Mémoire présenté devant l'UFR de Mathématique et Informatique

pour l'obtention du Master Actuariat

le 30/09/2022

Par : Jonas BUHLER

Titre: Quantification de l'impact de la loi Lemoine sur la mortalité en emprunteur via l'approche de la crédibilité.

Confidentialité : ☑ NON ☐ OUI Durée : ☐ 1 an ☐ 2 ans ☐ 3 ans ☐ 4 ans ☐ 5 ans

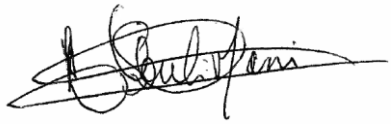
Membres du jury de l'Unistra :

J. BERARD
E. BIRMELE
A. COUSIN
P.-O. GOFFARD
M. MAUMY-BERTRAND

Entreprise : Sogecap

Directeur de mémoire en entreprise :
Nom : SOULIMANI Mohamed

Signature du responsable entreprise



Signature du candidat

Secrétariat : Mme Stéphanie Richard

Université de Strasbourg

Mémoire

Quantification de l'impact de la loi Lemoine sur la mortalité en emprunteur via l'approche de la crédibilité

Septembre 2021 — Septembre 2022

Jonas Buhler
Master 2

Entreprise : SOGECAP
Tuteurs : Bastien Laizet - Mohamed Soulimani
Actuaires associés

Résumé

La loi Lemoine, entrée en vigueur le 1er juin 2022, permet un meilleur accès au marché de l'emprunteur en France. La suppression des formalités médicales risque d'engendrer un impact sur la sinistralité observée de l'assureur. Cet impact doit être étudié et quantifié afin d'adapter au mieux les hypothèses techniques de projection. Le cas particulier du risque décès est traité dans cette étude et une nouvelle loi de mortalité décrivant la sinistralité du *new business* post loi Lemoine est construite. Les étapes suivies sont l'étude des contraintes liées aux données emprunteur, l'analyse de l'effet de la sélection médicale, l'explication de l'apport de la théorie de la crédibilité au sujet, le développement d'un algorithme de sélection de tranches d'âges venant supporter les limites de cette dernière théorie, et enfin la mise en place d'une méthodologie de construction d'une loi de mortalité moyenne d'un portefeuille *new business* post loi Lemoine. L'estimation de ces taux permet enfin le chiffrage de l'impact mortalité de la loi Lemoine ainsi que l'établissement d'un intervalle de confiance en réalisant des travaux de sensibilité.

Mots clefs : Mortalité, Loi Lemoine, Emprunteur, Théorie de la crédibilité, Effet sélection médicale, Sélection de tranches d'âges, Impact loi Lemoine, Contraintes maille prêt, Méthode bayésienne empirique de Bühlmann

Abstract

The Lemoine law, which came into force on June 1, 2022, allows better access to the borrower's market in France. The abolition of medical formalities risks having an impact on the insurer's observed Loss experience. This impact must be studied and quantified in order to best adapt the technical projection hypotheses. The particular case of death risk is treated in this study and a new mortality law describing the mortality rates of the new business post Lemoine law is built. The steps followed are the study of the constraints linked to the borrower data, the analysis of the medical selection effect, the explanation of the contribution of the credibility theory to the subject, the development of a selection algorithm for age groups coming to support the limits of this last theory, and finally the establishment of a methodology for constructing an average mortality law for a new business post Lemoine law portfolio. Finally, estimating these mortality rates allows quantification of the Lemoine law mortality impact as well as the establishment of a confidence interval by carrying out sensitivity work.

Key words : Mortality, Lemoine law, Borrower, Credibility theory, Medical selection effect, Selection of age groups, Lemoine law impact, Constraints Borrower database, Bühlmann's empirical Bayesian method

Note de synthèse

Problématique

Le marché de l'assurance emprunteur subit de forts changements en 2022 avec l'entrée en vigueur de la loi Lemoine. Le deuxième titre de la loi énonce la suppression des formalités médicales pour les assurés souhaitant souscrire une assurance emprunteur et respectant certains critères sur le capital emprunté et sur l'âge à la fin de prêt. La suppression des formalités médicales va entrainer un impact sur le risque assuré et il convient pour l'assureur de quantifier cet impact afin de maintenir sa solvabilité. Cette étude a pour objectif le développement de modèles statistiques qui permettront *in fine* la construction d'une loi de mortalité décrivant la sinistralité *new business*[1] d'une population assurée post loi Lemoine. L'impact sur la mortalité pourra ensuite se déduire de cette table construite.

Estimateur de Hoem appliqué au cas de l'emprunteur

L'étude d'une loi portant sur le domaine de l'emprunteur nécessite l'utilisation de données de bases de crédits. Cependant, ces dernières présentent certaines contraintes et leur bonne prise en compte est nécessaire pour garantir la précision des résultats de l'étude.

Dans le cas d'une base de crédits, les lignes peuvent présenter des dépendances. En effet, un même individu peut souscrire plusieurs prêts et peut donc avoir plusieurs lignes à son nom dans la base de données. Lors d'une construction de table de mortalité à la maille prêt, c'est-à-dire quand une ligne de la base correspond à un prêt et non à une tête, la non-prise en compte de cette dépendance dans la loi asymptotique des taux bruts peut engendrer un biais. La prise en compte de cette dépendance permet de déduire la convergence asymptotique de l'estimateur de Hoem suivante :

$$Loi(\widehat{q_x}) \underset{n_x \to +\infty}{\to} \mathcal{N}(q_x, \frac{q_x}{n_x^2} \sum_{p=1}^{p_{max}} p^2 E_{x,p})$$

où

- $E_{x,p}$ est la somme des expositions à l'âge x des individus ayant exactement p prêts distincts ;
- n_x est la somme des expositions à l'âge x de l'ensemble du portefeuille ;
- q_x est la probabilité de décès dans l'année d'un assuré d'âge x.

La connaissance plus fine de la loi de cet estimateur sous présence de dépendance peut permettre d'ajuster l'ensemble des modèles se basant sur la loi des taux de mortalité q_x.

Estimation de l'effet de sélection médicale

Les formalités médicales permettent à l'assureur d'avoir une meilleure connaissance de ses risques ainsi que de filtrer certains risques aggravés non assurables. De plus, un effet de sélection médicale est observable directement en portefeuille. Cet effet se traduit par une sous mortalité observée sur les premières années de chaque contrat, étant donné que l'individu vient de passer la sélection.

La loi Lemoine énonce la suppression des formalités médicales sous conditions d'éligibilité, réduisant la connaissance de l'assureur de son portefeuille et supprimant cet effet de sélection médicale observable. Dans un objectif de quantification de l'impact de la loi Lemoine sur la mortalité, la première étape est donc l'analyse de cet effet de sélection médicale.

1. Le terme *new business* représente les nouveaux arrivants en portefeuille. La formulation *new business* post loi Lemoine sera régulièrement employée dans cette étude et représente donc les nouveaux arrivants en portefeuille suite à la mise en vigueur de la réglementation.

Il est supposé que l'effet de sélection médicale est observé uniquement sur les λ premières années en portefeuille. Une méthodologie proposée pour estimer le facteur λ consiste à utiliser l'estimateur suivant :

$$\widehat{\lambda} = min\left\{ j \in [\![0; j_{max}]\!] \;\middle|\; {}^{j-1}\widehat{\Delta}^{j}_{obs} \in \overline{W}_j \right\} - 1$$

avec

$$\overline{W}_j = \left[q^j_{\frac{\alpha}{2}}; q^j_{1-\frac{\alpha}{2}} \right]$$

où

$$q^j_\alpha = z_\alpha * \sqrt{\frac{\sum_{x=x_{min}}^{x_{max}} n_x^2 \frac{\left(n_x^{j-1} - n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^j E^j_{x,p}\right) - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^{j-1} E^{B_{j-1} \setminus B_j}_{x,p}\right)}{(n_x^j n_x^{j-1})^2}}{\left(\sum_{x=x_{min}}^{x_{max}} n_x\right)^2}}$$

et

$${}^{j-1}\widehat{\Delta}^{j} = \frac{\sum_{x=x_{min}}^{x_{max}} n_x \left({}^{j}\widehat{q_x} - {}^{j-1}\widehat{q_x}\right)}{\sum_{x=x_{min}}^{x_{max}} n_x}$$

La robustesse de la méthodologie présentée ci-dessus est renforcée en l'appliquant dans une méthode de *bootstrap*.
L'application de ce test *bootstrap* sur les données du principal portefeuille emprunteur de l'entreprise donne l'estimation suivante de λ

$$\widehat{\lambda} = 2$$

Cela signifie que l'effet de sélection médicale n'est significativement [2] observable que sur les deux premières années de chaque contrat sur le produit. Avec la connaissance de ce facteur de sélection médicale, il est désormais possible de construire une loi de mortalité décrivant la sinistralité de la population assurée au-delà de deux ans. Cette nouvelle loi est notée ${}^{\lambda}q_x$ et représente donc les taux de mortalité d'une population sans effet de sélection médicale. La construction de cette loi se fait à partir de la base de données d'un produit emprunteur sur laquelle tous les sinistres et les expositions sur les deux premières années de contrat ont été tronqués. Or, le fait de supprimer une partie des données entraine un manque de volume et un éventuel risque d'échantillonnage dans l'estimation des taux de mortalité. Afin de limiter ce risque et de fiabiliser les estimateurs des q_x construits, la théorie de la crédibilité de Bühlmann, adaptée pour cette étude à la maille prêt, est appliquée.

Méthode bayésienne empirique de Bühlmann

La théorie de la crédibilité ici, appliquée à la mortalité, permet de construire un estimateur plus fiable des taux de mortalité en présence d'un faible volume de données et donc de risque d'échantillonnage. L'idée est l'utilisation de données de plusieurs portefeuilles au risque similaire pour la construction d'une loi de mortalité sur une plus petite base de données.
La table est construite par abattement sur une table de référence, i.e :

$${}^{\lambda}q_x = \widetilde{m_h} q_x^{ref}$$

avec un coefficient d'abattement calculé de la forme suivante :

$$\widetilde{m_h} = Z^h \widehat{m_h} + (1 - Z^h)\mu$$

où

$$Z^h = \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2 + \sigma^2)\sum_{p=1}^{p_{max}} {}^{p}C_h + \mu \sum_{p=1}^{p_{max}} p \; {}^{p}E_h\right)}$$

avec

2. Avec un niveau de confiance à 95%.

$$\tilde{\sigma}^2 = \frac{\sum_{h=1}^{r} E_h(\widehat{m_h} - \widehat{\mu})^2 - \widehat{\mu}^2 \left(\frac{\sum_{h=1}^{r}\sum_{p=1}^{pmax} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{pmax} {}^pC_h}{E_h} \right) - \widehat{\mu} \left(\sum_{h=1}^{r} \frac{\sum_{p=1}^{pmax} p \ {}^pE_h}{E_h} - \frac{\sum_{h=1}^{r}\sum_{p=1}^{pmax} p \ {}^pE_h}{\sum_{h=1}^{r} E_h} \right)}{\sum_{h=1}^{r} E_h - \frac{\sum_{h=1}^{r} E_h^2}{\sum_{h=1}^{r} E_h} + \frac{\sum_{h=1}^{r}\sum_{p=1}^{pmax} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{pmax} {}^pC_h}{E_h}}$$

et

$$\widehat{\mu} = \frac{\sum_{h=1}^{r} A_h}{\sum_{h=1}^{r} E_h}$$

Toutes les autres grandeurs ci-dessus dépendent uniquement de l'exposition des portefeuilles. Finalement, dans le sujet de la théorie de la crédibilité de Bühlmann appliquée à la mortalité, l'ensemble des portefeuilles qui seront retenus pour l'application du modèle ont été sélectionnés par une nouvelle méthodologie proposée basée sur une ACP. L'idée est représenter les portefeuilles sur un axe factoriel et réaliser un clustering qui permettra d'identifier des groupes homogènes de risque. Le résultat de l'ACP sur les 5 produits à disposition est le suivant :

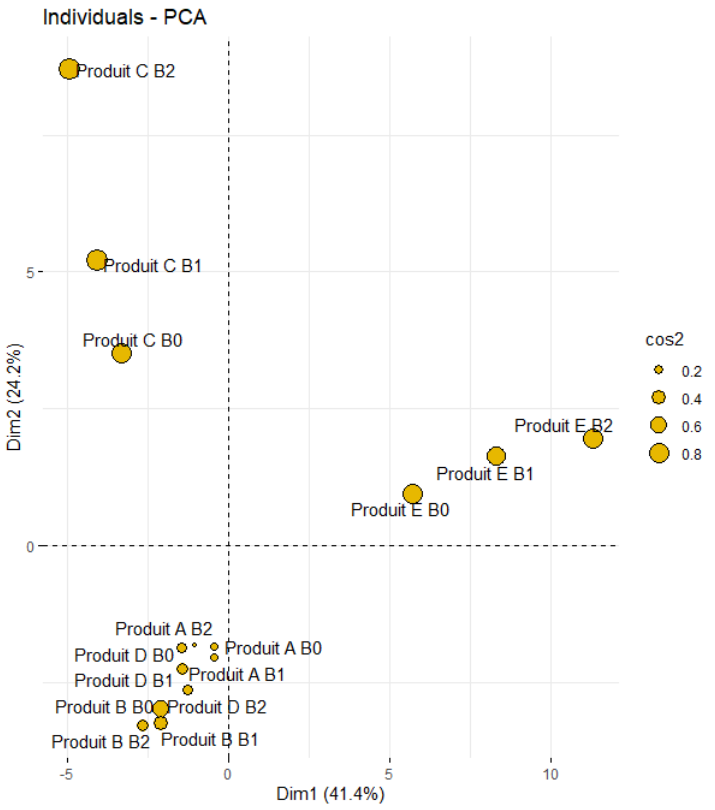


FIGURE 1 – Sélection de portefeuilles : résultat ACP sur données

L'ACP a amené à sélectionner les portefeuilles du cluster visible sous l'axe des abscisses pour l'application de la théorie de la crédibilité.

Finalement, comme la table construite par crédibilité est estimée sur base d'un abattement sur une table de référence q_x^{ref}, il peut être judicieux de ne pas appliquer un unique coefficient d'abattement sur toute la table q_x^{ref} mais de laisser plus de flexibilité au modèle. En effet, l'hypothèse du coefficient unique pour tous les âges est forte et il se peut que la déformation de la table de référence soit différente en fonction de la tranche d'âges. C'est pourquoi un modèle de sélection de tranches d'âges pour une construction de table de mortalité par abattement est mis en place dans la suite.

Sélection de tranches d'âges pour une construction de table par abattement

La construction d'une table de mortalité se fait généralement de deux manières différentes : par l'expérience ou par abattement. Dans ce deuxième cas, un coefficient d'abattement est estimé et est

appliqué à une table de référence pour obtenir une nouvelle table de mortalité sur un portefeuille généralement de faible volume. Le coefficient d'abattement est souvent sélectionné par la méthode du SMR :

$$SMR = \frac{\sum_x d_x}{\sum_x q_x^{ref} * n_x}$$

Où :

- d_x est le nombre de décès du portefeuille étudié à l'âge x ;
- q_x^{ref} est le taux de mortalité à l'âge x de la table de référence que l'on souhaite abattre.

Or, il est souvent utile de permettre une plus grande flexibilité au modèle en ne supposant pas que l'abattement soit le même pour tous les âges. Il convient alors de sélectionner des tranches d'âges sur lesquelles réaliser différents abattements. Dans la pratique, ces plages sont souvent sélectionnées à dire d'expert sans réelle justification mathématique.
Un algorithme, qui sera nommé dans la suite la méthode des plateaux, est proposé pour répondre à cette problématique. L'idée est d'étudier la fonction

$$SMR(x) = \frac{d_x}{q_x^{ref} * n_x}$$

Le nombre final de tranches qui sera retenu par l'algorithme sera calculé de la manière suivante :

$$NombreTranchesOptimal = k^* = \underset{i \in [\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}})$$

Le découpage idéal des âges est $A_1^{k^*}, A_2^{k^*}, ..., A_{k^*}^{k^*}$ avec :

$$\widehat{h^{(k^*)}} = \sum_{i=1}^{n} a_i \mathbf{1}_{A_1^{k^*}}(x)$$

avec

$$Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}}) + (i-1)\beta$$

où

$$\beta = \underset{x \in [\![2;n]\!]}{max} \frac{(SMR(x) - SMR(x-1))^2}{2}$$

et où

$$Loss(\widehat{h^{(i)}}) = \underset{h \in ESC(i)}{min} \sum_x (SMR(x) - h(x))^2$$

avec $ESC(n)$ l'ensemble des fonctions en escaliers à n étages.

L'application de cet algorithme sur les données du portefeuille emprunteur le plus volumineux de l'entreprise donne le résultat suivant :

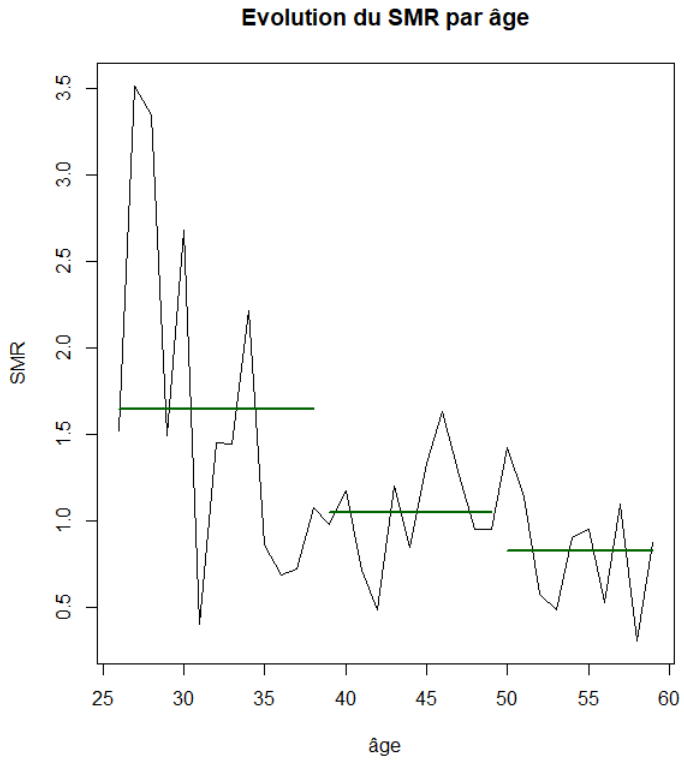


FIGURE 2 – Solution optimale, modèle des plateaux

Où les âges délimités par les plateaux verts correspondent aux 3 plateaux sélectionnés par l'algorithme pour une construction de table sur le portefeuille par abattement sur des taux de référence q_x^{ref}. Un deuxième algorithme, nommé l'algorithme des plateaux pondérés, reprend la même démarche en affinant le raisonnement. Cette démarche sera présentée plus en détail dans la suite.

Application de l'ensemble des méthodes proposées à l'estimation de l'impact loi Lemoine

La loi Lemoine est entrée en vigueur le 1er juin 2022 pour le *new business* et le 1er septembre 2022 pour le stock. Elle vise à permettre un meilleur accès au marché de l'assurance emprunteur (ainsi qu'aux prêts immobiliers) pour les risques aggravés en France.
Le deuxième titre de la loi énonce que les formalités médicales seront retirées au moment de la demande d'assurance sous certaines conditions d'éligibilité (portant sur le capital emprunté et sur l'âge de fin de prêt). La loi va donc impacter les différents risques en emprunteur, dont celui de la mortalité.

Afin de quantifier l'impact de la mise en application de cette nouvelle réglementation, une loi décrivant la mortalité du *new business* post loi Lemoine est construite. Ces taux sont estimés en analysant la composition probable d'un portefeuille emprunteur post loi Lemoine et en identifiant tous les sous groupes de risque qui composent ce portefeuille. L'allure d'un portefeuille emprunteur est le suivant :

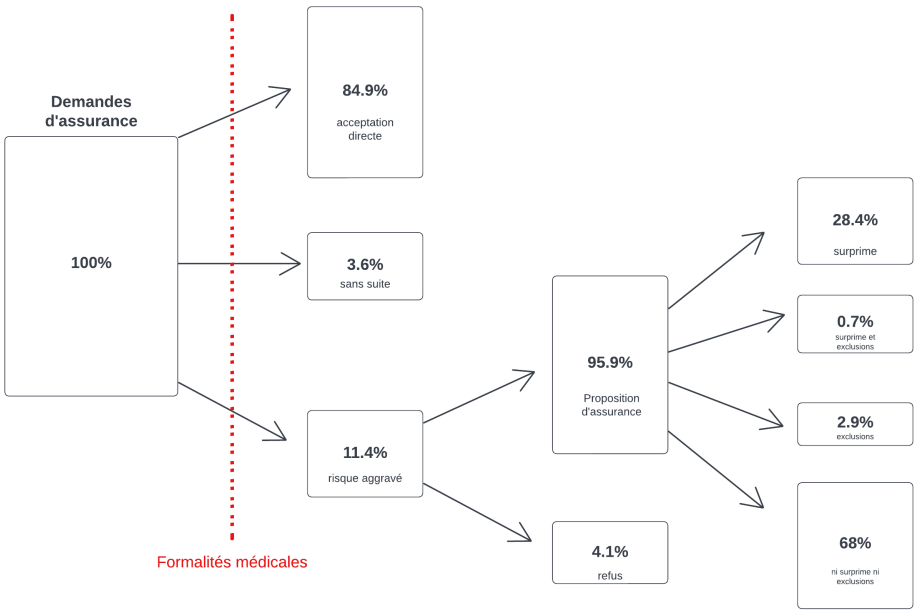


FIGURE 3 – Distribution des profils en demande d'assurance emprunteur, AERAS 2020

Les différents sous-groupes de risques identifiés sont les suivants :

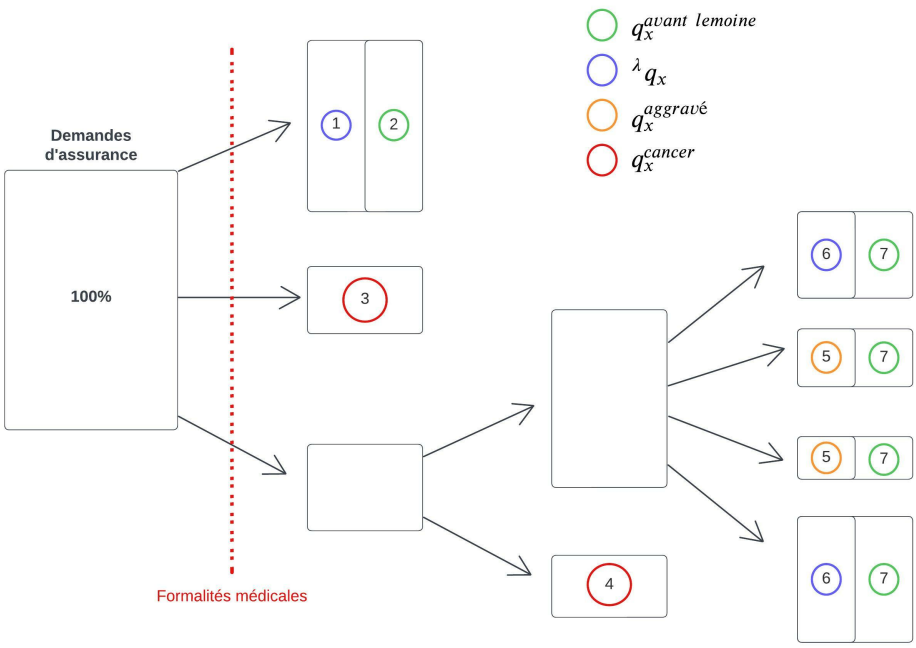


FIGURE 4 – Distribution des profils de risque dans un portefeuille emprunteur et mortalité associée

Pour chaque sous-groupe, une mortalité est estimée. La sinistralité d'un portefeuille de *new business* post loi Lemoine est donc la moyenne pondérée de la mortalité de tous ces groupes.
Un premier impact en nombre peut être estimé en comparant le nombre de décès attendu en appliquant à l'exposition les $q_x^{avant\ lemoine}$ avec le nombre de décès attendu en appliquant à l'exposition les $q_x^{apres\ lemoine}$:

$$\frac{Deces_{attendus}^{nouvelle\ loi} - Deces_{attendus}^{ancienne\ loi}}{Deces_{attendus}^{ancienne\ loi}} = 55\%$$

Ce résultat est à prendre avec du recul, car il est calculé à partir de l'hypothèse que toute la totalité de la catégorie « sans suite » reviendra en portefeuille avec un risque très aggravé (type pathologie cancéreuse). Une sensibilité sur cette hypothèse est faite :

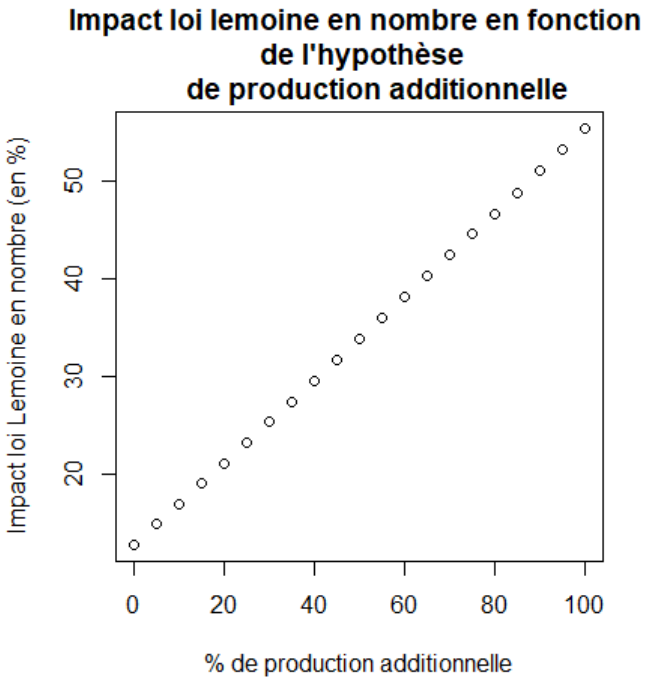


FIGURE 5 – Impact en nombre de la loi Lemoine en fonction de la part de production additionnelle dans la catégorie « sans suite »

Pour l'abscisse égale à 100%, autrement dit le cas où il est considéré que la totalité du groupe « sans suite » revient en portefeuille avec une pathologie de type cancer (mortalité extrêmement aggravée), l'ordonnée correspond bien au choc calculé précédemment : 55%.

Enfin, il convient d'encadrer l'impact sur la mortalité de la loi par un minorant et un majorant, correspondant à deux abscisses sur le schéma ci-dessus. Ces deux bornes représentent le meilleur scénario et le pire scénario pour l'assureur quant à la part d'individus du groupe « sans suite » revenant en portefeuille avec un risque très aggravé. Ce minorant et ce majorant sont calculés à partir de l'open data et il finalement estimé que l'impact sur la mortalité de la loi Lemoine se situera entre 13% et 19%.

Summary

Problematic

The borrower insurance market is undergoing major changes in 2022 with the entry into force of the Lemoine law. The second title of the law sets out the abolition of medical formalities for policyholders wishing to take out borrower insurance and respecting certain criteria on the capital borrowed and on the age at the end of the loan. The abolition of medical formalities may have an impact on the insured risk and the insurer should quantify this impact in order to maintain its solvency. This study aims to develop statistical models that will allow the construction of a mortality law describing the Loss experience of a insured new business post Lemoine law population. The impact on mortality can then be deduced from this built table.

Hoem estimator applied to the case of the borrower

The study of a law relating to the borrower's domain requires the use of credit database. However, the latter present certain constraints and their proper consideration is necessary to guarantee the accuracy of the study's results.

In the case of a credit database, the contracts can have dependencies. Indeed, the same individual can take out several loans and can therefore have several contracts in his name in the database. When constructing a mortality table on a credit database, i.e. when a base line corresponds to a loan and not to a head, the failure to take this dependence into account in the asymptotic law of gross rates can cause bias. Taking this dependence into account makes it possible to deduce the following asymptotic convergence of the Hoem estimator :

$$Loi(\widehat{q_x}) \underset{n_x \to +\infty}{\to} \mathcal{N}(q_x, \frac{q_x}{n_x^2} \sum_{p=1}^{p_{max}} p^2 E_{x,p})$$

where

- $E_{x,p}$ is the sum of exposures at age x of individuals with exactly p distinct loans.

The finer knowledge of the distribution of this estimator under the presence of dependence can make it possible to adjust all the models based on the distribution of mortality rates q_x.

Estimation of the medical selection effect

Medical formalities allow the insurer to have a better knowledge of its risks as well as to filter out certain uninsurable aggravated risks. In addition, a medical selection effect is directly observable in the portfolio. This effect results in a lower mortality observed over the first years of each contract, given that the individual has just passed the selection.

The Lemoine law stipulates the abolition of the medical formalities under eligibility conditions, reducing the insurer's knowledge of its portfolio and eliminating this observable medical selection effect. With the aim of quantifying the impact of the Lemoine law on mortality, the first step is therefore the analysis of this effect of medical selection.

It is assumed that the medical selection effect is observed only on the first λ years in the portfolio. A proposed methodology for estimating the λ factor is to use the following estimator :

$$\widehat{\lambda} = min \left\{ j \in [\![0; j_{max}]\!] \;\middle|\; {}^{j-1}\widehat{\Delta}^{j}_{obs} \in \overline{W}_j \right\} - 1$$

with

$$\overline{W}_j = \left[q^j_{\frac{\alpha}{2}}; q^j_{1-\frac{\alpha}{2}} \right]$$

where

$$q_\alpha^j = z_\alpha * \sqrt{\frac{\sum_{x=x_{min}}^{x_{max}} n_x^2 \frac{\left(n_x^{j-1}-n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2\left(q_x^j E_{x,p}^j\right)-(n_x^j)^2 \sum_{p=1}^{p_{max}} p^2\left(q_x^{j-1} E_{x,p}^{B_{j-1}\setminus B_j}\right)}{(n_x^j n_x^{j-1})^2}}{\left(\sum_{x=x_{min}}^{x_{max}} n_x\right)^2}}$$

and

$${}^{j-1}\widehat{\Delta}^j = \frac{\sum_{x=x_{min}}^{x_{max}} n_x \left({}^j\widehat{q_x} - {}^{j-1}\widehat{q_x}\right)}{\sum_{x=x_{min}}^{x_{max}} n_x}$$

The robustness of the methodology presented above is enhanced by applying it in a bootstrap method.
The application of this bootstrap test on the data of the company's main borrowing portfolio gives the following estimate of λ :

$$\widehat{\lambda} = 2$$

With the knowledge of this medical selection factor. It is now possible to construct a mortality law describing the mortality rates of the insured population beyond two years. This new law is denoted ${}^{\lambda}q_x$ and therefore represents the mortality rates of a population having removed the effect of medical selection. The construction of this law is done from the database of a borrower product on which we have truncated all claims and exposures over the first two years of the contract. However, the fact of deleting part of the data leads to a lack of volume and a possible risk of sampling in the estimation of mortality rates. In order to limit this risk and to make the estimators of the constructed q_x more reliable, Bühlmann's credibility theory, adapted for this study on a borrower database, is applied.

Empirical Bayesian method of Bühlmann

The credibility theory applied here to mortality makes it possible to construct a more reliable estimator of mortality rates in the presence of a small volume of data and therefore of sampling risk. The idea is to use data from several portfolios with similar risk to construct a mortality law on a smaller database.
The table will be built by reduction with a calculated coefficient of the following form :

$$\widetilde{m_h^*} = Z^{h^*}\widehat{m_h^*} + (1 - Z^{h^*})\mu$$

where

$$Z^h = \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h\right)}$$

with

$$\widetilde{\sigma^2} = \frac{\sum_{h=1}^r E_h(\widehat{m_h} - \widehat{\mu})^2 - \widehat{\mu}^2\left(\frac{\sum_{h=1}^r\sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^r E_h} - \sum_{h=1}^r \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h}\right) - \widehat{\mu}\left(\sum_{h=1}^r \frac{\sum_{p=1}^{p_{max}} p\ {}^pE_h}{E_h} - \frac{\sum_{h=1}^r\sum_{p=1}^{p_{max}} p\ {}^pE_h}{\sum_{h=1}^r E_h}\right)}{\sum_{h=1}^r E_h - \frac{\sum_{h=1}^r E_h^2}{\sum_{h=1}^r E_h} + \frac{\sum_{h=1}^r\sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^r E_h} - \sum_{h=1}^r \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h}}$$

and

$$\widehat{\mu} = \frac{\sum_{h=1}^r A_h}{\sum_{h=1}^r E_h}$$

All the other quantities above depend solely on the exposure of the portfolios.
Finally, in the subject of Bühlmann's credibility theory applied to mortality, the set of portfolios that will be choosed for the application of the model have been selected by a new proposed methodology based on a PCA. The idea is to represent the portfolios on a factorial axis and to carry out a clustering which will make it possible to identify homogeneous groups of risk.
The PCA result on the 5 products available is as follows :

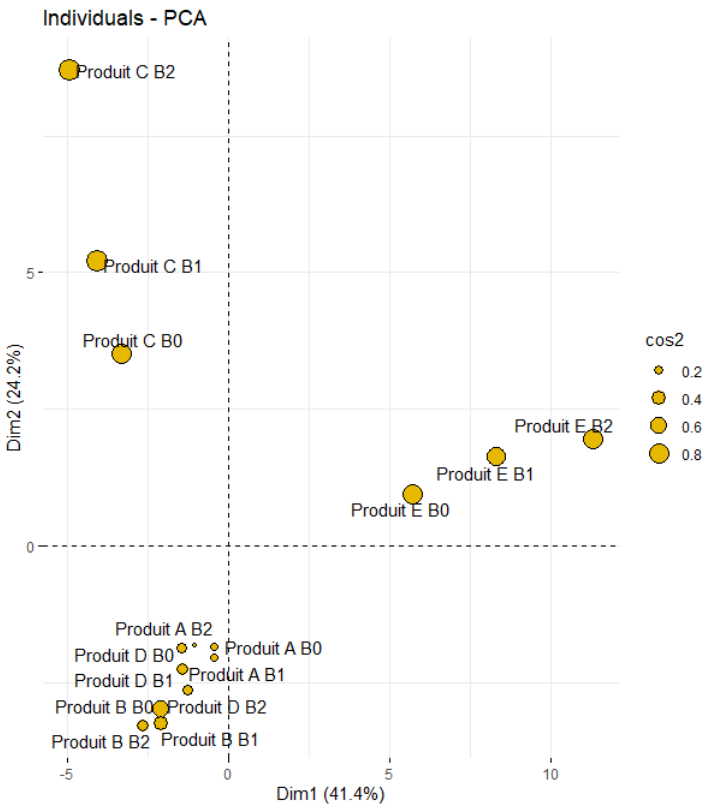


FIGURE 6 – Selection of portfolios : PCA results

The PCA led to the selection of the portfolios of the cluster visible under the x-axis for the application of the credibility theory.

Finally, as the table built by credibility is estimated on the basis of an abatement on a reference table q_x^{ref}, it may be wise not to apply a single abatement coefficient on the entire table q_x^{ref} but to leave more flexibility to the model. Indeed, the hypothesis of a single coefficient for all ages is strong and it is possible that the deformation of the reference table is different depending on the age group. This is why a model for selecting age groups for constructing a mortality table by abatement is constructed below.

Selection of age groups for a table construction by reduction

The construction of a mortality table is generally done in two different ways : by experience or by abatement. In this second case, an abatement coefficient is estimated and is applied to a reference table to obtain a new mortality table on a generally low-volume portfolio. The abatement coefficient is often selected by the SMR method :

$$SMR = \frac{\sum_x d_x}{\sum_x q_x^{ref} * n_x}$$

where :

- d_x is the number of deaths of the portfolio studied at age x ; ;
- n_x is the sum of the exposures of the portfolio studied at age x ;
- q_x^{ref} is the mortality rate of the reference table that we want to knock down.

However, it is often useful to allow greater flexibility in the model by not assuming that the reduction is the same for all ages. It is then necessary to select age groups on which to make different reductions. In practice, these ranges are often selected arbitrarily without any real mathematical justification.

An algorithm, called the plateau method, is proposed to answer this problem. The idea is to study the function

$$SMR(x) = \frac{d_x}{q_x^{ref} * n_x}$$

The final number of age groups that will be retained by the algorithm will be calculated as follows :

$$OptimalGroupsNumber = k^* = \underset{i \in [\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}})$$

The ideal division of ages is $A_1^{k^*}, A_2^{k^*}, ..., A_{k^*}^{k^*}$ with :

$$\widehat{h^{(k^*)}} = \sum_{i=1}^{n} a_i \mathbf{1}_{A_1^{k^*}}(x)$$

with

$$Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}}) + (i-1)\beta$$

where

$$\beta = \max_{x \in [\![2;n]\!]} \frac{(SMR(x) - SMR(x-1))^2}{2}$$

and where

$$Loss(\widehat{h^{(i)}}) = \min_{h \in ESC(i)} \sum_x (SMR(x) - h(x))^2$$

with $ESC(n)$ the set of step functions with n floors

The application of this algorithm on the data of the largest borrowing portfolio of the company gives the following result :

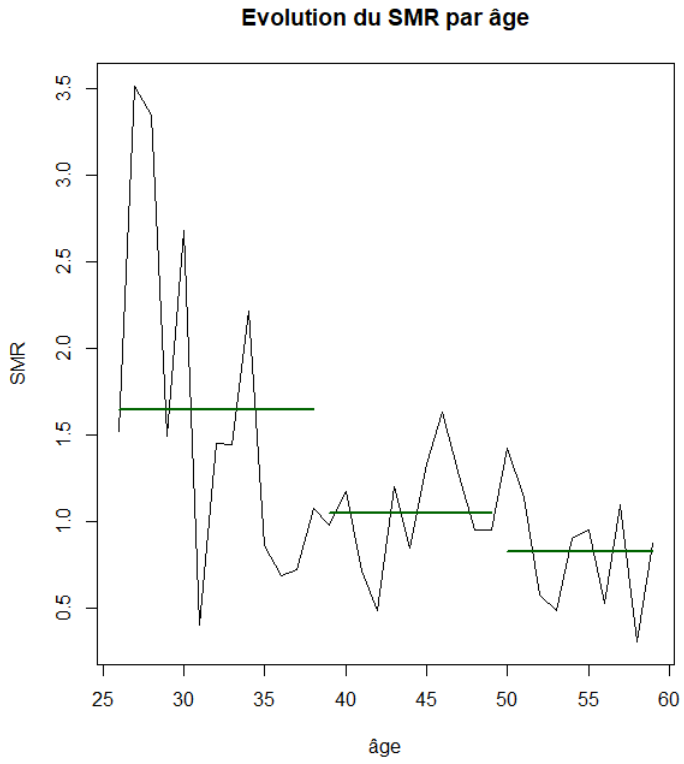


FIGURE 7 – Optimal solution, plateau model

Where the ages delimited by the green plateaus correspond to the 3 plateaus selected by the algorithm for a table construction on the portfolio by abatement on reference rates q_x^{ref}. A second algorithm, called the weighted plateau algorithm, uses the same approach by refining the reasoning. Its details will be presented in more detail in the following.

Application of all the proposed methods to the estimation of the Lemoine law impact

The Lemoine law came into force on June 1, 2022 for new business and September 1, 2022 for the stock. It aims to allow better access to the borrower insurance market for aggravated risks in France.
The second title of the law states that the medical formalities will be withdrawn at the time of the application for insurance under certain conditions of eligibility (concerning the capital borrowed and the age at which the loan ends). The law therefore risks impacting the various borrower risks, including that of mortality.
In order to quantify the impact of the application of this law. A law describing the mortality of the new business post Lemoine law is constructed. These rates are estimated by analyzing the composition of a borrower portfolio post Lemoine law and by identifying all the risk subgroups that make up this portfolio. The look of a borrowing portfolio is as follows :

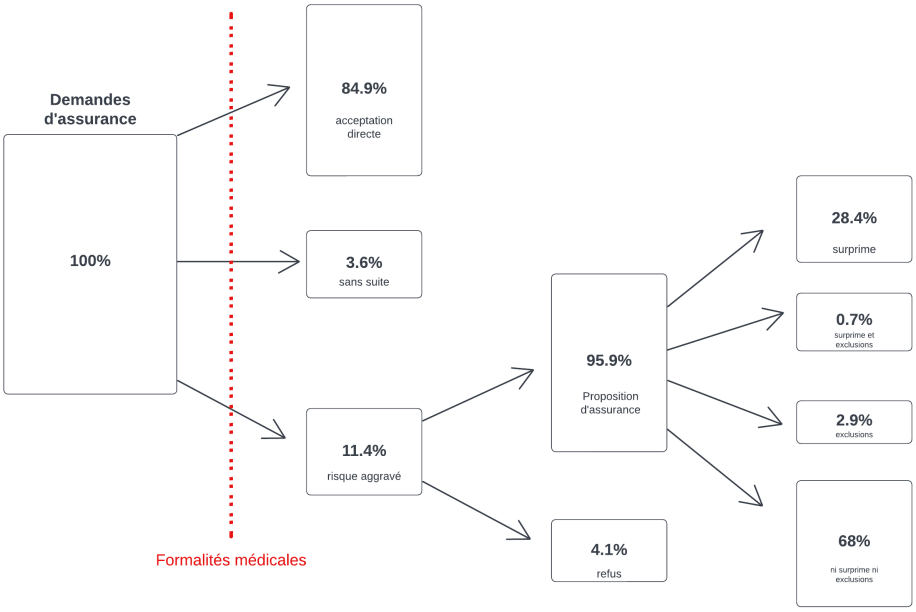


FIGURE 8 – Distribution of profiles applying for borrower insurance, AERAS 2020

The different risk sub-groups identified are as follows :

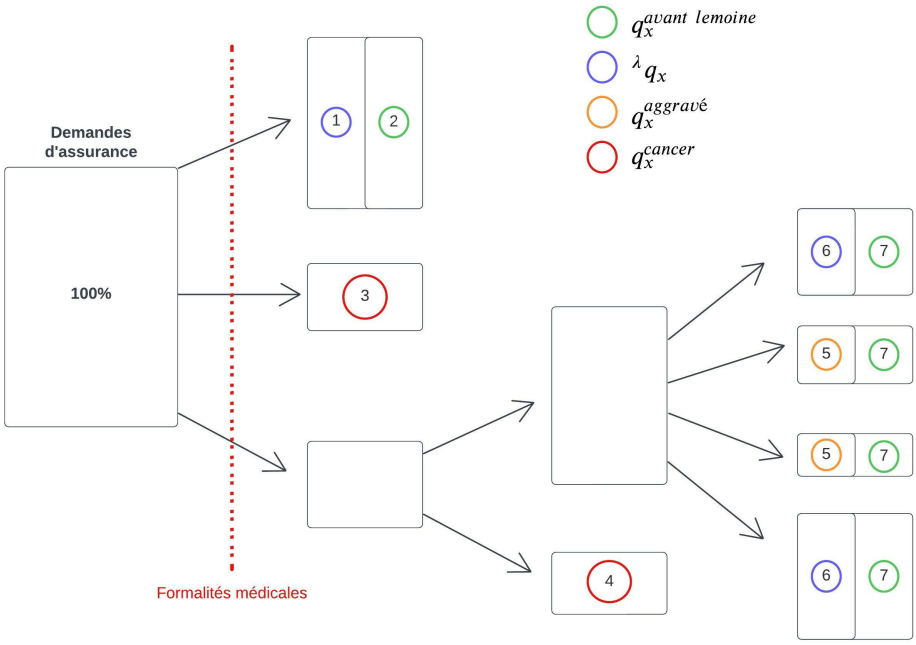


FIGURE 9 – Distribution of risk profiles in a borrowing portfolio

For each subgroup, a mortality is estimated. The mortality rates of a post-Lemoine law new business portfolio are therefore the weighted average of the mortality of all these groups.
An initial impact in terms of number can be estimated by comparing the expected number of deaths by applying the $q_x^{before\ lemoine}$ with the expected number of deaths by applying to the exposure the $q_x^{after\ lemoine}$:

$$\frac{deaths_{expected}^{new\ law} - deaths_{expected}^{old\ law}}{deaths_{expected}^{old\ law}} = 55\%$$

This result should be taken with a grain of salt because it is calculated on the assumption that the entire « category without follow-up » will return to the portfolio with a very high risk (cancer pathology type). A sensitivity on this assumption is made :

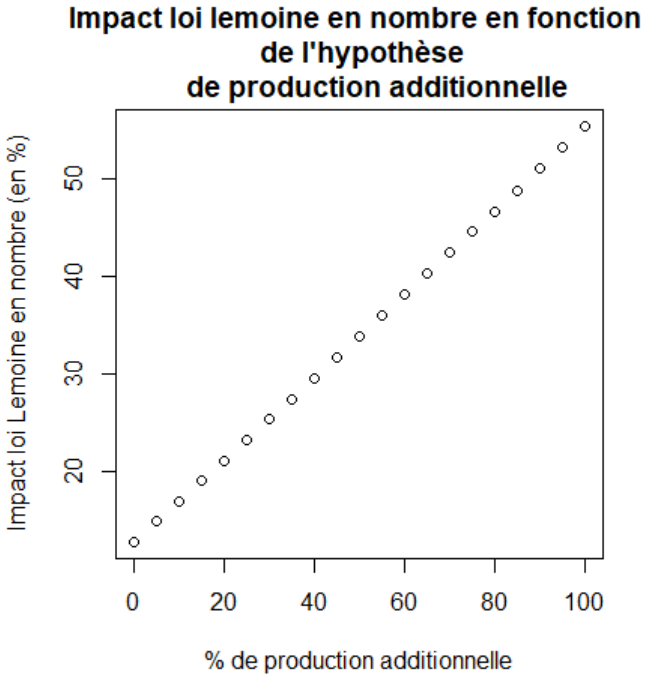


FIGURE 10 – Impact in number of the Lemoine law according to the share of additional production in the category « without follow-up »

For the abscissa equal to 100%, in other words the case where it is considered that the entire group « without continuation » returns to the portfolio with a pathology of the cancer type (extremely aggravated mortality), the ordinate corresponds well to the shock calculated previously : 55%.

Finally, the impact on mortality of the law should be framed by a lower bound and an upper bound, corresponding to two abscissae in the diagram above. These two bounds represent the best scenario and the worst scenario for the insurer in terms of the proportion of individuals in the group « without follow-up » returning to the portfolio with a very aggravated risk. This lowering and this uppering are calculated from open data and it finally estimated that the impact on mortality of the Lemoine law will be between 13% and 19%.

Remerciements

Je tiens à exprimer ma reconnaissance à toutes les personnes ayant contribué à la réalisation de mon mémoire :

Je remercie dans un premier temps **Mohamed SOULIMANI**, mon tuteur en entreprise, pour son suivi, sa bienveillance, ses conseils et son engagement.

Je remercie également **Bastien LAIZET**, mon tuteur en entreprise de début d'alternance, pour l'expertise technique qu'il m'a apportée ainsi que pour le temps qu'il m'a consacré.

J'adresse mes remerciements aux autres collaborateurs de Sogecap, en particulier à **Manon DAL PONT** et **Michael CHOUKROUN**, pour leurs conseils précieux et leurs relectures.

Je remercie mes camarades de promotion, particulièrement **Valentine POUTCOU**, qui m'a apporté un soutien inconditionnel ainsi que ses conseils précieux dans la réalisation de ce mémoire.

Enfin, je souhaite remercier l'équipe pédagogique du DUAS pour la qualité de leurs enseignements et ces trois années de formation. Je remercie en particulier **Areski COUSIN**, mon tuteur académique pour son suivi sur mon travail.

Table des matières

Introduction

La souscription d'un prêt auprès d'un établissement de crédit est systématiquement [3] accompagnée d'une assurance dite emprunteur. Cette dernière protège le prêteur face au risque de non-remboursement de la part de l'assuré lors de la survenance des risques couverts. Or, l'individu qui souhaite souscrire une assurance emprunteur doit auparavant se soumettre aux formalités médicales que lui impose l'organisme d'assurance. De cette manière, l'assureur a la possibilité de filtrer certains risques dont l'aléa est compromis ainsi que d'appliquer des surprimes ou des exclusions de garanties s'il juge que certains risques sont non diversifiables.

Le marché de l'assurance emprunteur subit actuellement de forts changements dus à l'entrée en vigueur de la loi Lemoine le premier juin 2022. Cette dernière vise à faciliter l'accès à l'assurance emprunteur pour tous et à supprimer les formalités médicales selon certains critères. De ce fait, la suppression de la sélection médicale va directement impacter la connaissance de l'assureur du profil de ses assurés. De plus, cela risque d'entrainer un comportement anti-sélectif chez ces derniers. En effet, certains individus atteints de risques très aggravés auront la possibilité de souscrire un prêt et d'être assurés, ce qui était impossible avant.

Il est donc nécessaire pour l'assureur de mener une étude sur les différents risques qu'il supporte afin de quantifier le surplus de sinistralité provoqué par cette réforme.

Parmi ces différents risques, celui qui doit obligatoirement être couvert sur un contrat d'assurance de prêt est le risque décès. Aussi ce risque est celui qui sera analysé et quantifié dans cette étude. L'objectif sera la construction d'une loi de mortalité modélisant la sinistralité des futurs assurés après mise en vigueur de la loi Lemoine. [4]

Dans un premier temps, le contexte de l'étude sera posé, à savoir la description de l'assurance emprunteur dans sa globalité, l'explication de la loi Lemoine et ses limites, ainsi que la description des données à disposition.

Par la suite, la description des différentes méthodes statistiques utilisées sera faite. Cela passera d'abord par des rappels sur les méthodes classiques de construction de table de mortalité. Ensuite, un chapitre sera dédié à l'analyse des estimateurs classiques de mortalité dans le cas particulier de l'assurance emprunteur. Dans un troisième chapitre une méthode d'analyse de l'effet de sélection médicale sera développée. Puis, la théorie de la crédibilité sera présentée et adaptée à son application sur une base emprunteur. Par la suite, une nouvelle méthode de sélection de tranches d'âges permettant de répondre aux limites de la théorie de la crédibilité sera proposée. Enfin, cette deuxième partie sera conclue par la construction de la loi décrivant la sinistralité d'un portefeuille auquel l'effet de sélection médicale aura été retiré.

Dans une dernière partie, l'étude d'impact sur la mortalité de la loi Lemoine pourra être menée

3. Légalement, la notion d'obligation n'est inscrite nulle part, mais la souscription d'une assurance est exigée par l'ensemble des établissements de crédits.

4. Dans la suite, ces individus seront nommés le ≪ *new business* post loi Lemoine ≫.

en analysant la mortalité des différents sous-groupes qui constitueront un portefeuille de futurs assurés après entrée en vigueur de la réforme.

Première partie

Contexte de l'étude

Il convient dans une première partie de décrire l'environnement de l'étude. La loi Lemoine s'appliquant à l'assurance emprunteur, il s'agit donc de définir ce périmètre d'activité. Il est également nécessaire de décrire de manière détaillée la nouvelle réglementation ainsi que ses interprétations actuelles. Enfin, les données qui serviront à mener l'étude sont décrites.

Chapitre 1

L'assurance emprunteur

Lors de la souscription d'un crédit, l'assurance emprunteur est une manière pour l'établissement prêteur de se protéger du risque que l'emprunteur ne puisse plus rembourser les échéances du prêt en cas d'accident de la vie. En cas de réalisation d'un risque entraînant la non solvabilité de l'assuré, l'assureur prend en charge, en totalité ou en partie, les échéances du prêt ou le capital restant dû à la banque.

1.1 Différents types de crédit

Les emprunts sont différenciés selon leur objet afin de fixer les taux d'intérêts. Cette classification est importante dans l'étude de la mortalité, car cette dernière varie selon le type de crédit.

1.1.1 Les crédits immobiliers

Il s'agit des emprunts destinés à financer, en partie ou en totalité, l'acquisition, la construction d'un bien immobilier ou des travaux sur ce même bien. Le crédit immobilier constitue le marché principal de l'assurance emprunteur.

1.1.2 Les crédits à la consommation

Ce type d'emprunt, également accordé à des particuliers, vise à financer des achats de biens et de services (achat d'une voiture, prêt étudiant, voyage, *etc.*). Le montant de ces prêts et leur durée sont généralement moins élevés que pour les crédits immobiliers, et la sélection médicale moins exigeante voir inexistante.

1.1.3 Les crédits professionnels

Il s'agit des emprunts destinés à financer les besoins des entreprises et des entrepreneurs dans le cadre strictement professionnel (achat de matériel et d'équipement, biens immobiliers, besoin en trésorerie, *etc.*).

1.2 Les différentes garanties du contrat

L'assureur prend des engagements, appelés garanties, vis-à-vis de l'assuré. Si un sinistre couvert par une des garanties se produit, l'assureur est dans l'obligation d'indemniser l'assuré. En France, la garantie décès est la seule garantie obligatoire dans les contrats d'assurance emprunteur. Elle peut être accompagnée d'autres engagements dont les modalités varient selon les contrats. De plus, toutes les garanties contiennent des exclusions. Elles concernent notamment les situations qui compromettent l'aléa.

1.2.1 La garantie décès

En cas de décès de l'emprunteur, le capital restant dû est remboursé par l'assureur à l'organisme prêteur. La dette n'est ainsi pas transmise aux héritiers de l'emprunteur. Cette garantie exclut en général le suicide lors de la première année du contrat, les meurtres, les situations de guerres, de terrorisme ou d'émeute.

1.2.2 La garantie PTIA

La Perte Totale et Irréversible d'Autonomie (PTIA) concerne toute invalidité physique ou mentale de l'emprunteur avant un âge de 65 ans, constatée par un médecin, causant une incapacité totale et définitive d'exercer toute activité rémunératrice, et l'obligeant à recourir à l'assistance d'une tierce personne pour accomplir les actes quotidiens de la vie (se nourrir, se laver, se déplacer, *etc.*). Dans ce cas, l'assureur, au même titre que pour la garantie décès, prend en charge le remboursement du capital restant dû.

1.2.3 La garantie invalidité permanente et totale (IPT) ou permanente et partielle (IPP)

À l'instar de la garantie PTIA, la garantie invalidité est facultative, et sa définition ainsi que ses limites peuvent varier en fonction des assureurs. L'article L.341-1 du Code de la Sécurité Sociale définit l'invalidité d'un individu comme l'état « le mettant hors d'état de se procurer un salaire supérieur à une fraction de la rémunération soumise à cotisations et contributions sociales qu'il percevait dans la profession qu'il exerçait avant la date de l'interruption de travail suivie d'invalidité ou la date de la constatation médicale de l'invalidité. » L'assureur peut ensuite évaluer le taux d'invalidité de l'assuré, à la suite d'examens médicaux. Si ce taux est supérieur à 66%, l'assuré est alors déclaré en état d'invalidité permanente et totale. L'assureur peut alors choisir de prendre en charge le remboursement du capital restant dû, ou de rembourser les échéances du prêt jusqu'à la fin du contrat si le taux d'invalidité de l'emprunteur se maintient au-dessus de 66%. Si le taux d'invalidité de l'assuré est ou devient inférieur à 66%, l'assuré est déclaré en état d'invalidité permanente et partielle. Dans ce cas, il existe des garanties qui permettent la prise en charge partielle des échéances du prêt tout au long du contrat.

1.2.4 La garantie incapacité temporaire de travail

La définition de la garantie incapacité temporaire de travail peut varier d'un assureur à l'autre. De manière générale, l'incapacité correspond à un état temporaire, dans lequel, à la suite d'un accident corporel ou d'une maladie, l'emprunteur n'est plus en mesure d'exercer, selon les contrats, une activité ou son activité professionnelle. Sur le même principe que la garantie invalidité, il existe différents degrès d'incapacité temporaire : partielle et totale. L'assuré ne peut donc pas rembourser les mensualités de son prêt à la suite de cette baisse de revenus. Dans ce cas, l'assureur prend en charge les remboursements

pendant la période d'incapacité. Si cette dernière dépasse le niveau maximal de prise en charge des échéances, qui est en général de 36 mois [1], l'assuré est considéré comme invalide et la garantie invalidité intervient alors. Certains arrêts maladie ne sont pas pris en charge par toutes les compagnies d'assurance. Par exemple, la dépression, la lombalgie, les arrêts maladie faisant suite à une tentative de suicide ou à une pratique de sports à risques peuvent être exclus ou faire l'objet d'options.

1.2.5 La garantie perte d'emploi

La garantie perte d'emploi, souscrite dans seulement 5% des contrats ([Echevin, 2019]) [2], peut compléter les garanties précédentes. En cas de licenciement économique de l'emprunteur, l'assureur prendra en charge le remboursement des échéances pendant une période de 24 mois au maximum. La démission, la rupture conventionnelle, le licenciement pour faute et les fins de Contrats à Durée Déterminée sont exclus de la garantie.

1.3 Co-emprunt

Il est possible pour plusieurs emprunteurs de souscrire ensemble un seul et même emprunt. Les co-emprunteurs sont soumis aux mêmes formalités (évaluation du profil de risque, questionnaire de santé...). De plus, chaque emprunteur paie une prime d'assurance et doit choisir son taux de couverture individuelle : sa quotité d'assurance. Cette dernière représente la part du capital emprunté couverte par l'assurance. Le taux de couverture total (les quotités de tous les emprunteurs réunies) doit être au minimum de 100% pour que l'établissement de crédit soit entièrement protégé. Il peut être réparti de la manière dont les emprunteurs le souhaitent. La quotité individuelle de chaque assuré ne peut, quant à elle, dépasser 100%.

1.4 Sélection médicale

Étape essentielle pour l'assureur, la sélection médicale en emprunteur permet l'appréciation des risques représentés par les assurés. Selon le niveau de capital emprunté et de l'âge, le niveau d'exigence des formalités médicales évolue. Pour les faibles capitaux et les individus plus jeunes, un simple questionnaire de la part de l'assuré suffit. À l'inverse, pour les emprunts importants et pour les individus plus âgés, des tests plus spécifiques tels que des analyses sanguines ou des rendez-vous médicaux peuvent être exigés par l'assureur.

Ces différences proviennent du fait que le risque encouru par l'assureur est beaucoup plus important dans le cas des hauts capitaux empruntés et des âges élevés et que ce dernier souhaite se couvrir contre les comportements anti-sélectifs des assurés. L'anti-sélection en assurance est l'existence d'une asymétrie d'informations entre l'assureur et l'assuré. Un assuré qui se sait à risque peut décider de s'assurer afin de se faire indemniser.

La sélection de risques peut sembler être en opposition avec le principe fondateur de l'assurance qu'est la mutualisation. Cependant, tous les risques ne peuvent être mutualisés et cela est particulièrement vrai dans le cas des risques aggravés en emprunteur. En effet, ces risques sont peu connus des assureurs en raison du faible volume de données et sont donc complexes à tarifer.

La sélection médicale apparaît alors comme un moyen de filtrer les risques aggravés à l'adhésion ou à minima d'ajuster la tarification et/ou d'imposer des exclusions de garanties si nécessaire.

1. Dans la réalité ce n'est pas aussi simple, l'assuré doit être classé comme invalide par la sécurité sociale, ce qui « normalement » est le cas après 36 mois d'arrêt de travail. Dans les faits, il est possible que cela perdure plus longtemps.

2. Cette garantie est peu développée en France et n'est pas proposée dans de nombreux contrats d'assurance emprunteur. De plus, c'est une garantie généralement onéreuse pour l'assuré.

1.5 Évolutions réglementaires avant loi Lemoine

Le marché de l'assurance emprunteur est en constante évolution depuis quelques années. Ces changements sont liés aux évolutions réglementaires depuis 2010. Ces différentes lois vont être présentées.

1.5.1 Loi Lagarde

La loi du 1er juillet 2010, dite loi Lagarde, interdit à l'établissement de crédit d'imposer son éventuel contrat d'assurance à l'emprunteur. En effet, cette réforme permet aux assurés de choisir une assurance d'emprunt dite individuelle auprès de l'organisme d'assurance de leur choix. Cette délégation d'assurance nécessite cependant que les garanties du contrat soient au moins équivalentes à celles incluses dans l'assurance de groupe proposée par le prêteur. Dès septembre 2010, il est également défendu aux établissements prêteurs de modifier les modalités du prêt en cas de choix d'assurance déléguée. Depuis le 27 janvier 2014, cette réforme a été renforcée afin d'empêcher les établissements bancaires de facturer des frais de délégation. En effet, l'article L.312-9 du Code de la Consommation indique qu'il leur est interdit d'≪ exiger le paiement de frais supplémentaires, y compris les frais liés aux travaux d'analyse de cet autre contrat d'assurance ≫.

1.5.2 Loi Hamon

La loi Hamon, adoptée le 17 mars 2014, permet à l'emprunteur de résilier son contrat d'assurance afin d'en changer. En effet ≪ l'assuré peut résilier le contrat dans un délai de douze mois à compter de la signature de l'offre de prêt ≫ (Article L.113-12-2 du Code des Assurances) et souscrire une nouvelle assurance, à condition que cette dernière propose des garanties équivalentes à celles du contrat initial.

1.5.3 Amendement Bourquin

En extension de la loi Hamon, depuis le 1er janvier 2018, l'amendement Bourquin rend tous les contrats résiliables une fois par an à chaque date d'anniversaire du contrat.

Chapitre 2

Loi Lemoine

Le périmètre emprunteur étant défini, il convient à présent d'expliquer le fonctionnement et les limites de la loi Lemoine.

2.1 Description de la loi

Dernière en date, cette proposition de loi a été déposée au Parlement en octobre 2021 par la députée Patricia Lemoine. Elle a été acceptée le 28 février 2022 et sa date d'application a été fixée au 1er juin 2022 pour les nouveaux contrats et au 1er septembre 2022 pour les contrats déjà souscrits. L'objectif de la loi [Lemoine, 2022] est de favoriser l'accès au marché de l'assurance emprunteur.

La loi Lemoine est séparée en deux titres :

1. « Droit de résiliation à tout moment de l'assurance emprunteur et autres mesures de simplifications » ;
2. « Droit à l'oubli et évolution de la grille de référence de la convention AERAS ».

Le premier titre établit que l'assuré a désormais la possibilité de résilier à tout moment son contrat d'assurance emprunteur pour en souscrire un autre (à garanties équivalentes) chez un autre assureur. Précédemment, l'assuré ayant souscrit un contrat d'assurance emprunteur n'avait la possibilité de résilier que la première année de souscription et ensuite à chaque date d'anniversaire du contrat.

Le deuxième titre porte sur deux points principaux et s'applique aux contrats immobiliers à usage d'habitation ou à usage mixte :

- **Le passage du droit à l'oubli de 10 à 5 ans** pour les anciens malades du cancer ainsi que pour les individus ayant été victimes d'une Hépatite C. Pour rappel, le droit à l'oubli est le droit pour l'assuré de ne pas déclarer une maladie dont il a souffert si la fin de son traitement s'est achevée au minimum 10 ans auparavant. À présent, ces 10 années sont donc ramenées à 5 pour les maladies citées ci-dessus.
- **La suppression des formalités médicales** pour tous les prêts/individus éligibles à la loi. Les conditions d'éligibilité sont les suivantes :

1. - « La part assurée sur l'encours cumulé des contrats de crédit n'excède pas 200 000 euros par assuré » ;
2. - « L'échéance de remboursement du crédit contracté est antérieure au soixantième anniversaire de l'assuré ».

Concernant les conditions d'application de la loi, certains points sont aujourd'hui sujets à interprétation. En effet, la loi soulève les questions suivantes :

- La loi mentionne un « encours cumulé » qui doit être inférieur à 200 000 euros. Faut-il prendre en compte tous les prêts immobiliers de l'individu au sein de toutes les banques ? Au sein d'une même banque tous produits réunis ? Au sein d'un même produit uniquement ?
- La loi mentionne la « part assurée » de l'individu, la quotité joue donc un rôle dans la détermination de l'éligibilité à la loi Lemoine. Comment traiter le cas d'un individu qui a deux quotités différentes sur deux risques différents ? (par exemple une quotité de 50% sur le risque décès et de 70% sur le risque arrêt de travail, qui regroupe toutes les garanties autres que le décès.)
- Si un individu souscrit deux prêts en même temps de durées différentes, quel âge à la fin de contrat est pris en compte pour déterminer les conditions d'éligibilité ? En supposant qu'un de ses deux prêts se termine avant 60 ans et l'autre après 60 ans.

Dans la suite de cette étude :

- Il sera supposé que l'encours cumulé concerne uniquement les prêts au sein d'un même produit ;
- La problématique de la quotité différente sur plusieurs risques ne sera pas traitée, car la mortalité sera la seule garantie qui sera analysée ;
- Lorsqu'un individu possède plusieurs durées différentes sur plusieurs contrats, l'âge de fin retenu pour déterminer les conditions d'éligibilité sera le maximum de tous ces âges de fin de prêt.

2.2 Risques potentiels liés à la loi Lemoine

Bien que la loi Lemoine soit une avancée majeure pour l'accessibilité au marché de l'assurance emprunteur à tous les types de risques, il en découle certaines limites. En effet, les conditions fixées par la loi pourraient être détournées. Un article de la [ligue contre le cancer, 2022] énumère certaines dérives qui pourraient être observées suite à la mise en application de cette loi :

- **Une appréciation laissée au banquier lors de la souscription d'un contrat**. Bien qu'officiellement les formalités médicales soient désormais proscrites pour la plupart des prêts, une sélection médicale « cachée » pourrait être réalisée par le banquier auprès duquel le contrat emprunteur est souscrit. Ce dernier pourrait décider de ne pas accorder de prêt à certains individus présentant un risque jugé aggravé ;
- **Risque de mise en place d'un délai de carence.** Pour contrer le comportement fortement anti-sélectif d'un assuré en fin de vie qui souscrirait un prêt inférieur à 200 000 €, l'assureur pourrait mettre en place un délai de carence pendant lequel les garanties du contrat ne feraient pas immédiatement effet ;
- **Risque de rajout de clauses d'exclusions**. Toujours dans l'optique de contrer les comportements anti-sélectifs des assurés ayant un risque aggravé, les contrats emprunteur pourraient tous faire l'objet d'exclusions de garanties liées à certaines pathologies ;
- **Risque de ne plus accéder à une assurance pour des prêts inférieurs à** 200 000 €. Dans le cas extrême, un assureur pourrait cesser de proposer des assurances pour les prêts ayant un capital éligible à la loi Lemoine, le risque étant trop incertain ;
- **Le risque d'augmentation des tarifs**. Ce dernier risque est celui ayant la plus forte probabilité d'occurrence . Du fait de l'arrêt de la sélection des risques pour une grande partie des contrats, la sinistralité moyenne des portefeuilles risque d'augmenter, et, si l'assureur ne souhaite pas modifier ses marges, la valeur de la prime augmentera naturellement.

À présent, le cadre réglementaire est posé. L'étape suivante dans le processus de quantification de l'impact de la loi est l'introduction des données. Étant donné la récente mise en vigueur de la loi, l'historique de données loi Lemoine à disposition est encore trop faible. Il est donc nécessaire de mener l'étude sur des données du passé. La méthodologie exacte de quantification sera présentée dans une partie ultérieure.

Chapitre 3

Données à disposition

La première étape dans l'étude de la mortalité d'une population est l'extraction des données. Une table de mortalité est construite à partir de données de deux types : l'exposition [1] et les sinistres. L'exposition décrit l'ensemble des adhésions d'un portefeuille, la date de début et de fin de chaque contrat. Les données de sinistres concernent, quant à elles, les informations sur les contrats sinistrés de la base d'exposition. Dans le cadre de cette étude, les bases sinistres concerneront le risque décès. Les extractions et retraitements des deux types de données sont maintenant présentés de façon indépendante.

3.1 Données d'exposition

3.1.1 Présentation des données

L'étude portera sur 5 portefeuilles emprunteur. Comme cela sera expliqué dans le chapitre 8, les méthodes d'estimation de la mortalité appliquées requerront l'utilisation de plusieurs portefeuilles différents.

Par souci de confidentialité, les 5 produits seront nommés produit A,B,C,D et E. Concernant la description de chaque produit :

- Le produit A est le produit emprunteur à usage immobilier le plus volumineux du groupe ;
- Les produits B et D sont également des produits emprunteur à usage immobilier ;
- Le produit C est un produit emprunteur à usage professionnel et immobilier ;
- Le produit E est un produit emprunteur assurant les crédits à la consommation.

Les données d'exposition de l'ensemble des portefeuilles sont extraites sur la période 2015-2018.

L'extraction des données d'exposition est réalisée sur le logiciel SAS à partir d'un serveur regroupant l'ensemble des bases d'adhésions sur le périmètre étudié. De façon mensuelle, la base est alimentée par une photographie de l'ensemble des portefeuilles à date.

Après un travail de jointure réalisé entre les différentes bases de données mises à disposition, les 5 bases d'expositions sont obtenues. Ces bases sont décrites à la maille prêt, c'est-à-dire qu'une ligne représente un contrat, avec une date de début et de fin d'observation. Le détail des variables extraites est le suivant :

1. L'exposition représente la durée (en années) passée en portefeuille d'un contrat.

Variable	Commentaire - Description
Id_Pret	Identifiant du prêt
Nom_Prenom	Concaténation du nom et prénom dans une variable
Nom_Assure	
Prenom_Assure	
Date acceptation offre	
Sexe	
Date de naissance	
Montant prêt	
Quotité	
date_debut	Date de début d'observation du prêt dans la période d'observation (voir censure et troncature)
date_fin	Date de fin d'observation du prêt dans la période d'observation (voir censure et troncature)
duree_init	Durée initiale du prêt au moment de sa souscription.
duree_diff	Durée du différé s'il existe.

TABLE 3.1 – Variables à disposition dans les données d'exposition

3.1.2 Travaux de qualité des données

L'utilisation directe des données d'exposition extraites n'est pas possible en raison des problèmes de qualité des données. La non-conformité de certaines variables, la création de doublons de contrats suite aux jointures, les variables vides, *etc.* sont des incohérences qu'il convient de retraiter. Une négligence de cette étape peut entrainer un fort biais sur les estimateurs des taux de mortalité construits à partir de ces données non fiables.

Bien que les bases de données soient extraites sur le même serveur, les problèmes de qualités des données diffèrent entre elles. Cela est dû par exemple à certaines variables qui ne sont renseignées que sur certains produits.

Les tests de qualité des données effectués sur les 5 bases d'exposition extraites sont les suivants :

- **Présence de doublons dans la base**. Un doublon est défini comme la présence de deux ou plusieurs lignes qui présentent la même clé primaire. Une clé primaire est une variable ou un ensemble de variables qui définit de façon unique une ligne de la base. Dans le cas présent, la clé primaire utilisée sur ces différentes bases emprunteur est la concaténation des variables : Id Pret, Nom assuré, Prénom assuré et date de naissance. Ces 3 dernières variables sont utilisées pour décrire un individu, car il n'existe pas de clé primaire à la maille tête qui définisse de façon unique un assuré dans les bases de données du groupe ;
- **Dates et âges cohérents**. Il doit être vérifié que, pour chaque portefeuille, tous les âges à l'adhésion sont conformes aux conditions générales du produit. De la même façon, les âges de sortie ne doivent pas excéder les âges maximums sous garanties définis dans ces mêmes conditions générales ;
- **Variables vides**. Il est vérifié que toutes les variables soient bien renseignées sur l'ensemble des 5 portefeuilles étudiés ;
- **Valeurs des variables non conformes**. Il est vérifié que certaines variables prennent des valeurs conformes. Par exemple la variable quotité doit être comprise entre 0 exclu et 100%. La variable Sexe doit quant à elle prendre des valeurs dans un ensemble à deux éléments (1 et 2 dans le cas présent).

Afin d'alléger la lecture, les résultats de ces différents tests sont présentés uniquement sur le produit A, les résultats des autres produits sont détaillés en Annexe. Par souci de confidentialité, le nombre de lignes des 5 portefeuilles n'est pas décrit.

Résultats des tests de qualité des données : Produit A

Test effectué	**Taux de non-conformité**	**Taux en exposition**
Variable Sexe mal renseignée	$0,70\%$	$0,70\%$
Variable Quotité nulle	$0,02\%$	$0,02\%$
Doublons	$0,21\%$	$0,23\%$
Age d'adhésion < Age min d'adhésion	$0,00\%$	$0,00\%$
Age d'adhésion > Age max d'adhésion	$1,10\%$	$0,99\%$
Age de sortie d'observation > Age max sous garanties	$0,43\%$	$0,44\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,10\%$	$0,10\%$
Nom vide	$0,01\%$	$0,01\%$
Prénom vide	$0,12\%$	$0,10\%$
Age début d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Date début observation = Date fin d'observation	$0,94\%$	$0,00\%$
Age début d'observation < 18	$0,00\%$	$0,00\%$

TABLE 3.2 – Résultats QDD produit A

Remarque : Les lignes non conformes liées aux prénoms vides concernent le cas des prénoms composés. Pour les individus ayant un prénom composé de plus d'un mot, leur prénom s'est concaténé avec leur nom dans la variable « Nom assuré », laissant la variable « Prénom assuré » vide.

Une première remarque générale est la bonne qualité des données globale sur les 5 produits [2]. Les tests qui ont amené à des taux de non-conformité non nuls sont traitables et n'impliquent pas nécessairement de suppressions de lignes. Suite à ces différents résultats, les retraitements effectués sont les suivants :

- Supprimer les lignes pour lesquelles la date de début est égale à la date de fin ;
- Conserver les lignes dont l'âge de sortie est supérieur à l'âge maximal sous garanties ;
- Supprimer les lignes où l'âge de l'individu à l'adhésion est inférieur à 18 ans ;
- Conserver les lignes où l'âge d'adhésion est supérieur à l'âge maximal théorique autorisé pour souscrire au contrat. En effet, ces lignes peuvent, au même titre que pour la condition précédente, être dues à des gestes et exceptions commerciales. De plus, notons que dans le cadre de la construction de la table de mortalité, ces individus n'influenceront pas les taux de décès, car il n'y a pas assez d'exposition et de décès sur les âges élevés pour pouvoir assurer la fiabilité des estimateurs construits sur ces plages d'âges ;
- Créer un algorithme basé sur une base d'apprentissage pour séparer les noms et prénoms en deux variables distinctes dans le cas des produits A, B et C ;

2. Voir en Annexe A.1 pour plus de détails sur les résultats des autres produits.

- Pour les individus ayant leur genre renseigné par un 3, rechercher s'ils ont un autre prêt dans la base et leur assigner le genre renseigné pour ce prêt. Dans le cas contraire, un algorithme basé sur une base d'apprentissage est crée pour assigner un genre à un individu en fonction de son prénom ;
- Supprimer les lignes où la quotité est nulle, car ces lignes sont souvent source de doublons ;
- Supprimer les doublons.

Le détail des suppressions de lignes liées à ces différents tests de qualité des données est présenté ci-dessous. À nouveau, les résultats sont uniquement présentés ici pour le produit A et les résultats des autres produits sont placés en Annexe A.2.

Suppressions de lignes : Produit A

Test	**Proportion de lignes supprimées**
Variable Quotité nulle	$0,02\%$
Doublons	$0,40\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,16\%$
Date début observation = Date fin d'observation	$1,24\%$
Age début d'observation < 18	$0,00\%$

TABLE 3.3 – Suppressions des lignes : produit A

Remarque : Au global, les tests ont amené à la suppression de $1,70\%$ des lignes de la base d'exposition du produit A.

Interlude : La distance de Damerau-Levenstein

Dans une problématique de qualité des données, surtout dans le cas où plusieurs sources de données sont utilisées, il est nécessaire d'utiliser un indice de similitude entre chaînes de caractères. En effet, les pratiques pouvant différer d'un gestionnaire de données à l'autre, il n'est pas rare que deux variables censées prendre la même valeur dans deux bases différentes ne soient pas exactement semblables. Par exemple, un individu peut se nommer « Jean-Marie » dans une base et « Jean-M » dans une autre.

Un indice de similitude doit alors être sélectionné pour gérer ces cas. En la matière, les distances entre chaînes de caractères les plus répandues sont :

- **La distance de Levenshtein** : Elle représente le nombre minimal d'opérations (Suppression-Ajout – Remplacement) à effectuer pour retrouver le mot 2 à partir du mot 1. Par exemple,

$$Distance_{Levenshtein}\,(LEMOINE, LEMMONE) = 2$$

Le résultat est de 2, car il a fallu supprimer une lettre M et rajouter une lettre I.

- **La distance de Damerau-Levenshtein** : Elle repose sur le même principe que la distance de Levenshtein à la différence qu'elle inclut une opération supplémentaire : la permutation de deux caractères. Par exemple,

$$Distance_{Levenshtein}\,(CREDIBILITE, CRE\mathbf{ID}BILITE) = 2$$
$$Distance_{Damerau-Levenshtein}\,(CREDIBILITE, CRE\mathbf{ID}BILITE) = 1$$

La distance de Damerau-Levenshtein est plus appropriée dans le cadre des bases de données en assurances, car elle tient davantage compte des fautes de frappe communes lors de la saisie des noms et prénoms.

Finalement, comme il est plus facile de travailler avec un indice de similitude prenant ses valeurs dans un ensemble fini, l'indice suivant est posé [3] :

$$I(mot_1, mot_2) = 1 - \frac{Distance_{Damerau-Levenshtein(mot_1,mot_2)}}{max\left(taille(mot_1), taille(mot_2)\right)}$$

Cet indice de similarité est inversement proportionnel à la distance de Damerau-Levenshtein. Il prend des valeurs entre 0 et 1, où la valeur 1 indique une similarité totale, et la valeur 0 une dissimilarité complète.

De plus, des cas d'exceptions sont rajoutés à la méthode de calcul de l'indice :

- I prend la valeur 1 (similarité totale) si l'un des deux mots est inclus dans l'autre. Par exemple,

$$Distance_{Damerau-Levenshtein}\left(\text{Jean-Philipe, Jean-P}\right) = 1$$

- Dans le cas où un des mots est composé, toutes les distances sont calculées et celle retenue est la plus grande.

  Par exemple,

$$\begin{aligned} &Distance_{Damerau-Levenshtein}\left(\text{Victor Martz, Viktor}\right) = \\ &max\left(Distance_{D-L}\left(\text{Victor, Viktor}\right), Distance_{D-L}\left(\text{Martz, Viktor}\right)\right) = \\ &max(0,83; 0) = 0,83 \end{aligned}$$

Finalement, il est important de noter que ce n'est pas parce que deux chaînes de caractères n'ont pas un indice de similarité de 1 qu'elles ne représentent pas le même mot. Dans le dernier cas exemple, l'indice était de $0,83$ mais il est imaginable de penser qu'une faute de frappe ait pu être faite.

Il convient donc de fixer un seuil de l'indice I à partir duquel la similarité des deux mots est acceptée. Dans la suite de cette étude, le seuil arbitraire de $0,80$ sera sélectionné. Sa validation découle de l'observation ligne à ligne de certains cas d'exemples pris sur les données.

3. Cette méthodologie est interne à l'entreprise.

3.2 Données des sinistres

3.2.1 Présentation des données

Pour rappel, les bases de données pour la sinistralité concernent dans cette étude le décès. De manière semblable aux données d'exposition, les données sont réparties en 5 portefeuilles qui représentent les 5 produits étudiés. Les données concernent les décès survenus sur la période de 2015 à 2018.

Les sources de ces 5 bases de données diffèrent, car ces produits sont distribués par différents organismes. Le détail des variables disponibles dans ces bases sinistres est le suivant :

Variable	Commentaire - Description
CODE PRODUIT	Identifie le produit lié au sinistre (il existe plusieurs codes produit par produit)
NOM ASSURE	
PRENOM ASSURE	
DT NAISS ASSURE	
LBL PRODUIT	Risque garanti par le contrat (DECES/PTIA dans le cas présent)
RISQUE COUVERT	Risque couvert par le sinistre
DEBUT EFFET GIE	Date de début de l'effet des garanties
FIN EFFET GIE	
ID SINISTRE	Identifie de façon unique chaque sinistre de la base
STATUT SINISTRE	Indique si le sinistre est payé, refusé, en attente, *etc.*
DT DECLARATION SIN	
DT SURVENANCE SIN	
DT REGLEMENT SIN	
ID PRET	
MONTANT REGLE	Par définition égal au produit entre le CRD et la quotité
CAPITAL RESTANT DU	
QUOTITE	

Table 3.4 – Variables à disposition dans les données sinistres

3.2.2 Travaux de qualité des données

De la même manière que pour les données d'exposition, les données sur les sinistres doivent être traitées pour en améliorer la fiabilité. Contrairement aux données d'exposition, la conformité des variables des bases sinistres est relativement bonne, car elles proviennent directement des délégataires et aucun travail d'extraction ou de jointure n'a dû être effectué au préalable.
Deux tests vont tout de même être menés pour s'assurer de la fiabilité des 5 bases sinistres :

1. Les montants réglés indiqués dans les bases sinistres vont être comparés aux montants enregistrés en comptabilité. Si ces montants coïncident, cela permettra de conforter l'utilisation de ces bases ;
2. Un travail de jointure entre les bases sinistres et les bases d'adhésions va être réalisé. Il s'agit d'associer à chaque sinistre son contrat dans la base d'exposition. Si la totalité des sinistres est retrouvée, cela permettra de garantir la bonne fiabilité des données sinistres.

Test 1 : Rapprochement avec la comptabilité

Les informations et données à disposition sont :

- Les informations comptables, à savoir les montants réglés par produit, par risque garanti, par année de règlement et année de survenance ;

- Les 5 bases sinistres des 5 produits.

Dans le cadre de cette validation, les données comptables serviront de référence au niveau de l'historique des montants des sinistres payés pour la comparaison avec les montants contenus dans les 5 bases sinistres. L'étude est menée produit par produit. Les montants réglés sur chaque produit par année de paiement et par année de survenance seront comparés avec ceux enregistrés en comptabilité, de façon à attester la fiabilité de la base des bases sinistres.

Les matrices suivantes représentent, pour chaque année de survenance (en ligne) et pour chaque année de paiement (en colonne) la différence relative entre le montant présent dans la base comptable et celui présent dans la base sinistres du produit. Cette différence relative est calculée à partir des montants de la base comptable. Afin d'alléger la lecture, le détail de ces matrices est donné pour le produit A uniquement. Les résultats relatifs aux 4 autres produits sont présentés en Annexe A.3 :

Comparaison comptable produit A

Année de survenance \ Année de paiement	2015	2016	2017	2018	2019	2020	total
2015	0,8%	1,0%	0,0%	0,0%	0,0%	0,0%	0,9%
2016	0,0%	0,0%	1,0%	0,0%	8,4%	−6,3%	0,2%
2017	0,0%	0,0%	−0,6%	0,3%	0,0%	0,0%	−0,1%
2018	0,0%	0,0%	0,0%	−1,7%	1,7%	17,7%	−0,2%
							0,4%

TABLE 3.5 – Différences relatives entre les montants réglés en comptabilité et dans la base sinistres : produit A

La différence relative entre les montants réglés enregistrés en comptabilité et ceux présents dans la base sinistres du produit A est de 0,4% sur la totalité de la période d'observation (2015-2018). Cet écart est jugé acceptable.

Finalement, étant donné les résultats, cette première étape de vérification vient conforter l'utilisation de ces 5 bases de données sinistres.

Test 2 : Jointure entre l'exposition et les sinistres

Une deuxième étape de fiabilisation des données sinistres est de vérifier que les contrats sinistrés présents dans la base peuvent être retrouvés dans la base d'exposition. Cette étape est un travail de jointure. Théoriquement, une unique jointure entre la base d'exposition d'un produit et sa base sinistres par la clé primaire (Id Pret/Nom/Prénom/Date de naissance) devrait permettre de joindre tous les sinistres à leur exposition. Cependant, les données sont de provenances différentes et les sources d'erreurs sont nombreuses. Par exemple, une lettre de différence entre le prénom d'un individu dans la base d'exposition et son prénom dans la base sinistres suffit à faire échouer la jointure.

Dans un premier temps, la jointure par la clé primaire est donc appliquée, cette étape permet tout de même de retrouver une grande partie des sinistres dans la base d'exposition. Puis, d'autres jointures sont faites en utilisant une partie de la clé primaire uniquement. Par exemple, la jointure entre les deux bases est faite sur le variable « Id Pret », « Nom » et « Date de naissance », puis il faut s'assurer que les deux prénoms présents dans les deux bases sont proches (utilisation de la distance de Damerau-Levenstein présentée précédemment).

Par exemple, en réalisant cette deuxième jointure, si un contrat possède le même Id prêt, nom et date de naissance dans les deux bases, et que cet individu se nomme « Jean-Pierre » dans la base d'exposition et « Jean P » dans la base sinistres, il est considéré que l'individu est retrouvé.

Ces travaux de jointure réalisés sur les 5 produits, les résultats sont les suivants :

- Produit A : $98,7\%$ des sinistres sont retrouvés dans la base d'exposition ;
- Produit B : $98,0\%$ des sinistres sont retrouvés dans la base d'exposition ;
- Produit C : $93,5\%$ des sinistres sont retrouvés dans la base d'exposition ;
- Produit D : $99,5\%$ des sinistres sont retrouvés dans la base d'exposition ;
- Produit E : $98,8\%$ des sinistres sont retrouvés dans la base d'exposition.

Il est à noter que la plupart des sinistres non retrouvés dans l'exposition possèdent des variables détériorées (absence de nom, variable Id Prêt tronquée, *etc.*), ce qui rend difficile le travail de jointure.

3.3 Conclusion

Finalement, les tests de qualité des données effectués sur les données d'exposition et sur les données sinistres ont permis de fiabiliser ces données et d'en conforter l'utilisation pour cette étude. En effet :

- La qualité des données faite sur les données d'exposition donne globalement des résultats satisfaisants sur les 5 produits étudiés ;
- Les montants réglés présents dans la base sinistres semblent être cohérents avec ceux enregistrés en comptabilité ;
- La majorité des sinistres de chaque produit ont pu être affectés à leur exposition. Les sinistres non retrouvés correspondent pour la plupart à des variables détériorées qui ont empêché la jointure.

Les données fiabilisées de ces 5 produits seront utilisées pour la construction d'une unique table de mortalité à partir de méthodes de crédibilité. La description de ces méthodes fait l'objet du chapitre 8.

Chapitre 4

Problématique(s) de l'étude

Le cadre de l'étude est maintenant posé. Ce dernier chapitre a pour objet la description linéaire des différentes problématiques qui seront rencontrées et qui devront être traitées au cours de ce travail de quantification de l'impact de la loi Lemoine.

4.1 Problématiques rencontrées

Pour rappel, la problématique centrale de cette étude est la suivante :

- **Quel est l'impact sur la mortalité en emprunteur de la mise en vigueur de la loi Lemoine ?**

Il convient dans un premier temps de répondre à cette question par d'autres questions.

Comme cela a été décrit dans le chapitre 2, la réforme énonce la suppression des formalités médicales, sous réserve d'éligibilité de la part de l'assuré. C'est pourquoi vouloir quantifier l'impact de la réforme sur la mortalité doit nécessairement passer par l'étude de l'effet de sélection médicale (SM) et la construction d'une loi de mortalité ne le prenant pas en compte. La première problématique est la suivante :

- **Comment décrire l'effet de sélection médicale et comment construire une table de mortalité ne le prenant pas en compte ?**

Cet effet étant observable sur les X premières années de chaque contrat, la construction de cette loi sans effet SM requiert de regarder les sinistres et l'exposition uniquement au-delà des X premières années de chaque contrat de la base. Autrement dit, les données devront être tronquées partiellement afin d'estimer les taux de mortalité sans effet SM. Dans la suite de cette étude, cette grandeur X, symbolisant le nombre d'années sur lesquelles l'effet de SM est visible, sera notée λ. Une partie sera dédiée à l'estimation de λ. Après avoir estimé ce facteur, la table sans effet SM sera construite à partir de données tronquées. Or, les estimateurs statistiques sont fonction des données, et réduire le volume de ces dernières impacte directement la confiance accordée en ces estimateurs. La seconde problématique est donc la suivante :

- **Comment construire des estimateurs fiables basés sur un faible volume de données ?**

La théorie de la crédibilité est un outil qui peut être utilisé en présence d'un manque de données, car elle permet la limitation du risque d'échantillonnage [1].

1. La description de ce risque sera faite ultérieurement dans une partie dédiée.

Enfin, après avoir pu estimer l'effet de sélection médicale, il s'agira de construire une loi de mortalité pouvant décrire la sinistralité moyenne d'un portefeuille de nouveaux arrivants après mise en vigueur de la loi Lemoine. Il sera nécessaire de considérer tous les types d'individus, allant des faibles risques aux risques très aggravés, qui se retrouveront en portefeuille. La troisième et dernière problématique est alors la suivante :

- **Comment construire une loi de mortalité décrivant la sinistralité moyenne d'une population non encore observée ?**

4.2 Portefeuille principal étudié

Les différentes étapes qui vont être mises en oeuvre dans cette étude ont été décrites ci-dessus : l'étude de l'effet de sélection médicale, la construction d'une loi de mortalité ne prenant pas en compte cet effet et la quantification de l'impact loi Lemoine.

Toutes ces études vont être menées sur le portefeuille du produit A. [2] Cette décision est prise car le portefeuille emprunteur A est le plus volumineux du groupe et, compte tenu de ses caractéristiques, est également celui qui risque d'être le plus impacté par la loi Lemoine. Ainsi, la table de mortalité décrivant la sinistralité d'une population *new business* post loi Lemoine sera construite en partie à partir des données du portefeuille A. Cependant, il est à noter que les données des autres portefeuilles seront également utilisées dans le cadre de la théorie de la crédibilité qui sera présentée dans une partie dédiée.

2. Pour plus de détails sur les caractéristiques de ce produit, voir le chapitre 3.

Deuxième partie

Méthodes d'estimation de la mortalité

L'objectif principal de cette partie est la construction d'une loi de mortalité ne prenant pas en compte l'effet de sélection médicale et pouvant être appliquée sur les individus éligibles à la loi Lemoine. [3] Pour atteindre ce but, les méthodes classiques d'estimation devront être rappelées. Puis, à partir de l'étude de l'effet de sélection médicale, une loi de mortalité candidate sera obtenue. Cette dernière sera enfin estimée par des méthodes de crédibilité.

3. Voir les critères d'éligibilité décrits dans le chapitre 2 .

Chapitre 5

Méthodes classiques

L'objectif de l'étude étant la construction d'une loi de mortalité, il convient de rappeler dans un premier temps les méthodes usuelles appliquées pour l'estimation des taux de sortie pour décès. Afin d'alléger le propos, les méthodes les plus classiques étudiées seront détaillées en Annexe B.1. Cela comprend : le lissage de Whittaker-Henderson, le lissage paramétrique par maximum de vraisemblance, le modèle de Gompert-Makeham, la fermeture par régression logit et le lissage géométrique. Il sera vu dans la suite que ces méthodes seront appliquées pour construire une table de mortalité qui servira de table de référence pour l'application d'un abattement dans le cadre de la théorie de la crédibilité.

5.1 Les phénomènes de censure et de troncature de données

5.1.1 L'estimation de taux de survie

La construction de taux de survie revient à l'étude d'une variable aléatoire T représentant une durée. Dans le cadre de la mortalité, la variable T représente une durée de vie. Estimer sa distribution à partir de l'ensemble des données d'un portefeuille permet de construire un taux moyen de décès à l'âge x noté q_x, qui peut être performant au niveau prédictif si appliqué sur un grand nombre de contrats. Cette performance prédictive résulte de la loi forte des grands nombres. Formellement, la définition du taux de sortie q_x est la suivante :

$$q_x = \mathbb{P}\left(T \in [x, x+1[| T > x\right)$$

Le taux q_x représente la probabilité de sortie (de décès dans le cas de la mortalité) d'un individu d'âge x entre l'âge x et l'âge $x+1$.

Il s'avère que les données utilisées pour estimer la distribution de T sont la plupart du temps incomplètes et il est nécessaire de prendre en compte ces manques afin de ne pas biaiser les taux de survie construits.

Deux problématiques principales ressortent lors de l'étude de T :

1. La censure à droite ;
2. La troncature à gauche.

Ces deux événements sont maintenant décrits.

5.1.2 La problématique de censure à droite

La problématique de censure à droite est liée au fait qu'un individu est rarement observé sur l'entièreté de sa vie. Généralement, une période d'observation $[t_0, t_1]$ est choisie et les réalisations de la variable T sont observées uniquement sur cette plage de temps. Cependant, les données sont incomplètes dans le sens où le décès d'un individu après t_1 est totalement invisible dans les données.

Ainsi, il convient de considérer la censure à droite dans l'estimation des taux de survie, car cela reviendrait dans le cas contraire à sous-estimer les q_x pour les âges très censurés (les âges élevés sont par exemple peu souvent observés en raison de données censurées).
Formellement, au lieu d'observer des réalisations de $T_1, T_2, ..., T_n$, les couples de variables $(X_1, \delta_1), ..., (X_n, \delta_n)$ sont observés avec :

- $X_i = min(T_i, C_i)$
- $\delta_i = \mathbf{1}_{T_i \leq C_i}$

Où la variable C représente la censure. Cette formulation signifie que seul le premier événement est observé :

- la sortie pour décès (dans le cas $T_i \leq C_i$) ;
- la sortie pour toutes autres causes, telles que le rachat, la fin d'observation, *etc.* (dans le cas $T_i > C_i$)

5.1.3 La problématique de la troncature à gauche

La problématique de la troncature à gauche provient du fait que les individus observés en portefeuille sont uniquement ceux qui ont survécu jusqu'à leur date d'entrée en période d'observation.

Dans le cas de l'assurance emprunteur, l'âge minimum possible de tous les assurés est de 18 ans. Il est donc impossible d'estimer la distribution de T pour les âges inférieurs à 18, car aucun contrat n'y a été observé.

Formellement, les observations ne correspondent pas à des réalisations de T mais bien à des réalisations de $T|T \geq \tau$. Avec τ une variable aléatoire de troncature qui peut être par exemple l'âge de l'individu.

Cette problématique est essentielle à prendre en compte dans l'étude de T, car ne pas considérer la troncature à gauche revient à sous-estimer la mortalité sur les âges peu observés car souvent tronqués (notamment pour les âges faibles).

5.2 Taux bruts et taux lisses

Les deux premières étapes de l'estimation de taux de sortie q_x sont la construction de taux bruts et de taux lissés.

Les taux bruts correspondent à la première estimation de q_x. Cette première étape se fait de façon indépendante pour tous les âges. Cependant, dans le cas de la mortalité, le risque étant faible en emprunteur, l'assureur dispose bien souvent de peu d'observations de décès. C'est pourquoi les taux bruts construits sont fortement soumis au risque d'échantillonnage (voir la section 7.6) et peuvent se retrouver biaisés.

Dans cette optique, des taux dits lissés sont construits à partir des taux bruts. Le principe est de lisser la courbe des taux bruts, qui est assez volatile, afin de diminuer les erreurs d'échantillonnage sur l'estimateur.

Des méthodes de calcul des taux bruts et lisses vont être présentées dans la suite.

5.3 La méthode de Hoem, cas classique i.i.d

La méthode de [Hoem, 1969] permet le calcul de taux bruts. Elle reprend l'estimateur naturel de q_x, à savoir la proportion de décès parmi toute une population à un certain âge. De plus, elle prend en compte les problématiques de censure à droite et de troncature à gauche, car il est considéré qu'un individu n'est à risque que sur la période où il est observé.

Les notations et les hypothèses suivantes sont introduites :

- d_i^x est la variable aléatoire qui prend la valeur 1 si l'individu i décède à l'âge x et 0 sinon. Cette variable aléatoire suit, par hypothèse, une loi de Bernoulli de paramètre q_x. Les variables $(d_i^x)_i$ sont supposées indépendantes ;
- n_x est la durée totale passée à l'âge x en personne/année de tous les individus du portefeuille.

La forme de l'estimateur de Hoem est la suivante :

$$\widehat{q_x} = \frac{\sum_{i=1}^{n_x} d_i^x}{n_x}$$

Propriétés de l'estimateur de Hoem

L'estimateur est sans biais :

$$\begin{aligned}\mathbb{E}(\widehat{q_x}) &= \mathbb{E}(\frac{\sum_{i=1}^{n_x} d_i^x}{n_x}) \\ &= \frac{n_x q_x}{n_x} \\ &= q_x\end{aligned}$$

Sa variance est égale à $\frac{q_x(1-q_x)}{n_x}$:

$$\begin{aligned}\mathbb{V}(\widehat{q_x}) &= \mathbb{V}(\frac{\sum_{i=1}^{n_x} d_i^x}{n_x}) \\ &= \frac{q_x(1-q_x)}{n_x}\end{aligned}$$

La loi asymptotique de l'estimateur est une loi normale. En effet, $\widehat{q_x}$ s'exprimant comme une somme de variables aléatoires $i.i.d$, le théorème central limite s'applique :

$$Loi(\widehat{q_x}) \underset{n_x \to +\infty}{\to} \mathcal{N}(q_x, \frac{q_x(1-q_x)}{n_x})$$

5.4 Le SMR

5.4.1 Principe de la méthode

La méthode la plus classique de construction de table de mortalité utilisant un modèle relationnel est la méthode SMR ou *Standardized mortality ratio*.
L'abattement est réalisé sur une table dite de référence, $(q_x^{ref})_x$. Il est supposé qu'il existe un coefficient α tel que les taux de mortalité du portefeuille étudié soient décrits de la façon suivante :

$$q_x = \alpha * q_x^{ref} \quad \forall x$$

Dans la pratique, ce coefficient α est sélectionné de manière à obtenir une table qui projette exactement la mortalité observée sur un portefeuille. Mathématiquement, cela s'écrit sous la forme suivante :

$$\widehat{q_x} = SMR * q_x^{ref} \quad \forall x$$

Le coefficient SMR est donc classiquement calculé de la manière suivante :

$$\widehat{\alpha} = SMR = \frac{\sum_x d_x}{\sum_x q_x^{ref} * n_x}$$

Où :

- d_x est le nombre de décès du portefeuille étudié à l'âge x ;
- n_x est la somme des expositions du portefeuille étudié à l'âge x ;
- q_x^{ref} est le taux de mortalité à l'âge x de la table de référence que l'on souhaite abattre.

Le SMR est donc en réalité un ratio observé sur attendu. En procédant de cette manière, les taux de mortalité sont construits de telle sorte à ce que la sinistralité observée soit égale à la sinistralité attendue. En effet, en nommant $\frac{O}{A}$ le ratio observé sur attendu du portefeuille sur lequel la table a été construite par SMR :

$$\begin{aligned}\frac{O}{A} &= \frac{\sum_x dx}{\sum_x q_x * n_x} \\ &= \frac{\sum_x dx}{\frac{\sum_x d_x}{\sum_x q_x^{ref} * n_x} \sum_x q_x^{ref} * n_x} \\ &= 1\end{aligned}$$

La méthode du SMR est utilisée pour la construction de tables de mortalité pour des portefeuilles au faible volume de données. Elle requiert néanmoins une table de référence à abattre, ainsi que la connaissance des décès et expositions à chaque âge sur le portefeuille.

Chapitre 6

La problématique de la maille prêt

L'étude étant menée sur un portefeuille emprunteur, il convient de prendre en compte les contraintes liées aux données dans l'estimation des taux de mortalité. En effet, une base de données emprunteur présente des dépendances interlignes, ce qui impacte directement la loi des estimateurs construits. Cette dépendance doit être prise en compte. Dans ce cadre, le théorème central limite (TCL) usuellement utilisé ne peut être appliqué. Une version modifiée du TCL est nécessaire, le théorème de Linderberg-Feller.

6.1 Le théorème de Lindeberg-Feller

Le théorème de Lindeberg-Feller est une version du théorème central limite en présence de variables aléatoires indépendantes mais non identiquement distribuées.

Énoncé

On suppose les v.a $X_1, X_2...$ indépendantes et telles que $\mathbb{E}(X_i^2) < +\infty \ \forall i$. La somme des n premières variables est notée S_n, i.e

$$S_n = \sum_{i=1}^{n} X_i$$

On suppose également qu'il ne s'agit pas du cas dégénéré où $\mathbb{V}(S_n) = 0 \ \forall n \geq 1$. Finalement, si la condition suivante est satisfaite :

$$\forall \epsilon > 0, \ \lim_{n \to +\infty} \frac{\sum_{i=1}^{n} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 \mathbf{1}_{|X_i - \mathbb{E}(X_i)| \geq \epsilon \sigma(S_n)}}{\mathbb{V}(S_n)} = 0$$

alors :

$$Loi(\frac{S_n - \mathbb{E}(S_n)}{\sigma(S_n)}) \underset{n \to +\infty}{\to} \mathcal{N}(0,1)$$

6.2 Estimateur de Hoem : présence de dépendance

Lorsque l'objectif est de construire une table de mortalité sur un produit emprunteur, les éléments de la section 5.3 ne sont pas tous immédiatement applicables. En effet, en assurance emprunteur, un individu peut avoir plusieurs prêts différents, et donc plusieurs lignes à son nom dans la base de données. Ainsi, l'indicatrice de décès dans l'estimateur de Hoem de ces individus se verra multipliée par leurs nombres de prêts. Cela engendre ainsi de la dépendance entre les variables $(d_i^x)_i$ et modifie les résultats démontrés dans le cas classique.

Cette partie a pour objectif la prise en compte de cette dépendance et la détermination des propriétés de ce nouvel estimateur.

Enfin, il convient d'expliquer pourquoi le nombre de prêts des individus est considéré et pourquoi le décès d'un individu qui possède k prêts est comptabilisé comme k décès dans l'estimateur de Hoem.

Les deux approches classiques de construction de table de mortalité en emprunteur sont les suivantes :

- **Construction à la maille tête** : tous les prêts de chaque individu sont concaténés en une seule ligne et ainsi la table est construite à partir d'une base où chaque ligne correspond à une tête ;
- **Construction à la maille prêt** : la table est construite à partir d'une base où une ligne correspond à un prêt et non à une tête. Un individu qui a plusieurs prêts aura plusieurs lignes à son nom dans la base de données et ainsi aura plus de poids dans le q_x. C'est l'approche utilisée dans cette étude, la justification est faite dans le paragraphe suivant.

Dans le cadre des méthodologies de projection, l'utilisation d'une table construite à la maille prêt s'avère préférable si ce sont des prêts/contrats qui sont projetés et non des individus. Or, dans le cadre des calculs Solvabilité 2, les équipes ALM projettent les portefeuilles à la maille prêt, ceci a motivé la décision de construire des tables à cette maille là.

En construisant une table à la maille prêt, il est supposé implicitement une corrélation non nulle entre le nombre de prêts et la probabilité de décéder. En quelque sorte, construire une table à la maille prêt revient à s'intéresser à la probabilité de décès sur un prêt à l'âge x. Là où une table à la maille tête donne la probabilité de décès d'un individu à l'âge x.

Les notations suivantes sont introduites :

- N le nombre total de prêts dans le portefeuille ;
- e_i^x la proportion de l'année passée par le prêt i à l'âge x. Si le prêt i est observé l'année complète à l'âge x, alors $e_i^x = 1$. Par définition $n_x = \sum_{i=1}^{N} e_i^x$;
- d_i^x la variable aléatoire indicatrice du décès du contrat i à l'âge x ;
- I_p l'ensemble de tous les individus ayant exactement p prêts distincts ;
- p_{max} le nombre maximal de prêts qu'un individu possède dans le portefeuille ;
- $E_{x,p}$ la somme des expositions à l'âge x au sein du groupe I_p, i.e $E_{x,p} = \sum_{i \in I_p} e_i^x$;
- $V_{x,p}$ le vecteur des expositions à l'âge x au sein du groupe I_p, i.e $V_{x,p} = (e_i^x)_{i \in I_p}$;
- $Y_{x,p}$ la variable aléatoire définie comme telle $Y_{x,p} = p\frac{\sum_{i \in I_p} d_i^x}{n_x}$.

L'hypothèse de linéarité du taux q_x sur un an est faite, i.e [1]

$$\forall x \; \forall t \in [0,1] \; {}_tq_x = t * q_x$$

Sous cette hypothèse, la loi de d_i^x peut être caractérisée :

$$d_i^x \sim Ber(e_i q_x)$$

Dans la suite, il sera supposé que tous les individus qui ont plusieurs prêts auront ces prêts sur la même période. Sous cette hypothèse, n_x peut se réécrire de la manière suivante :

$$n_x = \sum_{p=1}^{p_{max}} p \sum_{i \in I_p} e_i^x = \sum_{p=1}^{p_{max}} pE_{x,p}$$

Sous ces nouvelles notations, l'estimateur de Hoem peut se réécrire sous la forme suivante :

$$\begin{aligned} \widehat{q_x} &= \frac{\sum_{i=1}^{n_x} d_i^x}{n_x} \\ &= \sum_{p=1}^{p_{max}} p \frac{\sum_{i \in I_p} d_i^x}{n_x} \\ &= \sum_{p=1}^{p_{max}} Y_{x,p} \end{aligned}$$

Il convient également de remarquer que :

- $Y_{x,p}$ est indépendante de $Y_{x,k}$ $\forall p \neq k$ car $I_p \cap I_k = \varnothing$;
- $Y_{x,p}$ est une somme de variables aléatoires indépendantes par construction.

$Y_{x,p}$ étant une somme de variables aléatoires indépendantes et non identiquement distribuées, le théorème de Lindeberg-Feller s'applique donc ici (voir l'Annexe C.1 pour la démonstration de la condition d'application) :

$$Loi(Y_{x,p}) \underset{n_x \to +\infty}{\to} \mathcal{N}(\mathbb{E}(Y_{x,p}), \mathbb{V}(Y_{x,p}))$$

Et, par extension, par somme de lois normales indépendantes asymptotiques, la convergence en loi suivante s'applique également :

$$Loi(\widehat{q_x}) \underset{n_x \to +\infty}{\to} \mathcal{N}(\sum_{p=1}^{p_{max}} \mathbb{E}(Y_{x,p}), \sum_{p=1}^{p_{max}} \mathbb{V}(Y_{x,p}))$$

Les deux premiers moments de $Y_{x,p}$ et $\widehat{q_x}$ sont calculés ci-dessous.

1. Où, pour rappel, ${}_tq_x$ est la probabilité de décès d'un individu d'âge x entre x et $x+t$.

Espérance de $\widehat{q_x}$

Dans un premier temps :

$$\begin{aligned}
\mathbb{E}(Y_{x,p}) &= \mathbb{E}(p\frac{\sum_{i\in I_p} d_i^x}{n_x}) \\
&= p\frac{\sum_{i\in I_p} \mathbb{E}(d_i^x)}{n_x} \\
&= p\frac{\sum_{i\in I_p} q_x e_i^x}{n_x} \\
&= \frac{pq_x}{n_x}\sum_{i\in I_p} e_i^x \\
&= \frac{pq_x}{n_x}E_{x,p}
\end{aligned}$$

L'espérance de $\widehat{q_x}$ peut alors s'en déduire :

$$\begin{aligned}
\mathbb{E}(\widehat{q_x}) &= \sum_{p=1}^{p_{max}} \mathbb{E}(Y_{x,p}) \\
&= \sum_{p=1}^{p_{max}} \frac{pq_x}{n_x}E_{x,p} \\
&= \frac{q_x}{n_x}\sum_{p=1}^{p_{max}} pE_{x,p} \\
&= \frac{q_x}{n_x}n_x \\
&= q_x
\end{aligned}$$

$\widehat{q_x}$ est un estimateur sans biais, ce résultat était attendu. En revanche, étant donné la dépendance entre lignes, la variance de $\widehat{q_x}$ sera différente de celle de l'estimateur « classique » de Hoem.

Variance de $\widehat{q_x}$

Dans un premier temps :

$$\begin{aligned}
\mathbb{V}(Y_{x,p}) &= \mathbb{V}(p\frac{\sum_{i\in I_p} d_i^x}{n_x}) \\
&= p^2\frac{\sum_{i\in I_p} \mathbb{V}(d_i^x)}{n_x^2} \\
&= p^2\frac{\sum_{i\in I_p} e_i^x q_x(1-e_i^x q_x)}{n_x^2} \\
&= \frac{p^2}{n_x^2}(\sum_{i\in I_p} e_i^x q_x - \sum_{i\in I_p} q_x^2(e_i^x)^2) \\
&= \frac{p^2}{n_x^2}(q_x E_{x,p} - q_x^2||V_{x,p}||^2)
\end{aligned}$$

D'où la variance pour l'estimateur de Hoem sous dépendance :

$$\begin{aligned}\mathbb{V}(\widehat{q_x}) &= \sum_{p=1}^{p_{max}} \mathbb{V}(Y_{x,p}) \\ &= \sum_{p=1}^{p_{max}} \frac{p^2}{n_x^2}(q_x E_{x,p} - q_x^2 ||V_{x,p}||^2)\end{aligned}$$

Or, comme $e_{x,p} \leq 1 \ \forall x \ \forall p$, alors nécessairement $E_{x,p} \geq ||V_{x,p}||^2 \ \forall x \ \forall p$. De plus comme $q_x \ll 1$, alors $q_x^2 = o(q_x)$. Finalement, l'approximation suivante est obtenue :

$$\mathbb{V}(\widehat{q_x}) \approx \frac{q_x}{n_x^2} \sum_{p=1}^{p_{max}} p^2 E_{x,p}$$

Loi de l'estimateur de Hoem

En conclusion, l'estimateur de Hoem sous dépendance suit asymptotiquement une loi normale :

$$Loi(\widehat{q_x}) \underset{n_x \to +\infty}{\to} \mathcal{N}(q_x, \frac{q_x}{n_x^2} \sum_{p=1}^{p_{max}} p^2 E_{x,p})$$

Avec la prise en compte de la dépendance entre les lignes, un volume restreint d'informations est disponible pour le calcul de l'estimateur, il est donc attendu que la variance estimée ici soit supérieure à la variance de l'estimateur classique dans le cas *i.i.d.*

Chapitre 7

Effet sélection médicale

À présent que le contexte et les bases statistiques sont posés, une méthode permettant d'isoler et d'analyser l'effet de sélection médicale est proposée. C'est la première étude nécessaire pour quantifier la loi Lemoine, qui, pour rappel, induit la suppression des formalités médicales pour les assurés éligibles. [1]

7.1 Contexte

Comme il a été décrit dans la section 1.4, une sélection des risques est souvent opérée à l'entrée en assurance emprunteur. Cette sélection engendre alors une sous-mortalité observée sur les premières années de contrat d'un individu. Dans la suite, les notations suivantes seront posées :

- d_i^x la variable aléatoire indicatrice du décès de l'individu i à l'âge x ;
- e_i^x la fraction d'année passée à l'âge x de l'individu i ;
- A_i la variable aléatoire représentant l'ancienneté de l'individu i dans le portefeuille ;
- $(q_x)_{x\in[\![x_{min};x_{max}]\!]}$ la table de mortalité qui décrit la sinistralité d'une population ayant été soumise à une sélection médicale à l'entrée ;
- $(q_x^*)_{x\in[\![x_{min};x_{max}]\!]}$ la table de mortalité qui décrit la sinistralité de la même population en « retirant » l'effet de sélection médicale.

Naturellement, $q_x \leq q_x^* \ \forall x \in [\![x_{min}; x_{max}]\!]$, car la sélection médicale permet une meilleure sélection des risques et de diminuer la sinistralité moyenne observée.
Dans la suite, l'hypothèse fondatrice suivante sera faite :

Hypothèse sur la sélection médicale
L'effet sélection médicale disparait après λ années sous garanties.

Autrement dit :

$$\mathbb{E}[d_i^x | A_i \geq \lambda] = q_x^* e_i^x$$

ou alors :

$$Loi(d_i^x | A_i \geq \lambda) = Ber(q_x^* e_i^x)$$

1. Se référer au chapitre 2 pour les critères d'éligibilité.

Pour rappel :

$$\mathbb{E}[d_i^x] = q_x e_i^x$$

et :

$$Loi(d_i^x) = Ber(q_x e_i^x)$$

À présent, une démarche est présentée pour l'estimation de λ.

7.2 Test statistique : estimation du paramètre de la sélection médicale λ

Notations

Les notations suivantes sont posées :

- ${}^{j}\mathbf{q} = \left({}^{j}q_x\right)_{x\in[\![x_{min};x_{max}]\!]}$ la table de mortalité qui décrit la sinistralité de la population assurée après j années d'ancienneté ;
- ${}^{j}\widehat{\mathbf{q}} = \left({}^{j}\widehat{q_x}\right)_{x\in[\![x_{min};x_{max}]\!]}$ l'estimateur de la table $\left({}^{j}q_x\right)_{x\in[\![x_{min};x_{max}]\!]}$. Cet estimateur sera supposé sans biais dans la suite ;

Par définition :

$$\left({}^{0}q_x\right)_{x\in[\![x_{min};x_{max}]\!]} := (q_x)_{x\in[\![x_{min};x_{max}]\!]}$$

et

$$\left({}^{\lambda}q_x\right)_{x\in[\![x_{min};x_{max}]\!]} := (q_x^*)_{x\in[\![x_{min};x_{max}]\!]}$$

- ${}^{j-1}\widehat{\Delta}^{j}$ la moyenne pondérée par l'exposition des différences entre les $\left({}^{j}\widehat{q_x}\right)_{x\in[\![x_{min};x_{max}]\!]}$ et les $\left({}^{j-1}\widehat{q_x}\right)_{x\in[\![x_{min};x_{max}]\!]}$.

Ainsi :

$${}^{j-1}\widehat{\Delta}^{j} = \frac{\sum_{x=x_{min}}^{x_{max}} n_x \left({}^{j}\widehat{q_x} - {}^{j-1}\widehat{q_x}\right)}{\sum_{x=x_{min}}^{x_{max}} n_x}$$

et

$$\mathbb{E}[{}^{j-1}\widehat{\Delta}^{j}] = {}^{j-1}\Delta^{j} = \frac{\sum_{x=x_{min}}^{x_{max}} n_x \left({}^{j}q_x - {}^{j-1}q_x\right)}{\sum_{x=x_{min}}^{x_{max}} n_x}$$

Démarche pour l'estimation du paramètre λ

Premièrement, il est à noter que le paramètre λ n'est à priori pas global pour tout le marché et doit être estimé pour chaque portefeuille étudié.

Étape 1

Estimation de $\left(\left({}^{j}q_x \right)_{x \in [\![x_{min};x_{max}]\!]} \right)_{j \in [\![0;j_{max}]\!]}$ où j_{max} représente l'ancienneté maximum d'un prêt sur le portefeuille.

Étape 2

Sélection de l'estimateur de λ tel que :

$$\widehat{\lambda} = min \left\{ j \in [\![0; j_{max}]\!] \;\middle|\; {}^{j-1}\widehat{\Delta}^{j}_{obs} \in \overline{W}_j \right\} - 1$$

avec

$$\overline{W}_j = \left[q^j_{\frac{\alpha}{2}}; q^j_{1-\frac{\alpha}{2}} \right]$$

où

$$q^j_\alpha = z_\alpha * \sqrt{\frac{\sum_{x=x_{min}}^{x_{max}} n_x^2 \frac{\left(n_x^{j-1} - n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^j E^j_{x,p}\right) - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^{j-1} E^{B_{j-1} \setminus B_j}_{x,p}\right)}{(n_x^j n_x^{j-1})^2}}{\left(\sum_{x=x_{min}}^{x_{max}} n_x\right)^2}}$$

avec α le risque de première espèce et z_α le quantile de niveau α de la loi normale centrée réduite. Les autres notations introduites seront définies dans la partie suivante. Le détail de ces démarches et les preuves associées sont maintenant présentées.

Explications et détails des étapes de l'estimation de λ

Étape 1

La première étape est l'estimation de $\left(\left({}^{j}q_x \right)_{x \in [\![x_{min};x_{max}]\!]} \right)_{j \in [\![0;j_{max}]\!]}$. La table $\left({}^{j}q_x \right)_{x \in [\![x_{min};x_{max}]\!]}$ représente la mortalité de la population assurée [2] au-delà de j années passées en portefeuille. L'estimation de ces taux se fait donc en procédant à la troncature des expositions et à la suppression des sinistres du portefeuille sur les j premières années de chaque contrat. Après troncature, une partie de l'exposition et des sinistres a été retirée, et il ne reste plus qu'à estimer les taux de mortalité sur la base tronquée.

Ci-dessous, un schéma explicatif du processus de troncature sur les données :

2. La population étant ici pour rappel des contrats/prêts.

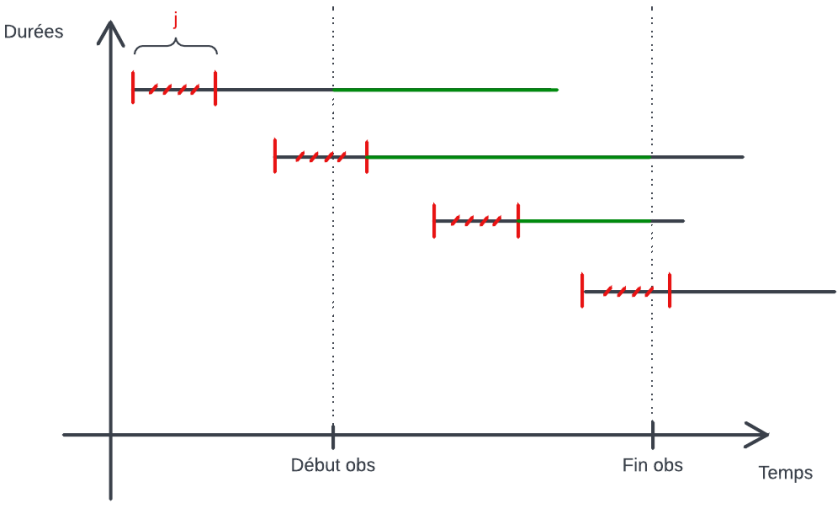


FIGURE 7.1 – Troncature des données pour l'estimation de la mortalité au-delà de j années

Pour chaque contrat, l'exposition est tronquée et les sinistres éventuels retirés sur les j premières années. La zone verte de la figure 7.1 représente la partie de l'information conservée sur chaque contrat pour la construction de la table.

Étape 2

Afin d'estimer λ, il est nécessaire de déterminer la durée j pour laquelle l'effet de sélection médicale s'est estompé, c'est-à-dire le moment où ${}^{j}\mathbf{q} \sim {}^{j-1}\mathbf{q}$. Pour pouvoir quantifier ce $\ll \sim \gg$, le test d'hypothèse suivant est effectué :

$$\mathcal{H}_0 : \mathbb{E}[{}^{j}\widehat{\mathbf{q}}] = \mathbb{E}[{}^{j-1}\widehat{\mathbf{q}}]$$

contre

$$\mathcal{H}_1 : \mathbb{E}[{}^{j}\widehat{\mathbf{q}}] \neq \mathbb{E}[{}^{j-1}\widehat{\mathbf{q}}]$$

La démarche qui va être effectuée pour mener à bien ce test est décrite dans
[Maumy-Bertrand et Bertrand, 2018, p.295]. Le test choisi est bilatéral, la zone d'acceptation définie va donc être encadrée par deux zones de rejet. La statistique de test et l'étude de sa loi sont maintenant détaillées.
La statistique de test sélectionnée est la suivante :

$$\begin{aligned} {}^{j-1}\widehat{\Delta}^{j} &= \frac{\sum_{x=x_{min}}^{x_{max}} n_x \left({}^{j}\widehat{q_x} - {}^{j-1}\widehat{q_x}\right)}{\sum_{x=x_{min}}^{x_{max}} n_x} \\ &= \frac{\sum_{x=x_{min}}^{x_{max}} n_x \, {}^{j-1}\widehat{\Delta}_x^{j}}{\sum_{x=x_{min}}^{x_{max}} n_x} \end{aligned}$$

Avec ${}^{j-1}\widehat{\Delta}_x^{j} = \left({}^{j}\widehat{q_x} - {}^{j-1}\widehat{q_x}\right)$.
Il convient à présent d'étudier la loi de la variable de décision ${}^{j-1}\widehat{\Delta}^{j}$. Les variables ${}^{j}\widehat{q_x}$ et ${}^{j-1}\widehat{q_x}$ étant fortement dépendantes par construction, le résultat n'est pas trivial.
Pour étudier cette variable aléatoire, les notations suivantes sont précisées :

$${}^{j}\widehat{q_x} = \frac{\sum_{i \in B_j} d_i^x}{n_x^j}$$

avec B_j l'ensemble des prêts tronqués sur les j premières années de contrat. Il est à noter que $B_j \subseteq B_{j-1}$. La variable ${}^{j-1}\widehat{\Delta}_x^j$ peut alors se ré-écrire de la manière suivante :

$$\begin{aligned}
{}^{j-1}\widehat{\Delta}_x^j &= {}^{j}\widehat{q_x} - {}^{j-1}\widehat{q_x} \\
&= \frac{\sum_{i\in B_j} d_i^x}{n_x^j} - \frac{\sum_{i\in B_{j-1}} d_i^x}{n_x^{j-1}} \\
&= \frac{n_x^{j-1}\sum_{i\in B_j} d_i^x - n_x^j \sum_{i\in B_{j-1}} d_i^x}{n_x^j n_x^{j-1}} \\
&= \frac{\left(n_x^{j-1} - n_x^j\right)\sum_{i\in B_j\cap B_{j-1}} d_i^x - n_x^j \sum_{i\in B_{j-1}\setminus B_j} d_i^x}{n_x^j n_x^{j-1}} \\
&= \frac{\left(n_x^{j-1} - n_x^j\right)\sum_{p=1}^{p_{max}} p\sum_{i\in B_j\cap I_p} d_i^x - n_x^j \sum_{p=1}^{p_{max}} p\sum_{i\in (B_{j-1}\setminus B_j)\cap I_p} d_i^x}{n_x^j n_x^{j-1}}
\end{aligned}$$

Par construction, une somme de variables aléatoires indépendantes apparaît. Par application du théorème de Lindeberg-Feller [3], cette variable suit asymptotiquement une loi normale. Pour rappel, sous $\mathcal{H}_0$, $\mathbb{E}[\widehat{\mathbf{q}}_j] = \mathbb{E}[\widehat{\mathbf{q}}_{j-1}]$, l'espérance de ${}^{j-1}\widehat{\Delta}_x^j$ est donc nulle sous $\mathcal{H}_0$.
La variance de cette variable aléatoire est maintenant développée :

$$\begin{aligned}
\mathbb{V}[{}^{j-1}\widehat{\Delta}_x^j] &= \mathbb{V}[\frac{\left(n_x^{j-1} - n_x^j\right)\sum_{p=1}^{p_{max}} p\sum_{i\in B_j\cap I_p} d_i^x - n_x^j \sum_{p=1}^{p_{max}} p\sum_{i\in (B_{j-1}\setminus B_j)\cap I_p} d_i^x}{n_x^j n_x^{j-1}}] \\
&= \frac{\left(n_x^{j-1} - n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2 \sum_{i\in B_j\cap I_p} \mathbb{V}[d_i^x] - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2 \sum_{i\in (B_{j-1}\setminus B_j)\cap I_p} \mathbb{V}[d_i^x]}{(n_x^j n_x^{j-1})^2}
\end{aligned}$$

par indépendance. D'où :

$$\begin{aligned}
\mathbb{V}[{}^{j-1}\widehat{\Delta}_x^j] &= \frac{\left(n_x^{j-1} - n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2 \sum_{i\in B_j\cap I_p} e_i q_x^j(1-e_i q_x^j) - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2 \sum_{i\in (B_{j-1}\setminus B_j)\cap I_p} e_i q_x^{j-1}(1-e_i q_x^{j-1})}{(n_x^j n_x^{j-1})^2} \\
&= \frac{\left(n_x^{j-1} - n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^j E_{x,p}^j - (q_x^j)^2 ||V_{x,p}^j||^2\right) - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^{j-1} E_{x,p}^{B_{j-1}\setminus B_j} - (q_x^{j-1})^2 ||V_{x,p}^{B_{j-1}\setminus B_j}||^2\right)}{(n_x^j n_x^{j-1})^2}
\end{aligned}$$

Avec :

- $E_{x,p}^j = \sum_{i\in B_j\cap I_p} e_i$
- $V_{x,p}^j = (e_i)_{i\in B_j\cap I_p}$
- $E_{x,p}^{B_{j-1}\setminus B_j} = \sum_{i\in (B_{j-1}\setminus B_j)\cap I_p} e_i$
- $V_{x,p}^{B_{j-1}\setminus B_j} = (e_i)_{i\in (B_{j-1}\setminus B_j)\cap I_p}$

De plus, l'approximation $(q_x^j)^2 << q_x^j$ est faite, il vient donc que $(q_x^j)^2 = o(q_x^j)$. Sous $\mathcal{H}_0$, ${}^{j-1}\widehat{\Delta}_x^j$ suit finalement asymptotiquement la loi suivante :

$${}^{j-1}\widehat{\Delta}_x^j \sim \mathcal{N}\left(0, \frac{\left(n_x^{j-1} - n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^j E_{x,p}^j\right) - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2 \left(q_x^{j-1} E_{x,p}^{B_{j-1}\setminus B_j}\right)}{(n_x^j n_x^{j-1})^2}\right)$$

Pour rappel,

$${}^{j-1}\widehat{\Delta}^j = \frac{\sum_{x=x_{min}}^{x_{max}} {n_x}^{j-1}\widehat{\Delta}_x^j}{\sum_{x=x_{min}}^{x_{max}} n_x}$$

3. La démonstration de la condition d'application est similaire à celle déjà présentée en Annexe C.1.

En supposant l'indépendance entre les âges, ce qui est une hypothèse forte, et par somme de lois normales asymptotiques indépendantes, ${}^{j-1}\widehat{\Delta}^{j}$ suit la loi suivante sous $\mathcal{H}_0$:

$$ {}^{j-1}\widehat{\Delta}^{j} \sim \mathcal{N}\left(0, \frac{\sum_{x=x_{min}}^{x_{max}} n_x^2 \frac{\left(n_x^{j-1}-n_x^j\right)^2 \sum_{p=1}^{pmax} p^2\left(q_x^j E_{x,p}^j\right) - (n_x^j)^2 \sum_{p=1}^{pmax} p^2\left(q_x^{j-1} E_{x,p}^{B_{j-1}\setminus B_j}\right)}{(n_x^j n_x^{j-1})^2}}{\left(\sum_{x=x_{min}}^{x_{max}} n_x\right)^2}\right) $$

Le nombre d'années au-delà duquel l'effet de sélection médicale s'estompe, noté λ, peut donc s'estimer comme le premier seuil $j-1$ où la réalisation de ${}^{j-1}\widehat{\Delta}^{j}$, notée ${}^{j-1}\widehat{\Delta}^{j}_{obs}$, appartient à la zone d'acceptation $\overline{W}_j$, déterminée en fonction de α et des paramètres la loi normale ci-dessus. Formellement, l'estimateur de λ s'exprime alors comme tel :

$$ \widehat{\lambda} = min\left\{ j \in [\![0; j_{max}]\!] \;\middle|\; {}^{j-1}\widehat{\Delta}^{j}_{obs} \in \overline{W}_j \right\} - 1 $$

avec

$$ \overline{W}_j = \left[q^j_{\frac{\alpha}{2}}; q^j_{1-\frac{\alpha}{2}} \right] $$

où

$$ q^j_\alpha = z_\alpha * \sqrt{\frac{\sum_{x=x_{min}}^{x_{max}} n_x^2 \frac{\left(n_x^{j-1}-n_x^j\right)^2 \sum_{p=1}^{pmax} p^2\left(q_x^j E_{x,p}^j\right) - (n_x^j)^2 \sum_{p=1}^{pmax} p^2\left(q_x^{j-1} E_{x,p}^{B_{j-1}\setminus B_j}\right)}{(n_x^j n_x^{j-1})^2}}{\left(\sum_{x=x_{min}}^{x_{max}} n_x\right)^2}} $$

Le « -1 » dans l'estimateur de λ provient du fait que s'il est observé que la mortalité est la même au-delà de $j-1$ années et au-delà de j années de contrat, alors l'effet de la sélection médicale est observable uniquement sur les $j-1$ premières années.

L'estimation de λ se fait donc en calculant successivement les grandeurs ${}^{0}\widehat{\Delta}^{1}_{obs}$, ${}^{1}\widehat{\Delta}^{2}_{obs}$... observées sur le portefeuille, puis en regardant si elles appartiennent aux zones de rejets successives $\overline{W}_1, \overline{W}_2, ...$ jusqu'à ce que $\mathcal{H}_0$ ne soit pas refusée. À ce moment, l'hypothèse $\mathcal{H}_0$ est conservée par défaut au seuil α et le risque associé à cette décision est un risque d'erreur de deuxième espèce. [4]

7.3 Approfondissement de la démarche : le *bootstrap*

La démarche du test statistique présentée ci-dessus peut être approfondie par une méthode de *bootstrap* pour en améliorer la précision.
Le nombre total de prêts dans la base de données étudiée est noté N. Le risque d'erreur de première espèce α est fixé. La démarche suivie sera la suivante :

- Tirer aléatoirement N prêts avec remise dans la base de données étudiée ;
- Calculer la zone de d'acceptation $\overline{W}_j$. La valeur de cette zone associée au tirage t sera notée $\overline{W}_j^{(t)}$;
- Calculer la valeur de la statistique de test observée ${}^{j-1}\widehat{\Delta}^{j}_{obs}$ sur l'échantillon tiré. La valeur observée de cette statistique associée au tirage t sera notée ${}^{j-1}\widehat{\Delta}^{j}_{obs}{}^{(t)}$;
- Répéter les deux premières étapes M fois ;

4. Ce risque est dans la pratique noté β, il est associé à la puissance du test $1-\beta$. Son calcul ne sera pas détaillé dans la suite. Pour plus d'informations, se référer à [Maumy-Bertrand et Bertrand, 2018].

- Calculer la p-valeur :

$$\text{p-valeur} = \frac{1}{M}\sum_{t=1}^{M} \mathbf{1}_{{}^{j-1}\widehat{\Delta}^{j}_{obs}{}^{(t)} \in \overline{W}_j^{(t)}}$$

- Si la p-valeur est inférieure ou égale à α, alors le test est significatif au seuil α. L'hypothèse nulle $\mathcal{H}_0$ est rejetée au seuil α et l'hypothèse $\mathcal{H}_1$ est décidée vraie au seuil α. Le risque associé à cette décision est un risque d'erreur de première espèce qui vaut α.
  Si la p-valeur est supérieure à α, alors le test n'est pas significatif au seuil α. L'hypothèse nulle $\mathcal{H}_0$ est conservée par défaut au seuil α. Le risque associé à cette décision est un risque d'erreur de deuxième espèce qui vaut β.

La méthodologie globale d'estimation de l'effet de sélection médicale est entièrement décrite, elle peut à présent être appliquée aux données.

7.4 Résultats de l'estimation du paramètre λ

À présent, la méthodologie d'estimation du paramètre λ, présentée dans le chapitre 7 est mise en œuvre. Pour rappel, le facteur λ représente, pour un portefeuille donné, le nombre d'années sur lesquelles l'effet de sélection médicale est observé.
Le risque de première espèce α est fixé à 5%.

Les résultats présentés ci-dessous sont issus de l'application des méthodes proposées sur les données du portefeuille A. [5]

La première étape de l'algorithme par *bootstrap* est faite en posant l'hypothèse :

$$\mathcal{H}_0 : \mathbb{E}[{}^{0}\widehat{q}] = \mathbb{E}[{}^{1}\widehat{q}]$$

contre

$$\mathcal{H}_1 : \mathbb{E}[{}^{0}\widehat{q}] \neq \mathbb{E}[{}^{1}\widehat{q}]$$

Le nombre de tirages M est fixé à 1000. En nommant N la taille de la base du produit A, M tirages avec remise de N lignes sont tirés dans la base de données. À chaque tirage, la condition suivante est regardée :

$${}^{j-1}\widehat{\Delta}^{j}_{obs} \in \overline{W}_j$$

avec

$$\overline{W}_j = \left[q^j_{\frac{\alpha}{2}}; q^j_{1-\frac{\alpha}{2}}\right]$$

où

$$q^j_\alpha = z_\alpha * \sqrt{\frac{\sum_{x=x_{min}}^{x_{max}} n_x^2 \frac{\left(n_x^{j-1}-n_x^j\right)^2 \sum_{p=1}^{p_{max}} p^2\left(q_x^j E^j_{x,p}\right) - (n_x^j)^2 \sum_{p=1}^{p_{max}} p^2\left(q_x^{j-1} E^{B_{j-1}\setminus B_j}_{x,p}\right)}{(n_x^j n_x^{j-1})^2}}{\left(\sum_{x=x_{min}}^{x_{max}} n_x\right)^2}}$$

Si la condition ci-dessus est vérifiée, alors $\mathcal{H}_0$ ne peut être rejetée. La p-valeur calculée ci-dessous représente le nombre de fois sur les M tirages où $\mathcal{H}_0$ a été conservée par défaut au seuil $\alpha = 5\%$:

$$\text{p-valeur} = \frac{1}{M}\sum_{t=1}^{M} \mathbf{1}_{{}^{j-1}\widehat{\Delta}^{j}_{obs}{}^{(t)} \in \overline{W}_j^{(t)}}$$

5. Se référer à la section 4.2 pour la justification.

Le résultat sur le produit A pour le premier test *bootstrap* est le suivant :

$$\text{p-valeur} = 0$$

La p-valeur est inférieure ou égale à $\alpha = 5\%$, le test est significatif au seuil 5%. L'hypothèse nulle $\mathcal{H}_0$ est rejetée au seuil 5% et l'hypothèse $\mathcal{H}_1$ est décidée vraie au seuil 5%. Le risque associé à cette décision est un risque d'erreur de première espèce qui vaut 5%. Cela signifie que l'effet de sélection médicale est significativement [6] observable la première année de chaque contrat du portefeuille. A présent, le test est réalisé sous la nouvelle hypothèse :

$$\mathcal{H}_0 : \mathbb{E}[{}^1\widehat{q}] = \mathbb{E}[{}^2\widehat{q}]$$

contre

$$\mathcal{H}_1 : \mathbb{E}[{}^1\widehat{q}] \neq \mathbb{E}[{}^2\widehat{q}]$$

La p-valeur calculée suite au *bootstrap* est la suivante :

$$\text{p-valeur} = 0,007$$

La p-valeur est inférieure ou égale à $\alpha = 5\%$, le test est significatif au seuil 5%. L'hypothèse nulle $\mathcal{H}_0$ est rejetée au seuil 5% et l'hypothèse $\mathcal{H}_1$ est décidée vraie au seuil 5%. Le risque associé à cette décision est un risque d'erreur de première espèce qui vaut 5%. L'effet de sélection médicale est significativement observable la deuxième année de chaque contrat sur le produit A. Finalement, le test est réalisé sous la nouvelle hypothèse :

$$\mathcal{H}_0 : \mathbb{E}[{}^2\widehat{q}] = \mathbb{E}[{}^3\widehat{q}]$$

contre

$$\mathcal{H}_1 : \mathbb{E}[{}^2\widehat{q}] \neq \mathbb{E}[{}^3\widehat{q}]$$

La p-valeur calculée suite au *bootstrap* est la suivante :

$$\text{p-valeur} = 0,899$$

La p-valeur est supérieure à $\alpha = 5\%$, le test n'est pas significatif au seuil 5%. L'hypothèse nulle $\mathcal{H}_0$ est conservée par défaut au seuil 5%. Le risque associé à cette décision est un risque d'erreur de deuxième espèce qui vaut β. L'effet de sélection médicale n'est significativement [7] visible que sur les deux premières années de chaque contrat sur le portefeuille du produit A.
Dans la suite de cette étude, il sera posé :

$$\widehat{\lambda} = 2$$

Pour plus de détails sur le comportement de la p-valeur en fonction du nombre de tirages pour chacun des trois tests *bootstrap* effectués, voir les graphiques en Annexe D.1.

Ce résultat signifie que l'hypothèse « l'effet de sélection médicale est visible uniquement sur les deux premières années » n'a pas pu être rejetée avec un niveau de confiance à 95%.

6. avec un niveau de confiance à 95%.
7. avec un niveau de confiance à 95%.

7.5 Apport de la crédibilité à l'estimation de l'impact loi Lemoine

Comme cela a été expliqué dans le chapitre 2, le deuxième titre de la loi énonce l'annulation des formalités médicales pour les individus remplissant certaines conditions d'éligibilités.

Pour rappel, dans les développements du chapitre 7, le facteur λ a été introduit et une méthode a été proposée pour son estimation. Le facteur λ est le nombre d'années sur lequel l'effet de sélection médicale est observé sur la sinistralité du portefeuille.

Or, la mesure de l'impact de l'absence future de sélection médicale doit passer par la projection de la future mortalité des portefeuilles emprunteur soumis à cette loi. Il convient donc de construire une loi qui n'intègre pas d'effet de sélection médicale. Le candidat idéal pour cette loi a été présenté dans le chapitre 7, il s'agit de la loi :

$$\left({}^{\lambda}q_x\right)_{x\in[\![x_{min};x_{max}]\!]}$$

Pour rappel, cette table de mortalité décrit la mortalité d'une population assurée après que l'effet de sélection médicale se soit estompé. Il est donc naturel de penser qu'elle pourra décrire une partie de la sinistralité future d'un portefeuille *new Business* post loi Lemoine, où une grande proportion des assurés n'auront pas passé de formalités médicales.

Or, après avoir estimé λ, il a été expliqué que la table $\left({}^{\widehat{\lambda}}q_x\right)$ était construite à partir de données de prêts tronqués sur les $\widehat{\lambda}$ premières années de contrat (voir la figure Troncature des données pour l'estimation de la mortalité au-delà de j années pour détails sur la troncature).

Cependant, le fait de tronquer les $\widehat{\lambda}$ premières années de chaque contrat revient à grandement diminuer le volume de données à disposition pour la construction de la table. C'est dans ce cadre que la théorie de la crédibilité s'avère utile : elle permet la construction de tables de mortalité en considérant un faible volume de données disponible. C'est l'apport de la théorie de la crédibilité à cette étude : permettre la construction d'une loi de mortalité à partir de données fortement tronquées, et donc faible en volume, pour modéliser le comportement futur d'une partie d'un portefeuille de *new business* post loi Lemoine. Les modèles de crédibilité utilisés seront développés dans la suite.

Le risque en construisant des estimateurs statistiques basés sur un faible volume de données se nomme le risque d'échantillonnage. Afin de mieux cerner l'utilité de la théorie de la crédibilité, il convient à présent de définir ce risque.

7.6 Le risque d'échantillonnage

Selon [Maumy-Bertrand et Bertrand, 2018], la statistique est définie comme « la discipline des mathématiques qui a pour objet les méthodes qui permettent de collecter et d'analyser les données empiriques et d'en extraire des statistiques ».

Une autre définition donnée par la même source est la suivante : « Les statistiques sont des données numériques qui interviennent pratiquement dans tous les domaines d'activité : gestion financière (états, banques, assurances, entreprises...), démographie, contrôles de qualité, études de marché, sciences expérimentales (biologie, psychologie...) ».

Quelle que soit la définition prise, l'objet qui est au cœur du concept de statistique est la donnée. Les données permettent l'inférence de par leur qualité et leur volume.

Cependant, le statisticien n'aura pas toujours accès à un volume suffisant de données pour ses études. Ce manque d'informations se répercute sur la précision des estimateurs qu'il construit. Un faible échantillon peut ne pas être représentatif de la population ou du phénomène étudié. Ce risque s'appelle le risque d'échantillonnage. Il convient de le prendre en compte, ou, à minima, de le quantifier par des intervalles de confiance, lors de toute étude statistique.

Le risque d'échantillonnage est maintenant illustré par un exemple simple. 500 réalisations d'une loi normale centrée réduite X sont simulées. Un sous-échantillon de taille 30 est tiré parmi ces 500 réalisations. L'objectif est d'estimer l'espérance de X uniquement à partir du sous-échantillon. Ici, la faible taille des données utilisée pour estimer $\mathbb{E}[X]$ entraîne une erreur d'échantillonnage. Le graphique ci-dessous représente les 500 réalisations de X, les 30 données du sous-échantillon [8], ainsi que la vraie espérance de X, à savoir 0, et la moyenne empirique sur le sous-échantillon de taille 30.

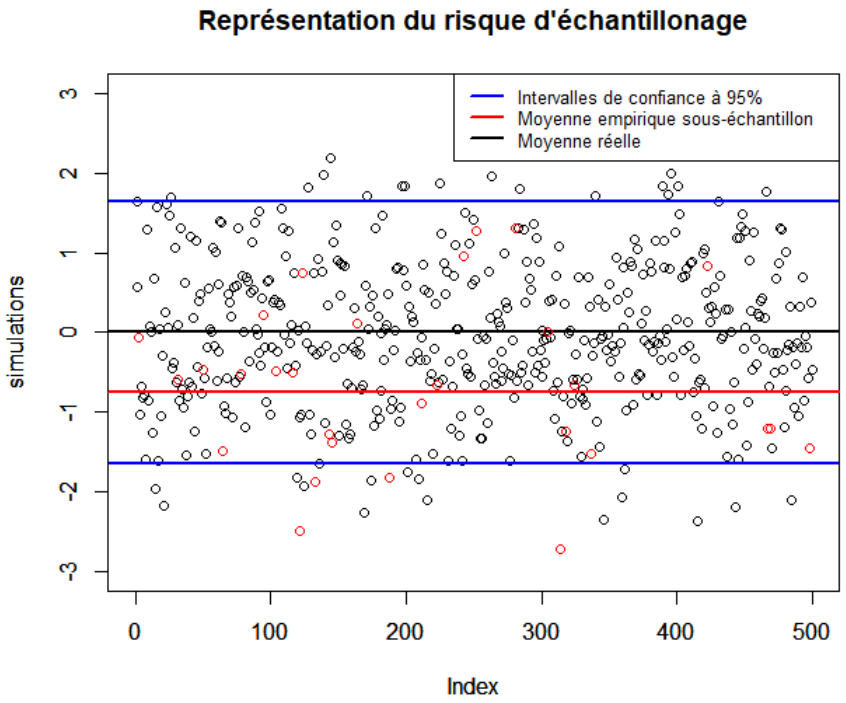


FIGURE 7.2 – Exemple : risque d'échantillonnage

Avec si peu de données pour estimer l'espérance de X, la moyenne empirique sur le sous-échantillon est de $-0,74$. Dans le cadre de l'estimation de lois en actuariat, il faut garder à l'esprit que le peu de données peut entraîner un biais non négligeable. Ce fait est d'autant plus vrai pour la construction de tables de mortalité. En effet, le risque étant relativement faible sur les populations assurées, le nombre de sinistres observé n'est parfois pas suffisant pour estimer de façon fiable les taux de mortalité.

La théorie de la crédibilité classique, présentée en Annexe E, permet de définir des seuils minimaux de données à atteindre afin de limiter cette erreur d'échantillonnage. Dans le cas de la mortalité, l'application de cette théorie indique qu'il est nécessaire d'observer au minimum 1082 sinistres pour minimiser le risque d'échantillonnage.

Dans le cadre de cette étude sur la loi Lemoine, étant donné la troncature partielle des données sur les 2 [9] premières années de contrat, le volume de données sera insuffisant et bien inférieur au seuil

8. Les 30 points du sous-échantillon sont représentés en rouge.
9. Il est rappelé que le facteur λ a été estimé à 2 dans la section 7.4.

minimal de 1082 sinistres. La théorie de la crédibilité de Bühlmann permet de répondre à ces limites et de construire des estimateurs des taux de mortalité plus fiables que ceux des méthodes classiques.

Chapitre 8

Méthode bayésienne empirique de Bühlmann

Comme cela a été décrit, la théorie de la crédibilité va être appliquée pour la construction d'une table de mortalité qui ne prend pas en compte l'effet de sélection médicale. Pour rappel, ces taux de mortalité sont notés ${}^{\lambda}q_x$, où λ a été estimé à 2 et représente le nombre d'années sur lesquelles l'effet de sélection médicale est significativement observable. Les méthodes présentées dans la suite vont permettre la construction de cette table de mortalité sur le portefeuille tronqué du produit A. [1].
Les grands principes du modèle décrit par [Klugman *et al.*, 2009] sont repris dans ce chapitre puis les estimateurs sont adaptés à la contrainte des données emprunteur. [2]. Stuart Klugman est un actuaire enseignant chercheur exerçant à l'heure actuelle aux États-Unis. Il reprend dans l'article mentionné ci-dessus le modèle de Bühlmann, en l'adaptant au cas de l'estimation de la mortalité. Hans Bühlmann est un mathématicien Suisse du 20e siècle, il est notamment réputé pour ses travaux sur la théorie de la crédibilité.

8.1 Philosophie de la méthode

L'objectif est l'estimation de la mortalité sur un portefeuille h^*, au sein duquel le volume de données est supposé faible. La construction de la table est réalisée par abattement sur une table de référence q_x^{ref}.

La philosophie du modèle est l'utilisation d'un ensemble de portefeuilles $1, 2, ..., r$ avec $h^* \leq r$ pour la construction de la table h^*. Il sera supposé que la mortalité de l'ensemble des portefeuilles peut s'exprimer comme le produit entre un certain coefficient (propre à chaque portefeuille) et la table de référence. i.e

$$q_x^h = m_h q_x^{ref} \quad \forall h \in [\![1; r]\!] \quad \forall x$$

De plus, le coefficient m_h est supposé aléatoire pour tout h, de moyenne μ et de variance σ^2. C'est par cette hypothèse que l'ensemble des portefeuilles pourront être utilisés pour l'estimation de la mortalité sur le portefeuille h^*. En pratique, les variables aléatoires m_h sont déjà réalisées et il est cherché à estimer leur réalisation pour chaque portefeuille grâce au ratio SMR (noté $\widehat{m_h}$ dans la suite). En calculant ce $\widehat{m_h}$ sur l'ensemble des r portefeuilles, les vraies réalisations des m_h peuvent être estimées et ainsi il est possible d'inférer sur les coefficients μ et σ^2.

Concernant le portefeuille h^*, le faible volume implique un risque d'échantillonnage dans l'estimation du m_h^* réalisé et potentiellement un écart entre $\widehat{m_h^*}$ et m_h^* réalisé. L'estimateur final du m_h^* réalisé utilisé

1. Se référer à la section 4.2.
2. Se référer au chapitre 6 pour plus de détails.

dans le modèle sera noté $\widetilde{m_h^*}$ et sera exprimé comme une combinaison linéaire entre l'estimateur biaisé par le manque de données $\widehat{m_h^*}$ (le SMR) et l'espérance de la variable m_h^* : μ.

Finalement, il sera préférable dans le modèle de choisir des portefeuilles avec un volume de données suffisant, de façon à limiter l'erreur d'échantillonnage dans le calcul des SMR : $(\widehat{m_h})_{h\in[\![1;r]\!]\setminus h^*}$. Le choix des portefeuilles fera l'objet d'une partie détaillée dans la suite.

Le schéma ci-dessous illustre la situation initiale à laquelle le statisticien est confronté dans le processus d'estimation de la mortalité sur le portefeuille h^*.

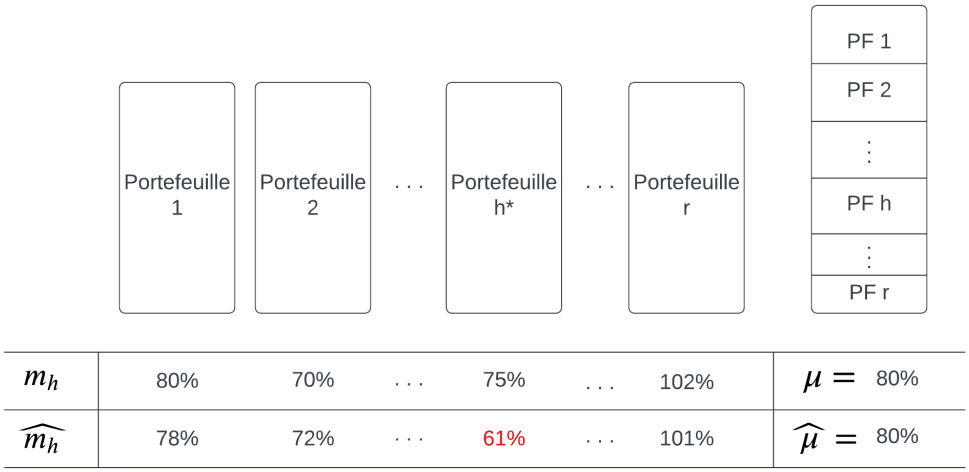


FIGURE 8.1 – Méthode bayésienne empirique de Bühlmann

où la ligne m_h représente les réalisations inconnues de la variable aléatoire m_h et où $\widehat{m_h}$ représente l'estimation par le SMR de ces réalisations.

Dans cet exemple, il est clair que retenir l'estimateur classique $\widehat{m_h^*}$ reviendrait à fortement sous-estimer la mortalité réelle sur le produit h^* (une sous-estimation de l'ordre de 14% ici). En effet, la confiance accordée au SMR sur le portefeuille h^* est relativement discutable dans la mesure où le risque d'échantillonnage est fort sur ces données.

L'estimateur final retenu dans la méthode bayésienne empirique de Bühlmann sera alors un point sur le segment reliant $\widehat{m_h^*}$ et la moyenne μ, .i.e

$$\widetilde{m_h^*} = Z^{h^*}\widehat{m_h^*} + (1 - Z^{h^*})\mu$$

La représentation des différentes valeurs que peut prendre l'estimateur $\widetilde{m_h^*}$ est faite ci-dessous.

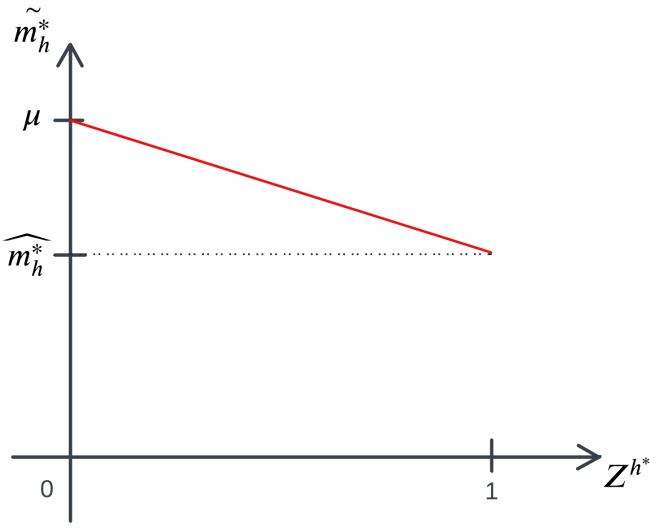


FIGURE 8.2 – Estimation du coefficient d'abattement dans le modèle de Bühlmann

Les questions principales qui seront posées dans la suite sont les suivantes :

- Comment calculer le paramètre Z^h (aussi appelé facteur de crédibilité) ?
- Quelles sont les propriétés de ce facteur ?
- Comment estimer les paramètres μ et σ^2 ?
- Quels critères utiliser pour la sélection de l'ensemble de portefeuilles $1, 2, ..., r$?

Une sous-partie sera dédiée à chacune de ces questions.

8.2 Hypothèses

Pour la suite les hypothèses suivantes sont posées [3] :

- r est le nombre de portefeuilles à disposition ;
- n_h est le nombre de prêts total du portefeuille h ;
- D_{hi} l'indicatrice de décès du contrat/prêt i du portefeuille h, d_{hi} est sa réalisation ;
- f_{hi} est la proportion d'année où le contrat/prêt i du portefeuille h est observé [4] ;
- I_p^h l'ensemble de tous les individus du portefeuille h ayant exactement p prêts distincts ;
- Pour simplifier les notations, q_{hi}^{ref} est la probabilité de décès dans l'année du contrat i du portefeuille h avec la table de référence q_x^{ref} ;
- la mortalité sur l'ensemble des portefeuilles est uniforme au cours d'une année ;
  i.e

$$\forall x \; \forall t \in [0,1] \; {}_tq_x = t * q_x$$

- A_h est le nombre de décès observé sur le portefeuille h, i.e

$$A_h = \sum_{i=1}^{n_h} D_{hi}$$

- E_h est l'attendu du portefeuille h selon la table de référence, i.e

$$E_h = \sum_{i=1}^{n_h} f_{hi} q_{hi}^{ref}$$

- $\widehat{m_h}$ est l'estimateur du ratio O/A, ou le SMR, du portefeuille, i.e

$$\widehat{m_h} = \frac{A_h}{E_h}$$

Il est supposé que la mortalité de l'ensemble des portefeuilles à disposition peut se décrire sous la forme suivante :

$$q_x^h = m_h * q_x^{ref}$$

avec m_h une variable aléatoire de caractéristiques suivantes : $\mathbb{E}[m_h] = \mu$ et $\mathbb{V}[m_h] = \sigma^2 \; \forall h \in [1, r]$.

Finalement, il est supposé que la loi, conditionnellement à m_h, des indicatrices de décès D_{hi} est une loi de $Ber(m_h f_{hi} q_i^{ref})$. Le paramètre μ peut également être interprété comme le taux d'abattement moyen entre l'ensemble des portefeuilles et la table de référence.

3. Ces notations sont reprises directement de l'article de [Klugman *et al.*, 2009], elles peuvent néanmoins représenter des objets déjà rencontrés dans cette étude mais nommés différemment.

4. Par simplification du modèle, il est supposé que les individus restent au maximum un an en portefeuille.

8.3 Estimation du facteur de crédibilité

Dans la suite, la mortalité du portefeuille h est cherchée à être estimée. Le volume du portefeuille h est supposé faible (sinon des techniques classiques d'estimation de la mortalité présentées dans le chapitre 5 seraient appliquées).
Par hypothèse du modèle, la mortalité du portefeuille h peut s'exprimer sous la forme suivante pour tout âge :

$$q_x^h = m_h q_x^{ref}$$

La méthode bayésienne empirique de Bühlmann propose un estimateur de la forme suivante :

$$\widetilde{m_h} = Z^h \widehat{m_h} + W^h$$

où

- $\widehat{m_h}$ est l'estimateur du ratio O/A du portefeuille h ;
- les quantités Z^h et W_h sont des grandeurs non aléatoires à estimer.

Dans le modèle, pour l'estimation de ces deux paramètres, la distance quadratique entre les variables aléatoires m_h et $\widetilde{m_h}$ est minimisée, i.e

$$(Z^h, W^h) = \underset{(Z^h, W^h)}{argmin}\ \mathbb{E}[(m_h - Z^h \widehat{m_h} - W^h)^2]$$

Afin d'alléger la lecture, les détails de la démonstration adaptée à la maille prêt (en prenant en compte la dépendance entre lignes de la même base) sont présentés en Annexe E.2. Cependant il est à noter que l'estimateur de Z^h construit et utilisé dans cette étude est différent de celui exprimé dans l'article de [Klugman *et al.*, 2009], car la contrainte de dépendance au sein des données a bien été prise en compte dans sa construction.

Après calculs, la forme finale pour l'expression du facteur de crédibilité de Bühlmann en présence de dépendance dans les données est la suivante :

$$Z^h = \frac{\mathbb{E}[m_h \widehat{m_h}] - \mu^2}{\mathbb{E}[\widehat{m_h}^2] - \mu^2} \tag{8.1}$$

$$= \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h\right)} \tag{8.2}$$

avec :

- ${}^pC_h = p^2 \sum_{z \in I_p^h} (f_{hz} q_z^{ref})^2$;
- ${}^pE_h = \sum_{z \in I_p^h} f_{hz} q_z^{ref}$.

8.4 Propriétés du facteur de crédibilité

Les propriétés attendues du facteur de crédibilité Z^h sont les suivantes :

- Z^h doit converger vers 1 quand l'exposition du portefeuille h augmente. En effet, il convient d'accorder plus de poids à l'expérience des données du portefeuille h quand le volume de ces dernières augmente ;
- Z^h doit être une fonction croissante de σ^2, la variance de m_h. En effet, lorsque la distance à la moyenne des m_h augmente, la distance théorique entre μ et le m_h réalisé augmente également. Il est donc préférable de ne pas accorder un trop grand poids à la moyenne des taux d'abattement μ dans la combinaison linéaire $\widetilde{m_h} = Z^h \widehat{m_h} + (1 - Z^h)\mu$.

L'étude de ces deux propriétés est maintenant réalisée.

Variation du facteur de crédibilité en fonction de l'exposition

Pour rappel :

$$Z^h = \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h\right)}$$

avec

- ${}^pE_h = \sum_{i \in I_p^h} p f_{hi} q_{hi}^{ref}$
- $E_h = \sum_{i=1}^{n_h} f_{hi} q_i^{ref} = \sum_{p=1}^{p_{max}} {}^pE_h$
- ${}^pC_h = p^2 \sum_{z \in I_p^h} (f_{hz} q_z^{ref})^2$

Ici, faire varier l'exposition revient à prendre la limite de Z^h quand n_h tend vers l'infini. Cette limite peut se décomposer en deux limites :

$$\lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2}$$

et

$$\lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p\ {}^pE_h}{E_h^2}$$

Si ces deux limites sont nulles, la propriété attendue : $\lim\limits_{n_h \to +\infty} Z^h = 1$ aura bien été démontrée.
En effet,

$$\begin{aligned}
\lim_{n_h \to +\infty} Z^h &= \lim_{n_h \to +\infty} \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h\right)} \\
&= \frac{\sigma^2}{\sigma^2 + \left(-(\mu^2+\sigma^2) \lim\limits_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2} + \mu \lim\limits_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p\ {}^pE_h}{E_h^2}\right)}
\end{aligned}$$

Avant de calculer ces limites, il convient de remarquer que :

$$\exists a \in \mathbb{R}^+ \ a \leq f_{hi}, q_i^{ref} \leq 1 \ \forall i \in [\![1; n_h]\!]$$

Il est également à noter que :

$$Card(I_p^h) \leq n_h \ \forall p \in [\![1; p_{max}]\!] \ \forall h \in [\![1; r]\!]$$

En effet, par construction des ensembles I_p^h :

$$\sum_{p=1}^{p_{max}} p * Card(I_p^h) = n_h$$

Les calculs des deux limites sont maintenant présentés :

$$\begin{aligned}
\lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2} &= \lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p^2 \sum_{z \in I_p^h} (f_{hz} q_z^{ref})^2}{(\sum_{i=1}^{n_h} f_{hi} q_i^{ref})^2} \\
&\leq \lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p^2 \sum_{z \in I_p^h} 1}{(\sum_{i=1}^{n_h} a^2)^2} \\
&= \lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p^2 Card(I_p^h)}{n_h^2 a^4} \\
&\leq \frac{\sum_{p=1}^{p_{max}} p^2}{a^4} \lim_{n_h \to +\infty} \frac{n_h}{n_h^2} \\
&= 0
\end{aligned}$$

Comme la quantité $\frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2}$ est positive par construction, la limite $\lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2}$ est égale à 0 par le théorème des gendarmes.
Le calcul de la seconde limite est maintenant présenté :

$$\begin{aligned}
\lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p \ {}^pE_h}{E_h^2} &= \lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p \ \sum_{i \in I_p^h} p f_{hi} q_{hi}^{ref}}{(\sum_{i=1}^{n_h} f_{hi} q_i^{ref})^2} \\
&\leq \lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p^2 \ \sum_{i \in I_p^h} 1}{(\sum_{i=1}^{n_h} a)^2} \\
&= \lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p^2 \ Card(I_p^h)}{n_h^2 a^2} \\
&\leq \frac{\sum_{p=1}^{p_{max}} p^2}{a^2} \lim_{n_h \to +\infty} \frac{n_h}{n_h^2} \\
&= 0
\end{aligned}$$

Comme la quantité $\frac{\sum_{p=1}^{p_{max}} p \ {}^pE_h}{E_h^2}$ est positive par construction, la limite $\lim_{n_h \to +\infty} \frac{\sum_{p=1}^{p_{max}} p \ {}^pE_h}{E_h^2}$ est égale à 0 par le théorème des gendarmes.
Il est donc possible de conclure sur la première propriété attendue du facteur de crédibilité, à savoir :

$$\lim_{n_h \to +\infty} Z^h = 1 \tag{8.3}$$

Variation du facteur de crédibilité en fonction de σ^2

Pour rappel :

$$Z^h = \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2 + \sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p \ {}^pE_h\right)}$$

Soit la fonction

$$Z(x) = \frac{x}{x - \alpha x + \beta}$$

avec

- $\alpha = \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2}$
- $\beta = -\mu^2 \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2} + \mu \frac{\sum_{p=1}^{p_{max}} p \ {}^pE_h}{E_h^2}$

L'égalité $Z(\sigma^2) = Z^h$ est satisfaite.
Les variations de la fonction Z sont maintenant étudiées. Pour cela, la dérivée de la fonction est calculée :

$$\begin{aligned} Z'(x) &= \frac{x - \alpha x + \beta - x(1-\alpha)}{(x - \alpha x + \beta)^2} \\ &= \frac{\beta}{(x - \alpha x + \beta)^2} \end{aligned}$$

Le signe de la dérivée de Z est gouverné par le signe de β. La propriété attendue du facteur de crédibilité Z^h est sa croissance en σ^2, cela revient, en considérant $\mu > 0$, à :

$$\begin{aligned} Z'(x) > 0 \\ &\Leftrightarrow \beta > 0 \\ &\Leftrightarrow -\mu^2 \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h^2} + \mu \frac{\sum_{p=1}^{p_{max}} p \ {}^pE_h}{E_h^2} > 0 \\ &\Leftrightarrow \sum_{p=1}^{p_{max}} p \ {}^pE_h > \mu \sum_{p=1}^{p_{max}} {}^pC_h \\ &\Leftrightarrow \sum_{p=1}^{p_{max}} p^2 \sum_{i \in I_p^h} f_{hi} q_{hi}^{ref} > \mu \sum_{p=1}^{p_{max}} p^2 \sum_{z \in I_p^h} (f_{hz} q_z^{ref})^2 \end{aligned}$$

Or, le produit $f_{hz} q_z^{ref}$ est toujours inférieur ou égal à 1, donc

$$\sum_{i \in I_p^h} f_{hi} q_{hi}^{ref} \geq \sum_{z \in I_p^h} (f_{hz} q_z^{ref})^2 \ \forall p \in [\![1; p_{max}]\!] \ \forall h \in [\![1; r]\!]$$

Finalement, une condition suffisante, mais non nécessaire, pour que le facteur de crédibilité Z^h soit une fonction croissante de σ^2 est la suivante :

$$\mu \leq 1 \tag{8.4}$$

Autrement dit, la mortalité moyenne sur l'ensemble des r portefeuilles doit être plus faible que la mortalité de la table à abattre q_x^{ref}.

8.5 Estimation des paramètres μ et σ^2

Dans le modèle exposé, les deux premiers moments de la variable m_h, μ et σ^2, sont supposés connus. Dans la réalité ils ne le sont pas et il faut les estimer.

Estimation du paramètre μ

L'estimateur de μ est l'estimateur naturel : le ratio O/A ou SMR. Le ratio est calculé sur l'ensemble des portefeuilles agrégés. La forme de l'estimateur est donc la suivante :

$$\widehat{\mu} = \frac{\sum_{h=1}^{r} A_h}{\sum_{h=1}^{r} E_h}$$

Cet estimateur est bien sans biais, i.e $\mathbb{E}[\widehat{\mu}] = \mu$:

$$\begin{aligned}
\mathbb{E}[\widehat{\mu}] &= \mathbb{E}[\frac{\sum_{h=1}^{r} A_h}{\sum_{h=1}^{r} E_h}] \\
&= \frac{1}{\sum_{h=1}^{r} E_h} \mathbb{E}[\sum_{h=1}^{r} A_h] \\
&= \frac{1}{\sum_{h=1}^{r} E_h} \mathbb{E}[\sum_{h=1}^{r} \sum_{i=1}^{n_h} D_{hi}] \\
&= \frac{1}{\sum_{h=1}^{r} E_h} \sum_{h=1}^{r} \sum_{i=1}^{n_h} \mathbb{E}[D_{hi}] \\
&= \frac{1}{\sum_{h=1}^{r} E_h} \sum_{h=1}^{r} \sum_{i=1}^{n_h} \mathbb{E}[\mathbb{E}[D_{hi}|m_h]] \\
&= \frac{1}{\sum_{h=1}^{r} E_h} \sum_{h=1}^{r} \sum_{i=1}^{n_h} \mathbb{E}[f_{hi} m_h q_{hi}^{ref}] \\
&= \frac{\mu}{\sum_{h=1}^{r} E_h} \sum_{h=1}^{r} \sum_{i=1}^{n_h} f_{hi} q_{hi}^{ref} \\
&= \frac{\mu}{\sum_{h=1}^{r} E_h} \sum_{h=1}^{r} E_h \\
&= \mu
\end{aligned}$$

Estimation du paramètre σ^2

La forme de l'estimateur de σ^2 est moins directe. Sa construction nécessite de nouveaux développements, car dans le cas présent, l'estimateur naturel :

$$\widehat{\sigma^2} = \sum_{h=1}^{r} E_h (\widehat{m_h} - \widehat{\mu})^2$$

n'est pas sans biais.
Afin d'alléger le propos, la démonstration de l'expression de l'estimateur de σ^2 est détaillée en Annexe E.3.
Enfin, comme c'était le cas pour le facteur de crédibilité, il convient de rappeler que les développements faits dans cette étude sont différents de ceux de l'article de Klugman, car ils intègrent la contrainte de dépendance entre les lignes à la maille prêt. [5]

La démonstration présentée en Annexe conclut à la forme suivante de l'estimateur sans biais de σ^2 adapté à la maille prêt

5. Voir le chapitre 6 pour plus de détails.

$$\widetilde{\sigma^2} = \frac{\sum_{h=1}^{r} E_h(\widehat{m_h} - \widehat{\mu})^2 - \widehat{\mu}^2 \left(\frac{\sum_{h=1}^{r}\sum_{p=1}^{pmax\ p}C_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{pmax\ p}C_h}{E_h} \right) - \widehat{\mu} \left(\sum_{h=1}^{r} \frac{\sum_{p=1}^{pmax}\ p\ {}^pE_h}{E_h} - \frac{\sum_{h=1}^{r}\sum_{p=1}^{pmax}\ p\ {}^pE_h}{\sum_{h=1}^{r} E_h} \right)}{\sum_{h=1}^{r} E_h - \frac{\sum_{h=1}^{r} E_h^2}{\sum_{h=1}^{r} E_h} + \frac{\sum_{h=1}^{r}\sum_{p=1}^{pmax\ p}C_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{pmax\ p}C_h}{E_h}}$$

8.6 Choix des portefeuilles

Le choix des portefeuilles sur lesquels appliquer le modèle décrit ci-dessus est primordial. En effet, retenir des données non « adaptées » au problème risquerait de biaiser l'estimateur $\widetilde{m_h}$. Cette partie a pour premier objectif la définition de critères pour le choix de ces r portefeuilles. Dans un second temps, trois méthodes seront proposées afin de mettre en place ces critères et choisir l'ensemble de portefeuille optimal.

8.6.1 Critères retenus

Dans la suite :

- le nombre de portefeuilles total à disposition sera noté n_{max} ;
- Le portefeuille i sera noté P^i ;
- L'ensemble de portefeuilles « optimal » sera noté P^*,
  avec $P^* = (P^i)_{i\in\Theta}$ $\Theta \subset [\![1; n_{max}]\!]$

Il est important de noter que la notion de portefeuille optimal introduite ici n'a de sens que si elle établit selon un certain critère. Ce critère dépendra de la méthode utilisée et sera donc évoqué plus tard.

Concernant le choix des portefeuilles, certaines propriétés sont souhaitables pour l'ensemble optimal P^* :

Critère 1 : un volume de données total suffisant

Le volume de données total, tous portefeuilles confondus, doit être suffisamment important pour pouvoir limiter au maximum l'erreur dans l'estimation du paramètre μ.

Critère 2 : un volume de données « homogène » entre les portefeuilles

Il est souhaitable que le volume d'un portefeuille n'écrase pas celui des autres. Dans le cas contraire, l'effet des autres portefeuilles serait fortement limité dans l'estimation des paramètres du modèle et μ tendrait vers le taux d'abattement du portefeuille volumineux.

Critère 3 : un nombre de portefeuilles suffisant

Il est souhaitable de retenir un nombre de portefeuilles suffisant, car une mesure d'hétérogénéité entre portefeuilles est estimée, σ^2, et un nombre trop faible de portefeuilles choisis pourrait entrainer un biais dans cette estimation.

Critère 4 : une certaine homogénéité des portefeuilles

Dans le modèle, l'homogénéité des portefeuilles est équivalente à une faible variance σ^2. Cette propriété du portefeuille optimal est souhaitable dans la mesure où une faible variance implique potentiellement un faible écart entre la réalisation de m_h, la quantité à estimer, et la moyenne μ. De cette façon, la combinaison linéaire $\widetilde{m_h} = Z^h\widehat{m_h} + (1 - Z^h)\mu$ apparait plus fiable.

8.6.2 Méthode 1 : Tous portefeuilles confondus

Une première méthode envisageable sera de prendre $P^* = (P^i)_{i\in[\![1;n_{max}]\!]}$. Soit l'ensemble des portefeuilles. En choisissant un tel ensemble de portefeuilles. Les critères 1, 2 et 3 seraient satisfaits.
En revanche, comme décrit dans le chapitre 3 :

- La mortalité sur certains portefeuilles est évidemment très différente que sur d'autres (opposition prêts à la consommation et prêts immobiliers par exemple).

Cette méthode semble *a priori* ne pas être la plus pertinente pour l'application du modèle bayésien empirique de Bühlmann.

8.6.3 Méthode 2 : In & Out

Une deuxième idée serait de rechercher la variance σ^2 minimale pour un nombre de portefeuilles donné. Soit n le nombre de portefeuilles à disposition et k le nombre de portefeuilles souhaités pour appliquer le modèle de crédibilité de Bühlmann. Il est possible de tester toutes les combinaisons de k portefeuilles parmi les n possibles, et de retenir la combinaison qui a amené à la variance σ^2 la plus faible. Cette méthode présente deux inconvénients :

- Elle peut être gourmande en temps si le nombre de portefeuilles n est grand ;
- Pour son application, il est nécessaire de poser le nombre de portefeuilles souhaité k. Dans la pratique, il n'y a aucun moyen de connaitre par avance le nombre k de portefeuilles optimal à sélectionner.

Afin de répondre à ces deux limites, une troisième méthode, basée sur une ACP sur les SMR des portefeuilles est maintenant proposée.

8.6.4 Méthode 3 : ACP sur SMR par tranches d'âges

L'idée est de représenter sur un plan l'ensemble des portefeuilles à disposition. Les proximités apparentes sur le plan seraient des indicateurs de proximité entre les risques des portefeuilles. L'utilisation d'analyse factorielle semble être une bonne piste pour atteindre cet objectif. L'étude du risque d'un portefeuille passe par ses données d'exposition et de décès, le SMR [6] par tranches d'âges peut être utilisé comme variable pour l'ACP.
La démarche suivie est la suivante :

- Pour chaque portefeuille, définir des tranches d'âges avec un pas régulier ;
- Pour chaque tranche de chaque portefeuille, calculer un SMR, ou observé sur attendu, avec une table de référence ;

6. Se référer à la section 5.4 pour voir les détails de la méthode.

- Concaténer toutes ces informations dans une matrice où une ligne représente un portefeuille, une colonne une tranche d'âge et un élément de la matrice un ratio SMR ;
- Réaliser une analyse en composantes principales sur la matrice ;
- Tracer le plan factoriel des individus ;
- Graphiquement ou à l'aide d'algorithme de *clustering*, rechercher des *clusters* sur le plan ;
- S'assurer que la proportion d'inertie expliquée par les deux axes factoriels est acceptable.

Un cas exemple est maintenant présenté. Sur le graphique ci-dessous, la notation P^j représente le portefeuille fictif numéro j.

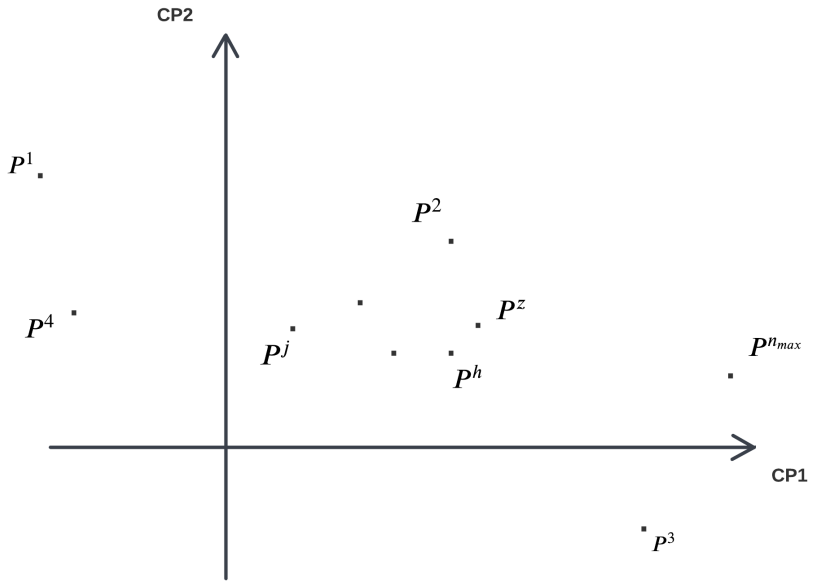


FIGURE 8.3 – Plan factoriel des portefeuilles

Sur le plan factoriel ci-dessus, un *cluster* est apparent. Au sein de ce dernier se trouve le portefeuille sur lequel une table de mortalité va être construite. De plus, les portefeuilles qui composent ce cluster semblent être homogènes au sens de leurs abattements sur la table de référence q_x^{ref}. Dans l'exemple, ce regroupement de portefeuilles est donc un bon candidat pour le portefeuille P^*.

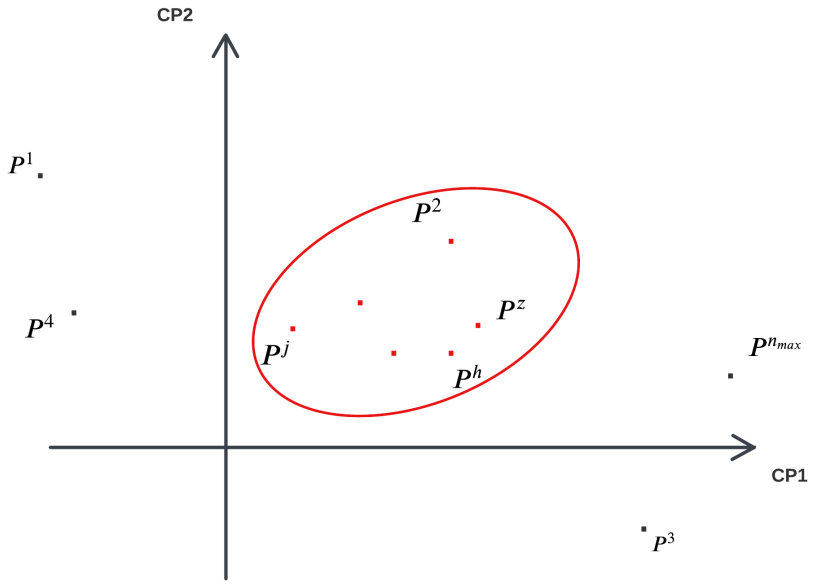


FIGURE 8.4 – Clustering sur plan factoriel des portefeuilles

Dans la méthode proposée, on recherchera donc par des méthodes de clustering (K-means, Mélanges gaussiens, *etc.*) des groupes de portefeuilles homogènes contenant le portefeuille dont on cherche à estimer la mortalité.

Remarque :

Cette méthode peut présenter des limites dans certains cas. Pour rappel, la théorie de la crédibilité est appliquée pour estimer la mortalité d'un portefeuille qui compte un faible volume de données. Le risque d'échantillonnage peut alors impliquer un écart entre la sinistralité observée et réelle sur ce portefeuille. Or, la méthode de sélection de portefeuilles par ACP présentée ci-dessus peut amener à un ensemble final de portefeuilles trop homogène. En effet, si le portefeuille au faible volume est proche d'autres portefeuilles sur le plan factoriel, leur risque n'est pas nécessairement le même en raison du risque d'échantillonnage. Le risque d'un choix d'un tel ensemble de portefeuilles trop homogène est que la variance du SMR du portefeuille à faible volume ne soit pas assez prise en compte.
Il est donc nécessaire de ne pas être trop restrictif à l'étape de clustering, de façon à se montrer prudent vis-à-vis du risque d'échantillonnage porté par le portefeuille à faible exposition.

8.7 Résultats du choix de portefeuilles

À présent, la méthode décrite ci-dessus peut être appliquée afin de déterminer l'ensemble de portefeuilles optimal à sélectionner pour appliquer la théorie de la crédibilité. L'ensemble des portefeuilles à disposition a été décrit dans le chapitre 3. Pour rappel, 5 bases de données sont disponibles : celles des produits emprunteur A,B,C,D et E.

Il convient également de rappeler que l'objectif de cette partie est la construction de la loi sans prise en compte de l'effet de sélection médicale : ${}^{\lambda}q_x$. Étant donné que le facteur λ a été estimé à 2 [7] et que la table va être construite sur les données du portefeuille A [8], la base à partir de laquelle les taux de mortalité vont être estimés est la base $B_2^{produit\ A}$. Pour rappel, la base $B_j^{produit\ k}$ représente la base de données du produit k tronquée sur les j premières années [9] .

Finalement, étant donné que la loi ${}^{\lambda}q_x$ servira à décrire la mortalité de la population éligible loi Lemoine, i.e les individus qui n'auront pas à passer la sélection médicale, il convient d'appliquer les « contraintes Lemoine » à la base $B_2^{produit\ A}$. En effet, la table ${}^{\lambda}q_x$ doit uniquement être construite à partir de contrats ayant une part assurée par tête inférieure à 200 000 € et dont la dernière échéance se situe avant les 60 ans de l'individu. La base de données du produit A dont les deux premières années de chaque contrat ont été tronquées et dont les lignes remplissent les deux conditions exprimées ci-dessus est notée ${}_{Lemoine}B_2^{produit\ A}$. C'est sur cette base de données que la table ${}^{\lambda}q_x$ sera construite. Il est à noter ici qu'appliquer le « filtre Lemoine » réduit encore davantage le volume de données de la base, ce qui conforte à nouveau l'utilisation de la théorie de la crédibilité.

À présent, la méthodologie proposée de choix de portefeuille pour l'application de la théorie de la crédibilité va être mise en oeuvre.

La méthode appliquée est celle de l'ACP. Les ratios SMR sont calculés par tranches d'âges de 5 ans sur l'ensemble des bases suivantes :

- Produit A : $B_0^{produit\ A}$, $B_1^{produit\ A}$ et ${}_{Lemoine}B_2^{produit\ A}$;
- Produit B : $B_0^{produit\ B}$, $B_1^{produit\ B}$ et $B_2^{produit\ B}$;
- Produit C : $B_0^{produit\ C}$, $B_1^{produit\ C}$ et $B_2^{produit\ C}$;
- Produit D : $B_0^{produit\ D}$, $B_1^{produit\ D}$ et $B_2^{produit\ D}$;
- Produit E : $B_0^{produit\ E}$, $B_1^{produit\ E}$ et $B_2^{produit\ E}$;

Une ACP avec centrage est réalisée sur l'ensemble des SMR par tranche d'âge de ces 15 bases de données. Les résultats sont les suivants :

7. Se référer à la section 7.4.
8. Se référer à la section 4.2 pour la justification du portefeuille.
9. Se référer au schéma Troncature des données pour l'estimation de la mortalité au-delà de j années pour plus de détails sur le sujet.

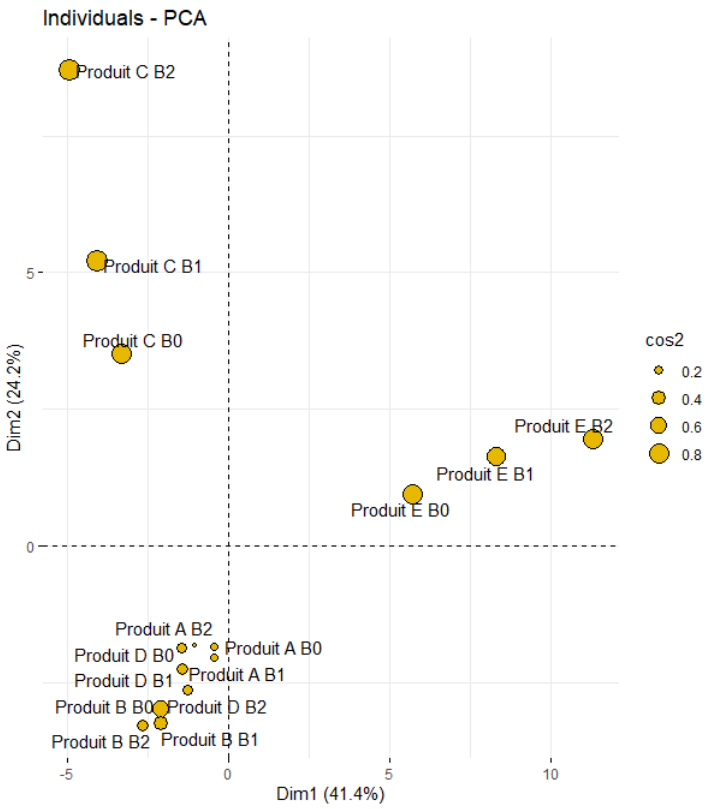


FIGURE 8.5 – Sélection de portefeuilles : résultat ACP sur données

La proportion d'inertie expliquée par les deux premiers axes factoriels est de $65,6\%$, ce qui est convenable pour l'étude menée. Graphiquement, les portefeuilles des produits A,B et D forment un cluster. L'utilisation d'un algorithme de clustering n'est ici pas utile.
Finalement, étant donné que la table ${}^{\lambda}q_x$ sera construite à partir de la base ${}_{Lemoine}B_2^{produit\ A}$, les bases finales retenues pour l'application du modèle de crédibilité de Bühlmann sont les suivantes :

- ${}_{Lemoine}B_2^{produit\ A}$;
- $B_2^{produit\ B}$;
- $B_2^{produit\ D}$;
- $B_0^{produit\ A}$.

Les bases tronquées sur 2 ans sont sélectionnées de façon à obtenir un risque homogène entre les portefeuilles. La base $B_0^{produit\ A}$ est retenue car elle a servi à construire la table de référence q_x^{ref} [10] qui sera abattue dans le modèle.
Finalement, le résultat est intuitif dans le sens où les produits sélectionnés correspondent à des produits d'assurance de prêt immobilier pour particuliers [11]. Là où les deux autres produits sont relatifs à des prêts professionnels et de consommation.

10. Cette étape sera détaillée dans la suite.
11. Se référer au chapitre 3 pour plus de détails sur les caractéristiques des portefeuilles.

8.8 Application de la crédibilité par tranches d'âges

Il est préférable de laisser une certaine flexibilité à un modèle fonctionnant sur un abattement d'une table de référence, de façon à permettre une déformation différente de la courbe de mortalité en fonction des âges. [12]

Le modèle de crédibilité de Bühlmann peut être appliqué par tranches d'âges. Après sélection du découpage des âges, les paramètres μ, σ^2, Z^{h^*} et le taux d'abattement final

$$\widetilde{m_h^*} = Z^{h^*}\widehat{m_h^*} + (1 - Z^{h^*})\mu$$

sont calculés indépendamment sur chaque plage d'âges.

Cependant, le choix des tranches d'âges sur lesquelles appliquer la théorie de la crédibilité est essentiel et ce dernier peut fortement influencer le résultat final. C'est pourquoi une méthode de sélection plus fine qu'un choix à dire d'expert doit être mise en œuvre. Le prochain chapitre est introduit en traitant des limites de l'application d'un coefficient d'abattement unique sur toute une table (aussi bien dans le cas du modèle SMR que dans le cas de la théorie de la crédibilité de Bühlmann). Après avoir établi ces limites, une méthode de sélection de tranches d'âges est proposée, nommée pour l'étude « la méthode des plateaux ».

12. la problématique sera détaillée dans la partie suivante.

Chapitre 9

Construction par abattements multiples d'une table de mortalité

Dans le chapitre précédent, il a été vu que la théorie de la crédibilité de [Klugman *et al.*, 2009] permettait la construction d'une table de mortalité par abattement étant donné un faible volume de données. La construction par abattement présente certaines contraintes et limites qu'il convient de détailler. De plus, il a été expliqué que la théorie de la crédibilité pourrait s'appliquer par tranches d'âges, ce qui soulève la question suivante : Comment sélectionner de manière mathématique les « bonnes » tranches d'âges sur lesquelles appliquer un modèle relationnel ? Ce chapitre s'attache à décrire une nouvelle méthode proposée de sélection de tranches d'âges.

9.1 Les limites de la méthode SMR

Comme cela a été décrit dans la chapitre 5, la méthode relationnelle du SMR repose sur l'hypothèse forte qu'il existe un coefficient α qui lie de façon proportionnelle la table de référence aux taux de mortalité réels du portefeuille étudié, i.e

$$q_x = \alpha * q_x^{ref} \quad \forall x$$

Cette hypothèse est forte dans la mesure où le coefficient α est supposé identique pour tous les âges. Dans la pratique, il n'est pas rare que la sinistralité des assurés sur certaines tranches d'âges se comporte de manière différente entre plusieurs portefeuilles. Cela devient un problème quand par exemple on souhaite projeter la sinistralité d'un portefeuille qui a évolué dans le temps et dont les proportions d'âges ont changé entre le moment de la construction de la table et le moment de la projection.

Il serait alors plus judicieux de pouvoir permettre au modèle d'être plus flexible et d'autoriser différentes déformations de la courbe $(q_x^{ref})_x$ en fonction de l'âge. Statistiquement, en imposant moins de contraintes au modèle, la qualité de l'estimateur final retenu est améliorée.

9.2 Algorithme de sélection des tranches d'âges

Dans le cas présent, imposer moins de contraintes au modèle soulève deux nouvelles questions :

- Quel est le bon nombre de tranches d'âges à sélectionner ?
- Quelles tranches d'âges sont les plus pertinentes ?

Pour répondre à ces deux questions, il convient d'analyser avec plus de précision la sinistralité observée et attendue sur le portefeuille (attendue avec la table de référence). Cette analyse est portée par âge et la fonction suivante est introduite :

$$SMR(x) = \frac{d_x}{q_x^{ref} * n_x}$$

Cette fonction renvoie, pour un âge x donné, le ratio observé sur attendu. Il est à noter que le ratio SMR est en réalité une moyenne pondérée de la famille $(SMR(x))_x$. :

$$SMR = \sum_x w_x * SMR(x)$$

Avec

$$w_x := \frac{q_x^{ref} * n_x}{\sum_x q_x^{ref} * n_x}$$

Concernant le choix des tranches d'âges, l'étude de la fonction $SMR(\cdot)$ peut s'avérer primordiale afin de déterminer les âges sur lesquels l'abattement semble être le même.
Pour illustrer ce dernier point, deux cas de figure sont maintenant présentés.

Cas 1 : un unique plateau

La fonction $SMR(\cdot)$ est tracée pour tout âge x :

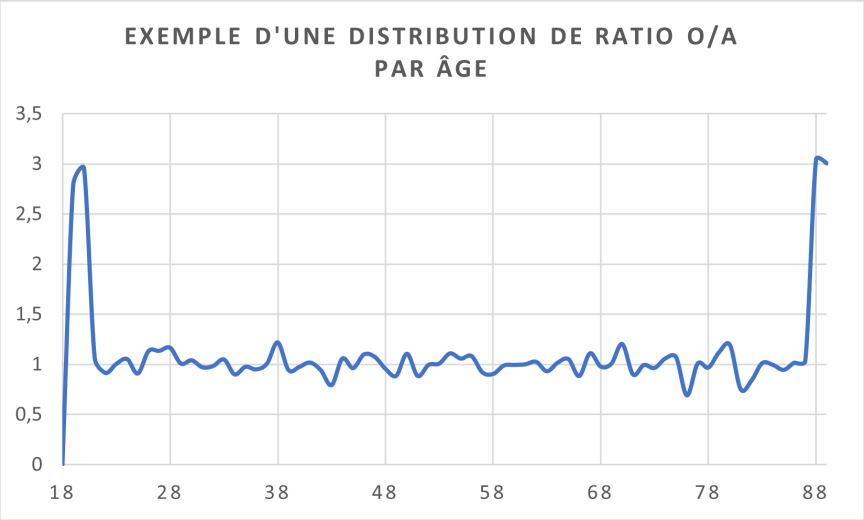


FIGURE 9.1 – Cas 1 exemple d'une distribution de ratio O/A par âge

Graphiquement, un plateau, associé à des fluctuations dues au faible volume de données, est observable. L'hypothèse du coefficient d'abattement constant α semble ne pas pouvoir être rejetée.

Cas 2 : Différents plateaux

La fonction $SMR(\cdot)$ est tracée pour tout âge x :

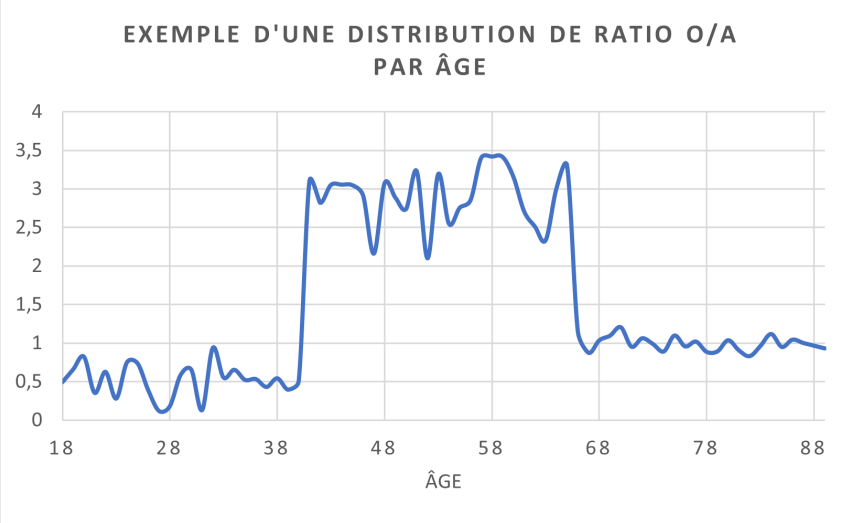


FIGURE 9.2 – Cas 2 exemple d'une distribution de ratio O/A par âge

Graphiquement, trois plateaux, associés à des fluctuations dues au faible volume de données, sont observables. L'hypothèse du coefficient d'abattement constant α semble trop forte. Il semblerait que la déformation de la courbe de référence soit différente en fonction des âges. Il serait donc préférable de considérer trois tranches d'âges pour réaliser plusieurs abattements.

Évidemment, la réalité des données ne présente presque jamais de cas aussi triviaux, mais l'identification de plateaux sur la fonction $SMR(\cdot)$ est une bonne piste pour la sélection de tranches.
Il est présenté dans la suite une proposition de méthode de sélection de tranches d'âges basée sur cette même idée.

9.2.1 Algorithme de sélection de tranches d'âges : la méthode des plateaux

Le principe est de reconnaitre, de façon mathématique, différents plateaux dans la représentation graphique d'une fonction. La méthode des plateaux peut ensuite être appliquée au cas de la sélection de tranches d'âges et à la fonction $SMR(\cdot)$.

Dans un premier temps, remarquons que rechercher des plateaux dans une fonction revient en réalité à un problème de minimisation des moindres carrés. La méthode suivie sera de minimiser la distance quadratique entre la fonction $SMR(\cdot)$ et des sous-ensembles de fonctions en escalier.

Interlude : les fonctions en escalier

Une fonction *en escalier*, ou *étagée*, est une fonction qui peut s'écrire sous la forme suivante :

$$h(x) = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}(x)$$

où

- $A_1, ...A_n$ est une suite finie d'ensemble d'intersection nulle, i.e : $\bigcap_k A_k = \emptyset$;
- $a_1, ..., a_n$ est une suite finie de valeurs dans $\mathbb{R}$.

Dans la suite, l'ensemble des fonctions à n étages sera noté $ESC(n)$.

En recherchant à minimiser la distance quadratique entre la fonction $SMR(\cdot)$ et l'ensemble des fonctions $ESC(n)$, le résultat est alors une fonction constante par palier, décrivant n tranches distinctes et dont la hauteur est la moyenne de la fonction $SMR(\cdot)$ sur chaque plage. Ceci décrit exactement l'idée de recherche des paliers exposée dans la partie précédente.

En effet, pour rappel, le problème des moindres carrés pour une constante revient à minimiser la *Loss* :

$$Loss(\alpha) = \sum_{x=1}^{n} (f(x) - \alpha)^2$$

La solution de ce problème est la moyenne des valeurs prises par la fonction f sur l'intervalle, i.e :

$$\widehat{\alpha} = \frac{1}{n} \sum_{x=1}^{n} f(x)$$

(*Voir l'Annexe F.1 pour le détail de ce calcul)*

Les détails de l'algorithme de sélection de tranches d'âges sont maintenant présentés :

Étape 1

Soit n le nombre d'âges sur lesquels la fonction SMR_x est non nulle. Pour tout i compris entre 1 et n, la quantité suivante est calculée :

$$\begin{aligned} \widehat{h^{(i)}} &= \underset{h \in ESC(i)}{argmin}\ Loss(h) \\ &= \underset{h \in ESC(i)}{argmin} \sum_{x} (SMR(x) - h(x))^2 \end{aligned}$$

La *Loss* associée au meilleur candidat de chaque palier est ensuite calculée :

$$Loss(\widehat{h^{(i)}}) = \underset{h \in ESC(i)}{min} \sum_{x} (SMR(x) - h(x))^2$$

Une fois cette étape validée, un « meilleur candidat » pour chaque palier a été déterminé.
Pour reprendre l'exemple précédent, les représentations de $\widehat{h^{(1)}}, \widehat{h^{(2)}}, \widehat{h^{(3)}}$ et $\widehat{h^{(4)}}$ sont faites :

Représentation de $\widehat{h^{(1)}}$

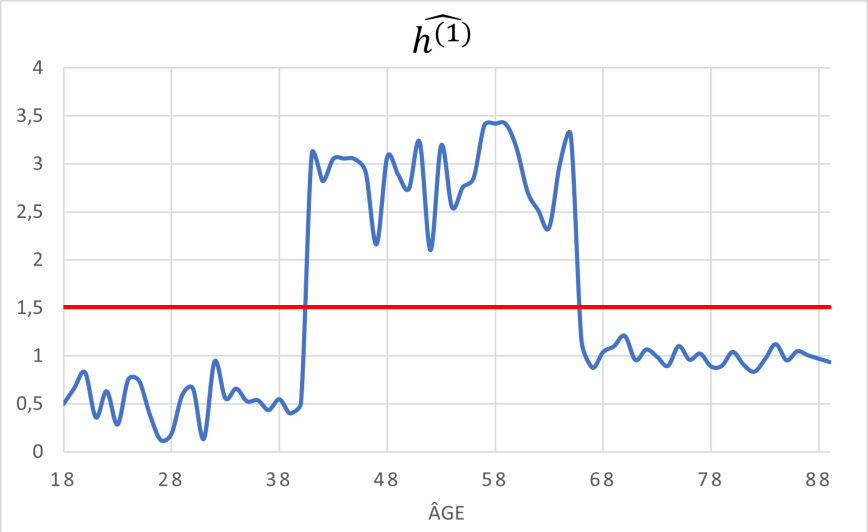


FIGURE 9.3 – Cas exemple : tranche optimale

Avec :

$$Loss(\widehat{h^{(1)}}) = 82,59$$

Représentation de $\widehat{h^{(2)}}$

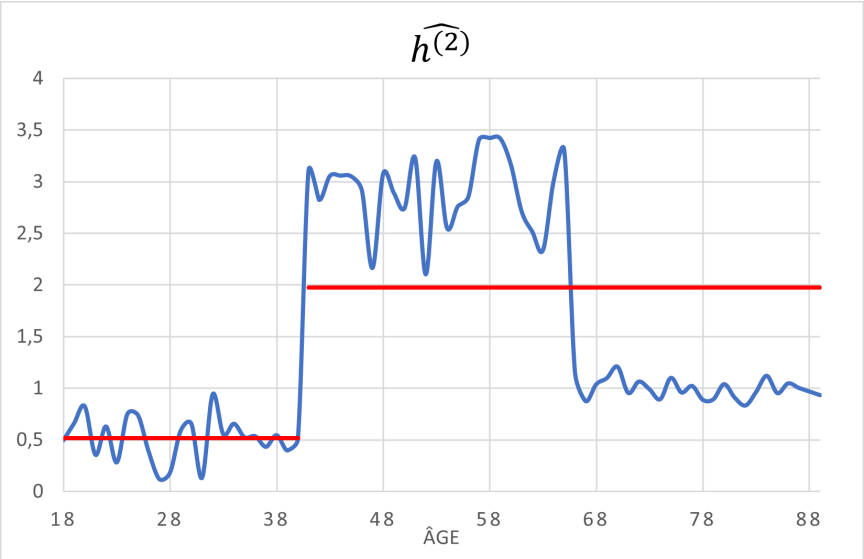


FIGURE 9.4 – Cas exemple : 2 tranches optimales

Avec :

$$Loss(\widehat{h^{(2)}}) = 49,4$$

Représentation de $\widehat{h^{(3)}}$

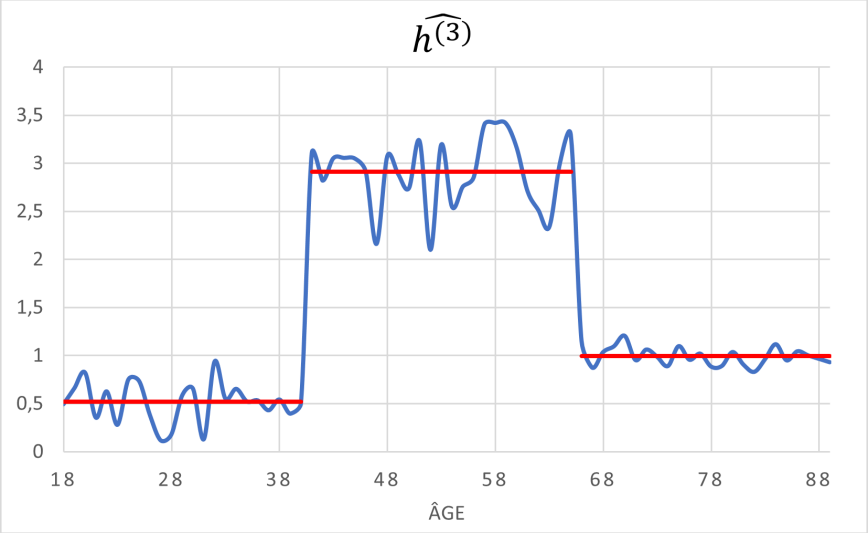


FIGURE 9.5 – Cas exemple : 3 tranches optimales

Avec :

$$Loss(\widehat{h^{(3)}}) = 9,36$$

Représentation de $\widehat{h^{(4)}}$

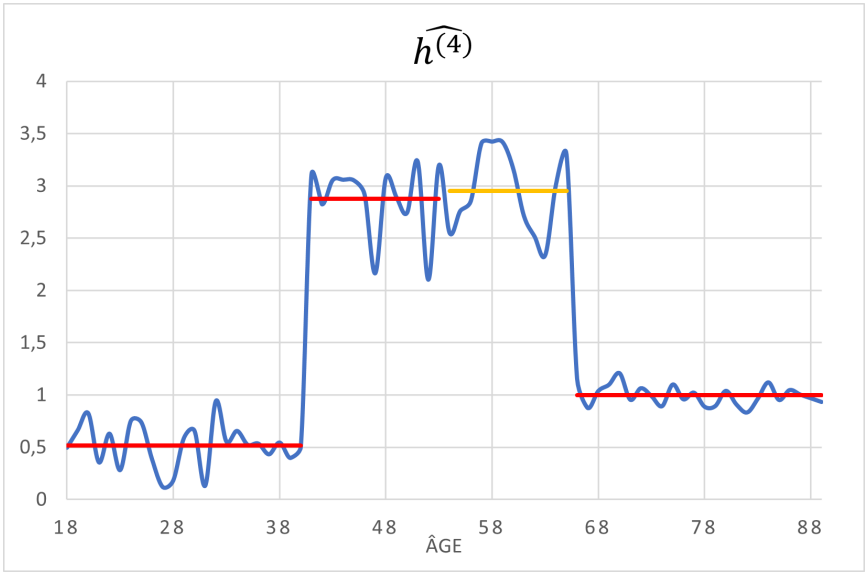


FIGURE 9.6 – Cas exemple : 4 tranches optimales

Avec :

$$Loss(\widehat{h^{(4)}}) = 6,94$$

Par construction, les *Loss* sont ordonnées de la manière suivante $Loss(\widehat{h^{(1)}}) \geq Loss(\widehat{h^{(2)}}) \geq Loss(\widehat{h^{(3)}}) \geq Loss(\widehat{h^{(4)}}) \geq .... \geq Loss(\widehat{h^{(n)}})$. En effet, cela découle de l'inclusion des ensembles $ESC(i)$:

$$ESC(1) \subset ESC(2)... \subset ESC(n)$$

Une fois le calcul de $(Loss(\widehat{h^{(i)}}))_{i\in[\![1;n]\!]}$ réalisé, le meilleur candidat si l'on souhaite obtenir k tranches distinctes est l'ensemble des tranches d'âges $A_1^k, A_2^k, ..., A_k^k$ avec :

$$\widehat{h^{(k)}} = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i^k}(x)$$

Cependant, afin de sélectionner le nombre de tranches optimal, il ne suffit pas de sélectionner la fonction $\widehat{h^{(i)}}$ qui a donné la *Loss* la plus faible. En effet, de par le caractère décroissant des $(Loss(\widehat{h^{(i)}}))_{i\in[\![1;n]\!]}$, la *Loss* la plus petite par construction est celle qui correspond au nombre de tranches égale au nombre d'âges n, i.

$$min((Loss(\widehat{h^{(i)}}))_{i\in[\![1;n]\!]}) = Loss(\widehat{h^{(n)}})) = 0$$

Il convient donc d'introduire un facteur de pénalisation, et ainsi une *Loss* pénalisée.

Étape 2

La *Loss* pénalisée suivante est introduite :

$$Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}}) + (i-1)\beta$$

où $\beta \in \mathbb{R}^+$.

De cette façon, le nombre de tranches est à présent pénalisé. Le facteur $i-1$ signifie que la tranche d'âges unique n'est pas pénalisée.

Dans cette deuxième étape, la famille $\left(Loss^p(\widehat{h^{(i)}})\right)_{i\in[\![1;n]\!]}$ est calculée. Le nombre de tranches optimal retenu est celui qui conduit à la *Loss* pénalisée la plus faible, i.e :

$$NombreTranchesOptimal = k^* = \underset{i\in[\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}})$$

Et ainsi le découpage idéal des âges est $A_1^{k^*}, A_2^{k^*}, ..., A_{k^*}^{k^*}$ avec :

$$\widehat{h^{(k^*)}} = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i^{k^*}}(x)$$

Remarque :

Ici, il a été décidé de pénaliser le nombre de tranches dans la *Loss* de façon linéaire (fonction $(i-1)$), car ce modèle est plus intuitif exprimé comme tel.
Cependant, il est possible d'exprimer la *Loss* sous la forme plus générale :

$$Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}}) + f(i)\beta$$

Où f est une fonction croissante du nombre de tranches.

Dans ce cas plus général, il est possible de pénaliser de moins en moins le nombre de tranches quand ce dernier augmente. De plus, les résultats peuvent varier selon le choix de la fonction f. Une fonction fortement croissante aura tendance à faire pencher l'algorithme vers le choix d'un faible nombre de tranches.

Dans la suite de cette étude, la pénalisation linéaire sera conservée.

9.2.2 Remarque sur le facteur β

Pour rappel, la *Loss* pénalisée introduite à la partie précédente est de la forme :

$$Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}}) + (i-1)\beta$$

Il est clair que le facteur β a du poids dans le problème d'optimisation. Les deux cas extrêmes le démontrent :

Cas 1 : $\beta = 0$

Dans le cas $\beta = 0$, alors $Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}})$ et ainsi

$$\underset{i\in[\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}}) = n$$

Ce cas trivial est à éviter, car il est évident que le découpage où chaque âge correspond à une « tranche d'âge » ne serait pas valide.

Cas 2 : $\beta \to +\infty$

Dans ce deuxième cas, la solution au problème d'optimisation serait la suivante :

$$\underset{i\in[\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}}) = 1$$

En effet, permettre plus d'une unique tranche pénaliserait trop fortement la *Loss*. Ce cas trivial est également à éviter.

Finalement, de la même manière que pour le facteur h dans le modèle de Whittacker-Henderson, il apparaît que le choix du facteur β doit être fait avec précaution.

Une méthode de calcul est maintenant proposée. Le facteur de pénalisation β qui sera retenu dans la suite est le suivant :

$$\beta = \underset{x\in[\![2;n]\!]}{max} \frac{(SMR(x) - SMR(x-1))^2}{2}$$

Le facteur β retenu est le saut le plus fort de la fonction $SMR(\cdot)$. Ce choix est fait de telle sorte à ce que l'algorithme décide de ne pas créer deux tranches supplémentaires au niveau d'un pic ponctuel de la fonction $SMR(\cdot)$.

9.2.3 L'optimisation dans la pratique

Pour rappel, l'étape 1 de l'algorithme consiste à calculer $(Loss(\widehat{h^{(i)}}))_{i\in[\![1;n]\!]}$. L'étape 2, quant à elle, consiste au calcul de la *Loss* pénalisée $\left(Loss^p(\widehat{h^{(i)}})\right)_{i\in[\![1;n]\!]}$.

Concernant cette première étape, l'optimisation sur chaque sous-ensemble $ESC(i)$ n'est pas directe. Deux approches sont maintenant présentées :

Approche 1 : Tester l'ensemble des fonctions

Dans cette méthode, il est proposé de calculer $Loss(h)\ \forall h \in ESC(i)\ \forall i \in [\![1;n]\!]$. Cependant, en termes de dénombrement, la taille de l'ensemble $ESC(i)$ est de l'ordre de $\binom{n}{i-1} = \frac{n!}{(i-1)!(n-i+1)!}$.

En termes de stockage, le logiciel R Studio ne supporte pas cette méthode pour l'optimisation sur les ensembles de fonctions excédents les 7 tranches, i.e $ESC(7)$.

Approche 2 : L'approche aléatoire

Une autre approche, plus efficace, est la recherche aléatoire. Cette méthode consiste au découpage aléatoire de l'ensemble des âges en k tranches si l'on cherche à calculer $h^{(k)}$. Un nombre M de découpages aléatoires est fixé et M ensembles de k de tranches d'âges sont tirés. La *Loss* est ensuite calculée sur chacun d'entre eux. Cette méthode fonctionne efficacement car il est possible d'imposer des contraintes au découpage aléatoire. Dans le cadre de cette étude, une taille minimale de chaque tranche est donnée en contrainte. Cela permet ainsi de ne pas tester un grand nombre de fonctions triviales.

9.2.4 Contraintes supplémentaires

La méthode présentée a l'avantage de permettre de rajouter des contraintes supplémentaires. Dans le cadre cette étude, une contrainte de proportion d'exposition minimale sur chaque tranche sera intégrée au problème. En effet, même en ne recherchant pas l'homogénéité parfaite de l'exposition entre les différentes tranches d'âges, il est nécessaire de forcer chaque tranche d'âges proposée par le modèle à dépasser un certain pourcentage de seuil d'exposition.

9.2.5 Résultats sur le cas exemple

Afin d'illustrer l'algorithme des plateaux, la méthode est appliquée au cas exemple présenté en début de partie. Pour rappel, l'allure de cette fonction exemple était la suivante :

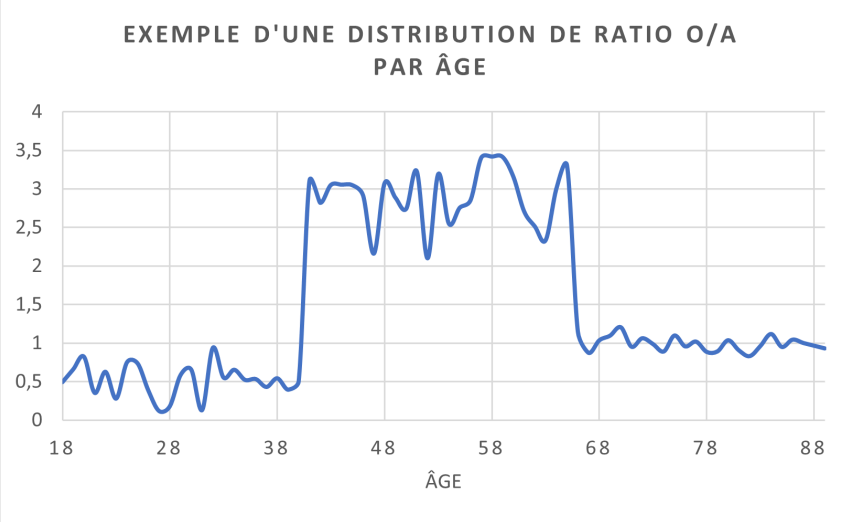


FIGURE 9.7 – Cas 2 exemple d'une distribution de ratio O/A par âge

L'optimisation est faite par découpage aléatoire et la famille $(Loss(\widehat{h^{(i)}}))_{i\in[\![1;n]\!]}$ est calculée. Suite à ça, le facteur de pénalisation est calculé par la formule suivante :

$$\beta = \max_{x\in[\![2;n]\!]} \frac{(SMR(x) - SMR(x-1))^2}{2} = 3,36$$

Enfin, la *Loss* pénalisée $\left(Loss^p(\widehat{h^{(i)}})\right)_{i\in[\![1;n]\!]}$ est calculée. Les représentations graphiques de ces deux *Loss* sont faites en fonction du nombre de tranches :

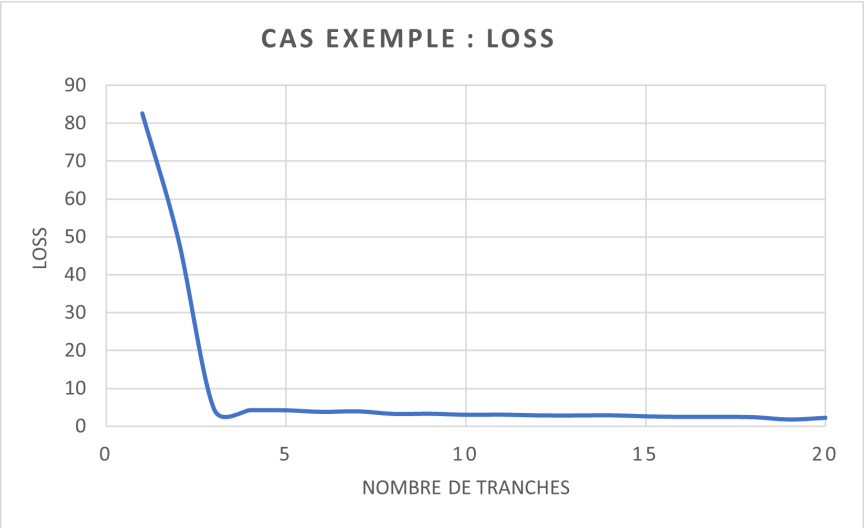


FIGURE 9.8 – Cas exemple : fonction *Loss*

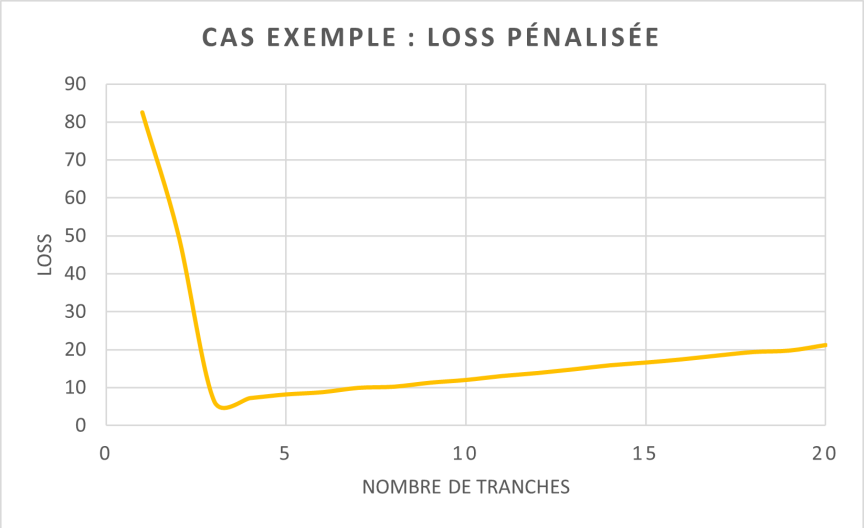


FIGURE 9.9 – Cas exemple : fonction *Loss* pénalisée

Comme attendu, la *Loss* pénalisée minimale est celle associée aux 3 tranches.

9.2.6 Une alternative : la méthode des plateaux pondérés

Dans les parties précédentes, il a été expliqué que la sélection de tranches d'âges pour la construction de table revenait à un problème d'optimisation des moindres carrés. Cette première méthode peut être approfondie en décrivant à présent le problème comme une optimisation des moindres carrés pondérés (MCP). La fonction *Loss* des MCP fait intervenir des poids w_x, et ainsi une fonction *Loss* généralisée :

$$Loss(\alpha) = \sum_x w_x \left(f(x) - \alpha\right)^2$$

La solution de ce problème d'optimisation est la moyenne pondérée par les poids $(w_x)_x$ des valeurs prises par la fonction $f(x)$:

$$\widehat{\alpha} = \frac{\sum_x w_x f(x)}{\sum_x w_x}$$

Voir la démonstration en deuxième partie de l'Annexe F.1

Dans le cadre de la méthode des plateaux, en sélectionnant les poids w_x égaux aux sinistralités attendues aux âges x, i.e

$$w_x = q_x^{ref} * n_x \quad \forall x$$

Alors, sur chaque sous-intervalle $[x_i, x_{i+1}]$, la solution au problème d'optimisation est le SMR sur cette tranche d'âges :

$$\begin{aligned}
\widehat{\alpha} &= \underset{\alpha}{argmin} \sum_{x\in[x_i,x_{i+1}]} \left(q_x^{ref} * n_x\right)\left(SMR(x) - \alpha\right)^2 \\
&= \frac{\sum_{x\in[x_i,x_{i+1}]} q_x^{ref} * n_x * SMR(x)}{\sum_{x\in[x_i,x_{i+1}]} q_x^{ref} * n_x} \\
&= \frac{\sum_{x\in[x_i,x_{i+1}]} q_x^{ref} * n_x * \frac{d_x}{q_x^{ref} * n_x}}{\sum_{x\in[x_i,x_{i+1}]} q_x^{ref} * n_x} \\
&= \frac{\sum_{x\in[x_i,x_{i+1}]} d_x}{\sum_{x\in[x_i,x_{i+1}]} q_x^{ref} * n_x} \\
&= SMR_{x\in[x_i,x_{i+1}]}
\end{aligned}$$

Finalement, cette méthode permet de limiter les fluctuations des plateaux retenus dues au manque de volume. Cet effet est particulièrement visible aux bords, dans le cas des âges extrêmes. En effet, un faible volume sur un âge x entraîne une forte volatilité de la quantité $SMR(x)$ et il devient risqué d'accorder autant de poids à ce coefficient dans le problème d'optimisation.

9.2.7 Alternatives à la *Loss* pénalisée

L'algorithme des plateaux et celui des plateaux pondérés reposent donc sur un facteur de pénalisation, β, pour l'optimisation du nombre de tranches dans un problème de construction de table. Cependant, comme exprimé plus haut, ces modèles sont relativement sensibles à la valeur de β et il se peut que les résultats produits par le modèle ne soient pas satisfaisants.
Dans ce cas, l'approche pénalisée peut être remplacée par deux autres méthodes :

Méthode 1 : la méthode du coude

L'étape 1 de l'algorithme reste la même, la *Loss* est calculée pour l'ensemble des nombres de tranches d'âges possibles (de 1 à n). La représentation graphique de ces *Loss* est ensuite réalisée. De manière similaire à ce qui est appliqué pour la détermination du nombre optimal d'axes factoriels dans une analyse en composantes principales (ACP), la règle du coude peut s'avérer utile dans le cas présent. La recherche visuelle d'un « coude » sur le graphique indique alors le nombre de tranches pour lequel la *Loss* a eu la plus nette amélioration. Pour en donner une illustration, cette règle peut s'appliquer au cas étudié depuis le début de cette partie. Pour rappel, la *Loss* (non pénalisée) est la suivante :

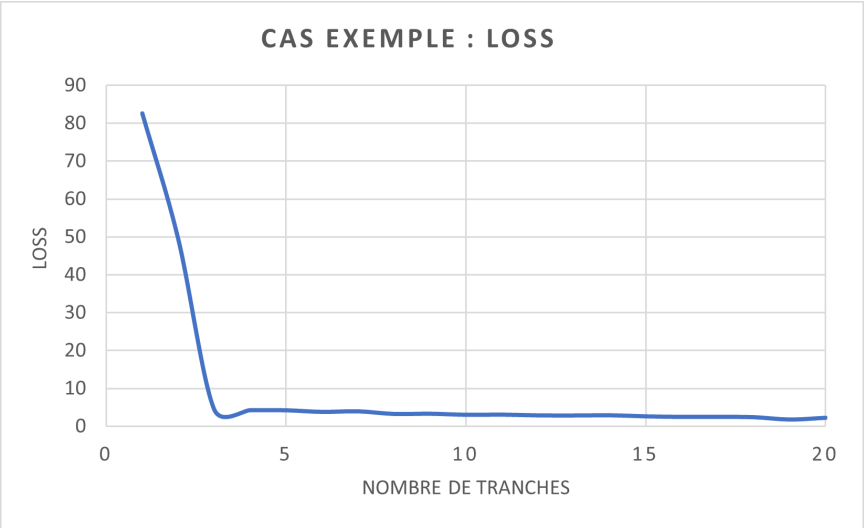


FIGURE 9.10 – Cas exemple : fonction loss

Dans ce cas, sans avoir eu à devoir calculer de facteur de pénalisation, le critère du coude indique clairement un nombre optimal de tranches de 3, ce qui est le résultat attendu.

Méthode 2 : Les dérivées successives

Pour une ouverture, il serait imaginable d'évaluer la « dérivée » de la fonction *Loss* et d'en déduire le nombre de tranches optimal à partir du point où la valeur absolue de cette dérivée serait la plus forte. En effet, la fonction *Loss* étant décroissante, sa dérivée est toujours négative et le point où la pente est la plus forte, symbolisant la meilleure amélioration de la *Loss*, est celui où la dérivée en valeur absolue est la plus grande.

Sur le graphique de la *Loss* en exemple, la pente est la plus forte entre 2 et 3, ce qui signifierait que le nombre optimal de tranches serait de 3, ce qui est le résultat attendu.

Néanmoins, il est nécessaire de rappeler que la fonction n'est pas toujours dérivable sur les entiers de par sa construction.

9.3 Résultats des algorithmes des plateaux

La méthodologie de choix de tranches d'âges pour la construction d'une table par abattement est maintenant mise en œuvre. Pour rappel, le facteur λ a été estimé à 2 ans [1] et les méthodes sont appliquées sur les données du portefeuille étudié : ${}_{Lemoine}B_2^{produit\ A}$ [2].

À présent, afin de construire la loi ${}^{\lambda}q_x$ applicable à une population éligible loi Lemoine, il faut commencer par tronquer les λ (2 dans le cas présent) premières années de chaque contrat de la base de données du portefeuille A, puis de ne conserver que les contrats remplissant les conditions d'éligibilité de la loi Lemoine [3]. Pour plus de détails sur la méthodologie de troncature et de calcul des taux, se référer au chapitre 7. Une fois cette partie de l'exposition et celle des sinistres de la base tronquées, la table de mortalité « sans sélection médicale » peut être construite par abattement de la table de référence en

1. Cela signifie que l'effet de sélection médicale est significativement observable sur les deux premières années de chaque contrat sur le portefeuille A.
2. Se référer aux chapitres 3 et 8.6 pour justification.
3. Se référer au chapitre 2 pour plus de détails.

utilisant les méthodes de crédibilité présentées dans le chapitre 8. La première étape avant de calculer les coefficients d'abattements par la crédibilité est donc la détermination des tranches d'âges.

Pour rappel, la méthodologie présentée dans le chapitre 9.2 se base sur l'étude du ratio observé sur attendu par âge :

$$SMR(x) = \frac{d_x}{q_x^{ref} * n_x}$$

La représentation graphique de cette fonction, calculée sur les données du portefeuille ${}_{Lemoine}B_2^{produit\ A}$ et en tous points non nuls, est la suivante :

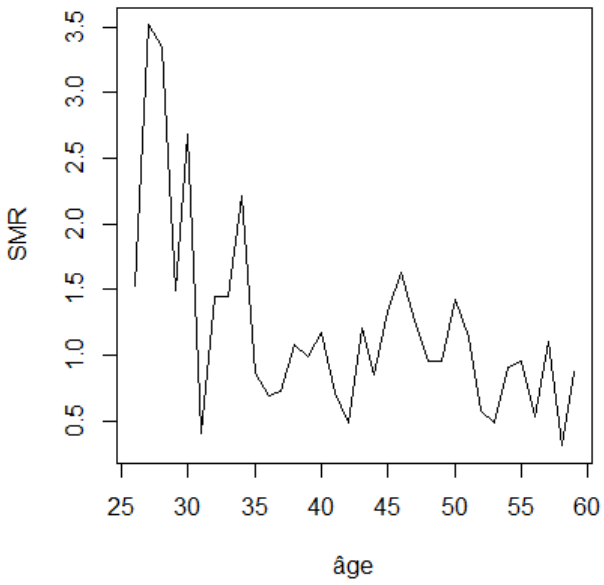


FIGURE 9.11 – SMR par âge sur la base ${}_{Lemoine}B_2^{produit\ A}$

La méthode des plateaux ainsi que la méthode des plateaux pondérés sont à présent appliquées.

Résultats : méthode des plateaux

Les fonctions suivantes sont tracées pour tout i.

$$\widehat{Loss(h^{(i)})} = \min_{h \in ESC(i)} \sum_x (SMR(x) - h(x))^2$$

et

$$\widehat{Loss^p(h^{(i)})} = \widehat{Loss(h^{(i)})} + (i-1)\beta$$

où

$$\beta = \max_{x \in [\![2;n]\!]} \frac{(SMR(x) - SMR(x-1))^2}{2}$$

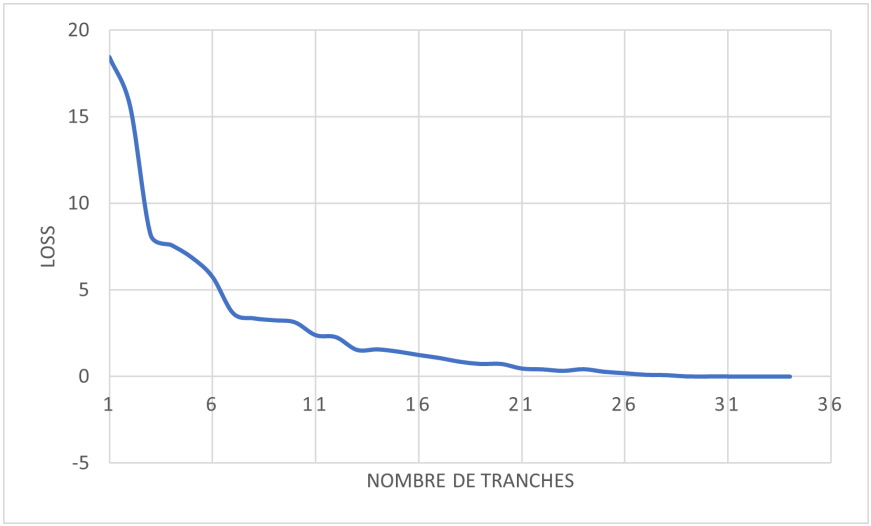


FIGURE 9.12 – Fonction *Loss*, modèle des plateaux

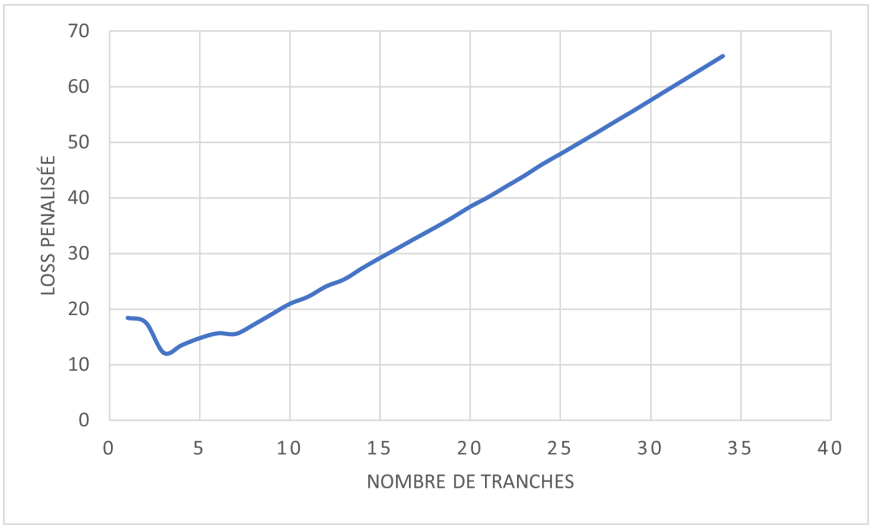


FIGURE 9.13 – Fonction *Loss* pénalisée, modèle des plateaux

Afin de discerner graphiquement le minimum de la *Loss* pénalisée, la même fonction est tracée uniquement pour les abscisses correspondant aux tranches allant de 1 à 10 :

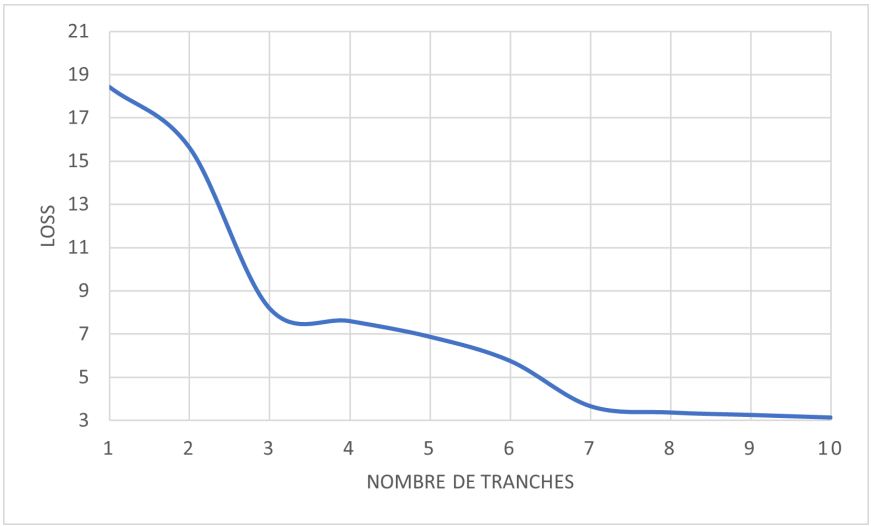


FIGURE 9.14 – Zoom fonction *Loss*, modèle des plateaux

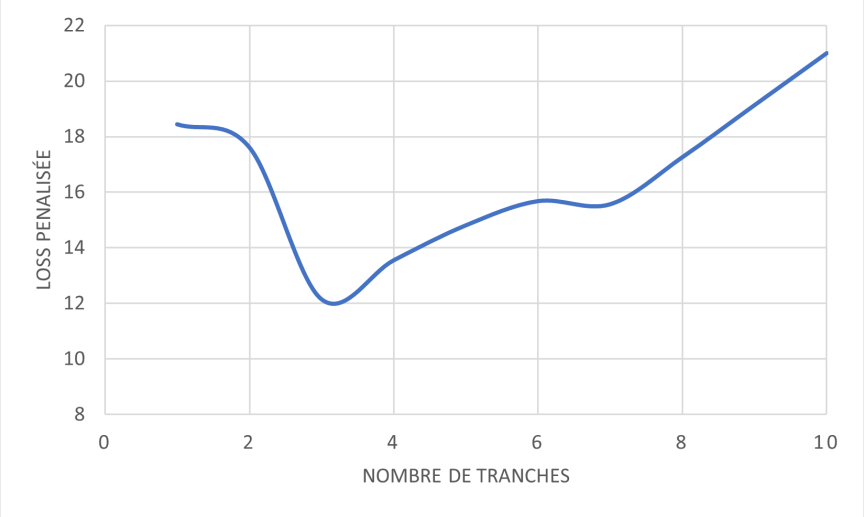


FIGURE 9.15 – Zoom fonction *Loss* pénalisée, modèle des plateaux

En appliquant la règle de décision établie dans la méthodologie, le nombre de tranches optimal est celui pour lequel la *Loss* pénalisée est la plus faible, i.e :

$$NombreTranchesOptimal = k^* = \underset{i \in [\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}})$$

Dans le cadre de cette étude, le nombre optimal de tranches est donc de trois. Il correspond aux tranches d'âges : 18-38 ans et 39-49 ans et 50-89 ans.
Le graphique associé du SMR par âge superposé à la fonction en escalier optimale de l'algorithme des plateaux est le suivant :

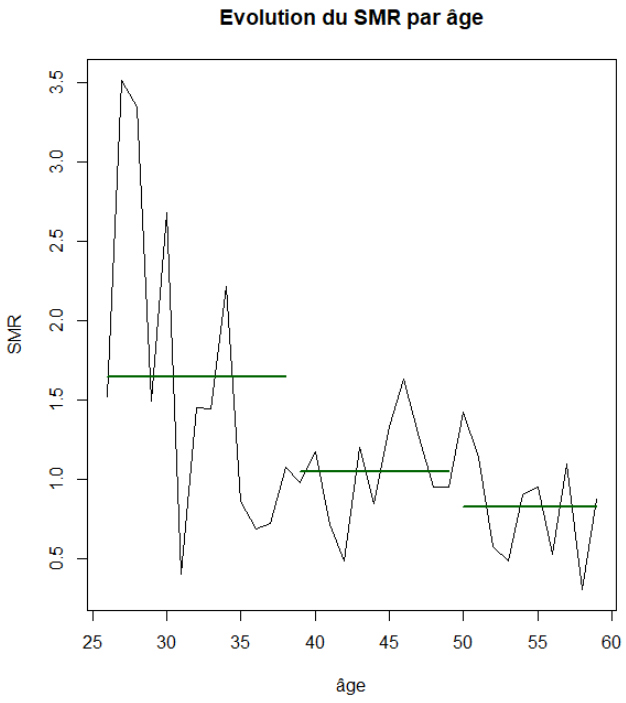


FIGURE 9.16 – Solution optimale, modèle des plateaux

Résultats : méthode des plateaux pondérés

Les fonctions suivantes sont tracées pour tout i.

$$Loss(\widehat{h^{(i)}}) = \min_{h \in ESC(i)} \sum_x w_x (SMR(x) - h(x))^2$$

où

$$w_x = q_x^{ref} * n_x$$

et

$$Loss^p(\widehat{h^{(i)}}) = Loss(\widehat{h^{(i)}}) + (i-1)\beta$$

où

$$\beta = \max_{x \in [\![2;n]\!]} w_x \frac{(SMR(x) - SMR(x-1))^2}{2}$$

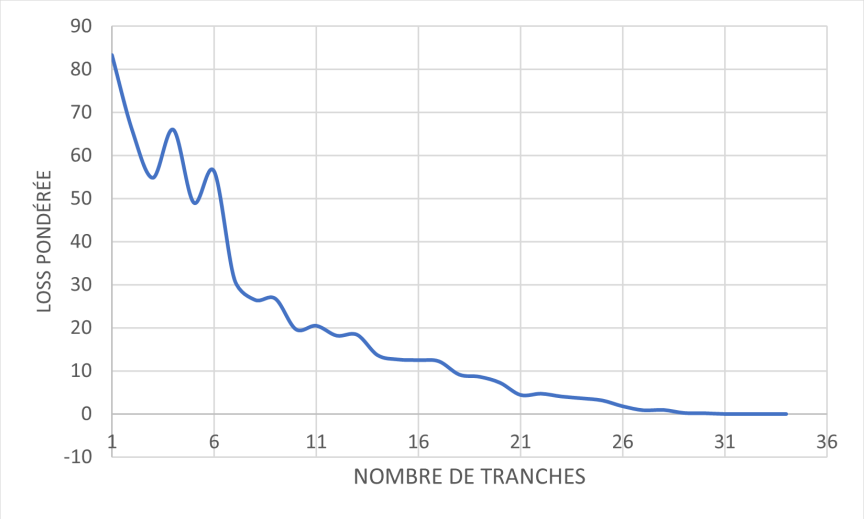


FIGURE 9.17 – Fonction *Loss*, modèle des plateaux pondérés

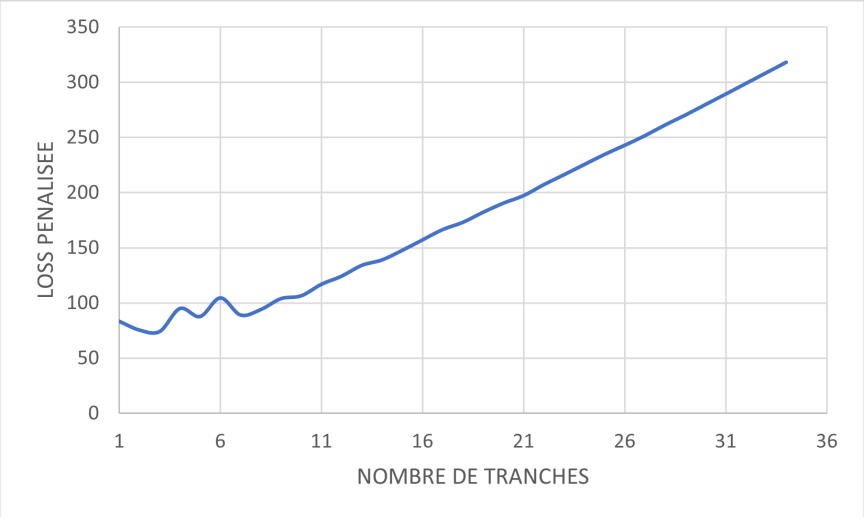


FIGURE 9.18 – Fonction *Loss* pénalisée, modèle des plateaux pondérés

Afin de discerner graphiquement le minimum de la *Loss* pénalisée, la même fonction est tracée uniquement pour les abscisses correspondant aux tranches allant de 1 à 10 :

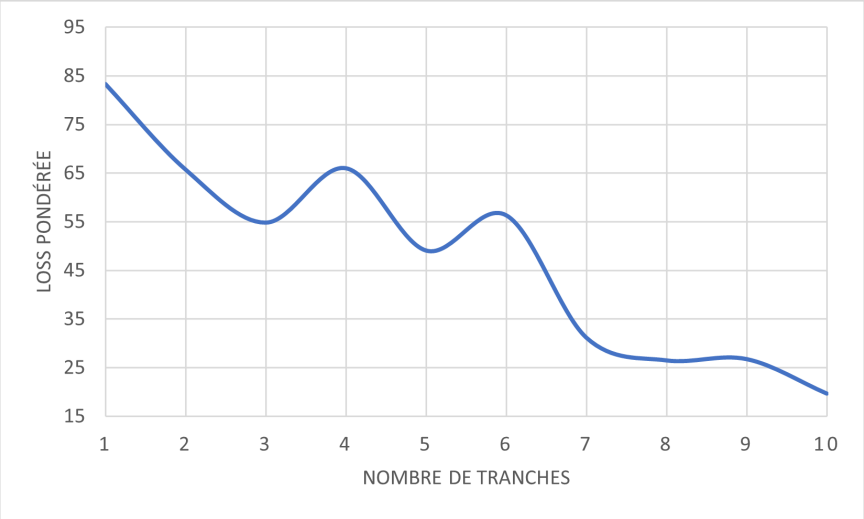


FIGURE 9.19 – Zoom fonction *Loss*, modèle des plateaux pondérés

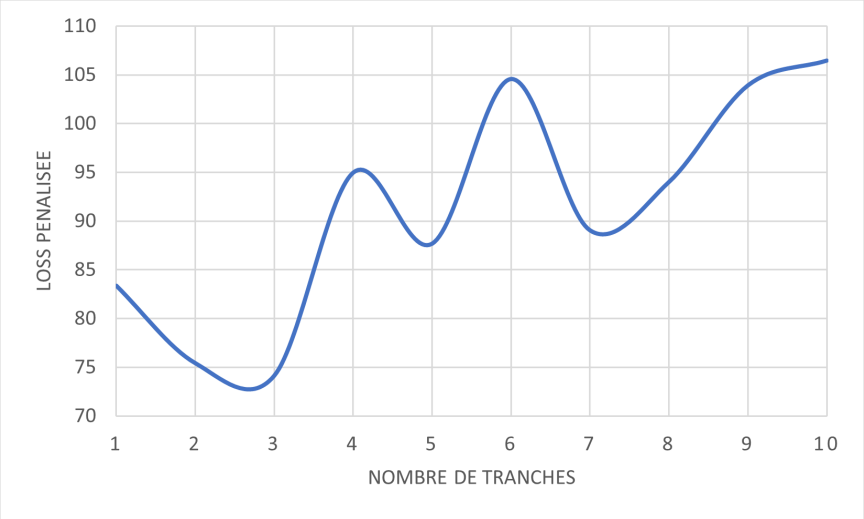


FIGURE 9.20 – Zoom fonction *Loss* pénalisée, modèle des plateaux pondérés

En appliquant la règle de décision établie dans la méthodologie, le nombre de tranches optimal est celui pour lequel la *Loss* pénalisée est la plus faible, i.e :

$$NombreTranchesOptimal = k^* = \underset{i \in [\![1;n]\!]}{argmin}\ Loss^p(\widehat{h^{(i)}})$$

Le nombre optimal de tranches est donc également de trois. Il correspond aux tranches d'âges : 18-34 ans , 35-44 ans et 45-89 ans .
Le graphique associé du SMR par âge superposé à la fonction en escalier optimale de l'algorithme des plateaux est le suivant :

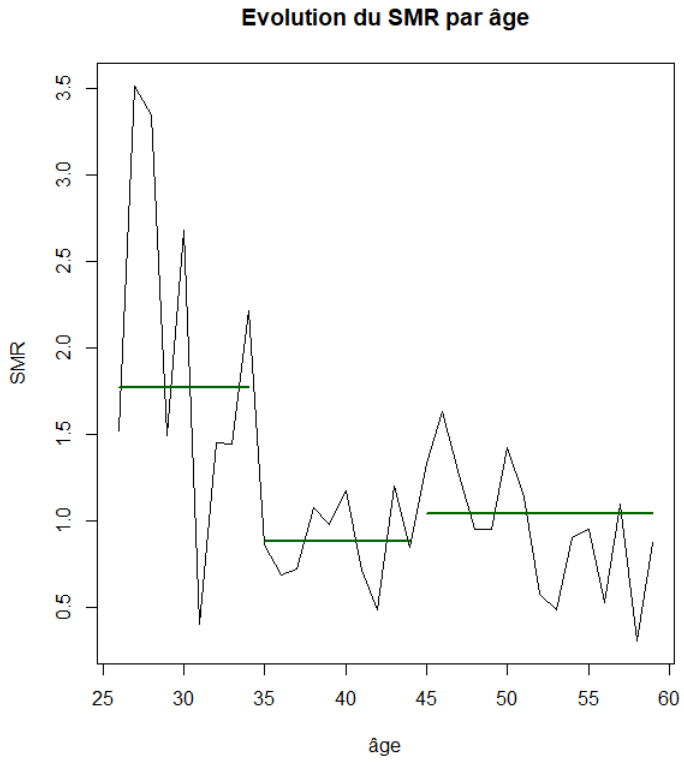


FIGURE 9.21 – Solution optimale, modèle des plateaux pondérés

Conclusions

Les résultats sont résumés dans le tableau suivant :

Méthode utilisée	**Tranches d'âges optimales**	**Répartition de l'exposition**
Méthode des plateaux	$18-38$ / $39-49$ / $50-89$	50% / 39% / 10%
Méthode des plateaux pondérés	$18-34$ / $35-44$ / $45-89$	28% / 47% / 25%

TABLE 9.1 – Résultats méthodes de sélection des tranches d'âge pour construction d'une table de mortalité par abattement

Le modèle qui sera retenu dans cette étude est celui des plateaux pondérés. En effet, les tranches optimales issues de ce modèle sont celles qui ont la meilleure répartition de l'exposition totale de la base de données.

Les méthodes d'estimation de la mortalité du modèle de crédibilité présenté dans le chapitre 8 vont finalement être appliquées sur les tranches $18-34$ ans, $35-44$ ans et $45-89$ ans pour la construction d'une table de mortalité sans prise en compte de l'effet sélection médicale sur le portefeuille ${}_{Lemoine}B_2^{produit\ A}$.

Chapitre 10

Application de la théorie de la crédibilité

La partie précédente a permis de définir une méthode de sélection de tranches d'âges pour un abattement multiple : la méthode des plateaux pondérés. Ce modèle a été introduit pour renforcer l'application de la théorie de la crédibilité de [Klugman *et al.*, 2009] et sa mise en œuvre a été présentée dans la section 9.3. À présent, les méthodes de crédibilité vont être appliquées pour la construction d'une table de mortalité sans effet de sélection médicale et applicable aux assurés éligibles à la loi Lemoine. Pour rappel, l'estimation de ces taux de mortalité est équivalente à la construction de la table sur la base de données ${}_{Lemoine}B_2^{produit\ A}$.

Pour rappel, les produits retenus pour l'application de la méthode bayésienne empirique de Bühlmann sont les produits A, B, et D. [1] De plus, considérant le fait que le paramètre λ a été estimé à 2 ans [2], les bases qui seront utilisées pour l'application du modèle seront les suivantes :

- La base du produit A tronquée sur les deux premières années (voir le graphique Troncature des données pour l'estimation de la mortalité au-delà de j années) et filtrée pour ne conserver que les contrats éligibles à la loi Lemoine. [3] C'est sur cette base de données que la table sans effet de sélection médicale sera construite ;
- La base du produit B tronquée sur les deux premières années ;
- La base du produit D tronquée sur les deux premières années ;
- La base du produit A non tronquée, car elle a servi à l'estimation de la table de mortalité qui sera abattue.

De plus, concernant le choix de la table de référence qui sera abattue dans le modèle, la table actuelle du produit A sera sélectionnée, autrement dit :

$$q_x^{ref} = q_x^{avant\ Lemoine}$$

Cette table a été construite sur les données du portefeuille A en suivant [lig, 2006] et en appliquant les techniques classiques d'estimation de la mortalité présentées dans le chapitre 5 et en Annexe B.1. Ainsi, les taux bruts ont été calculés par la méthode de Hoem, puis lissés consécutivement par le modèle

1. Voir la section 8.6 pour plus de détails.
2. Voir la section 7.4 pour plus de détails.
3. Les conditions d'éligibilité portant sur le capital assuré inférieur à 200 000 €et sur l'âge de la fin de prêt avant 60 ans.

de Whittacker-Henderson et par le modèle de Makeham. Enfin, les taux ont été prolongés par la méthode du logit et lissés une dernière fois par un lissage géométrique.

Le modèle de crédibilité est appliqué indépendamment sur les trois tranches construites par la méthode des plateaux pondérés [4], à savoir $18-34$ ans, $35-44$ ans et $45-89$ ans. Dans un premier temps, les deux facteurs μ et σ sont estimés pour chaque tranche par les formules démontrées dans le chapitre 8 :

$$\widehat{\mu} = \frac{\sum_{h=1}^{r} A_h}{\sum_{h=1}^{r} E_h}$$

et

$$\widetilde{\sigma^2} = \frac{\sum_{h=1}^{r} E_h(\widehat{m_h} - \widehat{\mu})^2 - \widehat{\mu}^2 \left(\frac{\sum_{h=1}^{r}\sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h} \right) - \widehat{\mu} \left(\sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} p\ {}^pE_h}{E_h} - \frac{\sum_{h=1}^{r}\sum_{p=1}^{p_{max}} p\ {}^pE_h}{\sum_{h=1}^{r} E_h} \right)}{\sum_{h=1}^{r} E_h - \frac{\sum_{h=1}^{r} E_h^2}{\sum_{h=1}^{r} E_h} + \frac{\sum_{h=1}^{r}\sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h}}$$

Les résultats sont les suivants :

Tranche d'âges	μ	σ^2
$18-34$ ans	1,04	0,85
$35-44$ ans	1,03	0,008
$45-89$ ans	0,93	0,015

TABLE 10.1 – Résultats sur les données des paramètres de crédibilité μ et σ^2

Une fois ces deux paramètres estimés, les facteurs de crédibilité Z_h sont calculés pour l'ensemble des portefeuilles et pour toutes les tranches d'âges par la formule suivante :

$$\begin{aligned} Z^h &= \frac{\mathbb{E}[m_h \widehat{m_h}] - \mu^2}{\mathbb{E}[\widehat{m_h}^2] - \mu^2} \\ &= \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h \right)} \end{aligned}$$

Les résultats sont les suivants :

Tranche d'âges	${}_{Lemoine}B_2^{produit\ A}$	$B_2^{produit\ B}$	$B_2^{produit\ D}$	$B_0^{produit\ A}$
$18-34$ ans	91,9%	75,3%	72,7%	99,9%
$35-44$ ans	32,3%	18,4%	10,8%	97,2%
$45-89$ ans	66,3%	88,4%	73,4%	94,8%

TABLE 10.2 – Résultats du calcul des facteurs de crédibilité Z_h

Les facteurs de crédibilité qui seront utilisés sont donc ceux de la première colonne du tableau ci-dessus. Pour rappel, le facteur de crédibilité varie entre 0 et 1 selon, entre autres, le volume de données à disposition. Plus il est proche de 1, plus la confiance accordée dans les données du portefeuille est importante. Dans le cas présent, il semblerait que la crédibilité ait un rôle non négligeable, étant donné les valeurs prises par Z_h sur les trois tranches du portefeuille ${}_{Lemoine}B_2^{produit\ A}$.

4. Se référer à la section 9.3.

Enfin, le taux d'abattement à appliquer sur la table de référence pour construire la table de mortalité sur chaque tranche de chaque portefeuille se calcule de la manière suivante :

$$\widetilde{m_h} = Z^h \widehat{m_h} + (1 - Z^h)\mu$$

Les résultats sont les suivants :

Tranche d'âges	${}_{Lemoine}B_2^{produit\ A}$	$B_2^{produit\ B}$	$B_2^{produit\ D}$	$B_0^{produit\ A}$
$18-34$ ans	224,2%	222,7%	191,7%	101,2%
$35-44$ ans	108,7%	108,7%	102,6%	101,2%
$45-89$ ans	100,1%	82,4%	82,2%	100,8%

TABLE 10.3 – Résultats du calcul des taux d'abattement m_h

Finalement, la table de mortalité modélisant le risque d'une population assurée loi Lemoine n'ayant pas passé de sélection médicale s'estime de la manière suivante :

$${}^{\lambda}q_x = \begin{cases} 224,2\% * q_x^{ref} & \forall x \in [\![18;34]\!] \\ 108,7\% * q_x^{ref} & \forall x \in [\![35;44]\!] \\ 100,1\% * q_x^{ref} & \forall x \in [\![45;89]\!] \end{cases}$$

Avec

$$q_x^{ref} = q_x^{avant\ Lemoine}$$

Cette table présente évidemment des irrégularités aux points de raccordement. Un lissage sera effectué, mais pas directement sur cette table. Cela sera décrit dans la partie suivante portant sur la modélisation de l'impact mortalité de la loi Lemoine.
À présent que la loi décrivant la sinistralité d'une population assurée éligible à la loi Lemoine sans effet de sélection médicale a été construite, les différents outils pour la modélisation de l'impact de la loi Lemoine sont à disposition et vont pouvoir être utilisés.

Troisième partie

Modélisation de l'impact sur la mortalité de la loi Lemoine

Cette dernière partie s'attache à développer une méthodologie de quantification de l'impact sur la mortalité de la loi Lemoine. La partie précédente a permis la construction d'une loi de mortalité décrivant la sinistralité sans effet de sélection médicale d'une population assurée éligible à la loi Lemoine. Or, ces taux de décès ne peuvent servir seuls à décrire la mortalité moyenne de tout un portefeuille de *new business* post loi Lemoine. En effet, il est nécessaire de prendre en compte les autres degrés de risques de mortalité composant ce futur portefeuille, tels que les individus atteints de pathologies graves, les individus ayant passé les formalités médicales, *etc.*

Chapitre 11

Méthodologie de construction d'une loi *new business* post loi Lemoine

11.1 Un fort impact attendu sur tous les risques

L'application des deux volets de la loi Lemoine va impacter les différents risques auxquels un assureur délivrant des contrats emprunteur peut faire face :

- **Risque de rachat** : Le titre 1 de la loi Lemoine, permettant la résiliation du contrat d'assurance de la part de l'assuré à tout moment, va impacter la modélisation du risque rachat. En effet, faciliter les démarches de résiliation risque de pousser les assurés à se tenir davantage au courant du marché emprunteur et des opportunités qui se présentent à eux ;
- **Risque de mortalité** : Le titre 2 de la loi Lemoine énonce que, sous conditions d'éligibilité, les individus n'auront plus à devoir passer de formalités médicales. Comme cela a été expliqué dans la section 1.4, ce processus permet à l'assureur une sélection des risques ainsi que de limiter la sinistralité sur ses portefeuille. Or, la garantie décès étant obligatoire sur un contrat emprunteur, le risque décès se verra impacté du fait de l'absence de formalités médicales à l'entrée ;
- **Risque de morbidité** : Pour les mêmes raisons que le risque décès, le risque de morbidité se verra impacté étant donné que les garanties incapacité et invalidité peuvent également être souscrites sur un contrat emprunteur.

11.2 Modélisation de l'impact sur la mortalité

Dans le cadre de cette étude, le seul impact de la loi Lemoine qui sera analysé est celui sur la mortalité. L'objectif est la construction d'une table de mortalité intégrant les effets de la loi Lemoine et décrivant la mortalité moyenne du *new business* sur les portefeuilles emprunteur. Le *new business* étant défini ici pour rappel comme l'ensemble des contrats souscrits après la mise en application de la loi Lemoine.

La méthodologie suivie pour quantifier l'impact sur la mortalité future des portefeuilles post loi Lemoine est maintenant présentée. Le principe est de segmenter une population assurée en fonction de la gravité de leur risque et d'en faire l'analyse avant/après la loi Lemoine.

Pour ce faire, les statistiques de [AERAS, 2020] sur les portefeuilles emprunteur sont utilisées :

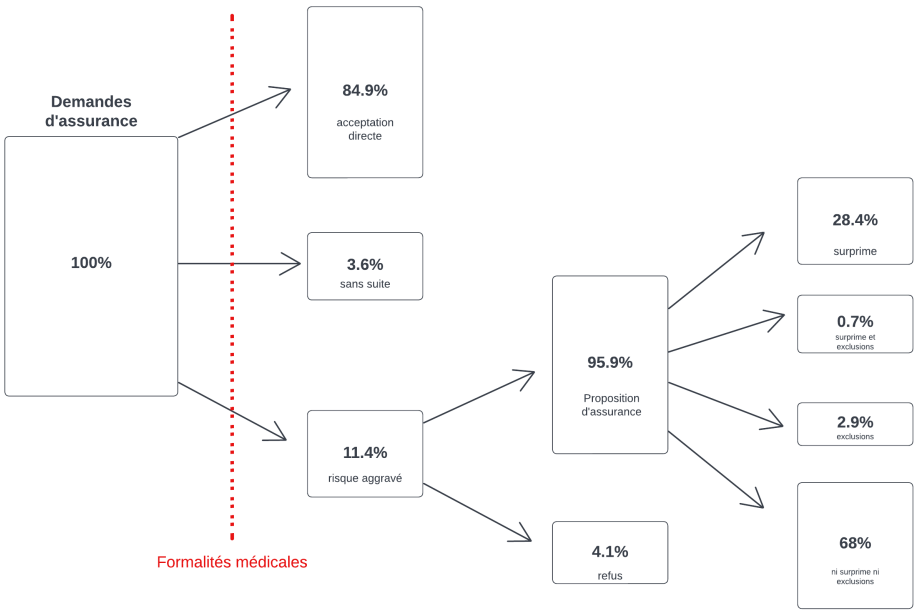


FIGURE 11.1 – Distribution des profils en demande d'assurance emprunteur, AERAS 2020

Pour cette étude, il sera nécessaire de faire des hypothèses de projection concernant les individus non visibles en portefeuille avant la mise en application de la loi Lemoine. Ainsi, le graphique ci-dessus peut être découpé selon deux cas :

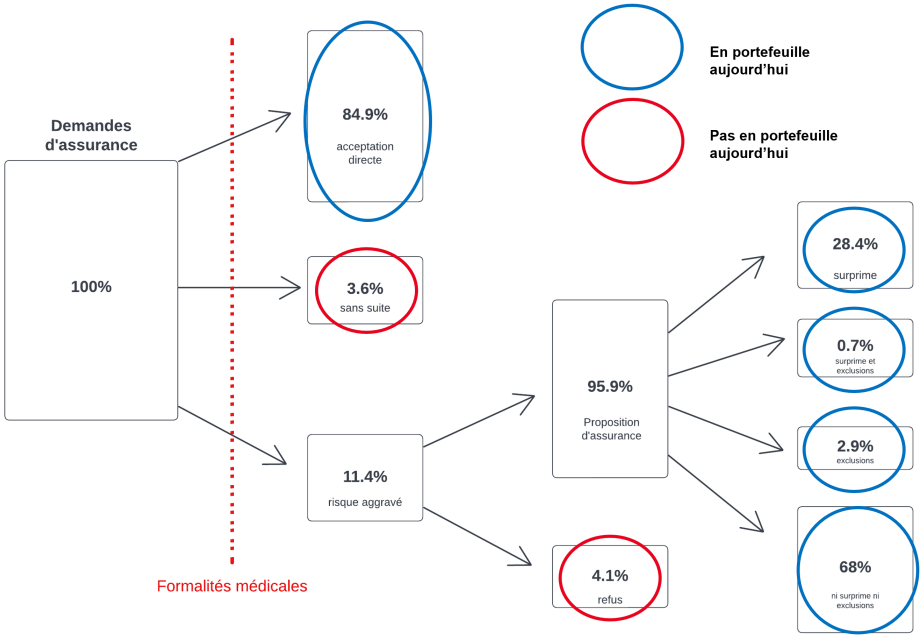


FIGURE 11.2 – Étude des profils de demande d'assurance emprunteur

Dans un premier temps, il est nécessaire de définir les sous-populations qui constitueront un portefeuille emprunteur post loi Lemoine. Suite à cela, il faudra décrire la mortalité de chacun de ces groupes, et enfin la table de mortalité finale sera construite par une moyenne pondérée entre ces différents risques.

La première étape consiste à déterminer l'allure d'un portefeuille après mise en application de la loi Lemoine. Selon les statistiques de l'AERAS ci-dessus, un portefeuille emprunteur sera composé de :

1. Une part d'individus n'ayant pas passé la sélection médicale, car ils remplissaient les conditions d'application de la loi Lemoine (sur le capital et l'âge) mais qui auraient été acceptés directement ;
2. Une part d'individus ayant passé la sélection médicale, car ils ne remplissaient pas les conditions d'application de la loi Lemoine (sur le capital et l'âge) mais qui ont été acceptés directement ;
3. Une production additionnelle, due au caractère attractif, ou « l'effet d'aubaine », de la loi Lemoine pour les risques très aggravés. Ces individus ne seraient jamais allés au bout de leurs démarches de demande d'assurance dans le passé car ils savaient qu'il n'auraient pas pu accéder à une proposition d'assurance. La catégorie « sans suite » est donc attribuée par hypothèse à cette catégorie de risques très aggravés ;
4. Une part d'individus refusés dans le passé à cause d'un risque très aggravé mais qui seraient aujourd'hui acceptés du fait de l'absence de formalités médicales ;
5. Une part de risques aggravés qui auraient eu accès à une proposition d'assurance moyennant des exclusions de garantie avant la loi Lemoine mais qui ont été acceptés directement aujourd'hui, car ils remplissent les conditions d'application de la loi (sur le capital et l'âge) ;
6. Une part d'individus n'ayant pas passé la sélection médicale, car ils remplissaient les conditions d'application de la loi Lemoine (sur le capital et l'âge) mais qui auraient été acceptés avec surprimes ou alors sans aucune forme de malus ;
7. Une part d'individu au risque aggravé ayant reçu une proposition d'assurance après avoir passé les formalités médicales, car ils ne remplissaient pas les conditions d'application de la loi Lemoine (sur le capital et l'âge) ;

Afin de décrire la mortalité moyenne d'un portefeuille post loi Lemoine, il faut donc dans un premier temps pouvoir décrire la mortalité de chacun de ces 7 groupes. Les taux de mortalité du groupe i sont notés $q_x^{(i)}$. La méthodologie appliquée sera la suivante :

Mortalité du premier groupe : les individus sans risque aggravé et n'ayant pas été soumis aux formalités médicales

Les individus composant cette première sous-population, bien qu'ils soient jugés « à faible risque », n'ont tout de même pas passé de formalités médicales. Or, l'effet de sélection médicale est bien observable sur l'ensemble d'un portefeuille. Il convient donc de retirer cet effet sélection médicale dans l'estimation de leur mortalité moyenne.

La mortalité de ce groupe est décrite par la table construite dans la partie précédente sur la base ${}_{Lemoine}B_2^{produit\ A}$. Les taux de mortalité obtenus $\left({}^{\lambda}q_x\right)_x$ par crédibilité représentent bien la sinistralité d'une population assurée éligible à la loi Lemoine et n'ayant pas passé de formalités médicales à l'entrée. Finalement :

$$q_x^{(1)} = {}^{\lambda}q_x$$

Mortalité du deuxième groupe : les individus sans risque aggravé et ayant été soumis aux formalités médicales

Ces individus représentent la partie « acceptation directe » du portefeuille emprunteur de *new business* avec un capital excédant 200 000 € par tête ou bien avec un âge à la fin de contrat dépassant les 60 ans ou bien les deux. Ils sont ainsi semblables aux individus déjà présents en portefeuille aujourd'hui,

car les deux groupes ont bien passé une sélection médicale. Pour décrire la mortalité future de ce deuxième groupe, les taux de mortalité actuels $q_x^{avant\ Lemoine}$ utilisés dans un contexte avant loi Lemoine seront appliqués [1]. Finalement :

$$q_x^{(2)} = q_x^{avant\ Lemoine}$$

Mortalité du troisième et quatrième groupe : les risques très aggravés non présents en portefeuille aujourd'hui

Afin de décrire la mortalité de ces individus, il convient de sélectionner de fortes hypothèses de mortalité car leur risque a été est jugé trop important jusqu'à présent pour être assuré. Le comportement de ces deux groupes est fortement anti-sélectif.

Des statistiques nationales sur les taux de mortalité des individus atteints de cancer vont être utilisées pour décrire la mortalité moyenne de ces deux groupes.

Cependant, cette étude soulève deux questions supplémentaires :

- La mortalité d'un individu assuré en emprunteur présentant un risque très aggravé tel qu'un cancer sera-t-elle la même qu'un autre individu de la population générale française atteint d'un cancer ?
- Dans le cas où la réponse est non, comment décrire la mortalité de l'individu assuré atteint d'un cancer ?

Ces deux réflexions ont été portées dans un article de [Planchet *et al.*, 2022]. C'est sur la base de cet article que la mortalité des groupes 3 et 4 sera estimée.

Le premier constat de l'article vient en réponse à la première question : à un même niveau de pathologie, la mortalité moyenne d'un individu assuré est plus faible que celle d'un individu de la population générale. Cela provient de variables cachées telles que le niveau de vie ou la catégorie socioprofessionnelles (CSP). En effet, la population assurée présente en moyenne de meilleures conditions de vie que la population générale. Le fait de pouvoir souscrire un prêt est souvent indicateur d'une CSP plus élevée que la moyenne. Ce meilleur niveau de vie moyen de la population assurée permet alors un meilleur accès aux soins et souvent une détection plus rapide de leur pathologie.

Ce constat fait, il reste la question de l'estimation de la mortalité de la population assurée atteinte d'une pathologie cancéreuse. Comme le pointe l'article, c'est principalement à cause de la méconnaissance de ce risque que les assureurs ont tendance à le refuser. Cette méconnaissance est liée au faible volume de données disponibles. L'idée est donc l'utilisation de l'open data.

La mortalité moyenne d'une population assurée atteinte d'une pathologie peut être approchée de la manière suivante :

$$q_x^{emprunteur,pathologie} = q_x^{emprunteur} * \frac{q_x(pop\ generale\ pathologie)}{q_x(pop\ generale)} * multiplicateurRR$$

où :

- multiplicateurRR est le chiffre indiquant à quel point la sinistralité de la population assurée est plus faible que celle de la population générale. Ce chiffre est estimé à $0,83$ dans l'article de Planchet 2022 pour le cas particulier du cancer du sein. Cette valeur sera reprise dans le cadre de cette étude ; [2]

1. Pour rappel, cette table a été construite sur le produit A par les méthodes classiques d'estimation de la mortalité présentées en Annexe B.1.

2. Ce choix est justifié par le fait que le cancer du sein est celui le plus prépondérant au monde([Espié *et al.*, 2012]).

- $q_x(pop\ \ generale)$ les taux de mortalité de la population générale en France, cette table sera sélectionnée comme la moyenne pondérée par les *Sexe Ratio* du produit A entre la TH-02 et la TF-02 ;

- $q_x(pop\ generale\ cancer)$ les taux de mortalité de la population générale en France atteinte de cancer. Ces taux vont être estimés à partir d'une étude qui sera présentée dans la partie des résultats.

Dans le cadre de l'étude de la mortalité des groupes 3 et 4, leurs taux de mortalité seront exprimés de la façon suivante :

$$q_x^{(3)} = q_x^{(4)} = q_x^{emprunteur,cancer} = q_x^{emprunteur} * \frac{q_x(pop\ generale\ cancer)}{q_x(pop\ generale)} * multiplicateurRR$$

Mortalité du cinquième groupe : les individus qui auraient dû avoir des exclusions

Il est visible sur les statistiques [AERAS, 2020] que les individus ayant un risque aggravé mais acceptés en portefeuille n'ont pas tous reçu de proposition d'assurance dans les mêmes termes. En effet, certains ont eu des surprimes, exclusions, les deux ou bien aucune forme de malus. Concernant la modélisation de la mortalité de ces individus, il est à noter que seule la mortalité moyenne de ceux qui possédaient des exclusions va évoluer post loi Lemoine. En effet, les individus surprimés auront un impact sur le volume de primes du portefeuille mais pas sur la sinistralité moyenne, étant donné que 100% de leur risque était déjà assumé par l'assureur.

La mortalité du deuxième groupe d'individus à estimer est donc celle des individus qui auraient dû avoir des exclusions de garantie mais qui n'en n'auront pas à l'avenir. Ce point doit faire l'objet d'hypothèses comme il repose sur un événement qui n'a jamais été observé. Ainsi, il sera supposé que la mortalité des individus présentant des exclusions de garantie est la même que les individus présentant des surprimes. De cette façon, avec la connaissance du taux de surprime, les taux de mortalité des contrats présentant des exclusions de garanties peuvent s'en déduire directement. En effet, le taux de surprime représente le surplus de risque par rapport à la population moyenne assurée.

$$PC = PP * (1 + c)$$

et

$$\begin{aligned} PC_{risques\ aggraves} &= PP_{risques\ aggraves} * (1 + c) \\ &= PP * (1 + \tau) * (1 + c) \end{aligned}$$

avec

- PC la prime commerciale ;
- PP la prime pure ;
- c le taux de chargement exprimé en pourcentage de la prime pure ;
- τ le taux de surprime.

L'espérance du risque, représenté par la prime pure, se déduit alors pour la sous-population des risques aggravés :

$$\mathbb{E}[Risque_{aggraves}] = \mathbb{E}[Risque_{classique}] * (1 + \tau)$$

La mortalité moyenne d'un individu présentant un risque aggravé sera donc approchée comme telle :

$$q_x^{risques\ aggraves} = q_x^{population\ assuree} * (1 + \tau)$$

Finalement, il est à noter que le taux de surprime n'est pas le même pour toute la population assurée présentant un risque aggravé. Sur ce sujet, les statistiques 2020 de l'AERAS présentent les résultats suivants :

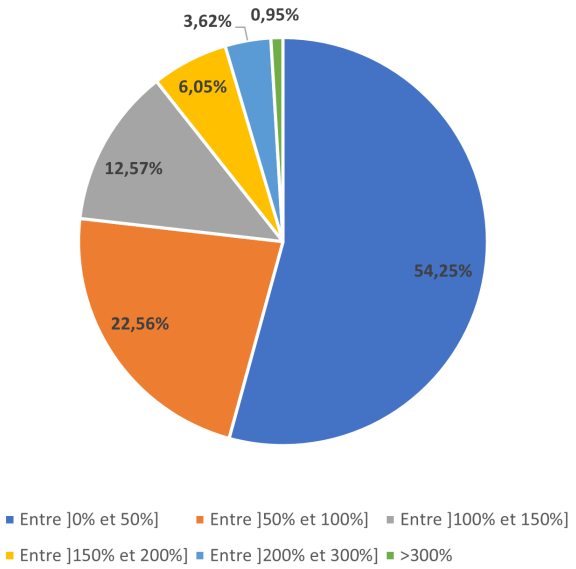


FIGURE 11.3 – AERAS 2020 : répartition des demandes d'assurance de prêts présentant un risque aggravé de santé ayant fait l'objet d'une proposition d'assurance avec surprime

Ainsi, à partir de ces statistiques il est possible de calculer un taux de surprime moyen sur les contrats emprunteur en France :

$$\overline{\tau} = \sum_{sous\_population} \alpha_{sous\_population} * \overline{\tau}_{sous\_population}$$

Où

- $\alpha_{sous\_population}$ représente la proportion de la sous-population par rapport au total, i.e $\alpha_{sous\_population} = \frac{taille_{sous\_population}}{taille_{population}}$;
- $\overline{\tau}_{sous\_population}$ est le taux de surprime moyen pour une sous-population donnée.

La mortalité du groupe 5, représentant les individus qui auraient dû présenter des exclusions de garantie s'ils avaient passé les formalités médicales, sera enfin décrite comme telle :

$$q_x^{(5)} = q_x^{population\ assuree} * (1 + \overline{\tau})$$

Où

- $q_x^{population\ assuree}$ est le taux de mortalité du portefeuille assuré ne présentant pas de risque aggravé. Dans cette étude, ce taux sera choisi comme celui du groupe 1.

Mortalité du sixième groupe : les individus de la catégorie « risque aggravé » qui auraient dû avoir des surprimes

Ces individus n'ont pas passé la sélection médicale car ils remplissaient les conditions d'application de la loi Lemoine (sur le capital et l'âge) mais ils auraient été acceptés avec surprimes ou alors sans forme de malus si la loi Lemoine n'avait pas vu le jour. Étant donné qu'ils n'ont pas passé de formalités médicales, leur mortalité sera estimée par la loi construite par crédibilité qui ne prend pas en compte l'effet de sélection médicale :

$$q_x^{(6)} = {}^{\lambda}q_x$$

Mortalité du septième groupe : les individus avec risques aggravés, mais acceptés avec formalités médicales et exclusions

Ces individus, bien qu'ayant un risque aggravé par rapport à la moyenne, sont observables avant la mise en vigueur de la loi Lemoine en portefeuille. De la même manière que ceux du groupe 2, ils ont servi à construire la loi de mortalité actuelle utilisée, $q_x^{avant\ Lemoine}$. Cette table intègre leur sinistralité et à ce titre elle peut être utilisée pour prédire la sinistralité du groupe 7. Ainsi :

$$q_x^{(7)} = q_x^{avant\ Lemoine}$$

Résumé des sous-groupes de risque sur un portefeuille emprunteur de *new business* post loi Lemoine

Finalement, les différents groupes de risques d'un portefeuille emprunteur post loi Lemoine peuvent être résumés sur le schéma suivant :

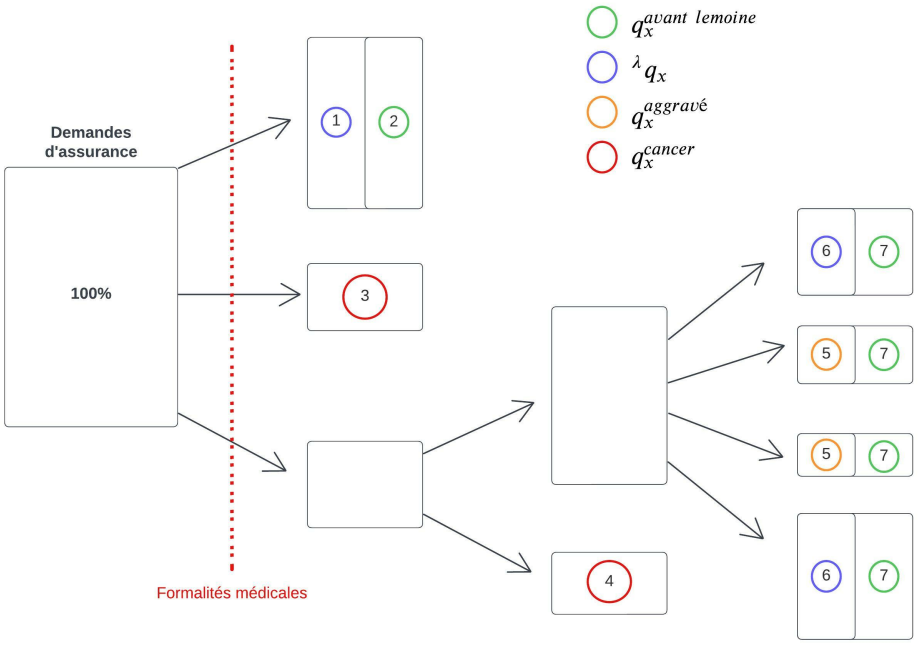


FIGURE 11.4 – Distribution des profils de risques dans un portefeuille emprunteur

11.3 Loi de mortalité moyenne sur un portefeuille *new businesss* post loi Lemoine

Finalement, après avoir estimé la mortalité de toutes les sous-populations qui vont constituer un portefeuille emprunteur de *new businesss* post loi Lemoine, il suffit d'en faire une moyenne pondérée pour obtenir un q_x moyen sur le portefeuille. Les poids sont les proportions que représentent chaque sous-population. Ces derniers sont estimés à partir des statistiques de l'AERAS présentées plus haut.

Finalement, le taux de mortalité moyen sur le portefeuille, q_x^{moyen}, est construit de la manière suivante :

$$q_x^{moyen} = \sum_{i=1}^{7} \alpha_x^{(i)} q_x^{(i)}$$

Remarque : Le poids $\alpha_x^{(i)}$ dépend de l'âge x car les sous-groupes ont été définis selon un critère d'éligibilité à la loi Lemoine. Il convient donc par exemple de ne pas comprendre dans le q_x^{moyen} la mortalité associée à la pathologie cancéreuse au-delà de 60 ans, étant donné que les individus devront obligatoirement passé des formalités médicales passé cet âge. Les poids choisis seront décrits dans le chapitre suivant traitant des résultats de l'application de ce modèle.

Chapitre 12

Application de la méthodologie

La méthodologie de construction de la table de mortalité décrivant la sinistralité moyenne d'une population *new business* post loi Lemoine est maintenant appliquée. **Par souci de confidentialité, les échelles des graphiques représentant les différents taux de mortalité sont retirées. De plus, les impacts sont uniquement donnés en pourcentage, sans préciser les valeurs calculées des nombres de décès et des charges ultimes.**

12.1 Résultats estimation de la mortalité des 7 groupes loi Lemoine

À présent, la mortalité des 7 groupes qui constitueront un portefeuille de *new business* post loi Lemoine va être estimée.

Groupe 1

La mortalité de ce groupe est estimée par la loi ne prenant pas en compte l'effet de sélection médicale présentée dans le chapitre 7 :

$$q_x^{(1)} = {}^{\lambda}q_x$$

Pour rappel, le facteur λ a été estimé à 2 par *bootstrap*. La table est construite par la théorie de la crédibilité, dont les résultats ont été présentés dans la section 10. Finalement :

$$q_x^{(1)} = \begin{cases} 224,2\% * q_x^{avant\ Lemoine} & \forall x \in [\![18;34]\!] \\ 108,7\% * q_x^{avant\ Lemoine} & \forall x \in [\![35;44]\!] \\ 100,1\% * q_x^{avant\ Lemoine} & \forall x \in [\![45;89]\!] \end{cases}$$

Groupe 2

La mortalité de ce groupe est celle des individus ayant passé une sélection médicale, elle sera estimée à partir de la table de mortalité actuelle du produit A, construite dans un contexte avant loi Lemoine. [1]

$$q_x^{(2)} = q_x^{avant\ Lemoine}$$

1. Voir le chapitre 10 pour plus de détails sur la construction de cette table.

Il est important de noter que faire cette hypothèse est une simplification de la réalité. En effet, la table de mortalité « avant Lemoine » a été construite sur l'ensemble du portefeuille du produit A. Or, dans le cadre de cette étude, ces taux de mortalité vont être appliqués à des individus non éligibles à la loi Lemoine, c'est-à-dire une grande part de contrats pour lesquels la part assurée va être supérieure à 200 000 euros. Or, rien ne justifie que la mortalité n'est pas corrélée avec le capital emprunté.

Cependant, il est souvent remarqué sur les portefeuilles emprunteur que la sinistralité diminue quand le montant emprunté augmente. L'hypothèse faite dans le cadre de cette étude est donc malgré tout prudente.

Groupe 3

Ce groupe est assimilé à la production additionnelle engendrée par la loi Lemoine. L'hypothèse est que le comportement de ces futurs assurés est fortement anti-sélectif, c'est pourquoi un taux de mortalité très aggravé, celui des individus atteints pathologies cancéreuses, leur est attribué. La méthodologie utilisée pour l'estimation d'une loi de mortalité d'une population assurée atteinte d'une certaine pathologie est celle développée par [Planchet *et al.*, 2022]. Pour rappel :

$$q_x^{(3)} = q_x^{emprunteur} * \frac{q_x(pop\ generale\ cancer)}{q_x(pop\ generale)} * multiplicateurRR$$

où :

- multiplicateurRR est le chiffre indiquant à quel point la sinistralité de la population assurée est plus faible que celle de la population générale. Ce chiffre est estimé à $0,83$ dans l'article de Planchet 2022 pour le cas particulier du cancer du sein. Cette valeur sera reprise dans le cadre de cette étude.
- $q_x(pop\ \ generale)$ les taux de mortalité de la population générale en France, cette table sera sélectionnée comme la moyenne pondérée par les *Sexe Ratio* du produit A entre la TH-02 et la TF-02.
- $q_x(pop\ \ generale\ \ cancer)$ les taux de mortalité de la population générale en France atteinte de cancer. Ces taux vont être estimés à partir de l'étude de 2017 présentée ci-dessous :

Selon le rapport technique sur la « Projection de l'incidence et de la mortalité par cancer en France métropolitaine en 2017 », présenté par [public France, 2017], la répartition des incidences et des décès pour causes de cancer en France en 2017 était la suivante :

	Incidence		Mortalité	
Classe d'âge	Homme	Femme	Homme	Femme
[00 ; 14]	919	764	96	78
[15 ; 49]	14 472	25 874	2 757	2 892
[50 ; 64]	57 690	50 616	17 355	11 967
[65 ; 74]	75 247	46 442	24 625	14 517
[75 ; 84]	45 728	35 881	22 207	16 504
[85; ++]	19 965	26 028	17 060	20 245
Total	214 021	185 605	84 100	66 203

FIGURE 12.1 – Incidence et mortalité des cancers en France en 2017

Pour une classe d'âges donnée, le taux de mortalité pour cause de cancer peut être estimé de la façon suivante :

$$q_x^{classe} = \frac{deces^{classe}}{incidence^{classe}}$$

La courbe de mortalité des individus de la population générale atteints de pathologies cancéreuses sera donc construite par palier.

Finalement, il est nécessaire de noter les limites de cette approche. L'utilisation de ces données permet une approximation des taux de mortalité réels des individus atteints de pathologies cancéreuses car l'espérance de vie de ces derniers est relativement faible. Cependant, les données d'incidence et de mortalité présentées pour l'année 2017 peuvent être séparées et les individus décédés ne sont pas tous comptabilisés dans l'incidence 2017. Il est aussi à noter que l'hypothèse de mortalité prise pour modéliser la sinistralité de ces individus, à savoir des taux de décès d'une pathologie cancéreuse, est très anti-sélective. Le modèle étant paramétrable, il est tout à fait possible de choisir un autre niveau de mortalité pour décrire le risque de ce groupe.

Groupe 4

Le groupe 4 est celui des individus refusés dans le passé suite à la sélection médicale. Il est supposé que ces individus reviendront en portefeuille avec le même risque aggravé, de type cancer par hypothèse. Leur mortalité est donc estimée de la même manière que celle du groupe 3.

$$q_x^{(4)} = q_x^{emprunteur} * \frac{q_x(pop\ generale\ cancer)}{q_x(pop\ generale)} * multiplicateurRR$$

Groupe 5

Le groupe 5 est celui des individus ayant précédemment eu des exclusions de garantie mais qui n'en auront plus dans le futur, car ils sont éligibles à la loi Lemoine. La méthodologie de calcul a été présentée dans la section 11.1. Le principe est de se baser sur un taux de surprime moyen, qui reflète le surplus de sinistralité des risques aggravés :

$$\begin{aligned} q_x^{(5)} &= \overline{\tau} * q_x^{(1)} \\ &= \sum_{sous\_population} \alpha_{sous\_population} * \overline{\tau}_{sous\_population} * q_x^{(1)} \\ &= (54,25\% * (1+25\%) + ... + 0,95\% * (1+300\%)) * q_x^{(1)} \\ &= 166\% * q_x^{(1)} \end{aligned}$$

Pour rappel, les chiffres utilisés proviennent des statistiques AERAS, voir la figure « AERAS 2020 : répartition des demandes d'assurance de prêts présentant un risque aggravé de santé ayant fait l'objet d'une proposition d'assurance avec surprime » pour les détails.

Groupe 6

Le groupe 6 est celui représentant les individus de la catégorie « risque aggravé » étant éligibles à la loi Lemoine. Étant donné qu'ils n'ont pas passé de formalités médicales, leurs taux de mortalité sont ceux de la table n'intégrant pas l'effet sélection médicale :

$$q_x^{(6)} = {}^{\lambda}q_x$$

Groupe 7

Le groupe 7 est celui représentant les individus de la catégorie « risque aggravé » n'étant pas éligibles à la loi Lemoine. Étant donné qu'ils passeront des formalités médicales, leurs taux de mortalité sont ceux de la table intégrant l'effet sélection médicale :

$$q_x^{(7)} = q_x^{avant\ Lemoine}$$

Concernant la modélisation de la mortalité de ce groupe, la remarque est la même que celle du groupe 2 : utiliser la loi de mortalité avant Lemoine est une hypothèse simplificatrice, mais prudente.

Comparaison globale des mortalités

La mortalité des 7 groupes décrits ci-dessus est donc estimée à partir de 4 lois différentes. À titre comparatif, les 4 lois sont représentées graphiquement :

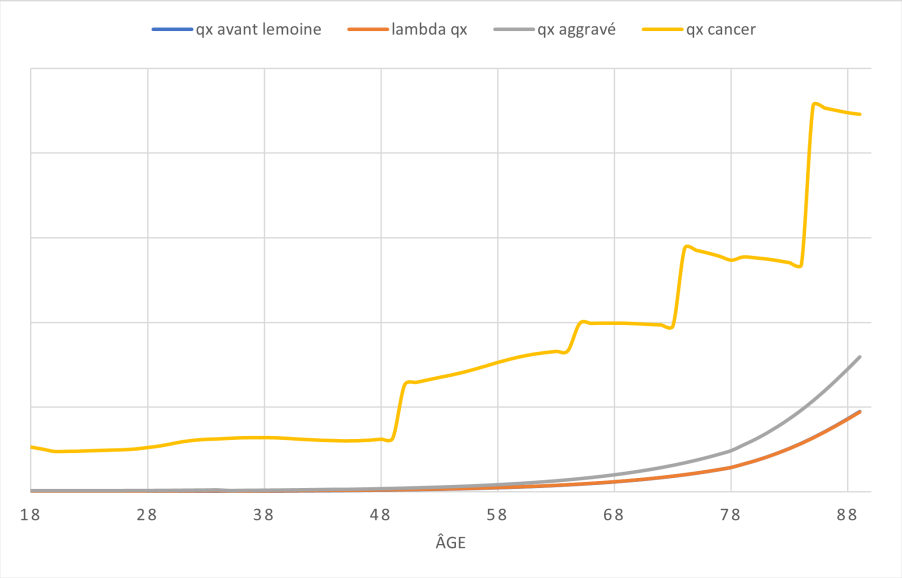


FIGURE 12.2 – Lois des 4 groupes de risques : portefeuille emprunteur

Afin de mieux discerner les différences entre les deux courbes, les tables $q_x^{avant\ Lemoine}$ et ${}^{\lambda}q_x$ sont tracées sur la plage d'âge 18-50 :

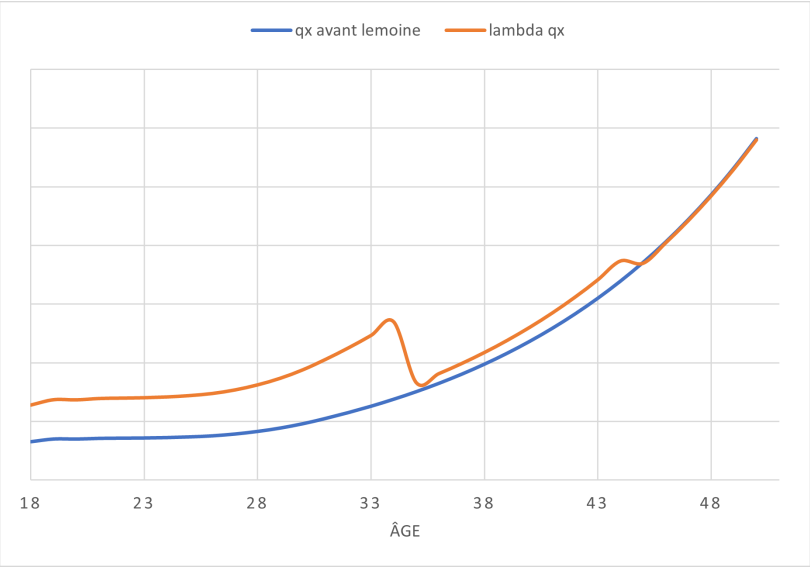


FIGURE 12.3 – Zoom sur les lois des groupes 1,2,6 et 7

Comme attendu, les inégalités suivantes sont vérifiées :

$$q_x^{avant\ Lemoine} < {}^{\lambda}q_x < q_x^{aggrave} < q_x^{cancer\ assures}$$

Il est à noter qu'à ce stade, les différentes lois estimées ne sont pas encore lissées, même si elles présentent certaines irrégularités. Le lissage des taux se fera en toute dernière étape, après avoir calculé la moyenne pondérée de ces différentes tables. Cela permettra d'appliquer un unique lissage et non plusieurs.

12.2 Résultats sur la mortalité moyenne d'un portefeuille *new business* post loi Lemoine

Pour rappel, les statistiques de l'AERAS sur la répartition des demandes d'assurance sont les suivantes :

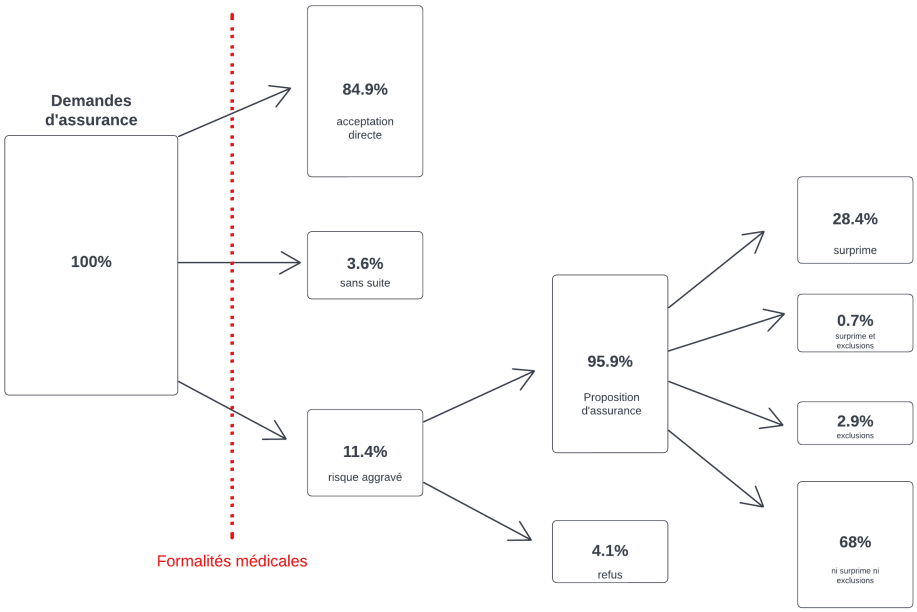


FIGURE 12.4 – Distribution des profils en demande d'assurance emprunteur, AERAS 2020

Les différentes sous-populations de risque ont été identifiées comme telles :

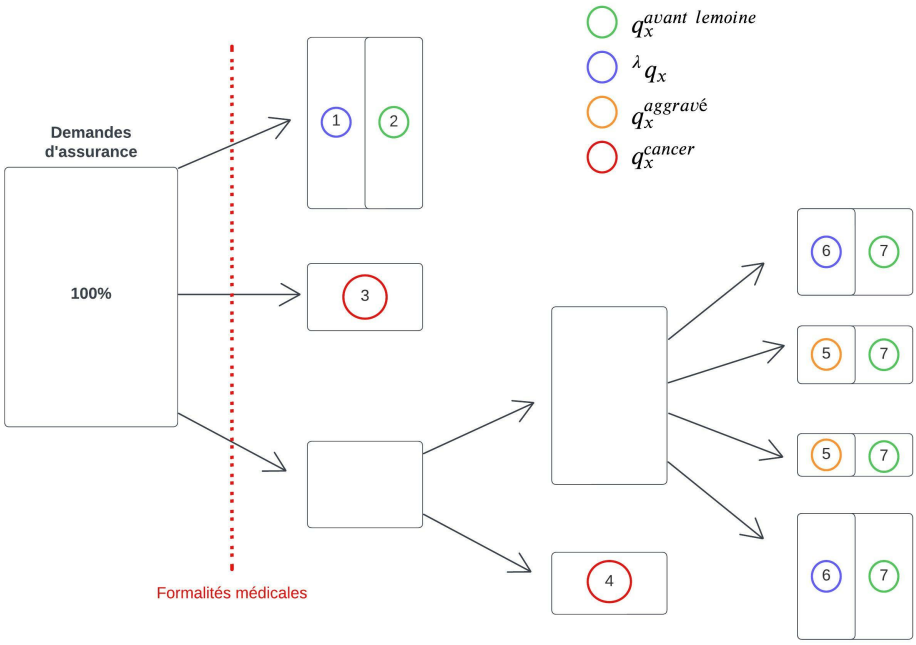


FIGURE 12.5 – Distribution des profils de risque dans un portefeuille emprunteur

À présent, une fois les 4 lois de mortalité estimées, la mortalité moyenne d'un portefeuille de *new business* post loi Lemoine peut être modélisée. Cette table est construite par moyenne pondérée des différentes lois de mortalité par les poids que représente chacun des 7 groupes.

Afin d'estimer quelle proportion représente chaque groupe, il est nécessaire de déterminer quelle part de chaque sous-population est éligible à la loi Lemoine. Pour rappel, les critères d'éligibilité portent sur la part de capital assuré au cumulé et sur l'âge à la fin de contrat. Un taux d'éligibilité par âge, noté TE_x est estimé à partir des données du portefeuille A comme tel :

$$TE_x = \frac{Nombre\ assures\ eligibles\ loi\ Lemoine\ (x)}{Nombre\ assures\ portefeuille\ A\ (x)}$$

Les proportions $\alpha_x^{(i)}$ que représente chacun des 7 groupes à l'âge x peuvent maintenant être calculées :

1. $\alpha_x^{(1)} = 84,9\% * TE_x$;
2. $\alpha_x^{(2)} = 84,9\% * (1 - TE_x)$;
3. $\alpha_x^{(3)} = 3,6\% * TE_x$;
4. $\alpha_x^{(4)} = 11,4\% * 4,1\% * TE_x$;
5. $\alpha_x^{(5)} = 11,4\% * 95,9\% * (0,7\% + 2,9\%) * TE_x$;
6. $\alpha_x^{(6)} = 11,4\% * 95,9\% * (28,4\% + 68,0\%) * TE_x$;
7. $\alpha_x^{(7)} = 11,4\% * 95,9\% * (1 - TE_x)$.

Finalement, étant donné que la somme de ces poids n'est pas égale à 1 [2] pour chaque âge donné, ils doivent être normalisés :

$$\alpha_x^{(i)} \leftarrow \frac{\alpha_x^{(i)}}{\sum_{i=1}^{7} \alpha_x^{(i)}}$$

Il vient :

$$q_x^{moyen} = \sum_{i=1}^{7} \alpha_x^{(i)} q_x^{(i)}$$

avec :

$$\sum_{i=1}^{7} \alpha^{(i)} = 1$$

Enfin, ces taux de mortalité moyens doivent être lissés étant donné les irrégularités des différentes tables qui les composent. Un lissage de Makeham, présenté en Annexe B.3.1, est réalisé sur la table $(q_x^{moyen})_x$.

Graphiquement, le résultat est le suivant :

2. La somme est différente de 1 en raison des individus non éligibles à la loi dans les groupes « sans suite » et « refusés ».

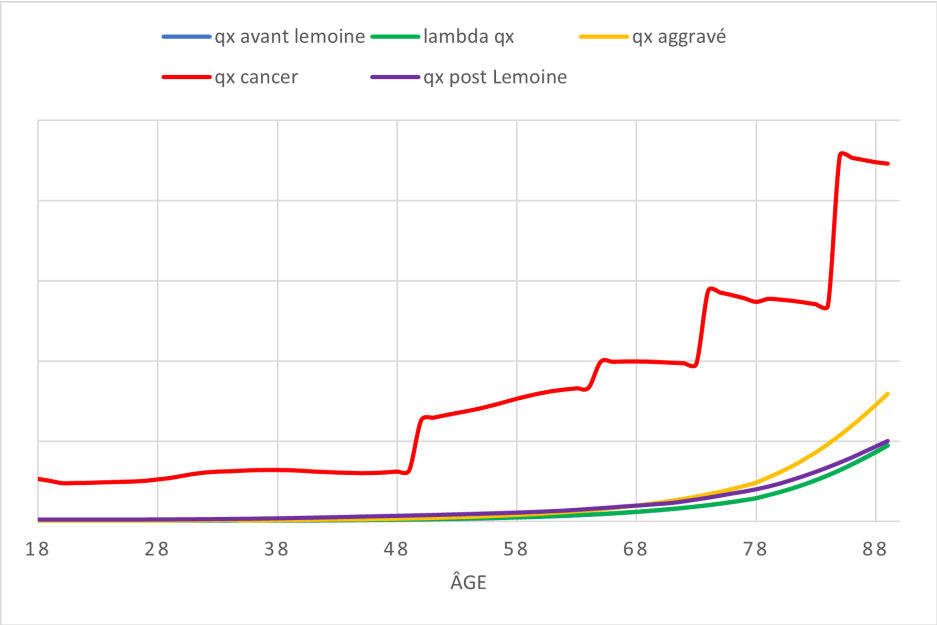


FIGURE 12.6 – Comparaison des 4 lois de mortalité des 7 groupes constituant un portefeuille post loi Lemoine

Afin de mieux discerner les différences entre les lois, le graphique sans la loi de mortalité des individus atteints de pathologies cancéreuses est également présenté :

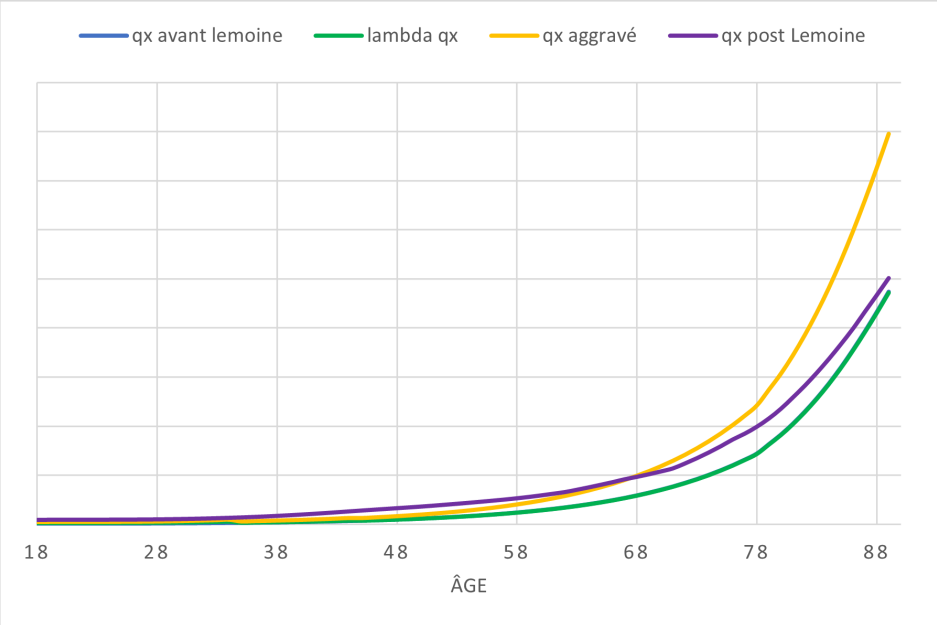


FIGURE 12.7 – Comparaison des lois de mortalité des différents groupes de risques or cancer

La courbe décrivant les q_x^{moyen}, la mortalité moyenne de la population *new business* post loi Lemoine, est bien encadrée entre les 4 différentes lois de mortalité décrivant la mortalité des 7 groupes de risque.

12.3 Estimation d'un premier impact de la loi Lemoine

À présent que la loi de mortalité du *new business* post loi Lemoine a été estimée, il est possible de quantifier un premier impact sur la mortalité de la mise en application de la réforme.

Dans un premier temps, le nombre de décès attendus sur le portefeuille du produit A avec l'ancienne loi est comparé au nombre de décès attendus avec la nouvelle table de mortalité post loi Lemoine. Le calcul du nombre de décès attendus est réalisé comme tel :

$$Deces_{attendus} = \sum_x n_x * q_x$$

En d'autres termes, les taux de mortalité sont appliqués à l'exposition du portefeuille. L'exposition prise est celle de la base qui a servi à estimer la loi de mortalité du produit A, à savoir les données sur la période 2015-2018.
Une fois les deux nombres de décès attendus calculés, leur différence relative est calculée :

$$\frac{Deces_{attendus}^{nouvelle\ loi} - Deces_{attendus}^{ancienne\ loi}}{Deces_{attendus}^{ancienne\ loi}} = 55\%$$

Ce résultat signifie qu'au lieu de prédire 100 sinistres, la nouvelle loi de mortalité post loi Lemoine en prédira 155.
Il est nécessaire de prendre du recul sur ce résultat qui semble très important. En effet, en appliquant la nouvelle loi de mortalité à l'exposition totale du portefeuille A, il est considéré que l'ensemble du portefeuille est constitué de *new business*. Cet impact en nombre est en réalité une projection de la sinistralité dans une vingtaine d'années, quand l'ensemble du portefeuille ne sera constitué que de « *new business* loi Lemoine ».

Afin d'obtenir un impact sur 1 an, il convient alors de considérer le stock. L'impact qui va être calculé à présent est un impact en montant. L'hypothèse est faite que les capitaux sous risques [3](CSR) 2022 seront similaires à ceux de l'année 2021. De cette manière, il est possible d'utiliser les CSR 2021 pour calculer un impact.
La charge ultime 2021 est calculée de deux manières :

- en appliquant aux CSR l'ancienne loi ;
- en appliquant aux CSR stock 2021 l'ancienne loi et en appliquant aux CSR *new business* (NB) la nouvelle loi post loi Lemoine.

Le calcul est le suivant :

$$Charge\ ultime_{2021}^{old} = \sum_x CSR_x * q_x^{avant\ Lemoine}$$

et

$$Charge\ ultime_{2021}^{new} = \sum_x CSR_x^{stock} * q_x^{avant\ Lemoine} + \sum_x CSR_x^{NB} * q_x^{post\ Lemoine}$$

A nouveau, la différence relative entre les deux charges ultimes est calculée :

$$Impact_{un\ an} = \frac{Charge\ ultime_{2021}^{new} - Charge\ ultime_{2021}^{old}}{Charge\ ultime_{2021}^{old}} = 5,27\%$$

3. En assurance emprunteur, les capitaux sous risque représentent les capitaux restant dus des prêts multipliés par leur quotité.

Il a été question dans cette section de « premier impact » car un dernier paramètre n'a encore été pris en compte dans la construction de la loi de mortalité décrivant la sinistralité du *new business* : la proportion d'individus atteints d'une pathologie cancéreuse au sein de la catégorie sans suite. En effet, jusqu'à présent, il était supposé que la totalité de cette catégorie était attitrée à la production additionnelle qui reviendrait en portefeuille par l'effet d'aubaine de la loi Lemoine. Cette hypothèse doit être calibrée car elle peut sembler trop forte, un travail de sensibilité sur cette dernière est donc nécéssaire.

12.4 Sensibilité sur la production additionnelle

Dans les résultats d'impacts sur la mortalité de la loi Lemoine présentés ci-dessus, l'hypothèse de production additionnelle était de supposer que la totalité des demandes « sans suite » reviendrait avec un risque très aggravé [4]. Cette hypothèse peut sembler trop prudente, il est nécessaire de calculer des sensibilités sur les impacts finaux estimés en faisant varier la proportion de risques très aggravés au sein du bloc « demandes sans suite », autrement nommé le sous-groupe de risque 3.

Le schéma ci-dessous représente le surplus de mortalité en nombre en faisant varier en abscisse le pourcentage d'individus de la catégorie « sans suite » étant atteints de pathologies cancéreuses.

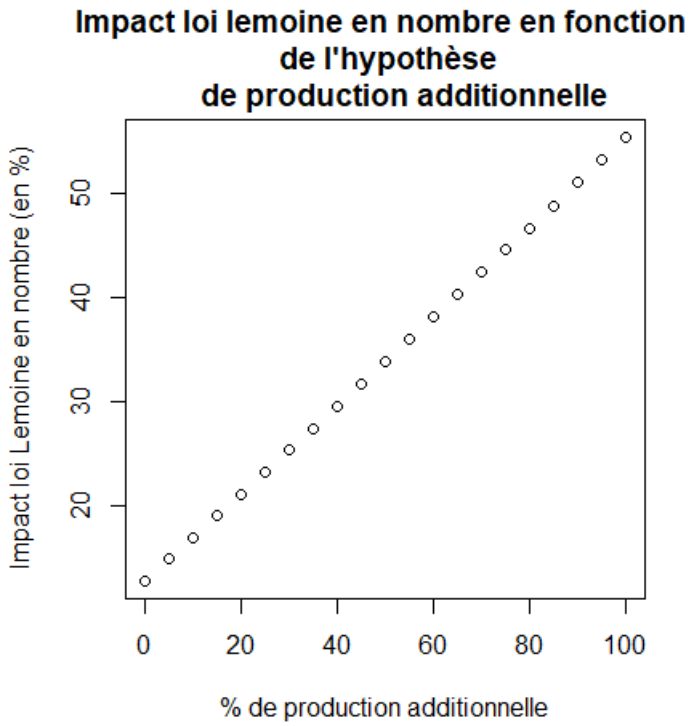


FIGURE 12.8 – Impact en nombre de la loi Lemoine en fonction de la part de production additionnelle dans la catégorie « sans suite »

L'impact varie de $12,8\%$ à $55,3\%$ selon l'hypothèse de production additionnelle prise.

Le deuxième schéma ci-dessous représente l'impact 1 an en montant en faisant varier en abscisse le pourcentage de production additionnelle au sein de la catégorie « sans suite ».

4. Type pathologie cancéreuse.

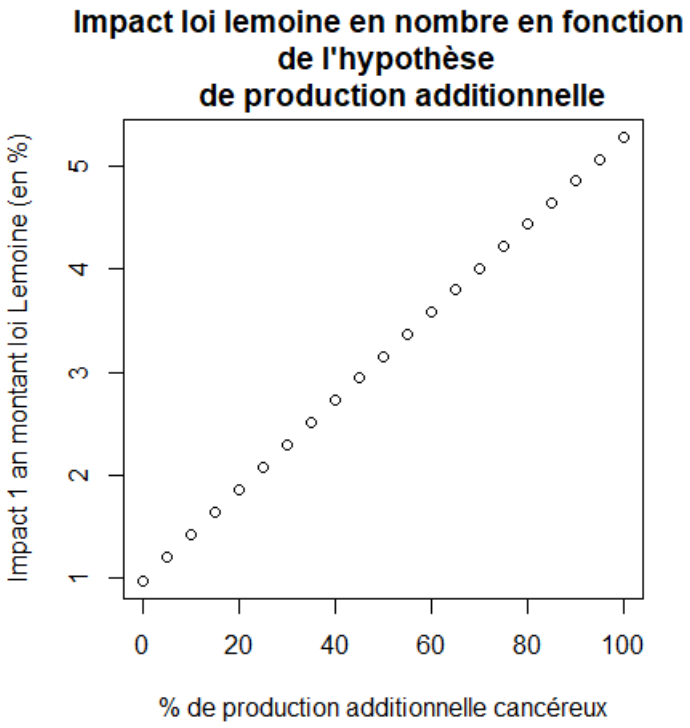


FIGURE 12.9 – Impact 1 an loi Lemoine en fonction de la part de production additionnelle au risque très aggravé

Pour l'abscisse égale à 100%, autrement dit le cas où il est considéré que la totalité de la production additionnelle revient en portefeuille avec une pathologie cancéreuse, l'ordonnée correspond bien au choc calculé précédemment : $5,27\%$.

L'impact 1 an de la loi Lemoine sur la mortalité varie donc entre $0,98\%$ et $5,27\%$ quand la part de risque très aggravé au sein de la production additionnelle varie entre 0% et 100%.

Étant donné que l'hypothèse maximale, correspondant au cas où tout le bloc « sans suite » reviendrait en portefeuille avec un cancer, peut être jugée trop forte, il convient de définir un majorant et un minorant pour cette hypothèse ainsi que de sélectionner un scénario central. Ce scénario consiste au choix d'une abscisse sur le graphique présenté ci-dessus.

12.5 Impact sur la mortalité de la loi Lemoine

Comme il a été vu précédemment, il est nécessaire d'encadrer l'impact sur la mortalité de la loi Lemoine en sélectionnant un majorant et un minorant sur l'hypothèse de proportion d'individus atteints d'un risque très aggravé au sein de la catégorie « sans suite » . Il s'agit bien ici d'hypothèses, car il est impossible de prévoir exactement la vraie valeur de ce chiffre. Le minorant et le majorant sont calibrés grâce à l'utilisation de l'open data :

Choix du minorant

Pour rappel, il a été vu sur les statistiques de [AERAS, 2020] que sur 100% de demandes d'assurances, $11,4\% * 4,1\% = 0,47\%$ d'entre elles présentent un risque aggravé. Ce chiffre peut être utilisé comme un minorant pour l'hypothèse de la proportion de production additionnelle au risque très aggravé. En effet, il ne comprend aucun effet d'aubaine et décrit simplement la « probabilité » qu'un contrat présente un risque très aggravé.

En sélectionnant l'abscisse égale à $0,47\%$ sur les deux graphiques de la partie précédente, les deux impacts minorants de la loi sont obtenus :

- L'impact en nombre, symbolisant le surplus de mortalité sur une population de *new business* post loi Lemoine, est de 13% ;
- L'impact 1 an en montant , symbolisant le surplus de charge sur une année en considérant la nouvelle mortalité du *new business*, est de 1%.

Choix du majorant

Il s'agit maintenant de supposer le pire scénario probable possible. Pour cela, la proportion maximale d'individus atteints d'une pathologie cancéreuse va être estimée de la manière suivante :

$$Prop_{majorante} = \frac{NB_{cancer} * PartM_{entreprise}}{3,6\% * N}$$

Où :

- NB_{cancer} est le nombre d'incidences de cancer en France entre 0 et 59 ans ;
- $PartM_{entreprise}$ est la part de marché de l'entreprise en assurance emprunteur ;
- $3,6\%$ est la part des dossiers « sans suite » sur 100% de demandes d'assurance ;
- N est la taille de la base de données du portefeuille emprunteur étudié.

Cette formule signifie qu'il est considéré que tous individus atteints d'une pathologie cancéreuse de France ayant moins de 60 ans, [5] souhaiteront contracter une assurance emprunteur après mise en vigueur de la loi Lemoine. La part captée de ses individus en portefeuille peut être estimée avec la part de marché de l'entreprise. Si, par exemple, 1000 individus entre 0 et 59 ans étaient atteints d'un cancer et que la part de marché de l'entreprise était de 10%, il serait supposé que l'effet d'aubaine de la loi Lemoine engendrerait 100 nouveaux contrats au risque très aggravé. Il convient ensuite de regarder ce que représente ce nombre de contrats avec pathologies cancéreuses par rapport au nombre de contrats « sans suite » en portefeuille.

5. L'incidence retenue est celle de tous les cancers avant 60 ans car la loi Lemoine ne s'applique plus pour les individus dont le prêt se termine après 60 ans. Il s'agit évidemment d'une hypothèse prudente car un individu âgé de 59 ans pourra difficilement contracter un prêt se finissant avant ses 60 ans.

C'est pour cette raison que la production additionnelle au risque aggravé est pondérée par le nombre de contrats total « sans suite » dans la formule ci-dessus.

Le chiffre NB_{cancer} est calculé à partir de [SPF, 2019], un rapport sur les incidences et mort pour cause de cancer en France, il est estimé dans ce document à 44 971. La part de marché de l'entreprise $PartM_{entreprise}$ est sélectionnée à partir de documents internes confidentiels. Le nombre de lignes N choisi est le nombre de lignes de la base du portefeuille A. [6] Finalement :

$$Prop_{majorante} = \frac{44\,971 * PartM_{entreprise}}{3,6\% * N} = 14,9\%$$

En sélectionnant l'abscisse égale à $14,9\%$ sur les deux graphiques de la partie précédente, les deux impacts majorants de la loi Lemoine sont obtenus :

- L'impact en nombre, symbolisant le surplus de mortalité sur une population de *new business* post loi Lemoine, est de 19% ;
- L'impact 1 an en montant, symbolisant le surplus de charge ultime sur une année en considérant la nouvelle mortalité du *new business*, est de $1,6\%$;

Intervalle de confiance et scénario central pour l'impact sur la mortalité de la loi Lemoine

En conclusion, il a été vu ci-dessus que la mise en vigueur de la loi Lemoine engendrerait un surplus de mortalité observée compris entre 13% et 19%. Ce chiffre peut être interprété comme un impact long terme : d'ici quelques années, quand la totalité des portefeuilles emprunteur seront constitués d'assurés *new business* post loi Lemoine, la mortalité moyenne de ces individus sera entre 13% et 19% supérieure à celle observée avant loi Lemoine.
L'impact 1 an en montant engendré par ce surplus de mortalité se situe entre 1% et $1,6\%$.

Concernant le scénario central retenu, une décision a été prise après discussions en interne avec les différentes directions et afin de présenter une cohérence globale des hypothèses loi Lemoine au niveau groupe.

Remarque : Il est nécessaire de prendre du recul par rapport à ces résultats. La construction d'une loi de mortalité sur une population non encore observée requiert de nombreuses hypothèses et il est impossible de capter tous les effets. En effet, les hypothèses prises peuvent être critiquées et l'étude de quantification menée n'a pas pris en compte la conjoncture économique et la montée des taux, ces deux facteurs pouvant fortement influencer le comportement des individus vis-à-vis des demandes de prêts. Néanmoins, tous les résultats présentés ont permis d'obtenir une idée relativement précise de l'ordre de grandeur général de l'impact sur la mortalité de la loi Lemoine.

6. Voir partie 4.2 pour la justification.

Conclusion

Cette étude s'est attachée à estimer l'impact de la loi Lemoine sur le risque de mortalité supporté par les assureurs. La problématique principale a été le développement de méthodes innovantes et leur application en finalité, pour la construction d'une loi de mortalité décrivant la sinistralité d'une population *new business* post loi Lemoine. Les différentes méthodologies proposées ont l'avantage de pouvoir être reprises indépendamment et appliquées sur d'autres données. *In fine*, le modèle construit est entièrement paramétrable et peut être utilisé à d'autres fins que l'étude de la loi Lemoine et sur n'importe quel portefeuille. Ces méthodologies sont brièvement résumées.

Une analyse poussée de la loi de l'estimateur des taux bruts de mortalité appliquée à l'assurance emprunteur a été réalisée. Cette étude pourra permettre la construction d'intervalles de confiance des taux de mortalité plus précis mais aussi l'ajustement de certains tests d'adéquation des q_x tel que celui du khi-deux.

Une méthode d'estimation du facteur λ, le nombre d'années au-delà duquel l'effet de sélection médicale n'est plus visible sur un portefeuille, a été proposée. Ce facteur étant potentiellement différent pour chaque portefeuille, la méthodologie présentée pourra s'appliquer à de nouvelles données et n'est pas propre à l'entreprise.

La méthode bayésienne empirique de Bühlmann, reprise par Klugman, a été poursuivie et modifiée pour être appliquée au cas de la mortalité en emprunteur. Concernant ce modèle de crédibilité, une méthode de sélection de portefeuilles a été proposée afin de répondre à la problématique laissée sans réponse dans l'article de Klugman. L'application de cette méthode, basée sur une analyse factorielle, a permis de limiter la prise de décisions subjectives vis-à-vis des hypothèses du modèle.

Deux algorithmes, celui des plateaux et celui des plateaux pondérés, ont été développés afin d'obtenir un critère objectif de sélection de tranches d'âges lors de la construction de table par abattements multiples. Ces algorithmes pourront permettre d'affiner la précision des taux de mortalité, même construits à partir de faibles volumes de données, que ça soit dans le cas classique du SMR ou bien dans le cas de l'application de la crédibilité de Bühlmann.

Finalement, l'ensemble des méthodes présentées ci-dessus ont été mises en application sur le cas de la loi Lemoine. Ainsi, une méthodologie basée sur des données de marché et non propres à l'entreprise a été proposée afin de quantifier l'impact sur le risque mortalité de la mise en application de la loi. La méthode présentée pourra être appliquée en utilisant les données propres à l'assureur pour la quantification d'un impact entreprise et non plus général. Concernant l'impact direct sur le surplus de mortalité du *new business*, il a été estimé qu'un portefeuille de nouveaux arrivants décéderait en moyenne 13% à 19% plus qu'un portefeuille d'assurés avant la loi Lemoine.

Il convient tout de même de prendre du recul sur ces résultats. L'impact est grandement influencé par l'hypothèse forte de présence en portefeuille d'assurés atteints de pathologies cancéreuses. En effet, la sinistralité très importante de ce sous-groupe a mécaniquement chargé la mortalité moyenne estimée du portefeuille *new business* post loi Lemoine. Les chiffres présentés reposent finalement bien sur des

hypothèses fortes, mais prudentes, et il s'agira d'étudier le comportement des populations emprunteur lorsque le volume de l'historique des données assurés emprunteur le permettra.

Bibliographie

[AERAS, 2020] AERAS (2020). Statistiques.

[Echevin, 2019] Echevin, D. (2019). *Comprendre et conseiller l'assurance emprunteur*. Argus de l'assurance.

[Espié *et al.*, 2012] Espié, M., Hamy, A., Eskenazy, S., Cuvier, C. et Giacchetti, S. (2012). Épidémiologie du cancer du sein. *EMC-Gynécologie*, 7(4):1–17.

[Gavin, 2008] Gavin, B. (2008). Selecting Mortality Tables : A Credibility Approach.

[Henderson, 1924] Henderson, R. (1924). A new method of graduation. *Transactions of the Actuarial Society of America*, vol 25:p.29–53.

[Hoem, 1969] Hoem, J. M. (1969). Markov chain models in life insurance. *Blätter der DGVFM*, 9(2):91–107.

[Klugman *et al.*, 2009] Klugman, S., Rhodes, T., Purushotham, M. et Gill, S. (2009). *Credibility Theory Practices*. Society Of Actuaries.

[Lemoine, 2022] Lemoine, P. (2022). LOI n° 2022-270 du 28 février 2022 pour un accès plus juste, plus simple et plus transparent au marché de l'assurance emprunteur (1).

[lig, 2006] lig (2006). *lignes directrices mortalité*. Commission d'agrément.

[ligue contre le cancer, 2022] ligue contre le cancer, L. (2022). Loi Lemoine : La Ligue contre le cancer salue cette avancée majeure mais reste vigilante face à de possibles dérives (Communiqué).

[Makeham, 1860] Makeham, W. (1860). on the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries*.

[Maumy-Bertrand et Bertrand, 2018] Maumy-Bertrand, M. et Bertrand, F. (2018). *Initiation à la statistique avec R-3e éd. : Cours, exemples, exercices et problèmes corrigés*. Dunod.

[Planchet *et al.*, 2022] Planchet, F., Debonneuil, É. et Péju, M. (2022). Proposal to extend access to loans for serious illnesses using open data. *Risks*, 10(3):51.

[public France, 2017] public France, S. (2017). Projection de l'incidence et de la mortalité par cancer en france métropolitaine en 2017.

[SPF, 2019] SPF (2019). Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018 - Tumeurs solides : Étude à partir des registres des cancers du réseau Francim.

Table des figures

Liste des tableaux

Annexes

Annexe A

Résultats des travaux de QDD sur les données

A.1 Résultats des travaux de QDD sur l'exposition des produits B,C,D et E

Résultats des tests de qualité des données : Produit B

Test effectué	Taux de non-conformité	Taux en exposition
Variable Sexe mal renseignée	$6,11\%$	$5,75\%$
Variable Quotité nulle	$0,02\%$	$0,02\%$
Doublons	$0,00\%$	$0,00\%$
Age d'adhésion < Age min d'adhésion	$0,01\%$	$0,01\%$
Age d'adhésion > Age max d'adhésion	$0,01\%$	$0,01\%$
Age de sortie d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,00\%$	$0,00\%$
Nom vide	$0,10\%$	$0,10\%$
Prénom vide	$6,11\%$	$5,75\%$
Age début d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Date début observation = Date fin d'observation	$0,11\%$	$0,00\%$
Age début d'observation < 18	$0,01\%$	$0,01\%$

Table A.1 – Résultats QDD produit B

Remarque : Les lignes non conformes liées aux prénoms vides concernent le cas des prénoms composés. Pour les individus ayant un prénom étant composé de plus d'un mot, leur prénom s'est concaténé avec leur nom dans la variable « Nom assuré », laissant la variable « Prénom assuré » vide.

Résultats des tests de qualité des données : Produit C

Test effectué	Taux de non-conformité	Taux en exposition
Variable Sexe mal renseignée	$0,14\%$	$0,12\%$
Variable Quotité nulle	$0,00\%$	$0,00\%$
Doublons	$0,00\%$	$0,00\%$
Age d'adhésion < Age min d'adhésion	$0,00\%$	$0,00\%$
Age d'adhésion > Age max d'adhésion	$1,35\%$	$1,22\%$
Age de sortie d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,00\%$	$0,00\%$
Nom vide	100%	100%
Prénom vide	100%	100%
Age début d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Date début observation = Date fin d'observation	$1,04\%$	$0,00\%$
Age début d'observation < 18	$0,00\%$	$0,00\%$

TABLE A.2 – Résultats QDD produit C

Remarque : Sur ce produit, les variables « Nom assuré » et « Prénom assuré » ne sont pas remplies. La variable « Nom/Prénom », qui est la concaténation des deux, sera utilisée.

Résultats des tests de qualité des données : Produit D

Test effectué	Taux de non-conformité	Taux en exposition
Variable Sexe mal renseignée	$43,0\%$	$41,8\%$
Variable Quotité nulle	$1,50\%$	$1,00\%$
Doublons	$5,80\%$	$7,00\%$
Age d'adhésion < Age min d'adhésion	$0,00\%$	$0,00\%$
Age d'adhésion > Age max d'adhésion	$0,10\%$	$0,10\%$
Age de sortie d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,00\%$	$0,00\%$
Nom vide	$0,00\%$	$0,00\%$
Prénom vide	$0,00\%$	$0,00\%$
Age début d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Date début observation = Date fin d'observation	$1,00\%$	$0,00\%$
Age début d'observation < 18	$0,00\%$	$0,00\%$

TABLE A.3 – Résultats QDD produit D

Résultats des tests de qualité des données : Produit E

Test effectué	Taux de non-conformité	Taux en exposition
Variable Sexe mal renseignée	$9,80\%$	$9,60\%$
Variable Quotité nulle	$0,60\%$	$0,50\%$
Doublons	$0,90\%$	$1,00\%$
Age d'adhésion < Age min d'adhésion	$0,00\%$	$0,00\%$
Age d'adhésion > Age max d'adhésion	$0,00\%$	$0,00\%$
Age de sortie d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,00\%$	$0,00\%$
Nom vide	$0,00\%$	$0,00\%$
Prénom vide	$0,00\%$	$0,00\%$
Age début d'observation > Age max sous garanties	$0,00\%$	$0,00\%$
Date début observation = Date fin d'observation	$1,90\%$	$0,00\%$
Age début d'observation < 18	$0,00\%$	$0,00\%$

TABLE A.4 – Résultats QDD produit E

A.2 Résultats suppressions de ligne de l'exposition des produits B,C,D et E

Suppressions de lignes : Produit B

Test	Proportion de lignes supprimées
Variable Quotité nulle	$0,02\%$
Age de sortie d'observation > Age max sous garanties + 5	$0,00\%$
Date début observation = Date fin d'observation	$0,11\%$
Age début d'observation < 18	$0,00\%$

TABLE A.5 – Suppressions des lignes : produit B

Remarque : Au global, les tests ont amené à la suppression de $0,11\%$ des lignes de la base d'exposition du produit B.

Suppressions de lignes : Produit C

Test	Proportion de lignes supprimées
Age de sortie d'observation > Age max sous garanties + 5	$0,00\%$
Date début observation = Date fin d'observation	$1,04\%$
Age début d'observation < 18	$0,00\%$

TABLE A.6 – Suppressions des lignes : produit C

Remarque : Au global, les tests ont amené à la suppression de $1,04\%$ des lignes de la base d'exposition du produit C.

Suppressions de lignes : Produit D

Test	**Proportion de lignes supprimées**
Variable Quotité nulle	$1,48\%$
Doublons	$4,50\%$
Date début observation = Date fin d'observation	$1,04\%$
Age début d'observation < 18	$0,00\%$

TABLE A.7 – Suppressions des lignes : produit D

Remarque : Au global, les tests ont amené à la suppression de $6,80\%$ des lignes de la base d'exposition du produit D.

Suppressions de lignes : Produit E

Test	**Proportion de lignes supprimées**
Variable Quotité nulle	$0,56\%$
Doublons	$0,68\%$
Date début observation = Date fin d'observation	$1,90\%$
Age début d'observation < 18	$0,00\%$

TABLE A.8 – Suppressions des lignes : produit E

Remarque : Au global, les tests ont amené à la suppression de $3,10\%$ des lignes de la base d'exposition du produit E.

A.3 Résultats de la QDD sinistre rapprochement comptable sur les produits B,C,D et E

Comparaison comptable produit B

Année de paiement / Année de survenance	2015	2016	2017	2018	2019	2020	2021	total
2015	$0,0\%$	$-1,2\%$	$3,3\%$	$0,0\%$	$0,0\%$	$0,0\%$	$0,9\%$	$-0,8\%$
2016	$-3,2\%$	$6,2\%$	$-25,6\%$	$0,0\%$	$0,0\%$	$0,0\%$	$0,0\%$	$0,2\%$
2017	$0,0\%$	$0,0\%$	$-0,6\%$	$-0,1\%$	$0,4\%$	$0,0\%$	$0,0\%$	$-0,3\%$
2018	$0,0\%$	$0,0\%$	$0,0\%$	$-0,4\%$	$-6,0\%$	$7,2\%$	$0,0\%$	$-1,7\%$
								$-0,7\%$

TABLE A.9 – Différences relatives entre les montants réglés en comptabilité et dans la base sinistres : produit B

La différence relative entre les montants réglés enregistrés en comptabilité et ceux présents dans la base sinistres du produit B est de $-0,7\%$ sur la totalité de la période d'observation (2015-2018). Cet écart est jugé acceptable.

Il est à noter que certaines différences relatives peuvent sembler importantes, mais en réalité représentent de faibles écarts en montant.

Comparaison comptable produit C

Année de survenance \ Année de paiement	2015	2016	2017	2018	2019	2020	2021	total
2015	0,0%	0,3%	0,0%	0,0%	−4,7%	0,0%	100,0%	−0,2%
2016	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2017	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2018	0,0%	0,0%	0,0%	−1,0%	−0,2%	0,0%	20,6%	−0,3%
								−1,9%

TABLE A.10 – Différences relatives entre les montants réglés en comptabilité et dans la base sinistres : produit C

La différence relative entre les montants réglés enregistrés en comptabilité et ceux présents dans la base sinistres du produit C est de $-1,9\%$ sur la totalité de la période d'observation (2015-2018). Cet écart est jugé acceptable.

Comparaison comptable produit D

Année de survenance \ Année de paiement	2015	2016	2017	2018	2019	2020	total
2015	46,6%	0,0%	0,0%	0,0%	0,0%	0,0%	12,6%
2016	0,0%	67,0%	−93,9%	0,0%	0,0%	0,0%	−28,40%
2017	0,0%	0,0%	11,4%	0,0%	0,0%	0,0%	4,3%
2018	0,0%	0,0%	0,0%	68,2%	−14,7%	0,0%	43,5%
							0,0%

TABLE A.11 – Différences relatives entre les montants réglés en comptabilité et dans la base sinistres : produit D

La différence relative entre les montants réglés enregistrés en comptabilité et ceux présents dans la base sinistres du produit D est de $0,0\%$ sur la totalité de la période d'observation (2015-2018). Cet écart est jugé acceptable.

Comparaison comptable produit E

Année de survenance \ Année de paiement	2015	2016	2017	2018	2019	2020	total
2015	24,0%	−67,8%	0,0%	0,0%	0,0%	0,0%	2,3%
2016	0,0%	15,1%	−447,5%	0,0%	0,0%	0,0%	−28,7%
2017	0,0%	0,0%	38,3%	0,3%	0,0%	0,0%	2,0%
2018	0,0%	0,0%	0,0%	33,1%	0,0%	0,0%	26,8%
							0,0%

TABLE A.12 – Différences relatives entre les montants réglés en comptabilité et dans la base sinistres : produit E

La différence relative entre les montants réglés enregistrés en comptabilité et ceux présents dans la base sinistres du produit E est de 0,0% sur la totalité de la période d'observation (2015-2018). Cet écart est jugé acceptable.

Il est à noté que certaines différences relatives peuvent sembler importantes, mais en réalité représentent de faibles écarts en montant.
De plus, une mauvaise ventilation des montants réglés en comptabilité est observée. Le problème a fait l'objet de retraitement en interne dans les services comptable depuis 2018.

Annexe B

Méthodes classiques d'estimation de la mortalité

B.1 Lissage de Whittaker-Henderson

Le lissage de [Henderson, 1924] est un modèle qui permet le passage de taux bruts à des taux lissés. Ce modèle s'attache à trouver un compromis entre deux contraintes : la fidélité aux taux bruts et la régularité. Pour ce faire,les deux critères suivants sont minimisés :

- **Le critère de fidélité :**

  La courbe de taux $c : x \to c_x$ est d'autant plus proche des taux bruts $\widehat{q}_x$ que la fonction suivante est faible :

  $$\mathcal{F}(c) = \sum_{x=x_{min}}^{x_{max}} w_x (c_x - \widehat{q}_x)^2$$

  Avec :

  - w_x une suite de poids positifs attribués à chaque observation ;
  - x_{min} et x_{max} les âges minimum et maximum sur lesquels les taux bruts sont calculés.

  La fonction $\mathcal{F}$ représente une distance entre les taux bruts et lisses, pondérée par les poids w_x.

  Pour résoudre le problème de minimisation plus aisément, le critère se réécrit sous forme matricielle :

  $$\mathcal{F}(c) = (c - \widehat{q})^T W (c - \widehat{q})$$

  Avec :

  $$c = \begin{pmatrix} c_{x_{min}} \\ ... \\ c_{x_{max}} \end{pmatrix} \qquad \widehat{q} = \begin{pmatrix} \widehat{q}_{x_{min}} \\ ... \\ \widehat{q}_{x_{max}} \end{pmatrix}$$

  et la matrice W, diagonale de dimension $x_{max} + 1 \times x_{max} + 1$.

  $$W = \begin{pmatrix} w_{x_{min}} & 0 & ... & ... & 0 \\ 0 & w_{x_{min}+1} & 0 & ... & ... \\ ... & ... & ... & ... & 0 \\ 0 & ... & ... & 0 & w_{x_{max}} \end{pmatrix}$$

  Le choix de la matrice des poids W repose sur l'approximation présentée précédemment :

$$\widehat{q}_x \sim \mathcal{N}(q_x, \frac{q_x(1-q_x)}{n_x})$$

Ainsi, plus le nombre d'individus n_x présents à l'âge x est grand, plus la variance de l'estimateur des taux bruts est faible, *i.e.* plus l'estimation est précise. Or, il est souhaitable d'accorder plus de poids aux individus ayant une meilleure précision, d'où le choix des $w_x = n_x$.

- **Le critère de régularité :**

Soit l'opérateur linéaire $\Delta : c_x \to c_{x+1} - c_x$. Il est possible de le définir par récurrence à des ordres supérieurs : $\Delta^z c = \Delta(\Delta^{z-1} c)$ avec z un entier naturel.

Soit z fixé, la courbe de taux $c : x \to c_x$ est d'autant plus régulière à l'ordre z que le critère suivant est faible :

$$\mathcal{S}(c) = \sum_{x=x_{min}}^{x_{max}-z} [(\Delta^z c)_x]^2$$

Habituellement, le lissage est réalisé pour une valeur de z égale à 2, si bien que le critère devient :

$$\mathcal{S}(c) = \sum_{x=x_{min}}^{x_{max}-1} (c_{x+2} - 2c_{x+1} + c_x)^2$$

Le critère de régularité s'exprime sous forme matricielle :

$$\mathcal{S}(c) = c^T K^T K c$$

Avec c défini précédemment et K la matrice telle que $\Delta^z c = K \times c$.

- **Critère final de Whittaker-Henderson :**

En combinant les critères de régularité et de fidélité, il vient :

$$WH_h(c) = F(c) + hS(c)$$

Le paramètre h indique l'importance accordée à la régularité du modèle. Plus h est grand, moins les taux lisses seront proches des taux bruts, mais plus ils seront réguliers.

La solution du problème de minimisation s'écrit :

$$q^l = (W + hK^T K)^{-1} W\widehat{q}$$

Avec q^l le vecteur des taux lissés.

B.2 Lissage paramétrique par maximum de la pseudo-vraisemblance

Dans le cadre de l'optimisation des paramètres dans un modèle de lissage, la méthode la plus utilisée est celle du maximum de la pseudo-vraisemblance.
Par hypothèse :

$$T \sim \mathbb{P}_{\theta_0} \; o\grave{u} \; \theta_0 \subset \mathbb{R}^k$$

Où T est une variable aléatoire de durée de vie.

Les taux de décès sont donc également fonction du jeu de paramètres θ_0 :

$$q_x = q(x, \theta_0)$$

De plus, pour rappel, l'estimateur de Hoem du taux de décès dans le cas classique suit asymptotiquement une loi normale :

$$\widehat{q_x} \sim \mathcal{N}\left(q_x, \frac{q_x(1-q_x)}{n_x}\right)$$

Les paramètres de la loi asymptotique de l'estimation de q_x sont donc eux-mêmes fonction du jeu de paramètre θ_0 :

$$\widehat{q_x} \sim \mathcal{N}\left(q(x,\theta_0), \frac{q(x,\theta_0)(1-q(x,\theta_0))}{n_x}\right)$$

En supposant des réalisations de $\widehat{q_x}$ observées pour tous les âges : $\widehat{q}^{obs}_{x=x_{min}}, ..., \widehat{q}^{obs}_{x=x_{max}}$.

Alors, pour un jeu de paramètre θ donné, la vraisemblance associée est donnée par :

$$\mathcal{L}(\theta) = \prod_{x=x_{min}}^{x_{max}} f_\theta(\widehat{q_x})$$

où f_θ est la densité de l'estimateur $\widehat{q_x}$, i.e :

$$f_\theta(\widehat{q_x}) = \frac{1}{\sqrt{2\pi}\sigma(x,\theta)} exp\left(-\frac{1}{2}\left(\frac{\widehat{q_x} - q(x,\theta)}{\sigma(x,\theta)}\right)^2\right)$$

avec

$$\sigma(x,\theta) = \frac{q(x,\theta_0)(1-q(x,\theta_0))}{n_x}$$

La vraisemblance s'écrit donc sous la forme suivante :

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{x=x_{min}}^{x_{max}} f_\theta(\widehat{q_x}) \\ &= \prod_{x=x_{min}}^{x_{max}} \frac{1}{\sqrt{2\pi}\sigma(x,\theta)} exp\left(-\frac{1}{2}\left(\frac{\widehat{q_x} - q(x,\theta)}{\sigma(x,\theta)}\right)^2\right)\end{aligned}$$

Étant donné que les sommes sont plus faciles à manipuler pour la dérivation, la log-vraisemblance est plus souvent utilisée. Dans le cas présent, sa forme est la suivante :

$$log(\mathcal{L}(\theta)) = \sum_{x=x_{min}}^{x_{max}} -log(\sqrt{2\pi}) - log(\sigma(x,\theta)) - \frac{1}{2}\left(\frac{\widehat{q_x} - q(x,\theta)}{\sigma(x,\theta)}\right)^2$$

Par simplification, l'écart type $\sigma(x,\theta)$ est remplacé par son estimateur.

Finalement, maximiser la vraisemblance $\mathcal{L}(\theta)$ revient à maximiser la pseudo-vraisemblance suivante :

$$L(\theta) = -\sum_{x=x_{min}}^{x_{max}} \frac{n_x}{\widehat{q_x}(1-\widehat{q_x})}\left(\widehat{q_x} - q(x,\theta)\right)^2$$

Cependant, selon l'algorithme d'optimisation utilisé pour maximiser cette pseudo-vraisemblance, il conviendra de sélectionner un jeu de paramètre initial adapté pour assurer la bonne convergence de l'algorithme.

Remarque : pseudo-vraisemblance adaptée aux données emprunteur

Comme décrit dans le chapitre 6, la dépendance dans les données en emprunteur modifie la forme des paramètres de l'approximation normale de l'estimateur de Hoem. Il sera montré que, sous présence de dépendance dans les données, l'estimateur de q_x suit asymptotiquement la loi suivante :

$$Loi(\widehat{q_x}) \underset{n_x \to +\infty}{\to} \mathcal{N}(q_x, \frac{q_x}{n_x^2}\sum_{p=1}^{p_{max}} p^2 E_{x,p})$$

où

- $E_{x,p}$ est la somme des expositions à l'âge x des individus ayant exactement p prêts ;
- p_{max} est le nombre maximal de prêts qu'un individu possède dans les données.

Ainsi, la pseudo-vraisemblance adaptée à la maille emprunteur devient :

$$L^{emprunteur}(\theta) = -\sum_{x=x_{min}}^{x_{max}} \frac{n_x^2}{\widehat{q_x}\sum_{p=1}^{p_{max}} p^2 E_{x,p}}\left(\widehat{q_x} - q(x,\theta)\right)^2$$

B.3 Le modèle de Gompertz-Makeham

B.3.1 Description du modèle et des paramètres

Le modèle de [Makeham, 1860] est un modèle paramétrique. Le modèle décrit les taux instantanés de mortalité sous la forme paramétrique suivante :

$$\mu_x = a + bc^x$$

Il est supposé qu'il existe seulement deux causes de mortalité : le décès dû à la vieillesse et le décès accidentel. Chacun de ces deux risques est représenté dans le modèle de Gompertz-Makeham :

- Le terme bc^x représente le risque du décès dû à l'âge, qui est bien un risque croissant et exponentiel avec le temps ;
- Le terme a représente la mortalité accidentelle, qui est par hypothèse constante avec le temps.

Concernant l'interprétation des paramètres du modèle :

- a est le niveau de référence de la fonction de hasard μ ;
- c représente la vitesse de croissance du risque de mortalité dû à l'âge ;
- b représente l'importance du risque de décès dû à l'âge par rapport au risque de décès global.

Ce modèle sera utilisé dans la suite pour lisser les taux bruts obtenus par la méthode de Hoem.

B.3.2 Optimisation des paramètres

La calibration des paramètres du modèle se fait en maximisant la pseudo-vraisemblance énoncée plus haut :

$$L(\theta) = -\sum_{x=x_{min}}^{x_{max}} \frac{n_x}{\widehat{q_x}(1-\widehat{q_x})} \left(\widehat{q_x} - q(x,\theta)\right)^2$$

Pour maximiser cette fonction, il convient de sélectionner un jeu de paramètres initiaux pertinent. Pour cela, il faut commencer par décrire les taux de mortalité dans le modèle :

$$q_x(\theta) = 1 - sg^{c^x(c-1)}$$

où

- $\theta = (a, b, c)$;
- $s = e^{-a}$;
- $g = exp\left(\frac{b}{ln(c)}\right)$;
- $c = c$.

La méthodologie de recherche du vecteur $(\widehat{s}_0, \widehat{g}_0, \widehat{c}_0)$ pour initialiser le problème d'optimisation est maintenant présentée.

Étant donné l'approximation $ln(1-q_x) \approx -q_x$ pour q_x petit :

$$\begin{aligned} ln(1-q_x) &= ln(1-(1-sg^{c^x(c-1)})) \\ &= ln(sg^{c^x(c-1)}) \\ &= ln(s) + ln(g)c^x(c-1) \\ &\approx -q_x \end{aligned}$$

Par le même raisonnement, pour q_x petit :

$$\begin{aligned} ln(1-q_{x+1}) &= ln(s) + ln(g)c^{x+1}(c-1) \\ &\approx -q_{x+1} \end{aligned}$$

D'où l'approximation, pour q_x petit :

$$\begin{aligned} q_x - q_{x+1} &\approx ln(g)c^{x+1}(c-1) - ln(g)c^x(c-1) \\ &= ln(g)c^x(c-1)(c-1) \\ &= ln(g)c^x(c-1)^2 \end{aligned}$$

En passant au log, l'approximation devient, pour q_x petit :

$$ln(|q_x - q_{x+1}|) \approx ln(ln(g)) + xln(c) + 2ln(c-1)$$

Cette étape fournit une méthode pour tester l'adéquation du modèle. En effet, d'après l'approximation, les points $(x, ln(|q_x - q_{x+1}|))_{x=x_{min},\ldots,x_{max}}$ sont supposés être alignés sur une droite de pente $ln(c)$ et d'ordonnée à l'origine $ln(ln(g)) + 2ln(c-1)$ si le modèle est adéquat.

Si c'est le cas, une régression linéaire simple est faite de la variable $ln(|q_x - q_{x+1}|)$ sur x. Le coefficient directeur de la droite et l'ordonnée à l'origine peuvent être approximés comme tels :

$$xln(c) + ln(ln(g)) + 2ln(c-1) = \widehat{\alpha} * x + \widehat{\beta}$$

Sont déduits les coefficients de l'approximation exprimée plus haut :

$$\widehat{\alpha} \approx ln(c)$$

et

$$\widehat{\beta} \approx ln(ln(g)) + 2ln(c-1)$$

Finalement, le jeu de paramètres $(\widehat{s}_0, \widehat{g}_0, \widehat{c}_0)$ pour initialiser l'algorithme est le suivant :

$$\begin{cases} \widehat{c}_0 = e^{\widehat{\alpha}} \\ \widehat{g}_0 = exp(exp(\widehat{\beta} - 2ln(e^{\widehat{\alpha}} - 1))) \\ \widehat{s}_0 = exp(-q_x - \widehat{c}_0^x(\widehat{c}_0 - 1)|ln(\widehat{g}_0)|) \end{cases}$$

Cette dernière égalité provient de l'approximation

$$ln(s) \approx -q_x - c^x(c-1)|ln(g)|$$

exprimée plus haut.

B.4 Fermeture par régression logit

Le modèle de Gompertz-Makeham présenté précédemment est efficace quand il s'agit de lisser des taux sur les âges où un fort volume de données est observé. Cependant, il présente certaines limites aux âges extrêmes. En effet, le modèle étant un lissage paramétrique, le peu de données mis à disposition sur ces plages extrêmes entraîne un biais dans l'estimation des paramètres. De plus, le modèle présuppose une allure exponentielle de la courbe, ce qui a tendance à surestimer les taux de mortalité réels dans la pratique.

Pour pallier ces limites, utiliser un modèle de fermeture de table semble être un choix indiqué. Le modèle présenté ci-dessous est la fermeture par régression logit. Ce dernier suppose l'existence d'une relation linéaire entre le logit du q_x à estimer et le logit du taux de mortalité d'une table de référence :

$$logit(q_x) = a * logit(q_x^{ref}) + b$$

Pour rappel, la fonction logit s'exprime comme telle :

$$logit(x) = ln\left(\frac{x}{1-x}\right)$$

Après avoir réalisé la régression linéaire des q_x sur les q_x^{ref} sur la plage d'âges où la table a été construite par des méthodes classiques d'estimation de la mortalité, il suffit d'inverser la fonction logit pour trouver la forme du q_x donnée par le modèle :

$$\begin{aligned}\widehat{q_x} &= logit^{-1}\left(\widehat{a} * logit(q_x^{ref}) + \widehat{b}\right) \\ &= \frac{e^b * (\frac{q_x^{ref}}{1-q_x^{ref}})^a}{1 + e^b * (\frac{q_x^{ref}}{1-q_x^{ref}})^a}\end{aligned}$$

Avec

$$logit^{-1}(y) = \frac{e^y}{1-e^y}$$

Dans le cadre de cette étude, la table de référence qui sera choisie sera une pondération des tables réglementaires homme et femme par les *sexe ratios* observés du portefeuille. Les *sexe ratios* sont les proportions d'hommes et de femmes observées sur les données.

Finalement, il est à noter que la régression linéaire peut ne pas être faite sur l'ensemble des âges sur lesquels la table a déjà été construite. Il est possible de se restreindre aux k derniers q_x construits de la table si par exemple l'objectif est la fermeture de la table aux âges élevés.

B.5 Le lissage géométrique

Le lissage géométrique est similaire au lissage par moyenne mobile. Sa formule est la suivante :

$$q_x = \sqrt[5]{q_{x-2} * q_{x-1} * q_x * q_{x+1} * q_{x+2}}$$

Le cas ci-dessus est le cas particulier du lissage géométrique « centré sur 5 observations ». C'est aussi le cas qui sera appliqué pour cette étude.

Dans la pratique, l'effet de ce lissage est marginal et il n'est utilisé que pour le lissage des âges où se produit un raccordement entre les taux de mortalité estimés de manière classique et les taux estimés par la fermeture de table.

Annexe C

La problématique de la maille prêt

C.1 Démonstration de la condition d'application du théorème de Lindeberg-Feller

Il va être montré que :

$$\forall \epsilon > 0, \ \lim_{n \to +\infty} \frac{\sum_{i=1}^{n} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 \mathbf{1}_{|X_i - \mathbb{E}(X_i)| \geq \epsilon \sigma(S_n)}}{\mathbb{V}(S_n)}$$

Avec $X_i \sim Ber(e_i q_x)$.

Remarquons premièrement que la quantité $|X_i - \mathbb{E}(X_i)|$ est bornée à 1, les variables X_i étant des variables de Bernoulli.

De plus, remarquons que $\sigma(S_n)$ diverge en l'infini :

$$\begin{aligned}
\sigma(S_n) &= \sqrt{\sum_{i=1}^{n} \mathbb{V}(X_i)} \\
&= \sqrt{\sum_{i=1}^{n} q_x e_i (1 - q_x e_i)} \\
&\underset{n \to +\infty}{\to} +\infty
\end{aligned}$$

Ainsi, $\forall \epsilon > 0 \ \exists N_0$ tq $\forall n \geq N_0 \ \epsilon \sigma(S_n) > 1$

Soit $\epsilon > 0$, il vient donc :

$$\begin{aligned}
&\lim_{n \to +\infty} \frac{\sum_{i=1}^{n} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 \mathbf{1}_{|X_i - \mathbb{E}(X_i)| \geq \epsilon \sigma(S_n)}}{\mathbb{V}(S_n)} = \\
&\lim_{n \to +\infty} \left(\frac{\sum_{i=1}^{N_0} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 \mathbf{1}_{|X_i - \mathbb{E}(X_i)| \geq \epsilon \sigma(S_n)}}{\mathbb{V}(S_n)} + \frac{\sum_{i=N_0+1}^{n} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 \mathbf{1}_{|X_i - \mathbb{E}(X_i)| \geq \epsilon \sigma(S_n)}}{\mathbb{V}(S_n)} \right) \\
&= \lim_{n \to +\infty} \frac{\sum_{i=1}^{N_0} \mathbb{E}(X_i - \mathbb{E}(X_i))^2 \mathbf{1}_{|X_i - \mathbb{E}(X_i)| \geq \epsilon \sigma(S_n)}}{\mathbb{V}(S_n)} \leq \lim_{n \to +\infty} \frac{N_0}{\mathbb{V}(S_n)} \\
&= 0
\end{aligned}$$

car $\mathbb{V}(S_n) \underset{n \to +\infty}{\to} +\infty$.

On conclut la démonstration en remarquant que les deux conditions de $\mathbb{E}(X_i^2)$ borné et que nous ne sommes pas dans le cas dégénéré $\mathbb{V}(S_n)$ sont trivialement satisfaites avec des variables de Bernoulli.

Annexe D

Effet de sélection médicale

D.1 Évolution de la p-valeur pour la détermination de λ

Les graphiques ci-dessous représentent l'évolution de la p-valeur en fonction du nombre de simulations pour chacun des 3 tests *bootstrap* effectués pour la détermination du facteur de sélection médicale λ.

Évolution de la première p-valeur : $\mathcal{H}_0 : \mathbb{E}[{}^0\widehat{q}] = \mathbb{E}[{}^1\widehat{q}]$

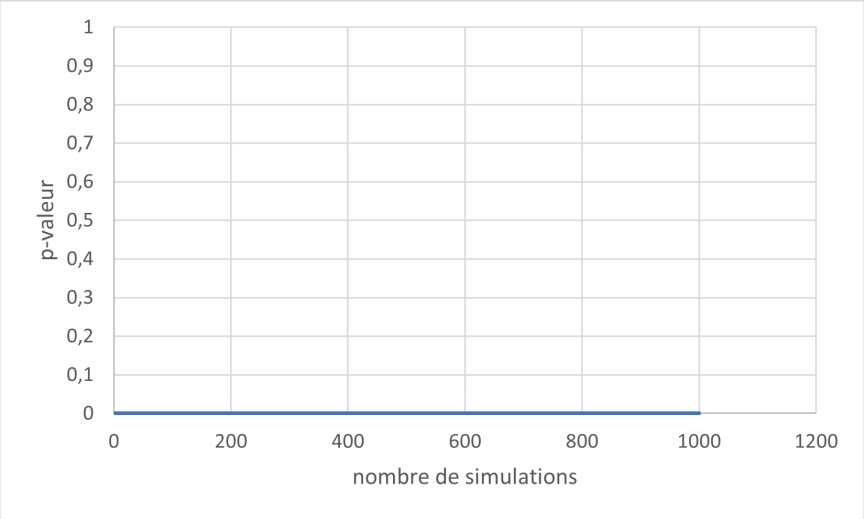


FIGURE D.1 – Évolution de la p-valeur : premier test *bootstrap* pour l'estimation de λ

Évolution de la deuxième p-valeur : $\mathcal{H}_0 : \mathbb{E}[^1\widehat{q}] = \mathbb{E}[^2\widehat{q}]$

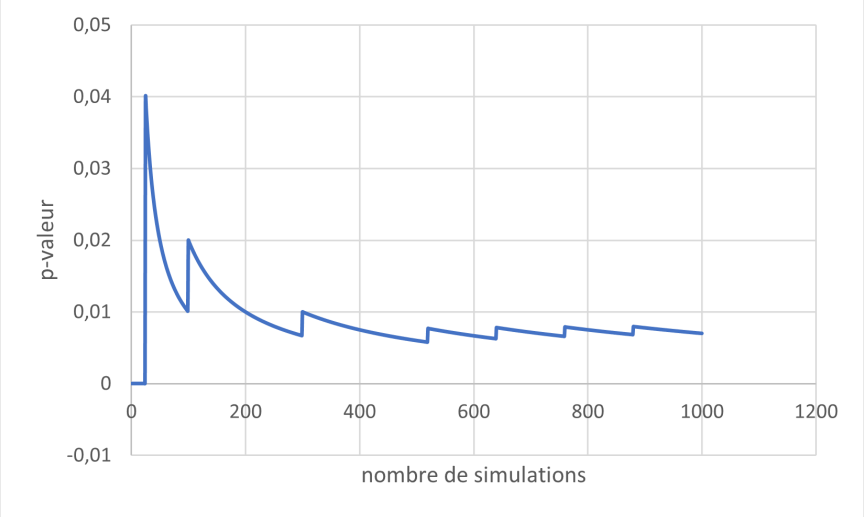


FIGURE D.2 – Évolution de la p-valeur : deuxième test *bootstrap* pour l'estimation de λ

Évolution de la troisième p-valeur : $\mathcal{H}_0 : \mathbb{E}[^2\widehat{q}] = \mathbb{E}[^3\widehat{q}]$

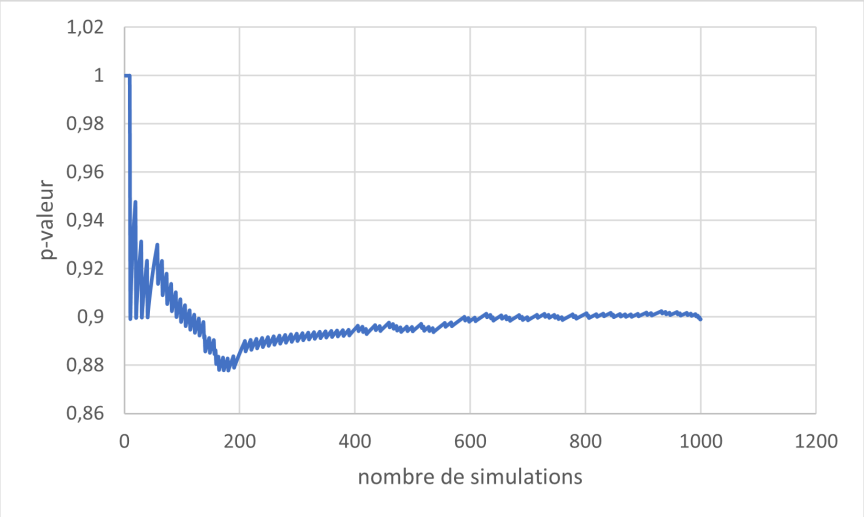


FIGURE D.3 – Évolution de la p-valeur : troisième test *bootstrap* pour l'estimation de λ

Annexe E

Théorie de la crédibilité

E.1 Théorie de la crédibilité classique

E.1.1 Crédibilité complète

Construction classique

Les détails de la théorie de la crédibilité complète présentés ici ont été développés par [Gavin, 2008].

L'objectif est d'estimer une probabilité de décès à l'âge x, notée q_x. Soit $\widehat{q_x}$ l'estimateur de q_x, determiné par des méthodes actuarielles. L'estimateur $\widehat{q_x}$ est crédible si la relation suivante est satisfaite :

$$\mathbb{P}(\frac{|\widehat{q_x} - q_x|}{q_x} \leq k) \geq p$$

Avec k et p des paramètres posés arbitrairement et avec k petit et p grand (proche de 1).

Cette relation s'interprète comme tel : $\widehat{q_x}$ est un estimateur crédible si la probabilité que sa différence relative avec q_x soit suffisamment faible est assez grande.

Sous certaines hypothèses sur l'estimateur $\widehat{q_x}$, cette relation peut être développée pour obtenir un critère sur la quantité de données nécessaire pour obtenir un estimateur crédible. Nous supposerons la normalité de $\widehat{q_x}$:

$$\widehat{q_x} \sim \mathcal{N}(q_x, \sigma_x^2)$$

Notons que dans le cas où $\widehat{q_x}$ est l'estimateur de Hoem, cette condition est asymptotiquement satisfaite.
Développons à présent la relation (E.1).

$$\begin{aligned}
\mathbb{P}(\frac{|\widehat{q_x} - q_x|}{q_x} \leq k) &= \mathbb{P}(|q_x - \widehat{q_x}| \leq kq_x) \\
&= \mathbb{P}(-kq_x \leq \widehat{q_x} - q_x \leq kq_x) \\
&= \mathbb{P}(-\frac{kq_x}{\sigma_x} \leq \frac{\widehat{q_x} - q_x}{\sigma_x} \leq \frac{kq_x}{\sigma_x}) \\
&= \mathbb{P}(-\frac{kq_x}{\sigma_x} \leq W \leq \frac{kq_x}{\sigma_x})
\end{aligned}$$

avec $W \sim \mathcal{N}(0,1)$.
En nommant $\phi(\cdot)$ la fonction de répartition d'une loi normale centrée réduite, alors :

$$\begin{aligned}
&\mathbb{P}(\frac{|\widehat{q_x} - q_x|}{q_x} \leq k) \geq p \\
\Leftrightarrow\ & 1 - 2\phi(-\frac{kq_x}{\sigma_x}) \geq p \\
\Leftrightarrow\ & \phi(-\frac{kq_x}{\sigma_x}) \leq \frac{1-p}{2} \\
\Leftrightarrow\ & -\frac{kq_x}{\sigma_x} \leq z_{\frac{1-p}{2}}
\end{aligned}$$

Avec $\phi^{-1}(\alpha) = z_\alpha$, la fonction quantile de la loi normale centrée réduite.
Dans le cas classique ($\widehat{q_x}$ est une somme de variables aléatoires $i.i.d$) :

$$\sigma_x^2 = \frac{q_x(1-q_x)}{n_x}$$

L'équation (E.1) devient alors :

$$\begin{aligned}
&-\sqrt{n_x}\frac{kq_x}{\sqrt{q_x(1-q_x)}} \leq z_{\frac{1-p}{2}} \\
\Leftrightarrow\ & \sqrt{n_x}\frac{kq_x}{\sqrt{q_x(1-q_x)}} \geq z_{1-\frac{1-p}{2}} \\
\Leftrightarrow\ & \sqrt{q_x n_x} \geq \frac{z_{1-\frac{1-p}{2}}\sqrt{1-q_x}}{k} \\
\Leftrightarrow\ & q_x n_x \geq (\frac{z_{1-\frac{1-p}{2}}\sqrt{1-q_x}}{k})^2
\end{aligned}$$

Or, l'approximation $q_x \sim 0$ peut être faite. Dès lors, $1 - q_x \sim 1$, l'inéquation devient :

$$q_x n_x \geq (\frac{z_{1-\frac{1-p}{2}}}{k})^2$$

Dans la suite, la quantité $(\frac{z_{1-\frac{1-p}{2}}}{k})^2$ sera nommée $\lambda_{(p,k)}$.

Finalement un critère sur le nombre de sinistres minimum nécessaire afin d'obtenir un estimateur de q_x crédible est obtenu.

Ce modèle présente pourtant ses limites. En effet, si, par exemple, p est fixé à $0,9$ et k est fixé à $0,05$ (choix classique dans la littérature actuarielle), alors le nombre de sinistres nécessaire pour obtenir un estimateur dit crédible pour un âge donné est de 1082. Dans un cas pratique, il est très rare d'avoir autant de données par âge, il convient alors de chercher une méthode alternative pour attester de la confiance de notre estimateur. La construction par abattement est un outil pour répondre à cette problématique.

Construction par abattement

Il sera supposé que la probabilité de décès q_x que l'on cherche à estimer peut s'écrire sous la forme $q_x = f q_x^A$.
Où :
-$f \in]0; +\infty[$ est le coefficient d'abattement.
-q_x^A est la table que l'on cherchera à abattre.

Un estimateur naturel pour f est le ratio des décès observés sur attendus, le SMR :

$$\widehat{f} = \frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} d_i^x}{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A}$$

En se rappelant que d_i^x suit par hypothèse une loi de Bernoulli de paramètre q_x, $\widehat{f}$ est bien un estimateur sans biais de f :

$$\begin{aligned}
\mathbb{E}(\widehat{f}) &= \mathbb{E}(\frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} d_i^x}{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A}) \\
&= \frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} \mathbb{E}(d_i^x)}{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A} \\
&= \frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x}{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A} \\
&= \frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} f q_x^A}{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A} \\
&= f
\end{aligned}$$

Pour obtenir un estimateur $\widehat{f}q_x^A$ crédible. Il est donc nécessaire que, pour k et p fixés :

$$\begin{aligned}
&\forall x \in (x_{min}, .., x_{max})\ \mathbb{P}(\frac{|\widehat{q_x} - q_x|}{q_x} \leq k) \geq p \\
\Leftrightarrow\ &\forall x \in (x_{min}, .., x_{max})\ \mathbb{P}(\frac{|\widehat{f}q_x^A - f q_x^A|}{f q_x^A} \leq k) \geq p \\
\Leftrightarrow\ &\mathbb{P}(\frac{|\widehat{f} - f|}{f} \leq k) \geq p
\end{aligned}$$

Or $\widehat{f}$ est par hypothèse une somme de variables aléatoires *i.i.d*, l'expression trouvée peut alors être développée de la même manière que dans la partie précédente, par l'approximation normale :

$$\begin{aligned}
&\mathbb{P}(\frac{|\widehat{f} - f|}{f} \leq k) \geq p \\
\Leftrightarrow\ &\mathbb{P}(-kf \leq \widehat{f} - f \leq kf) \geq p \\
\Leftrightarrow\ &\mathbb{P}(-\frac{kf}{\sigma_{\widehat{f}}} \leq \frac{\widehat{f} - f}{\sigma_f} \leq \frac{kf}{\sigma_{\widehat{f}}}) \geq p \\
\Leftrightarrow\ &\mathbb{P}(-\frac{kf}{\sigma_{\widehat{f}}} \leq W \leq \frac{kf}{\sigma_{\widehat{f}}}) \geq p \\
\Leftrightarrow\ &1 - 2\phi(-\frac{kf}{\sigma_{\widehat{f}}}) \geq p \\
\Leftrightarrow\ &\phi(-\frac{kf}{\sigma_{\widehat{f}}}) \leq \frac{1-p}{2} \\
\Leftrightarrow\ &-\frac{kf}{\sigma_{\widehat{f}}} \leq z_{\frac{1-p}{2}} \\
\Leftrightarrow\ &\frac{kf}{\sigma_{\widehat{f}}} \geq z_{1-\frac{1-p}{2}} \\
\Leftrightarrow\ &\frac{f^2}{\sigma_{\widehat{f}}^2} \geq (\frac{z_{1-\frac{1-p}{2}}}{k})^2 = \lambda_{(p,k)}
\end{aligned}$$

A ce stade, il convient de calculer la valeur de la variance de $\widehat{f}$:

$$\begin{aligned} Var(\widehat{f}) &= Var(\frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} d_i^x}{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A}) \\ &= \frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} Var(d_i^x)}{(\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A)^2} \\ &= \frac{\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x(1-q_x)}{(\sum_{x=x_{min}}^{x_{max}} \sum_{i=1}^{n_x} q_x^A)^2} \\ &= \frac{\sum_{x=x_{min}}^{x_{max}} n_x q_x(1-q_x)}{(\sum_{x=x_{min}}^{x_{max}} n_x q_x^A)^2} \end{aligned}$$

Le critère de crédibilité devient alors :

$$\begin{aligned} &\frac{f^2}{\frac{\sum_{x=x_{min}}^{x_{max}} n_x q_x(1-q_x)}{(\sum_{x=x_{min}}^{x_{max}} n_x q_x^A)^2}} \geq \lambda_{(p,k)} \\ &\Leftrightarrow (\sum_{x=x_{min}}^{x_{max}} n_x f q_x^A)^2 \geq \lambda_{(p,k)} \sum_{x=x_{min}}^{x_{max}} n_x q_x(1-q_x) \end{aligned}$$

Avec l'approximation $1 - q_x \sim 1$ et en se rappelant que $q_x = f q_x^A$, l'inégalité devient :

$$\begin{aligned} &(\sum_{x=x_{min}}^{x_{max}} n_x q_x)^2 \geq \lambda_{(p,k)} \sum_{x=x_{min}}^{x_{max}} n_x q_x \\ &\Leftrightarrow \sum_{x=x_{min}}^{x_{max}} n_x q_x \geq \lambda_{(p,k)} \end{aligned}$$

Ainsi, si la table est construite par abattement, le critère du nombre de décès minimum par âge peut se généraliser pour tout âge. Une table de mortalité construite par abattement sera donc dite crédible si elle est construite sur des données comptabilisant au minimum $\lambda_{(p,k)}$ sinistres.

Toutefois, pour $p = 0,9$ et $k = 0,05$, ce critère impose un nombre de sinistres total de 1082. Il peut arriver dans la pratique que les données disponibles soient insuffisantes. La théorie de la crédibilité présentée dans le chapitre « Méthode bayésienne empirique de Bühlmann » peut être appliquée.

E.2 Démonstration de l'expression du facteur de crédibilité Z^h adapté à la maille prêt

Calcul du paramètre W^h

Soit la *Loss* à minimiser $L = \mathbb{E}[(m_h - Z^h \widehat{m_h} - W^h)^2]$. Il est cherché à annuler la dérivée partielle selon W^h en 0 pour minimiser L. Il convient de remarquer premièrement que ce point correspond à un minimum local uniquement, car

$$\frac{\partial L^2}{\partial W_h^2} = 2 > 0$$

Le point W^h qui annule la dérivée partielle première est tel que :

$$\begin{aligned} \frac{\partial L}{\partial W^h} = 0 \Leftrightarrow \\ -2\mathbb{E}[(m_h - Z^h \widehat{m_h} - W^h)] = 0 \end{aligned}$$

d'où $W^h = \mathbb{E}[m^h] - Z^h\mathbb{E}[\widehat{m^h}]$

Or, $\mathbb{E}[m^h] = \mu$ par hypothèse, et, de plus :

$$\begin{aligned}
\mathbb{E}[\widehat{m^h}] &= \mathbb{E}[\frac{A_h}{E_h}] \\
&= \frac{1}{E_h}\mathbb{E}[A_h] \\
&= \frac{1}{E_h}\mathbb{E}[\sum_{i=1}^{n_h} D_{hi}] \\
&= \frac{1}{E_h}\mathbb{E}[\mathbb{E}[\sum_{i=1}^{n_h} D_{hi}|m_h]] \\
&= \frac{1}{E_h}\mathbb{E}[\sum_{i=1}^{n_h} \mathbb{E}[D_{hi}|m_h]] \\
&= \frac{1}{E_h}\mathbb{E}[\sum_{i=1}^{n_h} f_{hi} m_h q_{hi}^{ref}] \\
&= \frac{1}{E_h}\mathbb{E}[E_h m_h] \\
&= \mu
\end{aligned}$$

Finalement, $W^h = (1 - Z^h)\mu$. Il vient donc

$$\widetilde{m_h} = Z^h\widehat{m_h} + (1 - Z^h)\mu \tag{E.1}$$

On retrouve une forme d'écriture de m similaire à la forme proposée par la crédibilité partielle, à savoir une combinaison linéaire entre l'expérience du portefeuille et une référence. La différence étant ici la méthode de calcul du facteur de crédibilité Z^h.

Calcul du paramètre Z^h

Soit la *Loss* à minimiser $L = \mathbb{E}[(m_h - Z^h\widehat{m_h} - W^h)^2]$. Il est cherché à annuler la dérivée partielle selon Z^h en 0 pour minimiser L. Il convient premièrement de calculer la dérivée seconde en Z^h de L pour nous assurer que le zéro de la dérivée partielle simple corresponde à un minimum local (i.e $\frac{\partial L^2}{\partial Z_h^2} > 0$). Avec $L = \mathbb{E}[(m_h - Z^h\widehat{m_h} - (1 - Z^h)\mu)^2]$:

$$\frac{\partial L}{\partial Z_h} = 2\mathbb{E}[(m_h - Z^h\widehat{m_h} - (1 - Z^h)\mu)(-\widehat{m_h} + \mu)]$$

d'où

$$\begin{aligned}
\frac{\partial L^2}{\partial Z_h^2} &= 2\mathbb{E}[(-\widehat{m_h} + \mu)((-\widehat{m_h} + \mu))] \\
&= 2\mathbb{E}[(-\widehat{m_h} + \mu)^2] > 0
\end{aligned}$$

On cherche maintenant le point Z^h qui annule la dérivée partielle première :

$$\begin{gathered}
\tfrac{\partial L}{\partial Z^h} = 0 \Leftrightarrow \\
2\mathbb{E}[(m_h - Z^h\widehat{m_h} - (1 - Z^h)\mu)(-\widehat{m_h} + \mu)] = 0 \Leftrightarrow
\end{gathered}$$

$$Z^h\mathbb{E}[(-\widehat{m_h} + \mu)^2] = -\mathbb{E}[m_h(-\widehat{m_h} + \mu)] + \mu\mathbb{E}[(-\widehat{m_h} + \mu)]$$

d'où

$$\begin{aligned} Z^h &= \frac{-\mathbb{E}[m_h(-\widehat{m_h}+\mu)] + \mu\mathbb{E}[(-\widehat{m_h}+\mu)]}{\mathbb{E}[(-\widehat{m_h}+\mu)^2]} \\ &= \frac{\mathbb{E}[m_h\widehat{m_h}] - \mu\mathbb{E}[m_h] - \mu\mathbb{E}[\widehat{m_h}] + \mu^2}{\mu^2 - 2\mu\mathbb{E}[\widehat{m_h}] + \mathbb{E}[\widehat{m_h}^2]} \\ &= \frac{\mathbb{E}[m_h\widehat{m_h}] - \mu^2 - \mu^2 + \mu^2}{\mu^2 - 2\mu^2 + \mathbb{E}[\widehat{m_h}^2]} \\ &= \frac{\mathbb{E}[m_h\widehat{m_h}] - \mu^2}{\mathbb{E}[\widehat{m_h}^2] - \mu^2} \end{aligned}$$

À présent, les deux termes inconnus dans l'expression de Z^h sont développés, en commençant par le terme $\mathbb{E}[m_h\widehat{m_h}]$:

$$\begin{aligned} \mathbb{E}[m_h\widehat{m_h}] &= \mathbb{E}[m_h\frac{A_h}{E_h}] \\ &= \frac{1}{E_h}\mathbb{E}[\mathbb{E}[m_hA_h|m_h]] \\ &= \frac{1}{E_h}\mathbb{E}[m_h\mathbb{E}[\sum_{i=1}^{n_h} D_{hi}|m_h]] \\ &= \frac{1}{E_h}\mathbb{E}[m_h\sum_{i=1}^{n_h}\mathbb{E}[D_{hi}|m_h]] \\ &= \frac{1}{E_h}\mathbb{E}[m_h\sum_{i=1}^{n_h} f_{hi}m_hq_i^{ref}] \\ &= \frac{1}{E_h}\mathbb{E}[m_h^2E_h] \\ &= \mathbb{E}[m_h^2] \\ &= \mathbb{V}[m_h] + \mathbb{E}[m_h]^2 \\ &= \mu^2 + \sigma^2 \end{aligned}$$

finalement :

$$\mathbb{E}[m_h\widehat{m_h}] = \mu^2 + \sigma^2 \tag{E.2}$$

Le deuxième terme inconnu dans l'expression de Z^h est maintenant développé :

$$\begin{aligned} \mathbb{E}[\widehat{m_h}^2] &= \mathbb{E}[(\frac{A_h}{E_h})^2] \\ &= \frac{1}{E_h^2}\mathbb{E}[\mathbb{E}[A_h^2|m_h]] \end{aligned}$$

Pour un souci de clarté dans la démonstration, l'expression de A_h adaptée à la maille prêt est utilisée. i.e

$$A_h = \sum_{p=1}^{p_{max}} p\sum_{i\in I_p} D_{hi}$$

Notons qu'il s'agit uniquement d'un réarrangement de la somme A_h. Voir la partie La problématique de la maille prêt pour plus de détails à propos de cette expression.

En notant $S_p = p\sum_{i\in I_p} D_{hi}$:

$$\begin{aligned}
\mathbb{E}[\widehat{m_h}^2] &= \frac{1}{E_h^2}\mathbb{E}[\mathbb{E}[\sum_{p=1}^{p_{max}}\sum_{z=1}^{p_{max}} S_pS_z|m_h]] \\
&= \frac{1}{E_h^2}\left(\mathbb{E}[\mathbb{E}[\sum_{p\neq z} S_pS_z|m_h]] + \mathbb{E}[\mathbb{E}[\sum_{p=1}^{p_{max}} S_p^2|m_h]]\right) \\
&= \frac{1}{E_h^2}\left(\mathbb{E}[\sum_{p\neq z} \mathbb{E}[S_pS_z|m_h]] + \mathbb{E}[\sum_{p=1}^{p_{max}} \mathbb{E}[S_p^2|m_h]]\right)
\end{aligned}$$

Par souci de clarté, étudions ces deux termes séparément, calculons premièrement $\mathbb{E}[S_pS_z|m_h]]$ pour $p \neq z$:

$$\begin{aligned}
\mathbb{E}[S_pS_z|m_h]] &= \mathbb{E}[\sum_{i\in I_p}\sum_{j\in I_z} pzD_{hi}D_{hj}|m_h]] \\
&= pz\sum_{i\in I_p}\sum_{j\in I_z}\mathbb{E}[D_{hi}D_{hj}|m_h]]
\end{aligned}$$

Or les variables D_{hi} et D_{hj} sont les variables indicatrices de décès d'individus appartenant à des groupes distincts. Elles sont donc indépendantes conditionnellement à m_h. Il vient donc :

$$\begin{aligned}
\mathbb{E}[S_pS_z|m_h]] &= pz\sum_{i\in I_p}\sum_{j\in I_z}\mathbb{E}[D_{hi}|m_h]]\mathbb{E}[D_{hj}|m_h]] \\
&= \sum_{i\in I_p}\sum_{j\in I_z} pf_{hi}m_hq_{hi}^{ref}zf_{hj}m_hq_j^{ref}
\end{aligned}$$

En notant ${}^pE_h = \sum_{i\in I_p} pf_{hi}q_{hi}^{ref}$, il vient finalement :

$$\mathbb{E}[S_pS_z|m_h]] = {}^pE_h{}^zE_h\ m_h^2$$

Calculons à présent le second terme inconnu $\mathbb{E}[S_p^2|m_h]]$:

$$\begin{aligned}
\mathbb{E}[S_p^2|m_h]] &= \mathbb{E}[p^2\sum_{i\in I_p}\sum_{j\in I_p} D_{hi}D_{hj}|m_h]] \\
&= p^2\left(\mathbb{E}[\sum_{i\neq j} D_{hi}D_{hj}|m_h] + \mathbb{E}[\sum_{i\in I_p} D_{hi}^2|m_h]\right) \\
&= p^2\left(\sum_{i\neq j}\mathbb{E}[D_{hi}D_{hj}|m_h] + \sum_{i\in I_p}\mathbb{E}[D_{hi}^2|m_h]\right)
\end{aligned}$$

Or, au sein d'un groupe I_j, les individus sont indépendants par construction. De, plus, si $X \sim Ber(p)$,

alors $X^2 \sim Ber(p)$. Il vient alors :

$$\begin{aligned}
\mathbb{E}[S_p^2|m_h]] &= p^2 \left(\sum_{i \neq j} \mathbb{E}[D_{hi}|m_h]\mathbb{E}[D_{hj}|m_h] + \sum_{i \in I_p} \mathbb{E}[D_{hi}^2|m_h]] \right) \\
&= p^2 \left(\sum_{i \neq j} f_{hi} m_h q_{hi}^{ref} f_{hj} m_h q_j^{ref} + \sum_{i \in I_p} f_{hi} m_h q_{hi}^{ref} \right) \\
&= p^2 \left(\sum_{i \in I_p} \sum_{j \in I_p} f_{hi} m_h q_{hi}^{ref} f_{hj} m_h q_j^{ref} - \sum_{z \in I_p} (f_{hz} m_h q_z^{ref})^2 + \sum_{i \in I_p} f_{hi} m_h q_{hi}^{ref} \right)
\end{aligned}$$

En posant ${}^pC_h = p^2 \sum_{z \in I_p} (f_{hz} q_z^{ref})^2$, il vient :

$$\mathbb{E}[S_p^2|m_h]] = {}^pE_h^2 m_h^2 - {}^pC_h m_h^2 + p\ {}^pE_h m_h$$

Finalement, en revenant au calcul de $\mathbb{E}[\widehat{m_h}^2]$:

$$\begin{aligned}
\mathbb{E}[\widehat{m_h}^2] &= \frac{1}{E_h^2} \left(\mathbb{E}[\sum_{p \neq z} \mathbb{E}[S_p S_z|m_h]] + \mathbb{E}[\sum_{p=1}^{p_{max}} \mathbb{E}[S_p^2|m_h]] \right) \\
&= \frac{1}{E_h^2} \left(\mathbb{E}[\sum_{p \neq z} {}^pE_h {}^zE_h\ m_h^2] + \mathbb{E}[\sum_{p=1}^{p_{max}} {}^pE_h^2 m_h^2 - {}^pC_h m_h^2 + p\ {}^pE_h m_h] \right) \\
&= \frac{1}{E_h^2} \left(\sum_{p \neq z} {}^pE_h {}^zE_h\ \mathbb{E}[m_h^2] + \sum_{p=1}^{p_{max}} {}^pE_h^2 \mathbb{E}[m_h^2] - {}^pC_h \mathbb{E}[m_h^2] + p\ {}^pE_h \mathbb{E}[m_h] \right) \\
&= \frac{1}{E_h^2} \left((\mu^2 + \sigma^2) \sum_{p \neq z} {}^pE_h {}^zE_h + (\mu^2 + \sigma^2) \sum_{p=1}^{p_{max}} {}^pE_h^2 - (\mu^2 + \sigma^2) \sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h \right) \\
&= \frac{1}{E_h^2} \left((\mu^2 + \sigma^2)(\sum_{p=1}^{p_{max}} {}^pE_h)^2 - (\mu^2 + \sigma^2) \sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h \right)
\end{aligned}$$

Or, par construction, $\sum_{p=1}^{p_{max}} {}^pE_h = E_h$, donc :

$$\mathbb{E}[\widehat{m_h}^2] = \mu^2 + \sigma^2 + \frac{1}{E_h^2} \left(-(\mu^2 + \sigma^2) \sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h \right) \tag{E.3}$$

Enfin, la forme de du facteur de crédibilité de Bühlmann en présence de dépendance dans les données est telle que :

$$Z^h = \frac{\mathbb{E}[m_h \widehat{m_h}] - \mu^2}{\mathbb{E}[\widehat{m_h}^2] - \mu^2} \tag{E.4}$$

$$= \frac{\sigma^2}{\sigma^2 + \frac{1}{E_h^2} \left(-(\mu^2 + \sigma^2) \sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p\ {}^pE_h \right)} \tag{E.5}$$

E.3 Demonstration de l'expression du facteur σ^2

La forme naturelle de l'estimateur serait :

$$\widehat{\sigma^2} = \sum_{h=1}^{r} E_h(\widehat{m_h} - \widehat{\mu})^2$$

Dans cette expression, la variance est induite par deux sources différentes. En effet, l'estimateur $\widehat{m_h}$ est source de variance inter au sein de chaque portefeuille, là où l'estimateur $\widehat{\mu}$ est source de variance intra entre les différents portefeuilles.
Pour donner, un estimateur de la variance, il sera montré que :

$$\mathbb{E}[\widehat{\sigma^2}] = \alpha\sigma^2 + \beta$$

où α,$\beta \in \mathbb{R}^*$.
On pourra alors poser :

$$\widetilde{\sigma^2} = \frac{\widehat{\sigma^2} - \beta}{\alpha}$$

avec ainsi :

$$\begin{aligned}
\mathbb{E}[\widetilde{\sigma^2}] &= \mathbb{E}[\frac{\widehat{\sigma^2} - \beta}{\alpha}] \\
&= \frac{\mathbb{E}[\widehat{\sigma^2}] - \beta}{\alpha} \\
&= \frac{\alpha\sigma^2 + \beta - \beta}{\alpha} \\
&= \sigma^2
\end{aligned}$$

Calcul de $\mathbb{E}[\widehat{\sigma^2}]$

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma^2}] &= \mathbb{E}[\sum_{h=1}^{r} E_h(\widehat{m_h} - \widehat{\mu})^2] \\
&= \mathbb{E}[\sum_{h=1}^{r} E_h \left(\widehat{m_h}^2 - 2\widehat{m_h}\widehat{\mu} + \widehat{\mu}^2\right)] \\
&= \mathbb{E}[\sum_{h=1}^{r} E_h \widehat{m_h}^2] - 2\mathbb{E}[\sum_{h=1}^{r} E_h \widehat{m_h}\widehat{\mu}] + \mathbb{E}[\sum_{h=1}^{r} E_h \widehat{\mu}^2] \\
&= \mathbb{E}[\sum_{h=1}^{r} E_h \widehat{m_h}^2] - 2\mathbb{E}[\widehat{\mu} \sum_{h=1}^{r} A_h] + \mathbb{E}[\sum_{h=1}^{r} E_h \widehat{\mu}^2]
\end{aligned}$$

Or, $\sum_{h=1}^{r} A_h = \widehat{\mu} \sum_{h=1}^{r} E_h$

$$\begin{aligned}
&= \mathbb{E}[\sum_{h=1}^{r} E_h \widehat{m_h}^2] - 2\mathbb{E}[\widehat{\mu}^2 \sum_{h=1}^{r} E_h] + \mathbb{E}[\sum_{h=1}^{r} E_h \widehat{\mu}^2] \\
&= \sum_{h=1}^{r} E_h \mathbb{E}[\widehat{m_h}^2] - \mathbb{E}[\widehat{\mu}^2] \sum_{h=1}^{r} E_h
\end{aligned}$$

D'après l'équation (8) :

$$\mathbb{E}[\widehat{m_h}^2] = \mu^2 + \sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h\right)$$

Le calcul de $\mathbb{E}[\widehat{\mu}^2]$ est maintenant présenté

$$\begin{aligned}
\mathbb{E}[\widehat{\mu}^2] &= \mathbb{E}[(\frac{\sum_{h=1}^r A_h}{\sum_{h=1}^r E_h})^2] \\
&= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\mathbb{E}[(\sum_{h=1}^r A_h)^2] \\
&= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\mathbb{E}[\sum_{h\neq z} A_h A_z + \sum_{h=1}^r A_h^2] \\
&= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\sum_{h\neq z}\mathbb{E}[A_h A_z] + \sum_{h=1}^r \mathbb{E}[A_h^2]\right)
\end{aligned}$$

Or, d'après l'équation (8) :

$$\begin{aligned}
\mathbb{E}[\widehat{m_h}^2] &= \mu^2 + \sigma^2 + \frac{1}{E_h^2}\left(-(\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h\right) \\
&= \mathbb{E}[(\frac{A_h}{E_h})^2] = \frac{1}{E_h^2}\mathbb{E}[A_h^2]
\end{aligned}$$

D'où

$$\mathbb{E}[A_h^2] = E_h^2(\mu^2+\sigma^2) - (\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h$$

Étant donné l'indépendance entre les individus des différents portefeuilles et le fait que $\mathbb{E}[A_h] = \mu E_h$ (voir détails des calculs de l'équation (6)) :

$$\begin{aligned}
\mathbb{E}[\widehat{\mu}^2] &= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\sum_{h\neq z}\mathbb{E}[A_h A_z] + \sum_{h=1}^r \mathbb{E}[A_h^2]\right) \\
&= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\sum_{h\neq z}\mathbb{E}[A_h]\mathbb{E}[A_z] + \sum_{h=1}^r \mathbb{E}[A_h^2]\right) \\
&= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\sum_{h\neq z}\mu E_h \mu E_z + \sum_{h=1}^r\left(E_h^2(\mu^2+\sigma^2) - (\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h\right)\right) \\
&= \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\mu^2[\sum_{h=1}^r E_h]^2 + \sum_{h=1}^r\left(E_h^2\sigma^2 - (\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h\right)\right)
\end{aligned}$$

En effet $\sum_{h\neq z}\mu E_h \mu E_z = \mu^2\left([\sum_{h=1}^r E_h]^2 - \sum_{h=1}^r E_h^2\right)$.
Finalement :

$$\begin{aligned}
\mathbb{E}[\widehat{\mu}^2] &= \mu^2 + \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\sum_{h=1}^r\left(E_h^2\sigma^2 - (\mu^2+\sigma^2)\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{p=1}^{p_{max}} p\ {}^pE_h\right)\right) \\
&= \mu^2 + \frac{1}{\left(\sum_{h=1}^r E_h\right)^2}\left(\sigma^2\sum_{h=1}^r E_h^2 - (\mu^2+\sigma^2)\sum_{h=1}^r\sum_{p=1}^{p_{max}} {}^pC_h + \mu\sum_{h=1}^r\sum_{p=1}^{p_{max}} p\ {}^pE_h\right)
\end{aligned}$$

On peut à présent donner une forme pour l'estimateur de σ^2. On rappelle que

$$\widetilde{\sigma^2} = \frac{\widehat{\sigma^2} - \beta}{\alpha}$$

On a montré que

$$\mathbb{E}[\widetilde{\sigma^2}] = \sum_{h=1}^{r} E_h \left(\mu^2 + \sigma^2 + \frac{1}{E_h^2} \left(-(\mu^2 + \sigma^2) \sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{p=1}^{p_{max}} p \, {}^pE_h \right) \right) -$$
$$\left(\mu^2 + \frac{1}{\left(\sum_{h=1}^{r} E_h\right)^2} \left(\sigma^2 \sum_{h=1}^{r} E_h^2 - (\mu^2 + \sigma^2) \sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h + \mu \sum_{h=1}^{r} \sum_{p=1}^{p_{max}} p \, {}^pE_h \right) \right) \sum_{h=1}^{r} E_h$$

donc que

$$\begin{aligned} \alpha &= \sum_{h=1}^{r} E_h - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h} - \frac{\sum_{h=1}^{r} E_h^2}{\sum_{h=1}^{r} E_h} + \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} \\ &= \sum_{h=1}^{r} E_h - \frac{\sum_{h=1}^{r} E_h^2}{\sum_{h=1}^{r} E_h} + \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h} \end{aligned}$$

et

$$\begin{aligned} \beta &= -\mu^2 \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h} + \mu \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} p \, {}^pE_h}{E_h} + \mu^2 \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \mu \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} p \, {}^pE_h}{\sum_{h=1}^{r} E_h} \\ &= \mu^2 \left(\frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h} \right) + \mu \left(\sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} p \, {}^pE_h}{E_h} - \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} p \, {}^pE_h}{\sum_{h=1}^{r} E_h} \right) \end{aligned}$$

Finalement, en remplaçant μ par son estimateur, la forme de l'estimateur de σ^2 adapté à la maille prêt est la suivante :

$$\widetilde{\sigma^2} = \frac{\sum_{h=1}^{r} E_h (\widehat{m_h} - \widehat{\mu})^2 - \widehat{\mu}^2 \left(\frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h} \right) - \widehat{\mu} \left(\sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} p \, {}^pE_h}{E_h} - \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} p \, {}^pE_h}{\sum_{h=1}^{r} E_h} \right)}{\sum_{h=1}^{r} E_h - \frac{\sum_{h=1}^{r} E_h^2}{\sum_{h=1}^{r} E_h} + \frac{\sum_{h=1}^{r} \sum_{p=1}^{p_{max}} {}^pC_h}{\sum_{h=1}^{r} E_h} - \sum_{h=1}^{r} \frac{\sum_{p=1}^{p_{max}} {}^pC_h}{E_h}}$$

Annexe F

Construction par abattement d'une table de mortalité

F.1 Démonstration de la solution des moindres carrés dans un cas particulier

Dans le cadre étudié, la variable explicative n'est qu'une constante (par palier). La *Loss* à minimiser est donc la suivante :

$$Loss(\alpha) = \sum_{i=1}^{n} (y_i - \alpha)^2$$

Pour minimiser cette fonction de α, une solution est de chercher le point qui annule sa dérivée première. Pour vérifier que ce point correspond bien à un minimum, il convient premièrement de vérifier que la dérivée seconde de la fonction *Loss* est toujours positive :

$$Loss''(\alpha) = 2n > 0$$

À présent, le point qui annule est dérivée première est tel que :

$$Loss'(\widehat{\alpha}) = -2 \sum_{i=1}^{n} (y_i - \widehat{\alpha}) = 0$$

$$\Leftrightarrow \widehat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

La solution à ce cas particulier du problème des moindres carrés est correspond donc à la moyenne de la variable réponse observée y.

Démonstration de la solution des moindres carrés pondérés dans un cas particulier

Dans le cadre étudié, la variable explicative n'est qu'une constante (par palier). La *Loss* à minimiser est donc la suivante :

$$Loss(\alpha) = \sum_{i=1}^{n} w_i (y_i - \alpha)^2$$

Avec $w_i > 0 \; \forall i$.
Pour minimiser cette fonction de α, une solution est de chercher le point qui annule sa dérivée première. Pour vérifier que ce point correspond bien à un minimum, il convient premièrement de vérifier que la dérivée seconde de la fonction *Loss* est toujours positive :

$$Loss''(\alpha) = 2 \sum_{i=1}^{n} w_i > 0$$

À présent, le point qui annule est dérivée première est tel que :

$$Loss'(\widehat{\alpha}) = -2 \sum_{i=1}^{n} w_i (y_i - \widehat{\alpha}) = 0$$

$$\Leftrightarrow \widehat{\alpha} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i y_i$$

La solution à ce cas particulier du problème des moindres carrés est correspond donc à la moyenne pondérée par les poids w_i de la variable réponse observée y.