Modélisation de la durée de vie de contrats d'assurance

Birama Basse

le cnam
intd

Mémoire pour l'obtention du

Master Sciences humaines et sociales mention humanités numériques - Parcours Mégadonnées et analyse sociale
(MEDAS)

Modélisation de la durée de vie de contrats d'assurance

Birama BASSE

Date et lieu de la soutenance

- 05 Juillet 2022
- CFA CNAM Saint-Denis (93)

Membres du jury

- Ghislaine CHARTRON, Présidente du Jury
- Aurélien LATOUCHE, Tuteur Pédagogique

Promotion (2021-2022)

Le fait d'anticiper les résiliations des contrats demeure un enjeu à forte attente pour les organismes assureur, car il pourrait voir leur chiffre d'affaires chuter de manière significative, mais aussi perdre au profit des assurés qui représentent de bons risque.

L'objectif de ce mémoire est de modéliser la durée de vie de contrats d'assurance santé individuelle dans le but d'identifier un profil précis d'individus concernés par le risque de résiliation et d'expliquer ces résiliations par des variables qui peuvent avoir des influences sur celles-ci.

Mots clés :

Assurance santé,
Résiliations,
Modèles de durée,
Probabilité de Survie,
Estimateur de Kaplan-Meir,
Modèle de Cox,
Régression logistique

Anticipating contract terminations remains a high expectation issue for insurers, as they could see their turnover drop significantly, but also lose out to policyholders who represent good risks.

The objective of this thesis is to model the life of individual health insurance contracts in order to identify a precise profile of individuals concerned by the risk of cancellation and to explain these cancellations by variables that may have an influence on them.

Keywords

Health insurance,
Cancellations,
Duration models,
Probability of Survival,
Kaplan-Meir's estimator,
Cox model,
Logistic regression

Remerciements

Je tiens tout d'abord à adresser mes remerciements à tous ceux qui m'ont permis de réaliser mon apprentissage au sein de **GFP**.

Je tiens à témoigner toute ma reconnaissance à **Mathieu DES RIEUX**, tuteur en entreprise, de m'avoir accueilli au sein de leur équipe, de m'avoir fait confiance et je le remercie grandement pour sa disponibilité et toutes les connaissances qu'il a pu m'apporter au cours de ces deux années d'alternance.

Je remercie **M. Aurélien LATOUCHE**, tuteur pédagogique, pour sa disponibilité et ses précieux conseils.

Nous souhaitons également adresser nos remerciements les plus sincères au corps professoral et administratif du **Master MEDAS**, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Table des matières

Introduction

Le secteur de l'assurance est constitué principalement de multiples acteurs porteurs de risque des clients, généralement appelés **organismes d'assurance** qui peuvent être des sociétés d'assurance, des institutions de prévoyance ou des mutuelles. Ces organismes d'assurance procèdent soit à une gestion interne et directe de leurs contrats d'assurance, soit confient tout ou partie de cette gestion à un tiers externe : le délégataire de gestion.

L'augmentation du nombre de contrats Santé et/ou prévoyance, grâce notamment à l'obligation aux entreprises de couvrir leurs salariés a permis l'émergence de nombreux délégataires de gestion. Ces derniers gèrent pour leurs clients un grand nombre de données.
En exploitant et analysant ces big data grâce à des outils sécurisés et développés, le délégataire peut prévenir un grand nombre de risques.

En faisant appel à un délégataire de gestion, les entreprises ou les courtiers ou porteurs de risques peuvent bénéficier de nombreuses innovations technologiques et de l'expertise d'équipes qui leurs permettront alors de se concentrer pleinement sur le développement de leur activité. En optant pour une délégation totale ou partielle de la gestion de leurs contrats santé et prévoyance, en individuel comme en collectif, ces derniers optimisent ainsi leurs coûts et consacrent du temps à d'autres tâches pour gagner en agilité et en compétitivité.

La société **GFP** est un délégataire de gestion. Elle est spécialisé depuis plus de 30 ans, dans la gestion de contrats santé et/ou prévoyance pour le compte de tiers comme les assurances ou les courtiers. Son rôle consiste à gérer des portefeuilles de contrats collectifs ou individuels pour le compte de compagnies d'assurances, de Mutuelles ou d'Institutions de Prévoyance.

Parmi la gamme de contrats, **GFP** gère des contrats d'assurance complémentaire santé

destinés à garantir le remboursement des dépenses de santé en complément de l'assurance maladie obligatoire (frais médicaux, hospitalisation, soins dentaires, optique, radios,...). Ces contrats s'adressent à des particuliers.

L'objectif dans ce mémoire est de modéliser la durée des vie des contrats santé individuelle afin d'identifier un profil précis d'individus fragiles à la résiliation, dans le but de mettre en place des actions à mener pour diminuer ce risque de résiliation.

Nous allons dans un premier temps faire une présentation générale de l'environnement de travail en présentant l'entreprise d'une part et d'autre part, le contexte de l'étude et le traitement des données utilisées.
Ensuite nous présenterons dans la dernière partie, les modélisations effectuées : d'une part les modèles de survie et d'autre part par la régression logistique.

Première partie

Présentation de l'entreprise et contexte de l'étude

Chapitre 1

GFP | Gestionnaire Santé et Prévoyance

Afin de mieux comprendre l'environnement dans lequel j'ai effectué mon apprentissage, il est judicieux d'avoir une vue d'ensemble de l'entreprise au travers de son histoire, sa constitution et son fonctionnement. Nous procédons ainsi à une présentation de **GFP** dans sa globalité.

1.1 Présentation de GFP

GFP (Gestion - Formation - Prévoyance) est une société familiale créée en décembre 1988 par Gérard FEURTÉ, dont le siège social est situé à Chartres (28, Eure-et-Loir).

Son capital social détenu majoritairement par les membres de la famille du fondateur et le Comité de Direction marque son indépendance vis-à-vis de ses partenaires et du marché.

GFP est spécialisée dans la gestion déléguée de contrats Frais Médicaux et Prévoyance pour le compte de partenaires institutionnels (Mutuelles, Compagnies d'assurance, Institutions de prévoyance), de courtiers, de bancassureurs et de cabinets de conseil.

Le Groupe **GFP** est créé en 2014 et est constitué de 5 entités expertes dans leur métiers :

• Implantation géographique

Outre Chartres, le groupe a six autres implantations en France. Sur Chartres Métropole, il possède sept locaux. Au total, 13 sites d'activité sont gérés par **GFP**.

Le siège social est situé à Chartres (28).

Les Centres de gestion répartis sur le territoire sont autonomes. Ils possèdent leurs propres équipes affiliations, prestations et cotisations.

Les Centres de relation client sont également en France métropolitaine, à Chartres (28), Orange (84) et Orléans (45).

1.2 Secteur d'activité

L'activité de **GFP** consiste à gérer des portefeuilles de contrats collectifs (entreprises de 1 à 40 000 salariés), ainsi que leur population d'individuels et/ou d'allocataires pour le compte de Compagnies d'assurances, de Mutuelles, de bancassureurs et d'Institutions de Prévoyance.

Le cœur du métier réside dans la gestion complète des dossiers Frais Médicaux et Prévoyance et sur l'accompagnement de l'ensemble de la chaîne de valeur : affiliation, prestations, cotisations, gestion de la relation client, documentation, éditique, pilotage, gestion du risque et formation.

Le champ d'action est donc la gestion de contrats d'assurance de personnes physiques contre les accidents corporels, l'invalidité, la maladie, le décès.
Elle est souscrite soit :

— A titre individuel : qui est souscrite, de sa propre initiative, par un particulier en son nom

— A titre collectif : le contrat est souscrit par l'entreprise pour le compte de ses salariés

Dans un premier temps, le contact est effectué avec les partenaires (porteurs de risques, entreprises, courtiers...) afin de mettre en place les contrats pour gérer des portefeuilles collectifs et individuels Frais médicaux et Prévoyance. Les contrats sont, par la suite, enregistrés dans l'outil de gestion.
Ils sont alors informatisés, avec le paramétrage des informations (entreprises, CSP, garanties...), le suivi et la mise à jour des contrats, la vérification des garanties et la modification

éventuelle des contrats par les partenaires. À partir de cette étape, les adhésions et la gestion des contrats peuvent commencer.
Viennent donc l'enregistrement des adhésions, le remboursement des prestations, les mises à jour des données des bénéficiaires (RIB, coordonnées, choix d'option...) et la réponse aux mails des assurés, ainsi qu'aux appels pour répondre à leur demande. Et enfin, les données sont extraites et transmises aux partenaires chaque semaine, ou mensuellement, ou encore occasionnellement, en fonction de leur demande. Ils leur sont, par exemple, transmis la liste des salariée, les frais médicaux, le ratio sinistre/prime, les services...

1.3 Organisation de GFP

1.3.1 Organisation

L'organisation de **GFP** repose sur 3 processus :

— **Le processus de Pilotage**, C'est le processus où les lignes directrices sont établies et déployées, puis contrôler et corriger pour assurer la conformité des processus et enfin analysées et améliorées dans le but d'optimiser les processus.

— **Le processus de Réalisation**, qui repose sur deux axes : Le premier, où les contrats sont mis en place. Les besoins du marché sont identifiés et analysés, la part du marché augmentée et les offres de produit déterminées. Les besoins internes et externes d'évolutions fonctionnelles sont identifiés et développés. Les outils informatiques de sont paramétrés et les sites web, et les statistiques techniques et de production sont établies.

  Le second, qui concerne les flux entrants et sortants (réception des courriers, numérisation des dossiers reçus et traités...), la gestion déléguée (maintient à jour les informations des assurés et entreprises, cotisations, remboursements), la gestion SaaS (coordination des activités liées à la mise en gestion et à la relation client des clients SaaS), et la relation client.

— **Le processus de Support**, qui regroupe la comptabilité générale et partenaire (Gére la comptabilité générale, la comptabilité clients et fournisseurs,...), l' environnement de travail (assure la disponibilité et l'entretien des infrastructures, négocie et achète des logiciels,...), la maintenance informatique (Mettre en oeuvre, maintenir les matériels, les logiciels, programmes POWER/hors POWER,...) et enfin les ressources humaines.

1.3.2 Système d'information

Le système d'information (SI) est un élément central d'une entreprise ou d'une organisation qui permet aux différents acteurs de véhiculer des informations et de communiquer grâce à un ensemble de ressources matérielles, humaines et logicielles. Autrement dit, il permet de créer, collecter, stocker, traiter, modifier des informations sous divers formats.

GFP utilise l'IBM AS/400 (une gamme de mini-ordinateurs IBM) comme support pour son développement informatique, réputé pour sa fiabilité, sa polyvalence ainsi que sa souplesse d'évolution.
L'AS/400 fonctionne avec un système d'exploitation nommé 0S/400. L'architecture de ce système est faite de couches, les principaux éléments matériels (mémoire, entrées-sorties, gestion des processus...) et logiciels sont donc séparés.

C'est le système d'information sur lequel sont gérées et historisées les données des contrats gérés par le Groupe GFP. Il est interconnecté avec la GED, la messagerie ZIMBRA, le reporting et l'éditique.
Les données sont gérées par environnement de travail à partir de plusieurs serveurs de connexion.

• **Les outils de gestion**

L'ensemble des outils suivants constitue le système d'informe de GFP :

— **GED** (Gestion Electronique des Données) : est un outil qui permet la mise à disposition des documents de gestion dématérialisés. Les informations contenues dans ces documents sont vidéocodées pour permettre leur identification et réparties dans des bannettes de traitement (parexemple les Bulletins d'Adhésions)

— **OPEN :** est un outil de gestion développé en interne et sur lequel sont gérées et historisées les données des contrats. Il est interconnecté avec la GED, la messagerie ZIMBRA, le reporting et l'éditique.
Les données sont gérées par environnement de travail sur plusieurs serveurs (aussi appelés en interne « machines » ou « AS/400 »).

— **QONECT :** C'est l'outil de Gestion de la Relation Client (GRC) regroupant l'ensemble des dispositifs qui permet d'optimiser la qualité de la relation client. Il est développé par la société Prosodie.
Il gère l'ensemble des interactions avec nos clients quelque soit le média par lequel

ils nous contactent : téléphone, email, fax, courrier. Il est interconnecté avec notre base clients (Open) ainsi qu'avec la GED, la messagerie Zimbra, le reporting et l'éditique.

• **Comment sont intégrées les données**

Le système d'information de GFP est constitué d'applications et de données. Les informations sont spécialisées par services. Une architecture orientée (notée SOA pour Services Oriented Architecture) a été développée afin d'avoir une vision globale du système d'information.

L'architecture logicielle globale a donc été divisée en services correspondant aux processus métiers de GFP, et donc en plusieurs environnements différents.
Un service web est invoqué, envoyant des requêtes ODBC au serveur pour collecter des données afin d'intégrer les informations de contact et la GED à la base de données.

Les bases de données de GFP sont tout de même principalement alimentées par les gestionnaires, lorsque les données des contrats et des assurés sont historisées dans l'OPEN.

Pour l'échange de données informatisées, les données sont également intégrées par le biais de TALEND, un éditeur de logiciel spécialisé dans l'intégration des données (ETL). Il permet d'automatiser les extractions ou les imports, de collecter des données en provenance de sources multiples pour ensuite les convertir dans un format adapté à une Datawarehouse et les y transférer. Ceci intervient donc lors d'extractions automatisées par TALEND pour collecter les données (de la GED ou d'OPEN) et alimenter une base de SQL Server.

Les principales bases de données de GFP se situent donc dans OPEN, et peuvent être récupérées directement ou via des extractions automatisées par TALEND.

Les données pour chaque assuré reprennent :

- Le paramétrage du contrat :
  - Police (contrat collectif ou individuel)
  - Affectation (rattachement de la police à l'assureur du contrat
  - Garanties et tarifs - L'enregistrement de l'assuré (affiliation)
- L'appel, l'encaissement et la ventilation des cotisations

— Le paiement des prestations (liquidation)
— La relation client (appels, mails...)

Ces données sont transmises au courtier et à l'assureur au sein du service reporting.

• **Modes de gestion**

GFP propose à ses partenaires 3 modes collaboratifs différents :

— **Délégation de gestion** : **GFP** propose sa gestion déléguée de deux manières différentes
  — La gestion en marque blanche : La gestion est personnalisée aux couleurs du partenaire, GFP n'apparaît pas en tant que gestionnaire auprès des assurés.
  — La gestion en marque GFP : La gestion est réalisée sous les couleurs de GFP.
— **Solution SaaS** : Solution d'infogérance qui permet à nos partenaires de gérer leurs contrats totalement ou partiellement sur notre système d'information (ex : KLESIA)
— **Gestion BPO (Business Process Outsourcing)** : Solution d'externalisation qui permet à nos partenaires d'utiliser notre système d'information et de bénéficier d'une équipe dédiée et expérimentée (ex : Cabinet BESSE)

1.4 Délégation de gestion en assurance

• Définition :

La délégation de gestion consiste, dans le domaine assurantiel, à confier tout ou partie d'activités menant à l'exécution d'un contrat d'assurance à un autre organisme qui effectuera ses tâches de façon autonome en utilisant ses propres ressources, humaines, matérielles (notamment les systèmes d'informations) et financières. Le délégataire de gestion prend en charge l'intégralité du processus qui lui est confié. Ainsi, la délégation de gestion est une forme de sous-traitance.[4].

Cette délégation de gestion peut concerner tout ou partie du processus de la durée de vie du contrat d'assurance : souscription avec ou sans acceptation médicale, gestion des adhésions, encaissement des cotisations, gestion des droits des adhérents, gestion des prestations, etc.
Le délégant est l'organisme qui délègue et le délégataire est le tiers, à qui est confié tout

ou une partie des activités d'assurance.

• Les acteurs de la délégation de gestion en assurance :

Les principales catégories de délégataires de gestion présentes sur le marché des assurances de personnes sont :

— **Les courtiers** ayant développé leurs propre plate-forme de gestion ;
— **Les prestataires** spécialisés dans la gestion pour compte en assurance ;
— **Les organismes d'assurances** gérant une ou plusieurs activités d'un autre assureur (par exemple, les activités de souscription ou de gestion de sinistres) ;
— **Les GIE** (Groupe d'intérêts économiques) ou centre de gestion.

• Enjeux de la délégation de gestion :

Le marché de l'assurance (Santé/Prévoyance) est caractérisé par une part relativement importante d'externalisation (transfert d'activités d'une entreprise vers un prestataire externe spécialisé) des actes des de gestion.
Le recours à la délégation de gestion pour les organismes d'assurance peut intervenir dans trois cas de situation :

1. Une politique de l'organisme d'assurance :
   — en raison d'un savoir faire parce qu'il n'a pas une expertise complète (ex. gestion de contrats étrangers, représentation de collectivités locales, etc.) ;
   — ou se permettre de faire face à des charges qui ne peuvent être assumées par ses propres moyens (par exemple, systèmes d'information inefficaces pour gérer certains contrats ou garanties spécifiques etc.) ;
   — soit pour transformer les coûts fixes en coûts variables.
2. Une réponse à un cahier des charges lors de la réalisation de l'affaire :
   — Soit de la compagnie signataire du contrat d'assurance ;
   — soit du courtier qui apporte l' affaire.
3. Un partenariat de coassurance ou de réassurance avec une autre compagnie d'assurance lorsqu'elle prévoit la cogestion du contrat.

Par ailleurs, les intermédiaires d'assurances s'y intéressent également, notamment pour :

— offrir aux clients ce qu'ils perçoivent comme des services de meilleure qualité ;
— intégrer une partie de la chaîne de valeur ;
— réduire les contraintes lors des changements d'assureurs.

Pour les entreprises clientes, les principaux avantages associés à la gestion confiée reposent sur :

— l'accès à un service personnalisé et la mise à disposition d'un reporting spécifique ;

— la dissociation entre l'assureur, l'intermédiaire et le gestionnaire des prestation ;

— une meilleure transparence des coûts.

Chapitre 2

L'état de l'art, le système de santé français

Pour mieux cerner l'environnement dans lequel le sujet a été produit, il est nécessaire de comprendre quelques notions :

— L'assurance santé français
— Contrat complémentaire santé
— Résiliation de contrat d'assurance

2.1 La sécurité sociale en France

2.1.1 Présentation de la sécurité sociale

• Concepts et généralités

La sécurité sociale créée en 1945, désigne un ensemble de dispositifs et d'institutions majoritairement publique mis en place pour permettre à chaque individu ou ménage de faire face aux conséquences de la survenue d'un risque ou d'un besoin social tout au long de sa vie. Les risques sont des situations qui peuvent survenir et dégrader le bien être et le niveau de vie des ménages.
Nous avons plusieurs types de risques qui peuvent être de nature diverse :

— risques professionnels : accidents de travail, maladies professionnels
— risques non professionnels : maladie vieillesse, invalidité, maternité, décès, veuvage
— risques économiques : chômage

• Les différents régimes de le sécurité sociale

Les assurés de la sécurité sociale sont affiliés à des régimes (c'est à dire un statut qui permet à l'assuré de jouir de l'ensemble des ses droits). L'affiliation est obligatoire pour toute personne résidant en France et le régime de rattachement dépend de la situation professionnelle de l'assuré. Ce dernier participe au financement du régime auquel il est rattaché par les cotisations qu'il est tenu de verser.
Nous avons quatre régimes principaux de la sécurité sociale :

- **Le régime général** couvre l'ensemble des salariés du secteur privé ainsi que les travailleurs indépendants. Il compte plus de 62 millions de bénéficiaires, c'est à dire plus de 92% de la population française [5].
- **Le régime agricole** couvre les salariés exploitants dans le secteur agricole et de l'élevage en France.
- **Le régime social des indépendants** concerne les travailleurs non salariés (commerçants, artisants, professions libérales, etc) excepté ceux du secteur agricole.
- **Les régimes spéciaux** :
  - Le régime Alsace-Moselle : est un régime destiné aux départements du Haut-Rhin (68), du Bas-Rhin (67) et de la Moselle (57). En effet, ces départements regroupés sous le nom de Alsace-Moselle ont consevé le régime mis en palce alors ils étaient annexés par l'Allemagne.
  - Les autres régimes spéciaux : ils regroupent un certains corps de métier qui ont leur propre régime de sécurité sociale (les agents de la RATP et SNCF, les fonctionnaires, les marins, les militaires, les magistrats, opéra de Paris, etc).

• Les branches de la sécurité sociale

Le régime général de la sécurité sociale est divisé en plusieurs branches d'activités qui recouvrent chacun plusieurs risques :

- **La branche famille** prend en charge les prestations familiales et sociales (logements, handicap,...) et est gérée par la Caisse nationale des allocations familiales (Cnaf).
- **La branche Maladie** est gérée par la Caisse nationale d'assurance maladie (Cnam). Elle couvre les risques maladie, maternité, invalidité, décès.
- **Accidents du travail et Risques professionnels** couvre les risques accidents du travail et maladies professionnelles.
- **La branche vieillesse** est gérée par la Caisse nationale d'assurance vieillesse (Cnav). Cette branche verse les pensions de retraite de base.

Nous allons nous concentrer sur la branche qui nous concerne à savoir la branche maladie (ou assurance maladie).

2.1.2 Quelques mots sur l'assurance maladie

L'assurance maladie ou l'assurance santé prend en charge les dépenses de santé des assurés, garantit l'accès aux soins et favorise l'accès à la santé des plus démunis. Elle rembourse partiellement les dépenses de santé. Elle mène des programmes de prévention et participe à la régulation du système de santé français.

Gérée par la Caisse nationale de l'Assurance Maladie (CNAM), elle a principalement un réseau composée de :

— CPAM : Caisses primaires d'assurance maladie,
— DRSM : Directions régionales du service médical ;
— CARSAT : Caisse d'assurance retraite et de la santé au travail ;
— CGSS : Caisse générales de sécurité sociale dans les départements d'outre-mer ;
— UGECAM : Unions de gestion des établissements de caisse d'assurance maladie.

Les prestations de la branche maladie peuvent être soit :

— des **prestations en nature** correspondant à des remboursements de frais de santé (médecine, soins, frais pharmaceutiques, hospitalisation, vaccination...).
— des **prestations en espèces** (indemnités journalières de maladie en cas d'incapacité temporaire du travail).

Il y a un certain niveau de remboursement des frais de santé par la sécurité sociale obligatoire pour les assurés. Pour compenser les frais non couverts par la sécurité sociale, on peut souscrire à un contrat d'assurance complémentaire santé chez un organisme assureur.

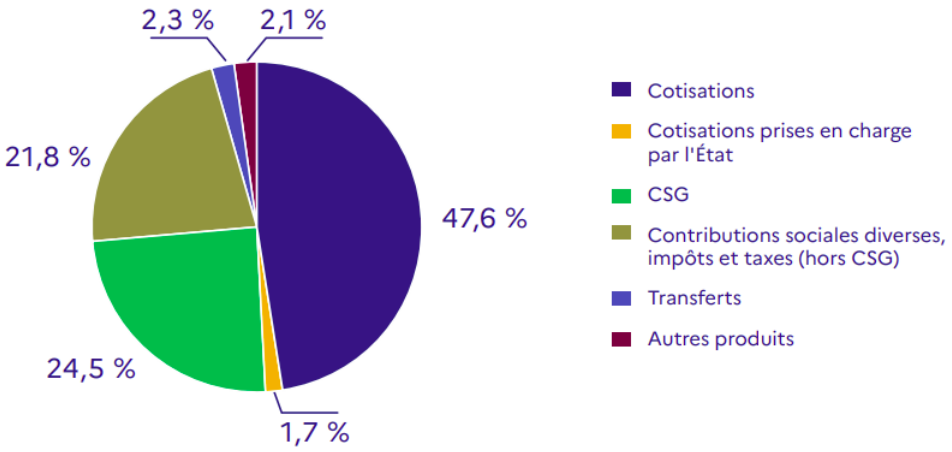
2.1.3 Les chiffres de la protection sociale

• Les financements de la sécurité sociale

Le financement de la sécurité est essentiellement assuré par :

— **Les cotisations sociales** : ce sont des prélèvements obligatoires sur les salaires. Elles sont à la fois à la charge de l'employeur et du salarié qui en payent chacun une part.
— **Les impôts et taxes affectés (ITAF)** qui sont des prélèvements obligatoires affectés au financement de la sécurité sociale.

— **La contribution publiques** : Ce sont des cotisations que l'état prend en charges. Selon la direction de la sécurité sociale [5], 528 milliards d'euros de recettes ont été recouvrées en 2020. Sur la figure suivante on peut voir les parts de chaque source de financement de la sécurité sociale.

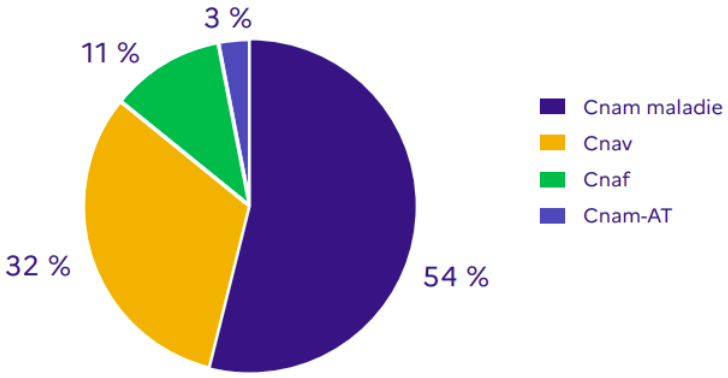


Source : Commission des comptes de la sécurité sociale, juin 2021.

FIGURE 2.1 – Recettes du régime général et du FSV en 2020

• **Les dépenses de la Sécurité sociale**

Les prestations annuelles de la sécurité sociale sont de 470 milliards d'euros et sont répartis comme suit :



Source : Commission des comptes de la sécurité sociale, juin 2021.

FIGURE 2.2 – Les dépenses du régime général en 2020

Plus de la moitié des dépenses (54%) de la sécurité sociale est occupé par la branche maladie gérée par le cnam.

2.2 L'assurance complémentaire santé

Une assurance complémentaire santé est un contrat d'assurance qui prend en charge tout ou une partie des dépenses de santé (concernant la maladie, l'accident et la maternité) non couvertes par l'assurance maladie obligatoire. Grâce à la complémentaire santé, l'assuré peut limiter des paiements directs qui peuvent rester élevés même si ces paiements ont été pris en charge par l'assurance maladie obligatoire.

2.2.1 Les différents types de contrats d'assurance complémentaire

On distingue principalement deux types de contrats complémentaires santé : les contrats individuels et les contrats collectifs.

- **Les contrats collectifs :** c'est un contrat conclu entre une entreprise et un organisme assureur pour faire bénéficier, à titre obligatoire, d'une couverture complémentaire santé à l'ensemble des salariés. Le salarié peut dans certains cas choisir de faire adhérer les membres de sa famille (conjoints, enfants).
- **Les contrats individuels :** ce sont des contrats souscrites à titre individuel. Ils s'adressent principalement aux personnes ne bénéficiant pas de complémentaire santé collectifs (étudiants, retraités, chômeurs,...).

La généralisation de la complémentaire santé depuis le 1er janvier 2016 avec l'entrée en vigueur de la loi ANI de 2013, les entreprises du secteur privé ont l'obligation de proposer une complémentaire de santé collective à l'ensemble de leurs salariés. Cette obligation s'applique à tous les salariés quelle que soit l'ancienneté dans l'entreprise.

2.2.2 Les différents acteurs

Un contrat d'assurance complémentaire santé peut être proposé par trois grands types d'organismes :

- **Les mutuelles :** ce sont des sociétés à but non lucratif. Elles sont régies par le code de la mutualité.
- **Les sociétés d'assurance :** ce sont des sociétés anonymes régies par le code des assurances.

— **Les institutions de prévoyance** : c'est des sociétés à but non lucratif et sont régies par le code de la sécurité sociale. Elles gèrent principalement des contrats collectifs couvrant les risques maladie, décès, incapacité, invalidité, dépendance.

En 2020, d'après les chiffres publiés par la DREES (Direction de la recherche, des études, de l'évaluation et des statistiques), 683 organismes d'assurance pratiquent des activités d'assurance de toute nature selon l'Autorité de contrôle prudentiel et de résolution (ACPR) : 369 mutuelles, 281 sociétés d'assurances et 33 institutions de prévoyance.(cf.[7]).

• **Répartition dans le payage français**

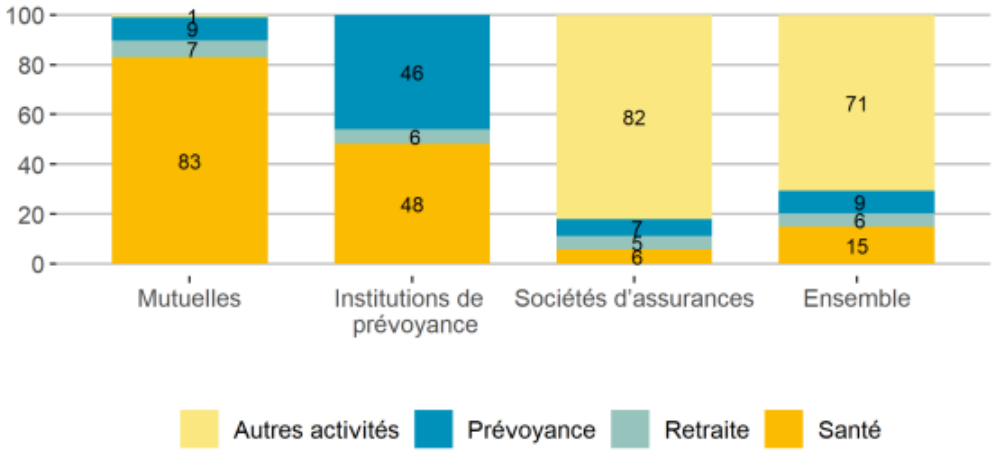


FIGURE 2.3 – Répartition des cotisations collectées en 2019 selon les types d'organismes complémentaires (Source : DREES)

En 2019, les organismes d'assurance santé ont collectés 37,5 milliards de cotisations. Ce sont les mutuelles qui dominent le marché de l'assurance santé avec 83% de cotisations collectés, ensuite les institutions de prévoyance avec 48% et enfin les sociétés d'assurances avec 6%. (Source DREES).

2.3 Résiliations chez les assurés

Plusieurs raisons peuvent pousser les assurés à la résiliation de leurs contrats :

— Des offres compétitives plus attractives ;

— Insatisfaction auprès de l'organisme assureur : par exemple, mauvais rapport qualité/prix selon l'avis des assurés, insatisfaction vis-à-vis de la gestion des contrats, insatisfaction des membres de la famille bénéficiaire etc.

— Changement de situation :par exemple, mariage, retraite ou déménagement etc.

2.3.1 Les modalités de résiliations en assurance santé

Les possibilités des assurés à résilier leurs contrats santé diffèrent selon le motif de résiliation.

• Résiliation à date d'anniversaire du contrat

Les contrats d'assurance complémentaire santé sont soumis par défaut au principe de tacite reconduction. Par conséquent, Ils sont automatiquement renouvelés chaque année, sauf indication contraire. Pour cela, la résiliation d'une complémentaire santé demande donc une action de la part l'assuré d'informer son assureur (organisme assureur : mutuelle ou institution de prévoyance).

• Résiliation pour changement de situation

L'assuré peut aussi de résilier son contrat en cas de changement de situation, dès la première avant la d'anniversaire du contrat et si le changement modifie le risque couvert. Les différents changements sont alors les suivants : mariage, divorce, retraite, changement de profession, déménagement, changement de régime social...
Cette résiliation ne peut intervenir que dans un délai de 3 mois à compter de la date de l'événement qui entraîne le changement de situation.

• Résiliation pour augmentation des cotisations injustifiées

Dans le cas d'une augmentation non justifié des cotisations non indiqué sur le contrat, l'assuré a la possibilité de demander la résiliation de son contrat complémentaire santé. Il dispose alors généralement d'un délai de 15 ou 30 jours pour adresser une demande de résiliation à son assureur.

• **Résiliation pour adhésion à une complémentaire santé collective obligatoire**

Dans le cadre de la loi ANI (Accord National Interprofessionnel) entré en vigueur le 1er Janvier 2016, les entreprises du secteur privé ont l'obligation de proposer à tout leurs salariés une complémentaire santé collective. Ainsi l'assuré peut peut résilier son contrat complémentaire santé individuelle hors date d'échéance, sous condition d'un préavis d'un mois.

2.3.2 Résiliations et motifs dans notre base de données

Nous distinguons plusieurs types de motifs de résiliation dans notre base. Ces motifs viennent en général de l'initiative du client qui sont le plus souvent des résiliations à échéance.

Le tableau ci-dessous indique les proportions de résiliations de notre base de données pour chaque motifs.

Motifs de résiliation	Proportion
ADHESION AU CMU	0,65%
ADHESION AU DISPOSITIF ACS	0,39%
DECES	9,45%
NON PAIEMENT	11,29%
RESILIATION DU CLIENT	64,63%
OBLIGATOIRE PASSAGE GROUPE	3,49%
CHANGEMENT DE SITUATION	1,09%
AUTRES MOTIFS	9,01%
TOTAL	100,00%

FIGURE 2.4 – Proportion des différents motifs de résiliation

Nous constatons ici que plus de la moitié des résiliations viennent de l'initiative de l'assuré. Suivi des radiations pour non paiement de cotisations qui représente $11,29$ des cas.

Chapitre 3

Cadre de l'étude

3.1 Construction de la base d'étude

3.1.1 Périmètre de l'étude

L'étude dans ce mémoire porte sur des contrats d'assurance santé individuelle gérés par GFP. En effet, GFP gère des contrats d'assurance santé ou prévoyance, que ce soit des contrats collectifs et individuels pour le compte de ces partenaires (assureurs, mutuelles, institutions de prévoyance etc).
Nous avons choisi le portefeuille de contrats d'assurance santé géré par **GFP** en prenant que les données relatives aux contrats d'assurance santé individuelle.
De plus, la date de début d'observations choisie est le 01 Janvier 2000 et la date de fin d'observation est le 31 Janvier 2022.

3.1.2 Construction de la base

Une fois le périmètre choisi, nous avons procédé à la construction de la base de données.
La plupart des données sont répertoriées dans des tables de notre DataShare. On extrait ces données via SQL server.
La phase d'extraction et le nettoyage des données a nécessite beaucoup temps car importante pour la suite notre étude.
Dans un premier temps, nous avons listés les différents variables nécessaire pour notre étude :

— Sur les contrats : type de contrats, date d'effet, date de fin des contrats,...

— Sur les assurés : sexe, date de naissance, situation familiale, ...

Ensuite nous avons repéré les tables qui contiennent ces différents variables.

Dans un un second temps, nous avons effectué un travail de gestion au niveau de ces différents tables afin de générer une base de données exploitables. Tout ce travail de Data Management a été réalisé sur notre système de gestion de base données relationnelles Microsoft SQL Server. Étant l'un des logiciels de ce type les plus utilisés dans les entreprises en particulier en Europe et aux États-Unis. Il offre une protection des données dynamiques, avec un accès unique pour les personnes autorisées aux données sensibles.

• **Présentation des bases utilisées**

Les données que nous avons récupérés sont regroupés au sein des tables suivantes :

— La table **"Bénéficiaires"** contient principalement les informations concernant les assurés à savoir nom, prénom, date de naissance, sexe, situation familiale, ...

— La table **"Polices"** qui contient toutes les informations relatives aux contrats. On y retrouve principalement la date d'effet du contrat, la date de résiliation du contrat, la nature du contrat, le numéro de contrat, ...

• **Nettoyage des données**

Dans cette partie, nous avons définie les types de données des variables, supprimer les observations aberrantes et les doublons, crée d'autres variables :

Nous avons :

— renommer certaines variables ;
— converti en type *numeric* les variables quantitatives dénombrables ;
— converti en type *factor* les variables qualitatives ;
— crée d'autres variables à partir des variables *dates* : *âge de l'assuré*, *ancienneté des contrats d'assurance*, etc ;
— recodé les variables *Régime* et *Motif*.

Nous avons aussi supprimer les contrats avec des données manquantes et les contrats sans effet (les contrats dont la date d'effet est égale à la date de fin du contrat)...

Après le traitement de la base, notre base de données finale contient 599415 observations, soit 599415 contrats à étudiés.

3.1.3 Présentation des variables

Nous allons présenter ici les différentes variables que nous allons utilisées dans notre étude.

- **L'âge** : L'âge de l'assuré. C'est une variable quantitative qui comporte 6 modalités : 'moins de 20ans', '20-29ans', '30-39ans', '40-49ans', '50-62ans' et '63ans et plus' ;
- **La civilité** : Cette variable correspond à l'état civil de l'assuré et comporte deux modalités : 'Madame', 'Monsieur' ;
- **La situation familiale** : Elle correspond à la situation familiale de l'assuré et contient 7 modalités : 'Célibataire', 'Concubin', 'Marié', 'Pacsé', 'Divorcé', 'Séparé', 'Veuf' ;
- **Nombre d'enfants** : Il s'agit du nombre d'enfants (de l'assuré) bénéficiaires du contrat en plus de l'assuré. Les différentes classes sont : '1', '2', '3', '4 et plus' ;
- **Région, département** : Il s'agit de la zone géographique de l'assuré ;
- **Date d'effet** : Date à partir de laquelle le contrats est souscrit ;
- **Date de fin** : Date à laquelle le contrat a été résilié s'il y a eu ;
- **Motif** : Elle correspond au motif de la résiliation. Elle indique si la résiliation provient de l'assuré ou de la compagnie ;
- **Ancienneté du contrat** : Cette variable indique le nombre d'années d'ancienneté du contrat.
- **Régime** : Elle correspond au régime de sécurité sociale de l'assuré. Il prend trois modalités : 'Régime général', 'Régime agricole' et les autres régimes (les régimes spéciaux).
- **Niveau de garantie** : correspond à la capacité de ce dernier à rembourser les dépenses de santé engagées par l'assuré. En d'autres termes, il s'agit de la performance de la couverture offerte par la garantie. Elle prend 3 modalités : 'Base', 'Option 1', 'Option 2'.
- **RESIL** : Cette variable comporte deux modalités : '1' si le contrat est résilié ou '0' si le contrat n'est pas résilié.

3.2 Analyse descriptive

Avant de faire la modélisation de la durée de vie des contrats, il est nécessaire d'avoir une bonne compréhension du portefeuille que nous examinons. Il est donc nécessaire de prendre le temps de décrire en détail les différentes variables et d'examiner la composition du portefeuille à analyser.

3.2.1 Répartition du portefeuille

Les 599415 observations de notre base de données se répartissent de la manière suivante :

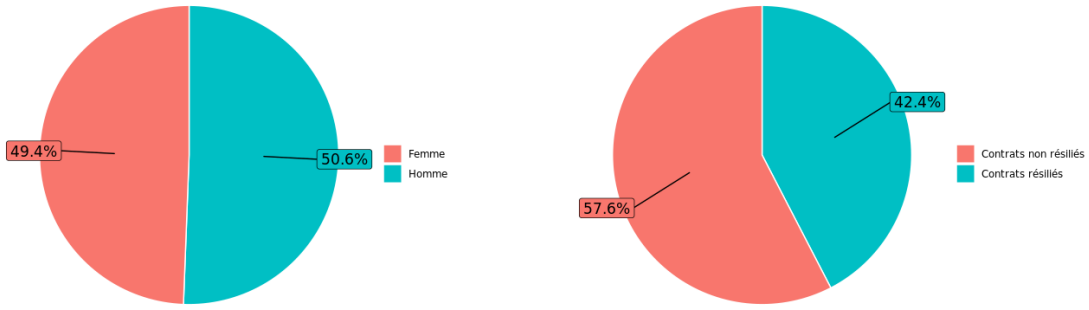


FIGURE 3.1 – Répartition du portefeuille

Dans le portefeuille, les hommes sont un peu majoritaires par rapport aux femmes. La proportion de contrats résiliés est de 42.4% contre 57.6% de contrats non résiliés.

L'évolution du taux de résiliation n'est cependant pas constante au fil des années. Le graphique ci-dessous présente la proportion de contrats résiliés pour chaque année. Nous constatons que nous avons un nombre de résiliations très élevé en 2016. Cette constatation peut être justifié avec la mise en vigueur de la **loi ANI** qui oblige les employeurs privés y compris les associations à proposer une mutuelle à l'ensemble de leurs salariés.

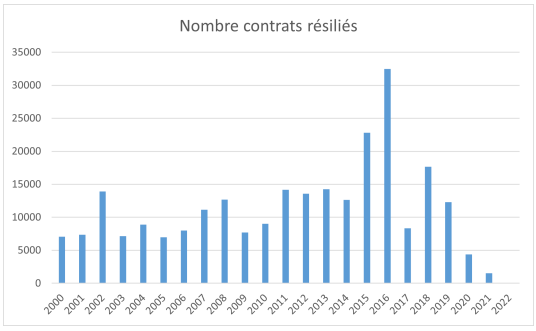


FIGURE 3.2 – Répartition du nombre de résiliations par année

3.2.2 Profil du portefeuille

• **Classe d'âge de l'assuré principal**

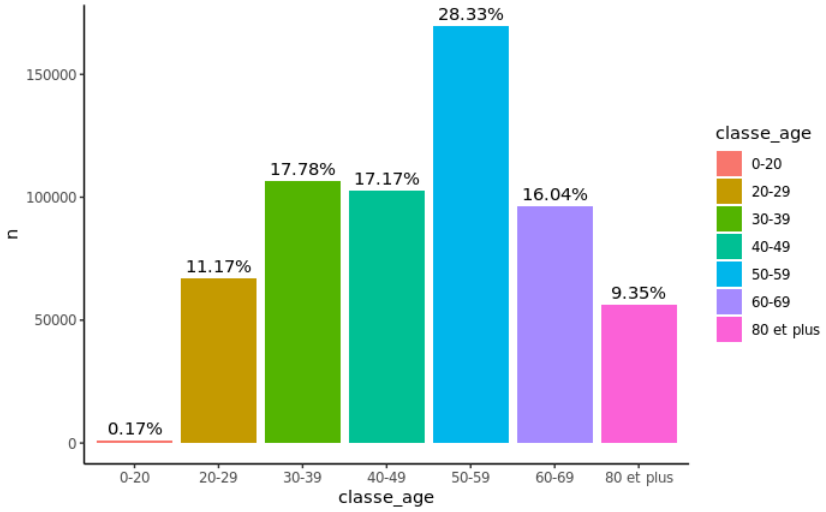


FIGURE 3.3 – Répartition du portefeuille par classe d'âge de l'assuré

La classe d'âge la plus représentée dans notre portefeuille est la classe 50-59 ans avec 28.33% des contrats. Suivi de la classe 30-39 ans avec 17.78%. La tranche d'âge 60-69 ans représente aussi une part importante avec 16.04% des contrats, cela peut s'expliquer par le fait que cette tranche d'âge correspond à l'entrée en retraite et la perte du contrat collectif d'entreprise, ce qui leur pousse à souscrire à un contrat individuel.

• **Le nombre d'enfants bénéficiaires du contrat**

On peut voir ici que 57.3% des contrats du portefeuille sont associés à une personne. Ces statistiques montrent que les assurés sont principalement des adultes vivant seuls et sans enfant. Selon leurs âges, ces individus sont principalement des salariés, au chômage ou retraités.

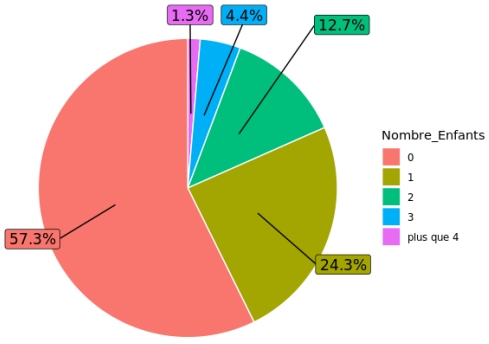


FIGURE 3.4 – Répartition du portefeuille en fonction du nombre d'enfants

- **Le régime de sécurité sociale**

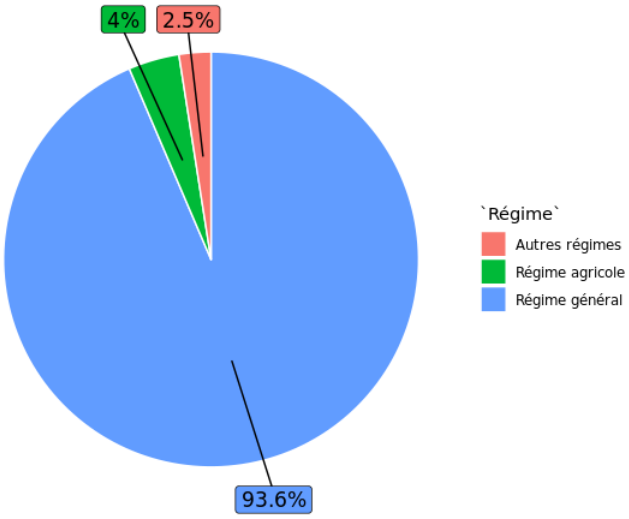


FIGURE 3.5 – Répartition du portefeuille en fonction du régime

Une part importante des assurés sont associés au régime général obligatoire avec 93.6% du nombre total d'assurés. Ce qui explique que la plupart des assurés sont des salariés.

- **Niveau de garantie**

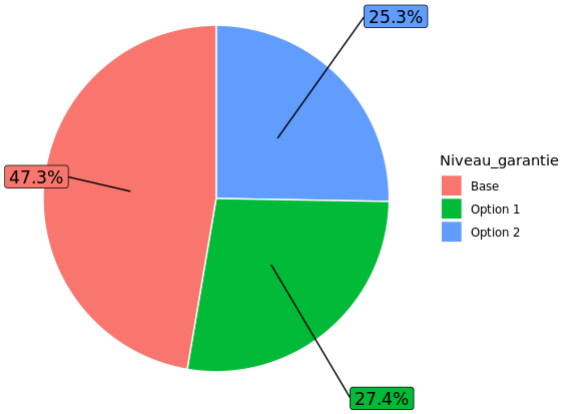


FIGURE 3.6 – Répartition du portefeuille par niveau de garanties

Nous pouvons constater que 47.3% des contrats ont un niveau de garanties "Base" et 27.4% ont un niveau de garanties "Option 1" et le reste des contrats ont un niveau de garanties "Option 2", soit 25.3%.

• **Situation familiale de l'assuré**

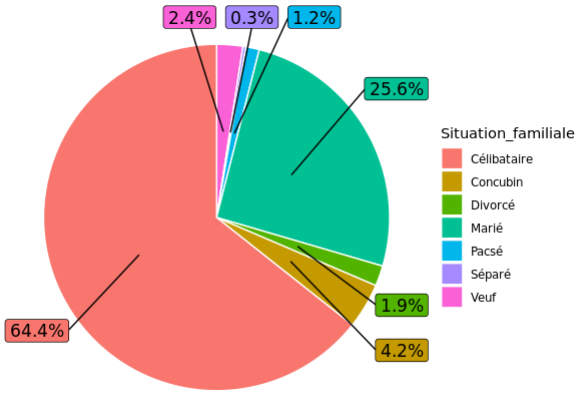


FIGURE 3.7 – Répartition du portefeuille en fonction de la situation familiale

Nous pouvons observer que plus de la moitié des assurés, 64.4%, sont des célibataires, suivie des mariés avec 25.6% et le reste des assurés ne représentent qu'environ 10%, qui sont soit concubin, divorcé, pacsé, séparé ou veuf.

• **Répartition du portefeuille en fonction de la région**

Région	Poids des assurés	Pourcentages
Auvergne-Rhône-Alpes	66812	11,15%
Bourgogne-Franche-Comté	22604	3,77%
Bretagne	23922	3,99%
Centre-Val de Loire	23586	3,93%
Corse	4748	0,79%
Grand Est	29216	4,87%
Hauts-de-France	45795	7,64%
Île-de-France	171884	28,67%
Normandie	26245	4,38%
Nouvelle-Aquitaine	53171	8,87%
Occitanie	52268	8,72%
Pays de la Loire	27053	4,51%
Provence-Alpes-Côte Azur	52160	8,70%
Total général	**599464**	**100,00%**

FIGURE 3.8 – Répartition du portefeuille en fonction du régime

On observe ici que la plupart des assurés résident dans la région îles-de-France avec 28.6% des cas. La région Auvergne-Rhône-Alpes vient en deuxième positon avec 11.15%.

La région Corse est la moins représentée avec 0.79%. Cette répartition semble cohérente vu la démographie de ses régions.

• Composition principale du portefeuille du portefeuille

Notre portefeuille est composé principalement :
— Une ancienneté moyenne de 5.7 ans des contrats dans le portefeuille ;
— L'âge moyenne de l'assuré principal est de 55 ans ;
— 93.6% des contrats sont du régime général obligatoire ;
— Un peu près de la moitié des contrats ont un niveau de garantie de base, soit 47.3% ;
— 28.6% des contrats sont localisés dans la région Île-de-France ;
— Plus de la moitié des assurés sont célibataires.

3.2.3 Études des corrélations entre les différentes variables

Dans cette section nous allons étudier la corrélation entre nos différentes variables. Elle va nous permettre d'identifier les variables ayant une forte corrélation entre eux et d'éviter l'utilisation de variables ayant la même information dans la partie modélisation. Elle se fera entre notre variable cible et les variables explicatives et ensuite entre les variables explicatives.
Pour ce faire, nous allons utiliser le test de V Cramer.

• Test de Khi-deux

Soit une population de taille n et considérons deux variables aléatoires X et Y telles que :
— X avec les modalités $(x_1, x_2, ..., x_k)$
— Y avec les modaalités $(y_1, y_2, ..., y_k)$
Soit N_{ij} le nombre d'individus avec les modalités x_i pour X et y_j pour Y.
L'objectif du test de khi_deux est de tester l'hypothèse H_0 contre l'hypothèse H_1 :
— H_0 : X et Y indépendante ;
— H_1 : X et Y non indépendante
Et la statistique du test est donnée sous H_0 par :

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(N_{ij}-\frac{N_iN_j}{n})^2}{\frac{N_iN_j}{n}}$$

Elle mesure la distance entre les effectifs de deux séries et les effectifs théoriques s'il y a indépendance de variable. Elle suit une loi de Khi-deux noté χ^2 de degré de liberté $(k-1)(m-1)$.

- **Le V Cramer**

Le V de cramer est basé sur le χ^2 maximal produit par le tableau de contingence théoriquement. Pour deux variables aléatoires X et Y, le V de Cramer est donnée par la formule suivante :

$$V = \sqrt{\frac{\chi^2}{\chi^2{}_{max}}} = \sqrt{\frac{\chi^2}{N \times [min(k,m)-1]}}$$

Avec

— k est le nombre de modalités de la variable X ;

— m est le nombre de modalités de la variable Y ;

— N est le nombre d'observations.

Nous avons obtenu les résultats suivantes avec le V de Cramer :

Variables	V Cramer/RESIL
Ancienneté	89,20%
Niveau_garantie	21,90%
Régime	10,80%
Situation_familiale	5,90%
Nombre_Enfants	5,80%
Classe_age	4,20%
Sexe	4,20%
Zone	3,90%

FIGURE 3.9 – Corrélation entre variables explicatives et variable cible

Nous voyons ici qu'il y a une forte corrélation entre les variables **Ancienneté** et **Résiliations** et une corrélation assez forte entre les variables **Régime** et **Résiliation**, et les variables **Niveau garantie** et **Résiliation**.

3.2.4 Échantillonnage

L'échantillonnage est une phase importante dans la modélisation. Le but l'échantillonnage est de construire deux bases d'étude, une base apprentissage et une base de validation que nous allons utiliser pour la modélisation dans le dernier chapitre de ce mémoire. Le modèle sera construit avec la base d'apprentissage et nous allons utiliser la base de validation pour vérifier la stabilité et la robustesse du modèle retenu.
Nous avons choisi de prendre 30% des observations pour la base de validation et 70% pour la base d'apprentissage.

Cette partie d'analyse descriptive nous a permis de mieux connaître le profil de notre portefeuille et la liaison entre nos différentes variables que nous allons utiliser pour la modélisation dans la suite de ce mémoire.

Deuxième partie

Modélisation

Chapitre 1

Modélisation de la durée de vie des contrats

La durée de survie correspond au temps qui s'écoule depuis un instant initial t jusqu'à la survenue d'un évènement précis (par exemple, décès). Les modèles de durée constituent aujourd'hui un outil souvent utilisé dans différents domaines de l'assurance : le temps passé en arrêt de travail, la durée avant la ruine, la durée de conservation d'un contrat mais aussi le temps d'attente entre deux sinistres. Cependant, ces modèles ont un champ d'application assez large comme dans la médecine (le temps avant une rechute ou un rejet de greffe), la démographie ou dans l'industrie (étudier la durée de vie d'une machine). Dans cette section, nous utiliserons ces modèles de durée pour modéliser la durée de vie des contrats santé. Nous chercherons ensuite à estimer la distribution des durées de survie, à comparer les fonctions de risques entre groupes et à analyser comment certaines variables explicatives modifient ces fonctions.

1.1 Les modèles de survie

1.1.1 Distribution de survie

Considérons une variable aléatoire T qui représente la durée de vie et prenant ses valeurs dans $[0, +\infty[$, F sa fonction de répartition et f sa densité.

— **Fonction de survie**

La fonction de survie est la probabilité de survivre jusqu'au temps t et défini par :

$$S(t) = 1 - F(t) = P(T > t)$$

$S(t)$ est une fonction monotone décroissante tel que $S(0) = 1$ et $\lim_{t \longrightarrow \infty} = 0$.

— **Survie conditionnelle**

On s'intéresse à la durée de survie après un instant t d'un individu sachant qu'il est en vie jusqu'à t.

Cette fonction est définie par :

$$S_u = (P > u + t | T > t) = \frac{S(u+t)}{S(u)}$$

— **Fonction de risque**

La fonction de risque est définie par :

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

et désigne la probabilité que l'évènement survienne dans l'intervalle de temps $[t, t + dt]$. Elle représente ici le taux de résiliation des contrats santé suivant une ancienneté donnée.

1.1.2 Censure et troncature

• **Censure**

Lors de la collecte de données de survie, la censure est le phénomène le plus couramment rencontré. Une durée de vie est dit censuré si l'on sait seulement qu'il commence ou se termine dans un intervalle de temps précis, et généralement en dehors de la période de suivi. Par conséquent, la durée exacte est inconnue. La censure est dite :

— **Censure à droite** : si à la fin de la période de suivi, l'évènement ne s'est pas encore produit c'est à dire qu'on ne connaît pas la durée de vie T, mais seulement que $T > t$.

— **Censure à gauche** : la censure à gauche se produit lorsque nous ne pouvons pas observer la durée à laquelle l'événement s'est produit. Pour des raisons évidentes, si l'événement est un décès, les données ne peuvent pas être censurées à gauche. Un bon exemple est le test de virus. Par exemple, si nous avons suivi un individu et enregistré un événement lorsque l'individu est testé positif pour un virus, mais nous ne connaissons pas la durée exacte à laquelle l'individu a été exposé à la maladie. Nous savons seulement qu'il y a eu une certaine exposition entre 0 et le moment où il a été testé.

Dans le domaine de l'assurance la censure à droite est le phénomène le plus rencontré (par exemple : durée de vie des contrats).

• **Troncature**

Les données tronquées sont assez différentes des données censurées. Elles correspondent à des durées où aucune observation n'est inférieure à un certain seuil (troncature à gauche) ou supérieure à un certain seuil (troncature à droite). Contrairement aux données censurées, nous n'avons pas d'informations sur l'existence d'une durée de vie minimale (ou maximale).

1.2 Modèle non paramétrique

Pour modéliser la durée de vie des contrats, on utilise principalement des modèles de durée de vie. Nous allons dans un premier temps faire recours à une estimation non paramétrique. Les modèles non paramétriques permettent d'estimer l'une des différentes fonctions caractérisant la distribution de la variable T sans faire aucune hypothèse a priori sur la loi de survie.

1.2.1 Estimateur de Kaplan-Meier

La construction de l'estimateur de Kaplan-Meier s'appuie sur : la probabilité de survivre après un temps $t > s$ et peut s'écrire :

$$S(t) = P(T > t|T > s)P(T > s) = P(T > t|T > s)S(s)$$

On renouvelle l'opération en faisant apparaître des produits de termes en $P(T > t|T > s)$.

Si on considère des temps d'évènements (sortie ou censure) distinctes $T_{(i)}(i = 1, ..., n)$ tel que $T_{(1)} < T_{(2)} < ... < T_{(n)}$, on peut estimer des probabilités de la forme :

$$p_i = P(T > T_{(i)}|T > T_{(i-1)})$$

avec p_i la probabilité de survivre sur l'intervalle $]T_{(i-1)}, T_{(i)}[$ sachant qu'on était vivant à l'instant $T_{(i-1)}$.

On note par $q_i = 1 - p_i$ qui peut être estimer de la manière suivante $\hat{q} = \frac{d_i}{r_i} = \frac{d_i}{n-i+1}$, avec r_i le nombre d'individu à risque de subir l'évènement avant le temps $T_{(i)}$ et d_i le nombre de sortie en $T_{(i)}$.
On définit la fonction $D_{(i)}$ par : $D_i = 1$ si $X_i < C$ et 0 sinon.

avec C une censure fixe et X_i la durée de survie observée.

A l'instant $T_{(i)}$, et en l'absence d'ex æquo, on observe alors, si $D(i) = 1$ une sortie par décès donc $d_i = 1$ et dans le cas contraire l'observation est censurée et $d_i = 0$.
L'estimateur de Kaplan-Meier s'écrit donc de la manière suivante :

$$\hat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{D_{(i)}}{r_i}\right) = \prod_{T_{(i)} \leq t} \left(1 - \frac{n - i + 1}{r_i}\right)^{D_{(i)}}$$

En présence d'ex æquo, si on suppose par convention que les observations non censurées précèdent toujours les observations censurées alors l'estimateur de Kaplan-Meir s'écrit comme suit :

$$\hat{S}(t) = \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

(cf. F. PLANCHET [2]).

1.2.2 Estimation de la fonction de survie de Kaplan-Meier

Nous allons utiliser dans cette partie notre base de données qui contient les contrats du 1er janvier 2000 jusqu'au 31 janvier 2022, avec la variable **ancienneté** que nous avons calculée comme étant la différence entre la date d'effet et la date de résiliation du contrat si le contrat est résilié ou bien la différence entre la date d'effet et la date de fin d'observation si le contrat est en cours.
Nous sommes ici en présence d'une censure à droite fixe, c'est-à-dire que la date de résiliation des contrats en cours de janvier 2022 n'est pas connue car la base est arrêtée à ce mois, donc l'ancienneté maximale ou censure est de 22 ans. Parmi les 599415 observations nous avons 254146 résiliations et 345269 censures soit 42.4% de contrats résiliés et 57.6% de censures (contrats non résiliés). L'ancienneté médiane des contrats est de 4ans. Le taux de censure 57.6% élevé peut s'expliquer par la croissance de l'activité qui induit que la que plupart des contrats du portefeuille sont des contrats récents.

Nous allons utiliser l'estimateur de Kaplan-Meier pour estimer la fonction de survie sur l'ensemble de notre base de données. Nous obtenons le graphique suivant :

La fonction de survie de notre portefeuille nous permet d'avoir une idée de la durée de survie des contrats de notre portefeuille. Nous remarquons sur ce graphique que la

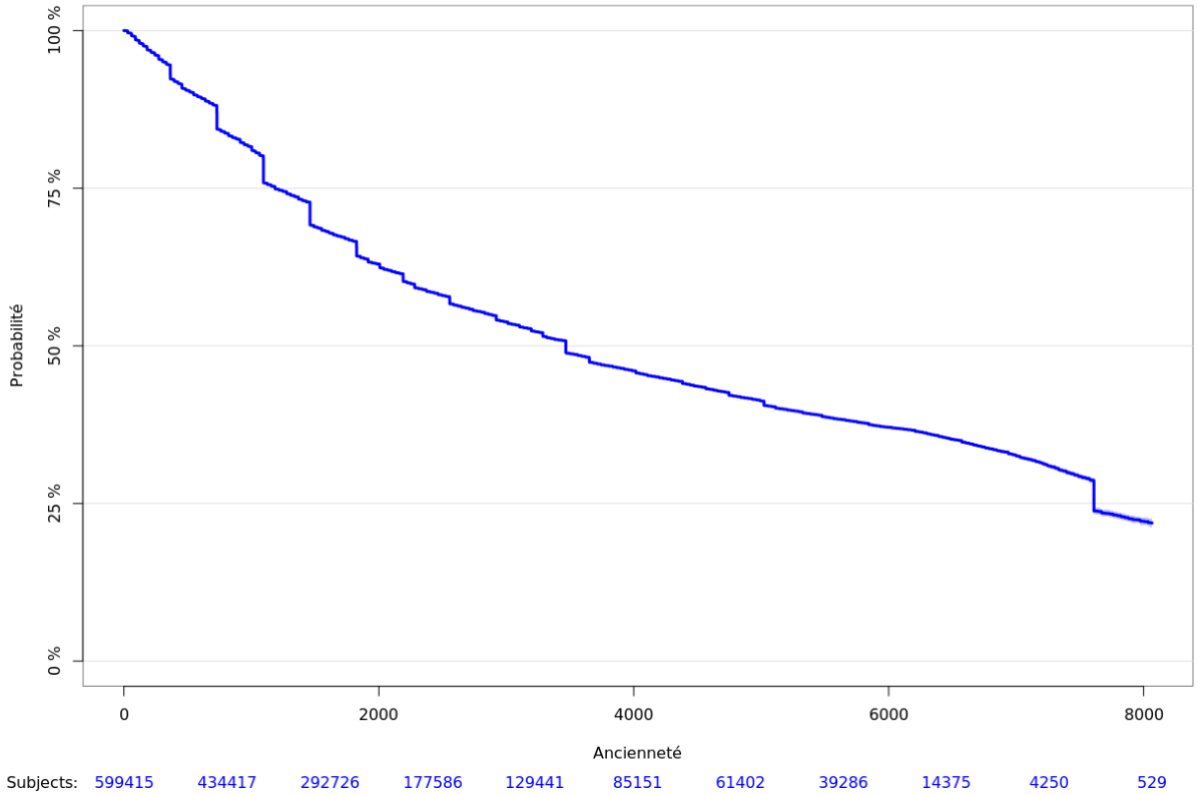


FIGURE 1.1 – Estimation non paramétrique de la fonction de survie du portefeuille par Kaplan-Meier

probabilité qu'on contrat reste en portefeuille diminue de manière presque linéaire suivant l'ancienneté du contrat.
Nous remarquons aussi que la probabilité qu'un contrat d'assurance santé reste encore portefeuille après 22 ans d'ancienneté est presque égale à 0.2.
Les sauts observés sur le graphique sont dus à des pics de résiliations annuelles (tous les 365 jours).

Nous avons donc zoomé ce graphique pour mieux observer les sauts pour des ancienetés inférieures à 1095 jours, soit 3ans.

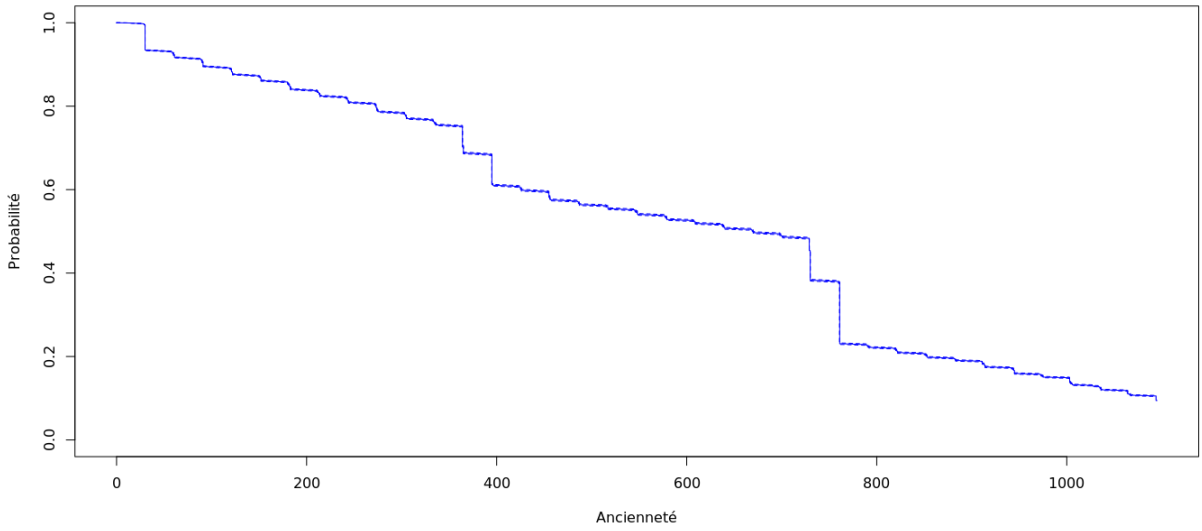


FIGURE 1.2 – Estimation de la fonction de survie du portefeuille par Kaplan-Meier pour les anciennetés inférieures à 3 ans

Ce zoom nous a permis de mieux observer les sauts de la fonction de survie tout les 12 mois. Cela peut s'expliquer par le fait que la plupart des contrats prennent effet le 1^{er} Janvier et peuvent être résilier le 31 Décembre, la date d'échéance.

Ces premiers graphiques nous ont permis d'identifier des sauts de cette estimation de la fonction de survie tous les 3655 jours d'ancienneté, soit tous les 12 mois.

∗ **Fonction de risque du portefeuille**

Nous allons représenter ici la fonction de risque définie dans **4.1.1**. Elle représente le taux de résiliation instantané entre t et l'instant d'après.

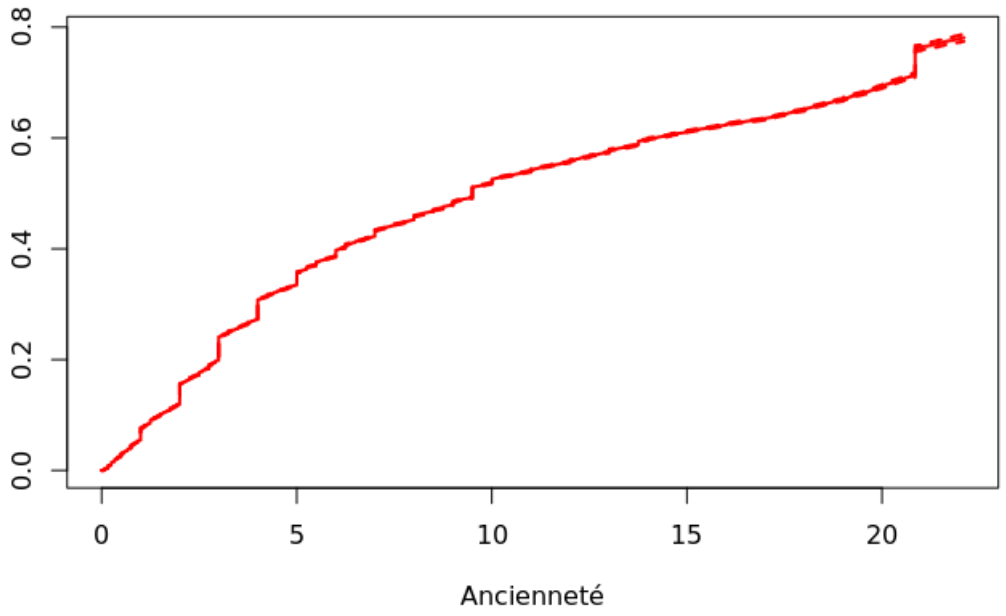


FIGURE 1.3 – Taux de résiliations instantané du portefeuille en années

On observe sur le graphique ci-dessus que la courbe croit très vite entre 0 et 10 ans et augmente lentement entre 10 et 20 ans.

— Le taux de résiliations durant la première année est de 6%, entre l'ancienneté 0 et 1 an.
— Entre l'ancienneté 0 et 5 ans, la probabilité de résiliation est 35%.
— Durant la dixième année, le taux de résiliation est de 52%.
— Et enfin entre 0 et 20 ans, le taux de résiliation est de 69%.

1.2.3 Analyse de la fonction de survie selon certains caractéristiques

L'enjeu de cette section est de comparer les fonctions de survie entre plusieurs groupes. Nous avons recodé certains de nos variables comme la classe d'âge en regroupant en trois classes d'âge (0-29, 30-62 et 63+), la variable Nombre enfants (avec enfants, sans enfants), la variable situation familiale (célibataire, marié et autres) et la variable niveau de garantie (base, option).

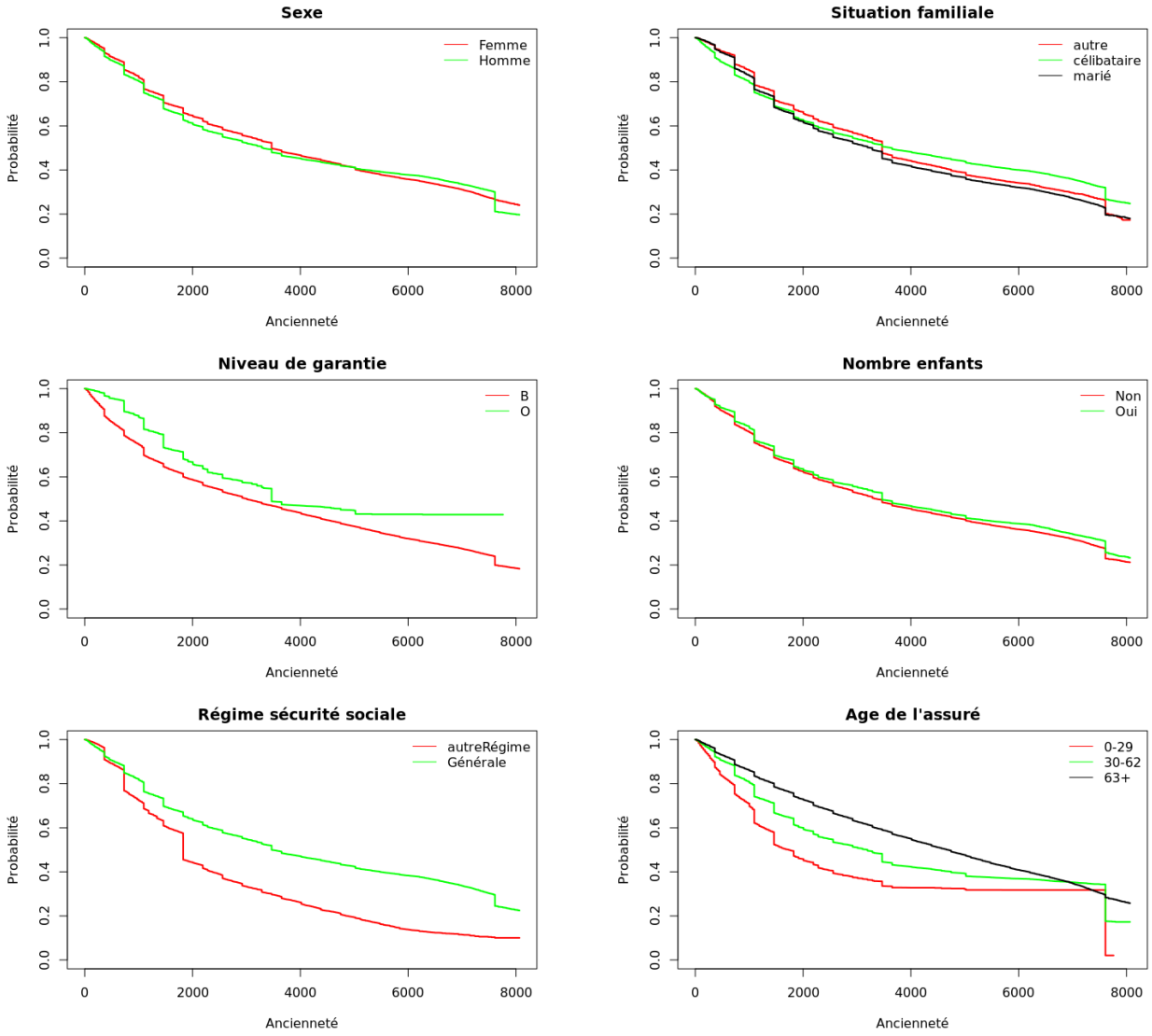


FIGURE 1.4 – Estimateur de Kaplan-Meir de la fonction de survie pour chacune des variables

L'effet de nos variables explicatives sur la durée de résiliation des contrats santé est illustré sur la figure ci-dessus.

• **Sexe** : Nous constatons une différence entre la courbe de survie des hommes et celle des femmes. La courbe de survie des femmes est au dessus de celle des hommes. Donc la durée de vie des contrats chez les femmes est plus grande. La statistique du Log-rank avec une valeur de 261, confirme une différence significative entre les deux courbes avec une p-value $p < 0.05$.

• **Situation familiale** : On observe une différence entre les trois courbes. Là encore la statistique du Log-rank confirme une différence significative entre les groupes liés à la situation familiale avec une p-value $p < 0.05$. Les assurés mariés ont tendance à résilier plus vite leurs contrats par rapport aux autres.

• **Niveau de garantie** : Nous constatons une différence entre les deux courbes de survie. Les contrats avec un niveau de garantie Option 2 ont une probabilité de survie plus importante que les contrats avec un niveau de garantie Base. Les contrats avec un niveau de garantie de base ont tendance a être résilié plus rapidement que les autres.
Il faut noter que le niveau **Base** s'agit d'une formule de couverture basique. Le niveau **Option 1** offre de meilleures couvertures sur les soins généraux, l'hospitalisation, les frais dentaires et les frais optiques et le niveau **Option 2** correspond à une formule de confort supérieure à la précédente, elle propose des garanties encore plus élevées sur les postes précédemment cités.

• **Nombre d'enfants** : Nous constatons aussi une différence entre les deux courbes de survie. La statistique du Log-rank prenant la valeur 238, confirme une différence significative entre les deux courbes avec une p-value $p < 0.05$. La courbe correspondant aux assurés ayant des enfants laisser à penser que ce groupe a tendance à résilier moins rapidement leurs contrats.

• **Classe d'âge** : La courbe de survie de la classe d'âge 0-29 croît plus vite par rapport aux courbes des autres classe d'âge. Cela laisse à penser aussi que la probabilité de survie de cette classe est plus petite que celle des autres groupes. Les assurés âgés (retraités) résilient moins vite que les autres plus jeune.

• **Régime** : Nous constatons que La courbe de la fonction de survie des assurés avec les autres régimes décroît plus vite. Donc les assurés ayant affiliés au régime générale obligatoire ont tendance à résilier moins vite leurs contrats.

Conclusion

L'estimation de la fonction de survie par Kaplan-Meir nous a permis d'avoir une première idée de la durée de survie des contrats de notre portefeuille.
• En effet, nous avons pu remarquer la diminution de la survie du portefeuille qui est due à des pics de résiliations à chaque 12 mois d'ancienneté (365 jours).
• Les assurés qui sont en retraites résilient moins vite que les autres assurés plus jeunes. De même, La probabilité de survie des contrats avec une formule de garantie de Base est moins importante par rapport aux formules de garanties Option.
• Le taux de résiliation est plus élevé durant la dixième année, entre 0 à 10 ans d'ancienneté. Tandis que, entre l'ancienneté 10 à 20 ans, le taux de résiliation est de 17%.

1.3 Modèle semi-paramétrique

Les modèles semi-paramétriques permet d'estimer une fonction de risque de la variable durée de vie T en tenant compte de l'influence des facteurs exogènes. Ce sont des modèles dits à risques proportionnels. Pour ces modèles, on retrouve la notion de hasard de base, qui donne une forme générale de hasard et est commune à tous les individus.

1.3.1 Le modèle de cox

Le modèle de Cox à risques proportionnels (Cox, 1972) est un modèle de régression statistique couramment utilisé dans la recherche médicale pour étudier l'association entre le temps de survie des patients et une ou plusieurs variables prédictives.
Son but est d'évaluer simultanément l'effet de plusieurs facteurs sur la survie. En d'autres termes, cela nous permet ici d'examiner comment nos différentes variables (caractéristiques des contrats santé) influencent le taux de résiliation à un moment donné.
Le modèle de Cox est exprimé par la fonction de risque notée $h(t)$. Cette fonction de risque peut être interprétée comme le risque de résiliation au temps t.
Il peut être estimé comme suit :

$$h(t) = h_0(t) \times \exp(b_1X_1 + b_2X_2 + ... + b_kX_k)$$

où,

— t est l'ancienneté du contrat
— $h(t)$ est la fonction de hasard déterminé avec un ensemble k de covariables

— les coefficients $b_1, b_2, ..., b_k$ mesurent l'impact des covariables (c'est à dire la taille de l'effet de résiliation).

— h_0 est appelé fonction de hasard de base

Les quantités $\exp(b_i)$ sont appelés rapports de risque (HR). Si b_i est positif ou de manière équivalente $HR > 1$ alors le risque de résiliations augmente et donc la durée de survie des contrats diminue.

Autrement dit, un $HR > 1$ indique une covariable positivement associée à la probabilité de résiliation, et donc négativement associé à la durée de survie.

— $HR = 1$: Aucun effet

— $HR < 1$: Réduction du risque

— $HR > 1$: Augmentation du risque

C'est un modèle à risque proportionnels c'est à dire que le rapport du risque défini par $HR = \frac{h(t|Y_i.\theta)}{h(t|Y_j.\theta)} = \frac{\exp(t|Y_i.\theta)}{\exp(t|Y_j.\theta)}$ est constant pour tout t.

Le modèle de cox repose sur deux hypothèses :

1. **L'hypothèse des risques proportionnels** c'est à dire que le rapport des risques instantanés entre deux individus est indépendant du temps ;
2. **L'hypothèse de log-linéarité** : c'est à dire que le logarithme du risque instantané est une fonction linéaire autrement dit, il existe une relation log-linéaire entre la fonction de risque et les covariables.

1.3.2 Régression de cox univariable

L'enjeu dans cette partie est d'étudier l'influence des variables (caractéristiques des contrats santé) sur la durée de vie des contrats. Nous avons commencé à faire des analyses de cox univariables pour chacune de nos variables et ensuite nous allons ajusté ces analyses de cox multivariées pour décrire l'influence des variables sur la survie des contrats.

Nous avons obtenu les résultats suivants pour la variable **Nombre enfant** :

Interprétation des résultats

Les résultats de la régression ci-dessus peuvent être interpréter comme suit :

1. **Signification statistique** : la colonne marquée "z" donne la valeur de la statistique de Wald. Il correspond au rapport de chaque coefficient de régression à son erreur type ($z = coef/se(coef)$). Cette statistique évalue β le coefficient d'une variable donnée

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Nombre_Enfants, data = df_MS)

  n= 599415, number of events= 254146

                      coef exp(coef)  se(coef)     z Pr(>|z|)
Nombre_EnfantsOui -0.060564  0.941233  0.004064 -14.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
Nombre_EnfantsOui    0.9412      1.062    0.9338    0.9488

Concordance= 0.508  (se = 0.001 )
Likelihood ratio test= 223  on 1 df,   p=<2e-16
Wald test            = 222.1  on 1 df,   p=<2e-16
Score (logrank) test = 222.1  on 1 df,   p=<2e-16
```

est statistiquement significativement différent de 0. Donc d'après le résultats ci-dessus, nous pouvons conclure que la variable **Nombre_Enfant** à des coefficients statistiquement significatifs.

2. **Les coefficients de régression** : la deuxième caractéristique à prendre en compte dans les résultats du modèle est le signe des coefficients de régression (coef). Un signe positif signifie que le danger (risque de résiliation) est plus élevé, et donc le pronostic moins bon, pour les sujets ayant des valeurs plus élevées de cette variable. La variable **Nombre_Enfants** à deux modalités : "oui" : pas d'enfants et "non" : au moins un enfant. Le résultat du modèle de Cox donne le rapport de risque (HR) entre les deux groupes, c'est-à-dire entre les assurés qui n'ont d'enfants et ceux qui ont au moins un enfant.
   Le coefficient β ici est égale à -0.06 indique le fait d'avoir un enfant diminue ($\beta < 0$) le risque de résiliation c'est à dire que les assurés ayant des enfants ont un risque de résiliation plus faible par rapport aux assurés sans enfants.
3. **Rapports de risque** : les coefficients exponentiels ($\exp(coef) = \exp(-0,06) = 0,94$) qu'on appelle également rapports de risque, donnent l'ampleur de l'effet des covariables. Par exemple ici, le fait d'avoir des enfants réduit le risque d'un facteur de 0.94.
4. **Intervalles de confiance des rapports de risque** : le résultat du modèle donne également des intervalles de confiance supérieurs et inférieurs à 95% pour le rapport de risque ($\exp(coef)$), borne inférieure à 95% = 0.9338, borne supérieure à 95% = 0.9488.
5. **Signification statistique globale du modèle** : enfin, le résultats donne des valeurs de p pour trois tests alternatifs de signification globale du modèle : le **test du rapport**

de vraisemblance, le **test de Wald** et les **statistiques du log-rang du score**. Ces trois méthodes sont asymptotiquement équivalentes. ils donneront des résultats similaires pour N assez grand et pour les petits N, ils peuvent différer quelque peu. Le test du rapport de vraisemblance a un meilleur comportement pour les petits échantillons, il est donc généralement préféré.

Nous avons ainsi effectuer ce calcul pour tout les autres variables de notre modèle.
Et nous concluons que nos variables ont des coefficients statistiquement significatifs au seuil de 5% c'est à dire que leurs p-value $p < 0.05$. (cf. **Annexe 1**)

1.3.3 Régression multivariable de Cox

L'enjeu de cette partie est de décrire comment les facteurs ont un impact conjoint sur la survie. Pour répondre à cette question, nous allons effectuer une analyse de régression multivariée de Cox. Nous avons inclus toutes nos variables ci-dessus dans le modèle et nous avons obtenu les résultats suivantes :

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Sexe + Niveau_garantie +
    Situation_familiale + Nombre_Enfants + Régime + classe_age,
    data = df_MS)

  n= 599415, number of events= 254146

                                   coef exp(coef)  se(coef)       z Pr(>|z|)
SexeHomme                      0.062635  1.064638  0.004013   15.61   <2e-16 ***
Niveau_garantieO              -0.786526  0.455424  0.004745 -165.77   <2e-16 ***
Situation_familialeCélibataire -0.078996  0.924043  0.006822  -11.58   <2e-16 ***
Situation_familialeMarié       0.100672  1.105914  0.007370   13.66   <2e-16 ***
Nombre_EnfantsOui             -0.067783  0.934463  0.004407  -15.38   <2e-16 ***
RégimeGénérale                -0.316592  0.728628  0.007615  -41.58   <2e-16 ***
classe_age30-62               -0.517503  0.596007  0.006612  -78.27   <2e-16 ***
classe_age63+                 -1.290500  0.275133  0.007512 -171.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                               exp(coef) exp(-coef) lower .95 upper .95
SexeHomme                         1.0646     0.9393    1.0563    1.0730
Niveau_garantieO                  0.4554     2.1958    0.4512    0.4597
Situation_familialeCélibataire    0.9240     1.0822    0.9118    0.9365
Situation_familialeMarié          1.1059     0.9042    1.0901    1.1220
Nombre_EnfantsOui                 0.9345     1.0701    0.9264    0.9426
RégimeGénérale                    0.7286     1.3724    0.7178    0.7396
classe_age30-62                   0.5960     1.6778    0.5883    0.6038
classe_age63+                     0.2751     3.6346    0.2711    0.2792
```

Nous constatons que toutes les covariables sont statistiquement significatif avec chacune un p-value ($p < 5\%$).

Nous avons ensuite calculer les statistiques et p-value de trois tests classiques (ratio de vraisemblance, Wald et score) pour estimer la significativité du modèle.

```
Concordance= 0.628  (se = 0.001 )
Likelihood ratio test= 44343  on 8 df,   p=<2e-16
Wald test            = 48691  on 8 df,   p=<2e-16
Score (logrank) test = 49387  on 8 df,   p=<2e-16
```

La p-value pour les trois tests est significative ($p < 0.05$), ce qui indique que le modèle est significatif. Ces tests évaluent l'hypothèse nulle selon laquelle tous les β sont égales à 0. Donc nous pouvons rejeter l'hypothèse nulle.

Interprétation des résultats

D'après les résultats ci-dessus, nous observons que :

• **Sexe** : nous avons un rapport de risque $HR = \exp(coef) = 1,06$. Les rapports de risque des covariables peuvent être interprétés comme des effets multiplicatifs sur le risque. Par exemple, en maintenant les autres covariables constantes, le fait d'être un assuré de sexe masculin augmente le risque de résiliation d'un facteur de $1,06$. Nous concluons que le fait d'être un homme est associé à une faible survie.

• **Nombre d'enfants** : Le fait d'avoir des enfants $\beta < 0$ diminue le risque de résiliation. Avec un rapport de risque de $HR = 0.93$, pour toutes choses égales par ailleurs, en maintenant les autres covariables constantes, les contrats des assurés n'ayant pas d'enfants ont un risque de résiliation plus élevé que celui des contrats des assurés avec enfants.

• **Niveau de garantie** : La valeur du ratio de risque pour la variable niveau de garantie est de 0.45, donc en considérant que les autres covariables sont constantes, nous pouvons dire que le risque de résiliation des contrats avec un niveau de garantie Option est 55% de moins que celui des contrats avec un niveau de garantie Base.

• **Situation familiale** : Le fait d'être un assuré de situation familiale célibataire $\beta < 0$ diminue le risque de résiliation alors que le fait d'être marié augmente le risque de résiliation.

• **Âge** : Les classes d'âge 30-62 et 63+ diminuent le risque de résiliation ($\beta < 0$). Donc en maintenant les autres covariables constantes, les assurés de la classe d'âge 0-29 ans ont un risque de résiliation plus élevé par rapport à la classe d'âge plus âgé.

• **Régime** : Nous voyons le coefficient $\beta < 0$, donc le fait d'être affilié au régime générale obligatoire diminue le risque de résiliation de 27%.

1.3.4 Vérification de la validité du modèle

Nous allons tester l'hypothèse de la proportionnalité des risques relatifs pour valider notre modèle. Cette hypothèse peut être vérifiée à l'aide de tests statistiques et de diagnostics graphiques basés sur les résidus de Schoenfeld.

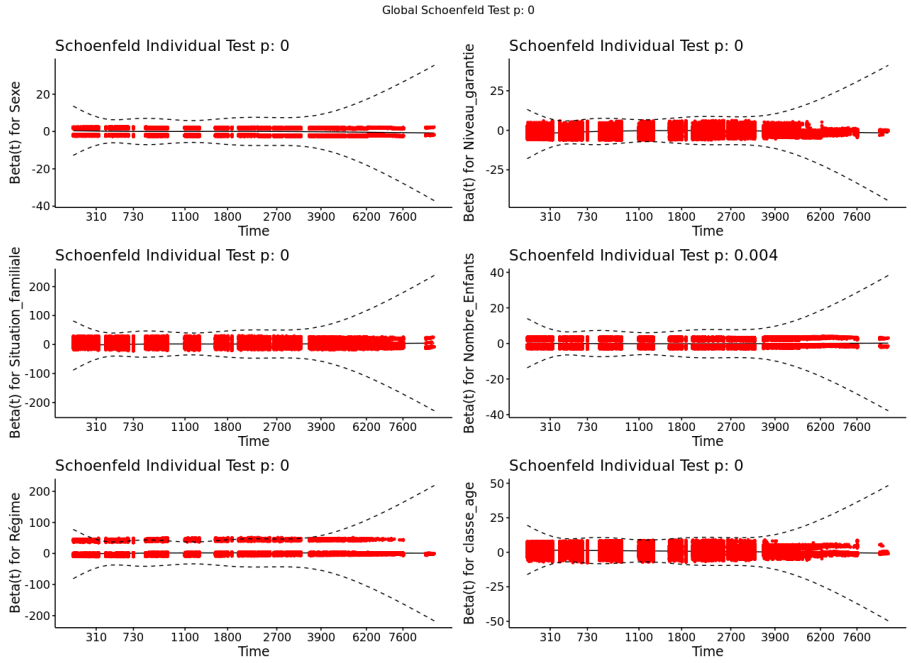


FIGURE 1.5 – Résidus de Schoenfeld

Nous obtenons ici, 6 graphiques pour nos six variables. Et, nous devons obtenir six courbes en traits continus qui sont le plus horizontal possible. Ce qui est globalement le cas ici. Donc, nous pouvons conclure que, que l'hypothèse des risques instantanés proportionnels est vérifiée.

Synthèse du modèle de Cox

Nous pouvons conclure que :

— Le fait d'être un homme augmente le risque de résiliation ;
— Le risque de résiliation diminue avec l'âge ;
— Les résiliations diminue avec le niveau de garantie ;
— Les résiliations diminue avec la composition familiale ;
— La situation familiale marié augmente le risque de résiliation

Ces conclusions sont cohérentes avec les résultats obtenus précédemment avec les courbes de survies.

Chapitre 2

La modélisation par la régression logistique

La régression logistique est un modèle statistique utilisé pour étudier la relation entre un ensemble de variables explicatives qui peuvent être quantitative ou qualitatives X_i et une variable cible à expliquée Y. Cette variable Y est le plus souvent binaire. Il s'agit d'un modèle linéaire généralisé utilisant la fonction logistique comme fonction de lien.

Un modèle de régression logistique permet également prédire la probabilité qu'un événement se produise (valeur 1) ou non (valeur 0) en optimisant les coefficients de régression. Ce résultat varie toujours entre 0 et 1. Un événement peut se produire lorsque la valeur prédite est supérieure à un seuil, mais pas lorsque la valeur est inférieure au même seuil.

L'enjeu dans cette parie est de modéliser la variable RESIL qui a deux modalités, 1 si le contrat est résilié et 0 si le contrat est en cours.

2.1 Théorie : Modèle linéaire généralisé

2.1.1 Modèle linéaire

L'intérêt du modèle est d'expliquer la variable Y à l'aide de n variables explicatives $X_1, X_2, ..., X_n$.

Le modèle linéaire est défini par la formule suivant :

$$Y = X^t\beta + \epsilon = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon$$

avec :

— Y la variable observée que l'on souhaite prédire et expliquée ;

— X_i, les variables explicatives appelés prédicteurs ;

— β_i les paramètres du modèle ;

— ϵ le terme d'erreur aléatoire appelé résidus et vérifie les propriétés suivantes : $E(\epsilon) = 0$ et $V(\epsilon) = \sigma^2$.

2.1.2 Modèle linéaire généralisé

Le modèle généralisé est constitué de trois composante :

• Distribution : Famille exponentielle

La variable Y à prédire est associé à une loi de probabilité qui à appartient à la famille exponentielle donc elle est de la forme :

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

avec :

— θ paramètre naturel ;

— ϕ paramètre de dispersion ;

— les fonctions a, b et c sont fixés.

• Prédicteur linéaire

Les variables $X_1, X_2, ..., X_n$ sont exprimées sous la forme $\beta_0 + \sum_{i=1}^{n} \beta_j X_j$, correspondent à la composante déterministe du modèle.

• Lien

Cette troisième composante exprime une relation fonctionnelle entre les deux premières composantes cités ci-dessus.
Soit $\mu_i = E(Y_i); i = 1, ..., n$, et on pose : $\eta = g(\mu_i) \quad i = 1, ..., n$ et g est appelé la fonction lien. Ce qui revient à écrire $g(\mu_i) = X_i^t\beta$. (cf. [8])

2.1.3 La régression logistique

Considérons $X = (X_1, X_2, ..., X_p)$ p variables explicatives de Y.
Soit $\pi(x) = E(Y|(X_i = x_1))$. Nous avons la relation linéaire de régression linéaire suivante :

$$g(\pi) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

avec g la fonction de lien.

Comme Y suit une loi de Bernouilli, on a $\pi(x) = P(Y = 1|X_i)$, la probabilité de survenance de l'évènement de résiliation sachant le facteur explicatif X_i soit égal à x.
Nous utilisons la fonction **logit** comme la fonction lien définit sur $[0, 1]$ et à valeurs dans R par :

$$g(x) = \ln(\frac{x}{1-x})$$

.

Ainsi

$$g(\pi) = \ln(\frac{\pi}{1-\pi}) \quad donc \quad \pi = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}$$

• Odds Ratio

Un odds ratio représente la probabilité qu'un résultat se produise compte tenu d'une exposition particulière, par rapport à la probabilité que le résultat se produise en l'absence de cette exposition. Il mesure l'évolution du rapport des probabilités d'apparition de l'événement $Y = 1$ par rapport à l'événement $Y = 0$ lorsque X_i (variable explicative) passe de x à $x + 1$ pour une variable continue et passe d'une modalité à une autre pour une variable qualitative.

$$OR = \frac{\pi(x+1)/[1 - \pi(x+1)]}{\pi x/[1 - \pi(x)]} = e^{\beta_i}$$

Dans le cas où X est binaire, nous :

$$P(Y = 1|X_i = 1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \quad et \quad P(Y = 1|X_i = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$OR = \frac{P(Y = 1|X_i = 1)P(Y = 0|X_i = 1)}{P(Y = 1|X_i = 0)P(Y = 1|X_i = 0)} = e^{\beta_i}$$

2.2 Résultats de la modélisation

2.2.1 Sélection des variables

Une analyse univariée et une étude d'association préalables nous ont permis de faire une première présélection des variables à intégrer dans le modèle. Par la suite, nous avons étudié l'importance de chacune de ces variables. Ensuite, les effets des interactions entre variables ont été testés pour une seconde présélection de variables à introduire dans la régression logistique. Par conséquent, plusieurs modèles ont été estimés. Nous avons utilisé le critères AIC (Akaike Information Criteria) pour la sélection des variables. Et nous avons retenu les variables suivantes comme étant les plus significatives de la variable explicative RESIL : Nombre enfants, Classe d'âge, Sexe, Situation ,familiale, Régime, Niveau de garantie, Ancienneté.

Nous avons donc affiner notre modèle en ne prenant en compte que ces variables.

2.2.2 Analyse des résultats du modèle retenu

- **Significativité des variables et coefficients**

Nous présentons d'abord dans le tableau ci-dessous les résultats du test de significativité. Les résultats dans ce tableau sont les coefficients estimés et les statistiques du test pour chacune des modalités de nos variables.

```
Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                     2.161019   0.029160   74.109  < 2e-16 ***
SexeHomme                       0.192923   0.006959   27.724  < 2e-16 ***
Niveau_garantie0               -1.248683   0.008625 -144.769  < 2e-16 ***
Situation_familialecélibataire -0.213052   0.011854  -17.973  < 2e-16 ***
Situation_familialemarié        0.068970   0.012972    5.317 1.05e-07 ***
Nombre_EnfantsOui              -0.079104   0.007559  -10.464  < 2e-16 ***
RégimeRégime agricole          -1.708174   0.030488  -56.028  < 2e-16 ***
RégimeRégime général           -1.017002   0.025080  -40.550  < 2e-16 ***
classe_age30-62                -0.095056   0.011536   -8.240  < 2e-16 ***
classe_age63+                  -0.213010   0.013623  -15.637  < 2e-16 ***
AncienneténcObs2               -1.393099   0.008633 -161.374  < 2e-16 ***
AncienneténcObs3               -1.728873   0.013847 -124.852  < 2e-16 ***
AncienneténcObs4               -2.718153   0.020728 -131.134  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 2.1 – Test de signicativité

On peut alors conclure que l'ensemble des variables sont significatives avec $p < 5\%$.

- **Importance variable** :

Nous avons également calculer l'importance de chaque variable prédictive dans le modèle en utilisant la fonction **varImp** du package **caret** de R :

	Overall <dbl>
SexeHomme	27.724224
Niveau_garantieO	144.769500
Situation_familialecélibataire	17.972858
Situation_familialemarié	5.317014
Nombre_EnfantsOui	10.464242
RégimeRégime agricole	56.027783
RégimeRégime général	40.549847
classe_age30-62	8.240036
classe_age63+	15.636518
AncienneténcObs2	161.373958

Les valeurs les plus élevées indiquent plus d'importance. L'ancienneté est de loin la variable prédictive la plus importante, suivi du du niveau de garantie.

- **Valeurs VIF** (facteur d'inflation de la variance) :

Les valeurs VIF mesure la corrélation et la force de la corrélation entre les variables prédictives dans un modèle de régression. Nous avons également calculer les valeurs VIF de chaque variable du modèle pour voir si la multicolinéarité cause problème :

	GVIF	Df	GVIF^(1/(2*Df))
Sexe	1.016445	1	1.008189
Niveau_garantie	1.559420	1	1.248768
Situation_familiale	1.065336	2	1.015948
Nombre_Enfants	1.172125	1	1.082647
Régime	1.040032	2	1.009861
classe_age	1.814746	2	1.160657
Ancienneté	1.217068	3	1.033283

Les valeurs VIF supérieures à 5 indiquent une multicolinéarité sévère en règle générale. Puisqu'aucune des variables prédictives de notre modèle n'a un VIF supérieur à 5, nous pouvons supposer que la multicolinéarité n'est pas un problème dans notre modèle.

2.2.3 Validation du modèle

Enfin, nous pouvons analyser les performances de notre modèle sur le jeux de données de test en calculant la matrice de confusion et en évaluant l'air sous la courbe ROC.

• Matrice de confusion

Il s'agit d'une matrice carrée $2x2$. Elle mesure la qualité de prédiction d'un modèle. Les deux lignes correspondent aux prédictions du modèle et les deux colonnes aux valeurs réelles.

Population		Valeur réelle	
		0	1
Valeur prédite	0	Vrai Négatif (VN)	Faux Négatif (FN)
	1	Faux Positif (FP)	Vrai Positif (VP)

Nous avons donc, avec notre modèle, obtenu les résultats suivants :

Base test	Prédit		Total
	1	0	
1	68,58%	31,42%	100,00%
0	27,25%	72,75%	100,00%

Base app	Prédit		Total
	1	0	
1	68,79%	31,21%	100,00%
0	27,06%	72,94%	100,00%

FIGURE 2.2 – Matrices de confusion

Sur la base d'apprentissage, le modèle prédit 68.79% des contrats résiliés contre 72.94% des contrats non résiliés. Et sur la base test nous avons 68.58% de bonnes prédiction pour les contrats résiliés et 72.75% pour les contrats non résiliés.

• Air sous la courbe ROC

La courbe ROC (Receiver Operating Characteristic) est une autre mesure de performance d'un modèle. Elle trace les valeurs du taux de vrais positifs et du taux de faux positifs. Elle donne le taux de vrai positifs en fonction du taux faux positifs.
Plus l'AUC (aire sous la courbe) est élevée, plus la précision est grande, notre modèle est capable de prédire les résultats.

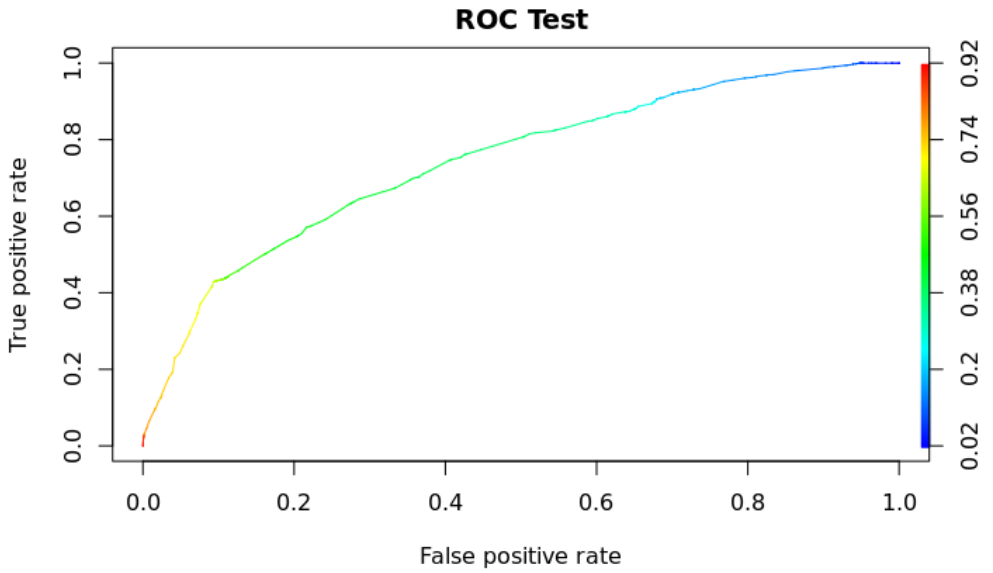


FIGURE 2.3 – Courbe ROC de l'échantillon test du modèle

Nous avons calculé l'aire sous la courbe ROC pour l'échantillon test qui est égale à 0.7426. En comparant cette valeur avec la valeur calculée à partir de l'échantillon d'apprentissage (0.7435), les prédictions du modèle obtenu peuvent être considérées comme fiables et les résultats seraient stables d'un échantillon à l'autre.

Comme l'AUC pour l'échantillon test est de 0.7426, ce qui est assez élevé. Nous pouvons dire que notre modèle fait un bon travail pour prédire si un assuré va résilier son contrat ou pas.

2.2.4 Interprétation des résultats obtenu avec les odds ratio du modèle

Nous avons ensuite calculé les odds ratio de notre modèle qui sont obtenus à partir des coefficients du modèle.
Ces odds ratio permettent de connaître le sens et la taille des effets des variables explicatives du modèle.

```
Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                     2.161019   0.029160   74.109  < 2e-16 ***
SexeHomme                       0.192923   0.006959   27.724  < 2e-16 ***
Niveau_garantie0               -1.248683   0.008625 -144.769  < 2e-16 ***
Situation_familialecélibataire -0.213052   0.011854  -17.973  < 2e-16 ***
Situation_familialemarié        0.068970   0.012972    5.317 1.05e-07 ***
Nombre_EnfantsOui              -0.079104   0.007559  -10.464  < 2e-16 ***
RégimeRégime agricole          -1.708174   0.030488  -56.028  < 2e-16 ***
RégimeRégime général           -1.017002   0.025080  -40.550  < 2e-16 ***
classe_age30-62                -0.095056   0.011536   -8.240  < 2e-16 ***
classe_age63+                  -0.213010   0.013623  -15.637  < 2e-16 ***
AncienneténcObs2               -1.393099   0.008633 -161.374  < 2e-16 ***
AncienneténcObs3               -1.728873   0.013847 -124.852  < 2e-16 ***
AncienneténcObs4               -2.718153   0.020728 -131.134  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 2.4 – Odds ratio du modèle avec les intervalles de confiance à 95%

Nous pouvons observer toutes choses étant égales par ailleurs que :

1. Les hommes ont un taux de résiliation plus élevé par rapport aux femmes.
2. Le taux de résiliation est élevé d'autant plus que le niveau de garantie sera faible.
3. Les assurés célibataires ont un taux de résiliations moins important par rapport aux autres. Le fait d'être marié implique une probabilité plus forte.
4. Les assurés les plus plus âgé sont ceux qui résilient le moins.
5. Les assurés n'ayant pas d'enfants résilient plus que les assurés avec des enfants.
6. Les contrats des assurés dont leur régime de sécurité sociale est le régime général sont les contrats ayant un taux de résiliation plus faibles.
7. Le taux de résiliations des contrats avec une ancienneté inférieur à 5 ans ont un résiliation plus forte. Et dès l'ancienneté dépasse 15 ans, la probabilité de résiliation est plus faible.

Conclusion

L'objet dans ce mémoire a été d'étudier la durée de vie des contrats d'assurance santé. Nous avons ainsi utilisé deux approches, d'une part les modèles de survie pour estimer la probabilité qu'un assuré résilie son contrat et d'autres nous avons modéliser le taux de résiliation par une méthode basée sur la régression logistique.

L'enjeu principal de ces modélisations est d'identifier les profils de clients les plus susceptibles de résilier leurs contrats d'assurance santé. Généralement les résultats obtenus confirment des faits observés ou des idées intuitives sur les phénomènes de résiliation.
Par exemple, avec l'estimation de la durée de vie des contrats par les modèles de survie, nous avons pu voir que les assurés non retraités résiliaient plus leurs contrats par rapport aux retraités, et que les contrats avec un niveau de garantie Base étaient plus résiliés que les contrats avec un niveau de garantie Option. Et ces résultats ont été confirmé par la modélisation par régression logistique.

L'étude a permis d'établir les effets des variables, à savoir si chaque variable a une influence positive ou négative sur le risque de résiliation. Elle nous a permis aussi de constater que la durée de vie des contrats n'est pas très élevée, que la probabilité de survie diminue rapidement suivant l'ancienneté des contrats et le risque de résilier son contrat diminue au fur et à mesure que l'ancienneté du contrat augmente.

Les résultats obtenus pourraient servir à d'autres services dans la mise en place d'actions ciblées sur les profils types qui résilient le plus leurs contrats, dans le but de limiter ces risques de résiliations et fidéliser les clients afin de pouvoir conserver ces assurés.

Notre étude comporte cependant certaines limites :
• Nous avons considéré que les assurés principaux sans leurs bénéficiaires. Ces derniers pourraient peut-être fournir des informations importantes.

• Nous avons utilisé que des variables ne dépendant pas du temps. Nous n'avons pas tenu compte des changements des assurés ayant lieu au cours de la période d'observation. Nous avons considéré que leur profil au début de leur contrat.

En plus des nouvelles études qui pourraient être mené dans le futur pour corriger ces limites, il pourrait être intéressant d'étendre notre étude par rapport aux cotisations et aux prestations liées aux contrats.

Bibliographie

[1] **Dirk F. Moore** (2016), *Applied Survival Analysis Using R*, Springer.

[2] **PLANCHET Frédéric** (2021), *Modèle de durée, application actuarielle*, http ://www.ressources-actuarielles.net.

[3] **Dominique Grandguillot** (2017), *L'essentiel du droit de la Sécurité Sociale*, Gualino éditeur, 16è édition.

[4] **Institut français de l'audit et du contrôle internes - IFACI** (2012), *La délégation de gestion en assurances de personnes : Pistes pour un contrôle interne efficace.*

[5] **Direction de la sécurité sociale** (2021), Les chiffres clés de la sécurité sociale 2020, https ://www.securite-sociale.fr/la-secu-cest-quoi/chiffres-cles.

[6] **Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)** (2017), *Les dépenses de santé en 2016. Résultats des comptes de la santé*, https ://drees.solidarites-sante.gouv.fr/sites/default/files/2020-10/pano_cns_2017.pdf

[7] **Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)** (2021),*La situation financière des organismes complémentaires assurant une couverture santé*, https ://drees.solidarites-sante.gouv.fr/.

[8] **Philippe BESSE**, *Introduction au modèle linéaire général*, Cours universitaire, https ://www.math.univ-toulouse.fr/˜besse/Wikistat/pdf/st-m-modlin-mlg.pdf.

[9] **Yohann LE FAOU** (2020), *Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé.* Thèse de doctorat, Sorbone Université.

[10] **Audrey PEYRILLER** (2019), *Prédiction des résiliations en santé individuelle.* Mémoire, Institut du risque management.

[11] **Manon BREFORT** (2018), *Analyse technique des résiliations dans le cadre de rétention commerciale.* Mémoire, ISFA (Institut de Science Financière et d'Assurances).

[12] **Aurélien LATOUCHE** (2022), *Analyse de survie*, Cours CNAM.

[13] **Francis MAHUT** (2020), *Délégation de gestion assurance santé, mutuelle et prévoyance*, https ://www.eficiens.com/delegation-de-gestion-assurance/.

[14] https ://www.gfpfrance.com/

[15] https ://www.axa.fr/complementaire-sante/regime-securite-sociale.html

Table des figures

Annexe 1 : Régression univariable de cox

Sexe

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Sexe, data = df_MS)

  n= 599415, number of events= 254146

              coef exp(coef) se(coef)     z Pr(>|z|)
SexeHomme 0.064551  1.066680 0.003978 16.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
SexeHomme     1.067     0.9375     1.058     1.075

Concordance= 0.514  (se = 0.001 )
Likelihood ratio test= 263.6  on 1 df,   p=<2e-16
Wald test            = 263.3  on 1 df,   p=<2e-16
Score (logrank) test = 263.4  on 1 df,   p=<2e-16
```

Niveau de garantie

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Niveau_garantie, data = df_MS)

  n= 599415, number of events= 254146

                    coef exp(coef)  se(coef)      z Pr(>|z|)
Niveau_garantieO -0.336032  0.714600  0.004085 -82.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
Niveau_garantieO    0.7146      1.399    0.7089    0.7203

Concordance= 0.558  (se = 0.001 )
Likelihood ratio test= 6868  on 1 df,   p=<2e-16
Wald test            = 6767  on 1 df,   p=<2e-16
Score (logrank) test = 6828  on 1 df,   p=<2e-16
```

Régime

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Régime, data = df_MS)

  n= 599415, number of events= 254146

                   coef exp(coef)  se(coef)      z Pr(>|z|)
RégimeGénérale -0.509290  0.600922  0.007503 -67.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
RégimeGénérale    0.6009      1.664    0.5921    0.6098

Concordance= 0.514  (se = 0 )
Likelihood ratio test= 4009  on 1 df,   p=<2e-16
Wald test            = 4607  on 1 df,   p=<2e-16
Score (logrank) test = 4707  on 1 df,   p=<2e-16
```

Situation familiale

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Situation_familiale,
    data = df_MS)

  n= 599415, number of events= 254146

                                    coef exp(coef) se(coef)      z Pr(>|z|)
Situation_familialecélibataire 0.037357  1.038064 0.006758  5.528 3.24e-08 ***
Situation_familialemarié       0.109331  1.115532 0.007320 14.937  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                               exp(coef) exp(-coef) lower .95 upper .95
Situation_familialecélibataire     1.038     0.9633     1.024     1.052
Situation_familialemarié           1.116     0.8964     1.100     1.132

Concordance= 0.503  (se = 0.001 )
Likelihood ratio test= 330.6  on 2 df,   p=<2e-16
Wald test            = 333.1  on 2 df,   p=<2e-16
Score (logrank) test = 333.3  on 2 df,   p=<2e-16
```

Nombre enfants

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ Nombre_Enfants, data = df_MS)

  n= 599415, number of events= 254146

                     coef exp(coef)  se(coef)      z Pr(>|z|)
Nombre_EnfantsOui -0.060564  0.941233  0.004064 -14.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
Nombre_EnfantsOui    0.9412      1.062    0.9338    0.9488

Concordance= 0.508  (se = 0.001 )
Likelihood ratio test= 223  on 1 df,   p=<2e-16
Wald test            = 222.1  on 1 df,   p=<2e-16
Score (logrank) test = 222.1  on 1 df,   p=<2e-16
```

Classe d’âge

```
Call:
coxph(formula = Surv(Ancienneté, RESIL) ~ classe_age, data = df_MS)

  n= 599415, number of events= 254146

                    coef exp(coef)  se(coef)       z Pr(>|z|)
classe_age30-62 -0.419459  0.657403  0.006269  -66.91   <2e-16 ***
classe_age63+   -0.741305  0.476492  0.006789 -109.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
classe_age30-62    0.6574      1.521    0.6494    0.6655
classe_age63+      0.4765      2.099    0.4702    0.4829

Concordance= 0.567  (se = 0.001 )
Likelihood ratio test= 12115  on 2 df,   p=<2e-16
Wald test            = 12616  on 2 df,   p=<2e-16
Score (logrank) test = 12913  on 2 df,   p=<2e-16
```

Annexe 2 : Test du Log-Rank

Sexe

```
Call:
survdiff(formula = Surv(Ancienneté, RESIL) ~ Sexe, data = df_MS)

                N Observed Expected (O-E)^2/E (O-E)^2/V
Sexe=Femme 296095   119323   123363       132       261
Sexe=Homme 303320   134823   130783       125       261

 Chisq= 261  on 1 degrees of freedom, p= <2e-16
```

Situation familiale

```
Call:
survdiff(formula = Surv(Ancienneté, RESIL) ~ Situation_familiale,
    data = df_MS)

                                 N Observed Expected (O-E)^2/E (O-E)^2/V
Situation_familiale=autre        60469    25420    26838      74.9      84.8
Situation_familiale=célibataire 385729   158347   160673      33.7      92.8
Situation_familiale=marié       153217    70379    66635     210.3     288.9

 Chisq= 323  on 2 degrees of freedom, p= <2e-16
```

Niveau de garantie

```
Call:
survdiff(formula = Surv(Ancienneté, RESIL) ~ Niveau_garantie,
    data = df_MS)

                          N Observed Expected (O-E)^2/E (O-E)^2/V
Niveau_garantie=B 283678   150670   129890      3324      7073
Niveau_garantie=O 315737   103476   124256      3475      7073

 Chisq= 7073  on 1 degrees of freedom, p= <2e-16
```

Nombre enfants

```
Call:
survdiff(formula = Surv(Ancienneté, RESIL) ~ Nombre_Enfants,
    data = df_MS)

                           N Observed Expected (O-E)^2/E (O-E)^2/V
Nombre_Enfants=Non 343509   153636   149839      96.2       238
Nombre_Enfants=Oui 255906   100510   104307     138.2       238

 Chisq= 238  on 1 degrees of freedom, p= <2e-16
```

Régime

```
Call:
survdiff(formula = Surv(Ancienneté, RESIL) ~ Régime, data = df_MS)

                              N Observed Expected (O-E)^2/E (O-E)^2/V
Régime=Autres régimes   14861    11100     5659      5231      5418
Régime=Régime agricole  23757     8290     6449       525       551
Régime=Régime général  560797   234756   242038       219      4681

 Chisq= 6059  on 2 degrees of freedom, p= <2e-16
```

Classe d'âge

```
Call:
survdiff(formula = Surv(Ancienneté, RESIL) ~ classe_age, data = df_MS)

                          N Observed Expected (O-E)^2/E (O-E)^2/V
classe_age=0-29   67924    31173    19396      7151      7921
classe_age=30-62 349013   143353   134076       642      1399
classe_age=63+   182478    79620   100674      4403      7638

 Chisq= 12754  on 2 degrees of freedom, p= <2e-16
```

Annexe 3 : Régression logistique

Courbes ROC