du modèle Prévoyance-Santé standard »

BESSE et *al.*, « Arbres binaires de décision », *Wikistat*, <http://wikistat.fr/pdf/st-m-app-cart.pdf> [En ligne ; consulté le 23 décembre 2016].

BESSE et *al.*, « Agrégation de modèles », *Wikistat*, <http://wikistat.fr/pdf/st-m-app-agreg.pdf> [En ligne ; consulté le 27 décembre 2016].

BREIMAN et *al.* [1984] « Classification and regresssion trees », Wadsworth & Brooks

BUZZI A. [2017] « Approximation du bilan économique sous Solvabilité II via des méthodes d'apprentissage automatique et application à l'ORSA », Mémoire d'actuaire sous la direction de Khalid JEBBARI, Université de Paris Dauphine

FRIEDMAN J. [2002] « Stochastic Tree Boosting », *Computational Statistics and Data Analysis*, vol. 38, 367-378

Règlement délégué 2015/35 de la commission européenne

## Liste des figures

Figure 1 : Bilan économique simplifié .....	3
Figure 2: calcul du SCR .....	4
Figure 3: Arbre de régression pour l'estimation du SCR Life .....	7
Figure 4: Détail des chocs appliqués dans le modèle prévoyance-santé .....	12
Figure 5: impact des simulations sur l'ensemble des hypothèses du modèle .....	15
Figure 6: histogrammes du SCR de défaut de type 1 .....	21
Figure 7: histogrammes du SCR de défaut de type 2 .....	23
Figure 8: Histogrammes pour le SCR de défaut .....	25
Figure 9: Histogrammes du SCR de mortalité (vie) .....	28
Figure 10: Histogrammes du SCR de longévité (Vie) .....	31
Figure 11: Histogrammes du SCR de morbidité (Vie) .....	34
Figure 12: Histogrammes du SCR de Frais (vie) .....	37
Figure 13: histogrammes du SCR révision (Life) .....	40
Figure 14: Histogrammes du SCR Vie .....	43
Figure 15: histogrammes du SCR longévité (HSLT) .....	46
Figure 16: Histogrammes de SCR frais (HSLT) .....	49
Figure 17: Histogrammes du SCR révision (HSLT) .....	52
Figure 18: histogrammes du SCR santé similaire à la vie .....	54
Figure 19: histogrammes du SCR HNSLT .....	58
Figure 20: histogrammes du SCR concentration (risque catastrophe en santé) .....	61
Figure 21: histogrammes du SCR Accident de masse (risque catastrophe en santé) .....	63
Figure 22: histogrammes du SCR pandémie (risque catastrophe en santé) .....	65
Figure 23 : histogrammes du SCR catastrophe en santé .....	67
Figure 24: histogrammes du SCR santé .....	69
Figure 25: Agrégation des estimations des sous-modules du SCR santé .....	71
Figure 26: histogrammes du SCR risque de taux d'intérêts .....	72
Figure 27: Histogrammes du SCR risque action .....	75
Figure 28 : histogrammes du SCR risque immobilier .....	77
Figure 29: Histogrammes du SCR risque de change .....	79
Figure 30 : histogrammes du SCR risque de concentration .....	80
Figure 31: histogrammes du SCR risque de spread .....	82
Figure 32 : histogrammes du SCR marché .....	84
Figure 33: histogrammes du BSCR .....	86
Figure 34: histogramme du BSCR après translation et réduction .....	87
Figure 35: Estimation du BSCR par agrégation des meilleurs modèles d'estimation des sous-modules .....	89
Figure 36 : histogrammes du SCR .....	90
Figure 37 : comparaison de la distribution du SCR transformé avec une loi Gamma .....	91
Figure 38 : histogrammes du ratio de solvabilité .....	93
Figure 39 : comparaison de la distribution du SCR avec la loi Gamma .....	93

## Liste des tableaux

Tableau 1: Fonctions de liens canoniques.....	10
Tableau 2: Détail des résultats pour l'estimation du SCR de défaut de type 1.....	22
Tableau 3: Détail des résultats pour l'estimation du SCR de défaut de type 2.....	24
Tableau 4: détail des résultats pour l'estimation du SCR de défaut .....	26
Tableau 5: comparaison de l'estimation directe et de l'agrégation pour le SCR de défaut .....	27
Tableau 6: coefficients du GLM d'estimation du SCR de défaut .....	27
Tableau 7: Détail des résultats des estimations du SCR mortalité (Vie) .....	29
Tableau 8: Coefficients du GLM d'estimation du SCR mortalité (Ve).....	30
Tableau 9: Détail des résultats pour l'estimation du SCR de longévité (Vie) .....	32
Tableau 10 : coefficients du GLM d'estimation du SCR de longévité (vie).....	33
Tableau 11: Détail des résultats pour l'estimation du SCR de morbidité (Vie).....	35
Tableau 12: coefficients du GLM d'estimation du SCR de morbidité (vie).....	36
Tableau 13: Détail des résultats des modèles d'estimation du SCR de Frais (vie).....	38
Tableau 14: coefficients du GLM d'estimation du SCR de frais (Vie) .....	39
Tableau 15: détail des résultats pour l'estimation du SCR révision (Vie) .....	41
Tableau 16: coefficients du GLM d'estimation du SCR révision (Vie) .....	42
Tableau 17: Détail des résultats de l'estimation du SCR vie .....	44
Tableau 18: Coefficients du GLM d'estimation du SCR Vie .....	45
Tableau 19: comparaison de l'agrégation et de l'estimation directe pour le SCR Vie.....	45
Tableau 20 : Détails des résultats pour l'estimation du SCR longévité (HSLT).....	48
Tableau 21: coefficients du GLM d'estimation du SCR Longévité (HSLT).....	48
Tableau 22: Résultats détaillés des estimations du SCR frais (HSLT) .....	50
Tableau 23: coefficients du GLM d'estimation du SCR de frais (HSLT) .....	51
Tableau 24: Détail des résultats de l'estimation du SCR révision (HSLT).....	53
Tableau 25: coefficients du GLM d'estimation du SCR révision (HSLT).....	53
Tableau 26: détail des résultats des estimations du SCR santé similaire à la vie .....	55
Tableau 27: coefficients du GLM d'estimation du SCR HSLT.....	56
Tableau 28: comparaison de l'agrégation et de l'estimation directe pour le SCR santé similaire à la vie .....	57
Tableau 29 : Détail des résultats de l'estimation du SCR HNSLT.....	59
Tableau 30: Coefficients du GLM d'estimation du SCR HNSLT.....	59
Tableau 31: détail des résultats de l'estimation du SCR concentration (catastrophe en santé) .....	62
Tableau 32: détail des résultats de l'estimation du SCR accident de masse.....	64
Tableau 33: détail des résultats des estimations du SCR Pandémie.....	66
Tableau 34: détail des résultats de l'estimation du SCR catastrophe en santé .....	68
Tableau 35 : coefficients du GLM d'estimation du SCR catastrophe en santé.....	68
Tableau 36: détail des résultats de l'estimation du SCR santé .....	70
Tableau 37: comparaison de l'estimation directe et de l'agrégation des sous-modules pour le SCR santé .....	71
Tableau 38 : coefficients du GLM d'estimation du SCR santé .....	71
Tableau 39: détail des résultats de l'estimation du SCR risque de taux d'intérêts.....	73
Tableau 40 : coefficients du GLM d'estimation du SCR risque de taux d'intérêts .....	74
Tableau 41: détail des résultats de l'estimation du SCR risque action .....	76

Tableau 42 : détail des résultats pour l'estimation du SCR risque immobilier .....	78
Tableau 43 : détail des résultats de l'estimation du SCR risque de change.....	79
Tableau 44 : détails des résultats de l'estimation du SCR risque de concentration .....	81
Tableau 45 : détails des résultats de l'estimation du SCR risque de spread .....	83
Tableau 46 : détail des résultats de l'estimation du SCR marché .....	85
Tableau 47: comparaison de l'estimation directe et de l'agrégation des sous-modules pour le SCR marché.....	85
Tableau 48 : détail des résultats de l'estimation du BSCR .....	88
Tableau 49 : comparaison de l'estimation directe et de l'agrégation des sous-modules pour le BSCR	89
Tableau 50 : détail des estimations du SCR.....	92
Tableau 51: détail des estimations du ratio de solvabilité.....	94
Tableau 52: résultats du backtesting sur des entités réelles .....	96

# ANNEXES

## Annexe A : Compléments sur les arbres binaires de décision

### Critère d'homogénéité :

Il convient de distinguer le cas des arbres de discrimination et des arbres de régression. Nous détaillerons ici que les arbres de régression, car nos variables d'intérêt sont quantitatives.

Soit une variable à expliquer quantitative  $Y$  avec une partition en  $J$  classes d'effectifs respectifs  $n_1, \dots, n_J$ . Pour chacune des classes, les individus sont notés  $i = 1, \dots, n_j$ . La valeur théorique (et inconnue) du  $i^{\text{ème}}$  individu de la classe  $j$  est  $\mu_{ij}$  et sa valeur observée est  $y_{ij}$ .

L'hétérogénéité de la classe  $j$  est égale à :

$$D_j = \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2 \text{ où } \mu_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{ij}$$

Cette hétérogénéité est donc homogène à la variance intra-classe. Elle est nulle si et seulement si les tous les éléments de la classe ont la même valeur, c'est-à-dire si le nœud est parfaitement homogène. L'hétérogénéité totale est la somme des hétérogénéités de chacune des  $J$  classes. Elle est homogène à la variance totale :

$$D = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2$$

L'hétérogénéité de l'ensemble non partagé est :

$$D_{tot} = \sum_{j=1}^J \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{..})^2 \text{ où } \mu_{..} = \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^{n_j} \mu_{ij}$$

Un arbre de décision binaire doit séparer au mieux les observations. Ainsi il doit à chaque itération maximiser la différence des hétérogénéités entre l'ensemble non scindé et l'ensemble divisé en  $J$  classes. Cette quantité est :

$$\Delta = D_{tot} - D$$

En développant le calcul, nous avons que :

$$\Delta = \sum_{j=1}^J n_j (\mu_{..} - \mu_{.j})^2$$

Cette quantité est homogène à la variance inter-classes. **À chaque étape, l'arbre doit trouver la division qui entraîne une partition en deux groupes bien distincts, chacun devant être le plus homogène possible.**

Cependant, les  $\mu_{ij}$  sont inconnus. L'hétérogénéité de chaque classe est donc estimée par

$$\hat{D}_j = \sum_{i=1}^{n_j} (\mu_{ij} - y_{ij})^2 \text{ où } y_{.j} = \sum_{i=1}^{n_j} y_{ij}$$

Puis l'hétérogénéité de la partition est estimée par plug-in :

$$\hat{D} = \sum_{j=1}^J \hat{D}_j$$

### Élagage et recherche de l'arbre optimal :

Les arbres obtenus par itération des étapes expliquées précédemment peuvent parfois être très raffinés et donc instables car ils dépendent grandement de l'échantillon qui a permis de les construire. Il est donc préférable de les élaguer afin d'obtenir des modèles plus parcimonieux et robustes. L'approche proposée par Breiman [1984] consiste à construire l'arbre maximal (c'est-à-dire un arbre où plus aucune division n'est possible) puis à construire une suite d'arbres emboîtés par élagage successif. Le choix du « meilleur » arbre se fait selon un critère à définir.

Soit un arbre  $A$  donné, contenant  $K$  nœuds terminaux. La qualité de discrimination de cet arbre est mesurée via un critère :

$$D(A) = \sum_{k=1}^K D_k(A)$$

$D_k(A)$  peut par exemple être le nombre d'observations mal classées dans la feuille  $k$ , l'hétérogénéité du nœud ou la déviance...

L'élagage de l'arbre repose sur un critère pénalisé du type :

$$C(A) = D(A) + \gamma K$$

Avec  $\gamma = 0$ , l'arbre maximal  $A_K$  minimise ce critère. Lorsque l'on augmente le paramètre  $\gamma$ , une des divisions de  $A$  apparaît comme superflue et est éliminée, les deux nœuds fils étant regroupés dans le nœud père qui devient ainsi terminal. L'arbre  $A_K$  devient l'arbre  $A_{K-1}$ . La division supprimée est celle pour laquelle l'amélioration de  $D$  est la plus faible.

En faisant croître le paramètre de pénalisation à chaque itération, nous pouvons obtenir une suite d'arbres emboîtés :

$$A_1 \subset \dots \subset A_{K-1} \subset A_K$$

Où  $A_1$  est l'arbre minimal, c'est-à-dire l'arbre où toutes les observations sont dans la racine.

Une fois cette séquence d'arbre construite, le choix de l'arbre optimal repose sur un critère. Il est par exemple possible de tracer l'évolution de la déviance en fonction du nombre de feuilles de l'arbre ou en fonction du critère  $\gamma$ . L'arbre optimal doit offrir un bon compromis entre complexité et qualité de discrimination.

## Annexe B : Gradient Tree Boosting

Soit une variable  $Y$  à prévoir et  $X$  un jeu de  $p$  variables explicatives, avec un échantillon de taille  $N$ . Notre but est de trouver une fonction  $F^*(x)$  reliant  $x$  à  $y$  et qui minimise une fonction de perte  $\psi(y, F(x))$  :

$$F^*(x) = \underset{F(x)}{\operatorname{argmin}} E[\psi(y, F(x))]$$

La fonction  $F^*(x)$  est approchée par une expression additive de la forme :

$$F(x) = \sum_{m=0}^M \beta_m h(x, a_m)$$

Où les fonctions  $h$  (les *base learner*) sont choisies comme des fonctions simples (telles que des combinaisons linéaires) de  $x$  et des paramètres  $a = (a_1, a_2, \dots)$ . L'ajustement des paramètres  $\{\beta_m\}_{m=0}^M$  et  $\{a_m\}_{m=0}^M$  est fait étape par étape :

$$(a_m, \beta_m) = \underset{\beta, a}{\operatorname{argmin}} \sum_{i=1}^N \psi(y_i, F_{m-1}(x_i) + \beta h(x_i, a))$$

À partir de cet ajustement, la fonction de régression  $F$  est ajustée :

$$F_m(x) = F_{m-1}(x) + \beta_m h(x, a_m)$$

Afin de déterminer les paramètres  $a_m$  et  $\beta_m$  optimaux à chaque étape, Friedman [2002] propose dans un premier temps d'ajuster une fonction  $h$  par moindres carrés au gradient de la fonction de perte (*pseudo-résidu*) puis de déterminer le paramètre  $\beta_m$  qui minimise la perte entre la variable à prédire et la fonction de régression agrégée. Ainsi :

$$a_m = \underset{a, \rho}{\operatorname{argmin}} \sum_{i=1}^N [\tilde{y}_{i,m} - \rho h(x_i, a)]^2 \text{ avec } \tilde{y}_{i,m} = - \left[ \frac{\partial \psi(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}$$

Et

$$\beta_m = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N \psi(y_i, F_{m-1}(x_i) + \beta h(x_i, a_m))$$

Dans le Gradient Tree Boosting,  $h$  est un arbre de régression à  $L$  nœuds terminaux. À l'itération  $m$ , un arbre de régression est ajusté et divise l'échantillon en  $L$  sous-ensembles notés  $\{R_{lm}\}_{l=1}^L$ . A chaque feuille est associée une valeur, correspondant à la moyenne des observations qui y sont regroupées, c'est-à-dire :

$$h(x, \{R_{lm}\}_1^L) = \sum_{l=1}^L \bar{y}_{lm} \mathbf{1}(x \in R_{lm})$$

avec  $\bar{y}_{lm}$  la moyenne dans chaque feuille des  $\tilde{y}_{i,m}$ .

Puisque la valeur associée à chaque feuille de l'arbre est constante, le calcul de  $\beta_m$  est fait à chaque nœud terminal :

$$\gamma_{lm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{lm}} \psi(y_i, F_{m-1}(x_i) + \gamma)$$

La fonction de régression est ensuite mise à jour :

$$F_m(x) = F_{m-1}(x) + \nu \gamma_{lm} 1(x \in R_{lm})$$

Le paramètre de rétrécissement  $\nu \in [0; 1]$  (*shrinkage parameter*) contrôle le taux d'apprentissage de l'algorithme en pénalisant l'ajout d'un nouveau modèle dans l'agrégation. De manière empirique, il a été montré que si  $\nu \leq 0.1$  la qualité de prévision du modèle finale est satisfaisante, bien que cela conduise à augmenter le nombre d'arbres.

Dans son article, Friedman propose différentes fonctions de pertes telles que :

- Les moindres carrés :  $\psi(y, F) = (y - F)^2$
- L'écart absolu :  $\psi(y, F) = |y - F|$
- Huber-M :  $\psi(y, F) = (y - F)^2 1(|y - F| \leq \delta) + 2\delta \left( |y - F| - \frac{\delta}{2} \right) 1(|y - F| > \delta)$

Nous utiliserons les moindres carrés.

## Annexe C : Théorie du modèle linéaire gaussien

### Estimation :

Le but est de déterminer une droite de régression  $\hat{Y} = X\hat{\beta}$  la plus proche possible des données observées. Les paramètres de la régression sont estimés par la méthode des moindres carrés :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} (Y - X\beta)'(Y - X\beta)$$

La résolution de ce problème d'optimisation conduit à :

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Cet estimateur est le même que celui obtenu par la méthode du maximum de vraisemblance dans le cas où les erreurs sont gaussiennes (comme nous l'avons supposé plus tôt). Cet estimateur vérifie :

$$E[\hat{\beta}] = \beta \text{ et } \operatorname{Var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$$

De plus, si les  $\epsilon_i$  sont indépendantes et identiquement distribuées selon une loi normale centrée, alors  $\hat{\beta}$  est le meilleur estimateur parmi tous les estimateurs sans biais de  $\beta$ .

La variance des erreurs  $\sigma^2$  est cependant inconnue, de même que les erreurs  $\epsilon$ . Celles-ci sont estimées par :

$$\hat{\epsilon} = \hat{Y} - Y \text{ avec } \hat{Y} = X\hat{\beta}$$

L'estimateur  $S^2$  défini par :

$$S^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{(n - p - 1)}$$

Est un estimateur sans biais de  $\sigma^2$ . De plus,  $S^2 \sim \sigma^2 \chi_{n-p-1}^2$ .

### Tests sur les paramètres du modèle :

Nous nous plaçons ici dans le cadre où les erreurs sont gaussiennes, centrées, homoscédastiques. Dans le cas contraire, les tests présentés dans cette partie ne peuvent être appliqués.

Un premier test à effectuer sur un modèle linéaire est celui de l'existence d'un sous-modèle. Dans le cadre d'un modèle avec  $p$  paramètres, cela revient à tester :

$$H_0: \exists i \in [0, \dots, j] \mid \beta_i = 0$$

Contre :

$$H_1: \forall j = 0, \dots, p \quad \beta_j \neq 0$$

Nous notons  $SCR$  la somme des carrés des résidus pour le modèle complet (avec  $p$  paramètres) :

$$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Et  $SCR_0$  la même quantité pour le sous-modèle avec  $p_0 < p$  paramètres.

Sous  $H_0$ , la variable :

$$\hat{F} = \frac{(SCR_0 - SCR)/(p - p_0)}{SCR/(n - p)}$$

Suit une loi de Fisher à  $p - p_0$  et  $n - p - 1$  degrés de liberté.

Intuitivement, si l'écart  $SCR_0 - SCR$  est très important, cela signifie que la somme des carrés résiduelle entre le sous-modèle et le modèle complet est importante, et il apparaît donc peu probable que le sous-modèle ajuste bien les données. En revanche, si cet écart est faible, cela signifie que le sous-modèle ajuste presque aussi bien les données que le modèle complet.

La règle de décision pour un test de niveau  $\alpha$  est donc :

Si  $\hat{F} \leq F_{p-p_0, n-p-1; 1-\alpha}$ , nous acceptons  $H_0$  et nous ne l'acceptons pas sinon.

Nous pouvons également tester la nullité d'une combinaison linéaire de certains paramètres. Les hypothèses du test sont alors :

$$H_0: C'\beta = 0 \quad \text{et} \quad H_1: C'\beta \neq 0$$

Où  $C$  est un vecteur de  $R^p$ .

Sous l'hypothèse  $H_0$ , la statistique :

$$\hat{T} = \frac{C'\hat{\beta}}{S\sqrt{C'(X'X)^{-1}C}}$$

Suit une loi de Student à  $n - p - 1$  degrés de liberté. Nous rejetons l'hypothèse nulle si  $t_{obs} > T_{n-p-1; 1-\frac{\alpha}{2}}$ .

Ce test peut être généralisé pour tester la nullité conjointe de plusieurs combinaisons linéaires. Dans ce cas, nous souhaitons tester :

$$H_0: C'\beta = 0 \quad \text{et} \quad H_1: C'\beta \neq 0$$

Où  $C$  est cette fois-ci une matrice de  $p$  lignes et  $m$  colonnes, où  $m$  correspond au nombre de colonnes à tester. Sous l'hypothèse nulle, nous avons :

$$\hat{F} = \frac{\hat{\beta}'C(C'(X'X)^{-1}C)^{-1}C'\hat{\beta}}{mS^2}$$

Suit une loi de Fisher à  $m$  et  $n - p - 1$  paramètres. L'hypothèse  $H_0$  est donc rejetée si  $f_{obs} > F_{m, n-p; 1-\alpha}$ .

#### Adéquation et validation du modèle :

Un premier critère de qualité du modèle est le critère d'ajustement. Il est défini par :

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

Avec :

$$SCT = SCE + SCR$$

$SCT$  représente la somme des carrés totales,  $SCE$  la somme des carrés de la régression et  $SCR$  la somme des carrés résiduelle :

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n \epsilon_i^2$$

Le  $R^2$  représente donc la part de variance des observations expliquée par le modèle. Idéalement, ce dernier doit être proche de 1.

Cependant, l'augmentation du nombre de variables explicatives entraîne systématiquement une augmentation du  $R^2$ . Ce critère a donc tendance à sur-ajuster. Afin de palier ce défaut du critère d'ajustement, une version ajustée a été proposée, en intégrant une pénalisation par rapport au nombre de variables du modèle :

$$R_a^2 = 1 - \frac{n-1}{n-p-1} (1 - R^2)$$

Nous pouvons également analyser la qualité de prévision d'un modèle. Soit  $y_{n+1}$  une observation de la variable à expliquer n'étant pas contenue dans l'échantillon ayant permis à estimer  $\hat{\beta}$  et  $(x_{n+1,1}, x_{n+1,2}, \dots, x_{n+1,p})$  le vecteur des observations des  $p$  variables explicatives associé.

La valeur prédite pour  $Y_{n+1}$  est :

$$\hat{Y}_{n+1}^p = X_{n+1} \hat{\beta}$$

L'erreur de prévision pour cette observation est définie par :

$$\epsilon_{n+1}^p = Y_{n+1} - \hat{Y}_{n+1}^p$$

Elle vérifie :

$$E[\epsilon_{n+1}^p] = 0 \text{ et } \text{Var}[\epsilon_{n+1}^p] = \sigma^2 (1 + X_{n+1}' (X'X)^{-1} X_{n+1})$$

Comme l'erreur de prévision est centrée, sa variance est en fait donnée par l'erreur quadratique moyenne de prévision (EQMP) :

$$\text{Var}[\epsilon_{n+1}^p] = E[(Y_{n+1} - \hat{Y}_{n+1}^p)^2]$$

La version empirique de l'EQMP définit le critère PRESS (*Predicted Residual Sum of Squares*) :

$$\text{PRESS} = \sum_{i=1}^n (\epsilon_{(i)}^p)^2$$

Avec :

$$\epsilon_{(i)}^p = Y_i - \hat{Y}_i^p = Y_i - X_i \hat{\beta}_{(i)}$$

Où  $\hat{\beta}_{(i)}$  est l'estimateur des moindres carrés de  $\beta$ , calculé pour l'échantillon privé de la  $i^{\text{ème}}$  observation. Plus le PRESS est petit, meilleur est la faculté de prévision du modèle.

La validation du modèle obtenu passe également par l'analyse des résidus. Ceux-ci sont centrés, hétéroscédastiques et corrélés :

$$E[\hat{\epsilon}] = 0 \quad \text{et} \quad \text{Var}[\hat{\epsilon}] = \sigma^2 M$$

Avec  $M = I - H = I_n - X(X'X)^{-1}X'$ .

Dans la suite, le  $i^{\text{ème}}$  élément diagonal de la matrice  $H$  sera noté  $h_{i,i}$ . Les résidus normalisés définis par :

$$r_i = \frac{\hat{\epsilon}_i}{\sigma \sqrt{1 - h_{i,i}}}$$

Sont homoscédastiques. Cependant, comme  $\sigma^2$  est inconnu et estimé par  $S^2$ , nous obtenons les résidus standardisés :

$$t_i = \frac{\hat{\epsilon}_i}{S \sqrt{1 - h_{i,i}}}$$

Toutefois, l'utilisation de ces résidus peut poser problème. En effet, si la  $i^{\text{ème}}$  observation est aberrante, la valeur du résidu  $\hat{\epsilon}_i$  sera grande. Par conséquent,  $S^2$  sera également grand. Ainsi, la valeur de  $t_i$  ne permet pas de détecter des valeurs aberrantes. Pour contourner ce problème, nous définissons les résidus studentisés :

$$t_i^* = \frac{\hat{\epsilon}_i}{S_{(i)} \sqrt{1 - h_{i,i}}}$$

Où  $S_{(i)}$  est l'estimateur de  $\sigma$  basé sur l'échantillon privé de la  $i^{\text{ème}}$  observation. Dans le cas d'une valeur aberrante, seule la valeur de  $\hat{\epsilon}_i$  sera impactée, et le résidu studentisé sera donc important.

Nous devons également contrôler la normalité des résidus. Ce contrôle peut se faire à l'aide d'un QQ-plot, représentant les quantiles empiriques de l'échantillon des résidus contre les quantiles théoriques d'une loi normale. Si les résidus sont bel et bien gaussiens, le QQ-plot doit avoir la forme d'une droite passant par l'origine et de pente égale à 1. Nous pouvons aussi effectuer un test de Shapiro-Wilk pour tester la normalité des résidus.

Il convient également de tester si les résidus studentisés sont homoscédastiques. Pour cela, il suffit de tracer la valeur de ces résidus contre celle de la variable à expliquer. Le graphe obtenu ne doit pas présenter de structure.

Enfin, l'analyse des valeurs aberrantes est une étape importante dans la validation d'un modèle. Une fois que celles-ci ont été détectées, il convient de voir leur influence sur l'estimation des paramètres du modèle.

Afin de déterminer si une observation est influente, nous pouvons analyser la matrice de projection  $H$ . En effet, la construction des valeurs ajustées est basée sur cette matrice :

$$\hat{Y}_i = \sum_{j=1}^n h_{ij} Y_j = h_{ii} Y_i + \sum_{\substack{j=1 \\ j \neq i}}^n h_{ij} Y_j$$

Grâce à la relation ci-dessus, nous pouvons voir que si  $h_{ii} = 1$  alors la valeur ajustée  $\hat{Y}_i$  est entièrement déterminée par la  $i^{\text{ème}}$  observation. De plus, comme  $\text{Var}[\hat{\epsilon}_i] = \sigma^2(1 - h_{ii})$ , la variance du résidu est d'autant plus faible que  $h_{ii}$  est proche de 1, ce qui signifie également que la  $i^{\text{ème}}$  observation a complètement « attiré » la droite de régression.

Comme  $\text{Tr}(H) = p + 1$ , la moyenne des  $h_{ii}$  est  $\bar{h} = \frac{p+1}{n}$ . Une règle empirique consiste à dire que la  $i^{\text{ème}}$  valeur est influente si  $h_{ii} > 2\bar{h}$ . Le cas idéal correspond à celui d'une influence équidistribuée, c'est-à-dire si  $\forall i = 1, \dots, n \quad h_{ii} = \bar{h}$ .

Nous pouvons également étudier la distance de Cook. Celle-ci est définie par :

$$\hat{C}_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{S^2(p + 1)}$$

La  $i^{\text{ème}}$  observation est considérée comme influente si  $\hat{C}_{i,obs} > f_{p+1, n-p-1; 1-\alpha}$ , où  $f_{p+1, n-p-1; 1-\alpha}$  désigne le quantile d'ordre  $1 - \alpha$  d'une loi de Fisher à  $p + 1$  et  $n - p - 1$  degrés de liberté.

## Annexe D : Rappels théoriques sur le modèle linéaire généralisé

### Estimation :

Une méthode « naturelle » d'estimation des paramètres  $\beta_1, \dots, \beta_p$  est l'estimation par maximum de vraisemblance.

Par indépendance des informations, on a

$$L(Y_1, \dots, Y_n) = \prod_{i=1}^n f_Y(y_i)$$

L'estimateur du maximum de vraisemblance vérifie :

$$\hat{\beta}_{MV} = \underset{\beta}{\operatorname{argmax}} L(Y, \beta) = \underset{\beta}{\operatorname{argmax}} l(Y, \beta)$$

Où  $l(Y, \beta)$  désigne la log-vraisemblance de l'échantillon.

Si l'on définit le vecteur de score :

$$S(Y, \beta) = \left( \frac{\partial}{\partial \beta_1} l(Y, \beta), \dots, \frac{\partial}{\partial \beta_p} l(Y, \beta) \right)$$

Alors l'EMV vérifie  $S(Y, \hat{\beta}_{MV}) = 0$ .

Les équations du score sont :

$$\sum_{i=1}^n \frac{y_i - \mu_i}{b''(\theta_i) a(\phi)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{pour tout } j = 0, \dots, p$$

Et il n'existe généralement pas de solution analytique. Une résolution par un algorithme de Newton-Raphson (utilisant la matrice Hessienne) ou la méthode des scores de Fisher (utilisant la matrice d'information de Fisher) ou est possible.

La matrice d'information de Fisher est la matrice de terme général :

$$I_{jk} = E \left[ - \frac{\partial^2 l}{\partial \beta_k \partial \beta_j} (Y, \beta) \right] = - \sum_{i=1}^n \frac{x_{ij} x_{ik}}{b''(\theta_i) a(\phi)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

Elle peut être écrite sous la forme :

$$I = X'WX$$

Avec  $W$  une matrice de pondération diagonale de terme général :

$$W_i = \frac{1}{b''(\theta_i) a(\phi)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

Dans le cas où le lien utilisé est canonique, nous avons

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

Et donc les équations du score se simplifient et deviennent :

$$\sum_{i=1}^n \frac{(y_i - \mu_i)}{a(\phi)} x_{ij} = 0 \quad \forall j = 0, \dots, p$$

Dans ce cas, nous avons également que  $\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}(Y, \beta)$  ne dépend pas de  $Y$  et donc la matrice hessienne et la matrice d'information de Fisher sont identiques. Ainsi les algorithmes des méthodes de Newton-Raphson et du score de Fisher sont les mêmes. Leur algorithme général est :

Initialisation :  $u^{(0)}$

Pour tout entier  $m$  :

$$u^{(m)} = u^{(m-1)} + [I^{(m-1)}]^{-1} S(Y, u^{(m-1)})$$

Arret quand :

$$|u^{(m-1)} - u^{(m)}| \leq \varepsilon$$

Sortie :  $\hat{\beta}_{MV} = u^{(m)}$

Nous pouvons remarquer qu'à chaque itération nous avons :

$$I^{(m-1)} u^{(m)} = I^{(m-1)} u^{(m-1)} + S(Y, u^{(m-1)})$$

$$\Leftrightarrow X'W_{(m-1)}Xu^{(m)} = X'W_{(m-1)}Xu^{(m-1)} + S(Y, u^{(m-1)})$$

Le membre de droite de cette équation est un vecteur de coordonnées :

$$\sum_{k=1}^p \sum_{i=1}^n \frac{x_{ij}x_{ik}}{b''(\theta_i)a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 u_k^{(m-1)} + \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{b''(\theta_i)a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

Il peut donc se réécrire  $X'Wz$  où :

$$z_i = \sum_{k=1}^p x_{ik} u_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^{-1}$$

Donc à chaque itération, l'algorithme calcule :

$$X'W_{(m-1)}Xu^{(m)} = X'W_{(m-1)}Z_{(m-1)}$$

Cette équation est une généralisation de l'équation normale établie dans le cadre du modèle linéaire.

**Remarque :** Nous pouvons constater que si nous utilisons une fonction de lien identité (cas du modèle linéaire généralisé gaussien) nous avons :

$$\mu_i = \eta_i = (X\beta)_i$$

Par conséquent :

$$\frac{\partial \mu_i}{\partial \eta_i} = 1$$

Donc la matrice de pondération  $W$  admet pour termes diagonaux :

$$W_i = \frac{1}{b''(\theta_i)a(\phi)} = \frac{1}{\text{Var}[Y_i]}$$

$$\Leftrightarrow W = \sigma^{-2}I_n$$

De plus, nous obtenons

$$z_i = \sum_{k=1}^p x_{ik}u_k^{(m-1)} + (Y_i - \mu_i)$$

$$\Leftrightarrow z_i = (X\beta^{(m-1)})_i + Y_i - \mu_i$$

$$\Leftrightarrow z_i = \mu_i + Y_i - \mu_i$$

$$\Leftrightarrow z_i = Y_i$$

Ainsi, l'équation générale établie précédemment devient :

$$\sigma^{-2}X'Xu = \sigma^{-2}X'Y$$

$$\Leftrightarrow u = (X'X)^{-1}X'Y$$

Nous retrouvons bien l'estimateur des moindres carrés obtenu dans le cadre du modèle linéaire.

#### Qualité d'ajustement :

Plusieurs critères permettent d'évaluer l'ajustement du modèle. Le premier critère est basé sur la déviance, définie par :

$$D = 2(\ln(L_{SAT}) - \ln(L)) = 2\ln(\lambda)$$

Où  $L_{SAT}$  et  $L$  désignent respectivement la vraisemblance du modèle saturé et la vraisemblance du modèle considéré, avec  $p + 1 < n$  paramètres. Le modèle saturé est le modèle qui compte autant d'observations que de paramètres à estimer. Pour un tel modèle, nous aurons donc :  $\hat{\mu}_i = y_i \quad \forall i = 1, \dots, n$ .

Intuitivement, si le modèle estimé décrit convenablement les données,  $L \approx L_{SAT}$ . Cela équivaut à  $\lambda \approx 1$  où  $\lambda$  est la statistique du rapport de vraisemblance :

$$\lambda = \frac{L_{SAT}}{L}$$

$D$  est la déviance réduite (ou normalisée) et  $D^* = \phi D$  la déviance non réduite. Comme  $D \sim \chi_{n-p-1}^2$ , le modèle estimé est considéré de mauvaise qualité si  $D_{obs} > \chi_{n-p-1;1-\alpha}^2$ . Une règle empirique consiste à dire que si  $\frac{D}{E[D]} = \frac{D}{n-p-1} \approx 1$  alors l'ajustement effectué par le modèle considéré est convenable.

Un autre critère de la qualité d'ajustement du modèle s'appuie sur la statistique de Pearson :

$$X^2 = \sum_{i=1}^n \frac{(Y_i - \mu_i)^2}{\text{Var}[Y_i]}$$

Comme pour la déviance, nous avons que  $X^2 \sim \chi_{n-p-1}^2$ , et la règle de décision quant à la qualité d'ajustement du modèle est la même.

#### Tests sur les paramètres du modèle :

De la même manière que dans le cadre du modèle linéaire général, il peut être intéressant d'effectuer plusieurs tests sur les paramètres du modèle estimé.

Pour tester l'existence d'un modèle emboîté, nous souhaitons tester :

$$H_0: \beta = (\beta_0, \beta_1, \dots, \beta_q) \text{ contre } H_1: \beta = (\beta_0, \beta_1, \dots, \beta_p) \text{ avec } p > q$$

Si l'on note  $D_0$  la déviance du modèle avec  $q + 1$  paramètres et  $D_1$  la déviance du modèle « complet » (avec  $p + 1$  paramètres) alors, sous  $H_0$  la statistique définie par :

$$\Delta = D_0 - D_1$$

Suit approximativement une loi du  $\chi^2$  à  $p - q$  degrés de libertés. Ainsi, l'hypothèse d'existence d'un sous-modèle est rejetée si  $\Delta_{obs} > \chi_{p-q;1-\alpha}^2$ . En effet, cela signifie que le rapport de vraisemblance entre le modèle complet et le sous-modèle est trop grand, cela signifie que l'ajout de variables apporte de l'information.

De façon plus générale, nous pouvons également effectuer des tests sur des combinaisons linéaires de paramètres, c'est-à-dire tester :

$$H_0: C\beta = r \text{ contre } H_1: C\beta \neq r$$

Où  $C$  est une matrice de dimension  $q \times (p + 1)$  et  $r$  un vecteur de dimension  $q$ . Différents tests sont alors possibles : le test du rapport de vraisemblance, le test de Wald et le test du score.

#### Test du rapport de vraisemblance :

Si nous notons  $\hat{L}$  la vraisemblance du modèle sans contrainte et  $\tilde{L}$  celle du modèle avec la contrainte, alors sous  $H_0$  la statistique :

$$2\ln(\lambda) = 2 \left( \ln(\hat{L}) - \ln(\tilde{L}) \right)$$

Suit une loi du  $\chi^2$  à  $q$  degrés de libertés. L'hypothèse nulle est rejetée si la valeur observée de la statistique dépasse le quantile d'ordre  $1 - \alpha$  d'une loi  $\chi^2_q$ .

#### Test de Wald :

Sous l'hypothèse nulle, la statistique :

$$(C\beta - r)'(CI^{-1}C')^{-1}(C\beta - r)$$

Suit asymptotiquement une loi du  $\chi^2$  à  $q$  degrés de libertés. L'hypothèse nulle est rejetée si la valeur observée de la statistique dépasse le quantile d'ordre  $1 - \alpha$  d'une loi  $\chi^2_q$ .

#### Sélection de modèle :

Une fois les différents modèles estimés, il convient de déterminer quel est celui « optimal ». La notion d'optimalité de modèle dépend notamment du type de critère utilisé. Nous en présenterons trois dans cette partie.

- Les coefficients d'ajustement

Déjà étudié dans le cadre du modèle linéaire général, ce coefficient est calculé comme suit :

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

Ce critère doit être le plus proche possible de 1, témoignant ainsi d'un bon ajustement du modèle. Toutefois, la sélection de modèle basée sur ce critère conduit souvent à un sur-ajustement. Nous préférons donc la version ajustée du coefficient de détermination :

$$R_a^2 = 1 - \frac{n-1}{n-p-1} (1 - R^2)$$

$R_a^2$  introduit une pénalisation du modèle étudié en fonction du nombre de paramètres, et permet donc un compromis entre l'adéquation aux données et la complexité du modèle.

- Le critère AIC

Le critère AIC permet quant à lui de minimiser l'écart entre le modèle estimé et le vrai modèle au sens de la dissemblance de Kullback. En notant  $m$  le nombre de paramètres du modèle estimé, le critère AIC est :

$$AIC(m) = \|Y - \hat{Y}_{(m)}\|^2 + 2\sigma_{(m)}^2|m| + n\sigma_{(m)}^2 \log(\sigma_{(m)}^2)$$

En posant  $\tilde{\sigma}_{(m)}^2 = \frac{\|Y - \hat{Y}_{(m)}\|^2}{n}$ , le critère peut se réécrire :

$$AIC(m) = \log(\tilde{\sigma}_{(m)}^2) + 2 \frac{|m|}{n}$$

Le modèle « optimal » est celui vérifiant :

$$\hat{m}_{AIC} = \underset{m \in M}{\operatorname{argmin}} AIC(m)$$

Le critère AIC fonctionne bien pour de petites collections de modèles, mais la qualité d'estimation a tendance à se dégrader lorsque  $m$  est trop grand. Pour éviter ce problème, il est possible d'utiliser le critère AIC corrigé :

$$AIC_c(m) = \log(\tilde{\sigma}_{(m)}^2) + \frac{n + |m| + 1}{n - |m| - 3}$$

Ce critère nous permet de sélectionner un modèle parcimonieux.

- Le critère BIC

Ce critère est une extension de l'AIC, mais le paramètre  $\beta$  n'est plus considéré comme un vecteur de  $R^k$  mais comme une variable aléatoire à valeurs dans  $R^k$ . Une loi *a priori* est donc placée sur  $\beta$  et cette information est par la suite exploitée pour l'estimation. Cela conduit au critère suivant :

$$BIC(m) = n \log(\hat{\sigma}_{(m)}^2) + \log(n) \times |m|$$

Et le modèle optimal au sens du BIC est :

$$\hat{m}_{BIC} = \underset{m \in M}{\operatorname{argmin}} BIC(m)$$

## **Annexe E : Démonstrations sur les GLM**

Proposition : l'estimateur des moindres carrés de  $\beta$  est donné par  $\hat{\beta} = (X'X)^{-1}X'Y$

Preuve :

On a  $\hat{\beta} = \operatorname{argmin}_{\beta} (Y - X\beta)'(Y - X\beta) = \operatorname{argmin}_{\beta} \Delta(\beta)$

Or :

$$\begin{aligned}\Delta(\beta) &= (Y' - \beta'X')(Y - X\beta) \\ \Delta(\beta) &= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta\end{aligned}$$

D'où

$$\frac{\partial \Delta}{\partial \beta}(\beta) = 2X'X\beta - 2X'Y$$

Et par conséquent

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Proposition : l'estimateur des moindres carrés de  $\beta$  est sans biais

Preuve :  $E[\hat{\beta}] = E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'X\beta = \beta$

Proposition : la variance de l'estimateur des moindres carrés est  $\operatorname{Var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$

Preuve : On a  $\hat{\beta} = AY = (X'X)^{-1}X'Y$

D'où  $\operatorname{Var}[\hat{\beta}] = A\sigma^2I_nA' = \sigma^2(X'X)^{-1}X'YI_nX(X'X)^{-1} = \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} = \sigma^2(X'X)^{-1}$

Théorème de Gauss-Markov : l'estimateur des moindres carrés  $\hat{\beta}$  est le meilleur estimateur parmi les estimateurs sans biais de  $\beta$

Preuve :

$$\text{Posons } \beta^* = A^*Y = (A + C)Y = ((X'X)^{-1}X' + C)Y = \hat{\beta} + CY$$

$$\text{On a } E[\beta^*] = E[\hat{\beta} + CY] = \beta + CE[Y] = \beta + CX\beta \text{ car } \hat{\beta} \text{ est sans biais.}$$

$$\text{D'où } E[\beta^*] = (I_p + CX)\beta \text{ où } I_p \text{ est la matrice identité de dimension } p.$$

$$\text{Donc l'estimateur } \beta^* \text{ est sans biais si et seulement si } (I_p + CX) = I_p \Leftrightarrow CX = 0$$

Supposons que l'on a  $CX = 0$ . Calculons la variance de  $\beta^*$  :

$$\text{Var}[\beta^*] = (A^*)\text{Var}[Y](A^*)'$$

$$\text{Var}[\beta^*] = \sigma^2(A + C)I_n(A' + C')$$

$$\text{Var}[\beta^*] = \sigma^2(AI_nA' + AI_nC' + CI_nA' + CI_nC')$$

$$\text{Var}[\beta^*] = \text{Var}[\hat{\beta}] + \sigma^2(AC' + CA' + CC')$$

Démontrons que la matrice  $AC' + CA' + CC'$  est positive :

$$AC' = (X'X)^{-1}X'C' = (X'X)^{-1}(CX)' = 0 \text{ car } CX = 0 \text{ par hypothèse}$$

$$\text{De même, } CA' = (AC')' = 0$$

De plus, la matrice symétrique  $CC'$  est positive.

Par conséquent :

$$\text{Var}[\beta^*] = \text{Var}[\hat{\beta}] + \sigma^2CC' > \text{Var}[\hat{\beta}]$$

Proposition : les résidus de la régression définis par  $\hat{\epsilon} = Y - \hat{Y}$  sont centrés et de variance  $\sigma^2 M$ , avec  $M = I_n - H = I_n - X(X'X)^{-1}X'$ .  $M$  est une matrice idempotente.

Preuve : Posons  $H = X(X'X)^{-1}X'$  de telle sorte que  $\hat{Y} = HY$ .

On a  $\hat{\epsilon} = Y - HY = (I_n - H)Y = MY$

Par conséquent,  $E[\hat{\epsilon}] = E[MX\beta] + E[M\epsilon] = E[MX\beta]$  car les erreurs sont centrées.

D'où :

$$E[\hat{\epsilon}] = MX\beta = (I_n - H)X\beta$$

$$E[\hat{\epsilon}] = X\beta - X(X'X)^{-1}X'X\beta$$

$$E[\hat{\epsilon}] = X\beta - X\beta$$

$$E[\hat{\epsilon}] = 0$$

Et :

$$\text{Var}[\hat{\epsilon}] = \text{Var}[(I_n - H)Y]$$

$$\text{Var}[\hat{\epsilon}] = \text{Var}[MY]$$

$$\text{Var}[\hat{\epsilon}] = M'\sigma^2 I_n M$$

$$\text{Var}[\hat{\epsilon}] = \sigma^2 M'M$$

$$\text{Var}[\hat{\epsilon}] = \sigma^2 M$$

Proposition :  $S^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-p-1}$  est un estimateur sans biais de  $\sigma^2$ . De plus,  $S^2 \sim \sigma^2 \chi_{n-p-1}^2$

Preuve :

On a  $\hat{\epsilon}'\hat{\epsilon} = (MY)'(MY) = Y'MMY = Y'MY$  car  $M$  est idempotente

Or :

$$Y'MY = (X\beta + \epsilon)'M(X\beta + \epsilon)$$

$$Y'MY = (\beta'X' + \epsilon')M(X\beta + \epsilon)$$

$$Y'MY = \beta'X'MX\beta + \beta'X'M\epsilon + \epsilon'MX\beta + \epsilon'M\epsilon$$

$$Y'MY = \beta'X'M\epsilon + \epsilon'MX\beta + \epsilon'M\epsilon$$

Ainsi :

$$E[\hat{\epsilon}'\hat{\epsilon}] = E[\epsilon'M\epsilon]$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = E[\text{Tr}(\epsilon'M\epsilon)]$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = \text{Tr}(E[\epsilon' M \epsilon])$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = \text{Tr}(E[\epsilon' \epsilon] M)$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = \text{Tr}(\sigma^2 M)$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = \sigma^2 \times (\text{Tr}(I_n) - \text{Tr}(H))$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = \sigma^2 n - \sigma^2 \times \text{Tr}(X(X'X)^{-1}X')$$

$$E[\hat{\epsilon}'\hat{\epsilon}] = \sigma^2 \times (n - p)$$

On a alors :

$$E\left[\frac{\hat{\epsilon}'\hat{\epsilon}}{(n - p - 1)}\right] = \sigma^2$$

Donc  $S^2$  est un estimateur sans biais de la variance des erreurs.

De plus, comme  $\hat{\epsilon} \sim N(0, \sigma^2(n - p - 1))$ , alors  $\frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2(n - p - 1)} \sim \chi_{n-p-1}^2$ .

Proposition : La statistique  $\hat{F}$  définie par :

$$\hat{F} = \frac{(SCR_0 - SCR)/(p - p_0)}{SCR/(n - p - 1)}$$

Suit une loi de Fisher à  $p - p_0$  et  $n - p - 1$  degrés de liberté.

Preuve : Nous avons, pour le modèle complet :

$$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Où  $Y \sim N(X\beta, \sigma^2 I_n)$  et  $E[\hat{Y}] = X\beta$ .

Par conséquent,  $Y - \hat{Y} \sim N(0, \sigma^2 I_n)$ , et  $\frac{Y - \hat{Y}}{\sigma} \sim N(0, 1)$ .

D'où :

$$\sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

De même pour le sous-modèle :

$$\sum_{i=1}^n \frac{(Y_i - \hat{Y}_{0,i})^2}{\sigma^2} \sim \chi_{n-p_0}^2$$

Ainsi,  $SCR_0 - SCR \sim \chi_{p-p_0}^2$  et par suite, nous obtenons que  $\hat{F} \sim F_{p-p_0, n-p-1}$ .

Proposition : Sous l'hypothèse  $H_0: C'\beta = r$ , la statistique :

$$\hat{T} = \frac{C'\hat{\beta} - r}{S\sqrt{C'(X'X)^{-1}C}}$$

Suit une loi de Student à  $n - p - 1$  degrés de libertés.

Preuve :

Nous avons  $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$  donc  $C'\hat{\beta} \sim N(C'\beta, \sigma^2 C'(X'X)^{-1}C)$ . Sous  $H_0$ ,  $C'\beta = r$ , par conséquent :

$$\frac{C'\hat{\beta} - r}{\sigma\sqrt{C'(X'X)^{-1}C}}$$

Cependant,  $\sigma$  est inconnu. Ce dernier est estimé par :

$$S^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n - p - 1}$$

Et

$$\frac{S^2}{\sigma^2}(n - p - 1) \sim \chi_{n-p-1}^2$$

Nous avons donc que :

$$\frac{\frac{C'\hat{\beta} - r}{\sigma\sqrt{C'(X'X)^{-1}C}}}{\sqrt{\frac{S^2(n - p - 1)}{\sigma^2(n - p - 1)}}} = \frac{C'\hat{\beta} - r}{S\sqrt{C'(X'X)^{-1}C}} \sim T_{n-p-1}$$

Proposition : sous l'hypothèse  $H_0: C'\beta = 0$  où  $C$  est une matrice à  $p$  lignes et  $m$  colonnes, nous avons :

$$\hat{F} = \frac{\hat{\beta}'C(C'(X'X)^{-1}C)^{-1}C'\hat{\beta}}{mS^2}$$

Suit une loi de Fisher à  $m$  et  $n - p - 1$  paramètres.

Preuve :

Nous avons  $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$  donc  $C'\hat{\beta} \sim N(C'\beta, \sigma^2 C'(X'X)^{-1}C)$ . Sous  $H_0$  :

$$\frac{C'\hat{\beta}}{\sigma\sqrt{C'(X'X)^{-1}C}} \sim N(0,1)$$

Et par conséquent :

$$\frac{(C'\hat{\beta})'(C'\hat{\beta})}{\sigma^2 C'(X'X)^{-1}C} = \frac{\hat{\beta}'C(C'(X'X)^{-1}C)^{-1}C'\hat{\beta}}{\sigma^2} \sim \chi_m^2$$

Or,  $\sigma^2$  étant inconnu, il est estimé par :

$$S^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-p-1}$$

Et nous avons :

$$\frac{S^2}{\sigma^2}(n-p-1) \sim \chi_{n-p-1}^2$$

Ainsi nous obtenons :

$$\frac{\frac{\hat{\beta}'C(C'(X'X)^{-1}C)^{-1}C'\hat{\beta}}{m\sigma^2}}{\frac{S^2}{\sigma^2(n-p-1)}(n-p-1)} = \frac{\hat{\beta}'C(C'(X'X)^{-1}C)^{-1}C'\hat{\beta}}{mS^2} \sim F_{m,n-p-1}$$

Proposition : l'erreur de prévision  $\epsilon_{n+1}^p$  vérifie :

$$E[\epsilon_{n+1}^p] = 0 \text{ et } \text{Var}[\epsilon_{n+1}^p] = \sigma^2(1 + X'_{n+1}(X'X)^{-1}X_{n+1})$$

Preuve :

$$E[\epsilon_{n+1}^p] = E[Y_{n+1} - \hat{Y}_{n+1}^p]$$

$$E[\epsilon_{n+1}^p] = E[Y_{n+1} - X_{n+1}\hat{\beta}]$$

$$E[\epsilon_{n+1}^p] = X_{n+1}\beta - X_{n+1}E[\hat{\beta}]$$

Et comme  $\hat{\beta}$  est un estimateur sans biais de  $\beta$  il vient :

$$E[\epsilon_{n+1}^p] = 0$$

Nous avons également que :

$$\text{Var}[\epsilon_{n+1}^p] = \text{Var}[Y_{n+1} - \hat{Y}_{n+1}^p]$$

$$\text{Var}[\epsilon_{n+1}^p] = \text{Var}[Y_{n+1}] + \text{Var}[\hat{Y}_{n+1}^p] - 2\text{Cov}(Y_{n+1}, \hat{Y}_{n+1}^p)$$

Avec :

$$\text{Var}[Y_{n+1}] = \sigma^2 I_n$$

$$\text{Var}[\hat{Y}_{n+1}^p] = \sigma^2 X_{n+1} (X'X)^{-1} X_{n+1}'$$

$$\text{Cov}(Y_{n+1}, \hat{Y}_{n+1}^p) = E[Y_{n+1} \hat{Y}_{n+1}^p] - E[Y_{n+1}] E[\hat{Y}_{n+1}^p] = (X_{n+1} \beta)^2 + X_{n+1} \beta E[\epsilon_{n+1}] - (X_{n+1} \beta)^2 = 0$$

D'où le résultat.

Proposition : si  $Y$  a une distribution appartenant à la famille exponentielle, alors  $E[Y] = b'(\theta)$

Preuve :

La densité de  $Y$  peut se mettre sous la forme :

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

Soit

$$l_Y(y) = \ln(f_Y(y)) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)$$

Alors :

$$\frac{\partial l_Y}{\partial \theta}(y) = \frac{y - b'(\theta)}{a(\phi)} = \frac{1}{f_Y(y)} \times \frac{\partial f_Y}{\partial \theta}(y)$$

De plus,

$$E\left[\frac{\partial l_Y}{\partial \theta}\right] = \int \frac{\partial l_Y}{\partial \theta}(y) \times f_Y(y) dy$$

$$E[Y] = \int \frac{\partial f_Y}{\partial \theta}(y) dy$$

$$E[Y] = \frac{\partial}{\partial \theta} \int f_Y(y) dy$$

$$E[Y] = 0$$

Or, nous avons

$$E\left[\frac{\partial l_Y}{\partial \theta}\right] = E\left[\frac{Y - b'(\theta)}{a(\phi)}\right]$$

D'où finalement :

$$E[Y] = b'(\theta)$$

Proposition : Si  $Y$  a une densité appartenant à la famille exponentielle alors  $Var[Y] = b''(\theta)a(\phi)$

Preuve :

La densité de  $Y$  peut se mettre sous la forme :

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

Soit

$$l_Y(y) = \ln(f_Y(y)) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)$$

Alors :

$$\frac{\partial l_Y}{\partial \theta}(y) = \frac{y - b'(\theta)}{a(\phi)} = \frac{1}{f_Y(y)} \times \frac{\partial f_Y}{\partial \theta}(y)$$

Et

$$\frac{\partial^2 l_Y}{\partial \theta^2}(y) = \frac{-b''(\theta)}{a(\phi)} = -\frac{1}{f_Y^2(y)} \times \frac{\partial f_Y}{\partial \theta}(y) + \frac{1}{f_Y(y)} \times \frac{\partial^2 f_Y}{\partial \theta^2}(y)$$

Nous avons

$$E\left[\frac{\partial^2 l_Y}{\partial \theta^2}(y)\right] = \int -\frac{1}{f_Y(y)} \times \frac{\partial f_Y}{\partial \theta}(y) dy + \int \frac{\partial^2 f_Y}{\partial \theta^2}(y) dy$$

$$E\left[\frac{\partial^2 l_Y}{\partial \theta^2}(y)\right] = -\int \frac{1}{f_Y(y)} \times \frac{\partial f_Y}{\partial \theta}(y) dy$$

$$E\left[\frac{\partial^2 l_Y}{\partial \theta^2}(y)\right] = -\int \frac{1}{f_Y^2(y)} \times \left(\frac{\partial f_Y}{\partial \theta}(y)\right)^2 f_Y(y) dy$$

$$E\left[\frac{\partial^2 l_Y}{\partial \theta^2}(y)\right] = -\int \left(\frac{1}{f_Y(y)} \times \frac{\partial f_Y}{\partial \theta}(y)\right)^2 dy$$

$$E\left[\frac{\partial^2 l_Y}{\partial \theta^2}(y)\right] = -E\left[\left(\frac{\partial l_Y}{\partial \theta}(y)\right)^2\right]$$

D'où en remplaçant les dérivées par leurs expressions :

$$-\frac{b''(\theta)}{a(\phi)} = -E\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right]$$

Or, comme nous avons  $E[Y] = b'(\theta)$ , il vient :

$$\frac{b''(\theta)}{a(\phi)} = \frac{\text{Var}[Y]}{a^2(\phi)}$$

D'où le résultat :

$$\text{Var}[Y] = b''(\theta)a(\phi)$$

Proposition : pour un vecteur  $Y$  dont la densité appartient à la famille exponentielle, les équations du score sont données par :

$$\sum_{i=1}^n \frac{y_i - \mu_i}{b''(\theta_i)a(\phi)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{pour tout } j = 1, \dots, p$$

Preuve :

La vraisemblance de l'échantillon s'écrit :

$$L(Y_1, \dots, Y_n) = \prod_{i=1}^n f_Y(y_i)$$

$$L(Y_1, \dots, Y_n) = \exp \left\{ \sum_{i=1}^n \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

Donc la log-vraisemblance de  $Y$  est :

$$l(Y_1, \dots, Y_n) = \sum_{i=1}^n \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) = \sum_{i=1}^n l_i(y)$$

Nous avons :

$$\frac{\partial l_i}{\partial \beta_j}(Y) = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}$$

Où :

$$\frac{\partial l_i}{\partial \theta_i} = \frac{\theta_i - b'(\theta_i)}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)} \quad \text{car } \mu_i = b'(\theta_i)$$

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

Donc nous obtenons

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{\theta_i - b'(\theta_i)}{a(\phi)b''(\theta_i)} x_{ij} \times \frac{\partial \mu_i}{\partial \eta_i}$$

Définition : Dissemblance de Kullback

La dissemblance de Kullback entre deux modèles  $m^*$  et  $m$  est définie par :

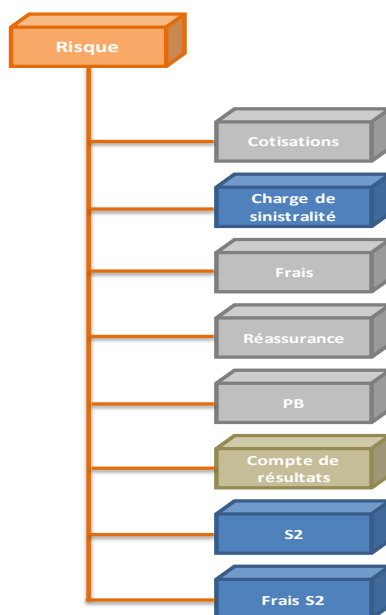
$$K(m^*, m) = \frac{n}{2} \left( \log \left( \frac{\sigma_{(m)}^{2*}}{\sigma_{(m)}^2} \right) + \frac{\sigma_{(m)}^{2*}}{\sigma_{(m)}^2} - 1 \right) + \frac{1}{2\sigma_{(m)}^2} \|\mu^* - \mu_{(m)}\|^2$$

Il s'agit d'une mesure de l'écart entre la loi du vrai modèle ( $m^*$ , inconnu) et celle du modèle étudié noté  $m$ .

## Annexe F : Présentation du modèle prévoyance-santé standard d'ACTUARIS®

### 1. Passif

La partie Passif est composée de plusieurs modèles principales qui correspondent aux garanties évoquées précédemment. Ces modèles disposent tous de la même architecture présentée ci-après :



*Architecture commune des modèles de la partie Passif*

Le modèle principal associé à la garantie est représenté en orange, les sous-modèles communs à chaque modèle principal en gris et les autres sous-modèles spécifiques à la garantie d'assurance sont représentés en bleus.

Le sous-modèle compte de résultats est différent selon que la garantie soit de l'assurance vie ou de l'assurance non-vie.

#### a. Cotisations

Ce sous-modèle permet de projeter l'évolution des cotisations et des provisions associées (PPNA et PANE). Le nombre de cotisants et la cotisation individuelle sont calculées à l'aide de chroniques de taux renseignées par l'utilisateur, puis le montant des cotisations appelées est déterminé en multipliant la cotisation individuelle par le nombre de cotisants. Enfin, les PPNA et PANE sont déterminées à l'aide de chroniques de taux appliquées au montant de cotisations appelées et les montants de cotisations émises et acquises sont obtenus en retranchant respectivement les variations de la PANE et de la PPNA.

#### b. Charge de sinistralité (Santé et Décès)

Ce sous-modèle permet de projeter l'évolution des provisions et des prestations. Les charges de prestations futures sont déterminées en multipliant les ratios S/P entrés par l'utilisateur par les

cotisations acquises. Les PSAP et les IBNR sont déterminés par année de survenance en appliquant la chronique de taux de liquidation renseignée par l'utilisateur à la charge de sinistre.

### c. Frais

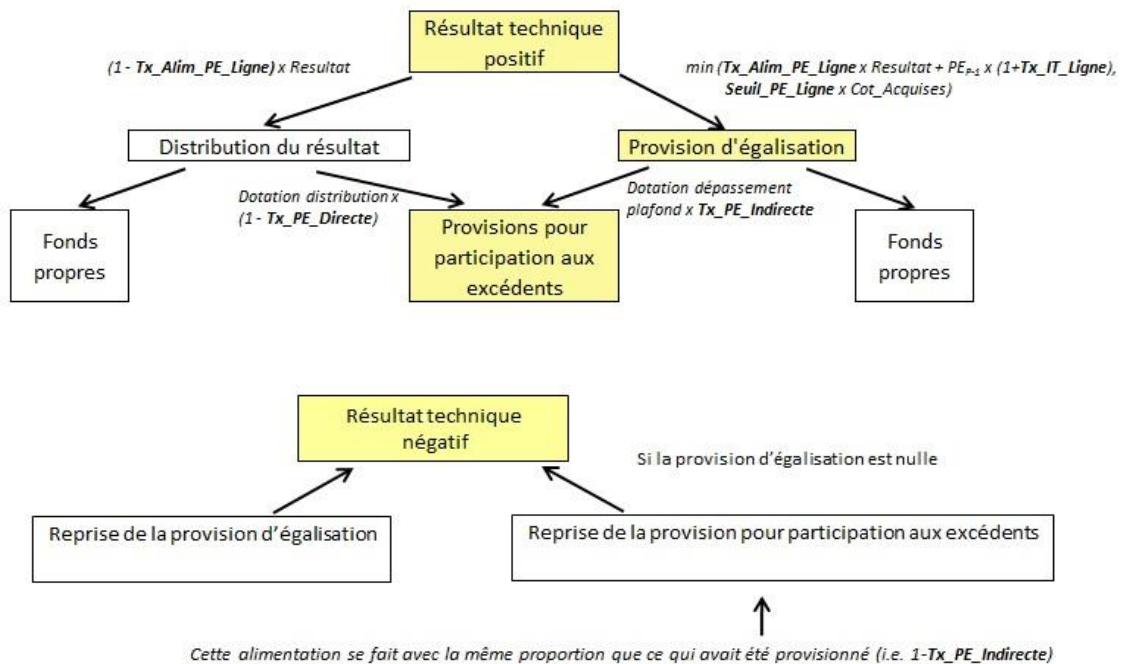
Ce sous-modèle permet de projeter l'évolution des frais comptables. Ces frais sont calculés à partir de chroniques de taux renseignés par l'utilisateur appliqués aux montants de prestations, de provisions et de cotisations.

### d. Réassurance

Ce sous-modèle permet de projeter l'évolution des montants cédés (provisions, prestations, cotisations) dans le cas de la réassurance en quote-part. Les tables contenant les caractéristiques des traités de réassurance passés et futurs sont renseignées par l'utilisateur. Les montants cédés sont calculés en appliquant les taux de cessions aux montants bruts des provisions, des prestations et des cotisations.

### e. Participations aux bénéficies

Ce sous-modèle permet de projeter le résultat technique et l'évolution des provisions d'égalisation. Tout d'abord les intérêts perçus sur les provisions de l'année précédente sont calculés en appliquant le taux technique renseigné par l'utilisateur à la demi-somme des provisions de clôture et d'ouverture. Le résultat technique est obtenu en ajoutant les cotisations acquises et les intérêts perçus sur les provisions de l'année précédente et en retranchant les prestations versées et les variations de provisions. Enfin la PEg et la PPE sont dotées des intérêts perçus sur les montants de l'année précédente, puis alimentées selon des taux renseignés par l'utilisateur comme le montre les schémas suivants :



Schémas d'alimentation de la PEg et la PPE

## **f. Compte de résultats**

Ce sous-modèle permet de projeter l'évolution des comptes techniques Vie et Non Vie au cours du temps. Les postes des comptes de résultats Vie et Non Vie sont complétés par les montants calculés dans les sous-modèles précédents. Par ailleurs le *Cash-Flow* de la garantie et le ratio S/P observé sont calculés.

## **g. Solvabilité 2**

Ce sous-modèle permet de calculer les différents chocs et de déterminer les provisions *Best Estimate* (brutes, cédées et nettes). Nous allons décrire le socle commun à l'ensemble des garanties d'assurance.

Dans un premier temps, les *BE* bruts (standards, choqués et sans actualisation) sont calculés par année de survenance afin d'obtenir les *BE* cédés en appliquant le taux de cession du traité de réassurance correspondant.

Ensuite les *BE* bruts et cédés de l'entité sont calculés selon la formule suivante :

$$BE = BE \text{ Sinistres Passés} + BE \text{ Sinistres Futurs} - \\ BE \text{ Cotisations} + BE \text{ Provision d'égalisation} + \\ BE \text{ Normes Actuelles}$$

- Les *BE* de Sinistres Passés et Futurs sont restituées à partir des *BE* calculés par survenance
- Les *BE* de Cotisation sont obtenus en considérant les cotisations qui font l'objet d'un engagement au 31/12/N
- Les *BE* de Normes actuelles correspondent à la somme des provisions S1 sans actualisation
- Le *BE* de Provision d'égalisation est la somme de la PPE et de la PEg en normes actuelles.

Enfin les *BE* sont regroupés par LoB dans différentes tables afin d'alimenter les calculs de la partie Solvabilité 2.

## **h. Frais Solvabilité 2**

Ce sous-modèle permet de calculer les *Best Estimate* de Frais. Son fonctionnement est similaire au sous-modèle Frais.

### **2. GSE et Actifs**

Le Générateur de Scénarios Économiques modélise l'environnement financier dans lequel se trouve l'organisme d'assurance modélisé et permet la projection de différents indices du marché. Le module actif permet de faire évoluer le portefeuille d'actifs de l'entité. La modélisation dépend de la classe d'actifs considérée (actions, obligations, immobilier, etc.)

### **3. Actif S2**

Cette brique permet de déterminer l'impact des différents chocs du SCR marché sur les placements en recalculant la valeur de marché dans chaque scénario

#### **4. Comptes**

Les modèles de la partie Comptes permettent de projeter le Bilan, le Compte de Résultats et le Résultat Fiscal. Les postes sont initialisés en fonction des hypothèses fournies par l'utilisateur. Comme nous l'avons vu, certains postes évoluent grâce au modèle au cours du temps. Les postes non modélisés sont quant à eux repris à l'identique par rapport aux paramètres fournis en input du modèle.

#### **5. Solvabilité 2**

Cette brique permet de calculer le SCR, le MCR et la risk margin de l'entité modélisée. Des calculs solvabilité 2 sont faits au sein de chaque garantie, et tous sont repris et agrégés au sein de ce module afin de calculer l'exigence en capital de l'organisme.

## ***Annexe G : Tableaux de sélections de variables***

**Tableau 1 : Sélection de variables pour l'estimation du SCR de défaut de type 1**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires				x	x	x	x	x	x	x	x	x		9
Coefficient d'évolution du CA (Santé)	x	x	x					x	x	x		x		7
Coefficient d'évolution du CA (Arrêt de travail)												x		1
Coefficient d'évolution du CA (Décès)	x	x	x					x	x	x		x		7
Proportion de garçons (RE)	x	x	x					x	x					5
Âge moyen (RC)	x													1
Âge moyen (AT)					x	x								2
Dispersion de l'âge (RE)								x						1
Taux de frais (RC)	x	x												2
Taux de frais (RE)		x						x						2
Taux de réassurance (Santé)	x	x	x				x	x	x	x	x			8
Taux de réassurance (Décès)	x	x	x				x	x	x	x	x			8
Taux de réassurance (RE)	x	x	x				x		x	x	x			7
Taux de réassurance (RC)	x	x	x				x		x	x	x			7
BE Life	x	x	x				x	x	x	x	x	x	x	10
BE NSLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Répartition des actifs		x												1
Ratio AT/MGDC		x	x	x	x	x		x			x		x	8
Ancienneté moyenne (Incapacité)		x						x						2
Ancienneté moyenne (Invalidité)								x						1
Duration de l'actif								x						1
Duration du passif										x	x			2
Duration moyenne (incapacité)		x	x					x	x		x		x	6
Duration moyenne (invalidité)					x	x					x			3
Ancienneté moyenne (incapacité)		x												1
Capitaux sous risques (Life CAT)		x	x	x	x	x		x						6
Coefficient d'absorption par le FDB (SLT Mortalité)		x												1
Taux de réassurance vie (AT)										x				1
Répartition du chiffre d'affaires (Rige et Elastic Net seulement)							x		x					2

**Tableau 2 : Sélection de variables pour l'estimation du SCR de défaut de type 2**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires				x	x	x	x	x	x	x	x	x	x	10
Coefficient d'évolution du CA (Santé)												x		1
Coefficient d'évolution du CA (Arrêt de travail)												x		1
Coefficient d'évolution du CA (RE)												x		1
Coefficient d'évolution du CA (RC)												x		1
Coefficient d'évolution du CA (Décès)												x		1
Créances réassureurs	x	x	x				x	x	x	x				7
S/P (RC)	x	x												2
Taux de réassurance (Vie - Arrêt de travail)												x		1
Âge moyen (RE)		x												1
Dispersion de l'âge (RC)											x		x	2
BE Life		x	x	x	x	x	x	x	x	x	x	x	x	12
BE NSLT											x	x	x	3
BE SLT	x	x	x	x	x	x	x	x			x	x	x	11
Duration de l'actif									x					1
Poids des obligations convertibles	x	x												2
Duration moyenne des OTF		x												1
Duration moyenne des OATI	x	x												2
Duration du passif	x	x							x		x	x	x	6
Duration moyenne (incapacité)		x	x		x	x		x		x				6
Capitaux sous risques (Life CAT)		x	x		x	x	x	x			x		x	8
Coefficient d'absorption par le FDB (risque de défaut)	x	x												2
Coefficient d'absorption par le FDB (Life CAT)	x	x												2
Nombre de personnes en invalidité permanente (Health CAT)	x	x												2
Exposition soins médicaux (Health CAT - Scénario concentration)	x													1
Taux de PANE (Santé)	x	x	x				x	x	x	x				7
Taux de PANE (AT)	x	x	x				x	x	x	x				7
Taux de PANE (RC)	x	x												2
Taux de PANE (Décès)	x	x	x				x	x	x	x				7
Taux de PPNA (AT)	x	x												2
Taux de PPNA (Décès)	x	x												2
Répartition du chiffre d'affaires (Rige et Elastic Net seulement)									x					1

**Tableau 3 : Sélection de variables pour l'estimation du SCR de défaut**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires		x		x	x	x	x	x		x	x	x	x	10
Créances réassureurs	x	x								x		x		4
Ratio AT/MGDC							x				x		x	3
Proportion d'hommes (RC)		x	x											2
Proportion de garçons (RE)	x	x							x					3
Dispersion de l'âge (RC)											x		x	2
Âge moyen (RC)				x	x	x								3
BE Life	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x	x	x	x	x	x	x	x	x		x		x	11
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Duration moyenne des OAT		x												1
Duration moyenne des OATi		x												1
Duration du passif					x					x		x	x	5
Duration moyenne (incapacité)								x	x		x		x	4
Duration moyenne (invalidité)		x	x		x	x	x				x		x	7
Ancienneté moyenne (invalidité)		x	x		x	x	x							5
Capitaux sous risques (Life CAT)									x		x		x	3
Coefficient d'absorption par le FDB (risque de défaut)	x	x												2
Taux de PANE (Santé)	x	x	x					x		x		x		6
Taux de PANE (AT)	x	x								x		x		4
Taux de PANE (Décès)	x	x	x				x	x	x	x	x	x	x	10
Taux de PPNA (Décès)	x													1
Répartition du chiffre d'affaires (Rige et Elastic Net seulement)							x		x					2

**Tableau 4 : Sélection de variables pour l'estimation du SCR mortalité (Life)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires		x	x	x	x	x	x	x	x	x	x	x	x	12
Proportion d'hommes (RC)											x			1
Âge moyen (RC)											x			1
Âge moyen (RE)											x			1
Âge moyen (AT)		x												1
Dispersion de l'âge (RE)				x	x	x					x		x	5
Dispersion de l'âge (RC)				x	x	x					x			4
Dispersion de l'âge (Incapacité)		x												1
Taux de frais (RE)	x	x												2
BE Life	x	x		x	x	x	x	x	x	x	x	x	x	12
BE NSLT	x	x	x	x	x			x			x		x	8
BE SLT	x		x	x	x	x			x	x	x	x	x	10
Poids des obligations		x												1
Poids du monétaire		x												1
Ratio AT/MGDC	x	x	x			x		x	x		x		x	8
Ancienneté moyenne (Invalidité)	x	x								x		x		4
Duration de l'actif		x												1
Duration du passif	x	x	x					x	x	x	x	x		8
Duration moyenne (incapacité)		x												1
Capitaux sous risques (Life CAT)	x	x	x				x	x	x	x	x		x	9
Nombre de personnes en incapacité d'au plus 12 mois (Health CAT)	x	x	x					x						4
Exposition concentration invalidité permanente (Health CAT)										x				1
Taux de réassurance(RE)	x		x											2
Taux de PPNA (Décès)	x													1
Répartition du chiffre d'affaires (Rige et Elastic Net seulement)							x		x					2

**Tableau 5 : Sélection de variables pour l'estimation du SCR longévité (Life)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires			x				x		x	x	x	x	6
Coefficient d'évolution du CA (RC)	x	x	x				x	x	x	x	x	x	9
Proportion d'hommes (RC)		x					x	x	x	x			5
Âge moyen (RC)	x	x	x	x	x	x	x	x	x	x	x	x	12
Dispersion de l'âge (RC)									x	x			2
Taux de frais (AT)	x												1
Taux de frais (RC)	x												1
Taux de frais (Décès)								x					1
BE Life	x	x	x						x	x	x	x	7
BE NSLT	x	x	x	x	x	x		x		x		x	9
BE SLT	x	x	x	x	x	x		x		x		x	9
Ratio AT/MGDC	x	x	x										3
Ancienneté moyenne (Incapacité)		x											1
Poids des actions		x											1
Poids des obligations		x											1
Poids de l'immobilier		x											1
Poids du monétaire		x											1
Poids des OPCVM		x											1
Duration du passif									x	x			2
Duration moyenne (incapacité)					x								1
Capitaux sous risques (Life CAT)					x								1
Exposition pour le risque d'invalidité permanente (scénario accident de masse - Health CAT)	x	x											2
Exposition invalidité d'au plus 10 ans (Scénario Concentration -Health CAT)	x	x											2
Coefficient d'absorption par le FDB (Risque de longévité - Life)	x	x											2
Taux de PPNA (Santé)		x											1
Taux de PPNA (RC)	x	x	x				x						4
Taux de PPNA (Décès)	x												1
Répartition du chiffre d'affaires (Rige et Elastic Net seulement)							x						1

**Tableau 6 : Sélection de variables pour l'estimation du SCR Morbidité (Life)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires										x	x	x	x	4
Coefficient d'évolution du CA (AT)	x	x	x				x	x	x	x		x		8
Coefficient d'évolution du CA (RE)	x	x												2
Âge moyen (AT)		x												1
Dispersion de l'âge (RE)											x			1
Dispersion de l'âge (RC)											x		x	2
BE Life										x	x	x	x	4
BE NSLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Ratio AT/MGDC	x	x	x	x	x	x		x		x		x		9
Ancienneté moyenne (Incapacité)	x	x	x											3
Ancienneté moyenne (Invalidité)		x												1
Duration du passif										x	x	x	x	4
Duration moyenne (incapacité)	x	x	x											3
Capitaux sous risques (Life CAT)	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Exposition soins médicaux (Health CAT)	x		x											2
Exposition incapacité d'au plus 12 mois (Health CAT - Scénario concentration)			x											1
Nombre de cotisants (Health CAT - Scénario Pandémie)			x											1
Exposition invalidité d'au plus 10 ans (Health CAT - Scénario concentration)			x											1
Taux de PPNA (RC)			x											1
Répartition du chiffre d'affaires (Rige et Elastic Net seulement)							x		x					2

**Tableau 7 : Sélection de variables pour l'estimation du SCR de frais (Life)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC (idem Step BIC)	Régression RIDGE	Régression LASSO	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
S/P (AT)	x	x	x										3
Chiffre d'affaires	x	x				x	x		x	x	x	x	8
Coefficient d'évolution du CA (AT)	x	x	x										3
Coefficient d'évolution du CA (Décès)	x	x	x										3
Âge moyen (AT)		x	x					x					3
Âge moyen (RC)										x		x	2
Âge moyen (RE)										x			1
Dispersion de l'âge (RC)						x				x			2
Dispersion de l'âge (RE)										x		x	2
Proportion d'hommes (RC)										x		x	2
Proportion de garçons (RE)										x		x	2
Dispersion de l'ancienneté (invalidité)					x								1
Ratio AT/MGDC	x	x	x	x	x		x	x					7
Taux de frais (AT)	x	x	x					x	x		x		6
Taux de frais (RC)	x	x	x						x		x		5
Taux de frais (Décès)									x		x		2
Duration du passif	x	x	x					x	x	x	x	x	8
Capitaux sous risques (Life CAT)	x	x	x	x	x	x	x	x	x	x	x	x	12
BE Life	x	x	x			x	x	x	x	x	x	x	10
BE SLT	x	x	x	x	x		x			x		x	8
BE NSLT	x	x	x	x	x		x			x		x	8
Taux de réassurance (RE)	x												1
Nombre de personnes en incapacité d'au plus 12 mois (Health CAT)		x	x										2
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)						x		x					2

**Tableau 8 : Sélection de variables pour l'estimation du SCR révision (Life)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (AT)	x	x	x				x	x	x	x		x		8
Coefficient d'évolution du chiffre d'affaires (Santé)	x	x	x					x	x					5
Chiffre d'affaires										x	x	x	x	4
Ratio AT/MGDC	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Dispersion de l'âge (AT)						x								1
Ancienneté moyenne (incapacité)					x									1
Ancienneté moyenne (invalidité)		x	x		x			x	x					5
Dispersion de l'ancienneté (invalidité)					x			x	x					3
Duration moyenne (incapacité)					x			x	x					3
Duration moyenne (invalidité)									x	x		x		3
Dispersion de l'âge (RC)				x	x	x	x				x		x	6
Dispersion de l'âge (RE)					x		x	x	x		x			5
Capitaux sous risques (Life CAT)	x	x	x		x	x	x	x	x	x	x	x	x	12
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE Life										x	x	x	x	4
Taux de PPNA (AT)									x					1
Taux de PPNA (RC)										x				1
Taux de frais (AT)									x	x				2
Taux de réassurance (AT - vie)									x					1
Duration du passif										x	x	x	x	4
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)							x		x					2

**Tableau 9 : Sélection de variables pour l'estimation du SCR Life**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC (idem step BIC)	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Chiffre d'affaires	x	x	x			x			x	x	x	x	8
Coefficient d'évolution du chiffre d'affaires (AT)	x	x	x				x	x	x		x		7
Coefficient d'évolution du chiffre d'affaires (RE)	x	x	x				x						4
Coefficient d'évolution du chiffre d'affaires (Décès)	x	x	x				x						4
Coefficient d'évolution du chiffre d'affaires (RC)									x				1
Dispersion de l'âge (RC)						x				x			2
Dispersion de l'âge (RE)						x				x			2
Âge moyen (RC)										x			1
Ratio MGDC/AT	x	x	x	x	x		x						6
Capitaux sous risque (Life CAT)	x	x	x	x	x	x		x	x	x	x	x	11
BE Life	x	x	x	x	x	x	x	x	x	x	x	x	12
BE NSLT	x	x	x				x		x		x		6
BE SLT	x	x	x				x	x	x	x	x	x	9
Duration du passif							x		x	x	x		4
Répartition du chiffre d'affaires (Ridge et elastic net seulement)						x		x					2

**Tableau 10 : Sélection de variables pour l'estimation du SCR longévité (Health SLT)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC (idem Step BIC)	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Répartition de l'actif	x	x											2
Ratio AT/MGDC	x	x	x			x	x	x	x	x	x	x	10
Dispersion de l'âge (Arrêt de travail)		x				x			x	x	x		5
Âge moyen (Arrêt de travail)						x		x	x	x			4
Duration du passif										x		x	2
Duration moyenne (incapacité)		x				x							2
Duration moyenne (invalidité)		x	x			x	x	x	x	x	x	x	9
Ancienneté moyenne (invalidité)				x	x	x		x	x	x	x		7
Ancienneté moyenne (incapacité)						x		x	x	x			4
Dispersion de l'ancienneté (invalidité)						x	x	x	x	x		x	6
Taux de frais (RC)	x	x											2
Taux de frais (RE)		x											1
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	12
BE NSLT	x	x	x	x	x		x	x	x	x	x	x	11
BE Life									x	x		x	3
Taux de PANE (Santé)	x	x											2
Capitaux sous risque (Life CAT)				x	x	x				x		x	5
Chiffre d'affaires										x		x	2

**Tableau 11 : Sélection de variables pour l'estimation du SCR frais (HSLT)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Ratio AT/MGDC	x	x	x	x	x	x	x	x			x		x	10
Âge moyen (Arrêt de travail)		x	x	x	x	x	x	x	x	x	x	x		11
Ancienneté moyenne (incapacité)				x	x		x	x	x	x	x			7
Ancienneté moyenne (invalidité)				x	x		x			x	x			5
Dispersion de l'ancienneté (incapacité)				x										1
Dispersion de l'ancienneté (invalidité)				x	x		x			x	x		x	6
Duration moyenne (incapacité)				x	x									2
Duration moyenne (invalidité)				x	x		x		x	x	x		x	7
Duration du passif											x		x	2
Taux de frais (Santé)	x													1
Taux de frais (Arrêt de travail)	x	x	x				x	x	x	x	x	x	x	10
Taux de frais (RC)	x	x					x		x					4
Taux de réassurance (RC)	x						x							2
Dispersion de l'âge (Arrêt de travail)							x							1
Dispersion de l'âge (RC)				x	x	x	x				x		x	6
Dispersion de l'âge (RE)				x	x	x	x				x			5
Chiffre d'affaires											x		x	2
Capitaux sous risque (Life CAT)	x	x	x				x	x	x		x		x	8
Nombre de personnes en invalidité d'au plus 10 ans (Helath CAT)	x	x	x				x	x						5
Nombre de personnes en invalidité permanente (Health CAT - Scénario Accident de masse)	x													1
Nombre de personnes en invalidité permanente (Health CAT - Scénario concentration)	x													1
BE NSLT	x	x	x	x	x	x	x	x	x		x		x	11
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE Life											x		x	2
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)							x		x					2

**Tableau 12 : Sélection de variables pour le SCR révision (HSLT)**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Taux de frais (RC)	x	x	x				x	x	x					6
Ratio AT/MGDC						x								1
Âge moyen (Arrêt de travail)					x				x					2
Ancienneté moyenne (incapacité)					x									1
Ancienneté moyenne (invalidité)		x			x									2
Duration Moyenne (incapacité)		x			x		x							3
Duration moyenne (invalidité)					x									1
Dispersion de l'ancienneté (incapacité)					x									1
Dispersion de l'ancienneté (invalidité)					x									1
Dispersion de l'âge (RC)					x	x					x		x	4
Dispersion de l'âge (RE)					x						x			2
BE Life											x			1
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x	x	x	x		x	x	x			x		x	9
Chiffre d'affaires											x			1
Capitaux sous risques (Life CAT)	x	x	x	x	x	x	x	x	x		x		x	11
Taux de PPNA (RC)	x	x												2
Duration du passif										x	x		x	3
Duration de l'actif										x				1
Absorption par le FDB (Life Frais)	x	x												2
Absorption par le FDB (Life Longévité)	x													1
Absorption par le FDB (SLT Rachat Up)	x													1
Duration moyenne des OTF	x	x	x											3
Duration moyenne des OTV	x	x	x				x	x	x					6
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)							x		x					2

**Tableau 13 : Sélection de variables pour l'estimation du SCR HSLT**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Taux de frais (Arrêt de travail)	x	x	x				x	x	x	x		x		8
Taux de Frais (RC)	x	x	x				x	x	x					6
Ratio AT/MGDC						x				x	x		x	4
Chiffre d'affaires											x			1
Capitaux sous risque (Life CAT)	x	x	x	x	x	x	x	x	x		x		x	11
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x		x			x	x	x			x		x	7
BE Life										x	x			2
Âge moyen (Arrêt de travail)		x												1
Ancienneté moyenne (incapacité)								x	x					2
Duration moyenne (incapacité)			x				x	x						3
Dispersion de l'âge (RE)							x				x		x	3
Dispersion de l'âge (RC)											x			1
Duration du passif										x	x	x	x	4
Duration moyenne des OTV							x		x					2
Absorption par le FDB (HSLT Révision)							x		x					2
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)							x		x					2

**Tableau 14 : Sélection de variables pour l'estimation du SCR HNSLT**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Evolution du CA (Santé)	x	x	x				x	x	x	x		x		8
Evolution du CA (Arrêt de travail)	x	x	x				x	x	x	x		x		8
Chiffre d'affaires	x		x							x	x	x	x	6
Taux de frais (Santé)	x	x												2
Taux de frais (Arrêt de travail)	x	x												2
Taux de réassurance (Santé)	x	x	x				x	x		x		x		7
Taux de réassurance (Arrêt de Travail - Non vie)	x	x	x				x	x		x		x		7
Taux de PPNA (Santé)	x	x							x	x		x		5
Taux de PPNA (Arrêt de travail)	x	x								x		x		4
Ratio AT/MGDC	x	x	x				x	x	x	x	x	x	x	10
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE Life	x	x	x				x			x	x	x	x	8
Ancienneté moyenne (invalidité)		x												1
Duration moyenne (incapacité)		x	x		x	x	x	x	x					7
Duration moyenne (invalidité)				x	x		x		x	x		x		6
Dispersion de l'âge (RC)				x			x				x			3
Dispersion de l'âge (RE)				x	x		x	x			x		x	6
Âge moyen (RC)											x			1
Âge moyen (RE)											x			1
Proportion d'hommes (RC)											x			1
Proportion de garçons (RE)											x			1
Duration du passif	x	x	x					x		x	x	x	x	8
Capitaux sous risques (Life CAT)	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Répartition de l'actif	x	x												2
Répartition du chiffre d'affaires (Ridge et elastic net seulement)							x		x					2

**Tableau 15 : Sélection de variables pour l'estimation du SCR HCAT Concentration**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC (idem step BIC)	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (RC)	x					x	x	x					4
Coefficient d'évolution du chiffre d'affaires (RE)									x	x			2
Taux de PPNA (santé)										x			1
Taux de PPNA (RE)										x			1
Exposition incapacité d'au plus 12 mois (Health CAT - Scénario concentration)	x	x	x			x	x	x	x	x	x	x	10
Exposition invalidité d'au plus 10 ans (Health CAT - Scénario concentration)	x	x				x	x	x					5
Exposition invalidité permanente (Health CAT - Scénario concentration)	x	x	x	x	x	x	x	x	x	x	x	x	12
Exposition Soins médicaux (Health CAT - Scénario Concentration)									x	x			2
Duration Passif									x				1
Âge moyen (Arrêt de travail)		x											1
BE Life								x					1
Absorption par le FDB (Scénario Pandémie)									x				1
Absorption par le FDB (Scénario Concentration)									x				1
Duration moyenne des OATi									x	x			2

**Tableau 16 : Sélection de variables pour l'estimation du SCR HCAT Accident de masse**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Exposition invalidité d'au plus 10 ans (Health CAT - Scénario accident de masse)	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Exposition incapacité d'au plus 12 mois (Health CAT - Scénario accident de masse)	x	x	x		x		x	x						6
Exposition invalidité permanente (Health CAT - Scénario accident de masse)	x	x					x	x		x	x	x	x	8
Exposition soins médicaux (Health CAT - Scénario Pandémie)	x	x			x		x				x			5
Exposition consultation (Health CAT - Scénario Pandémie)					x									1
Nombre de cotisants (Health CAT - Scénario pandémie)					x						x		x	3
Nombre de bénéficiaires pour incapacité d'au plus 12 mois (Health CAT - Scénario Accident de masse)		x			x			x	x					4
Coefficient d'évolution du chiffre d'affaires (Arrêt de travail)											x			1
Coefficient d'évolution du chiffre d'affaires (Santé)										x				1
Taux de PPNA (Santé)										x		x		2
Taux de PPNA (RC)										x	x			2
Taux de PPNA (RE)										x				1
Taux de PPNA (Arrêt de travail)										x				1
S/P (RC)											x		x	2
Duration du passif										x				1

**Tableau 17 : Sélection de variables pour l'estimation du SCR HCAT Pandémie**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Taux de frais (RE)	x	x						x						3
Taux de frais (RC)	x							x	x	x	x			5
Taux de frais (Santé)					x									1
Taux de réassurance (RC)	x	x						x		x				4
Taux de réassurance (Décès)									x					1
Taux de PANE (RC)	x							x		x	x			4
Taux de PPNA (RE)										x	x			2
Dispersion de l'âge (RE)	x								x					2
Ancienneté moyenne (incapacité)		x												1
Dispersion de l'ancienneté (invalidité)		x												1
Duration moyenne (invalidité)								x						1
Duration du passif										x	x			2
BE Life										x	x			2
BE NSLT										x	x			2
Coefficient d'évolution du chiffre d'affaires (RC)											x			1
Chiffre d'affaires										x	x			2
Exposition consultations (Health CAT - Scénario pandémie)	x	x	x	x	x	x	x	x	x		x			10
Exposition hospitalisation (Health CAT - Scénario pandémie)	x	x					x	x	x	x	x	x	x	9
Exposition soins médicaux (Health CAT - Scénario pandémie)	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Nombre de bénéficiaires pour incapacité d'au plus 12 mois (Health CAT)	x	x	x	x	x	x		x	x					8
Nombre de cotisants (Health CAT - Scénario pandémie)	x	x								x	x			4
Poids des obligations convertibles		x					x	x	x					4
Poids des obligations								x						1
Poids des obligations perpétuelles							x							1
Duration moyenne des OATi	x	x					x	x	x					5

**Tableau 18 : Sélection de variables pour l'estimation du SCR HCAT**

	GLM complet	GLM complet + Step AIC	GLM ACP	GLM ACP + Step AIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (AT)							x			1
Chiffre d'affaires						x	x			2
Exposition invalidité d'au plus 10 ans (Health CAT - Scénario accident de masse)	x	x	x	x	x	x	x			7
Exposition invalidité permanente (Health CAT - Scénario accident de masse)	x	x				x				3
Exposition invalidité permanente (Health CAT - Scénario concentration)	x	x	x	x	x	x	x	x	x	9
Exposition invalidité d'au plus 10 ans (Health CAT - Scénario concentration)								x	x	2
Nombre de bénéficiaires pour incapacité d'au plus 12 mois (Health CAT)					x		x			2
Nombre de bénéficiaires pour décès accidentel (Health CAT)							x			1
Nombre de bénéficiaires pour invalidité permanente (Health CAT)								x		1
Exposition consultation (Health CAT - Scénario pandémie)							x			1
Exposition hospitalisation (Health CAT - Scénario pandémie)				x			x			2
Exposition soins médicaux (Health CAT - Scénario pandémie)							x		x	2
Nombre de cotisants (Health CAT - Scénario Pandémie)									x	1
Capitaux sous risques (Health CAT - Scénario Pandémie)								x		1
Poids des obligations							x			1
Poids des obligations convertibles							x			1
Duration moyenne des OATI							x	x	x	3

**Tableau 19 : Sélection de variables pour le SCR Health**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (Santé)	x	x	x				x	x	x	x	x	x	x	10
Coefficient d'évolution du chiffre d'affaires (AT)	x	x	x				x	x	x	x	x	x	x	10
Chiffre d'affaires	x	x	x							x	x	x	x	7
Ratio AT/MGDC	x		x	x	x	x		x	x	x	x	x	x	11
Duration moyenne (incapacité)		x	x	x	x	x	x	x	x		x		x	10
Duration moyenne (invalidité)							x				x			2
Dispersion de l'âge (RC)				x	x	x					x			4
Dispersion de l'âge (RE)	x	x		x	x	x		x			x		x	8
Proportion d'hommes (RC)											x			1
Proportion de garçons (RE)											x			1
Taux de réassurance (Santé)	x		x					x		x		x		5
Taux de réassurance (AT - Non vie)	x	x	x					x				x		6
Taux de réassurance (AT - Vie)										x				1
Taux de PPNA (AT)	x	x								x				3
Taux de PPNA (Santé)										x				1
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE Life	x	x	x							x	x	x	x	7
Capitaux sous risques (Life CAT)	x	x	x				x	x	x	x	x	x	x	10
Exposition pour soins médicaux (Health CAT - Scénario pandémie)	x	x												2
Nombre de bénéficiaires pour incapacité d'au plus 12 mois (Heath CAT)	x	x	x				x	x	x					6
Exposition invalidité d'au plus 10 ans (Health CAT - Scénario accident de masse)	x	x												2
Exposition décès accidentel (Health CAT - Scénario concentration)								x						
Duration du passif	x	x	x					x		x	x	x	x	8
Poids des actions	x	x												2
Poids des obligations	x	x												2
Poids des obligations convertibles	x	x												2
Poids des obligations perpétuelles	x	x												2
Poids de l'immobilier	x	x												2
Poids du monétaire	x	x												2
Poids des OPCVM	x	x												2
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)							x		x					2

**Tableau 20 : Sélection de variables pour l'estimation du SCR risque de taux d'intérêt**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (Santé)	x	x	x					x	x	x		x		7
Coefficient d'évolution du chiffre d'affaires (RC)	x	x	x					x	x	x		x		7
Ratio MGDC/AT	x	x		x	x		x		x	x		x		8
Duration moyenne (incapacité)		x	x	x	x	x		x	x	x	x	x		10
Duration moyenne (invalidité)	x	x	x							x	x	x	x	7
Ancienneté moyenne (incapacité)								x	x					2
Ancienneté moyenne (invalidité)							x			x	x	x		4
Dispersion de l'ancienneté (invalidité)		x	x		x		x				x		x	6
Dispersion de l'âge (AT)							x				x		x	3
Âge moyen (AT)				x	x	x	x	x	x		x	x	x	9
Taux de frais (AT)	x	x						x						3
Âge moyen (RC)	x	x	x					x	x			x		6
Dispersion de l'âge (RC)				x	x	x								3
Dispersion de l'âge (RE)				x	x	x								3
Taux de réassurance (RE)	x		x					x	x					4
Capitaux sous risques (Life CAT)	x	x	x					x		x	x	x	x	8
BE SLT	x	x	x	x	x	x	x	x	x	x	x	x	x	13
BE NSLT	x	x	x	x	x	x	x	x	x	x		x		11
BE Life		x	x	x	x	x	x	x	x	x		x		10
Duration du passif										x	x	x	x	4
Duration de l'actif		x						x	x					3
Duration moyenne des OATi	x	x	x					x	x	x		x		7
Poids des actions		x												1
Poids de l'immobilier		x												1
Poids du monétaire		x	x											2
Poids des OPCVM		x	x						x					3
Poids des obligations		x						x	x					3
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)									x					1

**Tableau 21 : Sélection de variables pour l'estimation du SCR risque action**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (Santé)	x	x												2
Poids des actions	x	x	x				x	x	x	x	x	x	x	10
Poids des obligations	x	x	x				x	x	x	x	x	x	x	10
Poids des obligations convertibles	x	x	x	x	x	x		x	x	x	x	x	x	12
Poids des obligations perpétuelles	x	x	x					x	x	x	x		x	8
Poids de l'immobilier	x	x	x					x	x	x	x	x	x	9
Poids du monétaire	x	x	x								x		x	5
Poids des OPCVM	x	x	x				x	x	x	x	x	x	x	10
Âge moyen (AT)		x												1
Dispersion de l'âge (AT)		x												1
Ancienneté moyenne (invalidité)		x												1
Duration moyenne (incapacité)		x												1
Duration moyenne des OTF		x					x			x		x		4
Duration moyenne des OAT				x	x		x	x		x	x	x		7
Duration moyenne des OTV										x		x		2
Duration moyenne des OATi										x	x			2
Duration de l'actif		x		x	x		x			x		x		6

**Tableau 22 : Sélection de variables pour l'estimation du SCR risque immobilier**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Poids des actions	x	x	x	x			x	x			x			7
Poids des obligations	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Poids des obligations convertibles	x	x	x	x						x				5
Poids des obligations perpétuelles	x	x	x											3
Poids de l'immobilier	x	x	x	x	x	x	x	x	x	x	x	x	x	13
Poids du monétaire	x	x	x				x			x	x			6
Poids des OPCVM	x	x	x				x							4
Duration de l'actif								x						1
S/P (Santé)		x												1
BE Life		x												1

**Tableau 23 : Sélection de variables pour l'estimation du SCR risque de change**

	GLM complet	GLM complet + Step AIC	GLM ACP	GLM ACP + Step AIC	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
BE NSLT	x	x								2
Be Life		x								1
Taux de PPNA (Santé)	x									1
Taux de PANE (Santé)					x					1
Duration moyenne (invalidité)		x								1
Poids des actions		x				x	x	x	x	5
Poids des obligations		x				x		x		3
Poids des obligations perpétuelles						x	x	x		3
Poids des obligations convertibles							x		x	2
Poids de l'immobilier		x								1
Poids du monétaire		x				x		x		3
Poids des OPCVM		x								1
Duration moyenne des OTF		x								1
Duration moyenne des OATi			x	x		x	x	x	x	6
Duration moyenne des OTV							x		x	2

**Tableau 24 : Sélection de variables pour l'estimation du SCR risque de concentration**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC (idem step BIC)	Régression RIDGE	Régression LASSO	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Poids du monétaire	x	x	x	x	x	x	x	x	x	x	x	x	12
Poids des obligations	x	x	x	x	x	x	x	x	x	x	x	x	12
Poids des obligations convertibles	x	x	x			x	x		x	x			7
Poids des OPCVM		x	x			x				x		x	5
Poids de l'immobilier	x	x	x			x				x		x	6
Poids des obligations perpétuelles	x	x	x			x				x			5
Poids des actions	x	x	x			x							4
Duration de l'actif		x											1
Duration des OAT		x											1
Duration des OTV							x						1
Duration des OATi							x						1

**Tableau 25 : Sélection de variables pour l'estimation du SCR risque de spread**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Absorption par le FDB (Risque de spread)	x											1
Duration de l'actif	x	x	x	x	x	x	x	x	x	x	x	11
Duration des OAT	x	x	x	x	x	x						6
Duration des OTV	x							x		x		3
Duration des OATi		x						x	x	x	x	5
Duration des OTF				x	x	x	x		x			5
Poids des actions	x	x	x				x	x	x	x	x	8
Poids des obligations	x	x	x				x	x	x	x	x	8
Poids des obligations convertibles	x	x	x					x	x	x	x	7
Poids des obligations perpétuelles	x	x	x	x	x	x	x	x		x		9
Poids des OPCVM	x	x	x				x	x	x	x	x	8
Poids du monétaire	x	x	x					x	x	x	x	7
Poids de l'immobilier	x	x	x	x	x	x		x	x	x	x	10

**Tableau 26 : Sélection de variables pour l'estimation du SCR marché**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Régression Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (AT)	x	x						x						3
Coefficient d'évolution du chiffre d'affaires (RC)	x	x	x					x						4
Chiffre d'affaires	x	x								x	x	x		5
Ratio MGDC/AT	x	x	x											3
Duration moyenne (invalidité)		x	x		x	x		x						5
Dispersion de l'ancienneté (invalidité)		x	x					x						3
Âge moyen (RC)			x											1
Taux de PANE (AT)	x	x												2
Taux de réassurance (RE)		x						x						2
BE Life	x	x								x	x	x		5
BE SLT	x	x	x				x	x	x	x	x	x	x	10
BE NSLT	x	x	x					x		x		x		6
Capitaux sous risques (Life CAT)	x		x					x		x	x	x	x	7
Duration du passif										x	x			2
Poids des actions	x	x	x				x	x	x	x	x	x	x	10
Poids des obligations	x	x	x				x	x	x	x	x	x	x	10
Poids des obligations convertibles	x	x	x					x						4
Poids des obligations perpétuelles	x	x	x	x	x	x		x	x					8
Poids des OPCVM	x	x	x				x	x	x	x	x	x	x	10
Poids de l'immobilier	x	x	x											3
Poids du monétaire	x	x	x								x			4
Duration des OTF	x	x		x	x	x		x						6
Duration des OATi	x	x	x											3
Duration des OTV								x						1
Duration de l'actif	x	x	x				x	x	x	x		x		8
Répartition du chiffre d'affaires (Ridge et Elastic Net)							x		x					2

**Tableau 27 : Sélection de variables pour l'estimation du BSCR**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC (Idem step BIC)	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (Santé)	x	x	x				x	x	x	x	x	x	9
Coefficient d'évolution du chiffre d'affaires (AT)	x	x	x				x	x	x	x	x	x	9
Coefficient d'évolution du chiffre d'affaires (Décès)	x	x	x				x		x				5
Chiffre d'affaires	x	x	x						x	x		x	6
Âge moyen (AT)		x		x	x								3
Dispersion de l'âge (AT)		x								x		x	3
Duration moyenne (incapacité)		x	x				x			x		x	5
Duration moyenne (invalidité)		x								x			2
Ratio MGDC/AT	x	x	x	x	x		x			x		x	8
Dispersion de l'ancienneté (incapacité)	x												1
Dispersion de l'ancienneté (invalidité)				x	x								2
Ancienneté moyenne (invalidité)				x	x								2
Taux de frais (AT)	x	x				x	x						4
Taux de frais (Santé)						x							1
Taux de frais (RE)								x					1
Taux de frais (RC)								x					1
Taux de frais (Décès)								x					1
Taux de PANE (AT)								x					1
Taux de PANE (RC)								x					1
Taux de PANE (Décès)						x		x					2
Capitaux sous risques (Life CAT)	x	x		x	x				x	x	x	x	8
BE Life	x	x	x						x	x	x	x	7
BE SLT	x	x	x	x	x				x	x	x	x	9
BE NSLT	x	x	x	x	x				x	x	x	x	9
Duration du passif									x	x			2
Duration de l'actif		x	x				x						3
Poids des actions	x	x	x			x	x	x	x				7
Poids des obligations	x	x	x			x	x	x					6
Poids des obligations perpétuelles	x	x	x			x		x					5
Poids des obligations convertibles						x	x	x					3
Poids de l'immobilier	x	x	x										3
Poids du monétaire	x	x	x										3
Poids des OPCVM	x	x	x										3
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)						x		x					2

**Tableau 28 : Sélection de variables pour l'estimation du SCR**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	Elastic Net	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (Santé)	x	x	x					x		x	x	x	x	8
Coefficient d'évolution du chiffre d'affaires (AT)	x	x	x					x	x	x	x	x	x	9
Coefficient d'évolution du chiffre d'affaires (RE)	x	x						x						3
Coefficient d'évolution du chiffre d'affaires (Décès)	x	x	x					x	x	x				6
Chiffre d'affaires	x		x				x		x	x	x			6
Âge moyen (AT)				x	x	x			x		x			5
Dispersion de l'âge (AT)											x			1
Duration moyenne (incapacité)		x	x	x	x	x		x	x	x	x	x	x	11
Duration moyenne (invalidité)				x	x	x					x			4
Ratio MGDC/AT	x	x	x	x	x	x		x		x	x	x	x	11
Dispersion de l'ancienneté (incapacité)				x	x	x								3
Dispersion de l'ancienneté (invalidité)				x	x	x								3
Ancienneté moyenne (incapacité)				x	x	x					x			4
Dispersion de l'âge (RC)				x	x			x			x			4
Âge moyen (RC)									x		x			2
Proportion d'hommes (RC)	x	x					x	x	x					5
Proportion de garçons (RE)		x	x				x	x	x		x			6
Taux de frais (AT)							x							1
Taux de frais (Santé)							x							1
Taux de frais (RE)							x		x					2
Taux de PANE (Décès)							x							1
Taux de PPNA (Décès)									x					1
Taux de réassurance (AT - vie)		x						x	x					3
S/P (Décès)	x								x					2
S/P (RE)									x					1
Capitaux sous risques (Life CAT)	x	x	x	x	x	x		x		x	x	x	x	11
Exposition invalidité permanente (Health CAT - Scénario accident de masse)	x	x						x	x					4
BE Life	x	x	x	x	x	x		x	x	x	x	x	x	12
BE SLT	x	x	x	x	x	x		x	x	x	x	x	x	12
BE NSLT	x	x	x	x	x	x		x	x	x	x	x	x	12
Duration du passif								x	x	x	x			4
Absorption par le FDB (Life Longévité)	x	x						x	x					4
Absorption par le FDB (Life Frais)		x						x						2
Créances réassureurs									x					1
Duration de l'actif		x	x					x	x					4
Poids des actions	x	x	x				x			x				5
Poids des obligations	x	x	x				x	x	x	x		x		8
Poids des obligations perpétuelles	x	x	x				x							4
Poids des obligations convertibles							x							1
Poids de l'immobilier	x	x	x											3
Poids du monétaire	x	x	x											3
Poids des OPCVM	x	x	x											3
Répartition du chiffre d'affaires (Ridge et Elastic Net seulement)							x		x					2

**Tableau 29 : Sélection de variables pour l'estimation du ratio de solvabilité**

	GLM complet	GLM complet + Step AIC	GLM complet + Step BIC	GLM ACP	GLM ACP + Step AIC	GLM ACP + Step BIC	Régression RIDGE	Régression LASSO	GBM	Random Forest	GLM + GBM	GLM + Random Forest	Nombre de sélections
Coefficient d'évolution du chiffre d'affaires (AT)									x	x	x	x	4
Coefficient d'évolution du chiffre d'affaires (Décès)	x	x	x					x	x	x	x	x	8
Chiffre d'affaires	x	x	x					x	x	x	x	x	8
Âge moyen (AT)		x											1
Dispersion de l'âge (AT)		x											1
Duration moyenne (incapacité)	x	x			x			x		x		x	6
Duration moyenne (invalidité)	x	x		x	x	x				x		x	7
Ratio MGDC/AT	x	x	x	x	x			x	x	x	x	x	10
Ancienneté moyenne (invalidité)					x	x							2
Dispersion de l'âge (RC)										x		x	2
Âge moyen (RC)	x	x								x		x	4
Proportion d'hommes (RC)	x	x								x		x	4
Proportion de garçons (RE)										x		x	2
Taux de frais (Santé)									x		x		2
Taux de frais (Décès)									x		x		2
Taux de PPNA (Décès)									x		x		2
Taux de réassurance (AT - vie)		x							x				2
Capitaux sous risques (Life CAT)	x	x	x	x	x	x		x	x	x	x	x	11
BE Life	x	x	x	x	x	x		x	x	x	x	x	11
BE SLT	x	x	x						x	x	x	x	7
BE NSLT	x	x		x	x	x			x	x	x	x	9
Duration du passif	x	x	x					x	x	x	x	x	8
Créances réassureurs		x							x		x		3
Duration de l'actif	x							x					2
Poids des actions	x	x							x		x		4
Poids des obligations	x	x							x	x	x	x	7
Poids des obligations perpétuelles		x											1
Poids de l'immobilier	x	x							x				3
Poids du monétaire	x	x											2