Mémoire présenté le :

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

Par : Khalil Datsi

Titre Le risque incapacité de travail : étude de l'absentéisme sur un portefeuille d'assurance collective

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

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

Membres présents du jury de l'Institut des Actuaires

Entreprise :

Nom : Generali Vie

Signature :

Membres présents du jury de l'ISFA

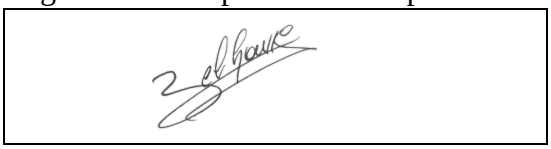
Directeur de mémoire en entreprise :

Nom : Selim Belhousse

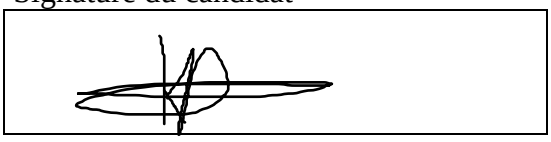
Signature :

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

Signature du responsable entreprise



Signature du candidat



Résumé

Ce mémoire d'actuaire explore la modélisation du risque incapacité et l'absentéisme dans le contexte de l'assurance collective en France. Il vise à approfondir la compréhension de ces risques et à contribuer à l'amélioration de leur gestion et de leur prévention pour les entreprises et les assureurs.

Ce mémoire débute par une étude du système de protection sociale en France, abordant l'histoire, l'organisation fonctionnelle et les offres proposées par le régime de base et le régime complémentaire. Il examine ensuite la prévoyance collective et le risque d'arrêt de travail, en considérant le contexte général et l'environnement juridique. L'absentéisme en entreprise en France est également analysé, en identifiant ses causes, ses conséquences et les méthodes de mesure et d'évaluation de son coût pour les entreprises.

Une base de données d'étude est construite à partir de différentes sources de données sur les sinistres, avec une description des étapes de traitement et de nettoyage des données ainsi que des variables d'étude à calculer. L'analyse descriptive du risque d'arrêt de travail est réalisée, en étudiant son évolution au cours des dernières années, l'impact de la saisonnalité et de la pandémie de COVID-19, ainsi que les tendances en matière de types d'arrêts et de risques psychosociaux chez les jeunes.

Puis, le mémoire aborde la modélisation de la durée des arrêts de travail à l'aide de modèles linéaires généralisés (GLM) et de modèles de durée. Il compare différentes distributions de durée, estime les paramètres des lois, valide les modèles et construit une loi de maintien en arrêt de travail.

Finalement, le mémoire s'intéresse à la prévention de l'absentéisme en entreprise en se basant sur la littérature spécialisée sur le sujet et en tentant de proposer une procédure de mise en place d’une stratégie de prévention en entreprise et d’évaluation des impacts de celle-ci.

En conclusion, ce mémoire offre un éclairage sur la modélisation du risque incapacité et l'absentéisme en assurance collective, en combinant une analyse théorique et empirique des données disponibles. Les résultats pourront servir pour les études de l’absentéisme à venir sur le portefeuille d’assurance collective de Generali Vie.

Abstract

This actuarial thesis explores the modeling of disability risk and absenteeism in the context of group insurance in France. It aims to deepen the understanding of these risks and contribute to the improvement of their management and prevention for both companies and insurers.

The thesis begins with a study of the social protection system in France, addressing its history, functional organization, and the offers proposed by the basic and complementary schemes. It then examines group foresight and the risk of work stoppage, considering the general context and legal environment. Absenteeism in companies in France is also analyzed, identifying its causes, consequences, and methods for measuring and evaluating its cost to companies.

A study database is constructed from various data sources on claims, with a description of the data processing and cleaning steps as well as study variables to be calculated. A descriptive analysis of the risk of work stoppage is carried out, studying its evolution over recent years, the impact of seasonality and the COVID-19 pandemic, as well as trends in types of work stoppages and psychosocial risks among young people.

Then, the thesis addresses the modeling of the duration of work stoppages using generalized linear models (GLM) and duration models. It compares different duration distributions, estimates the parameters of laws, validates the models, and constructs a law for maintaining work stoppages.

Finally, the thesis focuses on the prevention of absenteeism in companies based on specialized literature on the subject and attempts to propose a procedure for implementing a prevention strategy in companies and evaluating its impacts.

In conclusion, this thesis provides insights into the modeling of disability risk and absenteeism in group insurance, combining a theoretical and empirical analysis of the available data. The results will be useful for future studies on absenteeism in the group insurance portfolio of Generali Vie.

Remerciement

Je tiens tout d'abord à remercier particulièrement Bertrand FONTAINE mon manager pour m'avoir accueilli au sein de l'équipe grands comptes de Generali Vie, pour sa pédagogie, sa disponibilité et pour toutes les connaissances qu'il m'a transmises.

Je tiens ensuite à remercier chaleureusement mon tuteur Selim BELHOUSSE pour ses précieux conseils pour la réalisation de ce mémoire, son suivi et son aide tout au long de l'année sur l'ensemble des missions de souscription qui m'ont été confié.

Je remercie également mon tuteur pédagogique Denys POMMERET pour ses conseils avisés, sa disponibilité et son suivi.

J'adresse aussi mes remerciements à l'ensemble de mon équipe pour leur bienveillance, leurs conseils et leur bonne humeur au quotidien.

Enfin je tiens à remercier ceux qui sont pour moi ma motivation, mes parents, pour leur amour, leur soutien indéfectible et leurs encouragements depuis toujours.

Introduction

Dans le contexte économique actuel, les entreprises sont de plus en plus conscientes de l'importance de la gestion des risques, notamment en matière de santé et de sécurité au travail. L'assurance collective, qui vise à couvrir les salariés contre les risques d'incapacité et d'absentéisme, constitue un enjeu majeur pour les entreprises et les assureurs. La prévention et la gestion de ces risques sont essentielles non seulement pour préserver la santé et le bien-être des salariés, mais également pour limiter les coûts directs et indirects liés à l'absentéisme et à l'incapacité.

Le présent mémoire d'actuaire a pour objectif de modéliser le risque incapacité et l'absentéisme dans le contexte de l'assurance collective en France. Pour ce faire, nous commencerons par examiner le système de protection sociale en France, en explorant son histoire, son organisation fonctionnelle et les offres proposées par le régime de base et le régime complémentaire. Nous nous intéresserons ensuite à la prévoyance collective et au risque d'arrêt de travail en France, en étudiant le contexte général et l'environnement juridique.

Dans un deuxième temps, nous analyserons l'absentéisme en entreprise en France, en définissant ce phénomène, en identifiant ses causes et en présentant des méthodes de mesure et d'évaluation de son coût pour les entreprises. Cette analyse nous permettra de mieux comprendre l'impact de l'absentéisme sur la gestion des risques et la performance des entreprises.

Ensuite, nous nous concentrerons sur la construction d'une base de données d'étude à partir de différentes sources de données sur les sinistres, en décrivant les étapes de traitement et de nettoyage des données ainsi que les variables d'étude à calculer. Cette base de données servira de fondement pour l'analyse descriptive et la modélisation du risque d'arrêt de travail.

Nous procéderons alors à une analyse descriptive du risque d'arrêt de travail, en examinant son évolution au cours des dernières années, l'impact de la saisonnalité et de la pandémie de COVID-19, ainsi que les tendances en matière de types d'arrêts et de risques psychosociaux chez les jeunes.

Nous aborderons la modélisation de la durée des arrêts de travail à l'aide de modèles linéaires généralisés (GLM) et de modèles de durée pour la construction de tables de maintien en incapacité. Nous comparerons différentes distributions de durée, estimerons les paramètres des lois et construirons une loi de maintien en arrêt de travail. Cette modélisation nous permettra d'appréhender de manière plus précise et rigoureuse les risques liés à l'incapacité et à l'absentéisme dans le contexte de l'assurance collective en France.

Enfin, dans une dernière partie d’ouverture, nous tenterons d’aborder la problématique de la prévention de l'absentéisme en entreprise en présentant une procédure de mise en place d’une stratégie de prévention en nous appuyant sur la littérature spécialisée sur le sujet.

Table des matières

I. Le risque arrêt de travail dans le contexte de l'assurance collective 9

a. Présentation de la protection sociale en France : régime de base et régime complémentaire. 9

i. Histoire de la protection sociale 9

ii. Organisation fonctionnelle 10

iii. Détails de l'offre proposée par le régime de base 10

iv. Panorama de l'offre de couverture complémentaire 15

b. Qu'est-ce que la prévoyance collective ? 17

c. Le risque arrêt de travail en France 18

i. Contexte général 18

ii. Environnement juridique 19

II. L'absentéisme en entreprise en France 21

a. Qu'est-ce que l'absentéisme ? 21

b. Quelles sont les causes de l'absentéisme ? 21

c. Comment mesurer l'absentéisme en entreprise ? 23

d. Quel est son coût réel sur les entreprises ? 25

III. Construction d'une base de données d'étude : 26

a. Présentation des bases données 26

i. Base de données de sinistres AVT 26

ii. Base de sinistres DSN 27

b. Traitement de la donnée 28

i. Fusion des bases de données – inner join 28

ii. Ciblage sur le risque arrêt de travail 29

iii. Quels choix a-t-on effectué pour le nettoyage de la base de données ? 29

c. Variables d'étude à calculer 30

i. Durée de l'arrêt de travail 30

ii. Taux d'absentéisme 30

iii. Données d'exposition et limites à l'étude 30

IV. Analyse descriptive - Evolution du risque arrêt de travail sur les dernières années 32

i. Evolution globale du portefeuille arrêt de travail 32

ii. Saisonnalité des arrêts de travail et mise en valeur d'un effet COVID sur 2020 34

iii. Effet COVID sur l'année 2020 36

iv. Evolution des types d'arrêts par année de survenance 37
v. Arrêts de travail chez les jeunes – les risques psychosociaux 38
V. Modélisation de la durée des arrêts de travail 41
a. Modèle de durée : censures et troncatures 41
i. Censure à droite : 41
ii. Troncature à gauche : 42
b. Etude de la distribution de la durée des arrêts de travail 42
i. Comparaison des fonctions de répartition 42
ii. Estimation des paramètres de lois : 43
iii. Analyse graphique et comparaison des lois 45
c. Modélisation des arrêts grâce aux modèles linéaires généralisés 48
i. Objectif de la modélisation 48
ii. Données utilisées 48
iii. Présentation des variables explicatives 48
iv. Aspect théorique des modèles linéaires généralisés 51
v. Principe des GLM 53
vi. Validation d'un modèle GLM : 57
vii. Résultats 61
viii. Construction de loi de maintien en arrêt de travail 67
ix. Rappels théoriques sur les modèles de durée 67
x. Lissage des taux bruts – Méthode de Whittaker-Henderson 69
VI. Atténuation de l'absentéisme en entreprise – prévention du risque incapacité 76
a. Importance des stratégies de prévention 76
b. Analyse des différentes stratégies de prévention 77
i. Identification des stratégies de prévention possibles 77
ii. La meilleure stratégie à mettre en place : comparaison des différentes stratégies 77
c. Evaluation de l'impact d'un programme de prévention 78
i. Approche coût-efficacité (ACE) 78
ii. Approche coût-bénéfice (ACB) 79
iii. Indicateurs clés 79
d. Gains pour les entreprises et les assureurs 79
VII. Conclusion 82
VIII. Bibliographie 83
IX. Table des figures 85
X. Annexes 86

I. Le risque arrêt de travail dans le contexte de l'assurance collective

Dans le cadre de l'étude, nous avons pour objectif d'étudier et de modéliser l'absentéisme sur le portefeuille collectif de Generali Vie. Pour introduire ce vaste sujet, il est naturel de tout d'abord présenter le système de protection sociale français, présentation qui nous amènera à aborder les spécificités de la prévoyance collective, ensuite à l'aide d'une définition détaillé du concept d'absentéisme en entreprise et des notions actuarielles autour de l'arrêt de travail, nous expliciterons les enjeux et l'importance de l'établissement d'un baromètre de l'absentéisme sur un portefeuille d'assurance collective.

a. Présentation de la protection sociale en France : régime de base et régime complémentaire.

La protection sociale désigne l'ensemble des mécanismes qui permettent aux individus de faire face financièrement à certains risques sociaux, à des situations pouvant provoquer une baisse des ressources ou une hausse des dépenses. Ces risques sont les suivants :

- Santé : actes courants, hospitalisation, optique et dentaire : remboursement des frais médicaux
- Arrêt de Travail : incapacité, invalidité, maternité : remboursement d'indemnités journalières pour l'incapacité et la maternité et d'une rente pour l'invalidité
- Décès : versement d'un capital ou d'une rente aux ayants droits
- Retraite : constitution d'un capital ou d'une rente
- Dépendance : versement d'un capital ou d'une rente
- Chômage : versement d'indemnités de chômage

i. Histoire de la protection sociale

La Sécurité Sociale a été créé en 1945 via les ordonnances du 4 et 19 octobre en fusionnant toutes les anciennes assurances et avec comme triple objectif :

- Unité de la Sécurité Sociale
- Généralisation quant aux personnes
- Extension des risques couverts

L'ordonnance du 4 octobre 1945 prévoit un système coordonné de caisses se substituant à de multiples organismes existants. Les salariés des régimes spéciaux comme les fonctionnaires, les marins, les cheminots, etc vont refuser de s'intégrer dans le nouveau régime général et conserver dans un cadre « transitoire » qui dure encore, leurs régimes spécifiques.

> Régime : Ensemble de droits et obligations réciproques des Employés (et leurs « ayant-droit »), des Patrons, et d'une Caisse de Sécurité Sociale.

Les trois grands régimes introduits par les ordonnances des 4 et 19 octobre 1945 sont les suivants :

- Régime général : salariés et travailleurs assimilés à des salariés soit environ 80 % de la population.

- Régime des travailleurs non-salariés non agricoles ou Régime social des indépendants ou RSI : artisans, commerçants et professions libérales.
- Régime agricole : exploitants et salariés agricoles, ainsi que certains secteurs rattachés à l'agriculture (ex l'industrie agroalimentaire). Il est géré par la caisse centrale de la Mutualité Sociale Agricole, MSA.

La protection sociale était à l'origine financée exclusivement par des cotisations. Le financement s'est diversifié dans les années 1990 avec la CSG et d'autres impôts.

ii. Organisation fonctionnelle

D'un point de vue fonctionnel, l'organisation actuelle de la Sécurité Sociale résulte de l'ordonnance de 1967 qui instaure la séparation de la Sécurité Sociale en 4 branches autonomes. Chaque branche est alors responsable de ses ressources et de ses dépenses.

- Branche Maladie : maladie, maternité, paternité, invalidité, décès
- Branche Accidents du travail et Maladies professionnelles
- Branche Vieillesse et veuvage
- Branche Famille : handicap, logement, RSA ...

iii. Détails de l'offre proposée par le régime de base

1. L'assurance Maladie :

Définition de l'assurance maladie :

L'assurance maladie est définie dans le cadre de l'article L321-1 du CSS :

« L'assurance maladie comporte (…) l'octroi d'indemnités journalières à l'assuré qui se trouve dans l'incapacité physique constatée par le médecin traitant, (…) de continuer ou de reprendre le travail ; l'incapacité peut être également constatée, dans les mêmes conditions, par la sage-femme dans la limite de sa compétence professionnelle et pour une durée fixée par décret (…) ».

Conditions d'ouverture des droits de l'assurance maladie :

Des conditions d'ouverture des droits s'appliquent en fonction de la durée de l'arrêt de travail et de la situation de l'assuré :

- Si l'arrêt de travail est inférieur à 6 mois :
  - Avoir travaillé au moins 150 heures au cours des trois mois précédant l'arrêt de travail
  - Ou avoir cotisé sur un salaire au moins égal à 1 015 fois le montant du SMIC horaire au cours des six mois précédant l'arrêt
- Si l'arrêt de travail est supérieur à 6 mois :
  - Justifier à la date de l'arrêt, de douze mois d'immatriculation en tant qu'assuré social auprès de l'Assurance Maladie
  - Avoir travaillé au moins 600 heures au cours des douze derniers mois
  - Ou avoir cotisé sur un salaire au moins égal à 2 030 fois le montant du SMIC horaire au cours des douze mois précédant l'arrêt de travail, dont au moins 1 015 fois le montant du SMIC horaire au cours des six premiers mois

Les indemnités journalières ;

- Durée de service des indemnités journalières :

L'indemnité journalière est accordée à partir d'un délai de carence de 3 jours (1 jour pour les fonctionnaires), ce délai de carence ne s'applique pas dans les cas suivants :

- En cas de reprise d'activité entre deux prescriptions d'arrêt de travail qui ne dépasse pas 48 heures
- L'assuré est en affection de longue durée et les arrêts sont en rapport à cette maladie alors le délai de carence ne s'applique que pour le premier arrêt.

L'indemnité journalière est due au titre de chaque jour, ouvrable ou non.

Les causes de sorties de l'état d'incapacité sont : le rétablissement, le passage en invalidité et le décès.

- Montant de l'indemnité journalière :

Le montant de l'indemnité journalière versée par la Sécurité Sociale est calculé via le salaire journalier de base dont le calcul est détaillé ci-dessous et est égale à 50% du salaire journalier de base.

Salaire journalier de base : ce salaire est la moyenne des salaires bruts des 3 derniers mois travaillés précédant l'arrêt, plafonné à 1.8 SMIC (3 022,11 € au 1er aout 2022), c'est-à-dire :

$$Salaire\ journalier\ de\ base = \min\left(\frac{Somme\ des\ salaire\ bruts\ des\ 3\ derniers\ mois}{91{,}25}; 1{,}8 * SMIC\right)$$

(Décret n° 2010-1305 -29/10/10)

L'indemnité peut notamment être revalorisée en cas d'augmentation générale des salaires de l'entreprise durant l'arrêt.

2. L'assurance Maternité et congé Paternité

Conditions d'ouverture des droits de l'assurance maternité/paternité :

Les conditions d'ouverture des droits à l'assurance Maternité et le congé Paternité sont les suivantes :

- Pour l'assurance maternité :
  - Être immatriculée en tant qu'assurée sociale depuis au moins 10 mois, à la date prévue de l'accouchement
  - Avoir cotisé sur un salaire au moins égal à 1 015 fois la valeur du SMIC horaire au cours des six mois civils précédant la date du début de la grossesse ou du début
  - Ou avoir effectué au moins 150 heures de travail au cours des trois mois précédant l'arrêt de travail, à la date du début de la grossesse ou du début du congé prénatal
- Pour le congé paternité :
  - Justifier d'au moins 10 mois d'immatriculation en tant qu'assuré social à la date de début du congé
  - Avoir travaillé au moins 150 heures au cours des trois mois précédant la date de début du congé ou avoir cotisé sur un salaire au moins égal à 1 015 fois le montant du SMIC horaire au cours des six mois précédant le début du congé

Les indemnités journalières :

- Durée de service des indemnités journalières :

Le niveau de l'indemnité journalière due au titre du congé maternité/paternité est d'abord limité par la durée légale du congé :

- Congé Maternité : 16 semaines (peut être plus longue selon le nombre d'enfants de l'assuré et le nombre d'enfants attendus)
- Congé Paternité : 11 jours consécutifs maximum, 18 en cas de naissance multiple, ce congé est à prendre par le père dans les 4 mois suivants la naissance

Il n'y a aucun délai de carence.

- Montant de l'indemnité journalière :

- L'indemnité journalière Maternité : moyenne des salaires nets des prélèvements légaux et conventionnels et de la CSG (déduction forfaitaire 21%), limités à la Tranche A (= 1PMSS : 3 428 € en 2022) des trois mois précédant le congé prénatal.
- L'indemnité journalière Paternité : moyenne des salaires nets des prélèvements légaux et conventionnels et de la CSG, limités à la Tranche A des trois mois précédant le congé prénatal.

3. *L'assurance Invalidité*

Définition l'assurance invalidité :

L'assurance invalidité est définie dans le cadre des articles suivants comme étant le droit d'avoir accès à une pension lorsque l'assuré présente une invalidité réduisant sa capacité de travail ou de gains dans des proportions déterminées :

Article L341-1 du CSS : « L'assuré a droit à une pension d'invalidité lorsqu'il présente une invalidité réduisant dans des proportions déterminées sa capacité de travail ou de gain, c'est-à-dire le mettant hors d'état de se procurer un salaire supérieur à une fraction de la rémunération soumise à cotisations et contributions sociales qu'il percevait dans la profession qu'il exerçait avant la date de l'interruption de travail suivie d'invalidité ou la date de la constatation médicale de l'invalidité. »

Article R341-2 du CSS : « Pour l'application des dispositions de l'article L. 341-1 :

1) L'invalidité que présente l'assuré doit réduire au moins des 2/3 sa capacité de travail ou de gain
2) Le salaire de référence ne doit pas être supérieur au tiers de la rémunération normale mentionnée audit article. »

Article L341-4 du CSS : « En vue de la détermination du montant de la pension, les invalides sont classés comme suit :

1) Invalides capables d'exercer une activité rémunérée ;
2) Invalides absolument incapables d'exercer une profession quelconque ;

3) Invalides qui, étant absolument incapables d'exercer une profession, sont, en outre, dans l'obligation d'avoir recours à l'assistance d'une tierce personne pour effectuer les actes ordinaires de la vie. »

Conditions d'ouverture des droits de l'assurance Invalidité :

Les conditions d'ouverture des droits à l'assurance Invalidité sont les suivantes :

- Être immatriculé depuis au moins 12 mois au moment de l'arrêt de travail suite à l'invalidité ou au moment de la constatation de l'invalidité par le médecin conseil de la caisse d'Assurance Maladie ;
- Justifier, au cours des 12 mois qui précédent l'arrêt de travail pour invalidité ou constatation médicale de l'invalidité,
  - Soit avoir effectué au moins 600 heures de travail salarié ;
  - Soit avoir cotisé sur un salaire au moins égal à 2 030 fois le SMIC horaire.

La pension d'invalidité :

- Durée de service des indemnités journalières :

La pension invalidité est interrompue en totalité ou en partie en cas de reprise du travail, sinon elle prend fin à l'âge de la retraite prévu par la loi, elle est alors remplacée par la pension de vieillesse allouée en cas d'inaptitude au travail.

- Montant de la pension invalidité :

Le montant de la pension d'invalidité versée par la Sécurité Sociale est calculé via le salaire annuel moyen égal à la moyenne des salaires annuels bruts plafonnés à la tranche A des dix meilleures années d'activité. Le montant de la pension d'invalidité est le suivant :

- 1ère catégorie = 30 % du salaire annuel moyen
- 2ème catégorie = 50 % du salaire annuel moyen
- 3ème catégorie = 50 % du salaire annuel moyen majoré de l'allocation tierce personne

4. L'assurance Décès

Définition et condition d'ouverture des droits de l'assurance Décès :

D'après l'article L361-1 du CSS, l'assurance décès garantit aux ayants droit de l'assuré le paiement d'un capital égal à un multiple du gain journalier de base lorsque l'assuré exerçait une activité salariée moins de trois mois avant le décès, s'il était titulaire d'une pension d'invalidité ou d'une rente allouée en vertu de la législation sur les accidents du travail et maladies professionnelles ou lorsqu'il bénéficiait, au moment du décès, du maintien de ses droits à l'assurance décès.

Montant du capital décès :

Depuis la loi de financement de la sécurité social de 2015, il est égal à un montant forfaitaire fixé par décret et revalorisé chaque année (3 539€ en 2022).

5. L'assurance Accident du travail et Maladie Professionnelle

Définitions :

D'après l'article L411-1 du CSS, est considéré comme accident du travail un accident survenu par le fait ou à l'occasion du travail à toute personne salariée ou travaillant, à quelque titre ou en quelque lieu que ce soit, pour un ou plusieurs employeurs ou chefs d'entreprise. De plus, est aussi considéré comme accident du travail l'accident qui est survenu à un travailleur, pendant le trajet d'aller et de retour, entre :

- Tout lieu où le travailleur se rend de façon habituelle (résidence principale, résidence secondaire, …) et le lieu de travail.
- Le lieu où le travailleur prend habituellement ses repas dans la mesure où le parcours n'a pas été interrompu ou détourné pour un motif dicté par l'intérêt personnel et étranger aux nécessités essentielles de la vie courante ou indépendant de l'emploi et le lieu de travail.

Pour ce qui est des maladies professionnelles, d'après l'article L411-1, est présumée d'origine professionnelle toute maladie désignée dans un tableau de maladies professionnelles et contractée dans les conditions mentionnées à ce tableau.

Aucune condition n'est formulée pour l'ouverture des droits.

Les indemnités journalières :

- Délai de carence :

Aucun délai de carence n'est applicable : l'indemnité est due à partir du 1er jour suivant l'arrêt du travail, pendant toute la période d'incapacité de travail qui précède soit la guérison complète, soit la consolidation de l'état ou le décès.

- Montant de l'indemnisation :

Ce montant est calculé via le salaire journalier de base dont le calcul a été détaillé précédemment, le montant de l'indemnité est le suivant :

- 60 % du salaire journalier de base jusqu'au 28ème jour
- 80 % du salaire journalier de base à partir du 29ème jour d'arrêt

De plus, une des spécificités de cette assurance est le taux d'incapacité permanente, taux qui est proposé par un médecin conseil lors d'un contrôle médical organisé par la caisse d'Assurance Maladie de l'assuré.

Le montant de la prestation est alors dicté par ce taux :

- Si taux < 10% : Un capital est versé en fonction du taux d'incapacité de la victime sur la base d'un barème forfaitaire.
- Si taux ≥ 10% : Versement d'une rente viagère égale au salaire annuel multiplié par le taux d'incapacité qui peut être réduit ou augmenté en fonction de la gravité de celle-ci

iv. Panorama de l'offre de couverture complémentaire

1. Organisation du marché de l'assurance complémentaire

Le marché de l'assurance complémentaire est composé de quatre acteurs principaux :

- **Les organismes assureur :** mutuelle, institut de prévoyance, société d'assurance. Ces organismes sont régis par des réglementations différentes et ont un mode de fonctionnement qui leur est propre et chaque organisme à son domaine d'intervention privilégié.
- **Les courtiers :** c'est un travailleur indépendant qui n'est pas lié à une compagnie d'assurance, il est un intermédiaire au contrat.
- **Les gestionnaires :** soit rattachés de près ou de loin à un cabinet de courtage, soit indépendants, ces entreprises sont dédiées à la gestion administrative des contrats de Prévoyance, Santé, contrats Emprunteurs, …
- **Les entreprises de conseil :** elles portent assistance à l'entreprise et possède des domaines de compétences et d'interventions souvent variés (Retraite, Santé, Prévoyance, …)

2. Les garanties de prévoyance complémentaire

Dans un premier temps, il faut distinguer les deux secteurs principaux du périmètre de la protection sociale complémentaire en France, ces deux secteurs sont : L'épargne retraite et la Prévoyance.

Les garanties collectives d'un tel contrat ont objectif de prévoir pour les salariés, les anciens salariés et leurs ayants droit, la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité, des risques d'inaptitude et du risque chômage, ainsi que la constitution d'avantages sous forme de pensions de retraite, d'indemnités ou de primes de départ en retraite ou de fin de carrière.

- Les Frais de Santé

La garantie complémentaire frais de santé intervient en complément de la Sécurité Sociale, voire dans certains cas lorsque la Sécurité Sociale n'intervient pas.

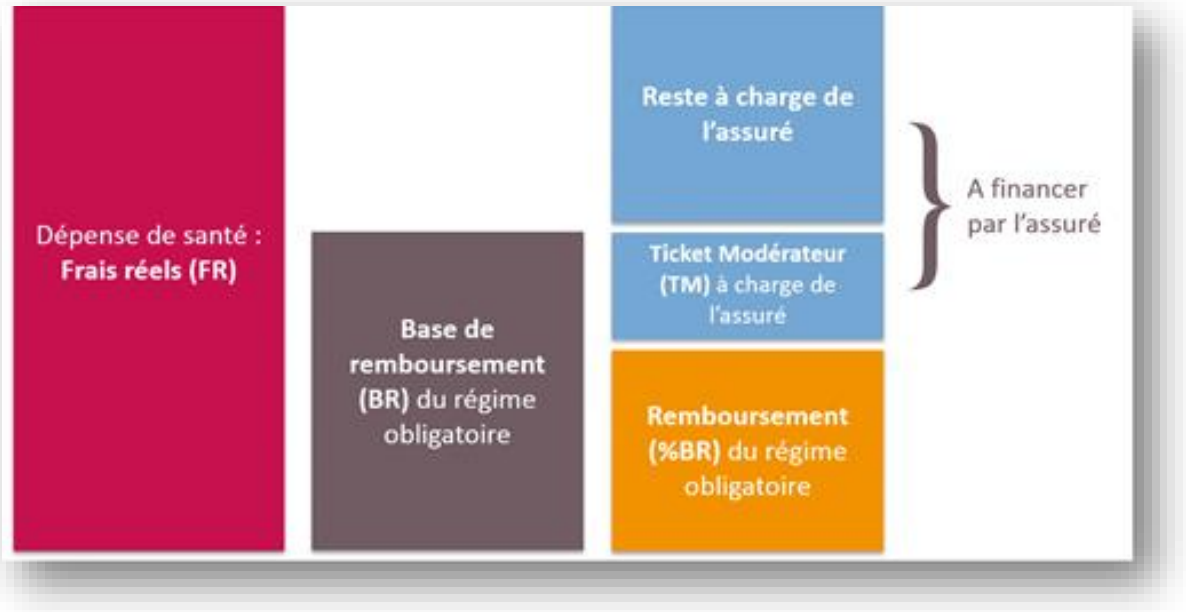


Figure 1: Intervention de la Sécurité Sociale sur les frais de santé

- La Prévoyance Lourde

➢ Capitaux Décès :

Le capital décès est versé par la complémentaire en cas de décès de l'assuré mais peut aussi être versée par anticipation en cas d'IAD (Invalidité Absolue et Définitive). Son montant dépend de la rémunération annuelle brute de l'assuré.
Ce capital est parfois accompagné de garanties annexes comme une garantie frais d'obsèques, majoration en cas décès accidentel, garantie « double effet » (en cas de décès du conjoint survivant, versement d'un capital aux enfants à charge), capital « pré-décès », …

➢ Rente de Conjoint :

La rente de Conjoint est destinée à compenser la disparition d'un revenu immédiat et/ou différé au sein de la famille, elle peut être de nature viagère ou temporaire dans l'attente de la pension de réversion des régimes de retraite obligatoire.
Elle est généralement calculée soit en fonction des droits de réversion acquis auprès des régimes de retraite complémentaire AGIRC-ARRCO, soit en fonction du dernier salaire de l'assuré décédé.

➢ Rente éducation :

La rente d'éducation est destinée à assurer le versement d'une rente aux enfants à charge de l'assuré décédé, au-delà d'un certain âge, elle peut être conditionnée à la poursuite des études par le bénéficiaire.
Le montant de la rente est généralement calculé au prorata du dernier salaire de l'assuré et est généralement exprimée par palier d'âge croissant du bénéficiaire.

➢ Incapacité de travail :

La garantie incapacité de travail permet au salarié en arrêt de travail de percevoir des indemnités journalières afin de compenser sa perte de salaire, ces indemnités viennent en complément de la Sécurité Sociale ainsi que du complément de revenu versé par l'employeur.
Les principaux paramètres définissant une garantie incapacité sont :

- La franchise : type et durée (ferme ou continue, rétroactive, discontinue)
- Le montant : en pourcentage du salaire de l'assuré
- L'assiette de calcul

L'employeur est tenu de verser un complément d'indemnisation appelé « Maintien de salaire » sous certaines conditions à un salarié en arrêt de travail. Cette obligation a été mise en place par la loi de mensualisation de 1978 puis améliorée par l'ANI du 11 janvier 2008.
Les conditions pour bénéficier du maintien de salaire pour un assuré sont :

- Condition d'ancienneté : le salarié doit avoir au moins 1 an d'ancienneté
- Franchise : 7 jours sauf en cas d'arrêt consécutif à un arrêt de travail ou une maladie professionnelle
- L'Employeur doit compléter les indemnités versées par la Sécurité Sociale à hauteur de 90 %, puis de 66,66 % de la rémunération brute que le salarié aurait gagnée s'il avait continué à travailler, et ce pour les durées définies ci-dessous :

Ancienneté	90% du salaire brut	66,6% du salaire brut
1 – 6 années	30 jours	30 jours
6 – 11 années	40 jours	40 jours
11 – 16 années	50 jours	50 jours
16 – 21 années	60 jours	60 jours
21 – 26 années	70 jours	70 jours
26 – 31 années	80 jours	80 jours
Plus de 31 années	90 jours	90 jours

➢ Invalidité :

La rente d'invalidité compense, en totalité ou en partie, la perte de revenu d'un assuré déclaré invalide. Elle vient compléter la pension d'invalidité versée par la Sécurité sociale. Le montant de cette renta peut-être fixe ou bien correspondre à un pourcentage de son salaire, elle dépend souvent aussi de la catégorie d'invalidité définie par la Sécurité sociale.

b. Qu'est-ce que la prévoyance collective ?

Dans le cadre de cette étude, nous avons la possibilité d'étudier les caractéristiques en matière d'absentéisme d'un portefeuille de prévoyance collective, celui de Generali Vie. La « prévoyance collective » est la couverture d'un groupe de personnes qui ont un lien objectif entre elles, ce groupe étant représenté par une personne qui va signer ce contrat, généralement l'entreprise dans laquelle travail un assuré.

Les différents intervenants dans un contrat de prévoyance collective sont le souscripteur, les affiliés, les assurés et les bénéficiaires.

- **Le souscripteur :** la personne morale qui signe le contrat et paie les cotisations (entreprise, association, établissement de crédit)
- **Les affiliés :** l'ensemble des personnes appartenant au groupe assurable
- **Les assurés :** l'ensemble des personnes soumises au risque
- **Les bénéficiaires :** l'ensemble des personnes susceptibles de recevoir des prestations (exemple : les enfants dans le cadre d'une rente éducation)

c. Le risque arrêt de travail en France

i. Contexte général

Pour l'assureur, la définition des prestations et de son intervention dans la prise en charge, est alignée sur la reconnaissance par la Sécurité Sociale d'un arrêt de travail vue précédemment. Les règles de base qui vont donc conduire à définir ce qu'est un arrêt de travail, avec la distinction entre l'incapacité et l'invalidité, sont donc celles de la Sécurité Sociale. On désigne en pratique par « incapacité » l'incapacité temporaire de travail et par « invalidité » l'incapacité permanente de travail.

1. Que désigne-t-on par la notion de risque arrêt de travail ?

Ce que nous désignerons dans cette étude par la notion de « risque arrêt de travail » constitue en pratique pour l'assureur l'association du risque incapacité et du risque invalidité. Cette association est nécessaire car l'incapacité temporaire de travail peut devenir permanente, soit par aggravement du facteur incapacitant, soit après la durée maximum légale de l'incapacité (36 mois), d'ailleurs cela constitue un réel enjeu pour l'actuaire qui doit prendre en compte la probabilité de passage en invalidité pour le calcul des provisions ITIP via les tables de maintien en incapacité.

Malgré qu'en pratique il aurait été plus simple de ne considérer qu'un seul type d'arrêt, ces deux risques sont très différents notamment via leur statut clairement et leur durée définis par la loi.

Parmi les différents types d'arrêts, on distingue les maladies ordinaires, que l'on classera dans la partie dédiée à l'absentéisme parmi les arrêts difficilement évitable, et le reste des types d'arrêts que nous trouverons dans le graphique ci-dessous issu du baromètre sur les arrêts de travail établie par REHALTO en 2019 et intitulé « Comprendre pour agir » :

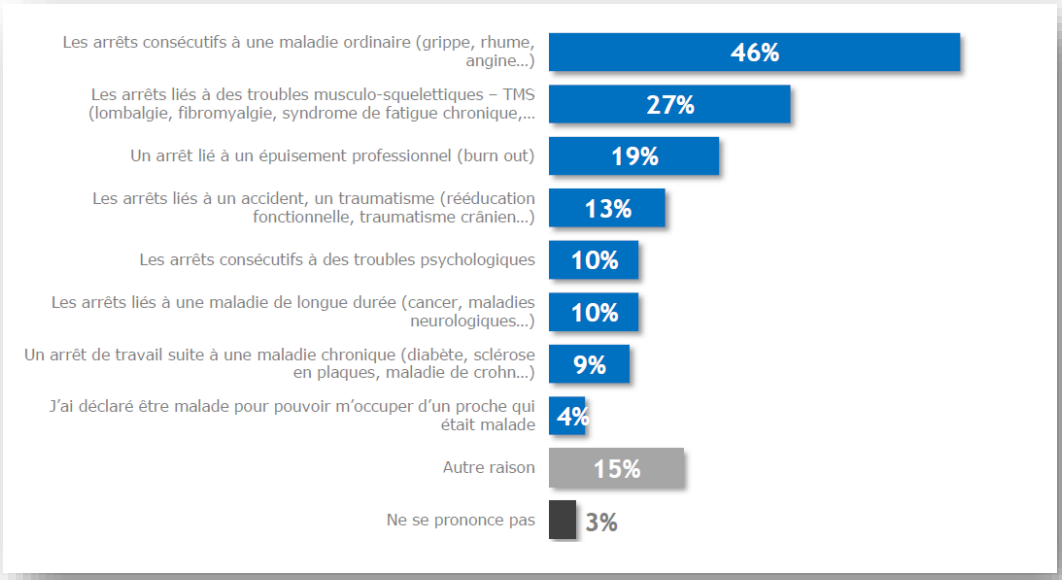


Figure 2 : Proportion des types d'arrêts connu par des salariés en entreprise

ii. Environnement juridique

L'environnement juridique autour de la notion du risque arrêt de travail est défini par les textes juridiques suivants :

- **Loi Evin (31 décembre 1989) :**

Son titre exact est : « Loi 89-1009 du 31 décembre 1989 renforçant les garanties aux personnes assurées contre certains risques »

C'est la première loi spécifique de Prévoyance, elle crée le premier ensemble de règles applicables à toutes les familles d'assureurs (les institutions de prévoyances, les mutuelles, les assureurs). Elle impose les trois points suivants :

- Évaluation des provisions techniques au niveau atteint.
- Reprise des encours (et évaluation des provisions pour revalorisation).
- Maintien des garanties décès aux invalides et aux incapables (décret du 17 juillet 2001).

Cette loi oblige l'ensemble des familles d'assureurs à provisionner les sinistres en cours pour les couvertures du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail, d'invalidité et du risque chômage.

- **Loi du 8 août 1994 :**

Cette loi arriva en complément de la loi Evin, dans le but de transposer dans le droit français une directive européenne en matière d'assurance, cette loi crée le premier corps de règles relatives à la protection sociale des salariés.

Cette loi pose les fondements juridiques de toute couverture collective, elle impose notamment une période de réexamen d'au plus 5 années pour les conventions collectives de branche ou d'entreprise qui désigne un organisme assureur et, en complément à la loi Evin, elle ajoute l'obligation d'organiser la poursuite de la revalorisation des prestations en cours de service en cas de résiliation du contrat ainsi que le maintien de la garantie décès.

- **Arrêté du 28 mars 1996 :**

Cet arrêté a pour objectif la définition de règles sur l'évaluation des provisions mathématiques d'inventaire sur les risques incapacité et invalidité :

  - Concernant le provisionnement incapacité : une provision de l'incapacité en cours doit être compléter par une provision de passage en invalidité ou encore appelée provision d'invalidité en attente.
  - Pour l'invalidité, une provision mathématique d'invalidité en cours doit être constituée
  - Les provisions mathématiques doivent être calculées en utilisant des tables réglementaires de maintien du BCAC, ou d'une table certifiée et en utilisant un taux actuariel inférieur à 75% du TME, majoré par 4,5%

II. L'absentéisme en entreprise en France

a. Qu'est-ce que l'absentéisme ?

L'absentéisme est une notion dont la définition n'est pas unanime selon les sources et diffère en fonction de la nature de l'absence, selon l'encyclopédie Larousse, l'absentéisme est le fait d'être absent du lieu de travail ou de tout lieu où pour des raisons de travail, de participation à une action ou autre, la présence est obligatoire, en entreprise on distingue les absences non évitables comme par exemple les congés maternité, pour événements familiaux, etc, et les arrêts évitables qui dépendrait autant de comportements personnels que de facteurs sociaux.

Selon le rapport de l'ANACT, l'absentéisme en entreprise désigne toute absence qui aurait pu être évitée par une prévention suffisante, faite en amont sur les conditions de travail au sens large, c'est-à-dire les conditions qui concernent les conditions physiques de travail (position de travail, matériel adapté à disposition du salarié, …), l'organisation du travail, la conciliation des temps professionnel et privé, et tout autre facteur pouvant améliorer la qualité de vie au travail. Selon ce rapport, les absences comme les congés de formation ou maternité font parti des absences qui ne sont pas de l'absentéisme. En bref, le choix de ce que l'on désigne ou pas comme de l'absentéisme appartient en pratique au responsable qui choisit ce que l'on considère ou non dans le calcul du taux d'absentéisme, qui est un indicateur que l'on détaillera dans la suite.

b. Quelles sont les causes de l'absentéisme ?

Les causes de l'absentéisme sont nombreuses et de nature très variées, en passant par des causes fréquentes et évidentes comme des conditions de travails pénibles ou inconfortable (ex : port de charge lourdes, mauvais matériel,…) et des causes un peu plus cachés comme l'ambiance au travail ou les exigences du travail qui peuvent avoir un impact psychologique important sur les salariés et ainsi être la source de troubles psychosociaux, nous allons à présent tenter de présenter certaines causes importante d'absentéisme en entreprise :

- **Les conditions matérielles de travail :** il faut dans un premier temps considérer les conditions matérielles de réalisation du travail comme étant une cause très importante de l'absentéisme en entreprise, cela passe par la configuration et l'état des locaux de travail, les horaires de travail, ces différents éléments augmentent la charge physique des salariés et sont souvent une des causes d'apparition de troubles comme les troubles musculo squelettiques ou psychosociaux. Hormis les maladies ordinaires, les troubles musculo squelettiques (TMS) et les troubles psychosociaux sont les causes principales d'arrêt de travail en France, selon Rehalto, respectivement 27% et 29%. Les TMS sont à l'origine de gènes et de douleurs dans la vie quotidienne et dans le travail, ils sont qualifiés par l'OMS comme des « maladies liées au travail » (« work-related diseases ») car la nature, le milieu et les conditions de travail ont une part importante dans l'étiologie des maladies multifactorielles. En effet, on identifie dans la littérature de nombreux facteurs d'origine professionnelle et personnelle, dont des facteurs non modifiables comme l'âge, le sexe ou la taille, et d'autres facteurs que l'on considérera évitable c'est-à-dire que l'on pourrait atténuer par de la prévention

comme la consommation de tabac, la sédentarité ou la posture adoptée au travail. L’entreprise a une responsabilité quant à la prévention des facteurs de risques de TMS de nature professionnelle ce sont en général des facteurs de risques biomécanique comme des mouvements trop répétitifs, des vibrations ou encore des séquelles dues au port de charges lourdes et des facteurs de risques psychosociaux.
Les risques psychosociaux sont réputés pour favoriser certaines pathologies de mal-être au travail comme le burn-out qui est synonyme d’épuisement lié à une charge de travail et d’une pression trop intense. La notion de risques psychosociaux est vaste et concerne l’ensemble des éléments inhérent impactant la charge mentale de l’employé comme les situations de stress, les violences internes ou externes à l’entreprise qui peuvent générer un mal-être de l’employé au travail puis dérivé sur des pathologies menant à de l’absentéisme.

- **L’engagement et le désengagement au travail :** Les causes de l’absentéisme ne sont pas uniquement dû à la santé physique ou psychologique des salariés, l’engagement au travail définit par LAWLER E.E. et HALL D.T. dans leur ouvrage *Journal of Applied Psychology* comme étant : « le degré selon lequel une personne perçoit son travail comme étant une partie importante de sa vie et de son identité, grâce aux opportunités qu’il offre de satisfaire des besoins importants » joue un rôle très important dans la vie d’une entreprise car il traduit la santé de la relation entre une personne et son employeur, particulièrement son degré d’implication dans son travail.

c. Comment mesurer l’absentéisme en entreprise ?

L’absentéisme comporte de multiples aspects qui s’entrecroisent : administratifs, économiques, sociaux et sanitaires. Il n’y a pas de cause unique à l’absentéisme des salariés. L’absentéisme est un révélateur du fonctionnement de l’organisation, de l’attrait que celle-ci recèle pour les salariés (en favorisant l’engagement au travail), mais aussi de l’état de santé global d’une population et de ses caractéristiques (âge et genre, etc.). Dans tous les cas, il est nécessaire, avant d’agir, de procéder à un bon diagnostic de la situation. Ensuite, un débat nourri entre les directions, l’encadrement, le personnel et ses représentants sera indispensable pour mettre en place un programme d’action efficace.

Pour pouvoir effectuer des actions de prévention sur l’absentéisme en entreprise, il faut d’abord être en mesure de le quantifier et de la caractériser.

Pour caractériser de l’absentéisme, il faut être en mesure d’identifier la nature des absences, la durée, la fréquence, potentiellement la rechute, etc.

Nous définissons ensuite les indicateurs que nous pouvons utiliser en pratique pour quantifier l’absentéisme en entreprise.

Le taux d’absentéisme est un indicateur fréquemment utilisé pour quantifier l’absentéisme, c’est le rapport entre nombre de jours arrêté sur une année et le nombre de jours de travail total :

$$Taux\ d'absentéisme = \frac{Nombre\ de\ jours\ d'absence\ total\ sur\ la\ période\ d'étude}{Nombre\ de\ jours\ sur\ la\ période\ d'étude}$$

Le taux de fréquence d'absentéisme : Il mesure la fréquence à laquelle les employés s'absentent. Ce taux est calculé en divisant le nombre total d'absences par le nombre total d'employés sur une période donnée.

$$Taux\ de\ fréquence\ absentéisme = \frac{Nombre\ total\ d'absences}{Nombre\ total\ d'employés}$$

La durée moyenne d'absence : Elle représente la durée moyenne d'une absence pour un employé. Elle se calcule en divisant le nombre total de jours d'absence par le nombre total d'absences sur une période donnée.

$$Durée\ moyenne\ d'absence = \frac{Nombre\ total\ de\ jours\ d'absence}{Nombre\ total\ d'absences}$$

L’indice de gravité de l'absentéisme : Il mesure l'impact des absences sur le temps de travail total. Il est calculé en divisant le nombre total de jours d'absence par le nombre total de jours travaillés théoriques multiplié par le nombre total d'employés.

$$Indice\ de\ gravité\ de\ l'absentéisme = \frac{Nombre\ total\ de\ jours\ d'absence}{\text{Nombre total de jours travaillés théoriques x Nombre total d'employés}}$$

	Définition	Rôles
Nombre de jours d'arrêts	Volume total des jours d'arrêts sur l'année	▪ Prise de conscience des volumes ▪ Identification des « masses » → focaliser l'action
Taux d'absentéisme	Part des jours de l'année passés en arrêt de travail sur toute la population	▪ Pilotage global ▪ Comparaison à la concurrence
Prévalence	Part des salariés ayant eu au moins un arrêt dans l'année	▪ Indicateur du contexte socio-économique d'une entreprise et des risques psychosociaux
Fréquence	Nombre moyen d'arrêts par salarié dans l'année	▪ Analyse et comparaison des populations entre elles ▪ Analyse et comparaison des arrêts par durée
Durée moyenne annuelle	Nombre moyen de jours d'arrêt dans l'année pour un salarié arrêté au moins une fois	▪ Indicateur notamment de la santé des salariés et du rôle de l'âge

d. Quel est son coût réel sur les entreprises ?

L'absentéisme engendre des coûts complexes en pratique pour l'entreprise. Il existe deux types de coûts selon qu'ils soient directs ou indirects :

- **Les coûts directs :**

Ce sont les plus simples à évaluer, puisqu'ils sont constitués principalement des montants de salaires et des cotisations sociales associés à la période d'absence d'un salarié. Ainsi, il s'agit de considérer le complément patronal durant l'absence du salarié par rapport aux indemnités journalières de la Sécurité sociale, et le maintien du salaire pendant le délai de carence.
Ce coût dépend essentiellement de la politique RH de l'entreprise et de la convention collective : en effet, une entreprise peut décider si oui ou non elle souhaite maintenir le salaire d'un employé pendant le délai de carence en cas de maladie lorsque son ancienneté est inférieure à un an.

- **Les coûts indirects**

Il existe certains coûts, dits *indirects*, reliés principalement à des risques sociaux engendrés par l'absence d'un salarié susceptibles de faire augmenter significativement le coût de l'absentéisme. Parmi ceux-ci peuvent être nommés :

- **Les coûts de remplacement :** un salarié absent peut être remplacé par un autre salarié (CDD, intérim, …) ce qui engendrera un coût supplémentaire à l'entreprise.
- **Les coûts de gestion** : la gestion des absences reste un processus long et coûteux pour les ressources humaines.
- **Les coûts liés aux dysfonctionnements organisationnels :** une absence provoque effectivement un dysfonctionnement dans son équipe puisque les tâches qui lui reviendraient en temps normal doivent être redistribuées à d'autres salariés. A ce titre, elles peuvent être redistribuées à des collègues, ce qui pourra engendrer des heures supplémentaires, ou elles peuvent être relayées à un remplaçant qu'il faudra alors former.
- **Les coûts d'improductivité :** Une personne remplaçant un salarié absent aura besoin de temps pour se former aux différentes tâches qui lui seront confiées, et seront moins efficaces que le salarié remplacé.
- **Les coûts sociaux :** les dépenses et la réorganisation du travail engendrées par une absence peuvent venir dégrader le climat social puisque les salariés présents peuvent voir leur charge de travail augmenter significativement.
- **Les coûts d'image :** le risque de perdre en productivité étant très prononcé.

III. Construction d'une base de données d'étude :

Pour effectuer notre étude, nous avons besoin d'une base de données exploitables pour y effectuer nos calculs, nos études statistiques et pour calibrer nos modèles. Pour ce faire, nous avons en notre disposition, deux bases de données : la première est une base de données de sinistres provenant du logiciel de gestion de Generali qui répertorie en ligne chaque sinistre avec les informations que nous présentons dans la suite, la deuxième est une base de données complémentaire issue de la DSN.

a. Présentation des bases données

i. Base de données de sinistres AVT

1. Extraction de la base de données

Les données de cette base sont des données de prestation en prévoyance collective du portefeuille de Generali VIE fournies par la direction de la Techniques Assurance.
Cette base de données est alimentée par les équipes de gestion, actualisée et mise à jour chaque mois puis importé dans le logiciel de gestion AVT, utilisé par l'ensemble des équipes de souscription et surveillance de portefeuille pour leurs opérations au quotidien.
Cette base de données nous donne accès à l'ensemble des sinistres prévoyance survenue sur le périmètre collectif depuis 2012.
Nous allons à présent nous focaliser dans le détail de la structure de cette base de données et des informations qu'elle pourra nous apporter dans notre démarche.

2. Quelle est la structure de la base de données d'étude ?

La base de données d'étude est structurée par ligne, chacune d'elle correspond à un sinistre prévoyance, voici une liste non-exhaustive des garanties couvertes :

Garantie	Fréquence
ITT	183 090
ITT_HPRO	144 514
INC	21 805
CDC_ASS	17 817
ITT_PRO	17 305
ITTM	14 273
ITTM_HPRO	11 815
ITT_HOSP	6 249
ITTA_PRO	6 207
IPT_HPRO	5 172
ALL_OBS	4 653
ITTA	4 561

Pour faciliter notre étude, nous regrouperons d'un côté les garanties d'incapacité ou incapacité de travail temporaire (ITT) et de l'autre côté les garanties d'invalidité ou incapacité de travail permanente (ITP).

La base de données initiale est composée de 63 variables pour chacun des sinistres, nous avons sélectionné parmi celles-ci, les variables suivantes :

Variable	Description
NUM_CONTRAT	Numéro de contrat
NOM_ASSURE	Nom de l'assuré
PRENOM_ASSURE	Prénom de l'assuré
GAR_ELEMENTAIRE	Intitulé de la garantie élémentaire
MOTIF_SITUATION	Motif du sinistre
DAT_NAISSANCE_ASS	Date de naissance de l'assuré
DAT_SURVENANCE_SIN	Date de survenance du sinistre
DEB_COUVERTURE	Date de début de couverture par l'assureur du sinistre
FIN_COUVERTURE	Date de fin de couverture par l'assureur du sinistre
DEB_PER_REGLEMENT	Date du premier règlement de l'assureur
FIN_PER_REGLEMENT	Date du dernier règlement de l'assureur
DT_MISE_INVALIDITE	Date de passage en invalidité
CD_OPTION	Code option

Après avoir sélectionné les informations intéressantes pour notre étude, un travail de data-cleaning était nécessaire pour différentes raisons, d'abord pour homogénéiser les données car elles ont été saisies par différentes équipes de gestions qui n'ont pas les mêmes process et les mêmes normes depuis 2012, puis enlever tout les doublons de la base et enfin il faut pouvoir être capable de capter les erreurs de saisie et de faire des choix quant au nettoyage ou au remplacement de celles-ci afin d'avoir la base de données la plus « propre » possible.

Une fois ce travail effectué, nous sommes passés d'une base de données de 711 420 observations à 107 205 observations.

ii. Base de sinistres DSN

1. Qu'est-ce qu'une base DSN pour un assureur ?

Une base DSN (Déclaration Sociale Nominative) est une base de données mise en place par les organismes de protection sociale pour collecter et centraliser les informations relatives aux salariés et à leurs contrats de travail.

Pour un assureur, la base DSN lui permet de disposer d'une source de données complète et à jour sur les employeurs et les salariés qui constituent ses clients potentiels. Cette base de données contient des informations telles que les noms et prénoms des salariés, leur numéro de

sécurité sociale, leur date de naissance, leur adresse, leur contrat de travail (type de contrat, durée, salaire, horaires de travail), ainsi que les cotisations sociales versées par l'employeur.

En utilisant les informations de la base DSN, l'assureur peut donc mieux cibler ses offres de produits d'assurance en fonction des profils et des besoins de chaque entreprise et de chaque salarié. Par exemple, si une entreprise a un grand nombre d'employés travaillant à l'extérieur, l'assureur peut proposer des assurances spécifiques couvrant les risques liés aux déplacements professionnels.

De plus, la base DSN est un outil qui permet à l'assureur de suivre l'évolution des données de ses clients, notamment en cas de changement de situation (nouvelles embauches, départs, modifications de contrat). Cela permet à l'assureur de maintenir des offres d'assurance adaptées et de s'assurer que les contrats sont toujours en conformité avec la réglementation en vigueur.

En résumé, une base DSN pour un assureur est une source de données précieuse qui lui permet de mieux connaître ses clients, d'adapter ses offres d'assurance à leurs besoins et d'assurer un suivi régulier de leur situation pour garantir une couverture adéquate.

2. Quel intérêt pour notre étude ?

Cette base de données est très importante pour notre étude car elle permet de compléter les informations que nous avons sur les assurés de notre base de données de sinistres, principalement grâce au numéro de sécurité sociale qui nous renseigne de précieuses informations comme le sexe de l’assuré et le département de naissance.

Ces informations sont primordiales afin de réaliser l’un des objectifs de ce mémoire qui est la réalisation d’un baromètre de l’absentéisme sur notre portefeuille, elles peuvent nous permettre de segmenter nos données en ayant une vision de l’absentéisme par sexe mais aussi par situation géographique.

b. Traitement de la donnée

i. Fusion des bases de données – inner join

Nous pouvons utiliser l'inner join comme opération de jointure pour combiner nos deux bases de données en ne conservant que les éléments qui ont des correspondances dans les deux ensembles. Dans le cas d'une inner join entre une base de sinistres d'arrêt de travail et une base DSN d'assureur, cela signifie que seuls les éléments communs aux deux bases seront conservés.

Notre base de sinistres d'arrêt de travail contient des informations sur les arrêts de travail des salariés d'une entreprise assurée par notre entreprise. Ces informations peuvent inclure les dates et durées des arrêts de travail, les diagnostics médicaux, les prescriptions médicamenteuses, les soins prodigués, etc.

Notre base DSN, quant à elle, contient des informations sur les employeurs et les salariés, telles que les noms et prénoms des salariés, leur numéro de sécurité sociale, leur date de naissance, leur adresse, leur contrat de travail, ainsi que les cotisations sociales versées par l'employeur.

En utilisant une inner join entre ces deux bases de données, nous pouvons combiner les informations disponibles dans les deux bases de données en ne conservant que les éléments communs. Par exemple, nous pouvons obtenir une liste des arrêts de travail survenus dans une entreprise donnée, en y associant les informations des salariés concernés, telles que leur nom, leur numéro de sécurité sociale, leur date de naissance, etc.

Cette opération de jointure peut être utile pour ce mémoire dans différents contextes, par exemple pour identifier les salariés ayant connu plusieurs arrêts de travail dans une période donnée, pour suivre l'évolution de la situation des salariés dans le temps, ou encore pour évaluer les risques liés à la santé des salariés d'une entreprise.

En résumé, l'inner join entre notre base de sinistres d'arrêt de travail et notre base DSN d'assureur nous permet de combiner les informations disponibles dans les deux bases de données en ne conservant que les éléments communs. Cela peut nous aider à mieux comprendre la situation de nos clients et à adapter nos offres d'assurance en conséquence.

ii. Ciblage sur le risque arrêt de travail

Pour étudier le risque incapacité de travail, il s'agit de sélectionner dans la base de données uniquement les sinistres de type ITT : Incapacité temporaire de travail. Pour ce faire, nous avons sélectionné les sinistres dont la garantie élémentaire couverte est relative aux risques "ITT".

De plus, parmi les données sélectionnées en incapacité, nous souhaitions ne considérer que les sinistres d'une durée inférieure à 3 ans, durée maximum légale de l'incapacité, bien que dans certains cas et pour certains contrats certains assureurs prennent en charge des incapacités dépassant le seuil des 36 mois. Nous calculons des variables à partir des données initiales pour en tirer les informations utiles à notre étude, ces variables vont être présentées et expliquées dans la suite.

iii. Quels choix a-t-on effectué pour le nettoyage de la base de données ?

Dans le processus de nettoyage d'une base de données différents cas de figures peuvent exister lorsque l'on parle d'une donnée que l'on juge impropre, celle-ci peut être erroné, décalée à cause d'une erreur de saisie manuelle par exemple, ou tout simplement manquante. Dans ces différents cas de figure plusieurs façons de résoudre le problème existent, on peut soit essayer de corriger les erreurs de saisie ce qui en pratique peut s'avérer très compliqué lorsque l'on travaille avec une base de données conséquente (notre base de données initiale contient 697000 lignes), soit trouver la source de l'erreur lorsqu'elle est automatique et la réparer, soit simplement supprimer la donnée. Dans le cas de notre étude, l'objectif était d'avoir la base de données la plus précise et donc pour ce faire nous avons fait les choix suivants :

- Supprimer les lignes qui possédaient des informations importantes manquantes (ex : le code option).
- Homogénéiser et simplifier les informations de certaines variables (ex : les garanties élémentaires, les codes options), par exemple en regroupant, dans le cas des garanties élémentaires.
- Corriger lorsque c'était possible les erreurs de saisie : par exemple lors de la saisie du nom et du prénom de l'assuré, dans de nombreux cas le nom et le prénom étaient saisis dans le même champ, celui du nom, ce qui fait qu'un grand nombre de prénoms d'assuré étaient manquants dans la variable prénom car saisi dans le champ nom. Cette anomalie par exemple était corrigeable pour la plupart des dossiers.

c. Variables d'étude à calculer

i. Durée de l'arrêt de travail

L'une des informations importantes que nous devons capter grâce à cette base de données pour notre étude est la durée de l'arrêt de travail, pour cela nous devons procéder avec les informations que nous possédons et en particulier grâce à la période de couverture, ainsi on calcule la durée d'un sinistre incapacité de la façon suivante :

$$Durée\ de\ l'arrêt = Date\ de\ sortie - Date\ de\ survenance + Franchise$$

Si nous avons la date de déclaration du sinistre dans nos données alors :

$$Durée\ de\ l'arrêt = Date\ de\ sortie - Date\ de\ déclaration + Franchise$$

Comme nous ne possédons pas dans notre base de données d'étude l'information sur la durée de la franchise pour chaque contrat, nous émettrons l'hypothèse dans nos calculs que cette durée est égale à 7 jours, délai réglementaire fixé par la loi de mensualisation de 1978 (cf partie I).

Donc dans le cadre de notre étude statistique, la durée de l'arrêt sera calculée via la formule suivante :

$$Durée\ de\ l'arrêt = Date\ de\ sortie - Date\ de\ survenance + 7$$

ii. Taux d'absentéisme

Cette variable à calculer est très importante et représente un enjeu important de cette étude, notamment pour la modélisation GLM dans la suite, elle est cependant assez complexe à calculer dans le contexte de l'étude avec les données en notre possession. La difficulté de ce calcul provient du fait que nous avons besoin de l'information du nombre de jour arrêté total sur une année pour chaque assuré et que cette base référence uniquement les sinistres un à un, la solution que nous avons choisie pour calculer ce taux dans la suite est la création d'une nouvelle base de données qui cette fois-ci sera une base de données agrégée par assurés et par année de survenance.

Rappelons à présent la formule de calcul de ce taux présenté dans la première partie :

$$Taux\ d'absentéisme = \frac{Nombre\ de\ jours\ d'absence\ total\ sur\ la\ période\ d'étude}{Nombre\ de\ jours\ sur\ la\ période\ d'étude}$$

iii. Données d'exposition et limites à l'étude

Nous segmenterons les périodes d'études annuellement, principalement sur la période 2017-2021 où l'on considère la donnée comme quasi complète. Le nombre de jours total d'une période correspondra alors au nombre de jours calendaires de l'année.

Un point d'attention important est l'exposition au risque de chaque assuré. Théoriquement pour avoir l'information de l'exposition au risque arrêt de travail en entreprise, les données adéquates nécessaires pour chaque assuré devraient être la date d'embauche ainsi que la date de fin de

contrat en cas de départ afin de connaître la période sous risque de chaque assuré et de calculer des taux d’absentéismes globaux pertinents. Dans ce mémoire le manque d’information sur les assurés dû principalement au contexte de l’assurance collective nous pousse à faire une hypothèse sur l’exposition qui se base sur le fait que les contrats sont renégociés chaque année, nous considérerons par défaut que les assurés ont une exposition égale à 1, c’est-à-dire que l’on considère les assurés dans l’entreprise sur l’entièreté de la période d’étude, ce qui en réalité pose problème et biaise l’étude.

De plus un autre frein à l’étude de l’absentéisme sur notre portefeuille et créant l’impossibilité de la mise en place d’un baromètre de l’absentéisme sur notre portefeuille est le manque d’information sur les assurés non-sinistrés.

IV. Analyse descriptive - Evolution du risque arrêt de travail sur les dernières années

Dans cette partie nous tâcherons d'analyser nos données en nous intéressant d'abord à la taille de notre portefeuille d'étude, puis à l'évolution du nombre d'arrêts de travail et sa saisonnalité.

i. Evolution globale du portefeuille arrêt de travail

Nous nous intéresserons dans un premier temps à l'évolution du nombre d'arrêts de travail pour incapacité sans prendre en compte l'évolution du portefeuille collectif. En effet l'évolution du portefeuille par année est une information importante pour que l'on puisse avoir des données comparables entre années.

Année de survenance	Nombre de sinistre
2011	383
2012	812
2013	1 760
2014	3 304
2015	4 246
2016	4 761
2017	5 832
2018	7 130
2019	9 437
2020	12 344
2021	14 845
2022	7 309

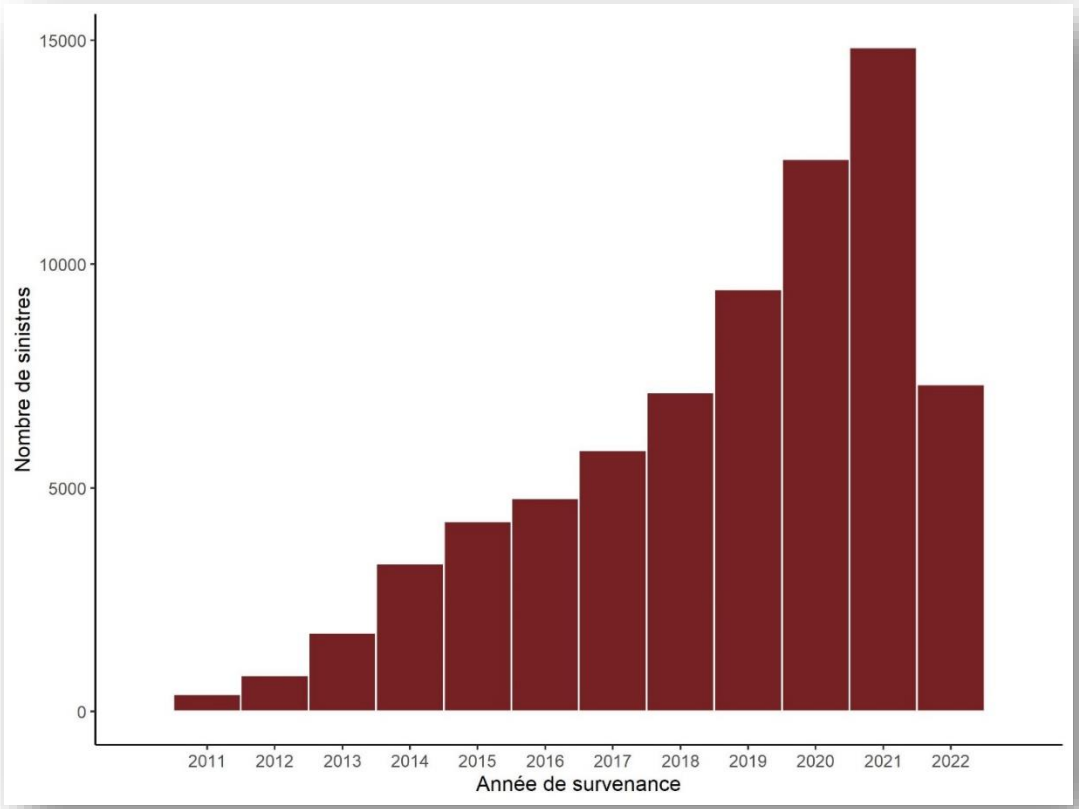


Figure 3 : Evolution du nombre d'arrêts par année de survenance

On observe une croissance relativement constante du nombre d'arrêt de travail entre 2011 et 2018 puis une forte hausse sur 2019-2021 suivi d'une baisse conséquente sur 2022 en raison de la date d'extraction des données courant 2022. Nous verrons dans la suite lorsque nous étudierons la saisonnalité que la forte hausse sur 2020 est une hausse appelée « effet COVID » par les assureurs du marché, on observe durant cette année une hausse significative des nombres d'arrêts durant les périodes de vague d'infection.

En divisant le nombre d'arrêts par la taille du portefeuille collectif, nous obtenons alors un taux d'arrêts maladie par salarié, ce qui semble très intéressant pour comparer les différentes années.

Année	Effectif couvert
2017	596 926
2018	596 951
2019	629 980
2020	666 012
2021	674 649
2022	674 649

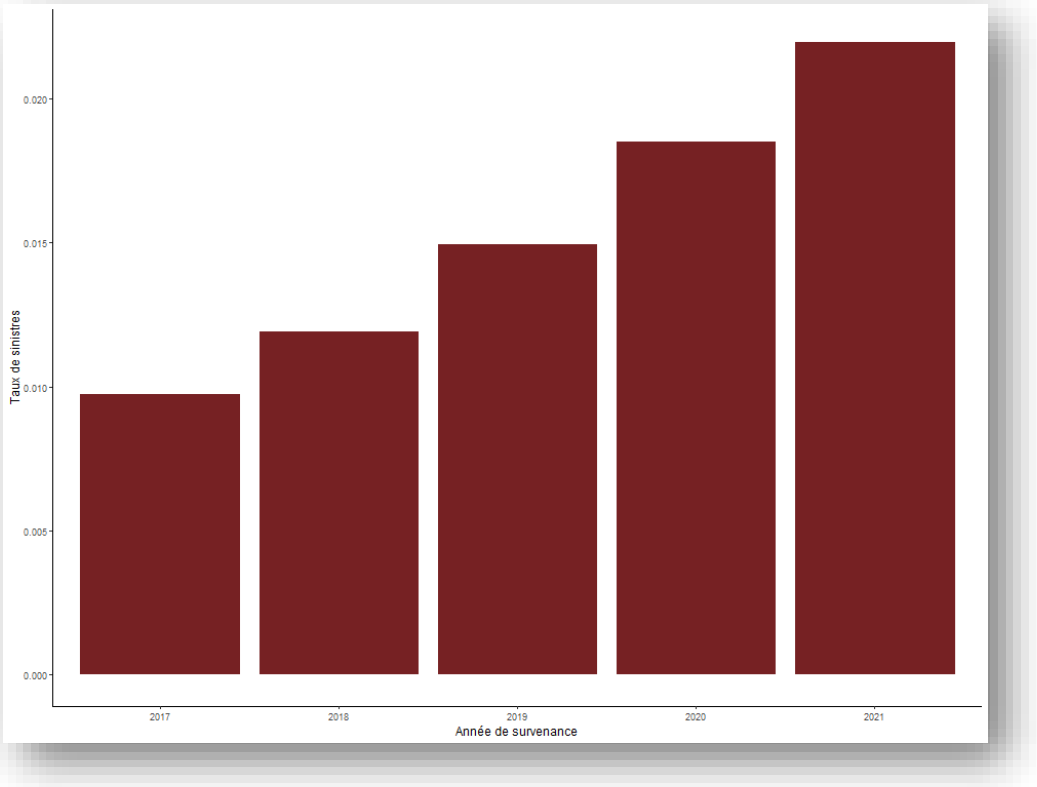


Figure 4: Evolution du taux d'arrêt de travail par assuré

On observe une croissance constante du taux d'arrêts de travail sur le portefeuille par assuré, cette croissance est en moyenne de 22,5%.

ii. Saisonnalité des arrêts de travail et mise en valeur d'un effet COVID sur 2020

Il s'agit ici de s'intéresser à la saisonnalité de la survenance des arrêts de travail sur notre portefeuille afin de constater les tendances sur les dernières années.

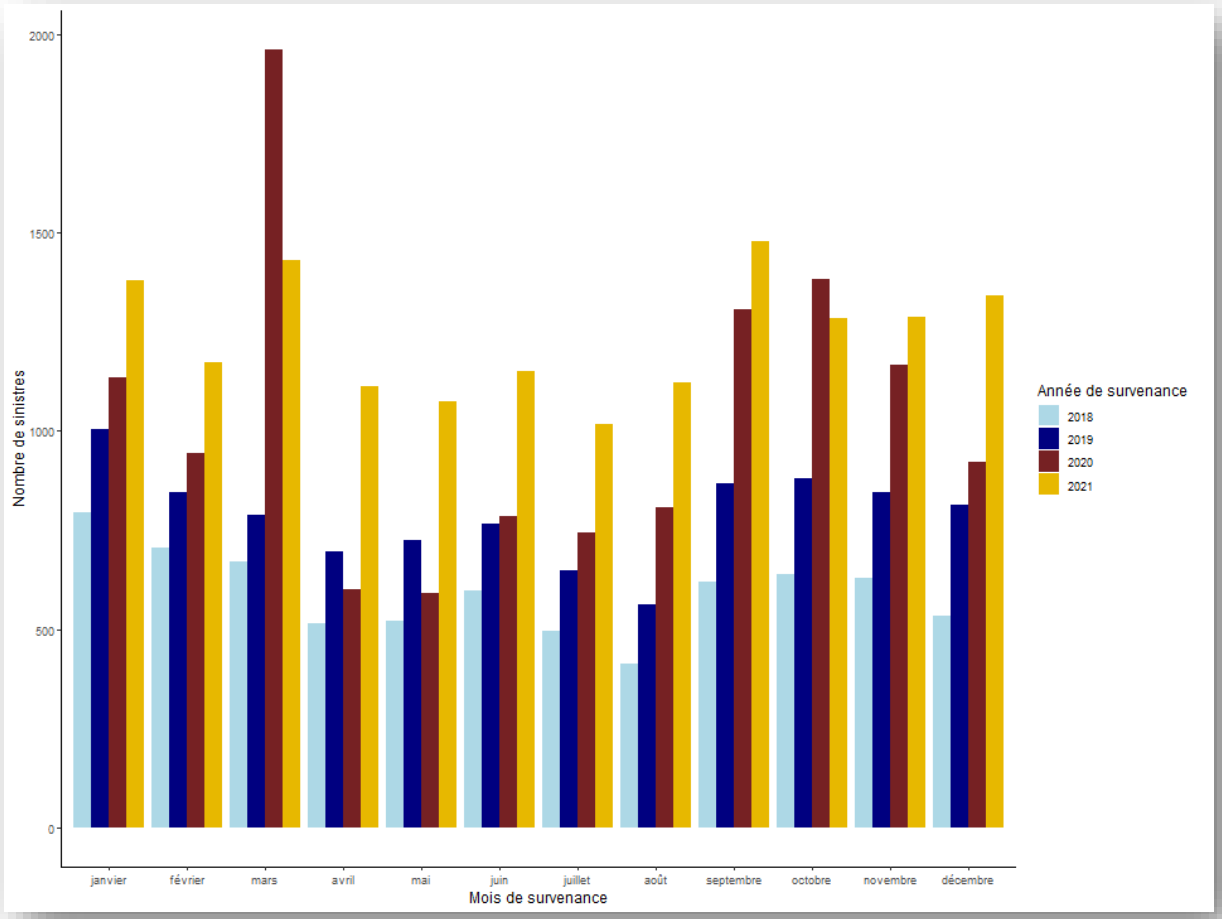


Figure 5 : Saisonnalité des arrêts de travail entre 2018 et 2021

Ce graphique est très intéressant car il nous permet de constater une réelle saisonnalité du nombre d'arrêts par année, de plus on observe le caractère exceptionnel des années 2020 et 2021, nous nous intéresserons à ces années dans la suite pour mettre en valeur l'effet Covid.

On remarque une baisse significative du nombre d'arrêts entre la période automne-hiver et la période printemps-été, cette baisse est dû en partie à la présence accrue des maladies ordinaires comme le rhume ou la grippe en hiver en sachant notamment que les maladies ordinaires représentent une partie importante du nombre d'arrêts maladie déclarés chaque année, 46% des arrêts maladie selon l'étude effectuée par REHALTO. Pour bien se rendre compte de cette différence, considérons le graphique suivant qui présente le nombre d'arrêts de travail par saison :

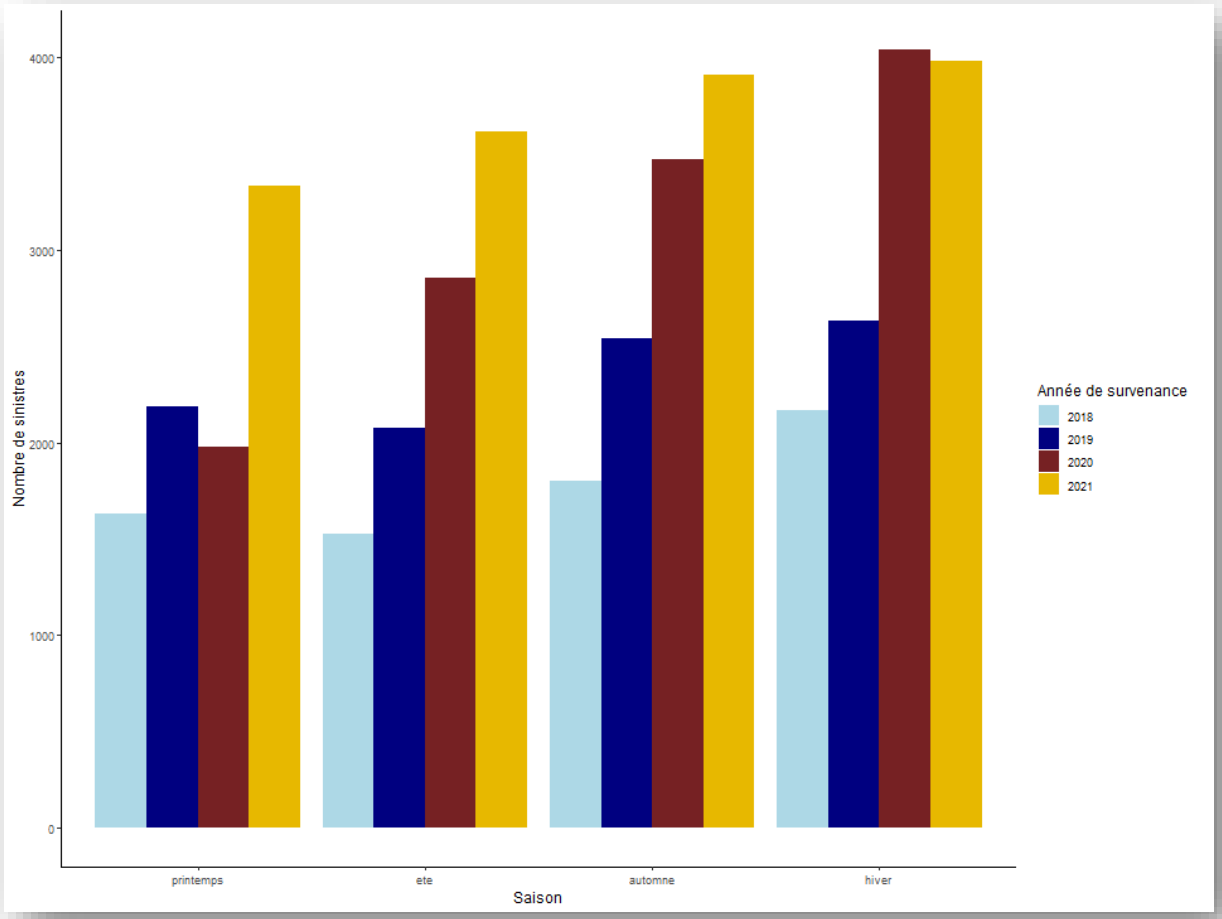


Figure 6 : Nombre d'arrêts de travail par saison

Sur 2018, nous observons une hausse de 34,7% du nombre d'arrêts entre l'été et l'hiver, une légère baisse de 7,2% entre printemps et été et une hausse 9% entre été et l'automne.

Intéressons-nous à présent à l'année 2020 et à l'effet COVID que l'on observe bien sur nos données.

iii. Effet COVID sur l'année 2020

Nous reconsidérons à présent le graphique d'évolution du nombre d'arrêts de travail par année de survenance depuis 2018 :

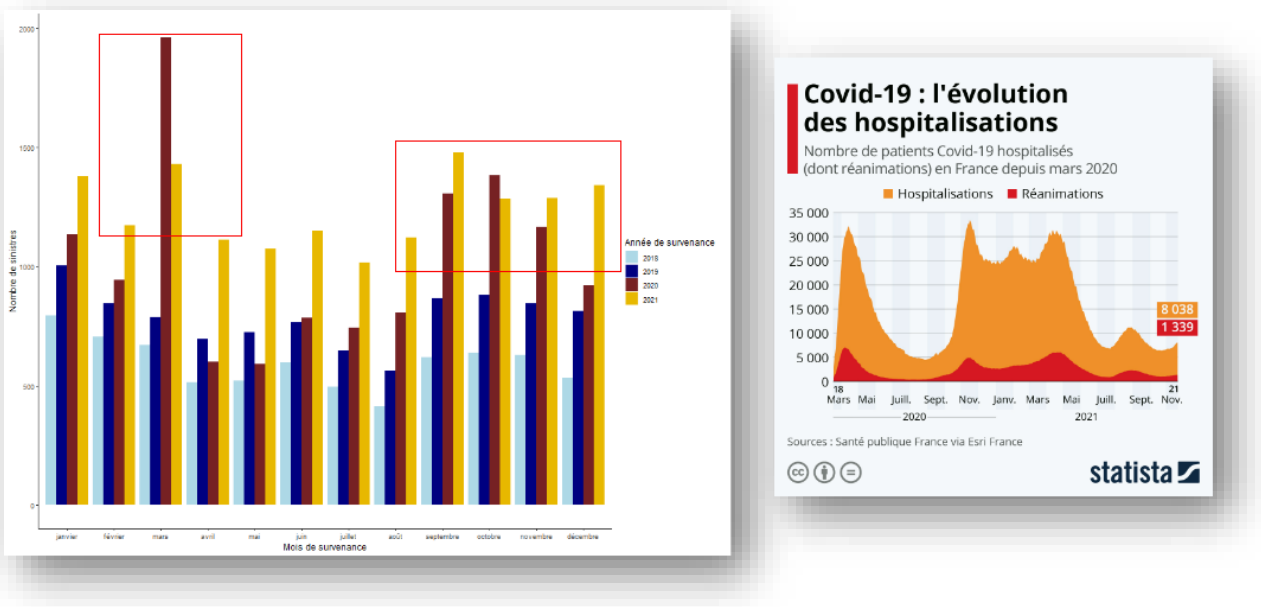


Figure 7 : Mise en valeur d'un « effet COVID » en comparaison avec la courbe d'évolution des hospitalisations sur 2020

Nous observons une hausse significative du nombre d'arrêts intervenus en mars 2020 (+192% par rapport à mars 2018), correspondant à la première vague importante d'infection de COVID-19 en France et du premier confinement.

Puis dû à la période de confinement et les mesures gouvernementales, un nombre d'arrêts inférieur à la moyenne sur la période avril-juin est survenu.

Nous observons ensuite une hausse significative du nombre d'arrêts sur la période septembre-novembre (+110% par rapport à 2018), ce qui semble correspondre à la seconde vague importante d'infection intervenue en 2020 et qui s'explique aussi par les mesures prises par le gouvernement pour simplifier la mise en arrêt de travail à la suite d'une infection au COVID.

On observe aussi et principalement une hausse significative constante sur l'année 2021 du nombre d'arrêts survenu, en comparaison à l'année 2019, la différence moyenne est de +58,8%.

Cette hausse significative du nombre d'arrêts de travail est à prendre très au sérieux pour l'assureur qui prend en charge une partie du salaire des assuré comme nous avons pu le présenter dans la première partie. Le suivi de ce risque est très important sur 2022 afin de savoir si nous assistons à un retour à la normal ou à une nouvelle tendance quant à mise en arrêt des travailleurs en France.

iv. Evolution des types d'arrêts par année de survenance

Une tendance à étudier lorsque l'on traite des arrêts de travail est la durée de ceux-ci, pour cela, nous allons distinguer trois différents types d'arrêts :

- Arrêt court : arrêt inférieur à 10 jours
- Arrêt moyen : arrêt durant entre 10 et 90 jours
- Arrêt long : arrêts d'une durée supérieure à 90 jours

Il semble très intéressant d'étudier l'évolution du type d'arrêt des salariés notamment pendant l'année 2020, mais aussi pour vérifier certaines tendances que nous pouvons lire dans la littérature comme le fait que les jeunes ont de plus en plus tendance à faire des arrêts court.

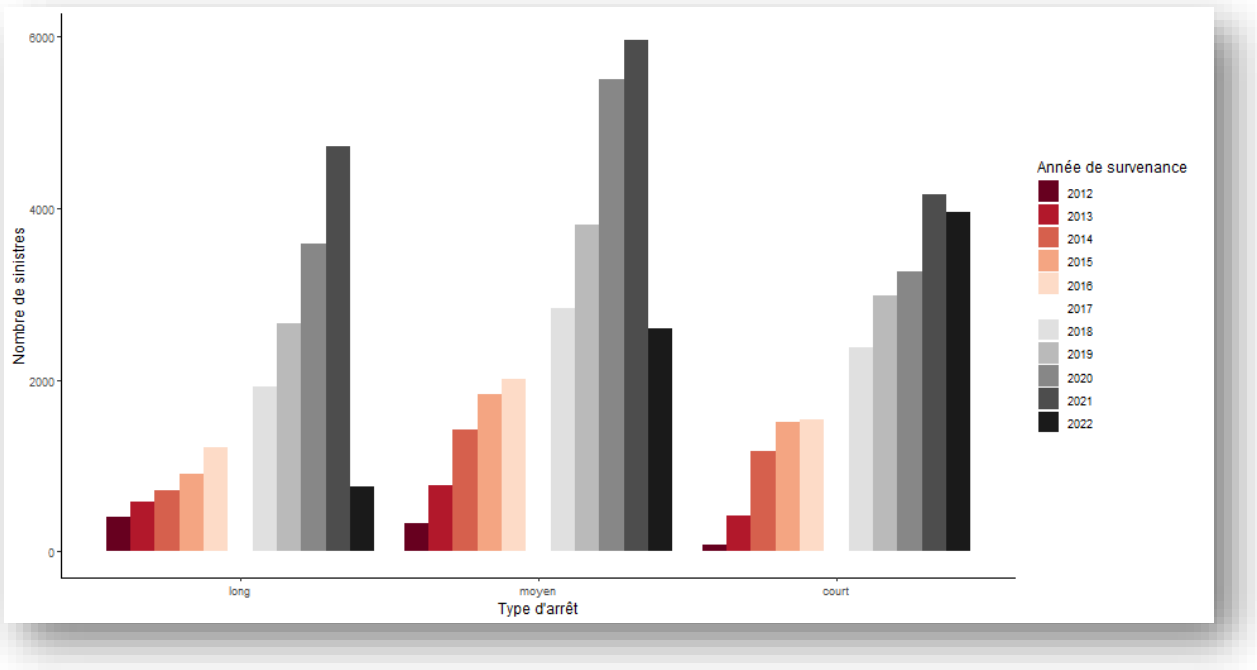


Figure 8 : Evolution du nombre d'arrêt par type

Nous observons ici une hausse significative du nombre d'arrêts moyen sur 2020, cette hausse est encore une fois sûrement dû aux dispositifs d'indemnisations des salariés lors des interruptions de travails à la suite d'un test positif au COVID-19, dans ces cas il était très fréquent que la période d'arrêt dépassait 10 jours qui correspondait au temps moyen d'atténuation des symptômes sauf pour les cas plus graves de COVID comme les COVID long.

Un point très intéressant quant à la tendance actuelle des arrêts de travail est l'évolution du nombre d'arrêts court sur les dernières années et particulièrement sur 2022, bien que l'on ait capté dans nos données qu'une partie des arrêts survenus sur 2022, on observe un nombre très important d'arrêts court, quasiment égal au total sur 2021. La hausse des arrêts courts est souvent assimilable à des arrêts dû aux troubles psychosociaux qui représentent selon une étude effectuée par REHALTO, à 19% des arrêts pour épuisement professionnel ou burn-out et à 10% des arrêts à la suite de troubles psychologiques.

Ces risques psychosociaux touchent de plus en plus les jeunes, c'est pourquoi nous allons à présent nous intéresser à l'évolution des types d'arrêts chez la classe d'âge 18-29 ans.

v. Arrêts de travail chez les jeunes – les risques psychosociaux

Ci-dessous le graphique présentant l'évolution du nombre d'arrêts de travail par type depuis 2012 :

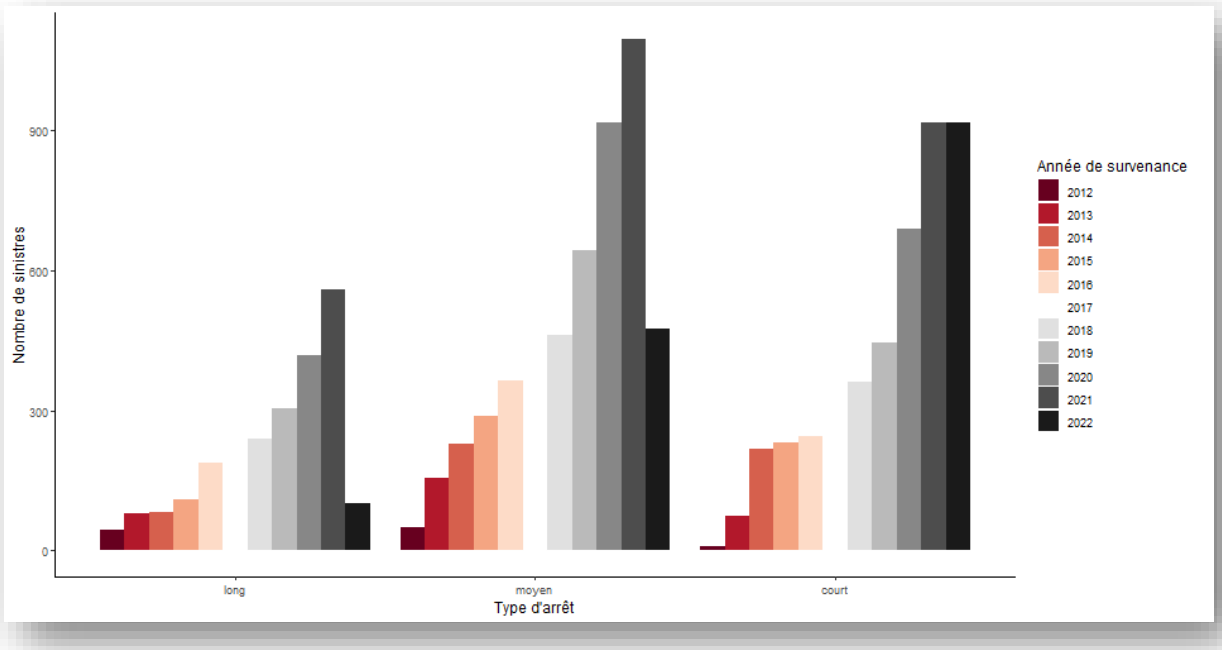


Figure 9: Evolution du nombre d'arrêt par type chez les jeunes

On observe depuis 2017, une tendance claire d'évolution du nombre d'arrêts chez les jeunes et particulièrement sur les arrêts de type court, on observe une croissance de +8,4% en 2018, +23,5% en 2019, +54,3% en 2020 et +33,1% en 2021. Les risques psychosociaux sont un des facteurs responsables de cette hausse chez les jeunes qui en sont de plus en plus victimes.

Les risques psychosociaux (RPS) sont des facteurs professionnels qui peuvent affecter la santé mentale, le bien-être et la performance des travailleurs. Ces risques peuvent avoir de nombreuses formes, notamment le stress, l'anxiété, la dépression, le harcèlement et la violence au travail.

Chez les jeunes travailleurs, les RPS peuvent être particulièrement préoccupants en raison de leur manque d'expérience professionnelle et de leur vulnérabilité aux pressions et aux conflits au travail. De plus, les jeunes travailleurs peuvent être confrontés à des conditions de travail précaires, telles que des contrats temporaires ou des horaires de travail flexibles, qui peuvent augmenter leur niveau de stress et d'insécurité au travail.

Les RPS peuvent également être à l'origine d'arrêts de travail chez les jeunes travailleurs, car ils peuvent entraîner des problèmes de santé mentale, tels que l'anxiété et la dépression. Les arrêts de travail liés aux RPS peuvent avoir un impact significatif sur la productivité de l'entreprise, la santé et le bien-être des travailleurs et le coût des prestations d'assurance.

Au fil des années, les RPS ont pris de plus en plus d'importance dans le monde professionnel. Les organisations ont commencé à mettre en place des mesures pour prévenir et gérer les RPS, telles que des programmes de soutien psychologique, des politiques de prévention du harcèlement et des programmes de formation pour les travailleurs et les gestionnaires.

Cependant, malgré les efforts de prévention et de gestion, les RPS continuent d'être un problème de santé au travail majeur, en particulier chez les jeunes travailleurs. Les entreprises doivent

donc continuer à investir dans des stratégies de prévention et de gestion des RPS pour protéger la santé et le bien-être de leurs employés, ainsi que pour améliorer leur productivité et leur rentabilité à long terme.

V. Modélisation de la durée des arrêts de travail

a. Modèle de durée : censures et troncatures

Pour modéliser le risque arrêt de travail sur notre portefeuille nous avons procédé à un travail de nettoyage et de cohérence minutieux sur nos données, cela est primordial pour avoir une modélisation qui soit la plus fiable possible. Cependant, l'information que nous avons actuellement dans notre base de données de sinistres arrêts de travail reste encore biaisée à cause de la nature des arrêts (cf. *Non Parametric Estimation from Incomplete Observation* E.L. Kaplan et P.Meier) et plus précisément de leur durée et de la période d'observation, on parle d'arrêts tronqués et/ou censurés, nous présenterons les types de censures et troncatures présentes dans nos données qui sont la troncature à gauche et la censure à droite de type I.

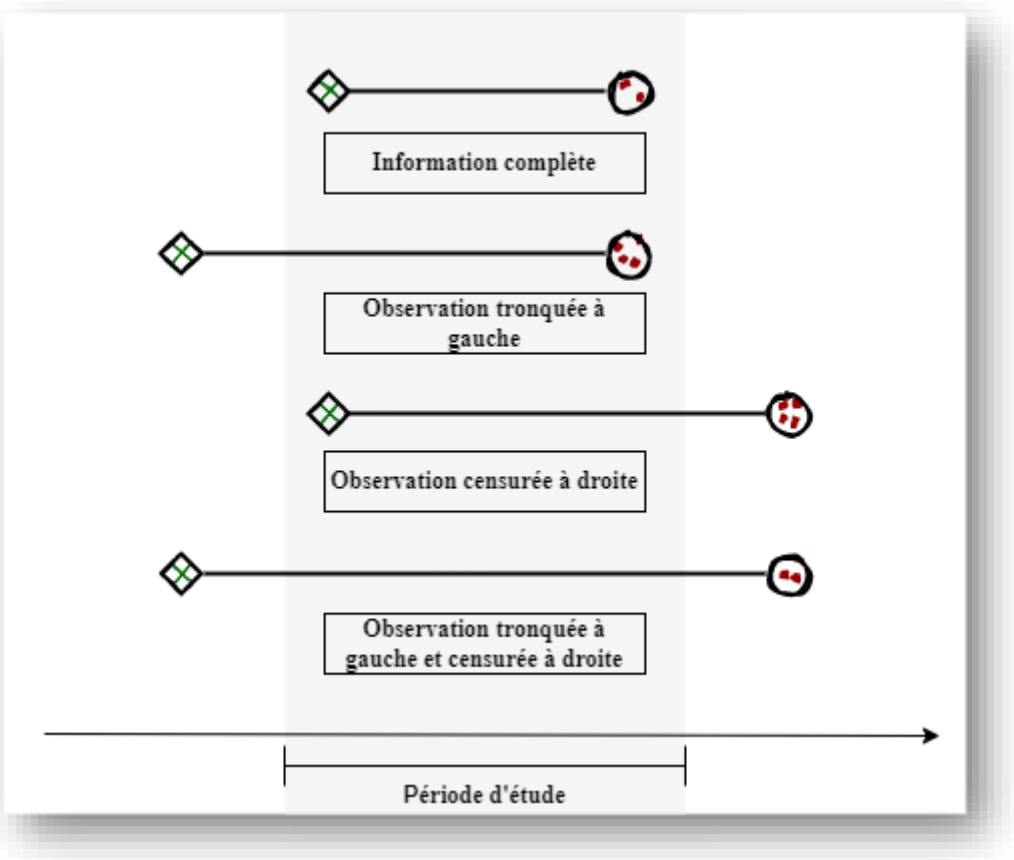


Figure 10 : Schéma représentant les phénomènes de censures et de troncatures

i. Censure à droite :

Certains arrêts de travail ne sont pas observés jusqu'au bout, on parle de censure à droite de l'information sur la durée de l'arrêt de travail, bien que nous ayons une date de fin d'observation dans nos données, la période réelle d'arrêts est plus longue que celle que nous captons dans nos données.

Considérons un échantillon de durée de survie $(X_1, \dots, X_n)$ qui dans notre cas représenterait la durée réelle des arrêts de travail dans notre portefeuille, ce que l'on observe dans notre base de données sont en fait les observations $(T_1, D_1), \dots, (T_n, D_n)$ avec :

$$T_i = X_i \wedge C_i \ \ et \ D_i = \begin{cases} 1 \ si \ X_i \leq C_i \\ 0 \ sinon \end{cases}$$

La variable D indique si l'observation est censurée ou non, et C représente la date de censure, dans la pratique, nous observons trois types de censures :

Nous observons deux causes principales de censures : la censure causée par l'observation sur un intervalle de temps : tout arrêt de travail continuant au-delà de fin 2022 est censuré dû à la date d'extraction des données, dans ce cas C_i=31/12/2022. La deuxième cause est la censure pour départ à la retraite, C_i=date de départ à la retraite de l'assuré.

La troisième et dernière cause traitée en tant que censure dans cette étude est la sortie au bout des 3 ans d'incapacité.

ii. Troncature à gauche :

La troncature gauche se produit lorsque la variable d'intérêt n'est pas observable en dessous d'un certain seuil c. Contrairement à la censure, la troncature signifie que l'information concernant les observations en dehors de cette plage est entièrement perdue, car on ne dispose d'aucune donnée concernant ces observations. En revanche, dans le cas de la censure, on sait qu'il existe une information, mais on ne dispose que d'une limite supérieure ou inférieure pour cette information.

La troncature peut être observée dans différentes situations, par exemple lors d'une migration informatique où seuls les sinistres en cours ont été transférés dans la nouvelle base de données, entraînant la perte d'informations sur les sinistres de durée plus courte pour les mêmes événements. Un autre exemple est celui du contrat d'arrêt de travail avec une franchise : les arrêts de durée inférieure à la franchise ne sont pas observés, ce qui signifie qu'aucune information n'est disponible sur eux.

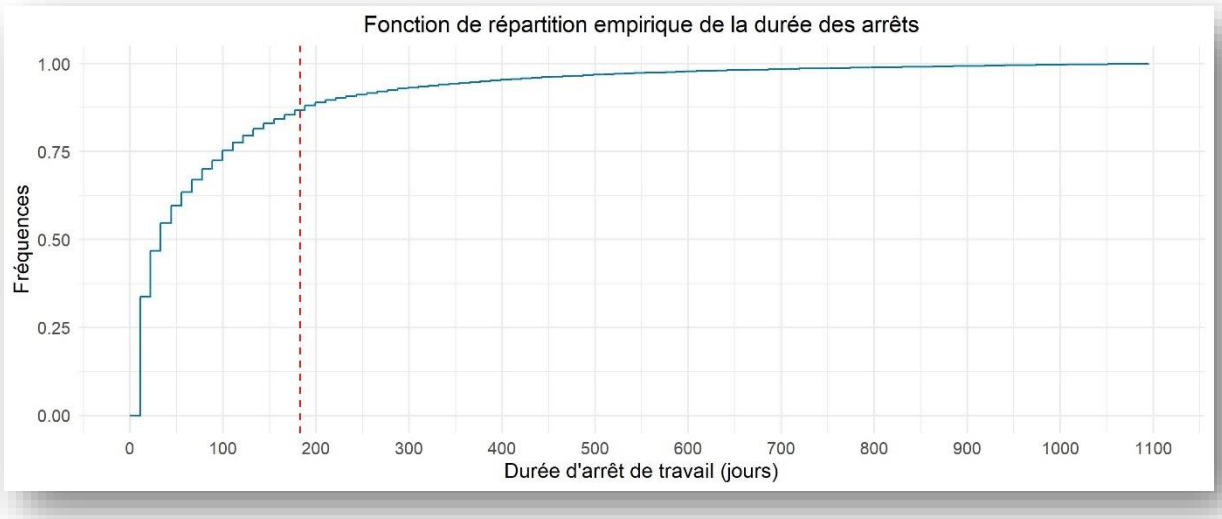
Nous avons sélectionné pour notre étude les observations qui ont eu lieu durant la période d'étude, c'est-à-dire celles qui ne sont ni tronquées ni censurées, ainsi que celles qui présentent une information partielle, les observations censurées à droite.

b. Etude de la distribution de la durée des arrêts de travail

L'idée de cette partie est de déterminer la loi continue qui s'ajustera au mieux à nos données, afin de calibrer au mieux nos modèles linéaires généralisés. Pour ce faire nous allons comparer la distribution de nos données avec la fonction de répartition de plusieurs lois usuelles, en estimant les paramètres des lois.

i. Comparaison des fonctions de répartition

Soit $\widehat{F_\theta}$ la fonction de répartition empirique de la durée d'incapacité dont le tracé se trouve ci-dessous, nous nous intéresserons aussi à la proportion des arrêts durant moins d'une demi-année afin de savoir s'il est intéressant ou non d'effectuer deux modélisations différentes pour les arrêts courts (< 6 mois) et les arrêts longs



Nous observons que plus de 80% des arrêts durent moins de 6 mois (représenté par l'axe vertical en point tillés sur le graphe), ce seuil est très important pour la sécurité sociale, en effet pour que l'indemnisation d'un arrêt de travail puisse dépasser 6 mois, il faut impérativement le service médical de l'Assurance Maladie donne son accord et que l'état de santé de la personne arrêtée le justifie. Nous verrons de notre côté dans la suite s'il est pertinent ou non mathématiquement d'affiner nos modèles en effectuant cette séparation des arrêts.

ii. Estimation des paramètres de lois :

Les lois usuelles que nous allons utiliser pour la comparaison avec notre distribution sont les lois suivantes :

- Loi Gamma $\Gamma(n, \gamma)$
- Loi Log-Normal $\mathcal{N}(\mu, \sigma^2)$
- Loi Inverse-Gaussienne $\mathcal{IG}(\eta, \theta)$
- Loi Exponentielle $\mathcal{E}(\lambda)$

Nous estimerons les paramètres de ces lois à partir des moyennes et des variances empiriques de nos échantillons que nous calculerons grâce aux estimateurs sans biais suivants :

$$\overline{X} = \frac{\sum_{i=1}^{m} X_i}{m} \text{ et } S^2 = \frac{\sum_{i=1}^{m} (X_i - \overline{X})}{m-1}$$

Avec $(X_1, \dots, X_m)$ un vecteur de variables iid de même loi que la variable à expliquer

Les paramètres des lois sont estimés de la façon suivante :

- Loi Gamma $\Gamma(n, \gamma)$:

$$\hat{n} = \frac{\overline{X}^2}{S^2} \text{ et } \hat{\gamma} = \frac{\overline{X}}{S^2}$$

- Loi Log-Normal $\mathcal{N}(\mu, \sigma^2)$:

$$\hat{\mu} = \frac{\tilde{X}^2}{\tilde{S}^2} \text{ et } \widehat{\sigma^2} = \tilde{S}^2$$

Avec $\tilde{X}$ et $\tilde{S}$ respectivement les estimateurs de la moyenne et la variance de la variable durée à expliquer composé par la fonction ln.

- Loi Inverse-Gaussienne $\mathcal{IG}(\eta, \theta)$:

$$\hat{\eta} = \overline{X} \text{ et } \hat{\theta} = \frac{\overline{X}^3}{S^2}$$

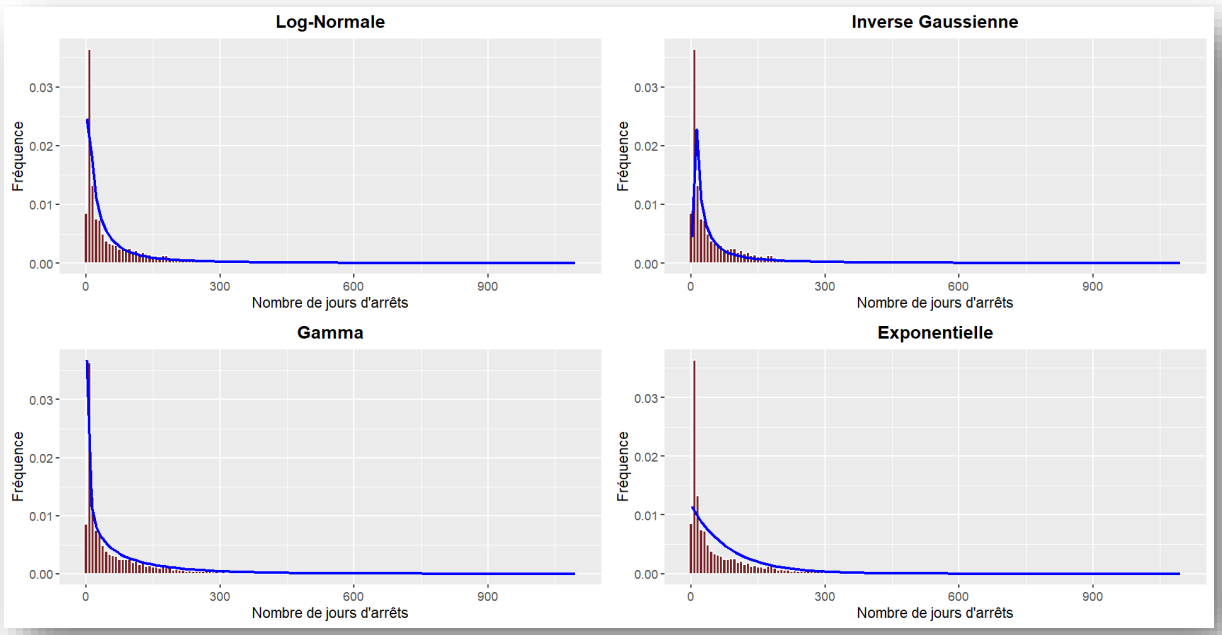
- Loi Exponentielle $\mathcal{E}(\lambda)$:

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

iii. Analyse graphique et comparaison des lois

1. Comparaison des densités

Nous nous baserons sur une analyse graphique dans un premier temps pour avoir une idée de la loi qui s'ajuste le mieux à nos données, pour ce faire nous allons nous baser sur l'histogramme des durées d'arrêts d'incapacité.



Nous avons tracé en bleu les densités des lois testées, on remarque que la loi gamma et la loi inverse gaussienne s'ajustent correctement à nos données. Cela reste une première impression graphique, nous confirmerons ou infirmerons cela dans la suite grâce à nos modèles linéaires généralisés.

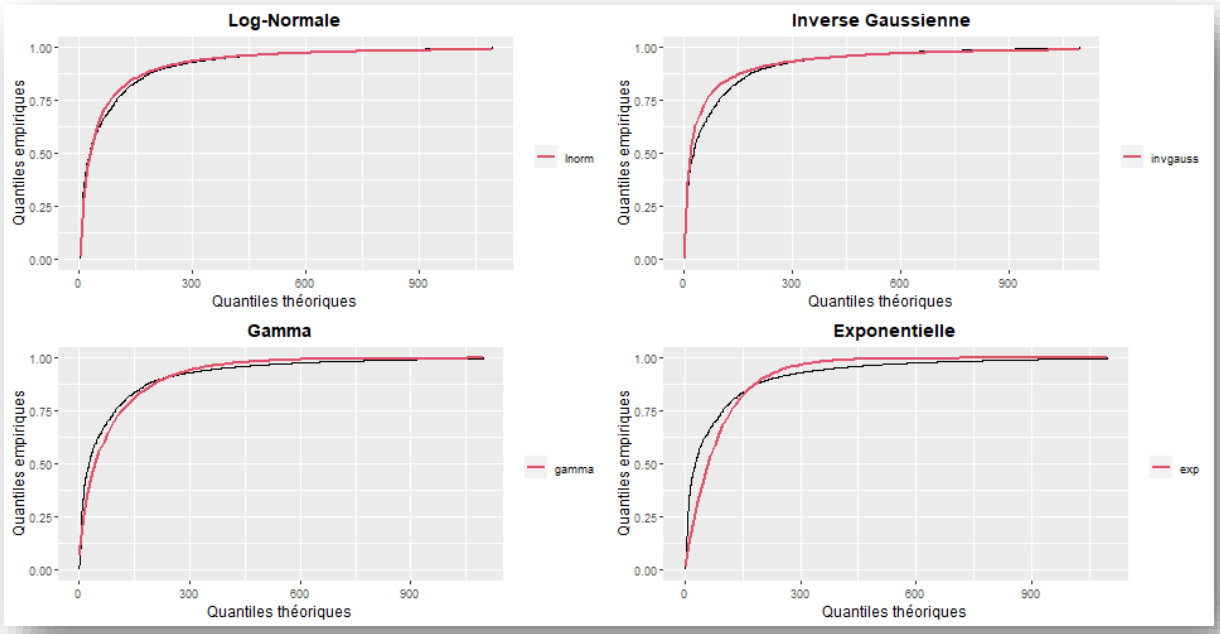
2. Comparaison des QQ-plot

Le Q-Q Plot (ou graphique quantile-quantile) est un outil qui permet de visualiser si une distribution de données suit une distribution théorique donnée, comme la distribution normale, par exemple. Nous construisons le graphique en représentant les quantiles empiriques des données en abscisse et les quantiles théoriques de la distribution en ordonnée.

Pour expliquer cette méthode, reprenons la notation précédentes $(X_{(1)}, \dots, X_{(m)})$ pour décrire notre échantillon de données triées par ordre croissant de la variable à expliquer (=durée des arrêts), F_θ la fonction de répartition de la loi usuelle testée. Nous nous basons sur l'hypothèse d'adéquation suivante :

$$\forall i \in [\![1, m]\!], \qquad F_\theta(X_{(i)}) \approx \frac{i}{m}$$

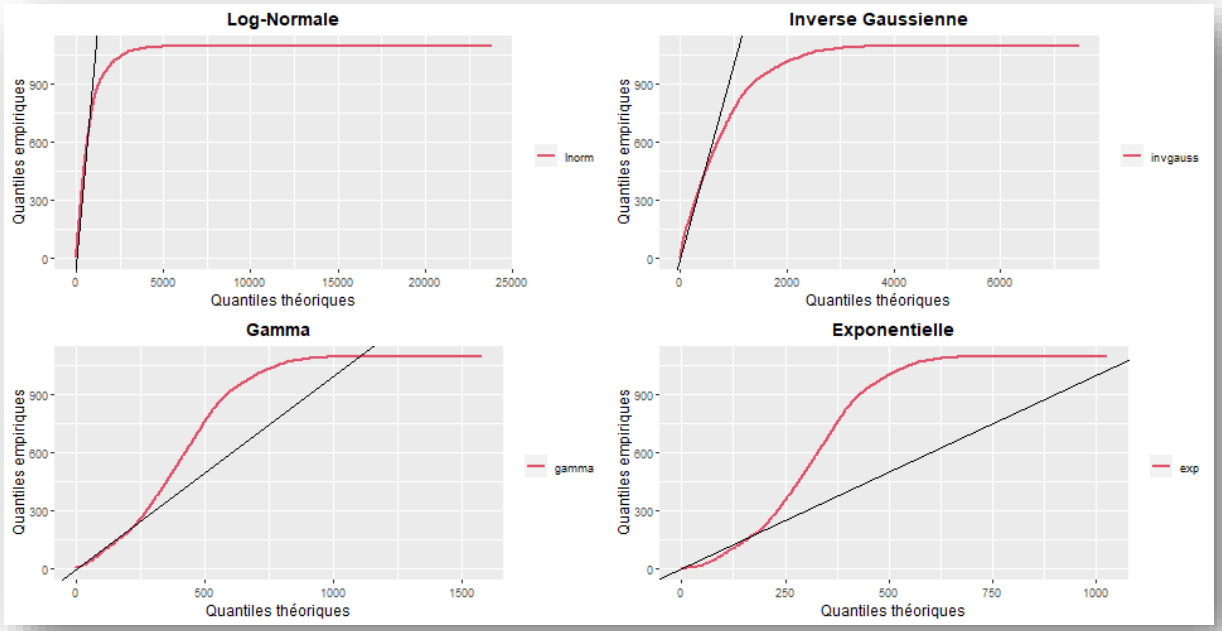
Dans un premier temps, nous allons comparer les fonctions de répartitions calibrées avec les paramètres estimés avec la fonction de répartition empirique de nos données :



Nous utilisons le Q-Q Plot en comparant les quantiles de la distribution théorique aux quantiles de la distribution empirique des données. Si la distribution des données suit parfaitement la distribution théorique, les points du graphique sont alignés sur une droite.

Nous pouvons également utiliser le Q-Q Plot pour comparer la distribution de deux ensembles de données différents en les superposant sur le même graphique. Cela nous permet de visualiser rapidement les différences entre les deux distributions.

Ci-dessous les QQ-plot des lois usuelles considérées :



Nous remarquons que les lois Log-Normale et Inverse Gaussienne s'ajustent particulièrement bien à la variable à expliquer. La loi Gamma reste cela dit intéressante nous l'utiliserons elle aussi dans la suite de l'étude alors que nous excluons la loi exponentielle de notre étude aux vues de cette analyse graphique.

c. Modélisation des arrêts grâce aux modèles linéaires généralisés

i. Objectif de la modélisation

L'objectif principal de cette étude est de modéliser la durée des arrêts de travail à partir des données de sinistres arrêt de travail stockées dans notre base de données. Plus précisément, nous cherchons à évaluer les facteurs qui influencent la durée des arrêts de travail et à déterminer les variables qui ont le plus d'impact sur cette durée.

Pour atteindre cet objectif, nous allons utiliser dans un premier temps des modèles de régression linéaire généralisée. Ces modèles sont une classe de modèles de régression qui permettent de modéliser des relations entre une variable réponse et des variables explicatives. Contrairement aux modèles de régression linéaire classiques, les GLM peuvent être utilisés pour modéliser des variables réponse qui ne suivent pas une distribution normale, telles que des variables binaires ou de comptage. Dans notre cas, nous allons utiliser des GLM pour modéliser la durée des arrêts de travail, qui est une variable de comptage.

En utilisant des modèles GLM, nous pourrons évaluer l'impact de différentes variables explicatives sur la durée des arrêts de travail, et déterminer celles qui ont le plus d'influence. Cela nous permettra de mieux comprendre les facteurs qui influencent la durée des arrêts de travail et de développer des stratégies pour réduire cette durée.

ii. Données utilisées

Les données utilisées dans cette étude proviennent du portefeuille d'assurance collective de Generali. Cette base de données contient des informations sur les arrêts de travail, telles que la date de début et de fin de l'arrêt, la durée de l'arrêt, l'âge et le sexe du travailleur, ainsi que sa catégorie socio-professionnelle.

Les données que nous utilisons ont été collectées sur une période de plus de 10 ans, depuis 2011. Les données sont anonymisées et ont été traitées afin de supprimer les valeurs aberrantes et les données manquantes.

Les variables explicatives que nous allons considérer dans nos modèles sont les suivantes :

- Age
- Sexe
- Catégorie socio-professionnelle

iii. Présentation des variables explicatives

Dans le cadre d'une modélisation GLM, il est important de segmenter les variables explicatives en différentes catégories, en fonction de leur nature et de leur impact potentiel sur la variable réponse. Cette segmentation permet de mieux comprendre les relations entre les variables et d'identifier les effets spécifiques de chaque variable sur la variable réponse.

Par exemple, en segmentant la variable continue de l'âge en différentes catégories, nous pouvons mieux comprendre comment l'impact de l'âge sur la durée des arrêts de travail varie en fonction de l'âge des travailleurs. De même, en segmentant la variable catégorielle de la catégorie socio-professionnelle en différentes catégories (cadre, non cadre, ensemble du

personnel), nous pouvons mieux comprendre comment l'impact de la catégorie socio-professionnelle sur la durée des arrêts de travail.

En segmentant les variables explicatives, nous pouvons également détecter des interactions entre les variables, c'est-à-dire des effets croisés entre les variables qui peuvent affecter la variable réponse de manière non linéaire. Ces interactions peuvent être modélisées à l'aide de termes d'interaction dans le modèle GLM, ce qui permet de mieux capturer la complexité de la relation entre les variables explicatives et la variable réponse.

1. *La variable Age*

Nous allons tenter de créer des classes d'âge homogènes en termes d'effectif, pour cela il s'agira de tracer dans un premier temps la distribution du nombre de travailleur arrêté en fonction de leur âge à la survenance du sinistre. Ci-dessous le graphique du nombre de sinistre par âge :

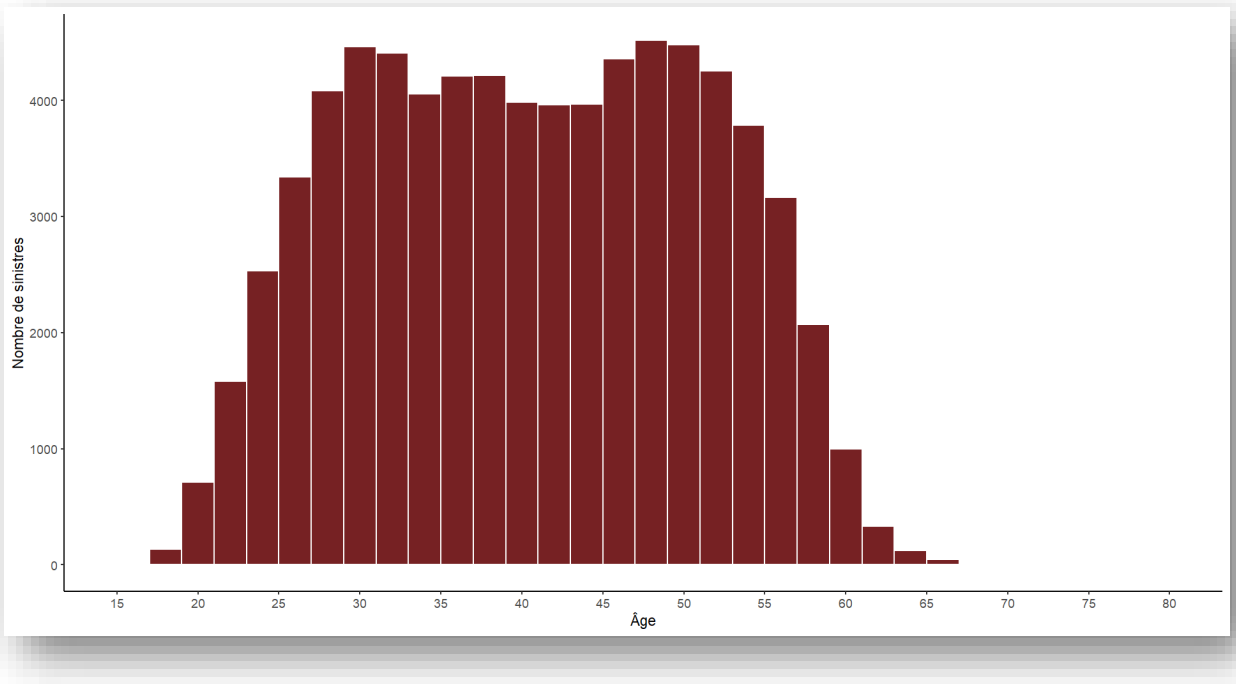


Figure 11 : Nombre de sinistres par âge à la survenance

On observe une distribution bimodale du nombre de sinistres sur notre portefeuille, nous cherchons à construire des classes homogènes.

Pour segmenter nos données en classes d'âges, nous calculons les quantiles de la distribution du nombre de sinistres d'ordre 20%, 40%, 60%, 80% et 100%. Nous obtenons alors 5 classes, [18,31[, [31,38[, [38,45[, [45,52[, [52,78]. Ci-dessous le graphique qui présente le nombre de sinistres observés par classe d'âge :

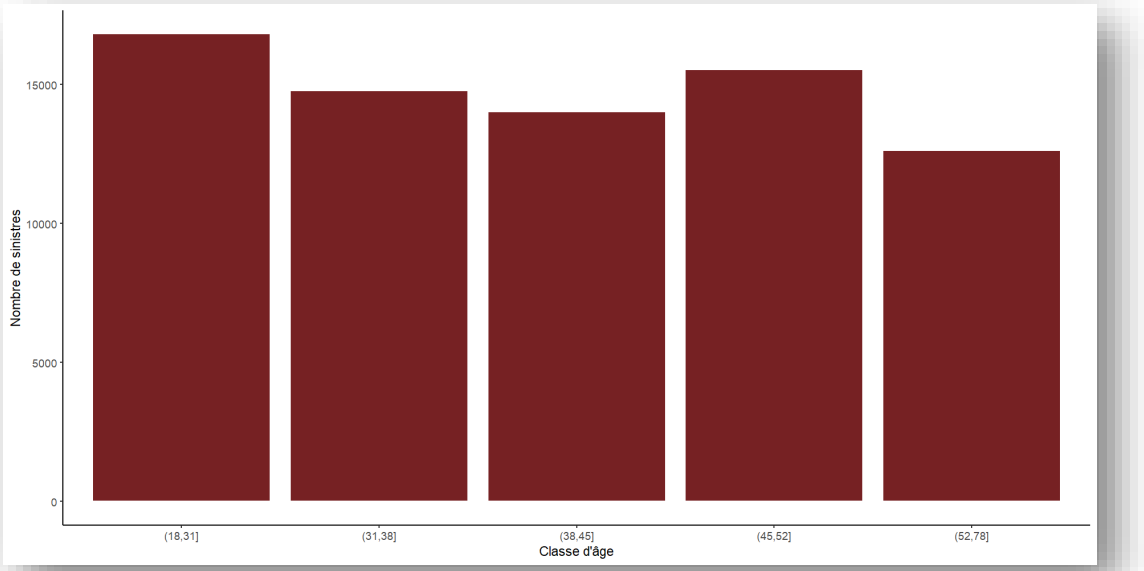


Figure 12: Nombre de sinistres par classe d'âge

2. *La variable Sexe*

La variable Sexe est naturellement segmentée sous les classes « M » et « F », ci-dessous la distribution des sinistres par sexe :

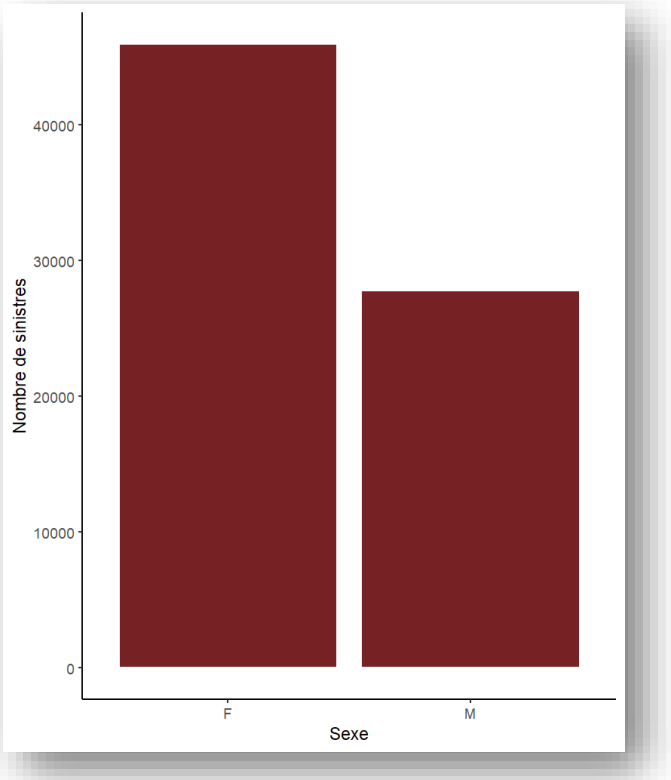


Figure 13: Nombre de sinistres par sexe

Le nombre de sinistres arrêt de travail est 65,6% plus élevé chez les femmes que chez les hommes. Nous nous intéresserons dans la suite à l'impact de la variable sexe sur nos modèles.

3. *La variable Catégorie Socio-Professionnelle*

Nous avons segmenté la variable catégorie socio-professionnelle en trois classes : les cadres, les non-cadres et l'ensemble du personnel. Ci-dessous la distribution du nombre d'arrêts par classe :

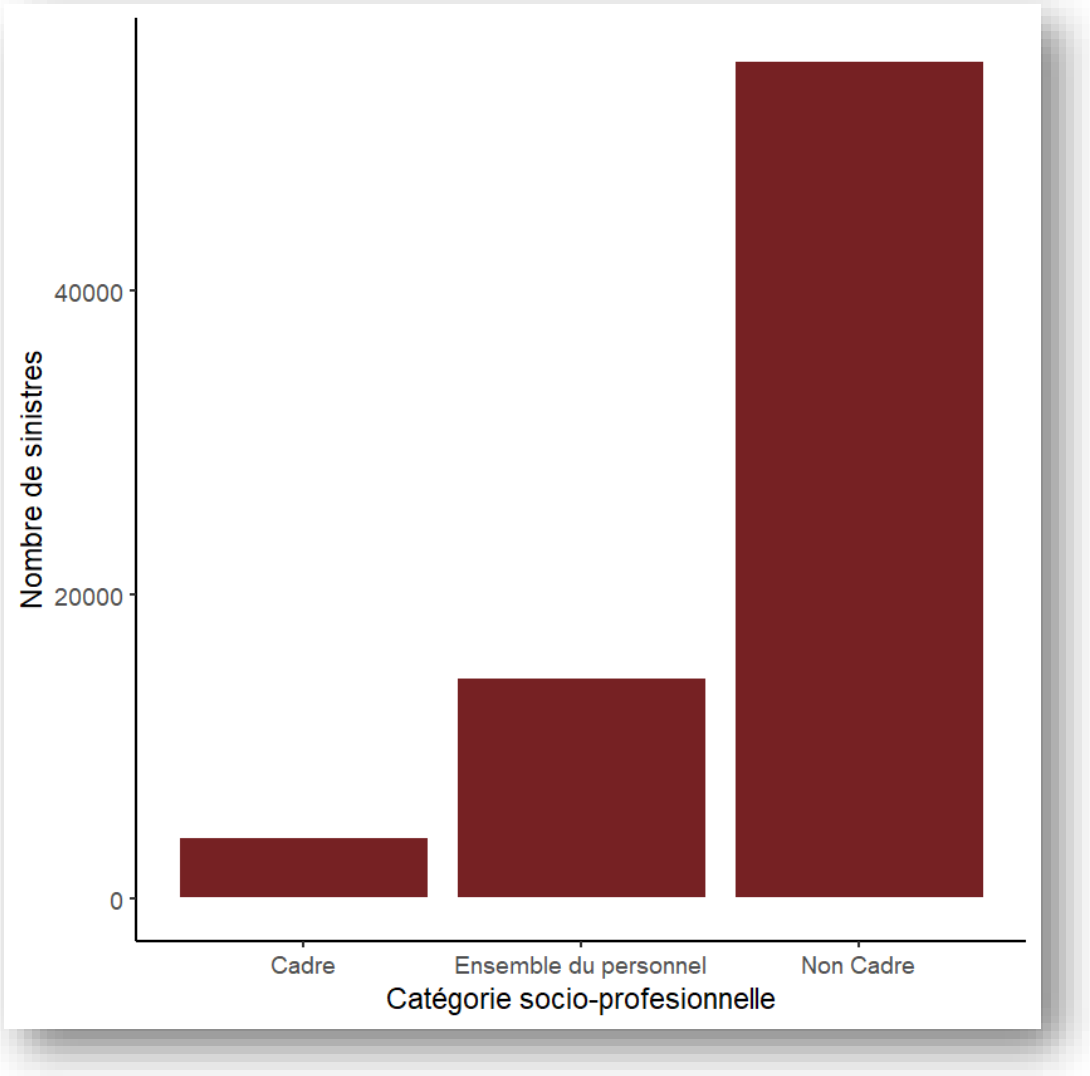


Figure 14: Nombre de sinistres par catégorie socio-professionnelle

Le nombre de sinistres généré par les salariés non-cadres est très largement supérieur au nombre de d'arrêts de travail par les cadres, cela dit cette statistique est à mettre en perspective avec la taille des effectifs cadres et non-cadres, d'où l'intérêt des modèles linéaires généralisés que nous allons présenter.

iv. Aspect théorique des modèles linéaires généralisés

Les modèles linéaires généralisés ont été introduits par Nelder et Wedderburn en 1972, ils constituent la base de référence pour modéliser l'effet des variables de segmentation sur un tarif

en assurance. Ces modèles sont adaptés à de nombreuses problématiques courante dans le domaine de la statistique et de l'actuariat.

1. *Intérêt des modèles linéaires généralisés*

Les modèles linéaires généralisés sont des modèles régulièrement utilisés en assurance que ce soit en assurance santé (remboursements soins, frais d'hospitalisation), en assurance auto (dommage matériel, vol, …), en assurance MRH (incendie, vol, dégâts des eaux, …).

Les GLM permettent de :

- Modéliser des réponses diverses
- Intégrer toute type d'information exogène susceptible d'influer sur la variable dépendante (réponse Y)
- Quantifier l'impact des facteurs de risque X (sens/intensité)
- Résidus hétéroscédastiques (la loi varie par profil)

Cependant, leur mise en place nécessite d'introduire deux hypothèses fondamentales :

- Les données que l'on cherche à expliquer sont indépendants entre elles
- Les variables explicatives X sont indépendantes deux à deux

Ces hypothèses sont très importantes notamment car l'un des intérêts des modèles linéaires généralisés en assurance est la tarification et donc le calcul d'une prime pure, calculée généralement de la façon suivante avec une approche fréquence-sinistre : $\mathbb{E}[S|X] = \mathbb{E}[N|X] * \mathbb{E}[Y_i|X]$

v. Principe des GLM

1. *Composant d'un GLM :*

Les différents composants d'un modèle linéaire généralisé sont :

- **La loi de la réponse aléatoire Y_i :** par hypothèse la distribution de cette loi appartient à la famille exponentielle

Rappel : La densité de probabilité d'une loi appartenant à la famille exponentielle s'écrit de la façon suivante :

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

Avec :

- Avec θ le paramètre naturel de la dispersion,
- φ le paramètre de dispersion, b une fonction définie sur $\mathbb{R}$ deux fois dérivable et de dérivée première injective,
- c une fonction définie sur $\mathbb{R}^2$

De plus on a en particulier avec les lois de cette famille :

$$\mathbb{E}[Y] = b'(\theta) \text{ ainsi que } \mathbb{V}[Y] = b''(\theta)\phi$$

- **Le prédicteur** noté η_i et défini tel que $\eta_i = \sum_{j=1}^{J} \beta_j X_{ij}$ est linéaire et déterministe : il est calculé via les facteurs de risques explicatifs
- **La fonction de lien g :** elle est par définition monotone, dérivable et inversible. Elle est définie et utilisée en pratique de la façon suivante :

$$g(\mathbb{E}[Y_i]) = \eta_i$$

En pratique, il faut adapter la fonction de lien au domaine de définition de Y, voici un tableau qui présente les différentes fonction lien utilisées en pratique :

Loi	Lien canonique	Moyenne
Normale $\mathcal{N}(\mu, \sigma^2)$	Identité : $\eta = \mu$	$\mu = X\beta$
Binomiale $\mathcal{B}(\mu)$	Logit : $\eta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$
Poisson $\mathcal{P}(\mu)$	Log : $\eta = \ln(\mu)$	$\mu = \exp(X\beta)$
Gamma $\mathcal{G}(\mu)$	Inverse : $\eta = \frac{1}{\mu}$	$\mu = \frac{1}{X\beta}$
Inverse-Gaussienne $\mathcal{IG}(\mu, \lambda)$	Inverse² : $\eta = \frac{1}{\mu^2}$	$\mu = \frac{1}{(X\beta)^2}$

La fonction de lien canonique est la fonction lien qui associe la moyenne μ au paramètre canonique θ. Elle est telle que :

$$g(\mu_i) = \theta_i$$

A présent, présentons quelques modèles fréquemment utilisés dans le contexte assurantiel :

2. *Modèle gaussien :*

Dans le cas d'un échantillon gaussien, les densités d'une famille de lois $N(\mu_i, \sigma^2)$ s'écrivent :

$$f(y_i, \mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right]$$

Et appartiennent à la famille exponentielle en posant :

$$\theta_i = E(Yi) = \mu_i$$
$$a(\theta_i) = \exp\left[-\frac{{\mu_i}^2}{2\sigma^2}\right]$$
$$b(\theta_i) = \frac{\theta_i}{\phi} = \frac{\mu_i}{\sigma^2}$$
$$c(y_i) = \exp\left[-\frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2}\right]$$

On remarque avec la première égalité que la famille gaussienne se met sous forme canonique

3. *Modèle inverse-gaussien :*

Le modèle GLM inverse-gaussien est un type de modèle de régression généralisé qui est utilisé pour modéliser des données avec une distribution inverse-gaussienne. La distribution inverse-gaussienne est caractérisée par des valeurs positives, des queues de distribution lourdes et une asymétrie positive. Elle est souvent utilisée pour modéliser des données de durée ou de temps jusqu'à un événement, ce qui correspond particulièrement bien avec l'objectif de ce mémoire.

Dans un modèle GLM inverse-gaussien, la fonction de lien est la fonction inverse. La variance est proportionnelle au carré de la moyenne de la variable dépendante. L

L'estimation des paramètres du modèle est effectuée à l'aide de la méthode du maximum de vraisemblance. Les prévisions du modèle peuvent être obtenues à l'aide de la fonction de lien inverse.

Le modèle GLM inverse-gaussien est souvent utilisé pour modéliser des données de durée d'arrêt maladie, de temps jusqu'à la guérison d'une maladie, ou de temps entre les événements de naissance ou de décès. Ce modèle est très intéressant car il peut permettre de prendre en compte la distribution asymétrique et à queue lourde des données de durée, ce qui peut être difficile à modéliser avec des méthodes de régression linéaire traditionnelles.

4. Modèle de Poisson :

Le modèle de Poisson est un type de modèle de régression utilisé pour modéliser des variables de réponse qui suivent une distribution de Poisson. La distribution de Poisson est souvent utilisée pour modéliser des variables qui comptent le nombre d'événements qui se produisent au cours d'une période donnée, comme le nombre de sinistres d'assurance sur une période donnée.

Le modèle GLM de Poisson utilise une fonction de lien logarithmique pour lier la variable de réponse à un ensemble de variables explicatives (également appelées variables prédictives ou variables indépendantes). Cela signifie que la relation entre la variable de réponse et les variables explicatives est modélisée comme une régression linéaire dans le logarithme de la variable de réponse. Cela permet de modéliser des variables de réponse qui peuvent prendre des valeurs entières uniquement et qui ont des variances qui augmentent avec la moyenne.

On considère n variables indépendantes Yi de loi de Poisson de paramètre μi = E(Yi). Les Yi sont par exemple les effectifs d'une table de contingence. Ces variables admettent pour densités :

$$f(y_i, \mu_i) = \frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!} = \exp(-\mu_i) * \frac{1}{y_i!} \exp(y_i \ln(\mu_i!))$$

$$\text{Avec : } \theta_i = \ln(\mu_i)$$

Et donc la fonction de lien canonique de ce modèle est la fonction logarithme népérien.

5. Modèle Log-Normal :

Le modèle log normal est un type de modèle de régression utilisé pour modéliser des variables de réponse qui suivent une distribution log normale. La distribution log normale est souvent utilisée pour modéliser des variables qui ont des valeurs positives et sont asymétriques, comme les revenus ou les tailles.

Le modèle GLM log normal utilise une fonction de lien logarithmique pour lier la variable de réponse à un ensemble de variables explicatives. Cela permet de modéliser des variables de réponse qui ont des valeurs positives et des variances qui augmentent avec la moyenne.

On considère n variables indépendantes Y_i de loi log normale de paramètre μ_i et σ^2. Les Y_i sont par exemple les revenus d'un échantillon. Ces variables admettent pour densités :

$$f(y_i, \mu_i, \sigma^2) = \left(\frac{1}{y_i \sigma \sqrt{2\pi}}\right) \exp\left(-\frac{(\ln(y_i) - \mu_i)^2}{2\sigma^2}\right)$$

$$\text{Avec : } \theta_i = \ln(\mu_i)$$

Et donc la fonction de lien canonique de ce modèle est la fonction logarithme népérien.

6. Modèle Gamma :

Le modèle GLM Gamma est un modèle de régression généralisé qui est souvent utilisé pour modéliser des données de coûts ou de temps de travail. La distribution Gamma est une distribution à valeur positive, caractérisée par une asymétrie positive et une queue de distribution lourde. Cette distribution est souvent utilisée pour modéliser des données de coûts, car elle prend en compte le fait que les coûts sont toujours positifs et souvent concentrés autour de zéro.

Dans un modèle GLM Gamma, la fonction de lien est la fonction logarithmique, qui permet de capturer les effets multiplicatifs des variables explicatives sur la variable dépendante. La variance est proportionnelle au carré de la moyenne de la variable dépendante, comme dans le modèle GLM inverse-gaussien.

L'estimation des paramètres du modèle se fait à l'aide de la méthode du maximum de vraisemblance, et les prévisions du modèle peuvent être obtenues à l'aide de la fonction de lien inverse exponentielle.

Dans le cadre de ce mémoire, le modèle GLM Gamma sera utile pour modéliser la durée des arrêts de travail sur notre portefeuille d'assurance collective. En utilisant ce modèle, nous pourrons identifier les variables explicatives qui ont le plus grand impact sur la durée des arrêts maladie, et évaluer l'efficacité des mesures de prévention de l'absentéisme. La fonction de lien logarithmique permettra de capturer les effets multiplicatifs des variables explicatives sur la durée de l'arrêt.

7. Modèle Logistique :

Le modèle logistique est un modèle de régression utilisé pour modéliser des variables de réponse binaires, c'est-à-dire des variables qui prennent seulement deux valeurs possibles, comme oui ou non, vrai ou faux, succès ou échec. Dans le contexte d'une étude sur les arrêts de travail, une application intéressante des modèles logistiques est la modélisation de l'incidence d'un arrêt de travail, cela nécessite d'avoir l'information sur l'ensemble des salariés en portefeuille et non pas uniquement les sinistrés. La méthode classique pour étudier l'incidence des arrêts de travail est l'utilisation d'un modèle de Cox pour l'estimation des taux d'incidence de façon analogue à l'estimation des taux de mortalité.

Ce modèle utilise une fonction de lien logistique pour lier la variable de réponse binaire à un ensemble de variables explicatives. Cette fonction de lien logistique permet de modéliser la probabilité de la variable de réponse en fonction des variables explicatives en transformant la réponse binaire en une probabilité qui varie de 0 à 1.

La fonction de lien canonique pour le modèle logistique est la fonction logit, présentée ci-dessus. La fonction logit transforme la probabilité de la variable de réponse en une variable continue qui varie de moins l'infini à plus l'infini. Cette variable continue est ensuite modélisée comme une régression linéaire des variables explicatives.

vi. Validation d'un modèle GLM :

Pour valider le modèle linéaire généralisé étudié, il existe les différentes méthodes suivantes :

1. Validation croisée (k-fold)

En apprentissage automatique, la validation croisée k-fold est une technique essentielle pour évaluer les performances d'un modèle et estimer son erreur. Cette méthode implique de diviser aléatoirement un ensemble de données en k sous-ensembles de taille égale. Chaque sous-ensemble est utilisé une fois comme ensemble de validation tandis que les autres sous-ensembles sont utilisés pour entraîner le modèle.

Cette procédure est répétée k fois pour que chaque sous-ensemble soit utilisé une fois comme ensemble de validation. Ainsi, on obtient k estimateurs différents du modèle, chacun entraîné sur des données différentes. En moyennant ces k estimateurs, on obtient l'estimateur final, qui est utilisé pour évaluer les performances du modèle en calculant l'erreur moyenne sur l'ensemble de validation.

La validation croisée k-fold est particulièrement avantageuse car elle permet d'estimer l'erreur du modèle de manière fiable et robuste, même avec un petit ensemble de données. Cela est particulièrement important dans les cas où l'on dispose de peu de données et où il est nécessaire de maximiser la précision des estimations. Ainsi, la validation croisée k-fold est une technique incontournable pour l'évaluation des modèles d'apprentissage automatique.

2. Validation de la significativité globale du modèle

Pour valider la significativité globale du modèle, il est judicieux de travailler avec des critères de qualité d'ajustement du modèle comme la déviance ou la statistique de Pearson que nous définissons ci-dessous :

- La déviance :

Pour définir la déviance du modèle, nous introduisons la notion de modèle saturé qui est le modèle qui possède autant de paramètres que d'observations. Il est caractérisé par l'égalité entre

les facteurs estimés et les observations, c'est-à-dire : $\hat{\mu}_i = y_i$. Puis nous comparerons ce modèle avec le modèle d'étude pour calculer la déviance.

Il faut aussi introduire la log-vraisemblance pour calculer la déviance :

Soit Y =$(Y_1, \dots, Y_n)$ le vecteur à expliquer (appartenant à la famille exponentielle) dont la densité s'écrit sous la forme :

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

La log-vraisemblance de ce modèle s'écrit :

$$l(y,\theta,\phi) = \sum_{i=1}^{n} \ln\left(f\left(y_i|\theta_i,\phi\right)\right)$$

$$= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + \sum_{i=1}^{n} c_i(y_i,\phi)$$

Nous utiliserons comme mesure de la qualité d'ajustement du modèle la statistique du rapport de vraisemblance suivante :

$$\lambda = \frac{L_{SAT}}{L} \qquad \text{ou} \qquad \ln(\lambda) = \ln(L_{SAT}) - \ln(L)$$

Car le modèle décrit bien les données lorsque L $\simeq L_{SAT}$

Enfin, on définit la statistique de déviance réduite ou normalisée de la façon suivante :

$$D = 2\ln(\lambda) = 2(\ln(L_{SAT}) - \ln(L))$$

La déviance non réduite est :

$$D^* = \phi D$$

De plus, $D \sim \chi^2_{n-p-1}$, c'est-à-dire que le modèle est de mauvaise qualité si $D_{obs} > \chi^2_{n-p-1;1-\alpha}$

- La statistique de Pearson :

La statistique du test de significativité globale de Pearson est définie de la façon suivante :

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(y_i)} = \frac{1}{\phi}\sum_{i=1}^{n} \omega_i \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

Comme pour la Déviance, nous avons $X^2 > \chi^2_{n-p-1;1-\alpha}$

3. *Validation de la significativité individuelle des coefficients de la régression*

Pour valider ou non le modèle en étudiant la significativité individuelle des coefficients de la régressions, nous pouvons utiliser des tests d'hypothèses sur les paramètres, pour cela nous introduirons l'hypothèse H_0 de la façon suivante :

$$H_0 : \boldsymbol{C}\beta = r$$

Avec : **C** une matrice connue possédant q lignes et r un ensemble de valeurs testées.

Ensuite nous pourrons considérer trois approches pour valider ou non cette hypothèse :

- Test du rapport de vraisemblance :

Nous avons défini le rapport de vraisemblance précédemment : λ mais dans ce cas nous n'utiliserons pas un modèle saturé mais un modèle sans contrainte de vraisemblance $\hat{L}$ et un modèle sous contrainte de vraisemblance $\tilde{L}$.

Ainsi nous le rapport de vraisemblance s'exprimera : $\lambda = \frac{\hat{L}}{\tilde{L}}$

La statistique de ce test est :

$$D = 2\ln(\lambda) \sim \chi_q^2$$

- Test de Wald :

Pour définir la statistique du test de Wald, il faut dans un premier temps définir l'estimateur du maximum de vraisemblance de β, $\hat{\beta}$.

$$\hat{\beta} \sim \aleph(\beta, \phi(X'WX)^{-1})$$

Avec W la matrice diagonale de pondération définie telle que :

$$[W]_{i,i} = \frac{1}{Var(Y_i)}\left(\frac{\partial y_i}{\partial \mu_i}\right)^2$$

Et donc avec les notations précédentes, sous H_0, $\boldsymbol{C}\hat{\beta} - r \sim \aleph(0, \phi C(X'WX)^{-1}C')$

Enfin, la statistique du test de Wald est définie comme :

$$\left(\boldsymbol{C}\hat{\beta} - r\right)'[\phi C(X'WX)^{-1}C']^{-1}(\boldsymbol{C}\hat{\beta} - r) \sim \chi_q^2$$

En pratique, nous pouvons tester un seul paramètre :

$$H_0 : \beta_j = r_j$$

On a alors, $C_i = 0, pour\ i \neq j\ et\ C_j = 1$ et $Var(\hat{\beta}_j) = \phi\psi_j$.

La statistique du test de Wald est dans ce cas égale à : $\frac{(\hat{\beta}_j - r_j)^2}{Var(Y_i)} \sim \chi_1^2$

- Test du Score :

Ce test est basé sur la dérivée de la fonction log-vraisemblance définie précédemment, cette dérivée est appelée le Score, on la défini en pratique dans ce cas par :

$$l'(\beta) = \phi^{-1} X' W G (y - \mu)$$

Avec G une matrice diagonale composée des éléments g'(μ_i) et W une matrice d'éléments $\left[\left(g'(\mu_i)\right)^2 V(\mu_i)\right]^{-1}$. De plus, on peut montrer que $\mathbb{E}[l'(\beta)] = 0$ et $Var\left(l'(\beta)\right) = \phi^{-1} X' W X$

Et enfin la statistique du test du score est donnée par :

$$(l'(\beta))^t [Var\left(l'(\beta)\right)]^t l'(\beta) \sim \chi_q^2$$

4. Etude des résidus

Pour évaluer l'erreur du modèle, nous pouvons calculer les résidus du modèle de différentes manières, les méthodes que l'on utilisera dans cette étude sont les principales utilisées pour calculer des résidus, les résidus de Pearson et les résidus de déviances définis tels que :

- Les résidus de Pearson :

Le résidu de Pearson est défini comme :

$$r_i{}^p = \frac{\sqrt{\omega_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

- Les résidus de Déviance :

On peut considérer que chaque observation y_i contribue à une quantité d_i à la déviance ($D = \sum_{i=1}^{n} d_i$), le résidu de déviance est défini comme :

$$r_i{}^D = \sqrt{d_i} \times signe(y_i - \mu_i)$$

La somme des carrés des résidus dans les deux cas est asymptotiquement une statistique du Khi-2 à n-p-1 degrés de liberté.

vii. Résultats

1. Présentation des résultats pour chaque modèle GLM

- **Significativité globale des modèles :**

Pour tester la significativité globale de nos modèles nous allons utiliser le critère de déviance, comme définit précédemment, ci-dessous un tableau présentant la déviance de chacun des modèles testés ainsi que la p-value issue du test du Khi-2 pour chaque modèle :

Loi	Déviance	p-value
Poisson	572 075,7	0
Log-Normal	50 391 578	0
Gamma	6 334,1	0
Inverse-Gaussienne	70,2	$1{,}36 \times 10^{-16}$

Nous observerons que la p-value de chaque modèle est inférieure à 0,05, ce qui nous permet de valider la significativité globale de chacun de nos modèles, cependant, il est important de garder à l'esprit que le fait d'obtenir une valeur de p exactement égale à 0 est très rare en pratique, en particulier pour les échantillons de taille finie. En effet, il est très probable que les tests statistiques rencontrent des limites numériques, qui peuvent conduire à des approximations numériques pour des valeurs très faibles de p.

Aux vues de la valeur de la déviance du modèle Poisson, nous décidons d'abandonner ce modèle pour le reste de l'étude.

- **Significativité individuelle des variables explicatives :**

Nous testerons la significativité individuelle des variables explicatives de chacun de nos modèles à l'aide de tests de Wald comme décrit précédemment.

- GLM Log-Normal :

Variable explicative	p-value
Intercept	0
SEXEM	$5{,}97 \times 10^{-16}$
AGE [31-38]	$8{,}65 \times 10^{-15}$
AGE [38-45]	$2{,}87 \times 10^{-40}$
AGE [45-52]	$6{,}33 \times 10^{-98}$
AGE [52-78]	$6{,}37 \times 10^{-149}$
CAT-PRO-Ensemble du personnel	$2{,}19 \times 10^{-17}$
CAT-PRO-Non-Cadre	$2{,}83 \times 10^{-174}$

- GLM Gamma :

Variable explicative	p-value
Intercept	0
SEXEM	$2{,}72 \times 10^{-15}$
AGE [31-38]	$2{,}15 \times 10^{-24}$
AGE [38-45]	$1{,}74 \times 10^{-56}$
AGE [45-52]	$4{,}63 \times 10^{-117}$
AGE [52-78]	$3{,}10 \times 10^{-155}$
CAT-PRO-Ensemble du personnel	$2{,}17 \times 10^{-07}$
CAT-PRO-Non-Cadre	$3{,}32 \times 10^{-124}$

- GLM Inverse-Gaussien :

Variable explicative	p-value
Intercept	0
SEXEM	$2{,}72 \times 10^{-15}$
AGE [31-38]	$2{,}15 \times 10^{-24}$
AGE [38-45]	$1{,}74 \times 10^{-56}$
AGE [45-52]	$4{,}63 \times 10^{-117}$
AGE [52-78]	$3{,}10 \times 10^{-155}$
CAT-PRO-Ensemble du personnel	$2{,}17 \times 10^{-07}$
CAT-PRO-Non-Cadre	$3{,}32 \times 10^{-124}$

La significativité individuelle des variables explicatives utilisées dans nos modèles est vérifiée à la vue des résultats des tests de Wald appliqués.

Nous allons à présent tenter d'interpréter les prédictions de ces modèles puis de juger de leur validité.

- **Résultats des modèles :**

Les modèles que nous avons calibrés nous donnent les prédictions suivantes :

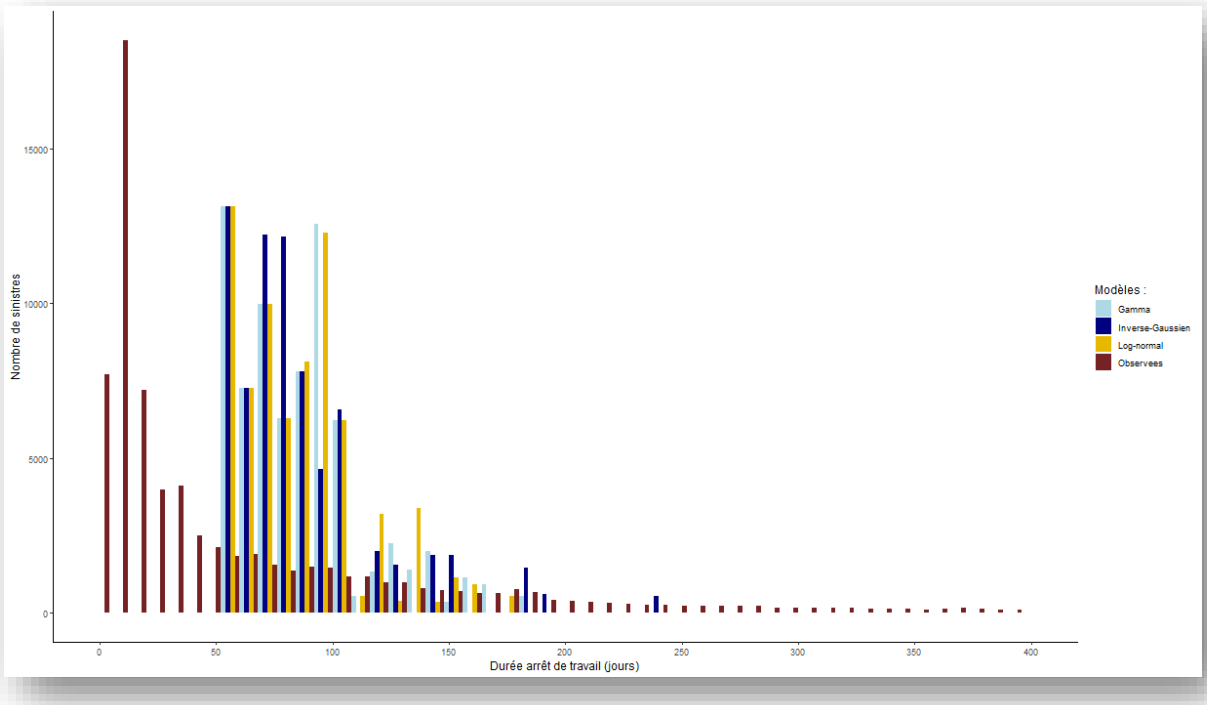


Figure 15 : Prédiction de la durée des arrêts de travail par nos GLM

Introduisons le **Standardized Mortality Ratio** adapté à l'étude des arrêts de travail qui est un indicateur de la qualité d'ajustement d'un modèle prédictif.

$$SMR_i = \frac{Nombre\ de\ jours\ d'arrêts\ observés\ par\ cohorte\ i}{Nombre\ de\ jours\ d'arrêts\ estimés\ par\ cohorte\ i}$$

Les prédictions des GLM semblent à première vue très mauvaises, cela est très certainement dû au nombre réduit de variables explicatives en notre possession rendant le modèle très robuste mais très peu efficace.

Si l'on calcule le SMR de nos modèles au global nous obtenons à priori de bons résultats :

GLM	SMR global
Gamma	99,8%
Inverse-Gaussien	98,9%
Log-Normal	99,8%

Nous estimons au global du portefeuille sinistré une durée qui correspond à l'observation cependant en réduisant l'estimation à la maille tête par tête, notre modélisation n'est pas du tout satisfaisante.

Essayons à présent de calibrer nos modèles en effectuant une séparation entre les arrêts de plus ou moins de six mois.

2. *Modélisations des arrêts de travail avec une distinction entre arrêt court et arrêt long*

Comme décrit précédemment, nous allons tenter de modéliser les arrêts court (moins de six mois) et les arrêts long (plus de six mois).

- **Résultats des GLM modélisant les arrêts court :**

La significativité globale des modèles et individuelles des variables explicatives a été vérifié de la même façon que précédemment, nous allons directement présenter les résultats, via l'histogramme ci-dessous :

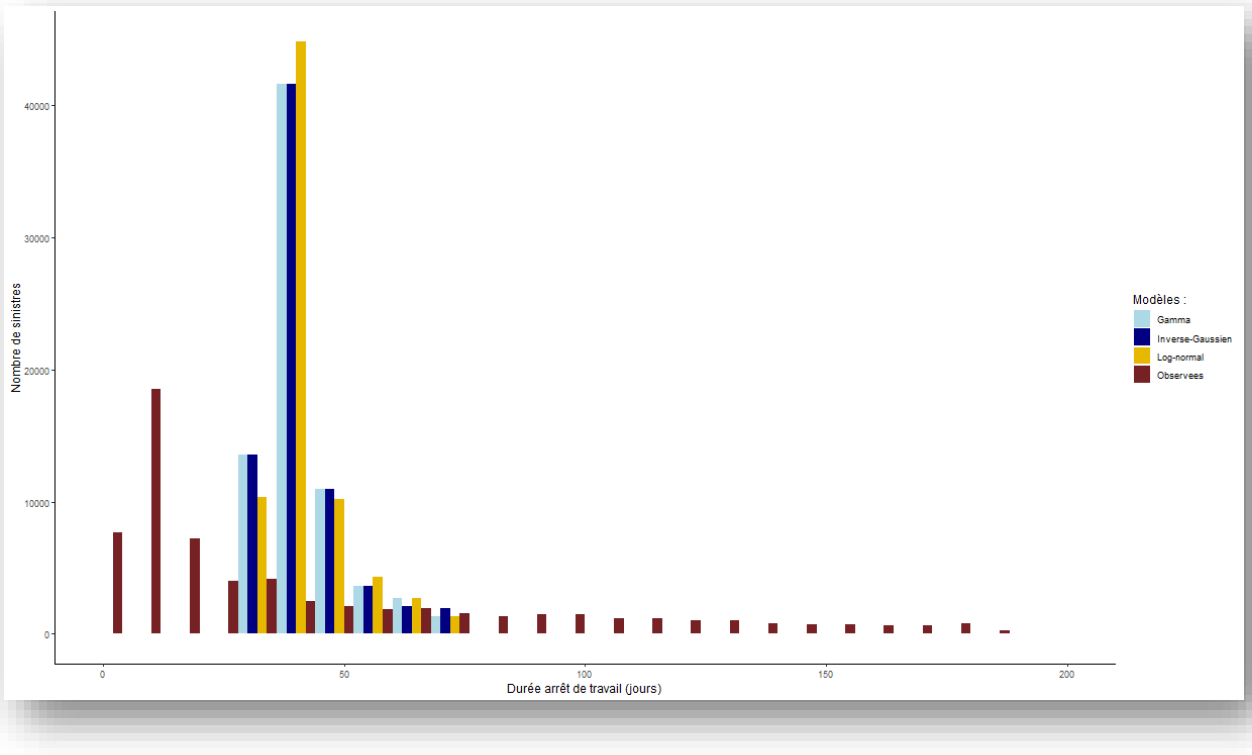


Figure 16:Prédiction de la durée des arrêts de travail de moins de six mois

Comme précédemment, peu importe le modèle utilisé nous ne parvenons pas à modéliser les arrêts très court (inférieur à 15 jours) qui correspondent à une partie importante des arrêts.

- **Résultats des GLM modélisant les arrêts longs :**

La significativité globale des modèles et individuelles des variables explicatives a été vérifié de la même façon que précédemment, nous allons directement présenter les résultats, via l'histogramme ci-dessous :

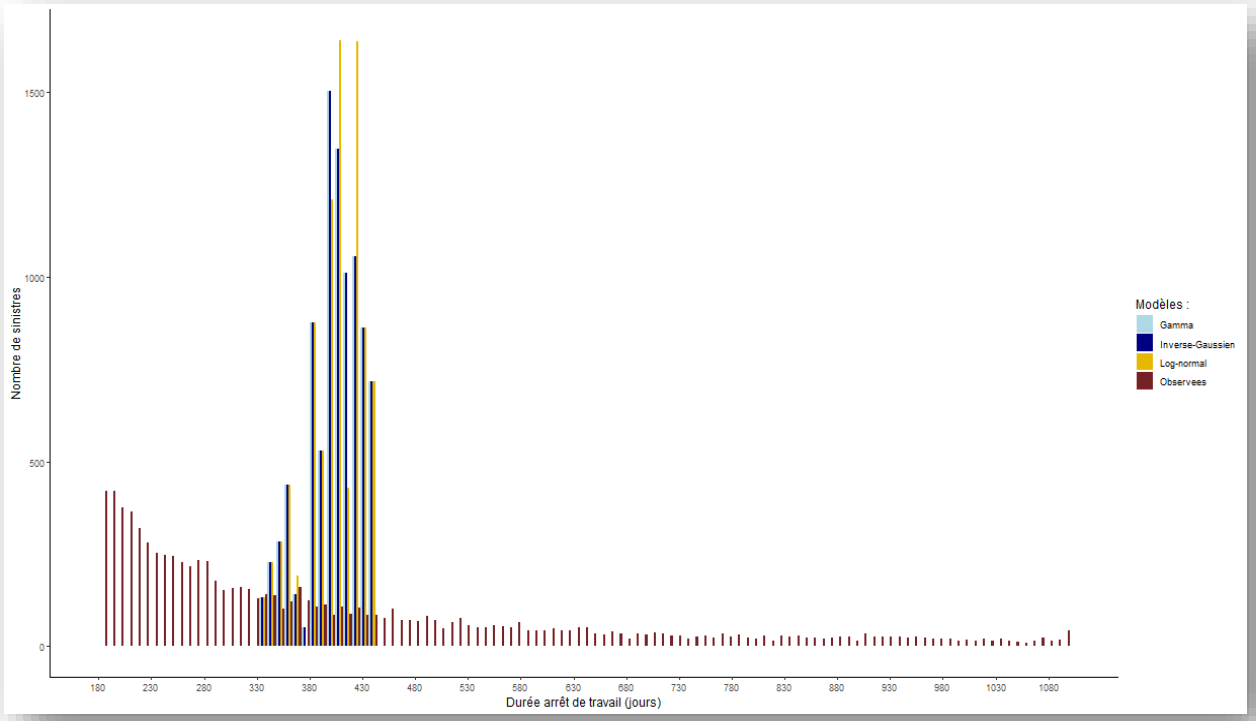


Figure 17: Prédiction de la durée des arrêts de travail de plus de six mois

On constate la même chose que précédemment malgré le split sur la durée des arrêts de travail, c'est-à-dire une prédiction pas du tout satisfaisante.

En conclusion de cette partie, un des intérêts des modèles linéaires généralisés appliqués aux sciences actuarielles est la possibilité d'effectuer une modélisation en tête par tête sur un portefeuille, or ce que l'on observe dans le contexte de la modélisation de la durée des arrêts de travail est l'inverse, c'est-à-dire que l'on a une modélisation au global qui est bonne mais à la maille tête par tête qui est très mauvaise. Nous concluons sur ce point que la modélisation effectuée dans le cadre de ce mémoire via des GLM n'est pas du tout efficace. Nous allons maintenant nous intéresser à une méthode de marché pour la modélisation de la durée des arrêts de travail qui est la construction d'une table de maintien en incapacité.

viii. Construction de loi de maintien en arrêt de travail

Cette sous-partie a pour objectif de modéliser la durée de maintien en arrêt de travail à l'aide de tables de maintien que l'on construira grâce aux données en notre possession.

ix. Rappels théoriques sur les modèles de durée

La durée de maintien en incapacité est la variable aléatoire T_x représente le temps passé en état d'incapacité d'un individu à partir de l'âge x. Cette durée est définie sur l'espace probabiliste $(\Omega, \mathcal{F}, \mathcal{P}) \to \mathbb{R}$.

La fonction de répartition F_x de la durée d'arrêt que nous avons utilisé précédemment, appelée fonction de répartition de T_x, c'est une mesure de probabilité définie sur la tribu borélienne $\mathcal{B}(\mathbb{R})$, elle donne la probabilité de sortir de l'état d'incapacité avant un certain moment t.

La fonction de répartition étant dérivable presque partout pour la mesure de Lebesgue, notons f_x sa dérivée, appelée densité telle que :

$$f_x(t) = \frac{dF_x}{dt} = \lim_{h \to 0} \frac{\mathbb{P}(t \leq T_x \leq t + h)}{h}$$

La fonction de survie de T_x est définie par :

$$S_x(t) = 1 - F_x(t) = \mathbb{P}(t < T_x)$$

Elle correspond à la probabilité que la durée de l'arrêt de travail d'un individu d'âge x soit supérieure à t.

La fonction de survie conditionnelle de T_x est définie par :

$$S_u(t) = \mathbb{P}(u + t < T_x | u < T_x) = \frac{\mathbb{P}(T_x > u + t)}{\mathbb{P}(T_x > u)} = \frac{S_x(u + t)}{S_x(u)}$$

Elle correspond à la probabilité que la durée de l'arrêt de travail actif depuis une durée u, d'un individu d'âge x, s'étale encore sur une période t.

La fonction de hasard est définie par :

$$h_x(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq T_x \leq t + h | t < T_x)}{h} = \frac{f_x(t)}{S_x(t)} = \frac{d(\ln(S_x(t)))}{dt}$$

Elle correspond à la probabilité de sortie instantanée de l'état d'incapacité. Cette fonction permet d'analyser si et quand un individu sort de l'état d'incapacité depuis le début de l'étude, et est donc essentielle pour l'analyse des durées de survie. En étudiant les variations de cette fonction, on peut identifier les moments où le risque de sortie de l'état d'incapacité est élevé pour l'individu, ainsi que l'évolution de ce risque au fil du temps. La fonction de hasard s'exprime formellement comme la limite d'un ratio, et est parfois appelée le taux de hasard instantané.

De plus,

$$S_x(t) = \mathrm{e}^{-\int_0^t h_x(s)\mathrm{d}s}$$

C'est grâce à cette propriété que nous allons pouvoir estimer la loi de T_x, en estimant la fonction de hasard de la loi.

Le taux de sortie est défini par :

$$ {}_uq_x(t) = \mathbb{P}(t \leq T_x \leq t+u) = \frac{S_x(t) - S_x(u+t)}{S_x(t)} = 1 - \frac{S_x(u+t)}{S_x(t)} $$

Ce rapport correspond à la probabilité qu'un individu en arrêt de travail à l'instant t sorte de l'état d'incapacité entre l'instant t et l'instant $t+u$. On en déduit la probabilité de maintien dans l'état tel que :

$$ {}_up_x(t) = 1 - {}_uq_x(t) = \frac{S_x(u+t)}{S_x(t)} $$

Ici nous ne considérons pas les causes de sorties de l'état d'incapacité qui sont le rétablissement, le passage en invalidité ou le décès car nous sommes dans une démarche de calcul des prestations liées au maintien en incapacité.

Nous allons à présent modéliser la loi via une estimation de la fonction de survie grâce à la méthode de Kaplan-Meier qui nous permettra d'obtenir les taux de sorties bruts que nous lisserons grâce à un lissage de Whittaker-Henderson que nous présenterons au préalable.

1. Estimation de la fonction de survie : estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur de la fonction de survie non paramétrique qui peut être utilisé en présence de censures et troncatures. L'estimateur de Kaplan-Meier est basé sur l'estimation des taux instantanés de survie, ce qui justifie son utilisation historique dans la tarification et le provisionnement liés à l'arrêt de travail.

Dans ce mémoire, nous présentons une version simplifiée de l'estimateur en raison de la présence de doublons dus aux données discrètes. Nous effectuerons nos calculs avec un pas quotidien, l'estimation se fait âge par âge de la façon suivante :

$$ \forall t \in [\![AncMin, \dots, AncMax]\!], \qquad \widehat{S_x}(t) = \prod_{T_i \leq t} [1 - \frac{d_x(T_i)}{n_x(T_i)}] $$

avec $n_x(t)$ l'échantillon sous risque à l'âge x juste avant le jour t et $d_x(t)$ le nombre de sorties d'incapacité pour l'âge x et l'instant T_i (par reprise d'activité, passage en invalidité, décès ou fin de garanties).

L'estimateur de Kaplan-Meier présente les propriétés suivantes :

- Unique estimateur cohérent de la fonction de survie,
- Estimateur du maximum de vraisemblance généralisée,
- C'est un estimateur robuste à condition que la survie et la censure n'aient aucune discontinuité commune,

- Il vérifie un théorème de normalité asymptotique à condition que la survie et la censure soient telles que leur fonction de répartition soient indépendantes et n'aient aucune discontinuité commune.

On obtient ainsi une estimation de la fonction de survie pour tous les âges :

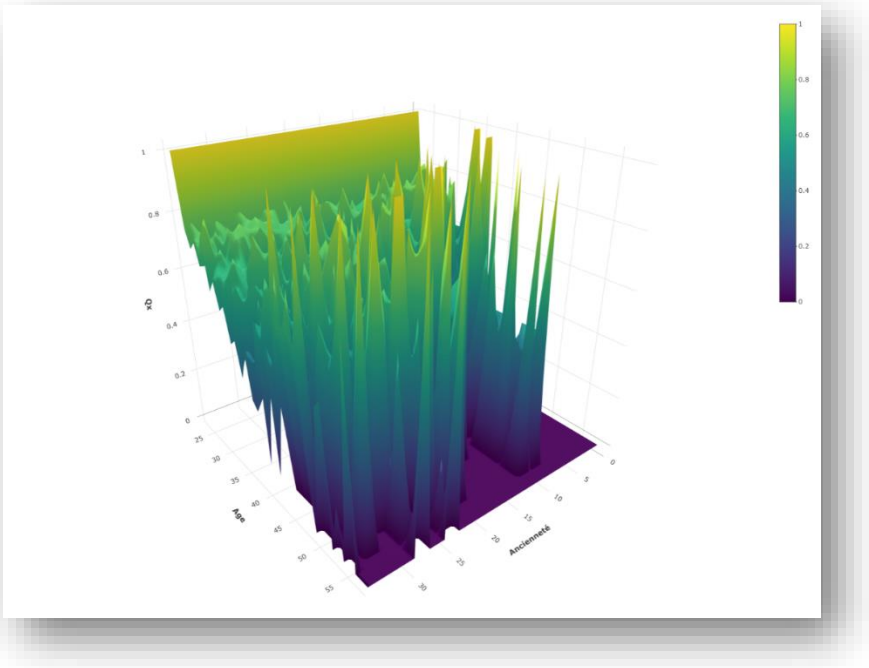


Figure 18 : Taux bruts Qxt pour le maintien en incapacité

x. Lissage des taux bruts – Méthode de Whittaker-Henderson

La méthode de Whittaker Henderson est une méthode de lissage non paramétrique qui permet de lisser des données en combinant un critère de fidélité et un critère de régularité. La fidélité fait référence à la capacité du modèle de s'ajuster aux données observées, tandis que la régularité vise à éviter un lissage excessif ou une sur-adaptation aux données.

Pour l'appliquer, on cherche à minimiser la somme de ces deux critères pour obtenir le lissage optimal des données. Cette méthode est similaire à un lissage bayésien, où les données sont considérées comme une réalisation d'une distribution de probabilité sous-jacente.

La méthode de Whittaker-Henderson est couramment utilisée pour le lissage de données unidimensionnelles telles que les tables de mortalité. Cependant, pour des données bidimensionnelles telles que la problématique du maintien en incapacité, l'application successive de deux séries de lissages unidimensionnels ne suffit pas car elle ne capture pas les dépendances entre les deux composantes des taux de sortie.

Ainsi, on cherche à appliquer un lissage à deux dimensions pour obtenir un lissage plus précis et plus adéquat pour les données bidimensionnelles. Cependant, l'adaptation de la méthode de Whittaker-Henderson dans ce cas pose des problèmes pratiques plus importants que théoriques, nécessitant des ajustements et des améliorations pour une utilisation efficace.

1. *Application au maintien en incapacité*

On pose :

- x : l'âge de l'assuré à la survenance
- $\widehat{q}_x(t)$: les taux bruts estimés selon Kaplan-Meier
- $q_x(t)$: les taux ajustés estimés avec le lissage de Whittaker-Henderson
- $w_x(t)$: les poids proportionnels aux effectifs sous risque

Notons $\Delta^n u(x)$ l'opérateur différence d'ordre n suivant :

$$\Delta^n u(x) = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j}\, u(x+j)$$

La méthode de Whittaker-Henderson utilise deux critères :

- **Le critère de fidélité** utilisé est un critère des moindres carrés ordinaires pondérés qui quantifie la qualité de l'ajustement, noté F et défini par :

$$F = \sum_{x=x_{min}}^{x_{max}} \sum_{t=t_{min}}^{t_{max}} w_x(t) * (q_x(t) - \widehat{q}_x(t))^2$$

- **Les critères de régularité** utilisés sont sous la forme de sommes d'opérateurs de différence, calculés pour chaque âge et chaque ancienneté, ils sont définis par :
  - Critère de régularité verticale :

$$S_v = \sum_{x=x_{min}}^{x_{max}} \sum_{t=t_{min}}^{t_{max}-z} (\Delta_v^z q_x(t))^2$$

  - Critère de régularité horizontale :

$$S_h = \sum_{x=t_{min}}^{t_{max}} \sum_{t=x_{min}}^{x_{max}-z} (\Delta_h^z q_x(t))^2$$

L'enjeu de cette optimisation est de minimiser la fonction $M = (1-\alpha-\beta)F + \alpha S_v + \beta S_h$, pour ce faire nous utiliserons la résolution matricielle suivante :

- Soit U un vecteur colonne de dimension $(t_{max} - t_{min} + 1) * (x_{max} - x_{min} + 1) = m = pq$, réorganisant les taux bruts $\widehat{q}_x(t)$ tels que :

$$u_{n(i-1)+j} = \widehat{q}_{1+x_{min}-1}\,(t_{min} + j - 1)$$

- Soit V le vecteur colonne des taux ajustés

- Soit W la matrice de poids de dimensions $m * m$
- Soit K_v^z la matrice définie par : $K_v^z . v = \begin{bmatrix} \Delta_v^z v_1 \\ \dots \\ \Delta_v^z v_{m-z} \end{bmatrix}$, K_v^z est la matrice à m-zq lignes et m colonnes contenant les coefficients binomiaux d'ordre z affectés de leur signe
- De même K_h^z la matrice définie par : $K_h^z . v = \begin{bmatrix} \Delta_h^z v_1 \\ \dots \\ \Delta_h^z v_{m-y} \end{bmatrix}$ de dimension $(p * (q-y), m)$

La fonction M peut alors s'écrire :

$$M = (1-\alpha-\beta)(V-U)'.W.(V-U) + \alpha V'.K_v^{z\prime}K_v^z.V + \beta V'.K_h^{z\prime}K_h^z.V$$

Soit X tel que :

$$\left(W + \alpha\, K_v^{z\prime}K_v^z + \beta\, K_h^{z\prime}K_h^z\right).X = W.U$$

Alors :

$$M = (V-X)'.\left(W + \alpha\, K_v^{z\prime}K_v^z + \beta\, K_h^{z\prime}K_h^z\right).(V-X) + E$$

Avec E un terme indépendant de X

La fonction M est alors minimale lorsque $(V-X) = 0 \Rightarrow V = X$

Les valeurs lissées via cette méthode s'obtiennent alors en posant :

$$\tilde{q} = V = \left(W + \alpha\, K_v^{z\prime}K_v^z + \beta\, K_h^{z\prime}K_h^z\right)^{-1}.W.U$$

Pour procéder au lissage des taux, nous devons fixer les paramètres du modèle de Whittaker-Henderson en entrée :

La matrice des taux bruts a été calculée grâce à la méthode de Kaplan-Meier, la matrice des poids est calculée proportionnellement aux effectifs en arrêt de travail, les indices de pondération pour les critères de fidélités α,β et les indices d'ordre de régularité verticale et horizontale z_v et z_h sont à déterminer.

La méthode de Whittaker-Henderson appliquée à nos données avec les paramètres $(\alpha, \beta, z_v, z_h) = (2,2,1000,1000)$ conduit à la représentation suivante des taux de sortie :

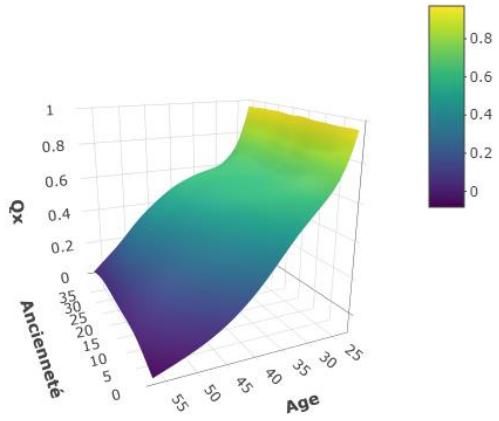


Figure 19 : Taux lissés avec les paramètres $(\alpha, \beta, z_v, z_h) = (2,2,1000,1000)$

Les taux semblent trop lissés, on observe trop peu d'irrégularité en fonction de l'âge et de l'ancienneté, nous allons tenter d'améliorer notre lissage en étudiant l'impact de chaque paramètre dans l'ajustement des taux.

Ci-dessous les lissages avec les paramètres $(\alpha, \beta, z_v, z_h) = (2,2,50,50)$ et $(\alpha, \beta, z_v, z_h) = (2,2,150,150)$:

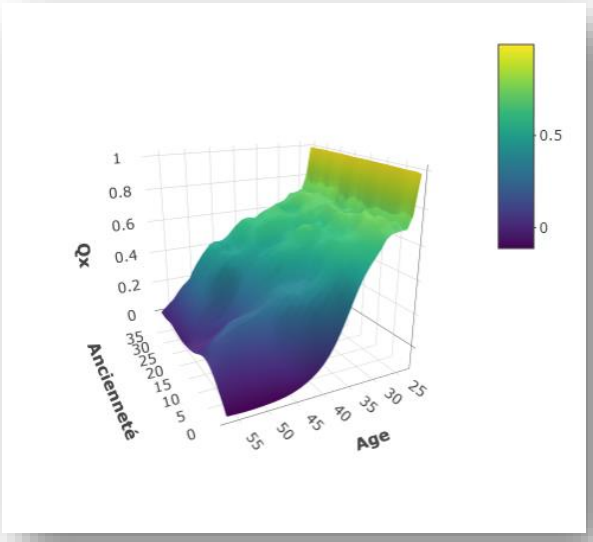


Figure 20: Taux lissés avec les paramètres $(\alpha, \beta, z_v, z_h) = (2,2,50,50)$

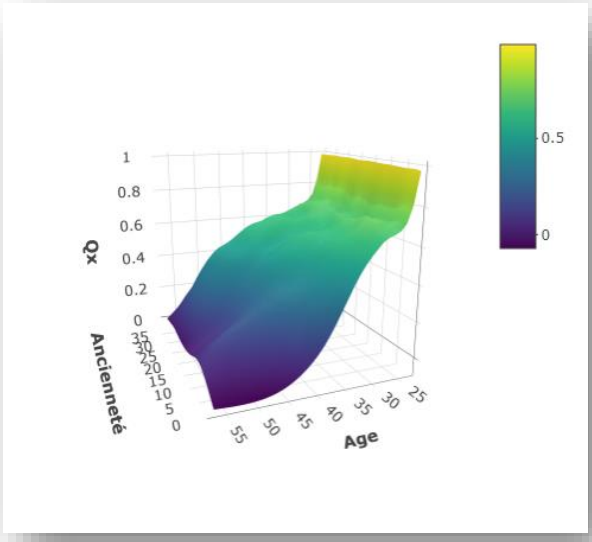


Figure 21: Taux lissés avec les paramètres $(\alpha, \beta, z_v, z_h) = (2,2,150,150)$

On constate qu'augmenter les paramètres d'ordres de régularité a pour impact de réduire les irrégularités sur les taux et de diminuer la valeur des taux bruts sur les anciennetés extrêmes, afin d'avoir une surface plus lissée nous allons augmenter ces paramètres et nous allons à présent observer l'impact des paramètres de fidélité.

Ci-dessous les lissages avec les paramètres $(\alpha, \beta, z_v, z_h) = (1,1,500,500)$, $(\alpha, \beta, z_v, z_h) = (2,2,500,500)$ et $(\alpha, \beta, z_v, z_h) = (3,3,500,500)$:

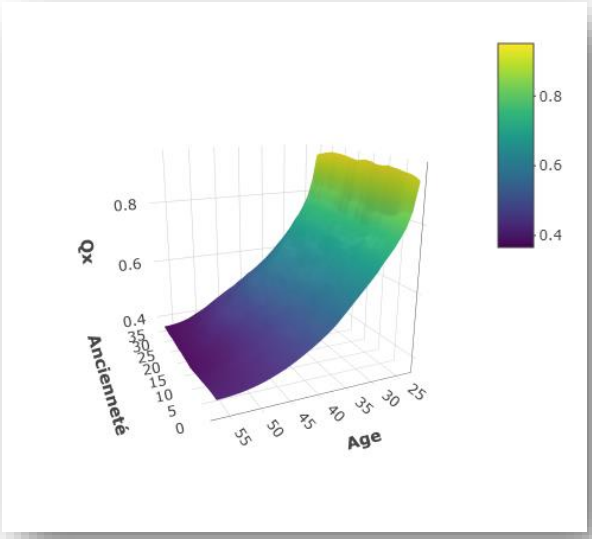


Figure 22: Taux lissés avec les paramètres $(\alpha, \beta, z_v, z_h) = (1,1,500,500)$

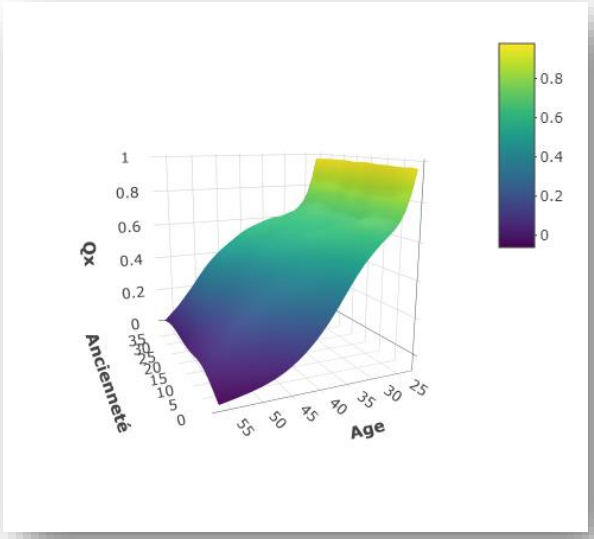


Figure 23 : Taux lissés avec les paramètres $(\alpha, \beta, z_v, z_h) = (2,2,500,500)$

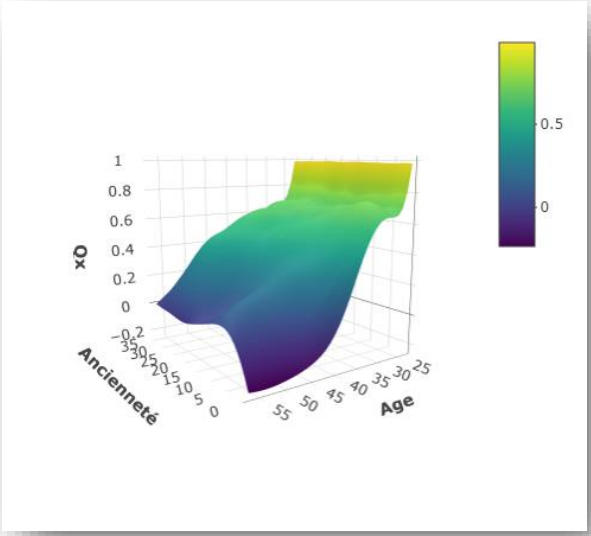


Figure 24 : Taux lissés avec les paramètres $(\alpha, \beta, z_v, z_h) = (3,3,500,500)$

On remarque que les taux obtenus sont plus lissés, il est aisément possible d'influencer significativement la qualité du lissage à l'aide de ces quatre paramètres.

A première vue, le lissage figure 22 n'est pas intéressant car trop lisse et pas du tout sensible à l'ancienneté, le lissage figure 23 semble bien plus intéressant car bien plus sensible à l'âge et l'ancienneté. Enfin, le lissage figure 24 semble être très satisfaisant car il semble mettre en

valeur des taux plus sensibles à l'âge et l'ancienneté tout en semblant être assez régulier. Nous allons vérifier ces remarques dans la suite grâce à des méthode de validation de notre lissage.

2. *Validation du lissage*

Test du Khi-2 :

Nous testons ici la validité de notre lissage avec une version ad hoc du test du Khi-2.

Nous effectuons pour cela des regroupements d'âges et d'anciennetés pour obtenir des classes bidimensionnelles. Le test est valable dans ce contexte car la somme de deux variables indépendantes avec une distribution du Khi-2 est une variable du Khi-2, de degré de liberté la somme des degrés de liberté.

Soit $D_{x,t}$ le nombre d'individus d'âge x sortis entre t et t+1 dans notre base de données, et $\widehat{D}_{x,t}$ le nombre d'individus d'âge x sortis entre t et t+1 prévu par notre modélisation.

La statistique du test est :

$$W_{k,l} = \sum_{x=1}^{l} \sum_{t=1}^{k} \frac{D_{x,t} - \widehat{D}_{x,t}}{\widehat{D}_{x,t}}$$

Pour mettre en place ce test, nous avons regroupé les âges en douze classes d'âge entre 23 et 62 ans et quinze classes d'ancienneté entre 0 et 1095 jours.

On peut montrer que si $L_{x,0} \to 0$, alors $W_{k,l}$ est asymptotiquement distribué comme une variable $\chi^2_{k+l-2} = \chi^2_{25}$. Pour comprendre le principe de ce test, schématisons l'hypothèse nulle :

Pour un test d'homogénéité du χ^2 où il s'agit de se demander si deux listes de nombres de même effectif total N peuvent dériver de la même loi de probabilité. L'hypothèse H_0 est la suivante : les deux échantillons proviennent de deux variables aléatoires de même loi.

Dans notre cas l'hypothèse H_0 sert à tester l'adéquation de notre échantillon de taux ajustés à notre échantillon de taux bruts. Cette hypothèse est rejetée pour un seuil à 95% dès lors que la statistique du test appartient à la région critique, c'est-à-dire $\{W_{k,l} > 14{,}61\}$.

La valeur de la statistique dans notre cas est de *9,42*, ce que nous conduit à accepter cet ajustement.

Taux SMR :

Nous avons calculé des SMR pour la modélisation des arrêts via des GLM, à présent recalculons ce ratio mais dans le contexte :

$$SMR = \frac{Nombre\ de\ sorties\ observées}{Nombre\ de\ sorties\ estimées} = \frac{64854{,}51}{64922{,}76} = 99{,}89\%$$

Le SMR de la modélisation étant proche de 100% et aux vues des résultats précédents, nous concluons que la modélisation des sorties via un table de maintien en incapacité est satisfaisante.

En conclusion, nous avons nettoyé les données et examiné la distribution de la durée des arrêts de travail, en comparant différentes lois. Les lois Gamma, Log-Normale et Inverse Gaussienne ont été retenues pour ajuster nos GLM, avec une exploration de la séparation des arrêts courts et longs. Les modèles GLM étaient significatifs, mais leur performance prédictive insatisfaisante. Nous avons donc étudié des tables de maintien en incapacité, utilisant la méthode de Kaplan-Meier pour estimer les taux de sortie bruts et les lisser avec Whittaker-Henderson. Le lissage est validé par un test du Khi-2 adapté et un SMR proche de 100%.

VI. Atténuation de l'absentéisme en entreprise – prévention du risque incapacité

a. Importance des stratégies de prévention

Les stratégies de prévention sont essentielles pour réduire les coûts liés aux arrêts de travail et améliorer la santé et le bien-être des travailleurs. Leur mise en place permet de :
Prévenir les risques professionnels et les maladies liées au travail, en limitant l'exposition des travailleurs à des facteurs de risque et en promouvant des pratiques de travail sûres et saines. Favoriser la détection précoce des problèmes de santé et la prise en charge adaptée des travailleurs concernés, en mettant en place des dispositifs de suivi médical et des programmes de sensibilisation et d'éducation. Faciliter la réinsertion professionnelle des travailleurs en arrêt de travail, en proposant des aménagements de poste, des formations adaptées et un accompagnement personnalisé. Contribuer à l'optimisation des coûts pour les entreprises et les assureurs, en réduisant la durée et la fréquence des arrêts de travail et en évitant les dépenses liées aux indemnités journalières, aux frais médicaux et aux pertes de productivité.

La prévention en entreprise est tout d'abord une obligation réglementaire, en effet, les articles L. 4121-1 à 4121-5 du Code du Travail obligent l'employeur à évaluer les risques pouvant porter atteinte à la santé mentale et physique de ses employés. L'article 4121-2 du Code du Travail propose 9 principes de prévention pouvant intervenir notamment dans la mise en place d'une stratégie :

- **Éviter les risques :** La priorité est donnée à l'élimination à la source des risques plutôt qu'à leur gestion a posteriori. Cela peut impliquer, par exemple, de choisir des machines moins dangereuses ou des produits chimiques moins toxiques.
- **Évaluer les risques qui ne peuvent pas être évités :** Si l'élimination du risque à la source n'est pas possible, l'employeur doit évaluer les risques restants et prendre les mesures appropriées pour les minimiser.
- **Combattre les risques à la source :** Il s'agit d'agir directement sur la source du risque plutôt que de mettre en place des mesures compensatoires ou protectrices pour les travailleurs exposés.
- **Adapter le travail à l'homme :** L'objectif est de prendre en compte les différences individuelles des travailleurs, notamment en matière d'ergonomie, pour réduire les effets du travail sur la santé.
- **Tenir compte de l'état d'évolution de la technique :** L'employeur doit prendre en compte les avancées technologiques et les meilleures pratiques disponibles pour minimiser les risques.
- **Remplacer ce qui est dangereux par ce qui est moins dangereux ou pas dangereux :** Si possible, des matériaux ou procédures moins dangereux doivent être utilisés à la place de ceux qui présentent un risque plus élevé.
- **Planifier la prévention :** Cela implique une approche globale intégrant dans un ensemble cohérent la technique, l'organisation du travail, les conditions de travail, les relations sociales et l'influence des facteurs ambiants.
- **Prendre des mesures de protection collective en priorité sur les mesures de protection individuelle :** Lorsque des mesures de protection sont nécessaires, l'accent doit être mis sur la protection de l'ensemble des travailleurs par des mesures collectives plutôt que de protéger chaque travailleur individuellement (par exemple, installer une ventilation générale plutôt que de fournir des masques respiratoires).

- **Donner les instructions appropriées aux travailleurs :** L'employeur doit fournir toutes les informations, instructions et formations nécessaires pour permettre aux travailleurs de prendre les mesures appropriées pour assurer leur propre sécurité et santé.

b. Analyse des différentes stratégies de prévention

i. Identification des stratégies de prévention possibles

Dans cette section, nous passerons en revue un éventail de stratégies de prévention potentielles pour réduire la durée des arrêts de travail, en 1948 l'Organisation Mondiale de la Santé a distingué trois niveaux de prévention :

- **Prévention primaire :** Il s'agit d'actions visant à prévenir l'apparition d'un arrêt de travail, en agissant sur les facteurs de risque en amont. Exemples : amélioration des conditions de travail, formation à la prévention des risques professionnels, programmes de santé au travail.

- **Prévention secondaire :** Ces stratégies visent à limiter la durée et la gravité des arrêts de travail existants, en facilitant la prise en charge médicale et en assurant un suivi adapté. Exemples : coordination des soins, télémédecine, accompagnement psychologique.

- **Prévention tertiaire :** Ces actions ont pour objectif de faciliter le retour au travail et d'éviter les rechutes. Exemples : aménagement du poste de travail, réadaptation professionnelle, soutien à la réinsertion.

L'Organisation Mondiale de la Santé distingue également :

- **La prévention individuelle :** Elle englobe toutes les actions entreprises pour inciter les individus à adopter des comportements favorables à leur santé. Ces actions peuvent être coercitives (comme les obligations vaccinales) ou basées sur des recommandations (telles que les campagnes d'information).
- **La prévention collective :** Cette approche vise à réduire les facteurs de risque liés à l'environnement et se situe généralement en amont du système de santé. Les mesures de sécurité sanitaire et d'hygiène publique, telles que le contrôle de la qualité de l'eau, la lutte contre la pollution et la sécurité alimentaire, relèvent de la prévention collective.

La prévention collective se divise généralement en actions comportementales (information, promotion, éducation à la santé) et actions environnementales, qui englobent entre autres prévention des risques psychosociaux.

ii. La meilleure stratégie à mettre en place : comparaison des différentes stratégies

Pour déterminer les approches les plus intéressantes, il faut ensuite comparer les résultats de l'évaluation des coûts et des bénéfices des différentes stratégies de prévention. Cette comparaison permettra d'identifier les approches les plus rentables et les plus efficaces pour réduire la durée des arrêts de travail et favoriser la réinsertion professionnelle des employés concernés.

Plusieurs facteurs permettent de déterminer les meilleures stratégies à mettre en œuvre, tels que :

- L'ampleur des bénéfices potentiels : Les stratégies qui génèrent des économies substantielles et une réduction significative des arrêts de travail seront privilégiées.
- La faisabilité et l'acceptabilité : Les stratégies les plus efficaces ne seront pas nécessairement les plus faciles à mettre en œuvre. Nous devrons tenir compte des contraintes pratiques, organisationnelles et financières pour chaque stratégie.
- Les effets sur la qualité de vie et la satisfaction des employés : Les stratégies de prévention doivent être bien acceptées par les employés et ne pas nuire à leur bien-être ou à leur motivation.

En combinant ces critères, il est possible d'établir des recommandations sur les stratégies de prévention à privilégier pour réduire les arrêts de travail et favoriser la réinsertion professionnelle.

c. Evaluation de l'impact d'un programme de prévention

La Haute Autorité de Santé (HAS) est un organisme public indépendant à vocation scientifique en France, ayant pour mission principale de contribuer à l'amélioration de la qualité du système de santé. Dans son guide intitulé "Choix méthodologiques pour l'évaluation économique à la HAS", cette institution fournit un cadre méthodologique détaillé pour évaluer l'efficacité économique des interventions en santé. Le document explore différentes approches pour mener des évaluations économiques, y compris l'analyse coût-efficacité et l'analyse coût-bénéfice. Ces méthodologies sont essentielles pour quantifier et comparer les bénéfices en termes de santé obtenus par rapport aux coûts engagés, fournissant ainsi des données précieuses pour les décideurs. Bien que centré sur le domaine clinique, ce guide peut être adapté pour évaluer une variété de mesures de prévention en entreprise, notamment les programmes de santé et de sécurité au travail.

i. Approche coût-efficacité (ACE)

L'approche coût-efficacité est une méthode d'évaluation économique qui compare les coûts et les effets des différentes stratégies de prévention, afin de déterminer laquelle offre le meilleur rapport entre les coûts engagés et les résultats obtenus. Cette approche se concentre sur l'efficacité des stratégies en termes de réduction de la durée et de la fréquence des arrêts de travail, sans prendre en compte les aspects monétaires des bénéfices obtenus. Les étapes clés de cette approche sont les suivantes :

- Identification et description des stratégies de prévention.
- Estimation des coûts associés à chaque stratégie, incluant les coûts de mise en place, de fonctionnement et de maintenance.
- Mesure des effets des stratégies sur la durée des arrêts de travail et la réinsertion professionnelle, en utilisant des indicateurs tels que le nombre de jours d'arrêt évités ou le taux de réinsertion.
- Calcul du rapport coût-efficacité pour chaque stratégie, en divisant les coûts par les effets obtenus.
- Analyse comparative et discussion des résultats, en identifiant les stratégies offrant le meilleur rapport coût-efficacité.

ii. Approche coût-bénéfice (ACB)

L'approche coût-bénéfice est une méthode d'évaluation économique qui va plus loin que l'approche coût-efficacité, en quantifiant également les bénéfices obtenus en termes monétaires. Cette approche permet d'évaluer les avantages économiques des stratégies de prévention, en comparant les coûts engagés aux bénéfices réalisés, tels que les économies sur les indemnités journalières, les frais médicaux et les pertes de productivité. Les étapes clés de cette approche sont les suivantes :

- Identification et description des stratégies de prévention.
- Estimation des coûts associés à chaque stratégie, incluant les coûts de mise en place, de fonctionnement et de maintenance.
- Quantification des bénéfices associés à chaque stratégie, en monétisant les effets sur la durée des arrêts de travail, la réinsertion professionnelle et les coûts évités.
- Calcul du retour sur investissement (ROI) pour chaque stratégie, en divisant les bénéfices nets (bénéfices moins coûts) par les coûts.
- Analyse comparative et discussion des résultats, en identifiant les stratégies offrant le meilleur retour sur investissement.

iii. Indicateurs clés

Afin de mener une analyse économique rigoureuse, il est essentiel de disposer d'indicateurs clés permettant de mesurer les coûts et les bénéfices associés à chaque stratégie de prévention. Voici quelques-uns des principaux indicateurs que nous utiliserons dans notre analyse :

- Coûts directs : Il s'agit des coûts liés à la mise en place, au fonctionnement et à la maintenance des stratégies de prévention. Ils incluent les dépenses en matériel, en personnel, en formation et en communication.
- Coûts indirects : Ces coûts représentent les conséquences économiques des arrêts de travail, comme les pertes de productivité, les frais de remplacement du personnel absent et les coûts administratifs.
- Économies réalisées : Il s'agit des économies engendrées par la réduction des coûts directs et indirects grâce à la mise en œuvre des stratégies de prévention. Elles incluent les économies sur les indemnités journalières, les frais médicaux et les pertes de productivité.

En utilisant ces indicateurs clés, il est possible d'évaluer et de comparer les différentes stratégies de prévention des arrêts de travail en termes de coûts et d'efficacité. Cela peut permettre d'identifier les approches les plus rentables et les plus efficaces pour réduire la durée des arrêts de travail et favoriser la réinsertion professionnelle des employés concernés.

d. Gains pour les entreprises et les assureurs

Réduction des coûts

L'identification et la mise en œuvre des stratégies de prévention les plus rentables et efficaces auront un impact significatif sur la réduction des coûts liés aux arrêts de travail pour les entreprises et les assureurs. Ces économies pourront se traduire par des gains en termes de coûts directs (indemnités journalières, frais médicaux) et indirects (perte de productivité, remplacement temporaire, coûts de réinsertion). En optimisant les investissements dans la

prévention, les entreprises et les assureurs pourront réaliser des économies importantes et améliorer leur rentabilité.

Amélioration de la gestion des risques

L'analyse des stratégies de prévention et de leurs impacts économiques permettra également aux entreprises et aux assureurs d'améliorer leur gestion des risques liés aux arrêts de travail. En identifiant les facteurs de risque et les leviers d'action les plus pertinents, ils pourront mettre en place des mesures ciblées et adaptées à leur contexte spécifique. Cette approche personnalisée de la prévention contribuera à réduire les risques et à minimiser les conséquences négatives des arrêts de travail pour les entreprises, les assureurs et les employés concernés.

Meilleure allocation des ressources

Enfin, l'évaluation des différentes stratégies de prévention et de leurs coûts-bénéfices permettra aux entreprises et aux assureurs de mieux allouer leurs ressources. Plutôt que de consacrer des moyens financiers et humains à des actions peu efficaces ou coûteuses, ils pourront investir dans des stratégies ayant un réel impact sur la réduction des arrêts de travail et la réinsertion professionnelle. Cette meilleure allocation des ressources contribuera non seulement à optimiser les dépenses en matière de prévention, mais aussi à renforcer l'efficacité globale des dispositifs mis en place pour gérer les arrêts de travail et favoriser le retour à l'emploi des employés concernés.

Résultats de l'enquête Rehalto sur l'absentéisme « Comprendre pour agir » :

Différents acteurs peuvent jouer un rôle dans la mise en place d'actions de prévention en entreprise pour lutter contre l'absentéisme. À cet égard, Rehalto a également interrogé les DRH pour identifier les différents interlocuteurs avec qui ils abordent la problématique de la prévention contre l'absentéisme au sein de leur entreprise.
D'une part, les résultats de l'enquête révèlent que plus de la moitié des DRH discutent de ces problématiques avec la médecine du travail, la direction générale de leur entreprise et/ou les membres du Comité d'Entreprise, du Comité d'Hygiène, de Sécurité et des Conditions de Travail (CHSCT) ou les Délégués du Personnel.
D'autre part, l'enquête dévoile qu'il existe des entreprises qui n'abordent pas ces problématiques.
Enfin, les résultats soulignent que les DRH évoquent très peu le sujet avec leur assureur (seulement 4%).
Pourtant, la prévention fait partie des responsabilités sociétales de l'assureur : en plus d'accompagner au mieux ses assurés pour atténuer les conséquences en cas de sinistre, l'assureur doit, en amont, les sensibiliser aux risques auxquels ils peuvent être exposés. De plus, les mutations sociétales placent la culture du risque au cœur du métier de l'assureur. En d'autres termes, il devient indispensable pour un assureur de maîtriser les risques et leurs possibles évolutions pour diminuer leur probabilité d'occurrence. De fait, si le coût de la prévention n'est pas supporté par les assureurs mais par les entreprises, il reste important pour eux de participer et d'accompagner les entreprises, par le biais de services supplémentaires, étant à l'origine de plus en plus de nouvelles offres d'assurance.
Par ailleurs, comme mentionné précédemment, les arrêts de travail engendrent un coût considérable pour les entreprises. De plus, la réduction de la probabilité d'exposition au risque entraînera également des conséquences sur les résultats des contrats portés par les assureurs et donc sur leur rentabilité face à une concurrence de plus en plus agressive. En d'autres termes, un potentiel rapprochement entre les assureurs et les entreprises dans un objectif d'instauration

de prévention des risques pour atténuer l'absentéisme permettrait la mise en place d'un modèle gagnant-gagnant.

VII. Conclusion

Enseignements

Ce mémoire d'actuaire a exploré la modélisation du risque incapacité et l'absentéisme dans le contexte de l'assurance collective en France. Il a réussi à fournir une analyse approfondie des différents aspects de ces risques et à mettre en avant des modèles statistiques permettant d'étudier la durée des arrêts de travail ainsi que leurs implications pour les entreprises et les assureurs. Notre travail a contribué à une meilleure compréhension du risque incapacité et de l'absentéisme pour le portefeuille collectif de Generali Vie, fournissant ainsi des informations et des recommandations à l'entreprise. Il a également proposé des voies pour proposer des stratégies de prévention et de gestion des risques liés à l'absentéisme en assurance collective.

Limites

Néanmoins, il est crucial de reconnaître les limites de cette étude. L'une des principales contraintes réside dans les données limitées en notre possession. Les données utilisées pour construire la base de données d'étude proviennent de différentes sources et peuvent présenter des incohérences ou des lacunes, ce qui affecte la précision et la généralisabilité de nos résultats. De plus, certaines informations pertinentes, telles que les facteurs de risque individuels ou organisationnels, pourraient ne pas être disponibles ou être difficilement accessibles en raison de problèmes de confidentialité ou de protection des données.

Pour pallier ces limites, de futures recherches pourraient se concentrer sur l'obtention de données plus complètes et représentatives. Il serait également bénéfique d'examiner les variations régionales ou sectorielles du risque incapacité et de l'absentéisme, ainsi que l'intégration de données longitudinales pour évaluer l'évolution des risques et l'impact des politiques au fil du temps.

En résumé, il reste du travail à faire pour combler les lacunes et améliorer la qualité des données, afin de fournir des informations encore plus précises et fiables pour la prise de décision.

VIII. Bibliographie

[1] AXA, «"Quels sont les risques de demain?",» 2019. [En ligne]. Available: https://www.axa.com/fr/magazine/rapport-risques-futurs-2019#:~:text=Selon%20le%20rapport%2C%20les%20principaux,et%20%C3%A0%20l'instabilit%C3%A9%20g%C3%A9opolitique..

[2] F. Planchet, «Ressources actuarielles,» 2020. [En ligne]. Available: http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance7.pdf?OpenElement.

[3] Rehalto, «Baromètre sur les arrêts de travail « Comprendre pour agir » 5ème édition,» 2019. [En ligne]. Available: https://entreprise-rh.com/wp-content/uploads/2019/10/R%C3%A9halto-BVA-Barom%C3%A8tre-arr%C3%AAts-de-travail-2019-4.pdf.

[4] N. Fouquet, L. Chérié-Challine, É. Rubion, A. Descatha et Y. Roquelaure, «TROUBLES MUSCULO-SQUELETTIQUES LIÉS AU TRAVAIL : NOMBRE DE CAS ÉVITABLES,» 2020.

[5] M. Humanis, «Rapport annuel sur l'absentéisme,» 2020.

[6] L. Brami, S. Damart et F. Kletz, «SANTÉ AU TRAVAIL ET TRAVAIL EN SANTÉ. LA PERFORMANCE DES ÉTABLISSEMENTS DE SANTÉ FACE À L'ABSENTÉISME ET AU BIEN-ÊTRE DES PERSONNELS SOIGNANTS,» 2013.

[7] S. Chaupain-Guillot et O. Guillot, «Les absences au travail : une analyse à partir des données françaises du Panel européen des ménages,» *INSEE*, 2007.

[8] A. GAUMET, «Construction de tables d'expérience pour l'entrée et le maintien en incapacité,» 2001.

[9] L. Giesecke, «USE OF THE CHI-SQUARE STATISTIC TO SET WHITTAKER-HENDERSON SMOOTHING COEFFICIENTS,» *Defense Manpower Data Center*, 1981.

[10] F. PLANCHET et P. WINTER, «L'utilisation des splines bidimensionnels pour l'estimation de lois de maintien en arrêt de travail,» 2007.

[11] F. PLANCHET et J. WINTER, «Les provisions techniques des contrats de prévoyance collective,» *Economica*, 2006.

[12] Ayming, «14ème Baromètre de l'Absentéisme® et de l'Engagement - édition 2022,» 2022. [En ligne]. Available: https://www.ayming.fr/insights/barometres-livres-blancs/barometre-de-labsenteisme-et-de-lengagement/#download.

[13] F. PLANCHET, «Méthodes de lissage et d'ajustement,» 2020.

[14] FÉDÉRATION FRANÇAISE DES ASSURANCES, «Coronavirus COVID-19 et assurance,» 2020. [En ligne]. Available: https://www.ffa-assurance.fr/infos-assures/coronavirus-covid-19-et-assurance#Sant%C3%A9%20et%20Pr%C3%A9voyance.

[15] Haute Autorité de Santé (HAS), «Choix méthodologiques pour l'évaluation économique à la HAS» 2011.

IX. Table des figures

Figure 1: Intervention de la Sécurité Sociale sur les frais de santé 16
Figure 2 : Proportion des types d'arrêts connu par des salariés en entreprise 19
Figure 3 : Evolution du nombre d'arrêts par année de survenance 33
Figure 4: Evolution du taux d'arrêt de travail par assuré 34
Figure 5 : Saisonnalité des arrêts de travail entre 2018 et 2021 35
Figure 6 : Nombre d'arrêts de travail par saison 36
Figure 7 : Mise en valeur d'un « effet COVID » en comparaison avec la courbe d'évolution des hospitalisations sur 2020 37
Figure 8 : Evolution du nombre d'arrêt par type 38
Figure 9: Evolution du nombre d'arrêt par type chez les jeunes 39
Figure 10 : Schéma représentant les phénomènes de censures et de troncatures 41
Figure 11 : Nombre de sinistres par âge à la survenance 49
Figure 12: Nombre de sinistres par classe d'âge 50
Figure 13: Nombre de sinistres par sexe 50
Figure 14: Nombre de sinistres par catégorie socio-professionnelle 51
Figure 15 : Prédiction de la durée des arrêts de travail par nos GLM 64
Figure 16:Prédiction de la durée des arrêts de travail de moins de six mois 65
Figure 17: Prédiction de la durée des arrêts de travail de plus de six mois 66
Figure 18 : Taux bruts Qxt pour le maintien en incapacité 69
Figure 19 : Taux lissés avec les paramètres $\alpha, \beta, zv, zh = (2,2,1000,1000)$ 72
Figure 20: Taux lissés avec les paramètres $\alpha, \beta, zv, zh = (2,2,50,50)$ 72
Figure 21: Taux lissés avec les paramètres $\alpha, \beta, zv, zh = (2,2,150,150)$ 72
Figure 22: Taux lissés avec les paramètres $\alpha, \beta, zv, zh = (1,1,500,500)$ 73
Figure 23 : Taux lissés avec les paramètres $\alpha, \beta, zv, zh = (2,2,500,500)$ 73
Figure 24 : Taux lissés avec les paramètres $\alpha, \beta, zv, zh = (3,3,500,500)$ 73
Figure 25: Sortie du GLM Poisson 86
Figure 26: Sortie du GLM Gamma 86
Figure 27: Sortie du GLM Inverse-Gaussien 87
Figure 28: Sortie du GLM Log-Normal 87
Figure 29: Exposition au risque 88
Figure 30: Comparaison modèle / observation par âge à la survenance 89

X. Annexes

A- Sorties des modèles linéaires généralisés :

```
Call:
glm(formula = Nb_jour_couvert ~ SEXE + AGE + CAT_PRO, family = poisson(),
    data = baseage)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.765  -10.060   -6.918    1.359   67.415

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  4.5216279  0.0017024  2656.1   <2e-16 ***
SEXEM                        0.1123143  0.0008132   138.1   <2e-16 ***
AGE(31,38]                   0.1928834  0.0013692   140.9   <2e-16 ***
AGE(38,45]                   0.3118002  0.0013482   231.3   <2e-16 ***
AGE(45,52]                   0.4670952  0.0012783   365.4   <2e-16 ***
AGE(52,78]                   0.5714885  0.0013030   438.6   <2e-16 ***
CAT_PROEnsemble du personnel -0.1755666  0.0015500  -113.3   <2e-16 ***
CAT_PRONon Cadre             -0.5498886  0.0014286  -384.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11466880  on 73770  degrees of freedom
Residual deviance: 10894804  on 73763  degrees of freedom
AIC: 11278939

Number of Fisher Scoring iterations: 6
```

Figure 25: Sortie du GLM Poisson

```
Call:
glm(formula = Nb_jour_couvert ~ SEXE + AGE + CAT_PRO, family = Gamma(link = "log"),
    data = baseage)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9119  -1.6649  -0.9376   0.1473   5.8163

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   4.53542    0.03102 146.222  < 2e-16 ***
SEXEM                         0.11988    0.01333   8.993  < 2e-16 ***
AGE(31,38]                    0.19842    0.01961  10.120  < 2e-16 ***
AGE(38,45]                    0.31526    0.01995  15.801  < 2e-16 ***
AGE(45,52]                    0.47773    0.01944  24.578  < 2e-16 ***
AGE(52,78]                    0.57883    0.02053  28.194  < 2e-16 ***
CAT_PROEnsemble du personnel -0.18304    0.03098  -5.908 3.49e-09 ***
CAT_PRONon Cadre             -0.57940    0.02848 -20.344  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.014802)

    Null deviance: 162689  on 73770  degrees of freedom
Residual deviance: 155980  on 73763  degrees of freedom
AIC: 783099

Number of Fisher Scoring iterations: 7
```

Figure 26: Sortie du GLM Gamma

```
Call:
glm(formula = Nb_jour_couvert ~ SEXE + AGE + CAT_PRO, family = inverse.gaussian(link = "inverse"),
    data = baseage)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.99582  -0.32259  -0.12590   0.01568   0.57548

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.0126553  0.0003341  37.883  < 2e-16 ***
SEXEM                       -0.0013343  0.0001611  -8.282  < 2e-16 ***
AGE(31,38]                  -0.0030088  0.0002733 -11.008  < 2e-16 ***
AGE(38,45]                  -0.0044663  0.0002690 -16.602  < 2e-16 ***
AGE(45,52]                  -0.0062021  0.0002546 -24.364  < 2e-16 ***
AGE(52,78]                  -0.0071364  0.0002593 -27.526  < 2e-16 ***
CAT_PROEnsemble du personnel  0.0013011  0.0002977   4.370 1.25e-05 ***
CAT_PRONon Cadre              0.0056490  0.0002756  20.494  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for inverse.gaussian family taken to be 0.03949169)

    Null deviance: 6242.9  on 73770  degrees of freedom
Residual deviance: 6165.7  on 73763  degrees of freedom
AIC: 768359

Number of Fisher Scoring iterations: 2
```

Figure 27: Sortie du GLM Inverse-Gaussien

```
Call:
glm(formula = Nb_jour_couvert ~ SEXE + AGE + CAT_PRO, family = gaussian(link = "log"),
    data = baseage)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-176.52   -71.38   -48.23    12.62  1040.77

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.51839    0.02540 177.917  < 2e-16 ***
SEXEM                          0.09940    0.01229   8.090 6.06e-16 ***
AGE(31,38]                     0.18635    0.02402   7.758 8.76e-15 ***
AGE(38,45]                     0.30656    0.02308  13.284  < 2e-16 ***
AGE(45,52]                     0.45468    0.02165  21.002  < 2e-16 ***
AGE(52,78]                     0.56129    0.02160  25.990  < 2e-16 ***
CAT_PROEnsemble du personnel -0.16913    0.01994  -8.483  < 2e-16 ***
CAT_PRONon Cadre             -0.52518    0.01866 -28.144  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 21802.93)

    Null deviance: 1658578570  on 73770  degrees of freedom
Residual deviance: 1608186992  on 73763  degrees of freedom
AIC: 946318

Number of Fisher Scoring iterations: 7
```

Figure 28: Sortie du GLM Log-Normal

B- Compléments sur le lissage de Whittaker-Henderson

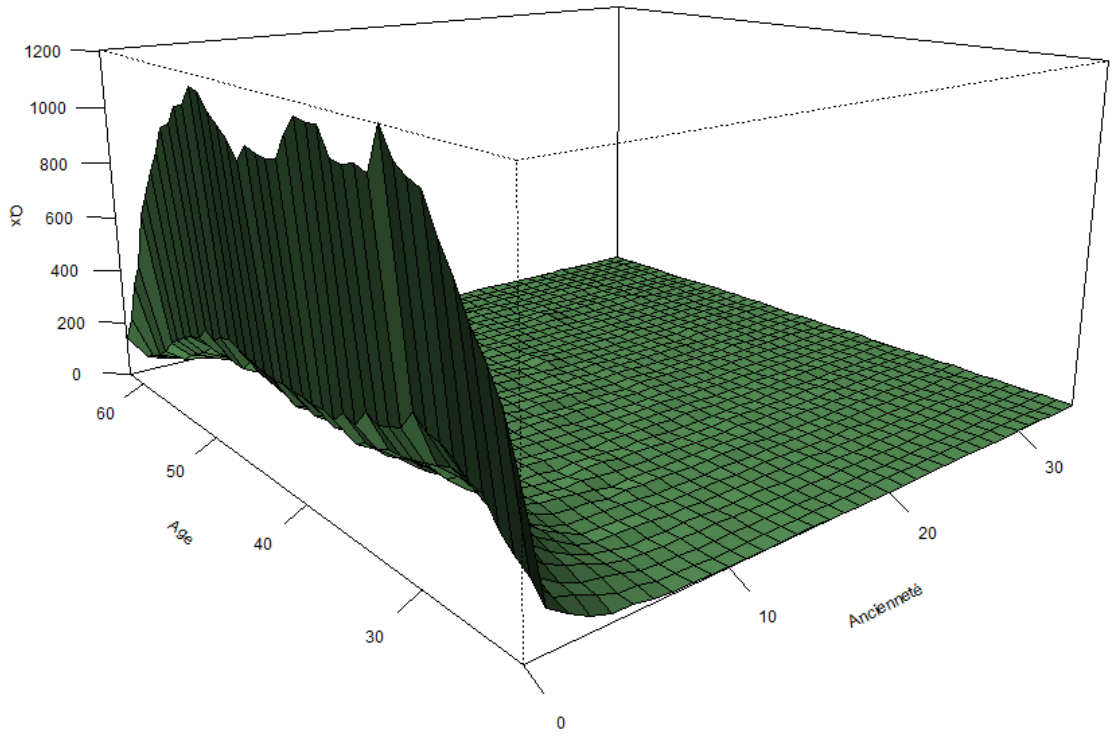


Figure 29: Exposition au risque

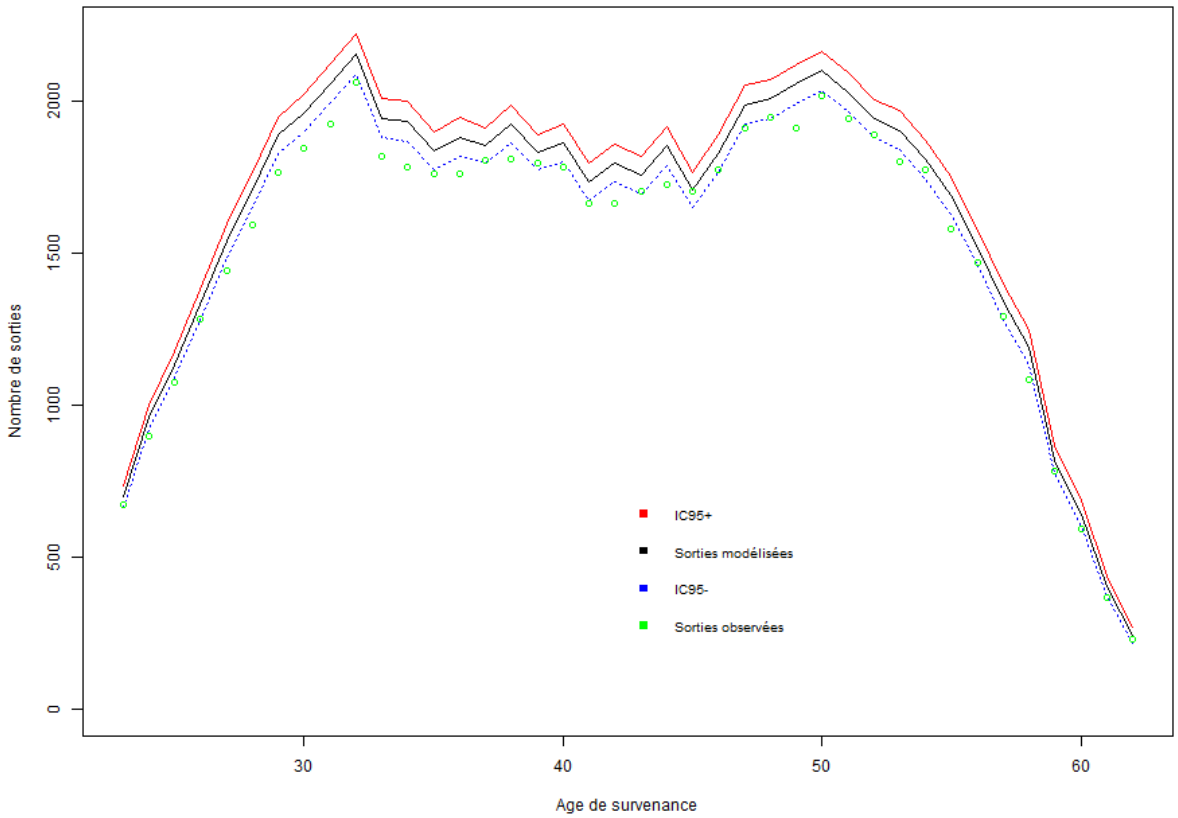


Figure 30: Comparaison modèle / observation par âge à la survenance