

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaire
le 14/03/2022

Par : **Albane GERONDEAU**

Titre : **Mortality level estimation for annuities and
underwritten annuities products combining
traditional and Machine Learning technics**

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

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

Membres présents du jury de la filière

Nom : Caroline HILLAIRET

Membres présents du jury de l'Institut
des Actuaire

BASTIEN Jonathan (visé)

MILOE Yann (visé)

Entreprise : **SCOR**

Signature :

Directeur du mémoire en entreprise :

Nom : Razvan IONESCU


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

Secrétariat:

Bibliothèque:



Signature du candidat



Abstract

Segmentation of the risk has always been of important topic for insurance companies to prevent themselves from adverse selection. Models used to estimate the risk are becoming more and more complex and exploit more and more variables.

Traditional methods such as Generalized Linear Model are not adapted to capture interaction effects within numerous input variables, and therefore uncertainty can remain in the efficiency of the fitted models.

We have seen in the past years an increasing use of Machine Learning models in a large variety of fields. Temptation to use those complex models in the insurance industry is no exception. However, insurance companies are faced with regulation constraints, requiring them to be in the capacity to explain their predictions. Therefore, the use of Machine Learning models are often limited.

This study shows a way to use Machine Learning models to supervise the calibration and segmentation performance of traditional Generalized Linear Model models, through the example of mortality level risk estimation using health status variables for enhanced annuity products.

Our study is based on UK health data originating from THIN database. We show that, when taking advantage of knowledge from medical experts as well as insights provided by sensibility analysis of a Machine Learning model, it is possible to have close predictive performances between a Generalized Linear Model and a Machine Learning model.

Key Words: Survival analysis, Generalized Linear Model, Machine Learning, Segmentation, Mortality Level, Enhanced Annuities

Résumé

La segmentation du risque a toujours été un sujet important pour les compagnies d'assurance afin de se prémunir contre l'antisélection. Les modèles utilisés pour estimer le risque deviennent ainsi de plus en plus complexes et contiennent de plus en plus de variables.

Dans le cas d'un nombre élevé de variables explicatives, les méthodes traditionnelles telles que le modèle linéaire généralisé (GLM) ne sont pas adaptées à la prise en compte d'effets d'interactions complexes, et donc l'incertitude peut subsister sur l'efficacité de ces modèles.

Nous avons constaté au cours des dernières années une utilisation croissante des modèles de Machine Learning dans une grande variété de domaines. La tentation d'utiliser ces modèles puissants dans le secteur de l'assurance ne fait pas exception. Cependant, les compagnies d'assurance sont confrontées à des contraintes réglementaires, les obligeant à être en mesure d'expliquer leurs tarifs. Par conséquent, l'utilisation des modèles d'apprentissage automatique est souvent limitée.

Notre étude montre un moyen d'utiliser des modèles de Machine Learning pour superviser les performances de calibrage et de segmentation des modèles GLM traditionnels, à travers l'exemple de l'estimation du risque de niveau de mortalité par de nombreuses variables relatives à l'état de santé pour des produits de rente sur risque aggravé.

Notre étude est basée sur des données de santé britanniques provenant de la base de données THIN. Nous montrons que, en tirant parti des connaissances d'experts médicaux ainsi que des informations fournies par l'analyse de sensibilité d'un modèle d'apprentissage automatique, il est possible d'avoir des performances prédictives proches entre un modèle linéaire généralisé et un modèle d'apprentissage automatique.

Mots-clés: Analyse de survie, Machine Learning, Segmentation, Niveau de mortalité, Rentes aggravées

Executive summary

Context

For insurance companies, longevity risk is the risk that life expectancies and survival rates exceed their predictions, resulting in greater-than-anticipated claims to pay. This longevity risk is the main risk in life annuities products.

Those products provide guaranteed income payments until death of the insured and are usually purchased at time of retirement. They are usually priced using few characteristics such as age and gender of the applicant. However, the last decades have seen the appearance of a new life annuities product, called underwritten or enhanced annuities, that propose annuities tailored to applicant's health status. Enhanced annuities seek at offering a better price, i.e. higher annuities amount, to applicants that have lower life expectancies than the average.

This master thesis focuses on the estimation of the level risk component of longevity risk, firstly to establish traditional life tables and secondly to establish models adapted to enhanced annuities. To do so, we considered THIN database, a large database representative of the UK population. This data set contains a large amount of information related to individual's health status, along with their survival.

Standard life tables estimation

The first part of the study consisted in constructing classic life tables, based on gender and a socio-economic variable called mosaic. Mosaic is a classification of households based on their residence postcode. The residencies are regrouped in 68 non-ordered groups.

In order to obtain a reasonable amount of produced life tables, without losing prediction performance, we constructed and used an agglomerative hierarchical clustering approach on the mosaic classes. This method consists in merging at each step the two modalities that led to the less loss in prediction power. We then obtained at each merging step a set of clusters of mosaics.

Once we determined the socio-economic clusters at each clustering step, we relied on a Generalized Additive Model (GAM) to construct our life tables per gender and cluster. In Figure 1 we show the AIC of the GAM model fitted on gender and mosaic clusters found at each clustering step.

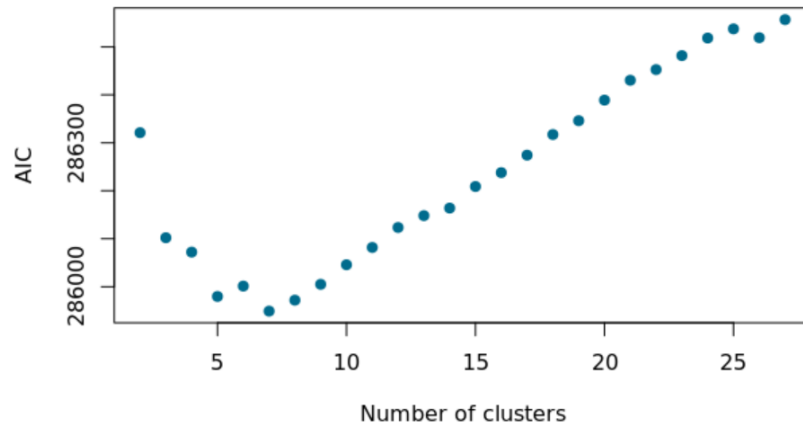


Figure 1: AIC of GAM model fitted on gender and mosaic clusters found at each clustering step

The optimal number of clusters is the one resulting in minimum AIC, namely 7. We thus obtained 14 life tables, one for each gender and mosaic cluster, as illustrated in Figure 2.

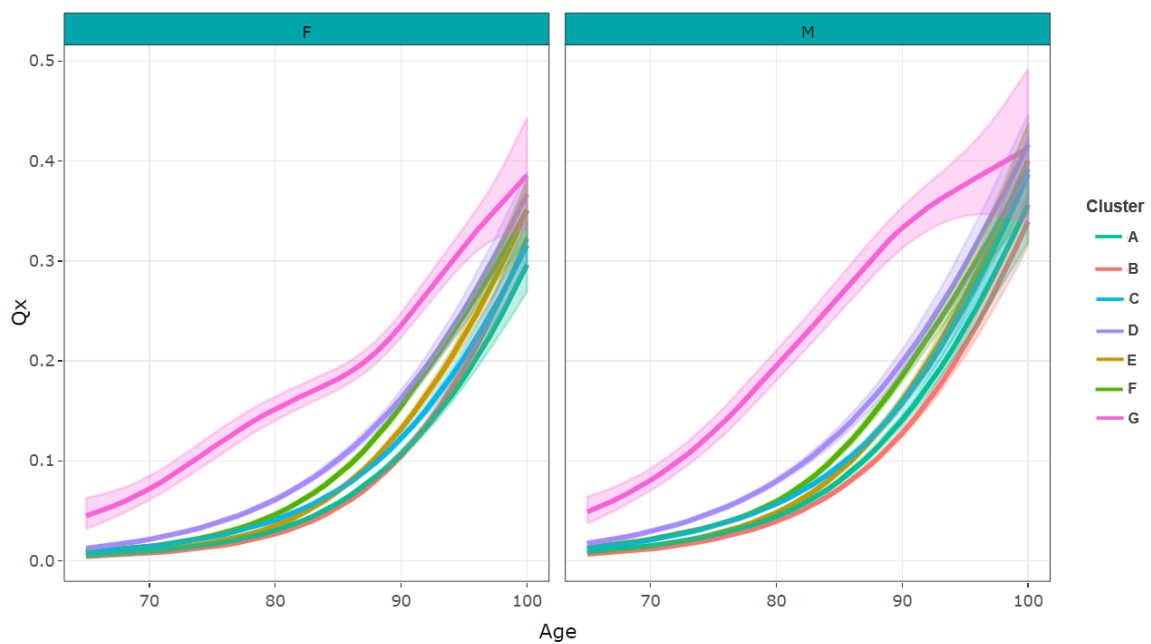


Figure 2: Fitted life tables per gender and socio-economic cluster using Binomial GAM model

Mortality rates estimation using health variables

Then, we considered health variables to predict mortality rates. We applied a traditional GLM model as well as a Machine Learning GBM model, using our fitted life tables as reference predictions.

When considering a classic approach as the GLM, one needs to assume a relationship between the response variable and the covariates. This relationship should reflect the impact of the health related variables on our mortality. We relied on descriptive statistic

analysis and medical knowledge during the feature engineering process in order to encompass all reasonable possible effects.

Afterwards, we applied a Lasso penalization in order to get rid of non-relevant terms and reduce the number of variables necessary to the model. Based on grid search on the penalization parameter λ , we decided to retain two penalization coefficients. The first penalization was given by λ_{1se} which is a commonly chosen penalization. The second penalization chosen was a little stronger and permeated to reduce a little more the complexity of the model without too much increase in deviance. In total, we fitted three models:

Model name	Penalization	Number of fitted coefficients	Number of variables used
GLM	0	143	38
GLMnet_long	$\lambda_{1se} = 8.51e-05$	77	32
GLMnet_small	$1.50e-03$	41	19

Table 1: Summary table of the GLM models fitted

In addition to those three models, we considered a Gradient Boosting Machine (GBM) model. We relied again on grid search to get the best hyperparameters in order to produce the best predictions.

The benefit of this model is that no relationship between the target variable and the covariates must be assumed. However, one limitation of these kinds of models is the limited interpretability. To avoid any issues with regulators, insurance companies aim to have the capacity to justify and explain the prices offered.

Some sensitivity analysis were completed on the GBM model, in the aim to unravel potential effects that had not been considered during the feature engineering process of the GLM. Partial dependencies plots confirmed the shape of impact of some variables. Also, Sobol indices estimation was performed on the GBM model in the desire to assess if we had taken into account the main explanatory variables interactions during the GLM feature engineering process. Unfortunately, estimation was computationally intensive, and our results were inconclusive and unstable.

Life expectancy predictions

After constructing our GAM life table model as well as the 3 GLMs and the GBM model using health variables, we computed residual life expectancies at age 65 based on the predicted mortality rates. We provide box plots showing the dispersion of period life expectancy predictions on our test data set per model:

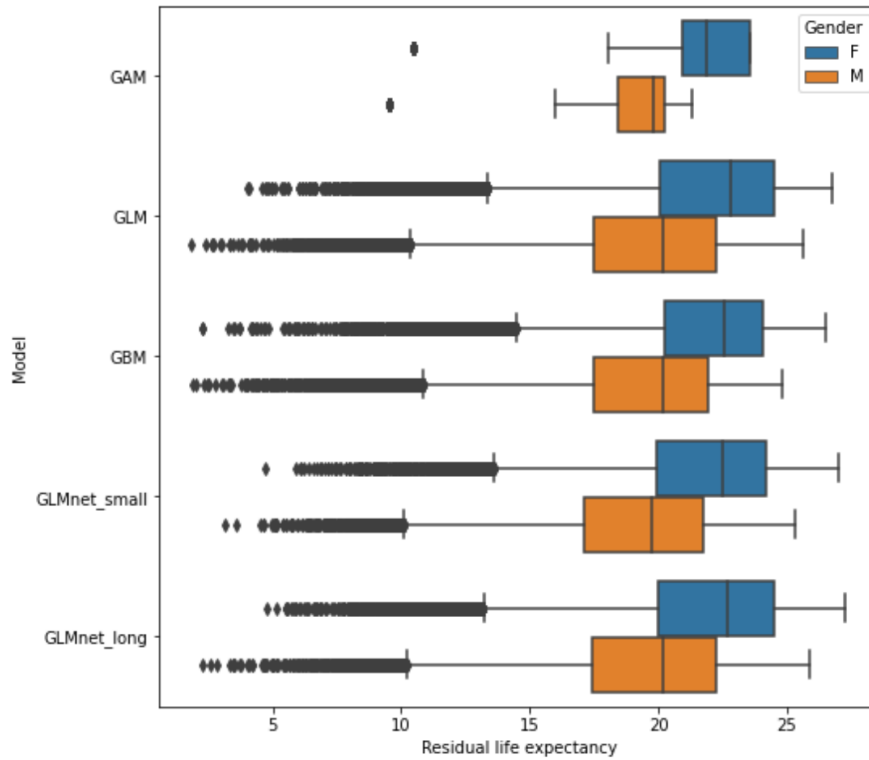


Figure 3: Distribution of predicted residual life expectancy at age 65 for each model

As expected, we visualize a significant increase of the segmentation, or discrimination, of the models that include health status variables compared to the GAM life table model.

Calibration performance

We assessed calibration accuracy using calibration plots and A/E¹ plots. Calibration metrics are statistics that partition a data set into groups and assess how the average predicted probability compares with the observed probability in each group.

We noticed that the models using health status variables were well calibrated, except for the GLMnet_small model which showed less accuracy in its calibration on some segments of the population.

Discrimination performance

We then gave interest in the ability of the models to discriminate the individuals. In survival models, discrimination performance refers to the ability of a model to predict whether an annuitant will live longer than another one's. This ability to discriminate is important for us, as the goal of enhanced life annuities is to offer more advantageous annuities to people that have a life expectancy shorter than others.

We relied on the C-index metric to assess discrimination performance:

¹Actual over Expected

Model	C-Index
GAM	58.37%
GBM	66.37%
GLM	66.20%
GLMnet_small	65.40%
GLMnet_long	65.92%

Table 2: C-Index discrimination metric of each model

We concluded that GBM had the best discrimination performance, however the GLM models have a close discrimination power.

Overall performance

On top of calibration and discrimination metrics, we assessed overall goodness of fit. We provide here the results of the metrics considered:

Model	pseudo- R^2	A/E
GAM	0.156	99.36%
GBM	0.180	99.44%
GLM	0.177	99.78%
GLMnet_small	0.170	100.23%
GLMnet_long	0.176	99.73%

Table 3: Overall performance metrics

Once again, the GLM models show a predictive performance close to the GBM one.

Conclusion

Based on the different calibration, discrimination and overall metrics, we concluded that the Gradient Boosting Model is the best predictive model. However, the GLM performances are very close. This means that the effect of all the main explanatory variable have been well taken into account during the GLM feature engineering process. Thanks to the data preparation and the insights on the underlying risk behaviour, we managed to reach the prediction power of a machine learning model.

We can also conclude that it would be reasonable to use one of the two GLMnet models. We could recommend the use of the GLMnet_long as it allows to reduce the complexity and the number of explanatory variables used with almost no loss in performances.

Note de synthèse

Contexte

Pour les compagnies d'assurance, le risque de longévité réside dans le risque que l'espérance de vie et les taux de survie dépassent leurs prévisions, entraînant des montants versés aux assurés plus importants que prévu. Le risque de longévité est le principal risque des produits de rentes viagères.

Ces produits permettent des versements de revenu garantis jusqu'au décès de l'assuré et sont généralement souscrits au moment de la retraite. Ils sont généralement tarifés en fonction de quelques caractéristiques telles que l'âge et le genre du demandeur.

Cependant, les dernières décennies ont vu l'apparition d'un nouveau produit de rentes viagères, dites rentes sur risque aggravé, qui proposent des rentes adaptées à l'état de santé du demandeur. Les rentes sur risque aggravé visent à offrir un meilleur prix, c'est-à-dire un montant de rente plus élevé, aux demandeurs de rente viagère dont l'espérance de vie est inférieure à la moyenne.

Ce mémoire s'intéresse à l'estimation du risque de niveau, composante du risque de longévité, d'une part pour établir des tables de mortalité traditionnelles et d'autre part pour établir des modèles adaptés aux rentes sur risque aggravé. Pour ce faire, nous avons utilisé la base de données THIN, qui est une grande base de données représentative de la population britannique. Ces données contiennent une grande quantité d'informations liées à l'état de santé de l'individu, ainsi qu'à sa survie.

Estimation de tables de mortalité classiques

La première partie de l'étude a consisté à construire des tables de mortalité classiques, basées sur l'âge, le genre ainsi qu'une variable socio-économique appelée *mosaic*. Le *mosaic* est une classification des ménages basée sur l'adresse de résidence. Les résidences sont classifiées selon 68 groupes non ordonnés.

Afin d'obtenir un nombre raisonnable de tables de mortalité produites, et sans perdre en performance de prédiction, nous avons construit et utilisé une approche de clustering hiérarchique agglomératif sur les classes *mosaic*. Cette méthode consiste à fusionner à chaque étape les deux modalités qui conduisent à la moindre perte de performance prédictive. Nous obtenons ainsi à chaque étape un jeu de clusters de *mosaics*.

Une fois les clusters socio-économiques déterminés à chaque étape du clustering, nous nous sommes appuyés sur un Modèle Additif Généralisé (GAM) pour construire nos tables de mortalité par genre et par cluster. Dans la figure 4, nous montrons l'AIC du modèle GAM ajusté par genre et par jeu de clusters obtenu à chaque étape.

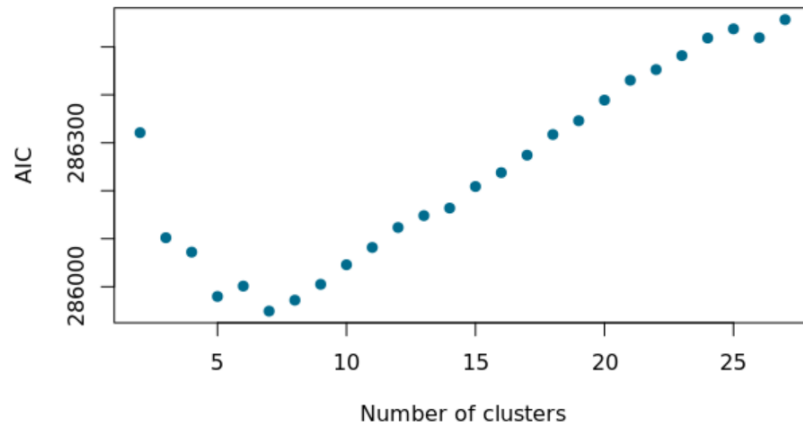


Figure 4: AIC du GAM ajusté sur le genre et les clusters de *mosaic* trouvés à chaque étape du clustering

Le nombre optimal de clusters de *mosaics* résultant en un AIC minimum est de 7. Nous avons ainsi obtenu 14 tables de mortalité, une par genre et cluster, comme illustré dans la Figure 5.

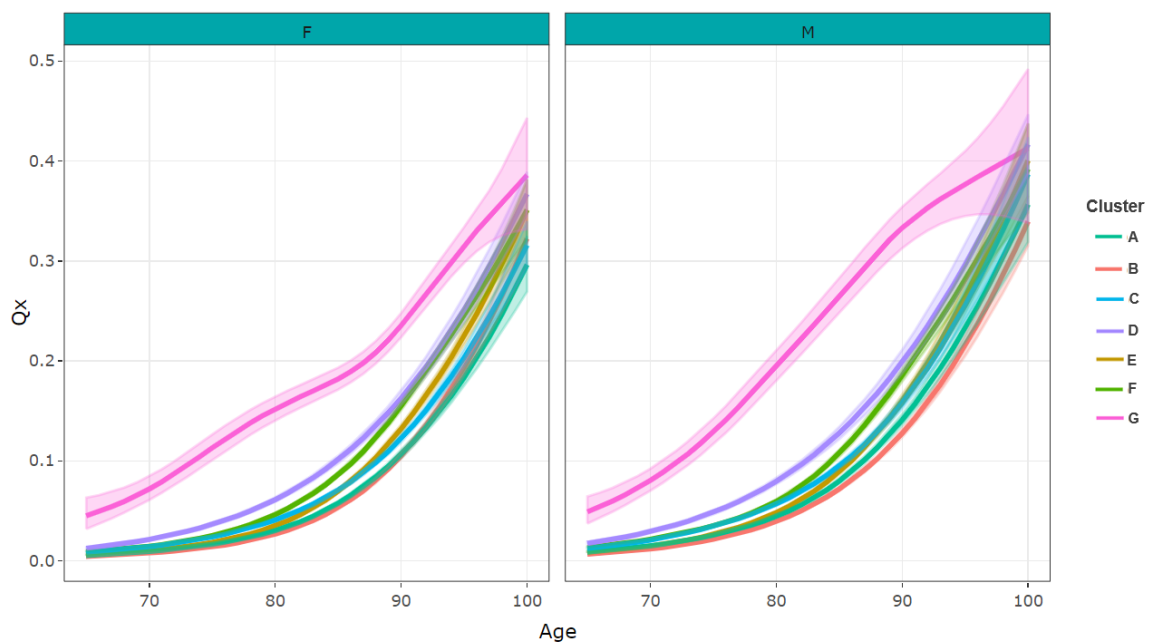


Figure 5: Tables de mortalité par genre et cluster socio-économique obtenues par un modèle GAM binomial

Estimation de taux de mortalité en fonction de l'état de santé

Ensuite, nous avons considéré des variables de santé pour prédire les taux de mortalité. Nous avons appliqué un modèle GLM traditionnel ainsi qu'un modèle de Machine Learning GBM, en utilisant nos tables de mortalité comme prédictions de référence.

Lorsque l'on considère une approche classique comme le GLM, il faut supposer une relation entre la variable de réponse et les variables explicatives. Cette relation devrait refléter l'impact des variables liées à la santé sur notre mortalité. Nous nous sommes

appuyés sur une analyse de statistiques descriptives et des connaissances médicales au cours du processus de *feature engineering* afin d’englober la majorité des impacts des variables explicatives.

Ensuite, nous avons appliqué une pénalisation Lasso afin de nous débarrasser des termes non pertinents et de réduire le nombre de variables nécessaires au modèle. Sur la base d’un *grid search* sur le paramètre de pénalisation λ , nous avons décidé de retenir deux coefficients de pénalisation. La première pénalisation a été donnée par λ_{1se} qui est une pénalisation couramment choisie. La deuxième pénalisation choisie était un peu plus forte afin de réduire un peu plus la complexité du modèle sans trop augmenter la déviance. Au total, nous avons équipé trois modèles :

Nom du modèle	Pénalisation	Nombre de coefficients	Nombre de variables utilisées
GLM	0	143	38
GLMnet_long	$\lambda_{1se} = 8.51e-05$	77	32
GLMnet_small	1.50e-03	41	19

Table 4: Tableau récapitulatif des modèles GLM

En plus de ces trois modèles, nous avons envisagé un modèle Gradient Boosting Machine. Nous nous sommes à nouveau appuyés sur un *grid search* pour obtenir les meilleurs hyperparamètres afin d’améliorer les prédictions.

L’avantage de ce modèle est qu’aucune relation entre la variable cible et les covariables ne doit être supposée. Cependant, l’une des limites de l’utilisation de ces types de modèles est leur interprétabilité limitée. Afin de respecter la réglementation, les compagnies d’assurance visent à avoir la capacité de justifier et d’expliquer les prix proposés.

Certaines analyses de sensibilité ont été réalisées sur le modèle GBM, dans le but de découvrir des effets potentiels qui n’avaient pas été pris en compte lors du processus de *feature engineering* du GLM. Une analyse de graphiques de dépendance partielle a confirmé des formes d’impact non-linéaire de certaines variables sur le taux de mortalité. De plus, une estimation des indices de Sobol a été effectuée sur le modèle GBM dans le but d’évaluer si nous avons pris en compte les principales interactions des variables explicatives. Malheureusement, l’estimation nécessitait des temps de calculs trop élevés et les résultats se sont révélés peu concluants et instables.

Prédiction de l’espérance de vie

Une fois avoir construit nos tables de mortalité GAM ainsi que les 3 GLM et le modèle GBM, nous avons calculé les espérances de vie résiduelles à 65 ans à partir des taux de mortalité prédits. Nous fournissons des boîtes à moustaches montrant la dispersion des prédictions d’espérance de vie périodique sur une partie de notre base de données de test pour chaque modèle :

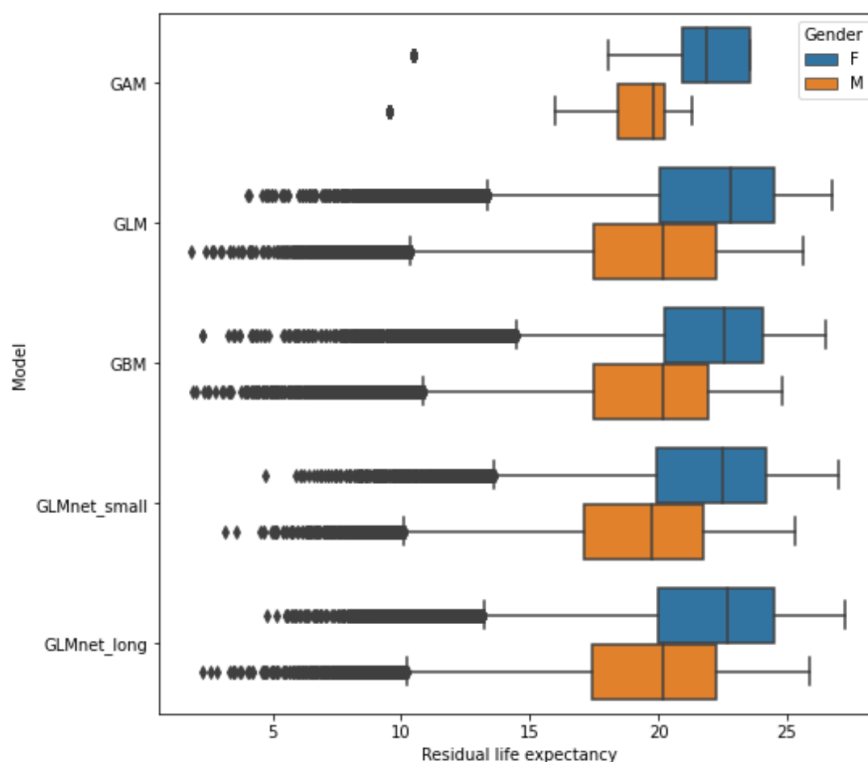


Figure 6: Distribution de l'espérance de vie résiduelle prédite à 65 ans pour chaque modèle

Comme anticipé, nous visualisons une augmentation significative de la segmentation, ou discrimination, des modèles qui incluent les variables d'état de santé par rapport au modèle de tables de mortalité GAM.

Performance de calibration

Nous avons ensuite évalué la précision de calibration de nos modèles à l'aide de *calibration plots* et de mesures de A/E^2 . Les métriques de calibration sont des statistiques qui divisent un ensemble de données en groupes et évaluent comment la probabilité prédite moyenne se compare à la probabilité observée dans chaque groupe.

Nous avons remarqué que les modèles utilisant des variables d'état de santé étaient bien calibrés, à l'exception du modèle GLMnet_small qui montrait un calibrage moins correct sur certains segments de la population.

Performance de discrimination

Nous nous sommes ensuite intéressés à la capacité des modèles à discriminer les individus. Dans les modèles de survie, la performance de discrimination fait référence à la capacité d'un modèle à prédire si un rentier vivra plus longtemps qu'un autre. Cette capacité de discrimination est importante pour nous, car le but des rentes viagères sur risque aggravé est d'offrir des rentes plus avantageuses aux personnes qui ont une espérance de vie plus courte que les autres.

²Actual over Expected

Nous nous sommes appuyés sur la métrique C-index pour évaluer les performances de discrimination :

Modèle	C-Index
GAM	58.37%
GBM	66.37%
GLM	66.20%
GLMnet_small	65.40%
GLMnet_long	65.92%

Table 5: Métrique de discrimination C-Index de chaque modèle

Nous avons conclu que le GBM avait les meilleures performances de discrimination, cependant les modèles GLM ont un pouvoir de discrimination proche.

Performance globale

En plus des mesures de calibration et de discrimination, nous avons évalué la qualité globale de l'ajustement. Nous donnons ici les résultats des métriques considérées:

Modèle	pseudo- R^2	A/E
GAM	0.156	99.36%
GBM	0.180	99.44%
GLM	0.177	99.78%
GLMnet_small	0.170	100.23%
GLMnet_long	0.176	99.73%

Table 6: Indicateurs de performances globales

Encore une fois, les modèles GLM montrent une performance prédictive proche de celle du GBM.

Conclusion

Sur la base des différentes métriques de calibration, de discrimination et de mesures globales, nous avons conclu que le modèle de Gradient Boosting est le meilleur modèle prédictif. Cependant, les performances du GLM sont très proches. Cela signifie que l'effet de toutes les principales variables explicatives a été bien pris en compte lors du processus de *feature engineering* du GLM. Le travail réalisé sur la compréhension des impacts non-linéaires et des interactions des variables explicatives nous a ainsi permis d'atteindre la puissance de prédiction d'un modèle de machine learning.

Nous pouvons également conclure qu'il serait raisonnable d'utiliser l'un des deux modèles GLMnet. On pourrait préconiser l'utilisation du GLMnet_long car il permet de réduire la complexité et le nombre de variables explicatives utilisées sans quasiment aucune perte de performance.

Remerciements

Je tiens à remercier en premier lieu Razvan Ionescu qui a été présent tout au long de mon travail. Je le remercie pour sa disponibilité et le temps consacré à l'encadrement de mon mémoire, ainsi que pour le partage de ses connaissances sur les modèles de durée. Et je le remercie avant tout pour sa bonne humeur qui m'a permis de conserver ma motivation et d'affronter les nombreux challenges qui ont pu ponctuer ce mémoire.

Je remercie également les membres de l'équipe BRM, Julien Tomas, Tiziana Torri, Agne Ulcinaite, Thomas Poinignon et Antoine Burg pour leurs conseils et l'équipe médicale, Manuel Plisson et les médecins, qui ont pu répondre à mes nombreuses questions.

Je tiens également à remercier mes collègues avec qui j'ai commencé mon aventure chez SCOR et qui ont su m'accueillir dans les meilleurs conditions, Antoine Moll, Denis Charles et Valentine Sarrazin.

Je remercie plus globalement les membres du Knowledge pour leur bienveillance et l'ambiance toujours positive et motivante.

Je tiens également à remercier Caroline Hillairet pour sa relecture et ses remarques qui m'ont permis d'améliorer la qualité de mon mémoire.

Contents

Abstract	i
Résumé	ii
Executive summary	iii
Note de synthèse	viii
Remerciements	xiii
Introduction	4
1 History on life annuities and mortality modeling evolution	5
1.1 Early ages - antiquity and middle age	5
1.2 First developments - 17 th century	5
1.2.1 First mortality table by J. Graunt	6
1.2.2 Life expectation calculation by the Huygen brothers	6
1.2.3 First annuity valuation method by Jan de Witt	7
1.2.4 Modern modelling by Edmond Halley	8
1.3 Importance of correct life annuity valuation - end 18 th century	8
1.4 Findings and theory about mortality modelling - from the 19 th century to nowadays	10
1.4.1 Modelling death rates from empirical life tables	10
1.4.2 Mortality forecasting and the construction of an annuity product	11
2 Life annuities products and embedded risks	12
2.1 Life annuity product presentation	12
2.1.1 Standard life annuity	12
2.1.2 Enhanced life annuity	13
2.2 Risks attached to this product	13
2.2.1 Longevity risk	13
2.2.2 Financial risk	14
2.2.3 Risks specific to enhanced annuities	14
3 Presentation of the data	15
3.1 THIN database	15
3.1.1 Presentation	15
3.1.2 Data collection	16
3.1.3 Available information in the database	16
3.2 Processing of the data and some statistics	16

3.2.1	Study period and selection of patients	16
3.2.2	Transformation into longitudinal data	17
3.3	Descriptive statistics	19
4	Estimation of raw death rates	25
4.1	Methodology	25
4.1.1	Lexis diagram	25
4.1.2	Mortality rate	26
4.2	Exposure to risk	27
4.2.1	Incomplete observations - censorship and truncation	27
4.2.2	Initial exposure to risk	27
4.3	Application	29
5	Construction of reference life tables by age, gender and socio-economic group	31
5.1	Life tables by age and gender	31
5.1.1	Methodology	32
5.1.2	Results	34
5.2	Life table by gender and socio-economic clusters	36
5.2.1	Mosaic : UK indicator of socio-economic group	36
5.2.2	Descriptive statistics	36
5.2.3	Impact of mosaic type on mortality	37
5.2.4	Clustering	38
5.2.5	Final life tables by gender and socio-economic cluster	40
6	From standard life tables to enhanced mortality rates using GLM and Machine Learning models	42
6.1	Modelling enhanced life annuities GLM	42
6.1.1	Model choice and inclusions of reference rates	42
6.1.2	Classic GLM	43
6.1.3	GLM-Lasso	45
6.1.4	Explaining mortality rate differences with GLMs	47
6.2	Modeling enhanced life annuities using a Machine Learning model	51
6.2.1	Gradient Boosting Machine	51
6.2.2	Grid search	52
6.2.3	Sensitivity analysis	52
7	Life expectancy prediction	56
7.1	Residual life expectancy	56
7.2	High Ages Extrapolation	57
7.3	Analysis of computed residual life expectancy	58
7.3.1	Average predicted residual life expectancy at age 65	58
7.3.2	Segmentation of prediction	58
7.4	Explaining life expectancy predictions with GLMs	61
8	Assessing models performance	64
8.1	Calibration performance metric	64
8.1.1	Calibration plot	65
8.1.2	A/E per risk factor	66
8.2	Discrimination performance metric	69

8.2.1	C-index	69
8.2.2	Computation difficulties	70
8.2.3	Discrimination metric results	71
8.3	Overall performance metrics	71
8.3.1	McFadden's Pseudo- R^2	71
8.3.2	Brier score	72
8.3.3	A/E	72
8.3.4	Application of overall metrics	72
8.4	Conclusion on metrics analysis	73
	Conclusion	74
	Dictionary	75
	Bibliography	77
	Appendix	78

Introduction

Life annuities on aggravated risk (impaired annuities or enhanced annuities) are life products offering more advantageous annuities for people whose life expectancy is lower than the average. Access to this type of product is favoured by persons suffering from one or several morbidities, chronic diseases such as diabetes or hypertension or more generally cardiovascular diseases. It can also depend on the lifestyle of the insured (regular smoker or not).

This type of insurance product is currently offered by 6 providers, including 5 in the UK and 1 in Canada. They offer increased annuities from 9% to 35% depending on annuitants characteristics. [24].

SCOR is a reinsurance company offering coverage for longevity risk. As of 2021, SCOR longevity risk exposure represented around 10% of total life and health risk portfolio. The majority of their longevity risk exposure are located in the UK market, and a part of them are covering enhanced annuity products. In this context, SCOR wants to update their mortality level estimates for the UK population.

The first part of this master thesis will focus on constructing traditional life tables using a socio-economic variable.

In the second part of this study, we will use those constructed life tables as reference predictions for models adapted to assess the level risk of enhanced annuity products. We will estimate the mortality rates of individuals using some of their lifestyle and health status variables. To meet with business constrains, the models that we want to use should be interpretable in order to be in the capacity to explain the offered prices. Therefore, we will focus on GLM models. However, in the context of adverse-selection threat, we will compare their performance with a Machine Learning model, in order to assess whether the GLM segmentation is as effective as a more complex black-box model.

The last part of the study will focus on assessing the performance of our model through a variety of metrics adapted for survival analysis.

The following work will focus on the level risk component of longevity risk.

Chapter 1

History on life annuities and mortality modeling evolution

In this chapter, we provide a brief history of life insurance with a focus on life annuities products. To fully understand historical facts, one must understand the background and the historical context. We will be very brief in this respect as our goal is to take a bird eye view on the mortality modelling evolution, the challenges triggered by life annuities products and the solution provided.

1.1 Early ages - antiquity and middle age

Life annuities is one of the oldest life insurance coverage. The first proof of existence of life annuity products come from the Egyptians in Antiquity [18]. But even though the Egyptians had started census, it remains uncertain that the premiums of those products were based on any mathematical tools.

Centuries later, around 200 AD, the Roman jurist Ulpian proposed a life annuity pricing grid where the value of a life annuity was decreasing with annuitant age. Some consider this pricing grid as the first life table ever created. However, it is most certainly more an *educated guess* than an accurate risk estimation. Indeed, the impact of age on life annuity price was supposed linear.

In the following centuries, selling life annuities was a popular and quick way for European kingdoms and governments to fill the state reserves. This practice triggered the need to assess annuity buyer's life expectancy.

We will have to wait the 17th century for improvements on life prediction theory.

1.2 First developments - 17th century

Life annuities found a new source of development during the 17th century to finance European numerous wars. European governments used annuities as a source of public finance. Selling life annuities was a simple and quick way to raise money for the ongoing wars. From there came the need to have solid mathematical tools to predict life expectancy. But as we will see, life contingent financial products were not priced following the mathematical theory that were in development at that time, for reasons that we will expose.

By the end of the 17th century, most of the problems concerning the valuation of fixed term annuities like discounting and compounding techniques had been resolved and permitted to determine the return of those products. But unlike the fixed term annuity, life annuity valuation technics were limited.

1.2.1 First mortality table by J. Graunt

No real improvement had been made concerning mortality modelling, until John Graunt introduced in 1662 the concept of life tables. Life tables are difficult to produce due to the amount of data necessary during a period avoiding wars or pandemics. However, during the 16th century, the Church of England started keeping records of marriages, burials and christenings. Those registers were called bills and were published every week. Those so-called Bills of London have been used by the London businessman Graunt. He published a book called "Natural and Political Observations on the Bills of Mortality" based on about 70 years of data. The Bills of London did not provide age at deaths but only number of deaths per cause each week. Depending on the nature of the cause of death (infant illness, elder illness) and some conjectures, Graunt distributed the deaths across some age bands for his study. He provided analysis on the evolution of deceases and elaborated a mortality table. His work founded the fields of demography and epidemiology, and contributed to actuarial sciences. Here is the life table he proposed:

Table 1. Graunt's Life Table.		
Age Interval	Prop. Deaths in Interval	Prop. Surviving til start of Interval
0-6	0.36	1.00
7-16	0.24	0.64
17-26	0.15	0.40
27-36	0.09	0.25
37-46	0.06	0.16
47-56	0.04	0.10
57-66	0.03	0.06
67-76	0.02	0.03
77-86	0.01	0.01

Figure 1.1: Graunt's life table based on the Bills of London

with prop.=proportion.

Although Graunt's table is now considered as incorrect, it gained the attention of some mathematicians as de Witt, Huygens, and Hudde that considered it and proposed improvements.

1.2.2 Life expectation calculation by the Huygen brothers

Huygens brothers found a use to Graunt's life table, they started a correspondence in 1669 on the expectation of life and an application to life annuities valuation. They developed the concept of mathematical expectation, but they stopped their work before applying it to life annuities valuation.

In their correspondence, the Huygens brothers provide the first correct probability based formulations of expected age at death and life expectancy. They depict a life

table as a continuous function, calculate median additional years of life, and draw a sharp distinction between median years of life and life expectancy.

Coming from the field of game theory, and with no probability theory developed at that time, the vocabulary they employed was the one of game of chances, considering the life table as a lottery [6]. Christiaan Huygens started working on the computation of mean and median life expectancy, as well as joint-life expectation. He considers Graunt's life table as a uniform distribution, creating the first survival function $S(x)$.

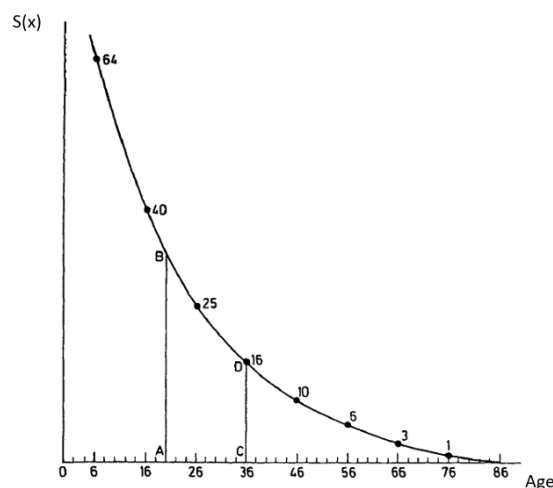


Figure 1.2: Huygens's survival function based on Graunt's life table [6]

He explains graphically how to get the median life expectancy for a person of a given age. For example, a person at age A has a median expectancy of AC . This is obtained by taking the point C such that there are half fewer individuals still alive at age C than at age A .

Even though the Huygens brothers were the firsts to do life expectancy calculation, it is later in 1746 that Antoine Deparcieux created and defined the notion of life expectancy.

Since the valuation of a life annuity only requires that the random variable of life duration T be replaced by the value of an annuity certain of fixed expected duration T , the Huygens brothers work has been a big step for life annuity valuation.

1.2.3 First annuity valuation method by Jan de Witt

Jan de Witt was a mathematician and the Grand Pensionary of Holland from 1653 to 1672. This period is called the golden age of Holland, even if it includes war with England (1665-1667) and hostilities with France. Like many other European contemporary government leaders, Jan de Witt relied on life annuities to finance the military needs. However, he was concerned about the possible miss-price of life annuities, a justified concern as life annuities premium was not age dependent [33].

De Witt corresponded with both Huygens and Hudde on topics evolving around mortality rates, mortality tables, and the valuation of annuities of one or more lives. In 1671, he published "Value of Life Annuities in Proportion to Redeemable Annuities", developing a method for relating the value of annuities to the age of the beneficiary. He showed

how to calculate the value of annuities using a constant interest rate and a hypothetical mortality table introducing thus the probabilities over the duration of human life [21].

In his approach, de Witt uses the concept of mathematical expectation developed by Huygens. His theoretical work has been upheld by the empirical work of Hudde on mortality statistics of life annuitant of Amsterdam from 1586 to 1590.

He calculated the price of a life annuity as the sum on every age t of the present value of an annuity of payment 1 until age t . Using modern notation, de Witt is the initiator of the now know following formula:

$$\begin{aligned} a_x &= \mathbb{E}[a_{\overline{T}|}] \\ &= \frac{1}{l_x} \sum_{t=1}^{w-x-1} a_{\overline{t}|} d_{x+t} \end{aligned}$$

with T the random variable of the remaining lifetime of someone aged x , l_x the number of person alive at age x , d_{x+t} the number of death observed at age $x + t$, and w the assumed maximum age.

Based on his recommendation, in 1672 the city of Amsterdam began offering life annuities with prices dependent on the age of the nominee. However, this practice did not become widespread.

1.2.4 Modern modelling by Edmond Halley

In 1693, Edmond Halley, the famous English astronomer, used the birth and death records of the city of Breslau (Poland) to construct a life table and used it to price life annuities. According to some historians, his work greatly influenced the development of actuarial science. The approaches and methods that Halley introduced are very close to the methods that actuaries are using today to price life annuities. Three centuries later, we are still relying on this work. [13]

1.3 Importance of correct life annuity valuation - end 18th century

During the 18th century, governments have used life contingents as a way to fill in quickly the treasury. However, these loans were not always priced correctly as we will see with the example of France.

Life annuities became the largest debt component of French during 18th century

Since 1522, French government used to fill their coffer by selling the so-called *rentes perpétuelles*, that one could translate to perpetual annuities. The annuity only obliged the debtor to make the annual payment of interest on the capital loaned and could in theory not include a maturity date: it was extinguished by a refund that could not be imposed or denied to debtors. These products made up most of France public debt in the 16th and 17th centuries. However, this started to change under Louis XIV.

Life annuities started to be used by the government of Louis XIV in 1653 as a new way to fill in the treasury. They were a way to get money without increasing government debt. Indeed, life annuities products are financial products with no refund of the capital, only payments of interest until death of the insured. Therefore, the French government having too much debt, they decided to reduce it by selling life annuities and using the cash to pay off part of the capital of their debt [16].

Life annuities remained a small part of interest costs until the death of Louis XIV in 1715. However, after the death of Louis XIV, they quickly became the largest component of French debt by 1789 [29]. During Louis XIV reign, French public debt had increased due to its involvement in costly wars and its aristocracy's extravagant spending. At his death, the debt was so high that they had no other choice than to reduce the return of the existing perpetual annuities. It is at that time that French government had to move away from perpetual annuities after 1720, replaced by life annuities [20].

The following graph shows the evolution of the nature of national public finances.

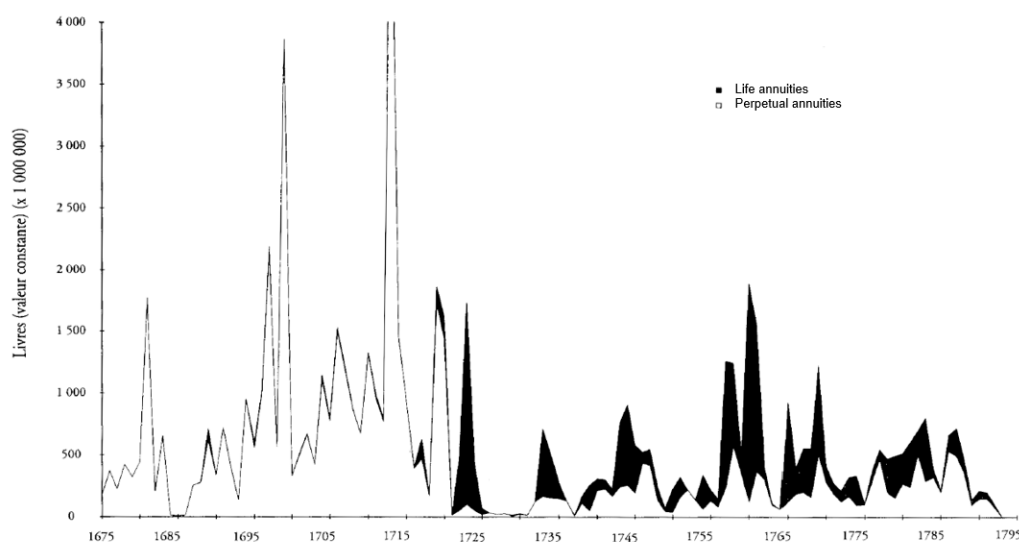


Figure 1.3: Annual volume of public borrowing, 1675-1795 (source [19])

From the debt burden to the political impasse that led to French Revolution

As we can see on Figure 1.3, in May 1789, the perpetual annuities and life annuities were roughly similar in terms of invested capital (more than 1.1 billion for each type), but the life annuities represented two-thirds (65%) of the interest, which amounted to more than 160 million per year for these two parts.

Life annuities mispricing increases the debt

Among the explanations put forward to understand the bad returns of tontines, it was assumed that those responsible for French finances underestimated the then booming actuarial science: as proof, the life annuities were still not correctly valued after the publication of the pricing tables of life annuities and tontines by Deparcieux (1746) [19]. Indeed, those life annuities were sold without distinction on age during the most desperate times (usually times of wars), in order to attract capital to finance wars. This

was particularly the case during the years 1780s, when Necker launched life annuities at the denier 10 (10%), regardless of the age of the subscriber. Even though progress had been made in actuarial sciences and first mortality tables had been computed, life annuities were not sold accordingly to their actuarial price.

The literature is divided around the causes of this miss-pricing. Here are possible explanations why. First, the need for finances in times of war was high and so the financial products needed to be attractive to be sold in big quantities. Second, looking at the frequent defaults in the history of French debt, a default risk premium had to be incorporated in the product as well. Finally, the mathematical tools for valuating contingency risk were still in development at that time and actuarial sciences seemed to not be trusted enough for the valuation.

This led to a massive anti-selection as life annuity buyers were well aware about the real value. This was aggravated as one could buy an annuity in someone else name. This feature, pushed Swiss bankers to create an investment vehicle, the *Thirty Maidens of Geneva*, where annuitants were 30 young ladies. The 30 annuities were pulled together into a single vehicle and shares of the vehicles were sold. This allowed investors to diminish the mortality risk and maximize their income. [14]

From the burden of the debt to the political impasse that led to French Revolution

Unlike Britain that paid for their debt by raising taxes, France favored to contain its debt by frequent partial defaults. When Louis XVI acceded to the throne in 1774, he announced his intentions not to default. However, to honor the debts without default, the government had to borrow even more, leading to increase even more the debt. Therefore, the other solution of raising taxes gained interest. But in order to do that, Louis XVI had to call the General Estates in July 1788. He eventually had to agree to the doubling of votes of the Third Estate in hope for a constitutional solution. This was the occasion for the Third Estate to speak out about the desire for increased control on public finances (among other things). But the resulting political impasse over budget reform ended up to the French Revolution [28] [29].

Epilogue

The case was made by Joseph Cambon, in charge of restructuring the French national debt, in 1794 after the French Revolution. To do so, he relied on the best actuarial science available at the time. He denounced French debt, and more specifically life annuities, as being "ruinous, impolitic and immoral" [28], especially the ones sold after 1770 at a flat rate not depending on age and at a yield higher than the market.

1.4 Findings and theory about mortality modelling - from the 19th century to nowadays

1.4.1 Modelling death rates from empirical life tables

Starting with de Moivre, an interest was to fit a model based on empirical observations of the death rate by age. The question was to understand the appropriate relationship of death rates and age.

De Moivre's linear assumption for interpolation of life tables

De Moivre assumes uniform deaths distribution between each discrete age to interpolate life table and that way obtain a continuous survival model from a discrete one. He is the first to formalize this model of a continuous model for human survival, even though his model is very simple.

Gompertz model of the force of mortality

Taking a biological approach to mathematical modelling, Gompertz (1825) assumed that the force of mortality starting at age 50/60 shows a nearly exponential increase, where the two parameters of his model are positive and vary with the level of mortality and the rate of increase in mortality with age. The Gompertz model has been modified by Makeham (1867), to include an additional constant to take into account mortality level unrelated to age [23]. Many models with various forms appeared following to Gompertz model.

The above models describe mortality at a fixed point in time; however, actual mortality is stochastic and evolving continuously. Thus, while the mortality models described above are static, the parameters must be fitted periodically to accommodate changes in mortality patterns.

1.4.2 Mortality forecasting and the construction of an annuity product

The presence of a downward trend in mortality from the turn of the 20th century encouraged annuity issuers to include future improvements in mortality rates in order to be protected against future losses. The first tables to take this phenomenon into account were those produced in the United Kingdom on the basis of data from insurance companies covering the period 1900-1920. The now famous Lee-Carter model is the reference model for prospective mortality table.

Today, the mortality assumptions of a portfolio of annuitants are constructed in two stages. First, it is necessary to measure the current mortality of annuitants. Then, it is necessary to apply a projection of future mortality in order to anticipate downward trends in mortality. It is this first component on which we will be working on in this study.

Chapter 2

Life annuities products and embedded risks

2.1 Life annuity product presentation

2.1.1 Standard life annuity

As said in the first chapter, life annuities are products that pay an annuity to the policyholder until they die. The goal of this product is to guarantee an income until death of the annuitant. Its price is usually a function of the age and gender, but it may also depend on other factors.

The annuitant pays a lump sum that corresponds to the price of the product, and in exchange receives an annuity amount at a regular interval until their death.

The following graphic shows the cashflows generated by a life annuity product in its simplest form:

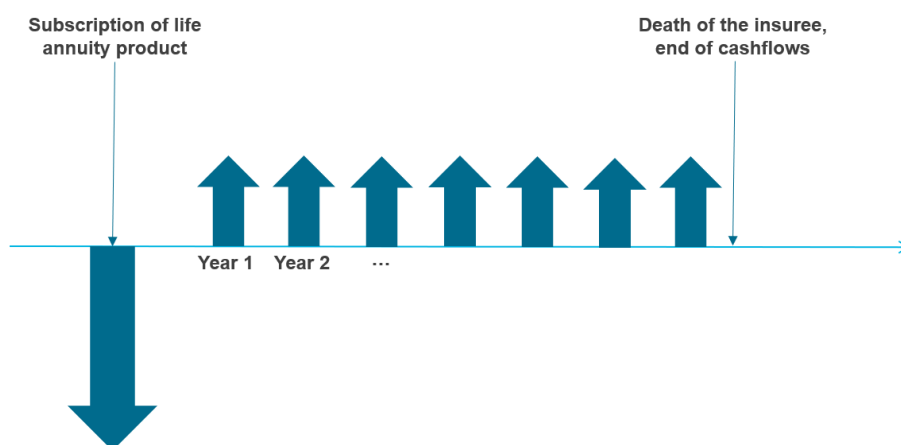


Figure 2.1: Cashflows of a life annuity product, simple case

Multiple forms of this product exist. As an example, deferred life annuities are annuities that start at a deferred date and for which the annuitant can contribute during the time of their career.

2.1.2 Enhanced life annuity

Enhanced annuities, also called underwritten annuities, allow insurers to take into account a person's lifestyle and medical history that may impact their life expectancy.

Enhanced annuities were developed around 1995 to enable to offer higher annuity payments to annuitants with a lower life expectancy than the average.

The applicants disclose their medical information to enable the insurer to estimate their life expectancy and to offer a life annuity price accordingly.

2.2 Risks attached to this product

Annuities, as every insurance products, are subject to risks. Here are presented the risks attached to this product.

2.2.1 Longevity risk

Longevity risk encompasses all the risk that are not market or financial risk. It is the risk that mortality of the annuitant is lower than the one expected by the insurer, or equivalently that annuitant lifespan is longer than anticipated by the insurer, on average on the insurer portfolio. If the insurer expected a lower life duration than the one observed, then it will result in higher pay-out ratios.

Longevity risk can be decomposed in three components as depicted in the following visualisation:

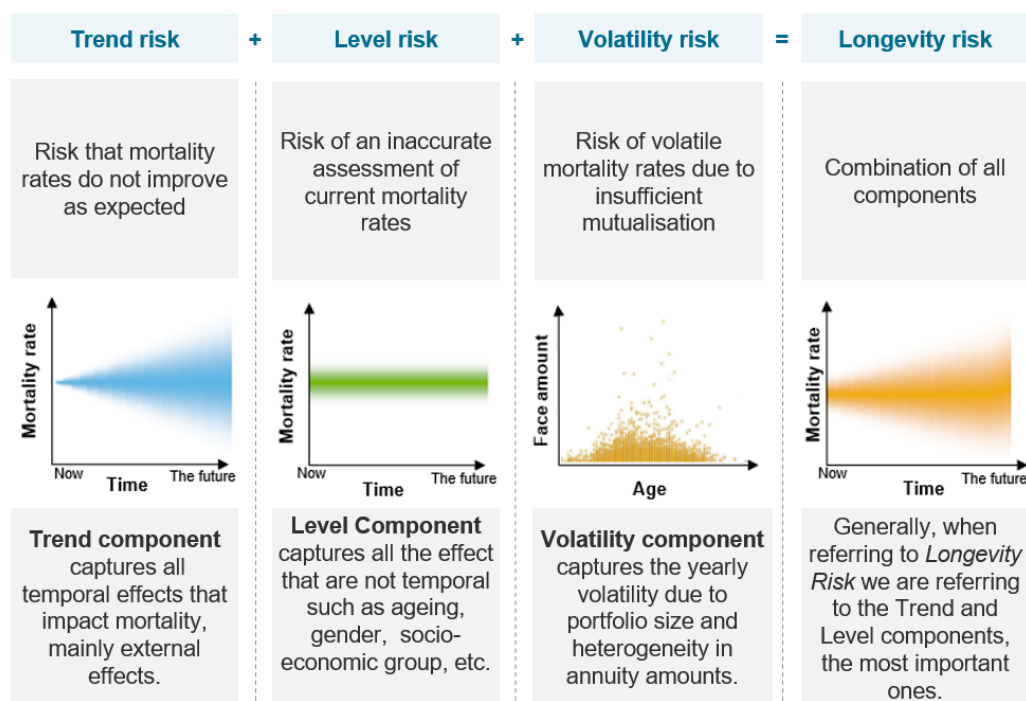


Figure 2.2: Decomposition of longevity risk (source : Internal)

- **Trend risk** is the risk that the future mortality rates differ from the predicted ones. It comes from a bad estimation of the evolution of mortality when mortality

is improving or reducing faster than expected. It is a systemic risk, if the improvement of mortality rates is underestimated, every life expectancy will be underestimated.

- **Level risk** is the risk of an estimation error on current mortality rates of the portfolio. This can occur when the sample population is too small or when this population is not representative of the portfolio (basis risk). For instance, when estimating mortality levels at high ages, only limited data is available. Therefore, level risk is usually high at high ages.
- **Volatility risk** is the risk due to the intrinsic volatility of each life duration composing the portfolio. High volatility risk may arise from a heterogeneous portfolio or from a too small portfolio. This risk can be reducing by pooling several portfolios using reinsurance.

2.2.2 Financial risk

When a life annuity product is subscribed, the received premiums (lump sum) are invested and a level of return is expected. The financial risk is the event that the financial income obtained is below the technical rate expected. The technical rate is defined as the rate used for updating the commitments of the insurer to the annuitant, and therefore constitutes the minimum reevaluation of the annuities that it guarantees each year to the annuitant.

2.2.3 Risks specific to enhanced annuities

In addition to these risks, and like any product with medical selection and asymmetric information, enhanced annuities may be subject to the risks of adverse selection and moral hazard.

- **Adverse selection** for this type of product results in a selection of policyholders who meet the conditions to subscribe to the product, but are in better health on average than the intended targeted annuitants. This type of adverse selection would consist for the applicant to "hide his good health".
- **Moral hazard** can occur in the event of an improvement in lifestyle declared by the insured at the time of subscription (change from smoker to non-smoker, for example).

We mention once again that the following study will focus on longevity risk only, and more specifically on the level risk component.

Chapter 3

Presentation of the data

3.1 THIN database

3.1.1 Presentation

SCOR's main longevity risk exposition being located in the UK, we are going to use a database recording information from the UK population.

The database we are using to conduct this study is called the THIN (The Health Improvement Network) database, which is a large UK primary care database [1]. It is provided through the IQVIA Medical Research Database (IMRD). It contains reliable data on the health status and medical treatments of the population and enables follow-up studies. Its primary purpose is to enable research in cardiovascular disease, mental health, pharmacoepidemiology and other areas of research in primary care.

THIN collects anonymized patient medical data from its member physicians. This means that data is collected at the level of the doctors, allowing a reliable transmission of information. It contains data collected from over 600 General Practitioner (GP) offices (an office can contain several practitioners) across the UK, covering 6% of the UK population. GPs treat all common medical conditions and refer patients to hospitals and other medical services for urgent and specialist treatment. They focus on the health of the whole person by combining physical, psychological and social aspects of care. GPs are often the first point of contact for anyone with a physical or mental health problem. They play a role in treatment of cardiovascular, metabolic and respiratory diseases and mental health problems, as well as chronic conditions such as asthma, hypertension and diabetes [3].

A limit of this database is the collection of information at the level of primary care practitioners only. The analysis of diseases treated by secondary care practitioners is therefore more difficult due to the collection of partial information transmitted by the patient's general practitioner. For instance, in cancer risk analysis, transmission of the information is not assured by the GP and no details on the severity of the disease or treatments used in secondary care will be provided. Cancer data is considered as being unreliable. This is why we will not be using cancer data in this study since other databases provide much more detailed information (eg. SEER database).

3.1.2 Data collection

As said previously, health data are collected through General Practitioners. GPs who are part of the program record information on each consultation. Each time a patient consults a GP, the GP inputs into a software the purpose of the visit, the drugs prescribed if any and the biometric measurements performed if any, such as height and weight or blood pressure. Additional information are also recorded as the date of consultation and the patient ID. THIN compile all those data and then proceeds to an anonymization of the patients.

In order to ensure that it is not possible to identify a patient, some variables are altered and lose some degree of precision. For example, for the date of birth only the year of birth is retained, and the postcode address of the patient is replaced by a socio-economic mosaic type (mosaic variable will be explained in the next section).

3.1.3 Available information in the database

The database includes 4 types of information collected:

- Information on social class: Based on the address of residence of the patient, a lifestyle profile is inferred. Three variables contribute to the lifestyle profile ; Townsend, IMD¹ and Mosaic class. Those three variables are all inferred based on the residential area.
- Additional Health Data (AHD) information: smoking status, duration of smoking, intensity of smoking, blood pressure, cholesterol level, BMI², waist circumference, alcoholism, alcohol intensity.
- Medical information: Dates of occurrence of major impairment events that includes: heart attack, stroke, angina, hypertension, heart failure, atrial fibrillation, diabetes, COPD³, asthma, kidney disease, Alzheimer's, dementia, high cholesterol, neuropathy, retinopathy, amputation, other circulatory diseases, etc. For some cases, the dates of recurrence are also provided. (e.g. heart attack)
- Therapy information: History of drugs prescribed for the patient by the GP, including drug names and prescription dates. As said previously, because the THIN database is a primary care dataset, this data does not have a good record history of cancer treatments. Chemotherapy and / or radiotherapy are performed in hospitals, and general practitioners do not record all treatments received outside primary care.

3.2 Processing of the data and some statistics

3.2.1 Study period and selection of patients

Study period: Our goal is to assess level risk. Therefore, we want to have enough data to reduce the estimation risk, however we want to avoid having too old mortality data that would not reflect the current mortality. Taking those two opposite constrains into account, we chose to consider the deaths observed between years 2014 to 2019.

¹Index for Multiple Deprivations

²Body Mass Index

³Chronic Obstructive Pulmonary Disease

Therefore we will consider only patients that were observed during part or all of those 6 years period.

However, the construction of our database will use records of medical events that could have occurred prior to 2014 in order to construct the health status variables of the patients.

Patients selection: Because we are interested in a life annuity product designed for newly retired persons, we are interested in the mortality risk after age 65. For that reason, we filtered our database and considered only patients 65 and older during the study period.

3.2.2 Transformation into longitudinal data

As said previously, information is recorded in databases when they are entered by the GPs during the consultation. This data contains the patient ID, the date of consultation and the collected medical information and are stored in different databases depending on the nature of the information.

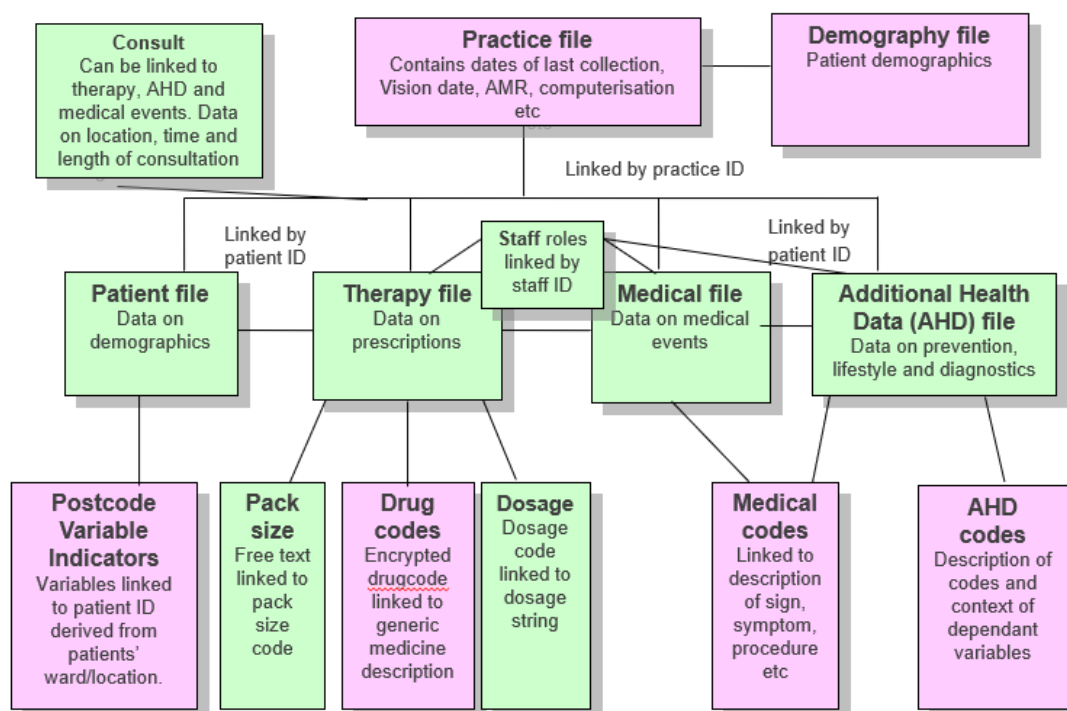


Figure 3.1: Structure and links between THIN databases

- Each row of the Patient file contains one patient with the year of birth, gender, start and end of observation, death date (if applicable),etc.
- Each row of the Therapy file contains one prescription of one prescribed drug, and quantity, with date of prescription and patient ID.
- Each row of the Medical file contains one medical event with event date and patient ID.
- Each row of the AHD file contains one lifestyle information (eg smoking status) or one measurement (eg BMI) with date and patient ID.

Preprocessing of the database consisted in transforming those raw data information from the different databases into a single longitudinal database. The final database contains for each row one patient with in columns all variable of interest, such as gender, year of birth, beginning and end of observation, death status, age at death (if applicable), event date of comorbidities (if applicable), lifestyle variables, measurements metrics and socio-economic variable.

Because we are interested in estimating the residual life expectancy of applicants given their health status at age 65, we construct the variables using information of the patients reflecting their records around that age. Therefore we use all the medical information recorded in order to estimate the medical health of the patient at age 65. For example, binary comorbidity variables, reflecting the event of comorbidity prior to age 65, were constructed using event dates information. In case of missing records for variables as smoking status, blood pressure, BMI, etc. simple imputation techniques were applied, imputing the mean values of the population or default values (eg non-smoker). Because only year of birth of patients was known (for anonymization purpose), we imputed all dates of birth as 1st of July of their year of birth.

Another limit of the database is that postcode variable indicators are not recorded continuously and therefore we will have to make the hypothesis that patients did not change postcodes. This can be a reasonable assumption for retired people after 65.

We show here the first few rows and columns of the constructed database:

Patient ID	Train	Follow-up start	Follow-up end	Date of birth	Death status	Death date	Gender	Mosaic type	smoker_status_uw	bin0_diabetes	diabetes_type	...
g778501mq	1	05/10/1995	19/10/2019	01/07/1950	0		F	33	SM	0	NONE	...
i995204T?	1	15/07/2005	05/07/2018	01/07/1943	1	05/07/2018	F	23	NS	0	NONE	...
d9868014F	1	18/02/2001	03/06/2020	01/07/1952	0		G	2	NS	0	TYPE1	...
c667400Ww	0	10/12/1998	02/10/2015	01/07/1936	0		F	64	EX	0	NONE	...

Figure 3.2: First columns and rows of the constructed patient-level database

We provide here the list of constructed columns, with the variable type or modalities if applicable. We provide next the list of the comorbidities columns in a separate table for convenience purpose. As said previously, all the comorbidity variables were estimated at age 65 of the patient.

Variable	Variable description	Value	Value description
ID	patient ID	Character	
train	train/test split	Binary variable	
followup_start	date of patient entry	Date	
followup_end	date of patient exit	Date	
dob	date of birth	Date	
death_status	death status	Binary variable	
death_date	death date (if applicable)	Date	
gender	gender	F,M	Female, Male
mosaic_type	socio-economic variable from Experian 2009 classification	68 classes of mosaics	See next Appendix
smoker_status	smoking status	SM, EX, NS	Smoker, Ex-smoker, Non-smoker
smoke_int_band	smoking intensity	Continuous variable	
quidur_band	duration since stopped smoking	Continuous variable	
alcohol_status	alcohol status	DR,NO	Drinker, Non-drinker
alc_int_band	alcohol consumption frequency	Continuous variable	
bmi	body mass index	Continuous variable	
sbp	systolic blood pressure	Continuous variable	
FEV1FVC	FEV1/FVC ratio, also called Tiffeneau-Pinelli index	Continuous variable	
HBA1C	glycated hemoglobin	Continuous variable	
diabetes_type	type of diabetes	None, Type1, Type2	
insulin	insulin prescription	Binary variable	
strokegroup	stroke classification (if applicable)	Group 1,2,3 or 4 (if applicable)	The higher the riskier

Figure 3.3: List of variables (except binary variable for comorbidities)

We note that we added a train/test variable randomly generated and splitting the dataset in 70% of train and 30% of test.

Below we provide the list of binary variables for comorbidities and some statistics: the number of cases and the corresponding percentage of patients.

Comorbidity variable	Number of positive cases (N = 1,050,693)	Percentage of positive cases	Comorbidity name
af	13944	1,3%	Atrial Fibrillation
angina	30318	2,9%	Angina
anydement	2043	0,2%	Any form of dementia
aortic	2040	0,2%	Aortic Aneurysm
asthma	89160	8,5%	Asthma
bronchiectasis	5463	0,5%	Bronchiectasis
ckd_d	20670	2,0%	CKD (diabetic definition)
ckd_nd	13806	1,3%	CKD (non-diabetic definition)
copd	31845	3,0%	Chronic Obstructive Pulmonary Disease (COPD)
crohns	2922	0,3%	Crohns Disease
cva	14676	1,4%	Cardiovascular Accident (Stroke) First event
diabetes	89403	8,5%	Diabetes
epilepsy	13980	1,3%	Epilepsy
heartvalve	9549	0,9%	Heart valve disorders
hf	7722	0,7%	Heart failure
hypertension	296652	28,0%	Hypertension
mi	29211	2,8%	Myocardial infarction (heart attack) first event
ms	3300	0,3%	Multiple Sclerosis
neuropathy	5040	0,5%	Neuropathy
othercerebro	7623	0,7%	Other cerebrovascular conditions
otherCv	28392	2,7%	Other cardiovascular conditions
otherimpairment	24117	2,3%	Other impairments
parkinson	2283	0,2%	Parkinson's disease
pvd	2304	0,2%	Peripheral Vascular Disease
ra	12114	1,2%	Rheumatoid Arthritis
retinopathy	16068	1,5%	Retinopathy
stroke	28389	2,7%	Stroke
tia	12828	1,2%	Transient Ischameic Attack

Figure 3.4: List of binary variables for comorbidities and statistics

3.3 Descriptive statistics

We have approximately 1 million patients in our final database. We provide here some statistics on our database.

- Number of patients : 1,050,693
- Number of columns : 129
- Number of medical variables : 59
- Number of deaths observed : 141,756

Smoking variables

We are provided with 3 variables related to smoking habits: smoking status, smoking intensity (applicable for smokers and ex-smokers) and smoking quit duration (applicable for ex-smokers).

Smoking status has 3 modalities :

- NS : Non-smoker representing 43% of patients
- EX : Ex-smoker representing 42% of patients
- SM : Smoker representing 15% of patients

Those statistics are coherent with the number published by the Office for National Statistics of the UK [34].

Smoking intensity variable informs of the smoking habits consumption for smokers or ex-habits for ex-smokers. Units is in cigarettes units per day. We provide distribution of the variable per group:

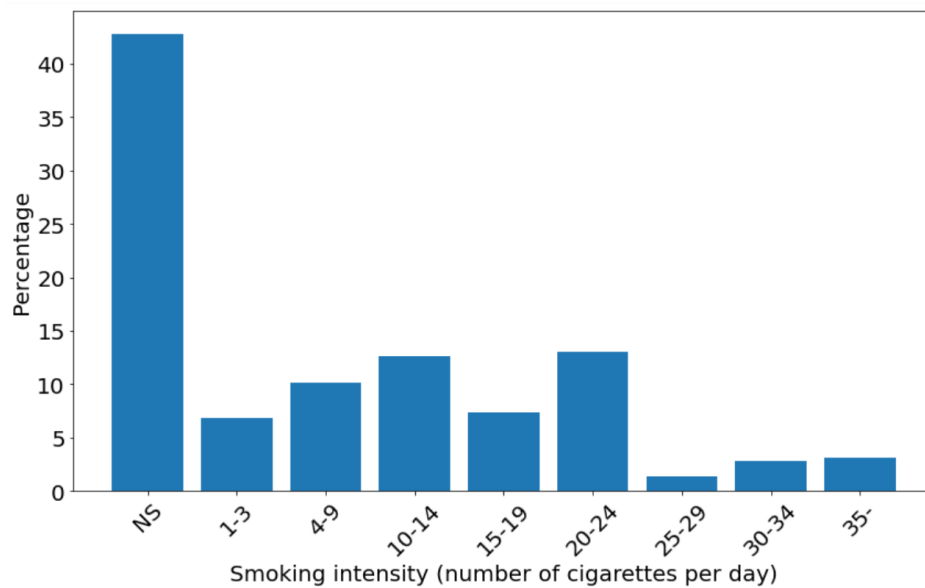


Figure 3.5: Smoker intensity modality frequency

For ex-smokers, we provide distribution of duration (in years) since smoke quitting:

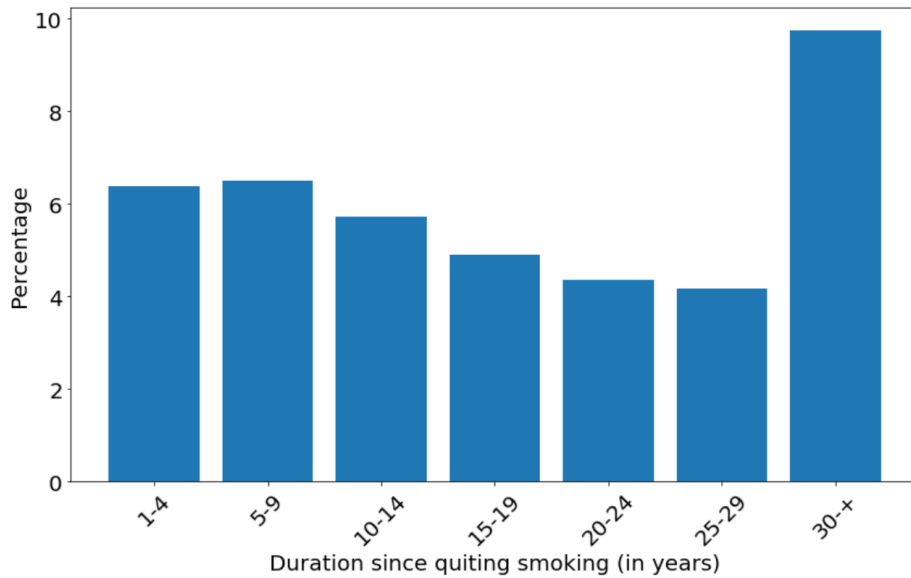


Figure 3.6: Smoker quit duration modality frequency

Body Mass Index

Body Mass Index is a quantity that allows to estimate the corpulence of a person. It is calculated based on height and body weight. Distribution of the BMI in our database is the following:

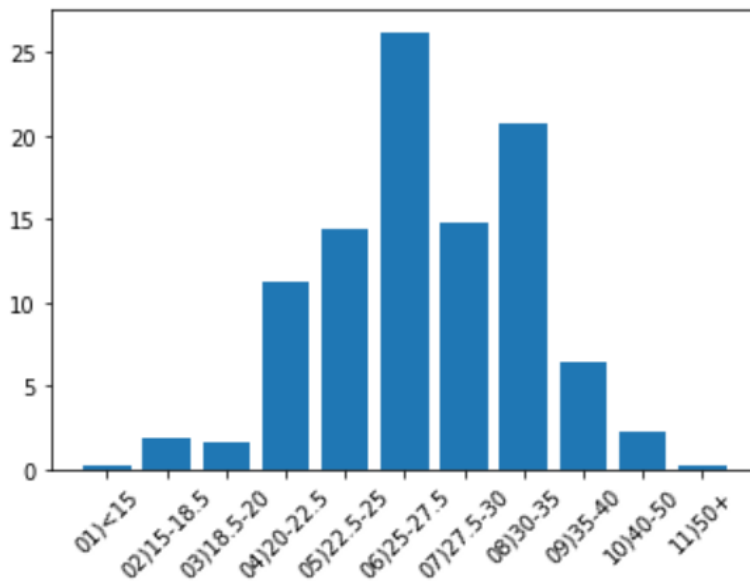


Figure 3.7: BMI modality frequency

This distribution is coherent with UK statistics [10].

Diabetes

For patients having diabetes, we provide statistics on modalities frequency of diabetes type and prescription of insulin treatments. There are 2 main forms of diabetes, type 1 diabetes and type 2 diabetes, both characterized by chronic hyperglycemia. Type

1 diabetes occurs in young people and often begins in childhood. It is caused by an autoimmune destruction of the pancreas which no longer produces insulin. The cause is poorly known and there is currently no possible prevention. Affected people are therefore dependent on insulin, which must be administered by injection. As for type 2 diabetes, which accounts for 90% of diabetes cases, it occurs later in life. It is mainly due to a state of insulin resistance and is associated with overweight. Insulin injection for type 2 diabetes is not always necessary. In our data set, 13.2% of patients with type 2 diabetes have an insulin prescription. Those statistics are coherent with the general UK population [5].

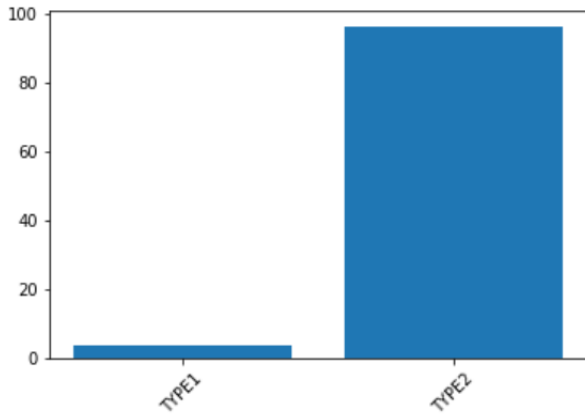


Figure 3.8: Diabetes type frequency

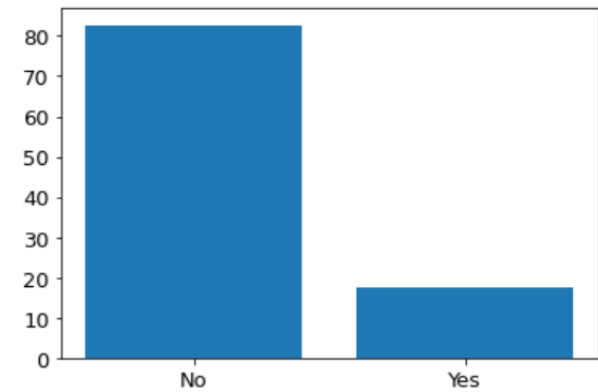


Figure 3.9: Insulin frequency

Stroke

For patients having had a stroke, we are provided with a group stroke variable informing on the severity of the event. The higher the group the more severe the event was.

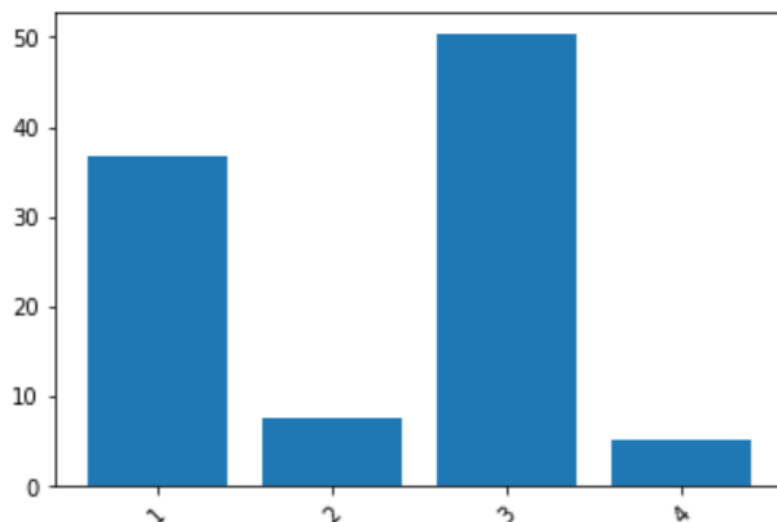


Figure 3.10: Stroke group frequency

Distribution of alcohol and blood pressure variables are provided in Appendix 8.2.

Correlation

In the chart below, we provide visual representation of the correlation within our explanatory variables. Correlation can give insights in the structure and relationship existing in our variables. It assesses the linear relation between two variables. If positive, the two variables tend to increase together, if negative then one tend to increase when the other decreases. A correlation close to zero means that there is no linear relation.

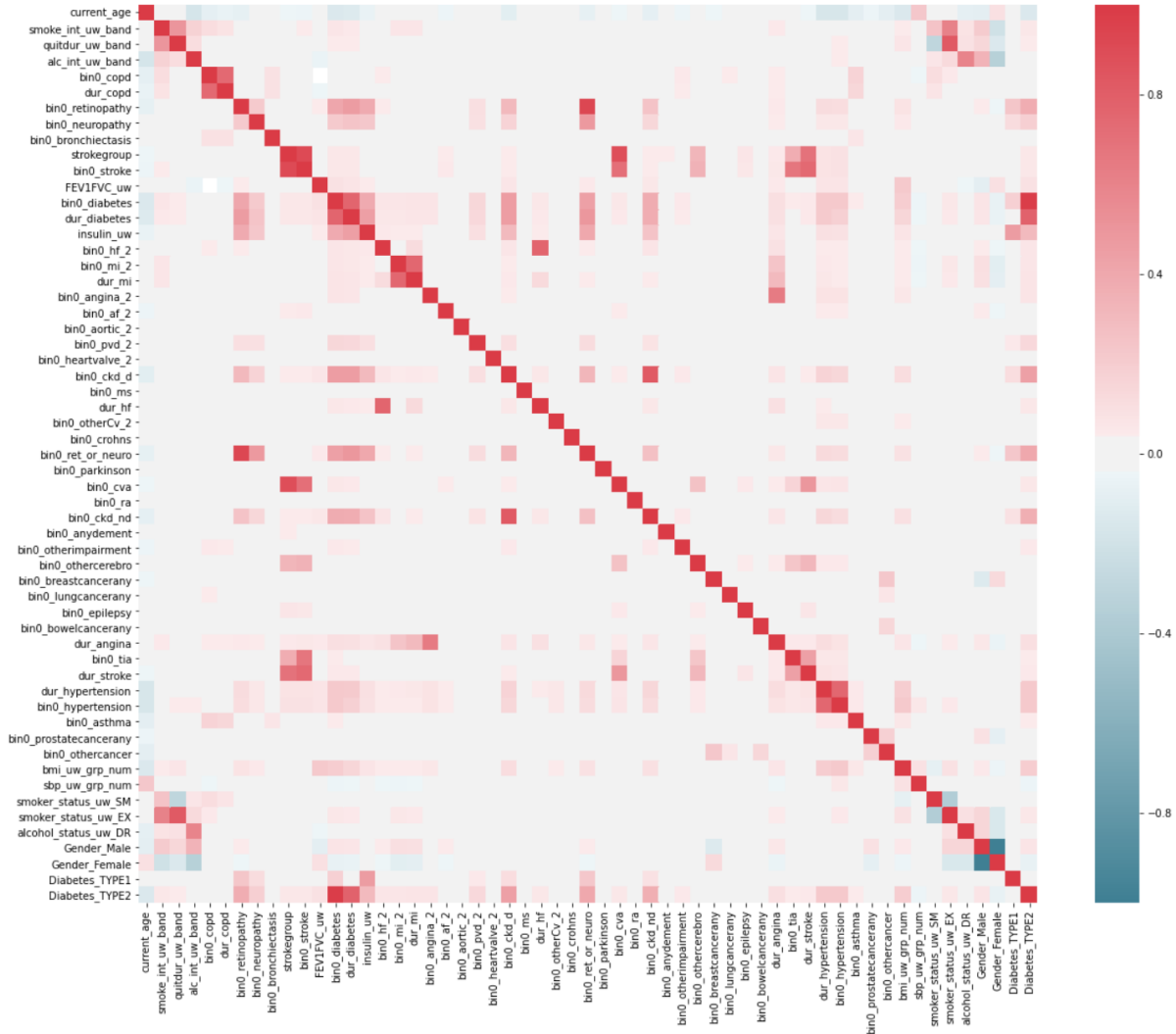


Figure 3.11: Correlation within explanatory variables

We can see that we have some correlated variables, which can be explained by two main reasons:

First, different level of precision of the same information: for example, several variables refer to diabetes: *bin0_diabetes* informs whether the patient has diabetes, while variable *dur_diabetes* inform the duration since diabetes diagnosis and *insulin* informs whether the patient is prescribed insulin. Similar explanation applied to stroke related variables: *TIA*, *CVA* and *stroke group*.

Second, one comorbidity may be caused by another, therefore we find significant correlation between the two variables. For example, some cases of retinopathy (damage

to the retina of the eyes) may be caused by diabetes. This is why some correlation between those two variables appear. We also see that *Diabetes_TYPE2* is correlated by numerous variables such as *BMI*, *hypertension* and *stroke*. There is also a causal relationship between those variables, as a high BMI may cause type 2 diabetes and hypertension. Another example is that hypertension and diabetes are risk factors for strokes, as shown by Chen al. in this paper [11]. Similar comments can be made about smoker status and COPD (Chronic Obstructive Pulmonary Disease) variables.

One should keep in mind these strong correlations when calibrating a model and when analyzing a model output, such as variable importance and significance (p-values). Indeed, correlation can add instability when estimating the parameters of the model (GLM for example). There are different ways to deal with correlated variables. When variables are correlated due to different levels of precision of the same information, one can choose to retain only the most relevant variable using regularization. Indeed, regularization penalizes the size of parameters fitted, and therefore can discard some redundant variables. In one of the following chapters, we will explain and apply the LASSO penalization when fitting a GLM. As we said, variables may also be correlated because there is a causal relation of one on the other, such as high BMI may cause on type 2 diabetes. Or, if there is a causal relation with a third variable, for instance a part of hypertension and type 2 diabetes correlation can be due to the fact that high BMI engender those comorbidities. For these cases, interaction terms between those variables can be considered to account for aggravating risk when formulating the model. Interaction terms can enable to consider the cumulative impact of several variables, as we will do during our study.

Chapter 4

Estimation of raw death rates

The goal of our study is to estimate the mortality level used for life expectancy and annuity pricing calculations. This study focuses on **death rates modeling**. We will use and challenge models that predict mortality rates base on explanatory variables. In this chapter, we focus on the theoretical framework and the way we deal with censorship and truncation. In the following chapters, we will apply different models to estimate mortality and construct life tables.

4.1 Methodology

4.1.1 Lexis diagram

We rely on observed mortality to estimate the mortality rates. Usually, we observe a population during a certain period of time, which we will call study period. We represent those observations using the Lexis diagram.

A Lexis diagram is a two-dimensional diagram used to represent follow-up time and events, in our case death events. Calendar time is represented on the horizontal axis, while age is represented on the vertical axis.

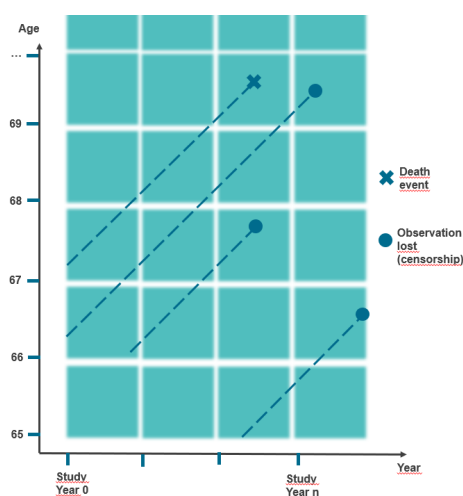


Figure 4.1: Lexis diagram of death events of observed population

This representation helps to visualize the path of each observation. Each line represents

the observation of an individual during the study period. The beginning of the line is the starting observation date, and the end of the line corresponds to the end of observation date. In this way, the length of the line represent the observed survival. The cross symbol represents a death event that causes the end of observation, the dot symbol represents the end of observation caused by any event other than death.

As already stated, we focus only on level risk and put the trend risk aside. Therefore, we make the hypothesis that mortality is stable over time within our study period. Thus, we choose to not take into account the calendar year impact. It corresponds to the actuarial practices, and we believe it is an acceptable approach as we have a limited period of study of only 6 years. Thus, the mortality trend should not bias our results. Graphically, in the Lexis diagram, this is equivalent to consider cells by age only as depicted in the following chart:

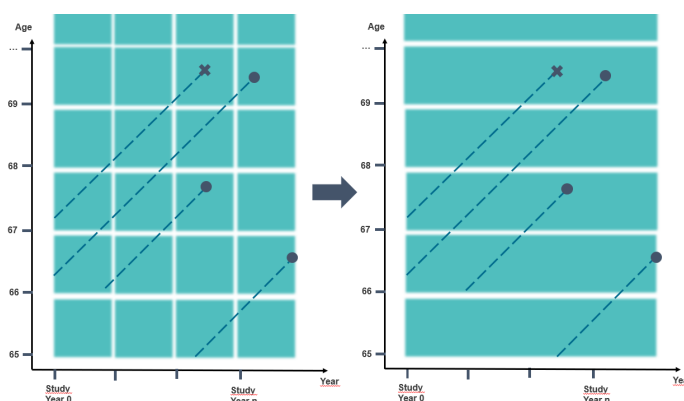


Figure 4.2: Hypothesis of null improvement trend within the study period years

4.1.2 Mortality rate

Theoretical mortality rate q_x

As a reminder, a mortality rate, or death rate, q_x is the probability for someone alive at an exact age x to die during the year. Mathematically,

$$q_x = \mathbb{P}(T \in [x, x + 1) \mid T \geq x)$$

with T the random variable of the age at death.

In this chapter, we show how one can estimate these probabilities from the survival observation. In a second stage, we will construct models that predict these probabilities for an individual with characteristics X :

$$q_x(X) = \mathbb{P}(T \in [x, x + 1) \mid \{T \geq x\} \cap X)$$

Mortality rate estimation \hat{q}_x

An intuitive way is to compute the total number of deaths over the number of persons observed at age x as follows.

$$\hat{q}_x = \frac{d_x}{l_x}$$

with \hat{q}_x the empirical death rate at age x , and d_x the number of death observed at age x and l_x the total number of observation. However, this approach is biased in case of truncation or censorship. Therefore, we will use what we call *initial exposure* E_x to replace l_x to correct for this bias. The exposure is defined in the next section.

4.2 Exposure to risk

4.2.1 Incomplete observations - censorship and truncation

As depicted in the Lexis diagram, we are facing two problems:

- *Truncation* - Let's consider the persons that turn 67 in 2014. Unless all these persons are born on the 1st of January, some people will turn 67 during the year and will then enter the cell midway (increments). Others, will be 67 at the start of observation and will then turn 68 and exit the cell during the year (decrements). Those movements across the cells creates truncation.
- *Censorship* - Some withdrawal can also happen. Withdrawal means that an observed individual may exit from the study for random reasons other than death. Those random decrements induce censorship because we do not know whether the individual died or not at age x .

Therefore, for each individual observation, one of the following three cases may arise:

- Full observation: the individual is observed from age x to age $x + 1$.
- Partial observation: the individual is only partially observed between age x and $x + 1$
- Death: the individual dies between age x and $x + 1$

In the next section, we introduce **exposure** E_x designed to take into account the observation above. The idea is to calculate individual exposure based on the amount of time each life was exposed to the risk of death at age x .

4.2.2 Initial exposure to risk

Let's consider only the individual observed between age x and $x + 1$. For each one of this observation, denoted i , we have at our disposal:

- $x + a_i$ age at start of observation
- $x + t_i$ age at end of observation
- $d_i = 1$ if a death is observed, 0 otherwise

By definition, we have the following: $0 \leq a_i$, $a_i < t_i$ and $t_i \leq 1$.

Assuming no censorship

Assuming there is no censorship and using actuarial notations, ${}_{1-a_i}q_{x+a_i}$ is the probability to observe a death between x and $x + 1$ for an individual i .

We define D_i , a Bernoulli random variable that takes value 1 with probability ${}_{1-a_i}q_{x+a_i}$.

Assuming there is no censorship, and relying on the law of large number, we have $\sum_i D_i \rightarrow \sum_i E[D_i] = \sum_i {}_{1-a_i}q_{x+a_i}$. Thus, we have:

$$\sum_i d_i = \sum_i {}_{1-a_i}q_{x+a_i} \quad (4.1)$$

Including censorship

While we dealt with the truncation, now we must cover the censorship. Indeed, a part of deaths haven't been observed due to censorship. The idea is to add all these missed deaths in the censored observations. When censorship occurs, we have $t_i < 1$ and $d_i = 0$, the probability to have a death observed during the censored period is: ${}_{1-t_i}q_{x+t_i}$. We correct equation 4.1 by adding to the observed death the expected missed deaths:

$$\sum_i {}_{1-t_i}q_{x+t_i}(1 - d_i) + \sum_i d_i = \sum_i {}_{1-a_i}q_{x+a_i} \quad (4.2)$$

By notation, we have ${}_{1-t_i}q_{x+t_i} = 0$ when $t_i = 1$.

Balducci hypothesis

In order to solve the previous equation, we must make an assumption about the mortality on each age interval $[x, x + 1)$, we consider Balducci's hypothesis [4].

Balducci hypothesis : mortality rate for the partial year is proportional to the rate for the full year

$${}_{1-t}q_{x+t} = (1 - t)q_x, \quad \forall t \in [0, 1]$$

Injecting this in the equation 4.2 allows us to obtain:

$$\sum_i (1 - t_i)q_x(1 - d_i) + \sum_i d_i = \sum_i (1 - a_i)q_x$$

After a few arrangements we have:

$$q_x = \frac{\sum_i d_i}{\sum_i (1 - d_i)(t_i - a_i) + d_i(1 - a_i)} \quad (4.3)$$

This obtained mortality estimator is recommended by the French Institute of Actuaries in this guidance paper [4].

Initial exposure to risk

The equation 4.3 gives us an estimation of mortality rates at age x . The numerator is the total number of deaths observed at age x denoted d_x and the denominator is called initial exposure, and it is denoted E_x .

$$E_x = \sum_i (1 - d_i)(t_i - a_i) + d_i(1 - a_i)$$

In practice, we compute the exposure for each individual at each observed age. We obtain pseudo observations that we will use in our modelling. The computed exposures allow us to get around the censorship and truncation problem and also allows applying classic models.

4.3 Application

To deal with truncation and censorship in our database, we apply the previous methodology and choose to discretize our patient-level observations by age. By computing exposures for each individual at each observed age, we create pseudo observations that correspond to one age of one patient per row.

Transformation from patient-level rows to pseudo observations

To apply the initial exposure approach, we have to discretize our patient level database presented in the previous chapter to pseudo observation (as explained in the Initial exposure to risk section).

We specify that we will keep the train/test binary variable we randomly imputed to each patient (in order to have a 70%/30% train/test division). For incoming model, this will prevent us from having pseudo-observations of the same patient that falls in the train and test pseudo observation data sets.

We show here the two first patient rows transformed into pseudo observation. Each row contains the observation of one patient at one age. Death status is set to 1 only at age of death (if applicable).

Patient ID	Train	Date of birth	Follow-up start	Follow-up end	Age	Death status	Death date	Gender	Mosaic type	smoker_status_uw	...	initial_exposure
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	63	0		F	33	SM	...	0.5
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	64	0		F	33	SM	...	1
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	65	0		F	33	SM	...	1
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	66	0		F	33	SM	...	1
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	67	0		F	33	SM	...	1
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	68	0		F	33	SM	...	1
g778501mq	1	01/07/1950	05/10/1995	19/10/2019	69	0		F	33	SM	...	0.21
i995204T?	1	01/07/1943	15/07/2005	05/07/2018	70	0	05/07/2018	F	23	NS	...	0.5
i995204T?	1	01/07/1943	15/07/2005	05/07/2018	71	0	05/07/2018	F	23	NS	...	1
i995204T?	1	01/07/1943	15/07/2005	05/07/2018	72	0	05/07/2018	F	23	NS	...	1
i995204T?	1	01/07/1943	15/07/2005	05/07/2018	73	0	05/07/2018	F	23	NS	...	1
i995204T?	1	01/07/1943	15/07/2005	05/07/2018	74	0	05/07/2018	F	23	NS	...	1
i995204T?	1	01/07/1943	15/07/2005	05/07/2018	75	1	05/07/2018	F	23	NS	...	1

Figure 4.3: First two patients transformed to pseudo observation to apply exposure to risk approach

As we can see, transformation into pseudo-observation has a cost in terms of the increase in the number of rows. Our produced pseudo-observation database contains therefore 4,708,748 rows.

Computing raw mortality rates

After having transformed our database to pseudo observations, we use the package *scor_survival* in Python to draw raw mortality rates and exposure:

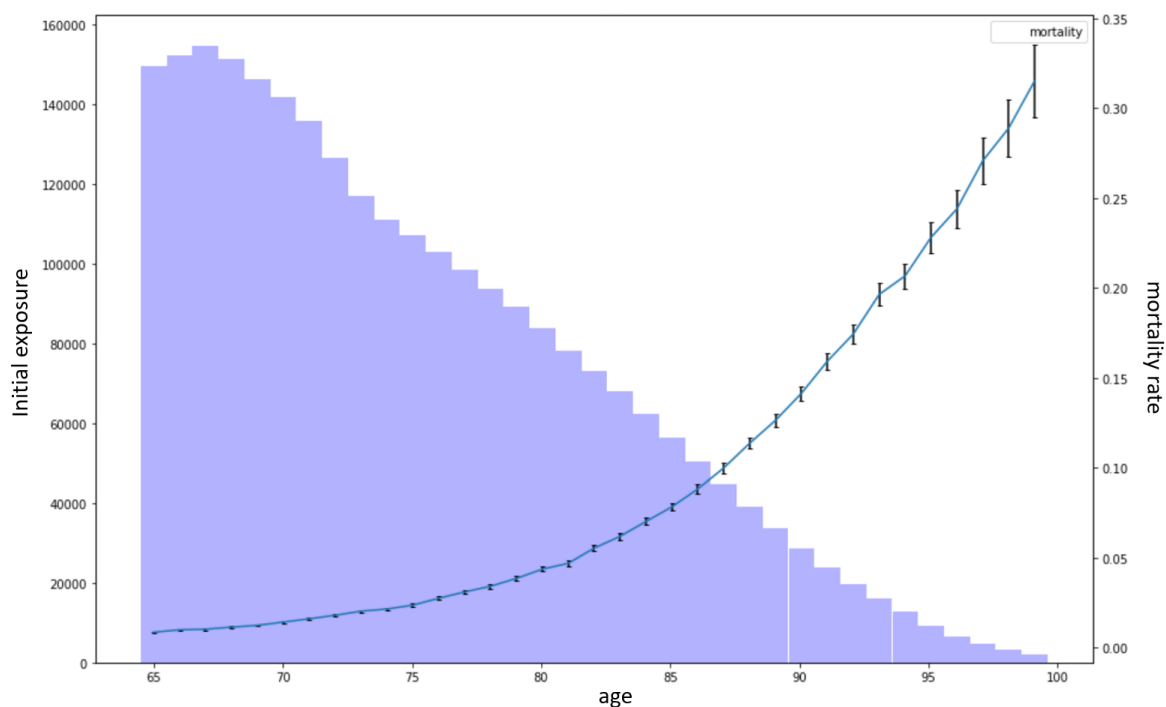


Figure 4.4: Raw death rates and initial exposure by age

The histogram shows initial exposure per age, and the curve gives the raw death rates per age. Those raw mortality rates are coherent with the UK mortality rates available in the Human Mortality Database.

Chapter 5

Construction of reference life tables by age, gender and socio-economic group

Before constructing models for enhanced life annuities product, we first construct life tables that can be used for classic annuities pricing and valuation. For these traditional products only limited information are available such as age, gender and socio-economic class (based on postcode address) of the annuity buyer. Therefore, we construct life tables that reflect the mortality according to these variables. In addition, due to operational limitations, only a limited number of life tables can be used in practice. Considering a life table per gender and mosaic socio-economic classes will lead to a too large number of tables. We will provide an approach to aggregate the socio-economic classes in a relevant way with minimum mortality prediction power loss.

In the first section, we construct life tables per gender, allowing us to introduce the overall mortality modelling approach. To achieve this construction, we consider GLM and GAM models. We provide a comparison between the performance of the two models. Afterwards, we construct life tables using the mosaic socio-economic variable as well.

In the next chapters, we will deal with enhanced annuities, we will use the mortality rates constructed in this chapter as reference tables in our mortality models. The idea is to rely on these reference mortality curves to construct more complex models that encompass comorbidity variables.

5.1 Life tables by age and gender

We apply a parametric model to construct the life tables from the crude mortality rates obtained in the previous chapter. One can notice that the raw mortality rate curve, plotted in Figure 4.4, is relatively smooth, thus one could use them directly. Our goal here is to introduce the modelling approach that would take on its full meaning when introducing the mosaic socio-economic groups in the next section.

Therefore, as a first step, we here construct life tables per gender in order to compare the performance of GLM and GAM models.

5.1.1 Methodology

Binomial distribution assumption

To apply parametric models, we must specify a distribution for the number of deaths.

To fix the ideas, let's consider a group of observations sharing the same characteristics X . Let l_x denote the number of observations at age x and D_x the random variable capturing the number of deaths occurred at this age.

A binomial model appears to be the natural approach, as a binomial distribution is a set of l_x *independent* Bernoulli trials, with for each trial predicts a death (1) or a survival (0) based on the probability of death $q_x(X)$. With $q_x(X)$ the death probability at age x for individual with features X .

However, as described in the previous chapter, we cannot use directly the number of observations l_x due to truncation and censorship in our survival observations. Therefore, as proposed by Delwarde and Denuit in [12], we consider instead the estimated initial exposure E_x . Indeed, the exposure E_x can be seen as the number of observations adjusted to deal with censorship and truncation.

We obtain the following mathematical representation:

$$D_x \sim \mathcal{B}(E_x, q_x(X)) \quad (5.1)$$

Estimation using GLM model

Now, let's consider a use General Linear Model to explain and predict mortality. As proposed by Delwarde and Denuit[12] we consider:

$$D_x \sim \mathcal{B}(E_x, \text{logit}^{-1}(\beta X))$$

with $\text{logit} : x \mapsto \log(x/(1-x))$ the canonical link function, $\beta = (\beta_0, \dots, \beta_p)$ the model parameter and the features $X = (1, X_1, \dots, X_n)$.

We are interested in the univariate response variable $Y = \text{logit}^{-1}(\beta X)$, it is $q_x(X)$, i.e. the mortality at age x for an individual with characteristics X . The expected value of Y of the predictor variables is given by:

$$\text{logit}(\mathbb{E}[Y | X]) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

The parameters β can be estimated by solving the maximum likelihood of the model based on the death observations. We can rely on **R** package *stats* to solve this optimization problem. The algorithm maximises the Binomial log-likelihood :

$$\begin{aligned} \log(\mathcal{L}) &= \log\left(\prod_{i=1}^n \mathbb{P}(D_i = d_i)\right) \\ &= \log\left(\prod_{i=1}^n \binom{E_i}{d_i} q(X_i)^{d_i} (1 - q(X_i))^{E_i - d_i}\right) \\ &= \sum_{i=1}^n d_i \log(q(X_i)) + (E_i - d_i) \log(1 - q(X_i)) + \text{constant} \end{aligned}$$

We note $(X_\kappa)_\kappa$ the list of all possible inputs X . We set $d_\kappa = \sum_{i, X_i=X_\kappa} d_i$ and $E_\kappa = \sum_{i, X_i=X_\kappa} E_i$ in order to group observations with identical covariates X_κ . Maximizing the previous equation is equivalent to maximize:

$$\begin{aligned} \log(\mathcal{L}) &= \sum_{\kappa} d_\kappa \log(q(X_\kappa)) + (E_\kappa - d_\kappa) \log(1 - q(X_\kappa)) \\ &= \sum_{\kappa} \left[\frac{d_\kappa}{E_\kappa} \log(q(X_\kappa)) + \left(1 - \frac{d_\kappa}{E_\kappa}\right) \log(1 - q(X_\kappa)) \right] E_\kappa \end{aligned}$$

From previous equation, we notice that solving optimization of model (5.1) is equivalent to solving the model predicting as target variable $\frac{d_\kappa}{E_\kappa}$ with weight variable E_κ .

As a results, in practice we aggregate our pseudo-observation database and use R package *stats* to fit a binomial model with target variable the raw mortality rate of each group (d_κ/E_κ) , and we put as weight parameter the sum of exposure of the group (E_κ) .

For our application, we consider only the Age and Gender. We have $X_1 = 1$ if *Female*, $X_1 = 0$ otherwise, plus for age $X_2 = x$. We allow for intercept and an interaction between age and gender, therefore we have the following relationship:

$$\text{logit}(q_x(X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2 + \epsilon$$

Estimation using GAM model

The main benefit of General Additive models (GAM) is to relax the linear relationship between the explanatory variables and the variable of interest. Thus, these models are better to capture the non linear effects, such as the effect of age on mortality, in our problem.

Indeed, GAMs are an adaptation of GLMs that allows to model a non-linear impact as it gets rid of the restriction that the relationship must be a weighted sum, and instead assumes that the outcome can be modelled by a sum of arbitrary functions (parametric or not parametric) of each feature.

As for GLMs, a GAM relates a univariate response variable, Y , to some predictor variables X and an exponential family distribution is specified for Y . As in the previous subsection, we consider the binomial distribution along with a canonical link function logit . The form of the GAM is the following:

$$Y = \text{logit}^{-1}(\beta_0 + f_1(X_1) + \dots + f_m(X_m)) + \epsilon$$

with β_0 the intercept and f_i some function to determine.

The expected value of Y to the predictor variables is given by:

$$\text{logit}(\mathbb{E}[Y | X]) = \beta_0 + f_1(X_1) + \dots + f_m(X_m)$$

The functions f_i may be functions with a specified parametric form (for example a polynomial, or an un-penalized regression spline of a variable) or may be specified non-parametrically, or semi-parametrically. This flexibility can relax assumptions on the actual relationship between response and predictor and provide better fits to data than purely parametric models. However, a drawback of GAMs is the loss of interpretability

because we are not provided with a single β parameter. In our study, we will use splines that combine several advantages. Details on B-splines and their fitting can be found in Appendix 8.4.

In the same way that we fitted different coefficients per gender in the GLM, we fit an intercept and a spline for each gender, in order to allow to capture different impact of age per gender.

5.1.2 Results

We display here analysis of the fitted models.

Fitted logit-mortality rates

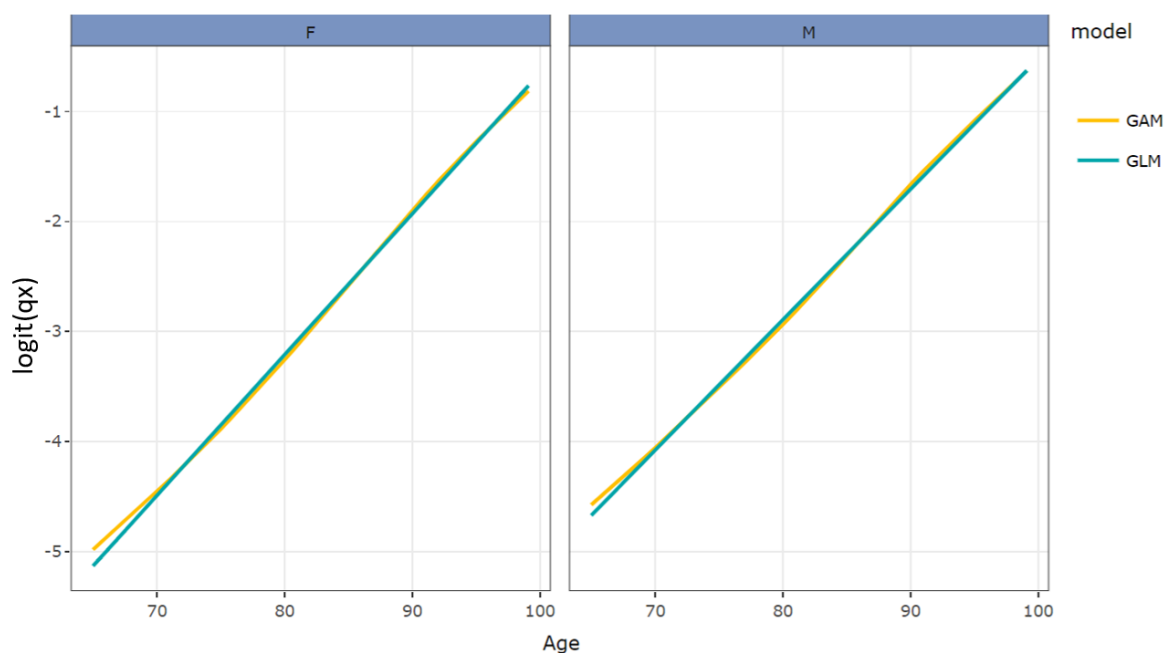


Figure 5.1: Mortality rate per age and gender for GLM and GAM (logit scale)

The GLM curve is a straight line, while the GAM curve is slightly curved. We provide in Appendix 8.4 an analysis of the fitted splines.

Standardized Mortality Ratio

The standardized mortality ratio (SMR) is the ratio of the number of observed deaths over the number of deaths that would be expected if the study population had the same age-specific rates as the standard population.

For each age band and gender, we compute SMR the following way:

$$SMR = \frac{\text{Actual deaths}}{\text{Expected deaths}}$$

The expected deaths for an age-band and gender are obtained by simply summing the exposure multiplied with predicted mortality rates of the model.

If the model captures correctly the mortality, we would expect an SMR close to 1. We computed the SMR on the **test dataset** and show the results in the following chart:

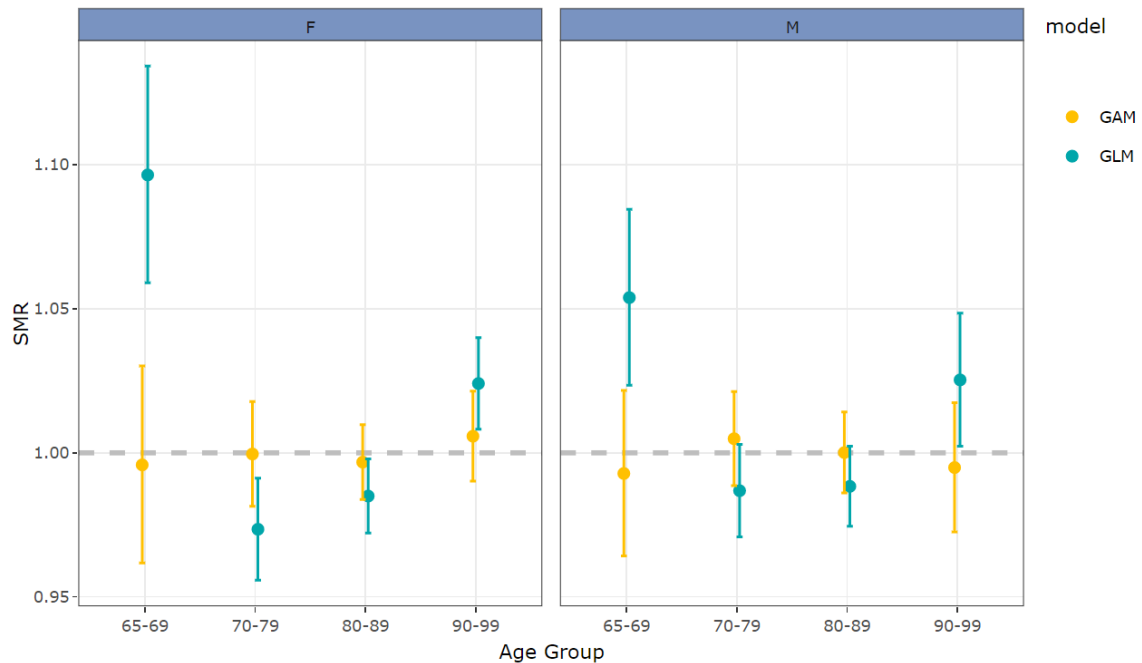


Figure 5.2: SMR per Gender and Age band obtained for GLM and GAM

As we can see, the GAM has captured correctly the impact of age for each gender. Indeed, confidence interval ranges of the SMR include 1. However, we see that the GLM underestimates the mortality between 65-69 years old, especially for women, and between 90-99. It also overestimates the mortality between 70-79 years old, especially for women.

This analysis shows that a GAM model is more appropriate than a GLM to capture the shape of mortality with age. In the next section, we will add mosaic socio-economic information to construct life tables, using GAM models.

5.2 Life table by gender and socio-economic clusters

Now that we have fitted our reference model on age and gender, we want to use an additional explanatory variable to explain mortality. In this section, we present how we integrated a socio-economic variable to explain mortality and obtain our final reference life tables.

5.2.1 Mosaic : UK indicator of socio-economic group

Mosaic is an indicator used in the UK for geo-demographic classification. This classification is done by Experian, a consumer credit reporting company. For each postcode, the company provides a classification, giving information on the socio-economic class the postcode resident is more likely to belong to. Experian usually sells their classification to companies that use them for retail purposes. Having the socio-economic information of a postcode allow them to be very specific with the content and tone of their campaigns and better fit their target group's preferences.

Different classification exists. Our data provided the type 2009 classification, which is composed of 68 categories. We provide the mosaic classification meanings in Appendix 8.7.

5.2.2 Descriptive statistics

Mosaic information is provided for 90% of the exposure of our database. We assume that a patient with no mosaic information available is independent of their mosaic type, i.e. observations with missing mosaic are proportionally distributed over all classes of mosaics.

In the chart below, we provide total exposure (left axis) and raw mortality rate (right axis) of each mosaic class.

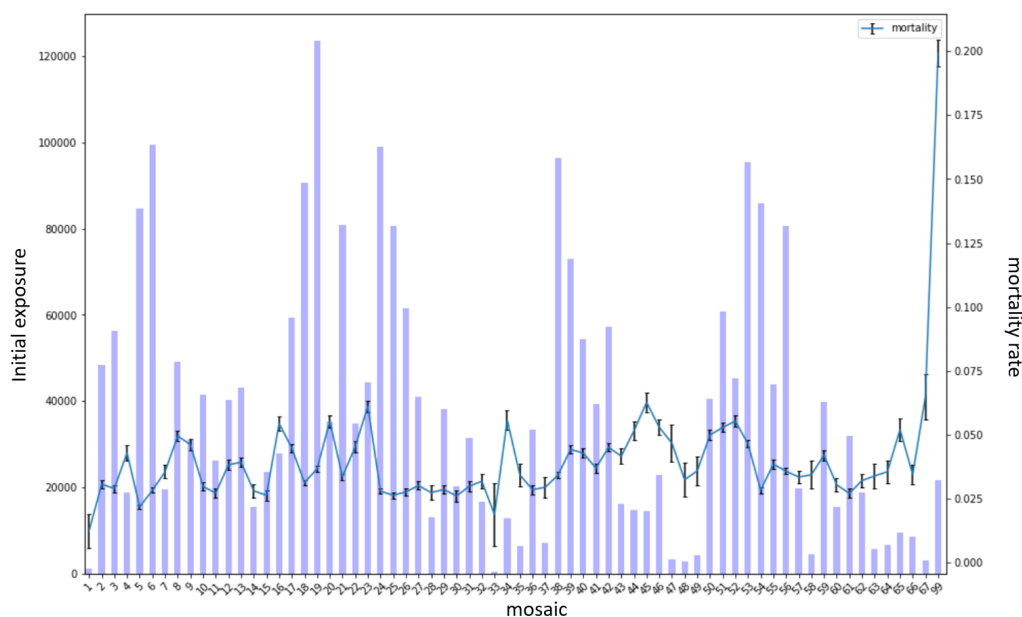


Figure 5.3: Distribution of exposure and mortality rate across the mosaic types

We notice that mosaic number 99 has a very high mortality rate. This is explained by the fact that mosaic 99 regroups nursing homes. We can also see that mosaics 1 and 33 have very low exposures.

The difference in death rates are mainly explained by the differences in age and gender distribution of the mosaic populations. However, this is not the only effect. Socio-economic factor have an important impact on mortality rate.

To capture the impact of mosaics controlling by age and gender, we will use the previously fitted GAM model.

5.2.3 Impact of mosaic type on mortality

SMR by mosaic type of fitted GAM model on age and gender

To visualize the impact of socio-economic mosaic type controlled by age and gender, we compute the SMR per mosaic using the prediction of the previous GAM model fitted on age and gender.

The Standardized Mortality Ratio of mosaic i is computed as follows:

$$SMR_i = \frac{d_i}{\delta_i^{GAM}}$$

with d_i the number of deaths of mosaic i and δ_i^{GAM} the predicted number of deaths of mosaic i using the GAM model:

$$\delta_i^{GAM} = \sum_{j \in \mathcal{M}_i} E_j q_j^{GAM}$$

with j all the pseudo-observations belonging to mosaic i .

The produced plot is the following:

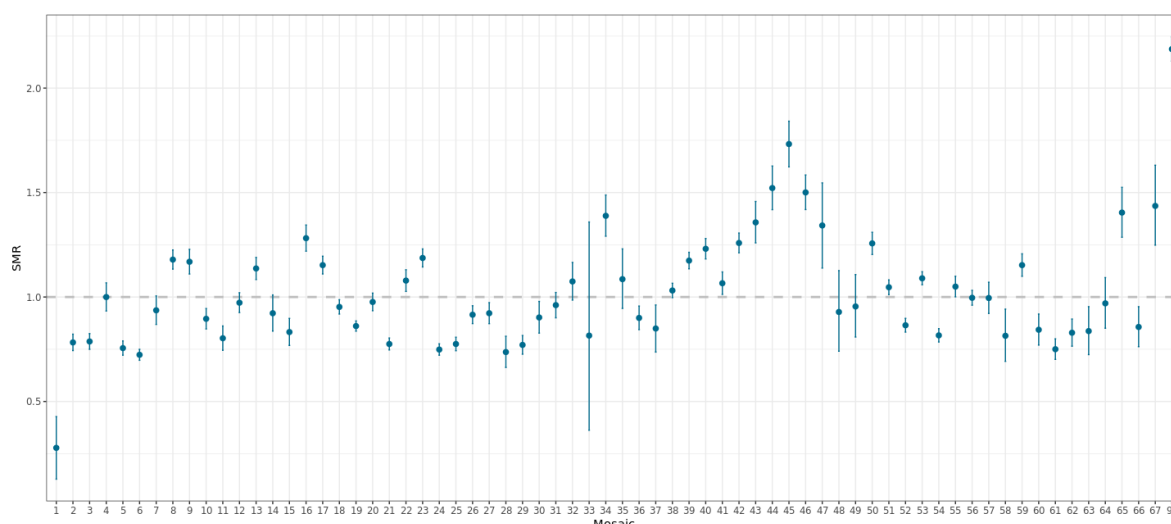


Figure 5.4: SMR per mosaic type with expected based on previously fitted GAM

The visual inspection of the chart indicates that some mosaic class have a significant lower or higher mortality than the one captured by the GAM model. As we controlled

for age and gender, these differences are not due to the age and gender. In other words, the mosaic has an impact on mortality. In the next section, we show how to construct life tables by socio-economic group.

Modeling the impact

A straightforward way to capture the additional impact of mosaic type on mortality is to enhance the model introduced previously by adding a parameter β that will capture the mosaic impact on mortality, as follows:

$$\text{logit}(\mathbb{E}[Y]) = \alpha_F + \alpha_M + f_F(\text{Age}) + f_M(\text{Age}) + \sum_i \beta_i \mathbf{1}_{\{\text{Mosaic_type}==i\}}$$

We can also rely on the previously fitted GAM model that provides death rate by age and gender, which we will note q_x^{GAM} . As a reminder, we have:

$$\text{logit}(q_x^{GAM}) = \alpha_F + \alpha_M + f_F(\text{Age}) + f_M(\text{Age})$$

We can keep the previously fitted intercepts and splines for our model with mosaic types:

$$\text{logit}(\mathbb{E}[Y]) = \text{logit}(q_x^{GAM}) + \sum_i \beta_i \mathbf{1}_{\{\text{Mosaic_type}==i\}}$$

Therefore, we fit a binomial GLM model on mosaic indicators using as offset $\text{logit}(q_x^{GAM})$. We specify that we do not want an intercept because it is already included in the q_x^{GAM} .

Thanks to this simple and rapid approach, we obtain a predicted mortality curve for each mosaic. However, it is not the retained approach as explained in the following subsection.

5.2.4 Clustering

For business constraints and better understanding, we want to reduce the number of class of mosaics. Therefore, we will have to cluster mosaic types into a lower number of groups. Several questions arise, such as: How many clusters should we choose? How can we cluster them in a relevant way?

These are not so easy questions, as we do not have at our disposal a way to measure how similar two mosaics are in term of mortality risk. The mosaic number does not provide any indication on mosaic definition proximity. In other words, **the mosaic variable modalities are not ordered**. In addition, because there are 68 mosaics (mosaic variable has 68 modalities), this creates a very large number of possible combinations when considering clustering mosaic modalities. Indeed, let's say that we aim to create 5 groups, then there would be more than 10 millions ways to aggregate the 68 mosaics into 5 clusters ($\binom{68}{5} = 10,424,128$). Therefore it would not be feasible to create clusters through testing all possible combinations. We present next the clustering methodology we adopted to solve our problem.

Clustering methodology

We aim to build a model with high explanatory power but using fewer categories of mosaic types. To do so, we will rely on an agglomerative hierarchical clustering approach. This means that, at the start of the clustering process, we will consider each mosaic type as its own cluster. We will then merge two clusters together at each iteration until having only two clusters left.

We consider a target function that reflects the performance of the model while penalizing on the complexity of the model (here we want to penalize on the number of clusters). Therefore, we consider as target function the Akaike information criterion (AIC). The AIC takes into account the log-likelihood of the model and the number of fitted parameters.

$$\text{AIC} = 2k - 2 \log(L)$$

In our case, L is the likelihood of our binomial model.

This definition of the AIC means that the lower the AIC value the better. Therefore, we will seek to minimize the AIC at each iteration.

The proposed methodology is the following :

1. For each combination of two clusters, we temporarily merge them and then fit the binomial model.
2. We collect AIC of each of the fitted models and merge the pair of clusters with the lower AIC.
3. We then reproduce this procedure using the updated clusters until we have only two remaining clusters.

Application and results

We implemented this methodology, which is computationally intensive. Indeed, for each iteration k , we have to fit a GLM for every combination of 2 clusters in the $68 - k$ left clusters. As there are 67 iterations, we have in total $\sum_{k=0}^{66} \binom{68-k}{2}$ models to fit.

After applying this methodology, we get at each step k a set of clusters, which contains $68 - k$ different classes. For each set of those clusters, we fit a GAM model to construct different splines for each gender and cluster. This allows us to take into account the mortality characteristics of socio-economic group. We store the AIC of the model fitted on each set of clusters and plot them in a graph with on the x -axis the number of classes of the corresponding set of clusters the model was fitted on. We do not plot the results for number of clusters over 30 as y scale becomes too wide.

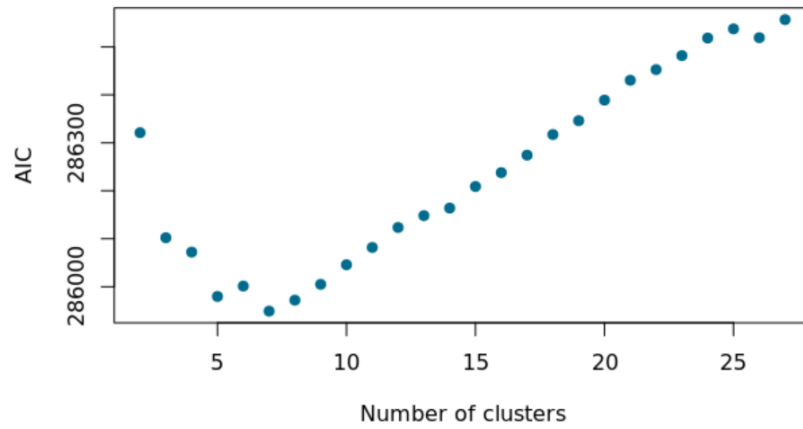


Figure 5.5: AIC of GAM model using gender and mosaic clusters found at each step k

The AIC graph shows that based on this metric, the optimal number of clusters is 7. We show in Appendix 8.8 the final 7 clusters and the name assigned to each of the cluster based on the description of the mosaic types they regroup.

5.2.5 Final life tables by gender and socio-economic cluster

Now that we have a reasonable amount of socio-economic groups, we can construct a set of mortality table by gender for each cluster. The produced mortality curves are represented here. In addition we provide logit-scale of the curves for more readability :

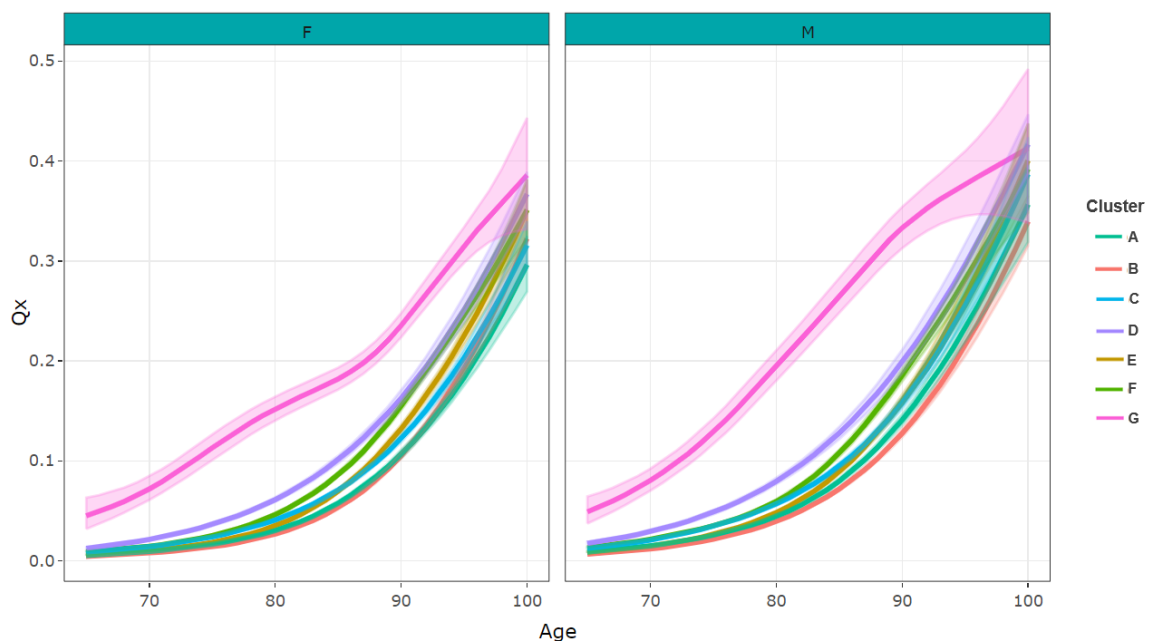


Figure 5.6: Fitted life tables per gender and socio-economic cluster using binomial GAM model

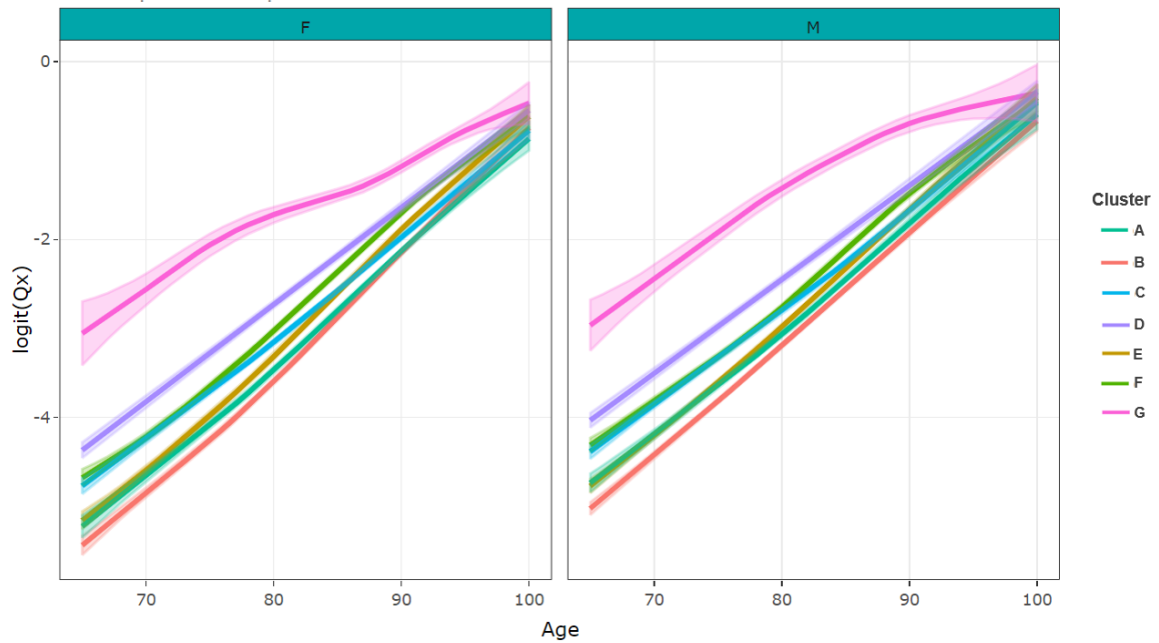


Figure 5.7: Fitted life tables per gender and socio-economic cluster using binomial GAM model with logit scale

The pink curve is the mortality of cluster G composed by only mosaic type 99 which is, as said previously, a specific mosaic type containing nursing homes. This is why we obtain very high mortality rates in this cluster.

We also notice that due to limited data, confidence intervals are larger between ages 95 to 100. To extend mortality rates, we will apply linear interpolation at high ages using UK national life table on ages from 95 to 100. We will explain this step further in the next chapters.

Chapter 6

From standard life tables to enhanced mortality rates using GLM and Machine Learning models

In the previous chapter, we created life tables per age, gender and mosaic/socio-economic cluster. In this chapter, we rely on these tables to reflect the annuitant mortality by using the comorbidity variables.

Additionally, we are facing the business constraints to construct an interpretable model. Indeed, to meet regulation constraints, the insurer must be able to justify the prices applied. As annuity prices depend on model output, to justify variations in prices, one need to understand the variation in model predictions. As an interpretable model, we selected a GLM model that we will introduce in the following section.

Moreover, from the underwriting operational perspective, we also want a model that considers a limited number of variables. Indeed, the higher the number of variables considered for the mortality prediction, the bigger the amount of information that the insurer would need to collect during the application process. We selected the GLM-Lasso model in order to reduce the complexity and the number of variables used in the final GLM model.

Finally, we would like to check if our model is as efficient as a Machine Learning one. Therefore, we will also consider a Gradient Boosting Model (GBM) considered to be among the most powerful models.

This chapter will focus on the theory behind each model and the application and fitting on our dataset. The next chapters will focus on assessing and comparing the performance of each model.

6.1 Modelling enhanced life annuities GLM

6.1.1 Model choice and inclusions of reference rates

Due to the peculiarities of our database and for technical constraints, we will use Poisson model as an approximation of Binomial model in this section. Indeed, incorporation of truncated data in Binomial models is limited while Poisson form allows to take

exposure into account through an offset term. We provide further explanations in Appendix 8.4.

In the previous chapter, using GAM we constructed references rates $q_{x,g,c}^{GAM}$ that vary by age x , g and mosaic cluster c .

Inspired by Cox's proportional hazard model [27], we assume that the mortality rates are:

$$q_x(X) = q_{x,g,c}^{GAM} \exp(X\beta)$$

with X the medical risk factor variables and β the unknown parameters to be fitted.

In this approach, we imply that mortality of individual with characters X is proportional to the references tables, whatever the age. This is a strong assumption well know not to be verified. However, we can extend our model by adding the age (and gender) to the explanatory variables X . This makes sense, as some comorbidity may impact differently the mortality based on age.

Assuming the death numbers D_x is Poisson distributed, one would need to fit the following model:

$$D_x \sim \mathcal{P}(q_{x,g,c}^{GAM} \exp(X\beta) \cdot E_x)$$

This formulation results in explaining death status D_x with link function log and with offset $\log(q_{x,g,c}^{GAM}) + \log(E_x)$.

Alternatively, based on the relational models, introduced by Hannerz [15] and used by Delwarde and Denuit in mortality prediction [12], we could build a model relying on these reference rates as follows:

$$\text{logit}(q_x(X)) = \text{logit}(q_{x,g,c}^{GAM}) + X\beta$$

Assuming that D_x is binomial distributed one would need to fit the following model:

$$D_x \sim \mathcal{B}(E_x, \text{logit}^{-1}(\text{logit}(q_{x,g,c}^{GAM}) + X\beta))$$

This formulation results in explaining death status D with link function logit and with offset $\text{logit}(q_{x,g,c}^{GAM})$.

6.1.2 Classic GLM

As described previously, with GLM, we model a relationship between the target variable Y et covariates X . Our model has the form of a linear combination between the covariates X that predicts Y . As we do not know the exact form of the relationship between our input vector and our target variable we must make an assumption.

We will establish our model based on all variables available, as well as any reasonable non-linear transformation or interaction effects. This transformation of input variables is called *features engineering*. As GLM are not able to automatically capture non-linear or interactions effect this step is essential.

For the feature engineering, we will rely on descriptive statistical analysis and medical knowledge in order to set a model that encompasses all reasonable possible effects. In the next section we will apply a penalization on the fitted coefficients in order to get rid of non-relevant terms.

A lot of our variables are binary and do not require non linear transformation. However, we have some continuous variables such as BMI, SBP, duration since quitting smoking

or intensity of alcohol consumption. We analyse raw mortality rates of BMI and SBP:

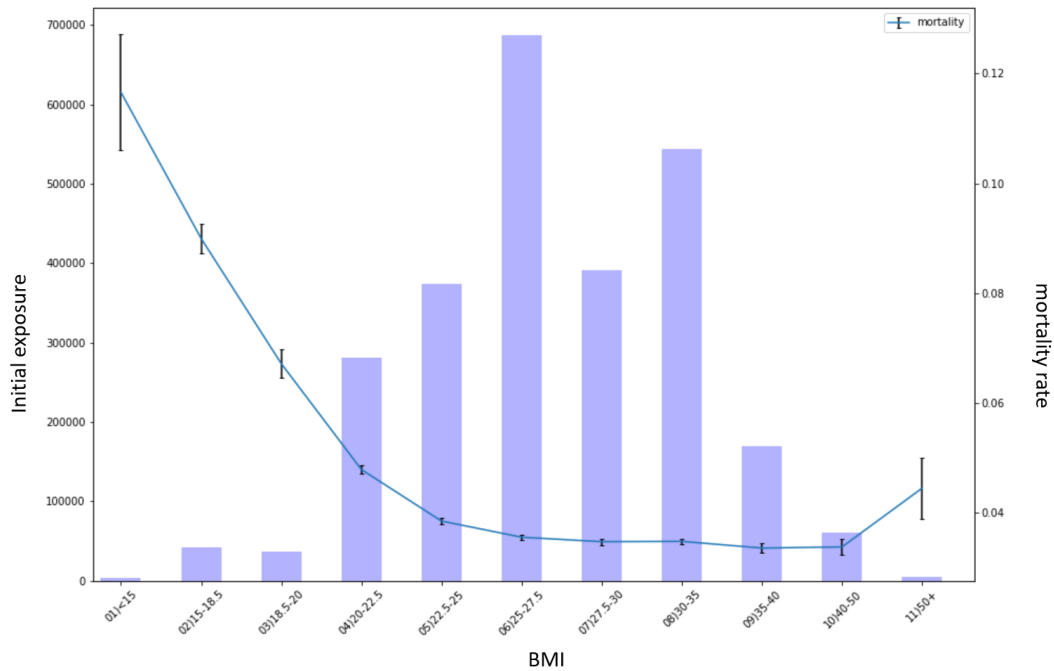


Figure 6.1: Raw mortality rate per BMI groups

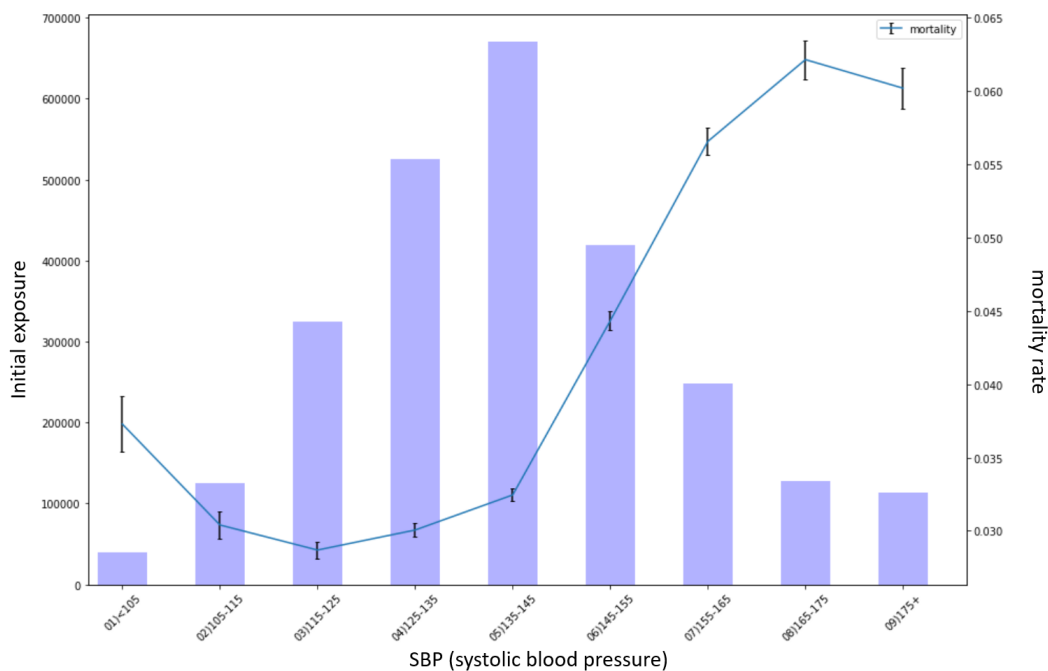


Figure 6.2: Raw mortality rate per SBP groups

The graphs above show a polynomial relationship with mortality rates, and therefore we will fit parameters on quadratic and cubic transformation of those variables. Similar conclusions can be drawn from the analysis of duration since quitting smoking and alcohol consumption variables (see graphs 8.9 8.10 in Appendix). Therefore, we will

add into our model quadratic or cubic effects for BMI by adding the following created variables: BMI, BMI² and BMI³.

We also create interactions of variables which are known to have a compounding effect. We relied on knowledge of SCOR medical experts to draw a list of variables having an aggravating impact. For example, an interaction is considered between the smoking status and blood pressure (SBP), because a causal relationship shows that smoking can increase blood pressure leading to potential malignant cases of hypertension [26]. We underline the fact that some correlated variables we noticed in chapter 3 are here considered as interaction terms, for example smoking status with COPD comorbidity. Also, numerous interactions between age and comorbidity variables are specified. Indeed, we believe that the increase of mortality risk of some comorbidities can vary with the age of the individual. Following the same reasoning, socio-economic clusters were featured with main comorbidity variables (diabetes, COPD and MI ¹), to capture potential differences in mortality risk impact depending on one's social class. Here are some other examples of interactions considered based on medical expertise: COPD and smoking status, stroke and smoking status, insulin and diabetes, retinopathy and diabetes, heart failure and cardiomyopathy, bronchiectasis and COPD, fev1fvc and COPD, etc.

Taking all those effects into consideration, we end up with a complex model that requires 143 coefficients to be fitted. We provide in Appendix 8.4 part of the results, however for confidentiality matters, we randomly deleted some rows.

Because a reference life table is included as base prediction, we chose to put aside the intercept term. Indeed, the average mortality should be already captured by the reference life table, making the intercept less relevant.

6.1.3 GLM-Lasso

Because of the complexity and the high number of fitted terms in our model and the existing correlation between some variables, we apply a GLM-Lasso model, adding a penalization on the L_1 norm of the fitted coefficients. L_1 penalization has the faculty to force some coefficients to zero and therefore reduce the number of explanatory variables.

The penalization is added when solving the maximum likelihood \mathcal{L} of the model :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}[\mathcal{L} + \lambda \sum_{i=1}^p |\beta_i|]$$

with \mathcal{L} the log-likelihood of our model and λ an arbitrary positive value reflecting the degree of penalization and therefore the resulting non-zero parameters.

Because we fit a Poisson model, we have

$$\mathcal{L} = \log\left(\prod_{i=1}^n \frac{\hat{y}_i^{y_i}}{y_i!} e^{-\hat{y}_i}\right)$$

with y_i the true value and \hat{y}_i the predicted value of the model.

¹Myocardial infarction

Usually, λ is chosen by grid-search maximising some target function, in our case we will focus on Poisson deviance. We use the `cv.glmnet` function from *glmnet* R package to select the best λ .

The `cv.glmnet` function operates a K-fold cross validation, meaning the data is divided into K subsets and K models are then trained on K-1 folds and validated on the last fold. Applying cross validation enables to avoid overfitting risk when selecting the best hyperparameters.

For each penalization λ and for each fold k , β is maximised, giving $\hat{\beta}_{\lambda,k}$ and the Poisson deviance is calculated. Residual deviance D is given by

$$D = 2(\mathcal{L}_{sat} - \mathcal{L}_{\lambda,k})$$

with \mathcal{L}_{sat} the log-likelihood of the saturated model, meaning of a model that fits the data perfectly, and with $\mathcal{L}_{\lambda,k}$ the log-likelihood of the fitted model on fold k with penalization λ . Because we use Poisson distribution, this gives :

$$\begin{aligned} D &= 2(\log(\prod_{i=1}^n \frac{y_i^{y_i}}{y_i!} e^{-y_i}) - \log(\prod_{i=1}^n \frac{\hat{y}_i^{y_i}}{y_i!} e^{-\hat{y}_i})) \\ &= 2 \sum_{i=1}^n (y_i \log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)) \end{aligned}$$

The mean Poisson deviance on the validation set of the K fitted models for each λ is represented on the following output:

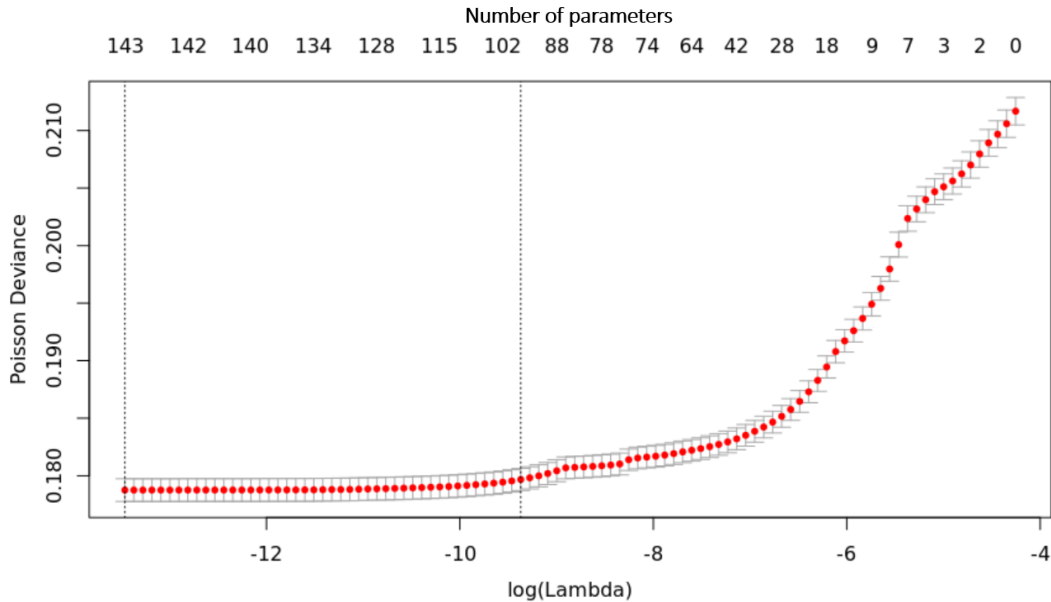


Figure 6.3: Grid search on GLM-Lasso λ parameter: Poisson deviance evolution using cross validation

The vertical lines show the locations of λ_{min} and λ_{1se} , which are two common choices for λ . λ_{min} is the penalization that results in the best out-of-sample performance of the model. λ_{1se} is the largest penalization within 1 standard error of λ_{min} .

We see that in our case, λ_{min} does not reduce the number of coefficients. Therefore we will choose λ_{1se} as penalization. This penalization reduces the number of fitted parameters from 147 to 77.

Also, we see that a $\log(\lambda)$ of -6.5 would enable to reduce the number of parameters to 41, without too much increase in deviance. This is why we will also fit a model with this penalization.

Therefore we fit a first GLM with a penalization of λ_{1se} , which we will call model *GLMNet_long* and a second GLM with a penalization of $\lambda = e^{-6.5}$ which we will call model *GLMNet_small*.

Model name	Penalization	Number of fitted coefficients	Number of variables used
GLM	0	143	38
GLMnet_long	$\lambda_{1se} = 8.51e-05$	77	32
GLMnet_small	$1.50e-03$	41	19

Table 6.1: Summary table of the GLM models fitted

We provide in table 8.1 the list of variables used by model and in 8.4 part of the fitted coefficients.

6.1.4 Explaining mortality rate differences with GLMs

In this section, we will take advantage of the interpretability power of GLM models to explain the predicted mortality rates deviation from our GAM life tables. We will also show how to explain the differences of two annuitants' mortality. The linear form of GLMs will enable us to decompose mortality rates by risk factors.

As an example, we will compare two applicants having age 65 and requesting underwritten annuities. The first applicant A has had a Type 2 diabetes for 10 years and the second applicant B suffers from Chronic Obstructive Pulmonary Disease (COPD).

In our example, both applicants are females belonging to mosaic cluster F. They have a BMI of 25, an SBP of 125, are smokers with consumption 20 cigarettes per day and alcohol consumption 10 glass per week. Applicants have no other comorbidities.

We use our trained GLMnet_small model to predict mortality curves. Because we will perform an extrapolation at high ages, we focus only on mortality under age 95.

We plot the predicted mortality rates for each annuitant A and B, as well as mortality rates of GAM life table of gender female and mosaic cluster F which we will refer to as ref:

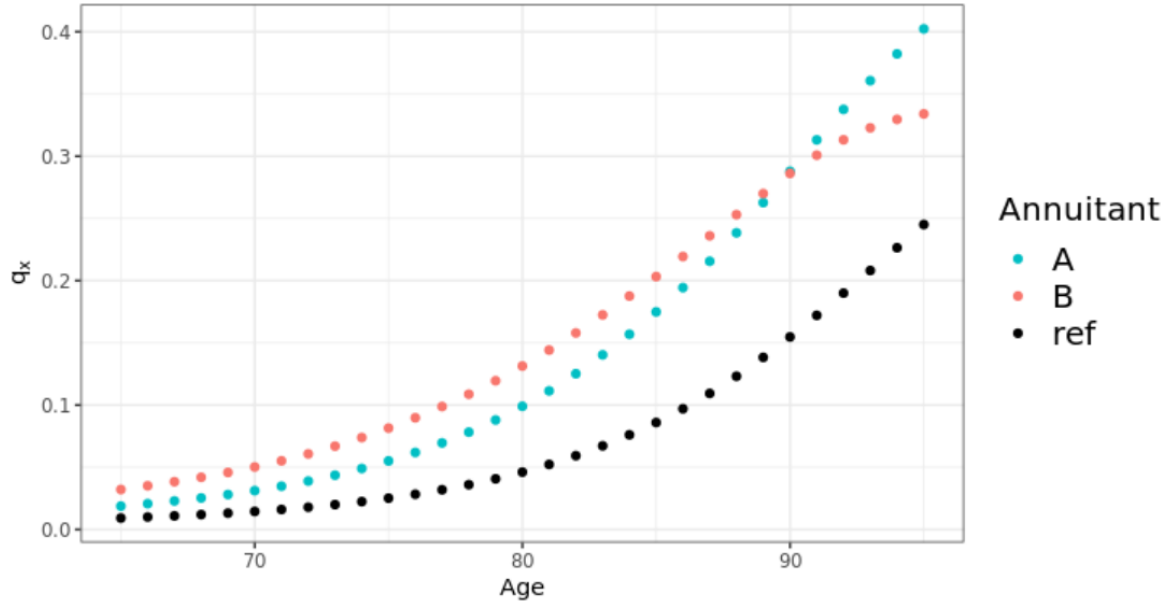


Figure 6.4: Predicted mortality rates for applicants A and B

As we can see, both annuitants have higher mortality rates than the average reference of their gender and mosaic cluster. In order to explain those differences, we use the linear form of the GLM.

Explaining deviation from reference mortality rates

We focus here on explaining the extra mortality of annuitant A compared to the reference mortality curve. We recall that $q_x(X^A) = q_{x,g^A,c^A}^{GAM} \exp(\beta X^A)$, with x the age, g^A the gender of applicant A, c^A their mosaic cluster and X^A their medical risk factors. Thus when we focus on the differences of the two predicted mortalities we have:

$$\frac{q_x(X^A)}{q_{x,g^A,c^A}^{GAM}} - 1 = \exp(\beta X^A) - 1$$

$$\approx \beta X^A$$

We see that the increase in mortality rates can be easily approximated by βX^A . This allows us to easily decompose the variation of the predicted mortality rates by each medical risk factor X_i . We plot in the following graph the coefficients $\beta_i X_i^A$ (we regrouped all terms related to diabetes together, as well as for smoking, bmi and sbp for more clarity):

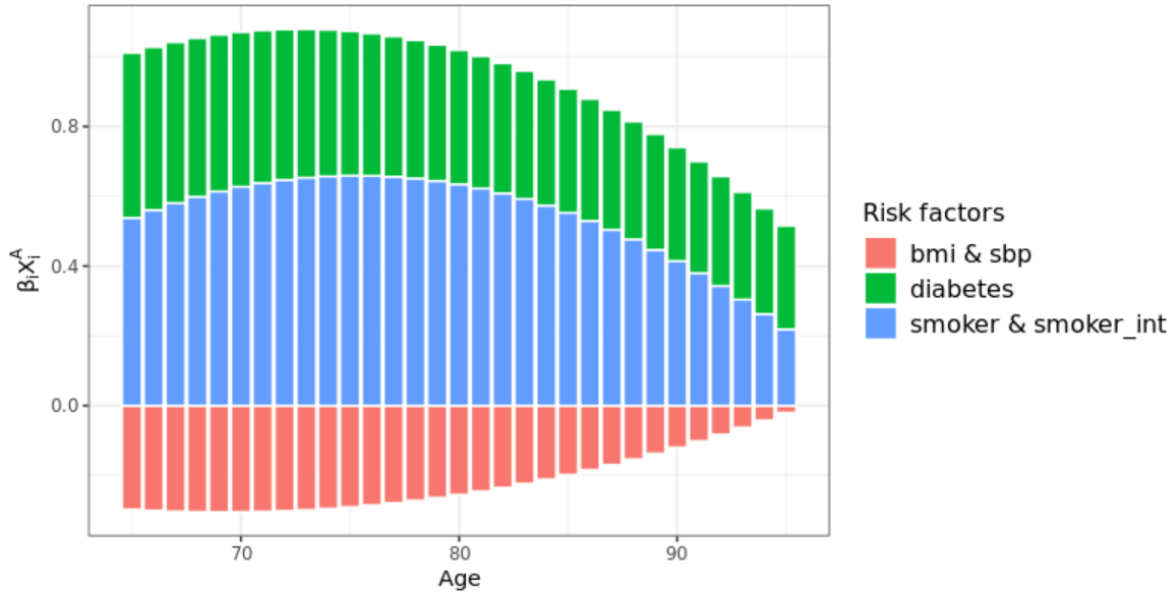


Figure 6.5: Decomposition of extra mortality by medical risk factors for annuitant A

We can see here that diabetes risk factor and smoking habits increases mortality rates of applicant A. We see however that their "good" bmi and sbp metrics decreases it. We can also notice that the impact of those medical risk factors decreases with age.

Explaining mortality rates difference between two annuitants

We are now interested in explaining the differences in mortality curves of applicant A and B. We recall again that $q_x(X) = q_{x,g,c}^{GAM} \exp(\beta X)$, thus when we focus on the differences of the two predicted mortalities we have:

$$\frac{q_x(X^A)}{q_x(X^B)} - 1 = \exp(\beta(X^A - X^B)) - 1$$

$$\approx \beta(X^A - X^B)$$

As $\beta(X^A - X^B) = \sum_i \beta_i(X_i^A - X_i^B)$, the total differences in mortality is obtain by summing $\beta_i(X_i^A - X_i^B)$ that reflects the contribution of risk factor i .

We plot $\beta_i(X_i^A - X_i^B)$ such as $X_i^A \neq X_i^B$:

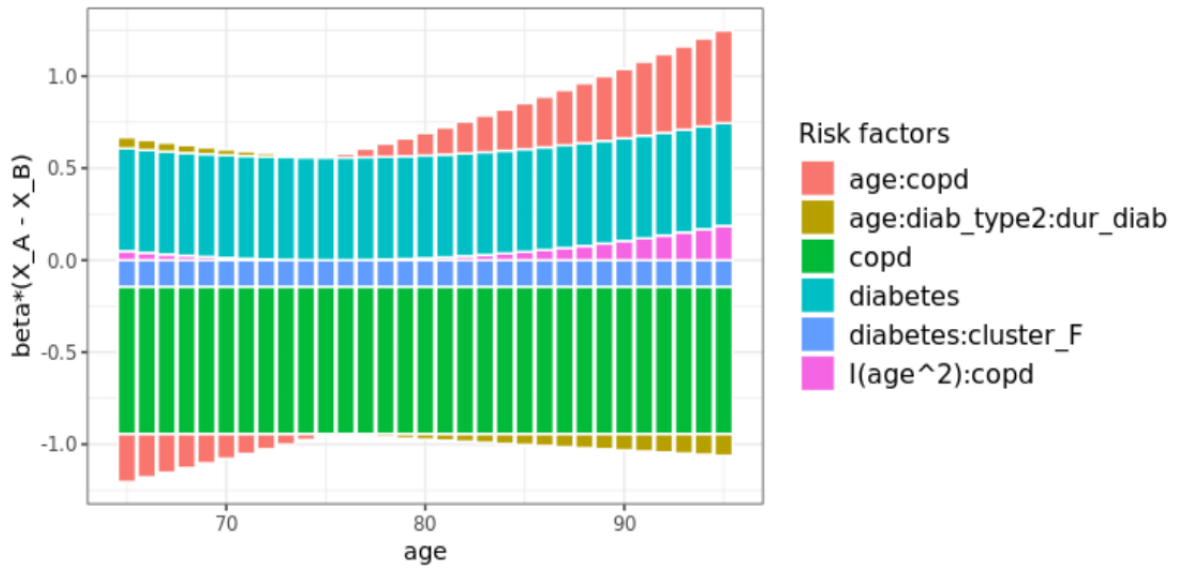


Figure 6.6: Detail on $\beta_i(X_i^A - X_i^B)$ values

Aggregating by medical risk factors, we get:

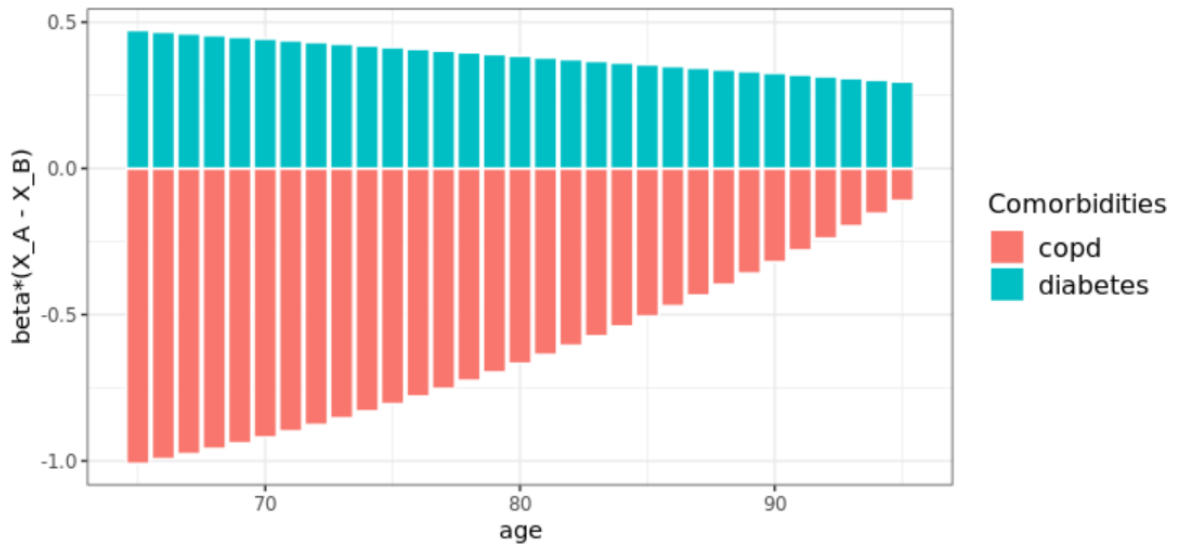


Figure 6.7: $\beta_i(X_i^A - X_i^B)$ values per comorbidity

As we can see, for both comorbidities diabetes and COPD, extra mortality risk decreases with age. We can also observe that COPD risk is higher at age 65 than diabetes, however it decreases with age faster than diabetes.

6.2 Modeling enhanced life annuities using a Machine Learning model

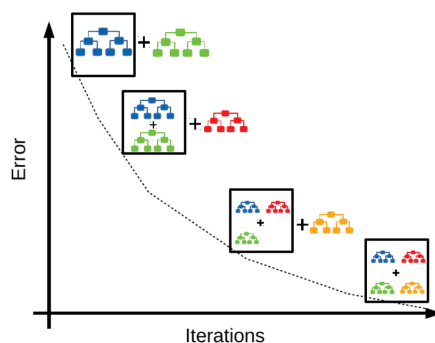
Now that we have fitted standard GLM models, we want to fit a Machine Learning model in order to compare predictions. We believe that an ML model could help detecting impact of explanatory variables that we did not take into account in the GLM. Indeed, ML models have the benefit not to need to have a primary idea of the impact shape of explanatory variable nor the pertinent variables interaction. The drawback of course is the loss of interpretability of those so called "black box" models.

We attempted several Machine Learning models, however we were limited due to the peculiarities of our database. For example, random forests are not adapted in presence of exposure, and GLMtrees were too computationally intensive to be solved. Therefore, we will focus on Gradient Boosting Machine only.

6.2.1 Gradient Boosting Machine

Gradient Boosting Machine is an ensemble method of weak learners (decision trees in our case). This means that it uses a lot of simple models (weak learners) in order to make its prediction. Boosting consists in constructing those weak learners iteratively, and adding them to a final strong model.

We can visualise the iterative construction of GBM the following way:



Iterative learning of GBM using decision trees as weak learners (source: Pal A. [7])

The algorithm followed by GBM is the following :

Initialization : $F_0 = \underset{c}{\operatorname{argmin}} \mathbb{E}[L(Y, c)]$ ▷ Start by fitting constant value
for $k=1$ to K **do**
 $r_k = -\frac{dL(F_{k-1}(X))}{F_{k-1}(X)}$ ▷ compute pseudo residuals
 fit h_k on vector (X, R_k) ▷ train weak learner on pseudo residuals
 $\gamma_k = \underset{\gamma}{\operatorname{argmin}} L(F_{k-1}(X) + \gamma \times h_k(X))$ ▷ Likelihood optimisation of γ multiplier
 $F_k = F_{k-1} + v \times \gamma_k h_k$ ▷ Adding fitted weak learner
end for

Likelihood L has to be specified, depending on the assumed model (Poisson in our case).

As for the GLM, we provide to the GBM the life table as basis prediction.

6.2.2 Grid search

Several hyper-parameters have to be provided before fitting the GBM. We present here some of them :

- K : the number of weak learners to construct
- v : learning rate parameter, enabling to reduce over-fitting.
- max_depth : the maximum depth of the weak learners (decision trees)

In order to determine the hyper-parameters that would enable the best model to be fitted on our data, we run a grid search. Simple grid search consists of testing all combination of hyper-parameters we want to test. We keep the set of hyper-parameters providing the minimum residual deviance. Models are trained on a train set and performance is estimated on a separate validation set in order to avoid over-fitting effect.

We use H2O package to test the following hyper-parameter values :

Hyper-parameter	Tested Values
K (n_trees)	200,500,1000,5000
v (learn_rate)	0.01,0.05,0.1,0.2
max_depth	1,2,3,4,5

Minimum residual deviance was found for the following set of hyper-parameters $\{K=1000, v=0.1, \text{max_depth}=3\}$.

6.2.3 Sensitivity analysis

GBM is what is called a "black box" model, meaning it cannot be interpreted easily because of the too large number of weak learners fitted. However some methods, called sensitivity analysis, can enable to get an idea of the impact of each feature on the model's prediction.

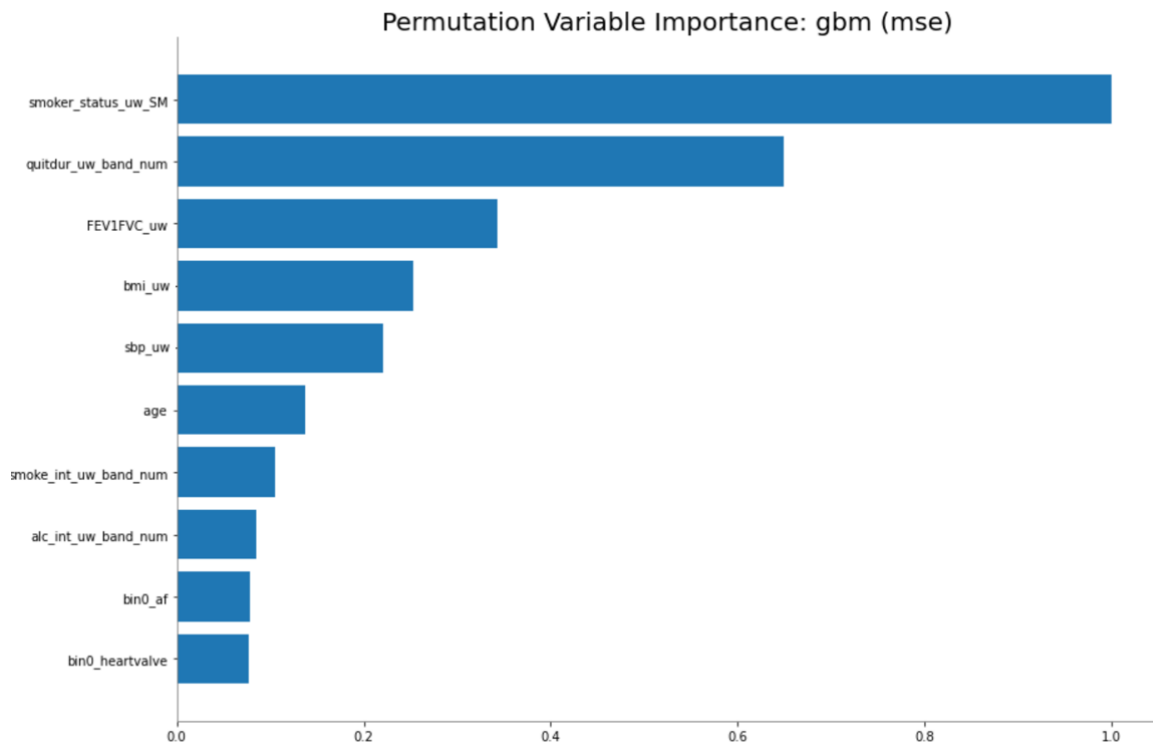
Permutation feature importance

We start by showing the variables that contains the most valuable information. Permutation feature importance measures the increase in the prediction error of the model after having altered the values of one explanatory variable [31]. The idea is to permute one column values and see what effect it has on the model performance. If the model performance has dropped significantly, then the explanatory variable must have an important prediction impact on the prediction. If the model performance does not decrease much, then the explanatory variable does not have a strong predictive value.

The detailed process of computing variable importance is the following :

```
Initialization :  $e_0 = \text{MSE}(y, f(X))$            ▷ Estimate initial mean square error
for  $j=1$  to  $p$  do                               ▷ Iterate on the  $p$  variables
     $\tilde{X}_j = \text{Perm}(X_j)$                            ▷ Permute values of vector  $X_j$ 
     $e_j = \text{MSE}(y, f(X_1, \dots, \tilde{X}_j, \dots, X_p))$    ▷ measure MSE of prediction with  $\tilde{X}_j$ 
     $FI_j = e_j - e_0$                                ▷ Compute relative feature importance
end for
```

We estimate permutation feature importance on the test set, giving :



Most important variables in the GBM model

We notice that age is part of the most important variables, despite the fact that our life table constructed with age, gender and cluster was provided as basis prediction. This means that these variable must have an additive impact once interacted with comorbidity variables. GLM model should therefore include interaction terms of age with comorbidity variable. We note that our feature engineering already takes into account a lot of those interactions.

Partial dependence plots

The partial dependence plot shows the marginal effect of one feature on the predicted outcome. It can give insights of the relationship shape of a feature of interest on the target variable, controlled by the other explanatory variables.

Partial dependence of a variable X_s at a given value v is given by the mean predicted output on the test data set when setting feature $X_s = v$ for all patients.

We show here the ones from BMI and SBP variables.

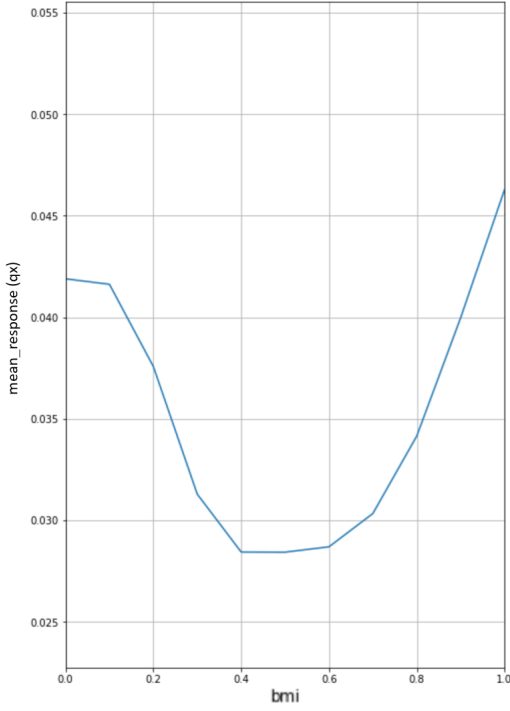


Figure 6.8: Body Mass Index

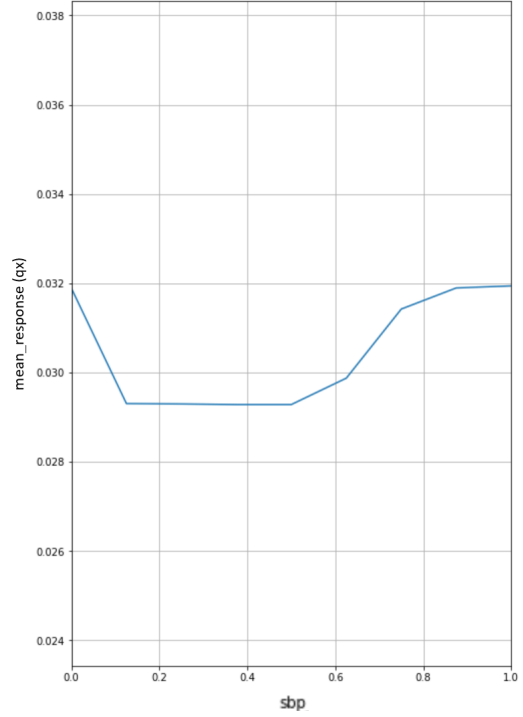


Figure 6.9: Systolic Blood Pressure

We see that the GBM succeeds to capture the U-shape impact of BMI and SBP on mortality rate, that we observed on figures 6.1 and 6.2. We provide additionally in Appendix 8.11 the partial dependence plots of smoking and alcohol intensities.

Sobol indices

A domain of interest in sensitivity analysis is the possibility to detect interaction effects between the explanatory factors. We follow the approach applied by Silvia Bucci in her master thesis *Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification Non-Vie. Application à la tarification à l'adresse*. [32], in which Sobol indices were used to detect interaction effects identified by complex black box models.

Sobol indices is a variance-based sensitivity analysis. The estimation of those indices consist of decomposing the conditional variance of the variable of interest Y with respect to the variance of the explanatory variables $X = (X_1, \dots, X_d)$ and their combinations of interaction terms, in order to input them in a traditional GLM model.

Mathematically, this decomposition of variance can be written as:

$$Var(Y) = \sum_{i=1}^d V_i + \sum_{i < j} V_{i,j} + \dots + V_{1,\dots,d}$$

with

$$\begin{aligned} V_i &= Var(\mathbb{E}(Y|X_i)) \\ V_{i,j} &= Var(\mathbb{E}(Y|X_{i,j})) - V_i - V_j \\ V_{i,j,k} &= Var(\mathbb{E}(Y|X_{i,j,k})) - V_i - V_j - V_k - V_{i,j} - V_{j,k} - V_{i,k} \end{aligned}$$

and so on.

Sobol indice S_{X_U} of the subset of explanatory variable X_U , $U \subset \{1, \dots, d\}$, is the percentage of variance explained by the interaction effect of the combined components of X_U .

$$S_{X_U} = \frac{V_U}{Var(Y)}$$

First order indices refer to the Sobol indice S_i of one explanatory variable. Second order refers to indices $S_{i,j}$ of pairs of explanatory variable.

By construction, the sum of all Sobol indices is equal to 1, which explained why we can interpret Sobol indices as a percentage of explained variance. The drawback of this method is that those indices are estimated by Monte Carlo simulations, which are computationally intensive. Numerous attempts to evaluate second order Sobol indices were conducted. Indeed, it would have been of value to detect possible interaction impacts that are not yet included in our specified GLM model.

However, due to the high quantity of explanatory variables and computational limit constraints, Sobol indices simulation results were unstable and inconclusive. To illustrate instability in results, we show in appendix 8.13 the output of significant interactions of two simulation of Sobol indices estimation. Research had to be dropped eventually.

Chapter 7

Life expectancy prediction

We make the choice to focus on life expectancy, as it is a good proxy of a life annuity value. Indeed, the present value discounted at 0% of 1 life annuity is equivalent to life expectancy. As the difference between interest rate and escalation is low, the value of 1 life annuity is close to life expectancy.

Thus, the predicted life expectancy provides a proxy of the annuity price that could be offered to annuitants depending on their health condition. The higher the predicted life expectancy, the less annuity amount will be offered. It enables us to get a proxy of the annuity price variations based on the several models constructed in the previous chapters.

On top of our reference life tables, we have fitted four models: one GLM using all explanatory variables, two GLMs constructed via a GLMNet using less explanatory variables and a GBM. In this chapter, we analyse their life expectancy predictions.

We present how to compute life expectancies from the predicted mortality rates and analyse them.

7.1 Residual life expectancy

Definition

We compute periodic life expectancies that measure the life expectancy of an individual as if there is no evolution in mortality into the future. In other words, we solely base our calculations on the estimated mortality rates at a given point in time.

To compute the real life expectancies, one would need to consider forecast of future mortality evolution, i.e. future mortality trend. As this master thesis focuses on mortality level, we choose to work with periodic life tables.

Even if periodic life expectancies doesn't provide an accurate estimation of life expectancy, it is an often used metric. For instance, demographers used it to compare differences in mortality from one group of social to another, such as differences in life expectancy between blue and white collar workers.

Residual life expectancy calculation

We are interested in computing residual life expectancy e_x at age $x = 65$. Relying on actuarial notations, one can compute the life expectancy as follows:

$$e_x = \sum_{k=0}^{\infty} {}_k p_x \times q_{x+k} \times \left(k + \frac{1}{2}\right)$$

with

- ${}_k p_x$ the probability of someone aged x to be alive at age $x + k$
- q_{x+k} the probability of someone aged $x + k$ to die within the year (mortality rate at age $x + k$)

For computation ease purpose and because our model provides us with mortality rates q_x , we convert the previous equation to the equivalent form that computes life expectancy in a descant sequence fashion:

$$e_x = 0.5q_x + (1 + e_{x+1})(1 - q_x)$$

with $e_{x=110} = 0$.

By fixing $e_{x=110} = 0$, we suppose that $q_{110} = 1$. As said previously, because we want to compute residual life expectancy at age 65, we apply this formula from $x = 110$ to $x = 65$.

7.2 High Ages Extrapolation

Our database did not enable us to have enough data to estimate mortality rates at age greater than 100. To compute life expectancy, we rely on mortality rates from ages 100 to 110 provided by the Human Mortality Database (HMD).

Human Mortality Database

The Human Mortality Database is an open source database providing death rates and life tables for national populations [2]. HMD constructs and share reliable mortality datasets by using death registration and census data. The dataset covers many developed and industrialized countries. Because our database is composed of people representative of the UK population, we will import the national mortality rates of the UK on period that correspond to the THIN data study period, namely 2015-2018.

Linear convergence at high ages

Several technics for high ages mortality extrapolation haven been proposed. Delwarde and Denuit provide a comparison between the several technics, considering the Belgium national population mortality data in [12].

HMD also rely on an extrapolation method to construct the mortality data on high ages (exceeding 90 on recent period) mortality. Their proposal is used by many researchers and could be considered as a field standard.

Due to lack of data at high ages, our model prediction can be considered as poor on highest ages. Indeed, because we have fewer data at high ages, confidence interval are

wider (as we observed when plotting mortality rates of GAM life tables). It is therefore more reasonable to combine those predicted mortality rates with national mortality rates.

We propose to apply a linear interpolation on predicted mortality rates from ages 96 to 100 using the rates constructed by HMD as follows:

$$q_x^{extrap} = \begin{cases} q_x^{pred} & x < 95 \\ (1 - \frac{x-95}{5})q_x^{pred} + \frac{x-95}{5}q_x^{HMD} & x \in [95, 100] \\ q_x^{HMD} & x > 100 \end{cases}$$

We apply this interpolation to all predicted rates from ages 96 to 100 for all models: GAM, GLM, GLMnet and GBM.

This ensures that the mortality assumption on ages 95 to 110 are consistent, whatever model considered for ages under 95.

7.3 Analysis of computed residual life expectancy

Having performing mortality extrapolation at high ages, we compute residual life expectancy at age 65 of all patients of our test database.

We consider only ages 65 to 68 at the start of the study period. We apply this filter on age in order to have a representative distribution of health status of the population at age 65. Indeed, the individual in our database that we observe at higher ages are more likely to be in good health.

7.3.1 Average predicted residual life expectancy at age 65

As a first analysis, we compute the mean residual life expectancy predicted for each model.

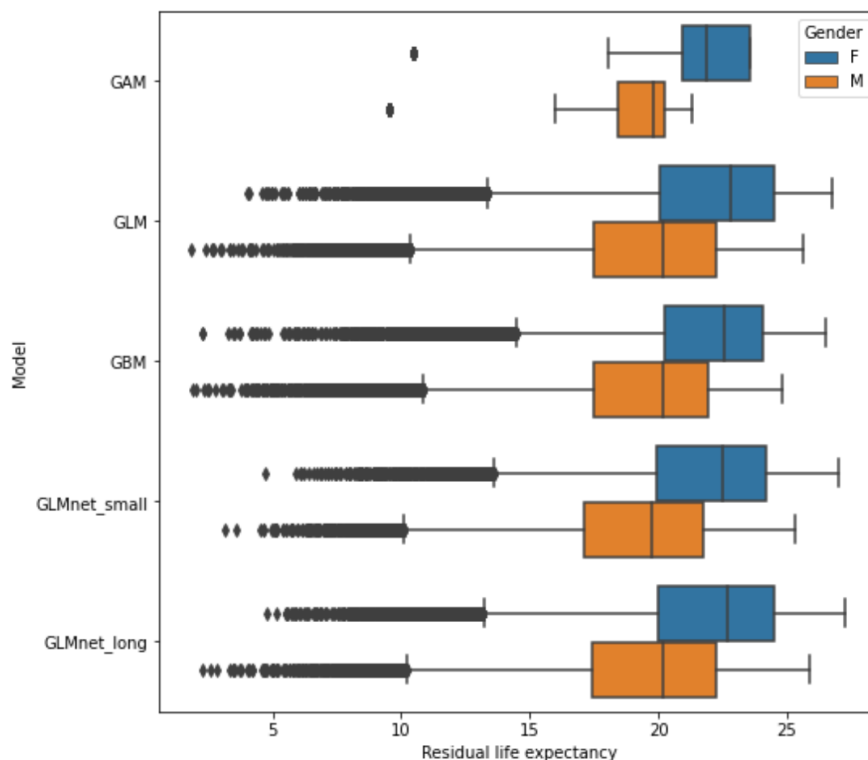
Model	Female	Male
GAM	21.74	19.45
GBM	21.78	19.37
GLM	21.92	19.55
GLMnet_small	21.79	19.17
GLMnet_long	21.86	19.53

Table 7.1: Mean predicted residual life expectancy at age 65 per gender

Overall, the average predicted life expectancies are similar from one model to another. However, we can notice that it is the GBM that is the closest to the mean prediction of the GAM reference life table. The GLM and GLMnet_long models seems to both slightly overestimate life expectancy for male and female compared to the reference life table.

7.3.2 Segmentation of prediction

We show here the distribution of estimated life expectancy for each model, through box plots showing minimum, first quartile, median, last quartile and maximum:



Distribution of predicted residual life expectancy at age 65 for each model

As expected, we can visualize here the increase of segmentation, or discrimination, of the models that includes health status variable compared to the GAM life table model. Indeed, we recall that GAM model does rely only on age, gender and socio-economic data to predict the mortality.

It seems that the GLMnet_small model segments a little less than the other 3 models, which would make sense since it has less explanatory variable and therefore is less able to segment patients.

We provide in table 8.14 an example of some life expectancy predictions for some patients.

To have a more accurate idea of the difference of prediction between models, we plot for each pair of models the prediction of one model vs the second for some randomly selected patients.

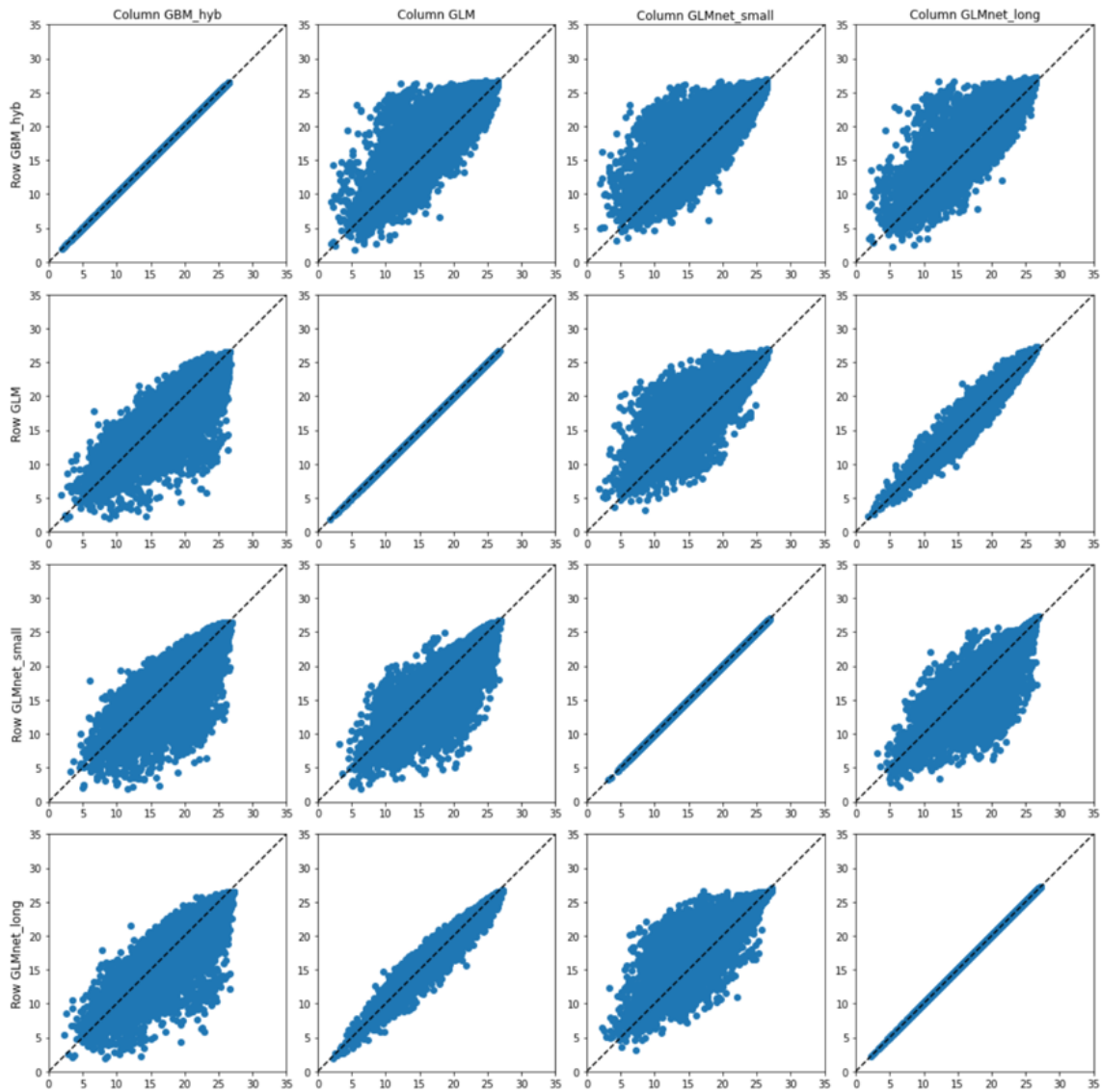


Figure 7.1: Comparison of patients life expectancy prediction of model i vs model j

Each point in each subplot represents one patient, with for x and y axis the residual life expectancy at age 65 of the two models.

We notice that the GLMnet_long model is really close to the prediction of the GLM one. However, we see that there is more difference with the prediction of the GBM model.

We support this graphical analysis with a table summarising the mean of absolute difference in life expectancy predictions of our models.

	GAM	GBM	GLM	GLMnet_small	GLMnet_long
GAM					
GBM	2.18				
GLM	2.38	0.94			
GLMnet_small	2.16	1.13	1.04		
GLMnet_long	2.39	0.90	0.40	0.96	

Table 7.2: Mean of absolute difference in life expectancy predictions

Now that we computed residual life expectancies and have graphically shown some of the differences in prediction of the models, we will use different metrics in order to assess the performance of each model. Life expectancy predictions will be of use for some (but not all) performance metrics.

7.4 Explaining life expectancy predictions with GLMs

Similarly to what we did in subsection 6.1.4 on predicted mortality rates of GLM models, we are here going to explain life expectancy predictions of GLM models.

Linear approximation of life expectancy using Taylor-Young formula

In this section, we are going to show how to take advantage of the linear form of GLM models in order to explain how risk factors impact on the predicted life expectancy. More specifically, we show how to evaluate the impact of risk factors on life expectancy and explain the deviation from the life expectancy of reference (reference means here the life expectancy that would be predicted using only gender and mosaic cluster). As in subsection 6.1.4 we consider the following approximation:

$$q_x(X) = q_{x,g,c}^{GAM} e^{\beta X} \quad (7.1)$$

For the following computation, we will consider an annuitant of fixed gender g and mosaic cluster c , and we will note $q_x^{ref} = q_{x,g,c}^{GAM}$ to simplify notations.

$$\begin{aligned} q_x(X) &= q_x^{ref} e^{\beta X} \\ &\approx q_x^{ref} (1 + \beta X) \\ q_x(X) - q_x^{ref} &\approx q_x^{ref} \beta X \end{aligned} \quad (7.2)$$

Let the residual life expectancy at 65 e_{65} be a function of the vector of mortality rates $Q = (q_{65}(X), q_{66}(X), \dots, q_x(X), \dots)$, we use Taylor-Young decomposition to approximate life expectancy of an annuitant e_{65} with the life expectancy of reference $e_{65}^{ref} = e_{65}(Q^{ref})$:

$$e_{65}(Q) = e_{65}(Q^{ref}) + \nabla e_{65}(Q^{ref})(Q - Q^{ref}) + o(Q^2) \quad (7.3)$$

with Q the vector of mortality rates $q_x(X)$ of the annuitant and Q^{ref} the vector of mortality rates from our life tables using only gender and mosaic of the annuitant.

We now want to estimate the gradient $\nabla e_{65}(Q^{ref})$ so that it can be easily computed. We start from the expression of e_y using survival function $S_y(t)$:

$$e_y = 0.5 + \sum_{t \geq 1} S_y(t)$$

with

$$S_y(t) = \prod_{k \geq 1}^t (1 - q_{y+k-1})$$

which leads to:

$$\frac{\partial S_y(t)}{\partial q_x} = \begin{cases} -\frac{1}{1-q_x} S_y(t), & \text{if } y \leq x < y+t \\ 0, & \text{otherwise} \end{cases}$$

From there, assuming $x \geq y$, we can express the derivative of e_y :

$$\begin{aligned} \frac{\partial e_y}{\partial q_x} &= \sum_{t \geq 1} \frac{\partial S_y(t)}{\partial q_x} \\ &= \sum_{t \geq x-y+1} -\frac{1}{1-q_x} S_y(t) \\ &= -\frac{1}{1-q_x} S_y(x-y) \sum_{t \geq x-y+1} S_x(t - (x-y)) \\ &= -\frac{1}{1-q_x} S_y(x-y) (e_x - 0.5) \end{aligned}$$

Coming back to equation 7.3, we can now express it by:

$$\begin{aligned} e_y(Q) &\approx e_y(Q^{ref}) + \nabla e_y(Q^{ref})(Q - Q^{ref}) = \\ &= e_y^{ref} - \sum_{x \geq y} \frac{1}{1-q_x^{ref}} S_y(x-y) (e_x^{ref} - 0.5) \underbrace{(q_x(X) - q_x^{ref})}_{\approx q_x^{ref} \beta X \text{ from 7.2}} \\ &= e_y^{ref} - \sum_{x \geq y} \frac{q_x^{ref}}{1-q_x^{ref}} S_y(x-y) (e_x^{ref} - 0.5) (\beta \cdot X) \end{aligned}$$

As $\beta \cdot X = \sum_i \beta_i X_i$, we can now express the deviation of $e_y(Q)$ from e_y^{ref} as a sum on each risk factor X_i :

$$e_y(Q) - e_y^{ref} \approx - \sum_i \sum_{x \geq y} \beta_i X_i \left(\frac{q_x^{ref}}{1-q_x^{ref}} S_y(x-y) (e_x^{ref} - 0.5) \right)$$

Thus, the contribution of risk factor X_i is: $\sum_{x \geq y} \beta_i X_i \left(\frac{q_x^{ref}}{1-q_x^{ref}} S_y(x-y) (e_x^{ref} - 0.5) \right)$

Illustrative example of life expectancy decomposition per risk factor

This approximation enables us to explain the gap of life expectancy between a specific annuitant and the general population. To illustrate this decomposition of life expectancy

per risk factor, we use the same annuitant A example as in section 6.1.4. As a reminder, our annuitant is a female from group socio-economic F, having type 2 diabetes. Based on the life table of female with socio-economic group F, the residual life expectancy at age 65 of this population is of 19.9 years. Using the glmnet_small model that takes into account their health status, the predicted residual life expectancy of annuitant A is 14.7. We now want to explain why annuitant A has a lower residual life expectancy than the one from their corresponding life table (female of socio economic group F). Using the linear approximation of life expectancy we just presented, we decompose the gap of life expectancy into gains and losses of life expectancy per risk factor. Here are the results:

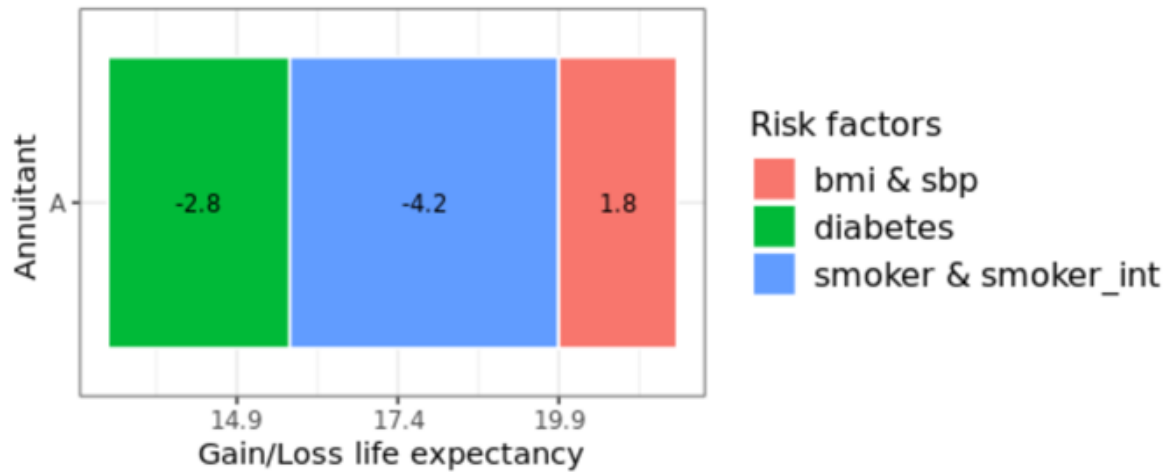


Figure 7.2: Gain/Loss of life expectancy per risk factor

We can then say that due to good bmi and sbp, life expectancy of annuitant A increases by 1.8 years compared to the general population. However, diabetes status reduces their predicted life expectancy by 2.8 years and smoker status reduces it by 4.2 years.

Chapter 8

Assessing models performance

We are interested in metrics to help determine which one of the models is the "best". We will present here different ways to assess the performance of our models and assess our models using those metrics on the *test set*.

Our goal here is to see whether it is possible to obtain good performance with a reduced number of explanatory variables. Indeed, the underwriting process would gain from reducing the number of data to collect from the annuitants. We will therefore seek whether GLMNet models are close to the GLM one.

Another goal is to check whether GLM models are outperformed by the GBM. The Gradient Boosting Model is a powerful ML approach, that is able to automatically capture interactions between variable along with non-linear effect. Thus, if GBM outperforms the GLM, that could mean that we missed some effect when setting the GLM.

Assessing the performance of survival models is not always straightforward. In this section, we follow the methodology explained and applied in *Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches)* written by Park S & al [25].

The methodology consists in assessing survival models through different types of metrics: overall metrics, calibration metrics and discrimination metrics. In this section we focus on each type of these categories, present some metrics and apply them on our test set.

8.1 Calibration performance metric

In this section, we present calibration metrics. Calibration evaluates whether the predicted probabilities fit with the observed outcome on average [25] [30].

Because we cannot observe the actual probabilities of each individual (we can only observe the outcome of survival or death), we group the pseudo observations and then assess how the average predicted probability compares with the observed probabilities in each group. The groups should be formed to gather pseudo observations with homogeneous death probabilities. Calibration metrics are statistics that partition a data set into groups and assess how the average predicted probability compares with the outcome in each group.

Because calibration metrics focus on the accuracy of a probability measure, we will estimate it on our pseudo-observation test set. Each line of our dataset will therefore provide us with a predicted mortality rate (one for each model), exposure, and an outcome death status.

8.1.1 Calibration plot

Calibration plot is a primary graphical method for evaluating calibration performance. It consists of creating groups by homogeneous mortality rates. Because it involves mortality rates, this metric will be assessed using our test database at pseudo-observation level.

Therefore, for each model, we will divide our pseudo observations by quantiles of the predicted probabilities. Because age has such a big impact on mortality rates, and we do not want to create bins that will end up regrouping pseudo observation of the same age, we divide our pseudo observations by quantiles of the predicted probabilities per age. This will enable to create bins having the same distribution of age.

As we have a large amount of data at our disposal, we can afford to create our bins using percentiles. We therefore divide our test pseudo observation database into 100 groups.

In Figure 8.1, we provide a scatter plot chart for each model. A point corresponds to one of the 100 bins created. We plot the observed empirical mortality rate inside the bin (y axis) versus the mean predicted mortality rate (x axis).

A model that performs well, will have predictions close to observation. Thus, the closer the curve to the $y=x$ axis, the better model calibration. If the bins are on the left side of the $y=x$ axis, then the model has underestimated the empirical mortality rate of the bin. If the bins are on the right side of the $y=x$ axis, it overestimated it.

As we are dealing with longevity risk, we want to avoid overestimating mortality rates. Indeed, overestimating mortality rate will result in predicting underestimate life expectancy. In the case of life annuity products, this would lead to an unanticipated and underestimated amount of annuity cash flows.

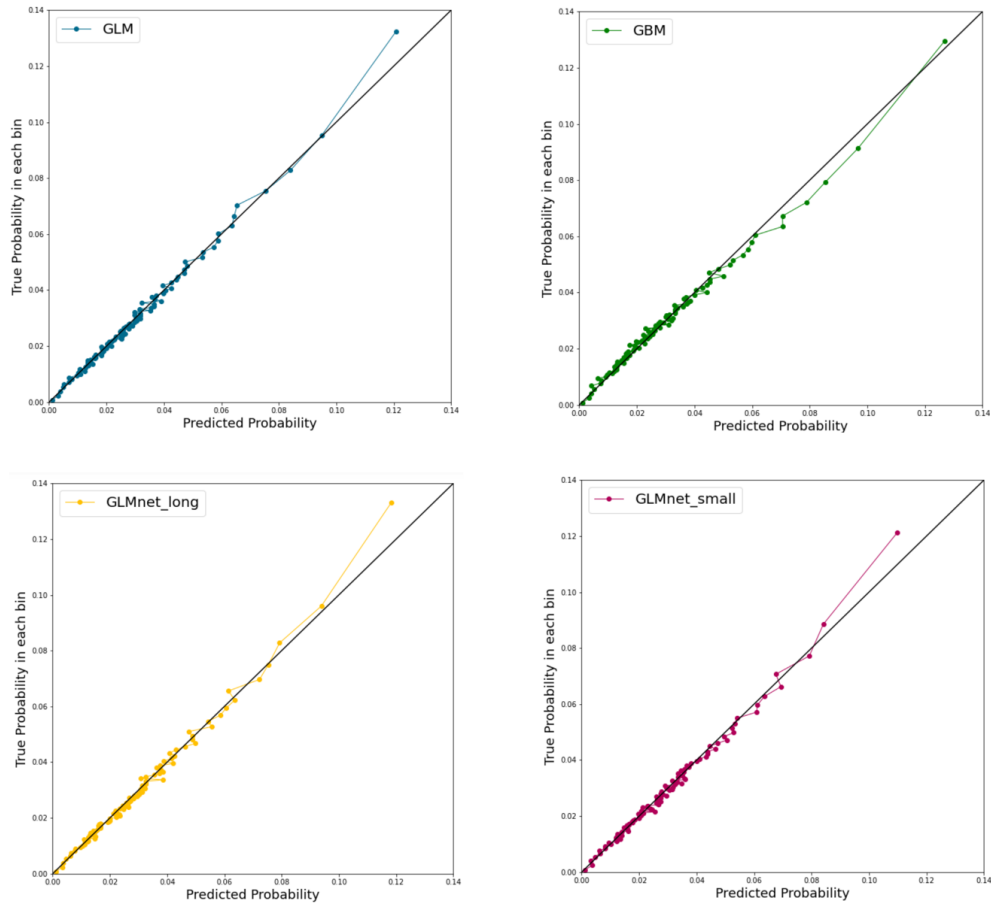


Figure 8.1: Calibration plot per model with 100 bins

The calibration plot confirms some remarks we made while analysing life expectancy predictions. Namely, models GLM and GLMnet_long seem to have similar behaviour and we notice once again the smaller discrimination power of the GLMnet_small model. Indeed, the last percentile, and therefore the pseudo observations considered as being the most at risk, has a mean mortality rate smaller than the 3 other models, meaning there is more mutualization of the risk or in equivalently less segmentation.

Apart from those confirmations, we can say that models seem to be rather well calibrated. However, we can see that all the models seem to underestimate the mortality risk in the last percentile bin (the riskier pseudo observations), apart from the GBM model.

8.1.2 A/E per risk factor

Another way to assess calibration is to group pseudo-observation per risk factors (instead of quantiles, or percentiles as done previously). This enables to verify whether models are well calibrated with respect to one or several risk factors.

A/E is a ratio of actual deaths in a specific population to those predicted by a life table or model. The closest it is to 1, the best calibrated is the model. In the case of an A/E exceeding to 1, the model is overestimating the mortality. A/E is a metric similar to SMR that does not necessarily focus on age groups.

It enables to measure whether the model was able to capture the mortality specificities with respect to one or several characteristics (for example age, gender, diabetes status,

etc...).

We compute total number of deaths observed in a group G , it is called the *Actual*. The *Expected* corresponds to the expected number of deaths, it is obtained by summing the predicted mortality rates of pseudo observation i times their exposure:

$$\frac{A^G}{E^G} = \frac{\sum_{i \in G} D_i}{\sum_{i \in G} \hat{q}_i * E_i}$$

with i all pseudo observations within group G .

To assess if A/E's gap from 1 is due to volatility or model misestimation, we provide also the confidence intervals at 95% level.

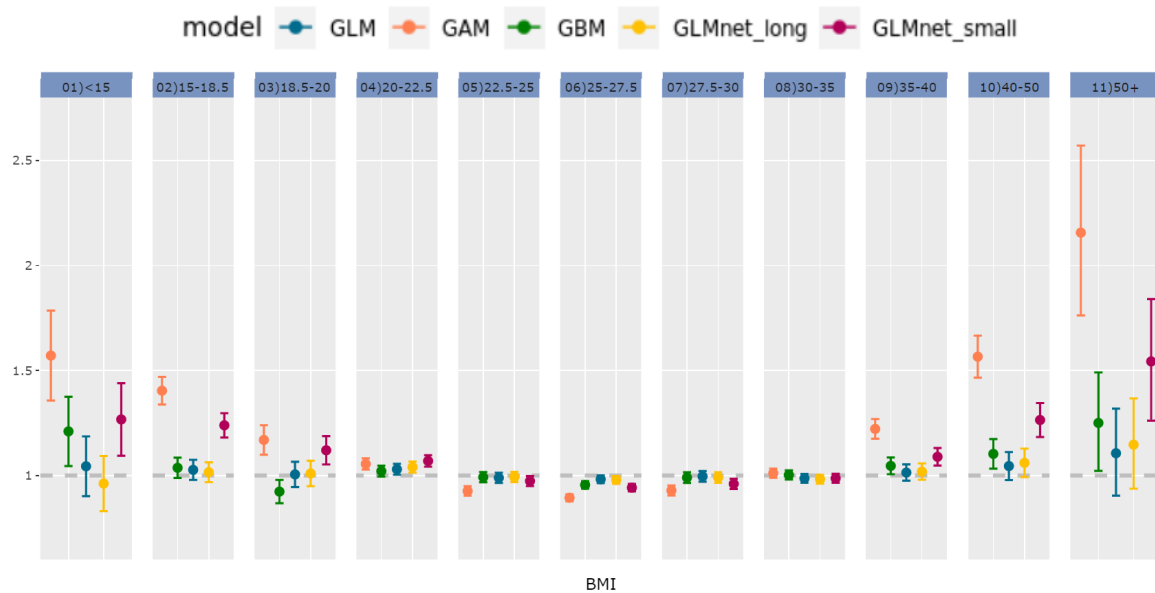
We compute it using the assumption that deaths follows a binomial distribution with mean the predicted mortality rate of group G and size the total exposure of the group. Under this null hypothesis, we get that

$$A^G \sim \mathcal{B}\left(\sum_{i \in G} E_i, \frac{\sum_{i \in G} \hat{q}_i * E_i}{\sum_{i \in G} E_i}\right)$$

Lower and upper bounds of SMRs are computed as quantiles 2.5% and 97.5% of this binomial distribution, divides by expected number of deaths E^G . If 1 is not part of the confidence interval of the SMR of the model for group G , we reject the null hypothesis. This would mean that the actual number of deaths does not follow such a distribution and therefore the model is not well calibrated on the group G .

We decided to plot the Actual over Expected metric for the GAM life tables (using age, gender and mosaic clusters) as well. As the GAM model does not take into account any health related variable, it will mainly reflect the increase (or decrease) in mortality of the risk factors studied. One can see the A/E's obtained with the GAM model as the average impact of a risk factor on mortality.

Body Mass Index



Obtained A/E per BMI group

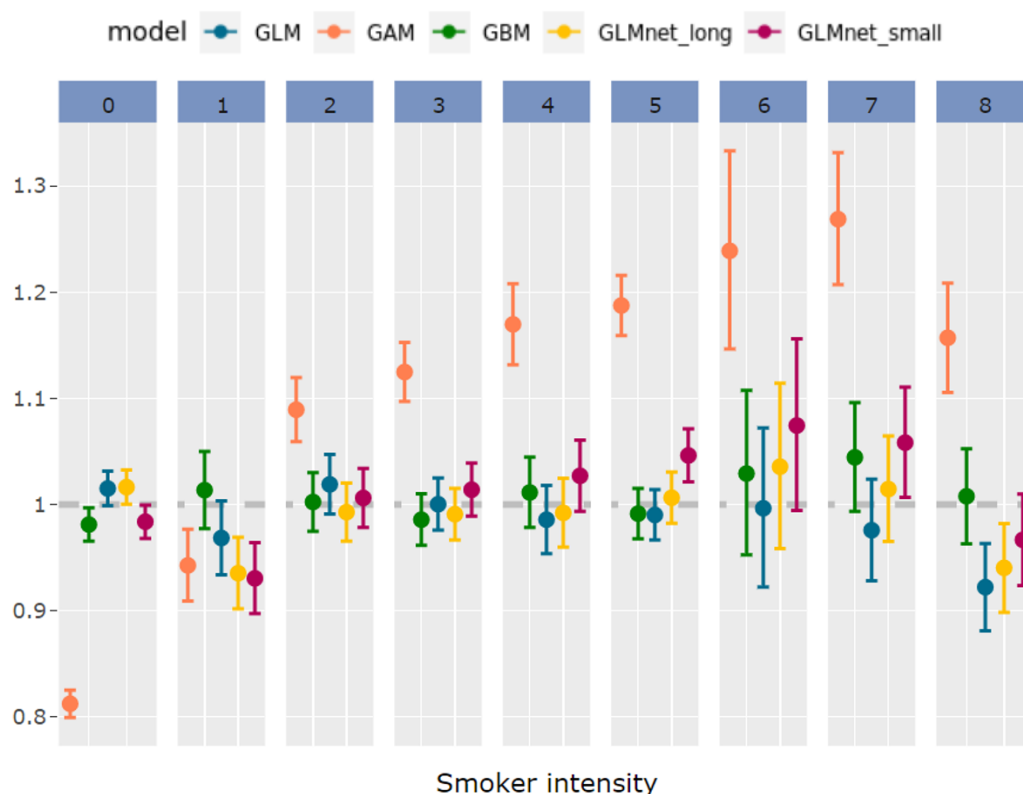
As mentioned previously, A/E of our GAM life tables show the increase of risk. As we could expect, we see that having a BMI too low or too high results in an increased mortality risk.

The first two BMI group corresponds to a BMI below 18.5, that is person who are greatly underweighted. This can be caused by anorexia or due to another pathology. This group exhibits mortality 50% higher than average.

At the other extreme, the last two groups correspond to morbid obesity. This group exhibits even a higher extra mortality, between 50% to 100% higher than average.

Regarding the models trained on comorbidity variables, we can say that they seem to capture well the impact of BMI on mortality. It appears that they are well calibrated with respect to BMI risk factor, except for the GLMnet_small model that seems to be less accurately calibrated at low and high BMI values.

Smoker intensity



Obtained A/E per smoking intensity

As one could expect, the GAM A/Es shows that smoking intensity habits increases the mortality risk.

Intensity 0 is composed of non-smokers, while intensity 8 is composed of the heaviest smokers. Non-smokers have a mortality around 20% lower than average, while heavy smokers (intensities 4 and higher) have a mortality 20% higher.

We reach the same conclusion as for the BMI risk factor. The models are rather well calibrated with respect to smoking intensity risk factor, except for the GLMnet_small model that shows a slightly less accurate calibration.

Other risk factors

In Appendix 8.15, we provide other A/E plots, with respect to comorbidity variables diabetes, asthma, COPD and stroke. We also provide a plot showing combining effect of systolic blood pressure depending on and diabetes status.

Once again, we can conclude that the GLM, GLM_long and GBM are better calibrated than the other two.

8.2 Discrimination performance metric

In previous section we assessed whether the models have an accurate average prediction. In other words, we checked if models are subject to bias on some specific group.

In this section, we are interested in the ability of the models to discriminate the individuals accurately.

In survival models, discrimination performance refers to the ability of a model to predict whether an annuitant will live longer than another one. This ability to discriminate is important for us, as the goal of enhanced life annuities is to offer more advantageous annuities to people that have a life expectancy shorter than others. It is therefore necessary to construct a model that discriminates annuitants accurately.

We note that calibration metrics and discrimination metrics are complementary analysis, one does not insure the other and both metrics should be evaluated.

In the previous chapter, we saw that some of our fitted models segmented more than others. Indeed, the Gradient Boosting Model seems to be the more discriminating models, whereas the GLMnet_small model segments less the patients. We will therefore want to assess which segmentation is the most accurate.

8.2.1 C-index

The most common metric to assess discrimination performance is the Concordance-Index (or C-Index) [25]. C-Index is a metric that has been introduced by Harrell and al. [17] and was first used for biomarker performance analysis in the medical field.

C-Index is the estimate of a probability of concordance. It is a conditional probability that compares whether the model life expectancy predictions (\hat{E}_i, \hat{E}_j) of two individuals i and j are ordered in the same way as their observed survival time (O_i, O_j).

$$\text{C-Index} = P([\hat{E}_i < \hat{E}_j | O_i < O_j])$$

Because the C-Index is a probability, it ranges from 0 to 1. If all pairs of individuals are rightly order, then we would have a C-Index equal to 1. Random guess would have a C-Index of 0.5. In the case of a model that orders wrongly all pairs of individuals, then we would have a C-Index equal to 0.

This probability is estimated by considering all patients in our test set and ordering them by their predicted residual life expectancy at age 65. We then calculate the C-Index as the proportion of concordant pairs divided by the total number of possible evaluation pairs [8].

Because we are in the case of censorship, we cannot compare all patients of our test data set. Indeed, we do not observe the age of death of everyone and therefore cannot

conclude on the survival prediction order accuracy in all cases. We present here the different cases that can arise when comparing two patients i and j and comment on the possibility or impossibility of comparison.

- The easier case occurs when both patients i and j have an observed age at death. In that case, patients i and j are considered as being a "*comparable pair*" because we can compare the predicted life expectancies and conclude of the ranking accuracy. If the ranking is accurate, the pair is said to be "*concordant*". If not, the pair is "*discordant*".
- When both patients have no known age at death, then comparison is not possible and we consider i and j as being an "*incomparable pair*". Nothing can be done to conclude on the ranking accuracy of the model and we will not be able to use this pair of patients to estimate the C-Index.
- The final case arises when death of one of the patients is observed (say patient i) but not the one of the other patient (patient j). Assessment of discrimination accuracy then depends on the predicted ranking of the assessed model:
 - if patient i was predicted as having a longer life expectancy than patient j , and
 - * if patient j is older than the age at death of patient i , then this pair is *comparable* and we can conclude that the pair is *discordant*.
 - * if patient j is not known to be older than the age at death of patient i , then we cannot conclude on the ranking accuracy of the model. These pair is *incomparable* because of censoring.
 - if patient i was predicted as having a shorter life expectancy than patient j , and
 - * if patient j is older than the age at death of patient i , then this pair is *comparable* and we can conclude that the pair is *concordant*.
 - * if patient j is not known to be older than the age at death of patient i , then this pair is *incomparable*.

8.2.2 Computation difficulties

Because the C-Index is estimated on the life expectancy prediction, we will assess it using our test data set at patient level. This data set is composed of 350,231 patients. Therefore, testing all the combinations of pairs of patients is not feasible.

We could simply assess the C-Index on a subset of our test database, however we chose to evaluate the C-Index in a more reliable way which had the benefit to reduce the numbers of pairs to evaluate.

The idea is the following. Due to the survival data structure available, it appears to us that it is more relevant to work by cohorts.

Because we are predicting period life expectancy, comparing patients of different cohorts would not make sense. Indeed, individuals of different cohorts have different Mortality Improvement (MI) rates that would have to be applied. Therefore, resulting period life expectancy cannot be compared. In order to bypass this problem, we compared pairs of patients of same cohort.

As the data is composed of an aggregation of several cohorts, most of the observation couples considered in the C-index calculation would be individual with different age. Considering cohorts, allows us to compute the C-index on couples of individual with the same age. Thus, the C-index will measure the model ability to predict differences in mortalities that are not due to age.

For each cohort h , we computed the number P_h of *comparable* pairs as well as the number V_h of valid / *concordant* pairs. We can therefore compute the C-Index as:

$$\text{C-Index} = \frac{\sum_h V_h}{\sum_h P_h}$$

Comparing patients within cohorts reduced the number of pairs to assess enough to be able to evaluate the C-Index in a reasonable time (a few hours).

8.2.3 Discrimination metric results

As well as for previous metrics results, we provide the GAM as a reference model. We show on the following table the results per model:

Model	Concordant pairs	Comparable pairs	C-Index
GAM	201,566,041	345,304,591	58.37%
GBM	248,876,977	374,979,194	66.37%
GLM	248,227,125	374,951,574	66.20%
GLMnet_small	245,178,677	374,895,628	65.40%
GLMnet_long	247,095,796	374,825,903	65.92%

C-Index discrimination metric of each model

As it should, we see that taking into account health status increases significantly the discrimination performance of our models. Once again, the GBM model has the best C-Index score, follow by the GLM, the GLMnet_long and then the GLMnet_small.

We conclude that our GLM models are not too far off the discrimination power of the Machine Learning one.

8.3 Overall performance metrics

On top of calibration and discrimination metrics, we can assess metrics of overall goodness of fit.

8.3.1 McFadden's Pseudo- R^2

We start by introducing the McFadden (1973) pseudo- R^2 metric [9]. McFadden pseudo- R^2 assesses the improvement from a null model to our fitted model:

$$\text{pseudo-}R^2 = 1 - \frac{\log(L_{M_{\text{model}}})}{\log(L_{M_{\text{intercept}}})}$$

with L_M the likelihood of model M , M_{model} our fitted model and $M_{\text{intercept}}$ the model with only intercept as predictor.

The higher the pseudo- R^2 is, the better the decrease of deviance. Pseudo- R^2 are usually considered as the equivalent of the traditional R^2 in ordinary least square regression. However, we note that we should not expect similar ranges of values as the R^2 usually takes in the presence of good predictive models.

Indeed, as stated by McFadden himself in a footnote of his paper *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments* (pages 34-35) [9]: "Those unfamiliar with the [pseudo- R^2] index should be forewarned that its values tend to be considerably lower than those of the R^2 index and should not be judged by the standards for a "good fit" in ordinary regression analysis. For example, values of .2 to .4 for [pseudo- R^2] represent an excellent fit."

8.3.2 Brier score

The Brier score is a metric equivalent to the Mean Square Error (MSE) adapted for models with binary outcomes. It computes the mean squared difference between the predicted death probability and the outcome status across all the pseudo observations [25].

$$\text{Brier score} = \frac{1}{n} \sum_i^n (\hat{q}_i - \mathbb{1}_{D_i=1})^2$$

with D_i equal to 1 in case of death and 0 in case of survival. A score closer to 0 indicates a better predictive performance

Because we are in the case of censorship and truncation, we must adapt this measure to include this specificity. Therefore, we add exposure as weight and compute a weighted mean.

$$\text{Weighted Brier score} = \frac{1}{\sum_i E_i} \sum_i^n E_i (\hat{q}_i * E_i - \mathbb{1}_{D_i=1})^2$$

Note that Brier scores are not informative by themselves but should be used to compare and rank the performance of different models.

8.3.3 A/E

We can also rely on the A/E measure, but this time computed on the whole test set. We restate that the Actual over Expected ratio is a ratio assessing whether the model tends to over or under-estimate the predicted outcome, in our case mortality.

$$\frac{A}{E} = \frac{\sum_i D_i}{\sum_i \hat{q}_i * E_i} * 100$$

A model predicting accurately will have a $\frac{A}{E}$ score close to 100%. A score inferior to 100% will result in overestimating mortality, and a score bigger than 100% will result in underestimating mortality.

8.3.4 Application of overall metrics

Now that we have presented some overall goodness of fit metrics, we assess them on our test dataset at pseudo-observation level and show the results in the following table:

Model	pseudo- R^2	A/E	Brier score
GAM	0.156	99.36%	0.030867
GBM	0.180	99.44%	0.030597
GLM	0.177	99.78%	0.030635
GLMnet_small	0.170	100.23%	0.030756
GLMnet_long	0.176	99.73%	0.030697

Overall performance metrics

First, McFadden pseudo- R^2 values are indeed, as explained earlier, considerably lower than values that we could have expected from a traditional OLS R^2 . Even though interpretation of the McFadden pseudo- R^2 is not as straightforward as a R^2 , we can use this metric to compare goodness of fit between our models. With respect to McFadden pseudo- R^2 , the GBM appears to be the model that fits best.

Second, the A/E metric informs us that all models overestimate slightly mortality rates which is a risk when dealing with longevity risk, except for the GLMnet_small model that tend to overestimate slightly mortality rates.

Finally, Brier score shows scores that seem to not be different enough from one model to another to draw any conclusions.

8.4 Conclusion on metrics analysis

Based on the different calibration, discrimination and overall metrics, we can conclude that the Gradient Boosting Model is the best predictive model. However, the General Linear Model performance are very close. This means that all the main explanatory variable impacts have been taken in account in the GLM model. Thanks to the data preparation and the insights on the underlying risk behaviour, we manage to reach the prediction power of a machine learning model.

We can also conclude that it would be reasonable to use one of the two GLMnet models. We could recommend the use of the GLMnet_long as it allows to reduces the number of fitted terms with almost no loss in performances.

Conclusion

The aim of this master thesis was to construct mortality level rates, first for standard life annuities products and then for underwritten life annuities products.

For a standard life annuities basis construction, and to respect business operational constraints, we developed and applied a custom ascendant hierarchical clustering method. It allowed us to perform an aggregation of socio-economic classes on a relevant way from mortality risk perspective. We managed to provide smoothed life tables with limited dimensions that captures well the age, gender and socio-economic impact on annuitant mortalities.

For underwritten annuities, we had to use more complex models to estimate mortality risk based on their health status. However, interpretability of models is an important topic as we must be able to explain the differences in the proposed premiums. This is why GLM models were chosen to estimate level mortality risk.

Because of the unknown relationship form existing between the predictors and the target variable, a complex model based on medical expertise and statistical analysis was constructed, featuring a lot of non-linear transformations and interaction terms enabling to encompass a majority of possible effects. In order to retain only relevant features, two additional GLM using LASSO penalization were fitted to reduce the complexity of the model and the number of variables used.

At a time when Machine Learning models are shown to often outperform traditional models, we drew an interest in verifying that our GLM prediction performance was not out-powered by a Machine Learning model. Indeed, the fear for an insurance company to be anti-selected is more than ever present, and it felt necessary to check that our fitted GLM model could segment well enough the annuitants. After considering several Machine Learning models and their compatibility with truncated and censored data, we settle on constructing a Gradient Boosting Model.

Sensitivity analysis was performed on the GBM model in the attempt to unravel potential unconsidered patterns in the data that might not have been considered by the GLM features. Partial dependence plots confirmed non-linear relationship of some variables. Also, Sobol indices estimation was performed on the GBM model in the desire to assess if we had taken into account the main explanatory variables interactions during the GLM feature engineering process. Unfortunately, estimation was computationally intensive, and our results were inconclusive and unstable.

Using several metrics adapted for survival analysis studies allowed us to conclude that the GLM had a predictive performance very close to the GBM model. We also concluded that it felt reasonable to reduce the number of terms in the GLM with a limited loss of performance.

Dictionary

A/E Actual versus Expected.

ADL Activities of Daily Living.

AIC Akaike Information Criterion.

BMI Body Mass Index.

C-Index Concordance-Index.

COPD Chronic Obstructive Pulmonary Disease.

GAM Generalized Additive Model.

GBM Gradient Boosting Machine.

GLM Generalized Linear Model.

GP General Practitioners.

HMD Human Mortality Database.

IMRD IQVIA Medical Research Database.

MI Mortality Improvement.

ML Machine Learning.

MSE Mean Square Error.

OLS Ordinary Least Square.

SBP Systolic Blood Pressure.

SMR Standardized Mortality Ratio.

THIN The Health Improvement Network.

UK United Kingdom.

Bibliography

- [1] The health improvement network. <https://www.the-health-improvement-network.com/>.
- [2] Human mortality database. <https://www.mortality.org/>.
- [3] Nhs website - presentation of general practitioner.
- [4] Lignes directrices mortalité de la commission d'agrément. *Institut des Actuaire*s, 2006.
- [5] diabetes statistics. *The British Diabetic Association*, 2020.
- [6] Hald A. *A history of probability and statistics and their applications before 1750*. 1990.
- [7] Pal A. Gradient boosting trees for classification: A beginner's guide. *Medium*, 2020.
- [8] Silva A. Concordance index as an evaluation metric. *Medium*, 2019.
- [9] Hu B., Shao J., and Palta M. Pseudo-r² in logistic regression model. *Statistica Sinica*, 16:847–860, 07 2006.
- [10] Baker C. Obesity statistics. *UK Parliament*, 2022.
- [11] Feng W. Chen R., Ovbiagele B. Diabetes and stroke: Epidemiology, pathophysiology, pharmaceuticals and outcomes. *Am J Med Sci*, 2016.
- [12] Delwarde A. Denuit M. *Construction de tables de mortalité périodiques et prospectives*. 2005.
- [13] Ciecka J. E. Edmond halley's life table and its uses. 2008.
- [14] Hassan F. Lessons from history ii: The “thirty maidens of geneva” and the french revolution. *Finance Watch*, 2013.
- [15] Hannerz H. An extension of relational methods in mortality estimations. *Max Planck Institute for Demographic Research*, 2001.
- [16] Maurice Harbulot. Les emprunts viagers de l'ancien régime. *Journal de la société française de statistique*, 32:288–309, 1891.
- [17] Pryor D. Lee K. Harrell F., Califf R. and Rosati R. Evaluating the yield of medical test. *American Medicine Association*, 1982.
- [18] Pirenne J. *Histoire de la civilisation de l'Égypte ancienne*. 1963.

- [19] Béguin K. and Pradier P-C. Emprunts souverains et vulnérabilité financière de la monarchie d'Ancien Régime. Tout s'est-il joué sous Louis XIV ? September 2011.
- [20] Béguin K. *Financer la guerre au XVIIe siècle : la dette publique et les rentiers de l'absolutisme / Katia Béguin*. Époques. Champ Vallon, Seyssel (Ain), DL 2012.
- [21] Daxenberger L. Johan de witt - the first calculation on the valuation of life annuities. *Mac Tutor*, 2015.
- [22] Hlavac M. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. R package version 5.2.2.
- [23] Pascariu M. Modelling and forecasting mortality. *University of Southern Denmark*, 2018.
- [24] Davies P. Enhanced annuities - *Understand how enhanced annuities can give you a larger income if you've suffered from a medical condition*. May 2020.
- [25] Kim H. Park S-H. Park S-Y., Park J-E. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *The Korean Society of Radiology*, 2021.
- [26] Gupta S. Marmot M. Primatesta P., Falaschetti E. and Poulter N. Association between smoking and blood pressure. *AHA Journals*, 2001.
- [27] Cox D. R. Regression models and life tables. *Journal of the Royal Statistical Society*, 1972.
- [28] Velde F. R. and Weir D. R. The financial market and government debt policy in france, 1746–1793. *The Journal of Economic History*, 52(1):1–39, 1992.
- [29] Weir D. R. Tontines, public finance, and revolution in france and england, 1688–1789. *The Journal of Economic History*, 49(1):95–124, 1989.
- [30] D'Agostino R.B. and Nam B.H. Evaluation of the performance of survival analysis models: Discrimination and calibration measures. *Handbook of Statistics*, 2004.
- [31] Billiau S. From scratch: Permutation feature importance for ml interpretability. *Towards Data Science*, 2021.
- [32] Bucci S. Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification non-vie. application à la tarification à l'adresse. *Master Thesis for Institut des Actuaire*s, 2021.
- [33] Goetzmann W. and Rouwenhorst K., editors. *The Origins of Value: The Financial Innovations that Created Modern Capital Markets*. Oxford University Press, 2005.
- [34] Scanlon S. Windsor-Shellard B., Horton M. and Manders B. Adult smoking habits in the uk: 2019. *Office for National Statistics*, 2021.

Appendix

Appendix A : Descriptive statistics

Alcohol consumption intensity

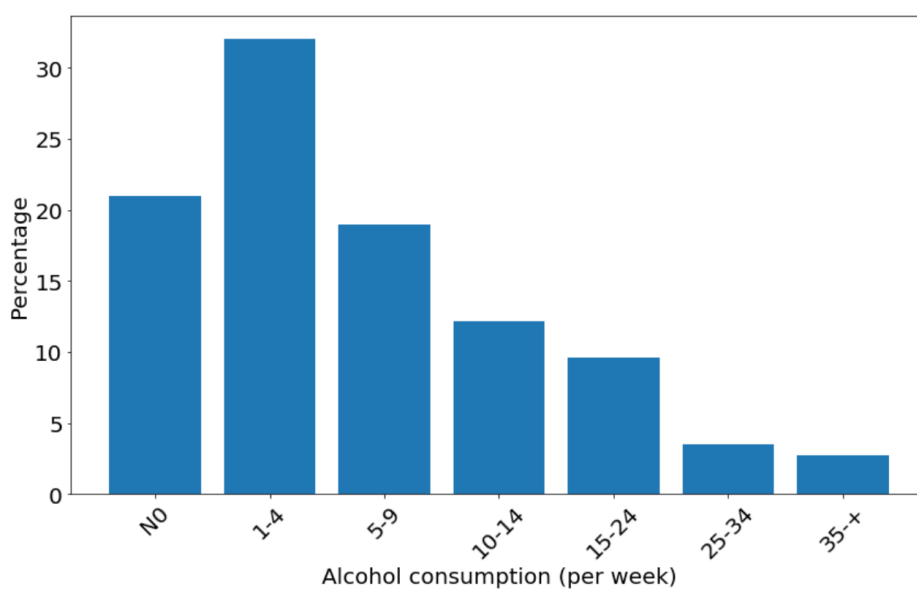


Figure 8.2: Alcohol consumption distribution (number of drinks per day)

Systolic blood pressure

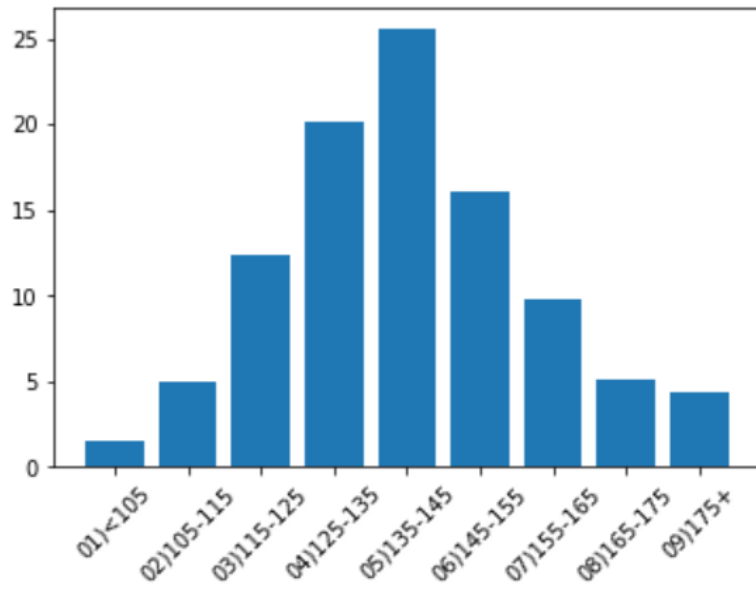


Figure 8.3: SBP modality frequency

Appendix B : GAM using B-splines

B-Splines

The term 'spline' refers to a craftsman's tool, a flexible thin strip of wood or metal, used to draft smooth curves. Several weights would be applied on various positions so the strip would bend according to their number and position. Spline functions uses the same concept for fitting the impact of an explanatory variable.

Splines are piece-wise polynomial functions. First, we divide our x-axis into uniform intervals using m knots. Those knots will define the intervals the polynomial functions will be fitted on. We will use cubic splines, which are splines with third degree polynomial functions.

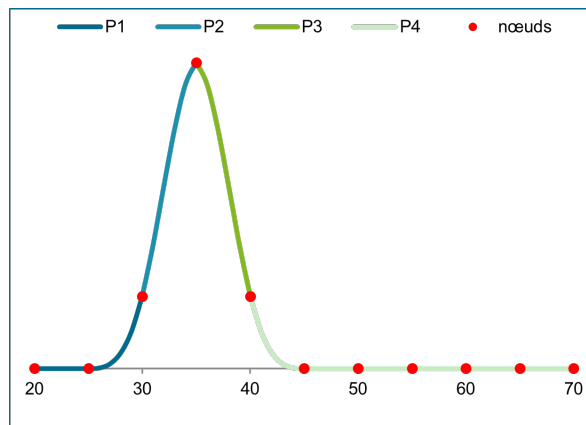


Figure 8.4: A cubic spline fitted on 5 knots

Construction of one B-spline

Each spline is fitted on n consecutive knots $[x_j, x_{j+n}]$ on which we interpolate cubic polynomials:

- Successive polynomials go through the junction points: $P_i(x_i) = y_i = P_{i-1}(x_i)$, $i = j, \dots, j + n - 1$,
- Successive polynomials have first derivatives (i.e. slopes) equal to the junction points: $P'_i(x_i) = P'_{i-1}(x_i)$, $i = j, \dots, j + n - 1$,
- The successive polynomials also have second derivatives (curvatures) equal to the junction points: $P''_i(x_i) = P''_{i-1}(x_i)$, $i = j, \dots, j + n - 1$,
- Two more constraints are added on the second derivatives of the extreme points: $P''_j(x_j) = P''_{j+n-1}(x_{j+n}) = 0$.

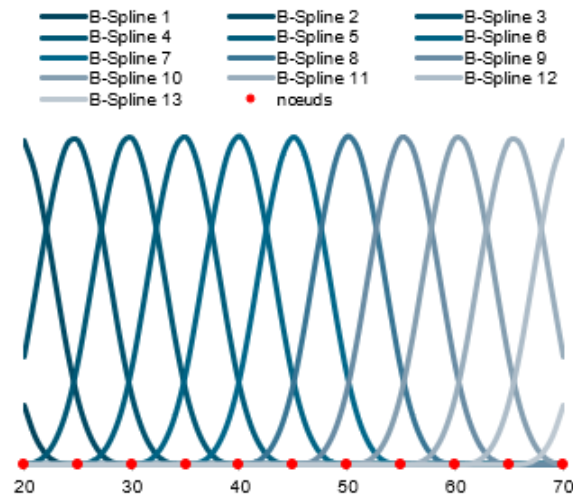
To get the spline represented in figure 8.4, we fix $n=5$ which leads to fit 4 polynomials. The spline is the sum of those cubic functions:

$$B_j(x) = P_{1,j}(x) + P_{2,j}(x) + P_{3,j}(x) + P_{4,j}(x)$$

There are 16 degrees of freedom (4 polynomials of degree 3) and 15 constraints (5 nodes with 3 constraints), so there is only one degree of freedom left. This last degree of freedom will be used to set the "high" of the spline to 1.

Field of B-splines

We create splines for each $j \in [0, m - n]$, which leads to a field of splines:



Field of cubic splines

Those splines will then be used as regressors on which we will fit the model

Fitting of the model

We define the following function:

$$g(x) = \sum_{j=0}^{m-n} \alpha_j B_j(x)$$

We then search the parameters α_j that minimise the objective function:

$$\sum_x (q_x - g(x))$$

So this leads to perform a classical linear regression where:

- the explanatory variables are the B-splines
- the explained variable is the death rate q_x

Using vector notations, the optimisation problem gives the solution:

$$\alpha = (G^t G)^{-1} G^t q$$

Appendix C : Analyses of the fitted splines

Now that the splines are fitted, we analyse the difference between the two splines (not including intercepts, nor the logit link function). We plot the two splines. In the second graph we show the difference between the male spline and the female spline:

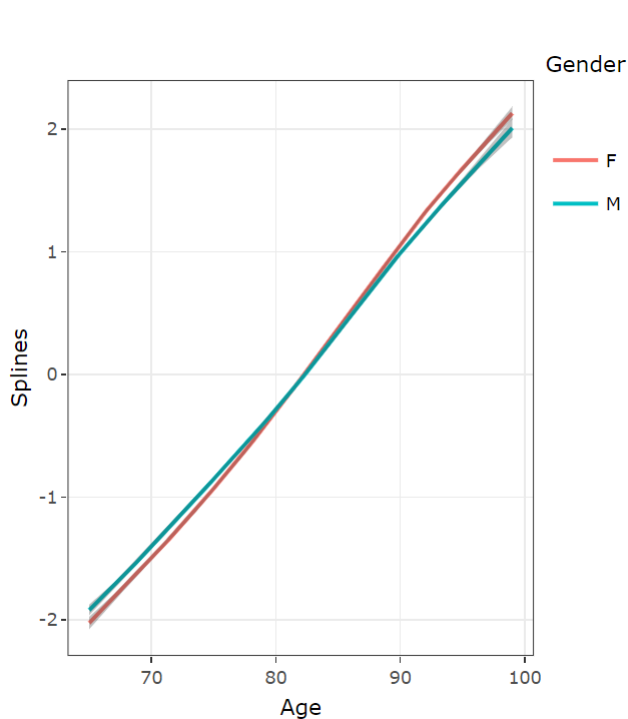


Figure 8.5: Fitted splines

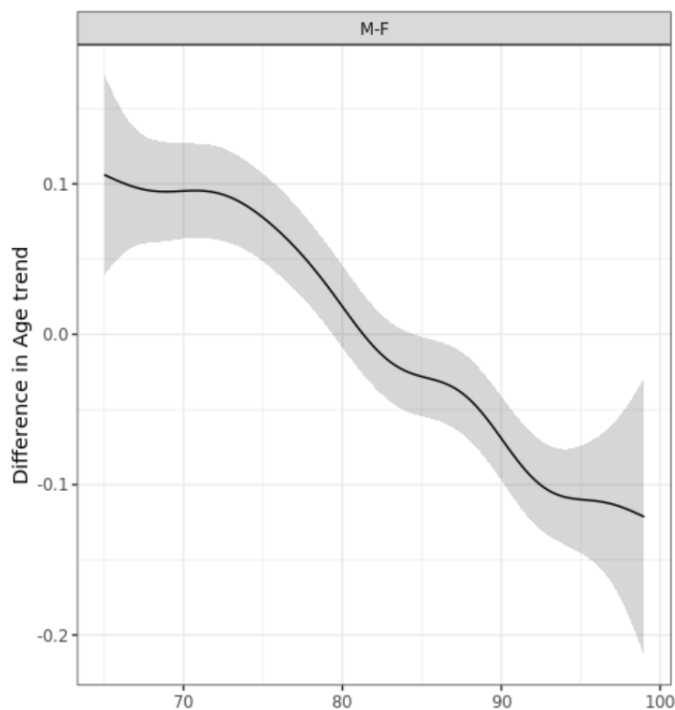


Figure 8.6: Splines difference between men and women

We can see that the impact of getting older is significantly different at some ages between men and women. This justifies the need to fit different splines for each gender as we did.

Appendix D : Mosaic codes

Mosaic number	Mosaic name	Description
1	Global Power Brokers	Wealthy and ambitious high flyers living in the very best urban
2	Voices of Authority	Influential thought leaders in comfortable and spacious city homes
3	Business Class	Business leaders approaching retirement, living in large family homes in the most prestigious residential suburbs
4	Serious Money	Families with considerable wealth living in large, exclusive detached houses where money is seemingly no object
5	Mid-Career Climbers	Families enjoying the fruits of career success in pleasant detached houses
6	Yesterday's Captains	Retired couples, enjoying pensions gained from successful careers, still living in the homes where they raised families
7	Distinctive Success	Successful business people, often self-made, living in large detached houses in semi-rural locations
8	Dormitory Villagers	Comfortably off families in spacious homes in pleasant settings but within easy reach of jobs
9	Escape to the Country	Families choosing to give their children a country lifestyle while commuting to urban jobs or running businesses from home
10	Parish Guardians	Couples approaching retirement age with ample income living in very pleasant rural locations
11	Squires Among Locals	Affluent families bringing city wealth into a countryside setting in order to enjoy quality of life
12	Country Loving Elders	Elders enjoying retirement on a comfortable pension in the countryside
13	Modern Agribusiness	Farmers and workers in agriculture or related industries, living at or close to their workplace
14	Farming Today	Older couples in more remote areas, earning a reasonable income from the land
15	Upland Struggle	Families in the remotest areas of the country, working hard in agriculture or tourism
16	Side Street Singles	Young singles in cramped rented town flats, with little disposable income
17	Jacks of All Trades	Blue collar workers and traders, serving the needs of small market towns
18	Hardworking Families	Married couples approaching retirement age, in not especially fashionable small town locations
19	Innate Conservatives	Pillars of local society who are chiefly recent retirees in low density estates on town fringes
20	Golden Retirement	People in their 60s and 70s with considerable assets behind them, living in their ideal houses for retirement
21	Bungalow Quietude	Elderly people owning their own bungalow and drawing a modest pension
22	Beachcombers	Pensioners with good incomes living in holiday areas often close to attractive coastal scenery
23	Balcony Downsizees	Elders, generally single, who have downsized to flats more suited to their income and capabilities
24	Garden Suburbia	Mid-life families with above average incomes living in the nicer middle ring suburbs of larger cities
25	Production Managers	Middle income married couples owning unpretentious semi-detached housing
26	Mid-Market Families	Families with many grown-up children still living at home, living in cheaper suburban semis
27	Shop Floor Affluence	Employees earning reasonable incomes, living with their families in relatively inexpensive semis in industrial towns and cities
28	Asian Attainment	Comfortable middle-aged families with school age and older children, predominantly from an Asian background
29	Footloose Managers	Well paid singles in small but relatively expensive housing in dormitory towns
30	Soccer Dads and Mums	Parents of school age children, owning large recently built detached houses with mortgages funded by their successful careers
31	Domestic Comfort	Families with high incomes derived from managerial positions and considerable property wealth in their suburban detached houses
32	Childcare Years	Young, well educated and well paid couples, either married or cohabiting, most of whom are starting families
33	Military Dependents	Servicemen and their families renting quarters from the Ministry of Defence
34	Buy-to-Let Territory	Young newcomers to the workforce, renting urban flats often from professional private landlords
35	Brownfield Pioneers	Young people living in affordable city housing, much of which is brownfield infill
36	Foot on the Ladder	Young singles and couples who have recently bought their first small house which consumes a large proportion of their income
37	First to Move In	People living in the most recently built, brand new housing
38	Settled Ex-Tenants	Older couples whose children have flown the nest working in low skilled occupations and living in ex-council housing
39	Choice Right to Buy	Middle aged couples, some with older children still at home, living in the more desirable ex-council areas
40	Legacy of Labour	Older families on low incomes living on council estates in areas where industry was once prevalent
41	Stressed Borrowers	Middle aged people renting or owning in council areas, many of whom are over-stretched with debt
42	Worn-Out Workers	Older workers employed in low skilled work or unemployed with low prospects
43	Streetwise Kids	Large young families with many single parents, often unemployed and claiming benefits, living on deprived council estates
44	New Parents in Need	Young parents, often single, bringing up young children in barely adequate council terraces facing considerable disadvantage
45	Small Block Singles	Disadvantaged young singles in small flats rented from the council in the midst of urban sprawl
46	Tenement Living	Singles with poor employment prospects renting very small flats in mid rise blocks from the council
47	Deprived View	Singles living in high rise council tower blocks who have very low incomes and high dependence on benefits
48	Multicultural Towers	Flat-dwellers from a wide range of ethnic backgrounds, renting mostly from the council in large purpose built blocks
49	Re-Housed Migrants	People from diverse ethnic backgrounds surviving in low standard small flats mostly rented from inner London councils
50	Pensioners in Blocks	Low income, older singles, renting small flats from the council with poor facilities
51	Sheltered Seniors	Elderly people, mostly single who are housed in specially built flats and supported on basic pensions
52	Meals on Wheels	Some of the oldest people in society who have reasonable pensions, living in accommodation where they receive appropriate care
53	Low Spending Elders	Low income elders in council bungalows and semis suited to their declining mobility, surviving on modest pensions
54	Clocking Off	Older couples close to retirement owning spacious semis who have earned reasonable incomes in skilled, industrial occupations
55	Backyard Regeneration	Singles and families in affordable but respectable terraces which for the young are a stepping stone to better things
56	Small Wage Owners	Owners living in inexpensive private terraces in a range of relatively low paid occupations
57	Back-to-Back Basics	Young sharers and couples with young children, starting out in low price, older terraces
58	Asian Identities	Traditional South Asian families owning relatively small terraces for their many family members
59	Low-Key Starters	Low income young singles in urban terraces offering cheap rents from either private landlords or the council
60	Global Fusion	Young working people living in metropolitan terraces from a wide variety of ethnic backgrounds
61	Convivial Homeowners	Well paid professional couples, often with children, choosing to live in diverse urban areas rather than the suburbs
62	Crash Pad Professionals	Young, well paid, mostly single professionals, who have chosen flats suitable for commuting to urban jobs
63	Urban Cool	Successful city dwellers owning or renting expensive flats in trendy inner urban locations
64	Bright Young Things	Well-educated young singles paying high rents to live in smart inner city apartments
65	Anti-Materialists	Sharers and singles, many on benefits, renting cheap bedsits and flats in town centres
66	University Fringe	A mix of students living alongside graduates who are starting out on careers whilst still enjoying the city student lifestyle
67	Study Buddies	Students living in halls of residence and vibrant but unloved private student housing

Figure 8.7: Description of mosaic classes

Appendix E : Socio-economic clusters of Mosaics

Cluster letter	Mosaic types in cluster	Cluster description
A	11,15,19,37,52,54,58,60,62,63,66	Young graduates and rural classes
B	2,3,5,6,21,24,25,28,29,61	Business class and retirement quietude
C	22,32,35,38,41,51,53,55	Family starters and young elders
D	34,43,44,45,46,47,65,67	Limited incomes, renters
E	4,7,10,12,14,18,20,26,27,30,31,36,49,56,57,64	Hard working classes
F	8,9,13,16,17,23,39,40,42,50,59	Low spending classes, usually in countryside areas
G	99	Retirement home

Figure 8.8: Cluster description of mosaics

Appendix F : GLM models

Switching from binomial to Poisson modelling

Due to truncation and in cases when there are too many variables to group the pseudo-observations, some additional difficulties arise when fitting a binomial model.

Indeed, working at the pseudo-observation level and with cases of exposure inferior to 1 for pseudo observation with death $D = 1$ (pseudo observation that could not be observed during their whole age year due to start and end dates of the pseudo observation), it is not feasible to fit a binomial model using traditional implemented packages.

The problem is to be able to take into account correctly those rows having an exposure $E < 1$ and a death status $D = 1$.

As explain in Chapter 5, to fit binomial models for mortality rate estimation, we consider as target variable $Y_i = \frac{D_i}{E_i}$ and as weight $w_i = E_i$. However, in our dataset, we have a significant number of truncated pseudo-observations due to our data format, resulting from deaths that occurred during the first half year and last half year of the study period. Those pseudo-observations would then have a target variable $Y > 1$, which blocks us when fitting the model.

A solution could be to limit the study period by putting aside the first half year and last half year. This would enable to limit the problem, as a very large part of pseudo-observation where a death occurs ($D = 1$) with exposure lower than 1 ($E < 1$) are coming from two half years. For the remaining few cases, one could force the exposure to 1 when death occurs.

However, this solution would imply to put aside a significant amount of data. We estimated that with this approach we would have to put aside around 42% of pseudo-observations, corresponding to 29% of total exposure.

Therefore, in order to bypass this technical difficulty, we decided to use Poisson model as approximation of Binomial model.

An analysis of the impact of this approximation has been done to measure its validity. To assess the impact, we compared predictions of a Binomial model and a Poisson model fitted on the same pseudo-observation dataset obtained by leaving the first and the last half years out of the study period. To insure that we can fit the Binomial model, we forced to 1 the exposures all the remaining problematic pseudo-observations.

Results indicate that the mean of absolute difference in mortality rate predictions of the Binomial q_x^B and Poisson q_x^P on the trained data set is of 0.0008, meaning that on average we have $q_x^P = q_x^B \pm 0.0008$. This result makes us conclude that the approximation has very limited impact on the prediction.

Analysis of form of impact of some numerical variables on mortality risk

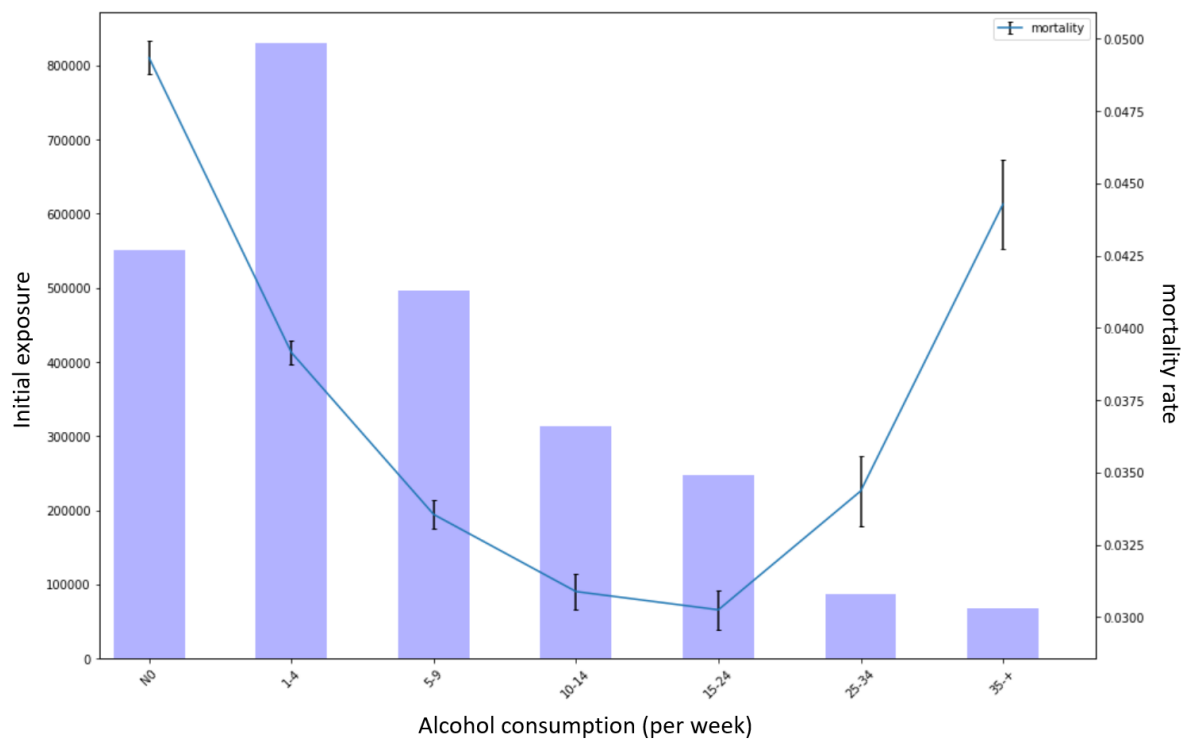


Figure 8.9: Raw death rates and initial exposure by alcohol consumption intensity

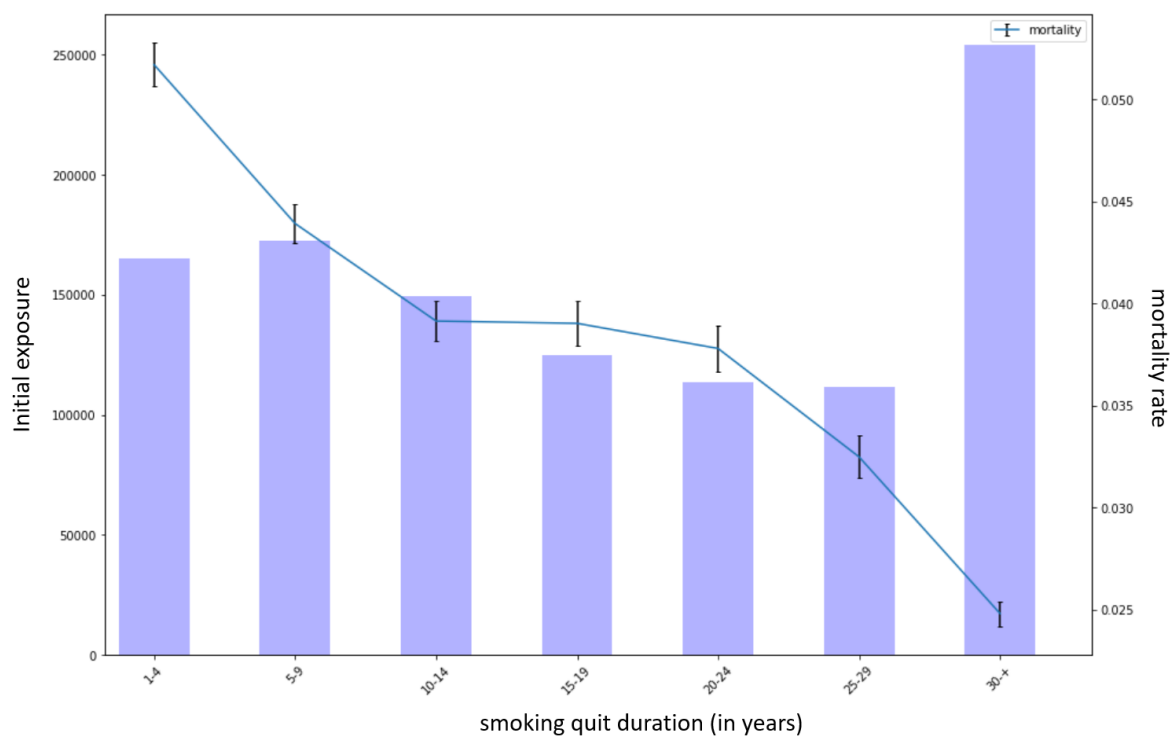


Figure 8.10: Raw death rates and initial exposure for ex-smokers by duration since quitting smoking

GLM variables used

Information	GLM	GLMnet_long	GLMnet_small
age	X	X	X
gender	X	X	X
cluster	X	X	X
smoker_status	X	X	X
smoke_intensity	X	X	X
quit_duration	X	X	X
sbp	X	X	X
bmi	X	X	X
alcohol_status	X	X	X
alcohol_intensity	X	X	X
fev1fvc	X	X	
hba1c	X	X	
diabetes	X	X	X
diabetes_type	X	X	X
duration_diabetes	X	X	X
insulin	X	X	X
hypertension	X	X	X
copd	X	X	X
duration_copd	X	X	
pneumoconiosis	X		
bronchiectasis	X		
stroke	X	X	X
stroke group	X	X	
mi	X	X	X
mi_status	X	X	
duration_mi	X	X	
tia	X		
retinopathy	X	X	
neuropathy	X	X	
amputation	X		
af	X	X	
heartvalve	X	X	
hf	X		X
cardiomyo	X	X	
othercv	X	X	
pvd	X	X	
aortic	X		
angina	X	X	

Table 8.1: List of used information for each GLM model

GLM fitted coefficients

For confidentiality purpose, we only show part of the results.

Output tables were produced using stargazer library [22].

Table 8.2: Results

	<i>Dependent variable:</i>		
		mortality_status	
	GLM	GLMnet_long	GLMnet_small
smokerEX	0.500***	0.577***	0.201***
smokerSM	0.824***	0.906***	0.652***
I(sbp ²)	0.005***		
I(sbp ³)	-0.0002	0.0003***	
I(bmi ²)	0.025***	0.024***	
bmi	-0.229***	-0.206***	-0.047***
pneumoconiosis	0.233**		
bronchiectasis	0.406***		
copd	1.080***	0.816***	
diabetes	0.582***	0.476***	0.548***
hypertension	0.476***	0.561***	0.908***
smokerEX:I(age ²)	0.002	-0.031***	
smokerSM:I(age ²)	-0.070***		
smokerEX:sbp	-0.007	-0.010**	
smokerSM:sbp	0.002	-0.007	
age:sbp	-0.002	-0.003	0.005***
sbp:male	-0.014**		
I(age ²):I(bmi ²)	-0.001***	-0.001***	0.001***
XXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXX	XXXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXX	XXXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXX	XXXXXX	XXXXXX
bmi:alcoholDR	-0.028***	-0.025***	-0.049***
smokerEX:I(quitdur ²)	-0.035***	-0.009***	-0.010***
alcoholDR:alc_int	0.035***	0.055***	0.060***
age:pneumoconiosis	-0.045***		
age:bronchiectasis	-0.015*		
bronchiectasis:copd	-0.241*		
copd:dur_copd	0.005*	0.004	
stroke:diabetes	0.068	0.179***	0.209***
XXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXX	XXXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXX	XXXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXX	XXXXXX	XXXXXX	XXXXXX
cluster_B:diabetes	0.059	0.055	0.043
cluster_C:diabetes	-0.169***	-0.138***	-0.131**
cluster_D:diabetes	-0.340***	-0.288***	-0.317***

cluster_E:diabetes	0.018	0.024	0.027
cluster_F:diabetes	-0.183***	-0.148***	-0.134***
cluster_G:diabetes	-0.581***	-0.560***	-0.574***
diabetes:amputation	0.552***		
age:hf	0.008***	0.009***	0.010***
hf:heartvalve	0.264**		
hf:cardiomyo	0.151		
sbp:hf	0.004	-0.035*	
I(bmi^2):mi	0.003		
age:angina	0.002		
smokerEX:angina	-0.014		
smokerSM:angina	0.013		
af:angina	0.268**		
pvd:angina	0.322**	0.372***	
I(sbp^2):angina	0.0004		
age:af	0.006***	0.005***	
smokerEX:male:quitdur	0.005		
smokerEX:age:quitdur	-0.001***		
smokerEX:age:I(quitdur^2)	0.0005***		
XXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXX
smokerEX:male:alcoholDR	-0.087***		
smokerSM:male:alcoholDR	-0.131***		
male:alcoholDR:alc_int	0.020**		
age:male:copd	-0.006	-0.004***	
I(age^2):male:copd	0.00004		0.0001***
smokerEX:copd:fev1fcv	-0.013	0.018**	
smokerSM:copd:fev1fcv	-0.093		
age:copd:fev1fcv	-0.031***	-0.035***	
I(age^2):copd:fev1fcv	0.002		
smokerEX:copd:I(fev1fcv^2)	0.002		
smokerSM:copd:I(fev1fcv^2)	0.009		
I(age^2):copd:I(fev1fcv^2)	-0.001		
age:copd:diabetes	-0.003***	-0.003***	0.002***
diabetes:diabType2:dur_diab	0.066***	0.009***	
male:diabetes:diabType2	-0.139		
XXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXX
XXXXXXXXXXXXXXXXXXXXXXX	XXXXX	XXXXX	XXXXXX
I(age^2):male:diabetes	-0.019		-0.102***
age:male:diabetes	-0.016	-0.051*	
diabType2:dur_diab:insulin	0.017***	0.020***	0.028***
age:diabType2:dur_diab	-0.001***		0.0001**
smokerEX:male:alcoholDR:alc_int	-0.006	-0.015***	

smokerSM:male:alcoholDR:alc_int	0.012	0.006	
smokerEX:age:diabetes:diabType2	-0.001***	-0.001	-0.0004
smokerSM:age:diabetes:diabType2	-0.003***		
diabetes:insulin:retino:neuro	0.426***	0.484***	
Observations	1,979,753	1,979,753	1,979,753
Log Likelihood	-229,838.200	-230,296.100	-231,539.400
Akaike Inf. Crit.	459,962.500	460,744.300	463,160.800

Note:

*p<0.1; **p<0.05; ***p<0.01

Partial dependence plots

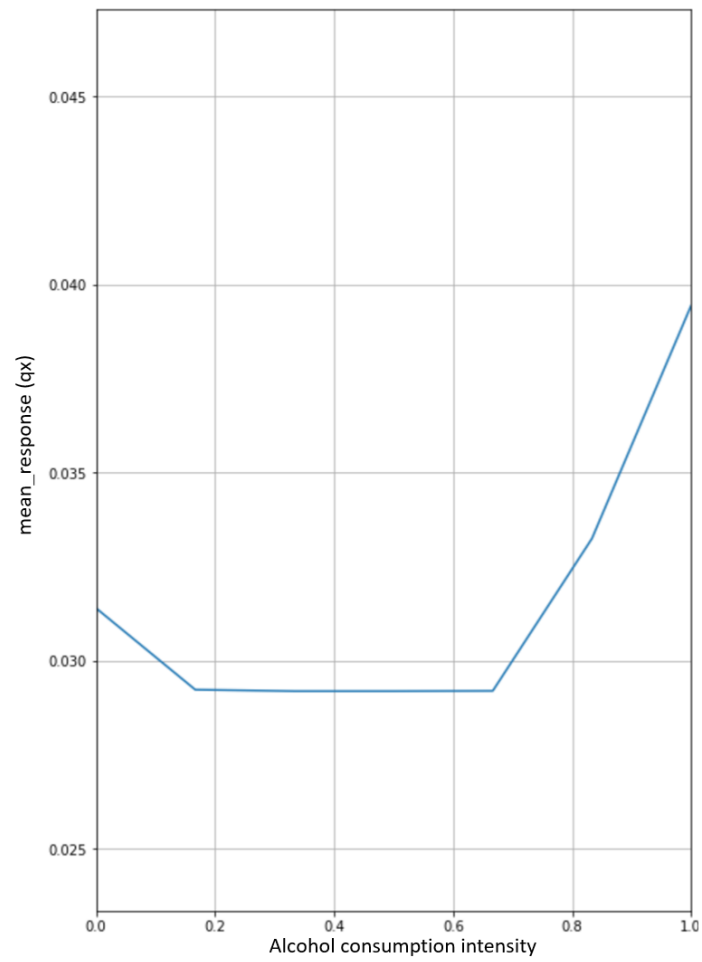


Figure 8.11: Alcohol intensity

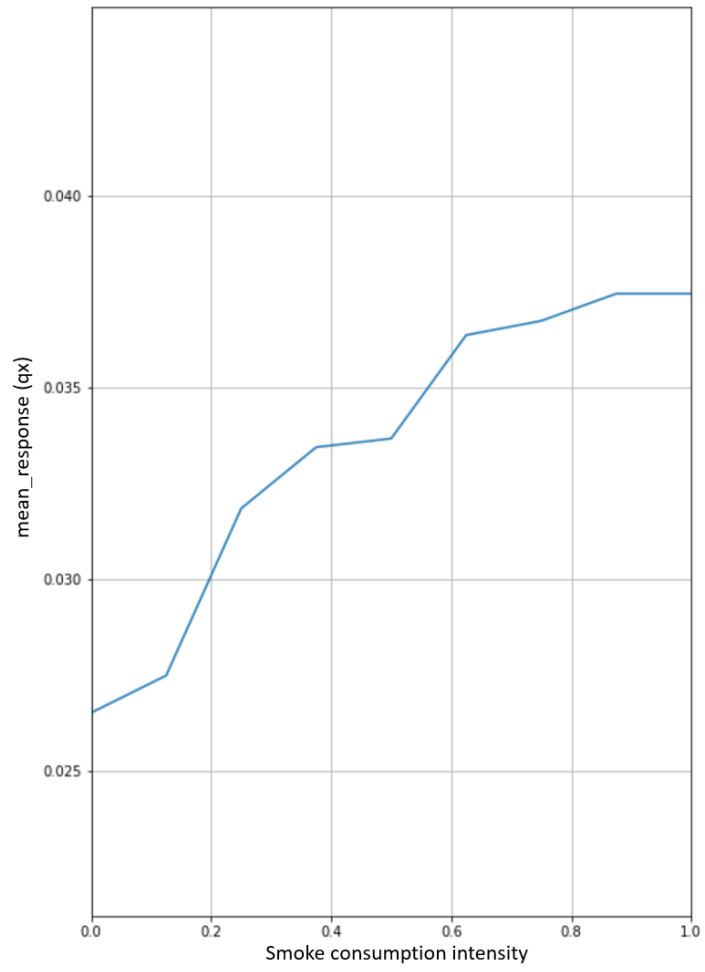


Figure 8.12: Smoking intensity

Appendix G : Sobol indices estimation

We show here two trials of Sobol indices estimation of interaction of 2 variables. Only indices which were statistically significant are shown :

var1	var2	bias	std.	error	min. c.i.	max. c.i.
<i>Trial 1</i>						
age	ms	0.011250171	-0.0024776936	0.00462799	0.006783727	0.02385711
age	alc_intensity	0.006083213	-0.0005622116	0.003544078	0.0007676575	0.01457626
FEV1FVC	ms	0.006139946	-0.0015600773	0.004750803	0.000402484	0.01878863
diabetes	ms	0.003609153	-0.0009533923	0.003162622	0.0005824086	0.01197714
insulin	bmi	0.005593142	-0.0013184237	0.003869996	0.001454393	0.01720457
insulin	Gender_Male	0.005740616	-0.000723981	0.003182367	0.002080052	0.01446672
hf	ms	0.003331235	-0.0006955901	0.002909534	0.00008922588	0.01021082
ms	ra	0.004802342	-0.0014393711	0.003564213	0.001643779	0.01448383
ms	tia	0.003968219	-0.0012045978	0.003106743	0.0009808656	0.01208533
ms	cluster_F	0.003752597	-0.0015096706	0.003386932	0.00111108	0.01364524
ms	bmi	0.008375782	-0.0024159424	0.003990414	0.005432473	0.01988247
ms	Gender_Male	0.007136052	-0.0021418646	0.003593263	0.003784697	0.01732323
<i>Trial 2</i>						
strokegroup	FEV1FVC	0.003728423	-4.003752e-04	0.002369775	0.001519279	0.007863623
FEV1FVC	insulin	0.005071523	-1.239178e-03	0.003221828	0.002166893	0.010398840
FEV1FVC	af	0.004900609	5.095821e-05	0.002745056	0.001170501	0.008593657
FEV1FVC	ms	0.003432635	-1.146918e-03	0.002388513	0.001137339	0.007515480
FEV1FVC	ra	0.003754826	-1.046158e-03	0.002087583	0.002003515	0.008226078
insulin	alc_intensity	0.005067548	-8.517901e-04	0.002750398	0.001890117	0.010316258
strokegroup	ms	0.003422084	5.706487e-05	0.001614264	0.0002472710	0.006634395
FEV1FVC	angina	0.003117382	3.437843e-05	0.001548871	0.0001537934	0.006113801
FEV1FVC	af	0.004900609	1.461408e-05	0.001962592	0.0009180545	0.008601897

Figure 8.13: Residual Life expectancy at age 65 predictions per model

As we can see, the two trials gave very different outputs. Moreover, some interactions of variables shown as relevant by the estimated Sobol indices were sometimes doubtful.

Numerous attempts were run in order to stabilize results. However we concluded that due to lack of computational power, results were not conclusive.

Appendix H : Life expectancy prediction example

We provide examples of life expectancy predictions for a reference population when tuning some parameters such as smoking status, alcohol status, BMI, SBP, diabetes and COPD:

gender	cluster	Smoker status	smoke_int	quidur	alcohol_int	BMI	SBP	COPD	Diabetes	Insulin	...	GAM	GBM	GLM	GLMnet_small	GLMnet_long
F	E	EX	4	25	1	35	145	0	Type2	0	...	21,86	19,45	19,45	19,36	19,43
F	E	NS	0	0	1	20	115	0	Type1	1	...	21,86	19,14	20,61	19,44	20,89
F	E	NS	0	0	5	30	135	0	Type2	0	...	21,86	18,06	18,15	19,12	18,07
F	E	EX	4	20	0	22.5	115	0	Type2	1	...	21,86	18,03	18,44	17,28	18,15
F	E	EX	15	25	1	30	125	0	Type2	1	...	21,86	18,73	19,55	18,79	19,50
F	E	NS	0	0	10	20	105	0	Type1	1	...	21,86	19,21	20,27	18,93	20,39
M	E	SM	15	0	1	27.5	115	0	Type2	0	...	19,78	15,67	16,68	14,17	15,13
M	E	EX	20	25	15	25	125	0	Type2	1	...	19,78	17,11	17,75	15,92	17,54
M	E	NS	0	0	0	25	125	0	Type2	0	...	19,78	18,63	18,76	18,31	18,52
M	E	NS	0	0	1	30	135	0	Type2	1	...	19,78	15,61	16,92	18,83	17,49
F	E	EX	10	10	15	27.5	115	1	Type2	0	...	21,86	16,92	13,64	15,77	12,31
F	E	NS	0	0	1	22.5	135	0	None	0	...	21,86	23,84	24,56	24,09	24,65
F	E	EX	20	10	1	30	135	0	Type2	1	...	21,86	14,75	12,52	14,02	12,24
F	B	SM	3	0	5	22.5	135	0	None	0	...	23,55	23,07	22,96	20,95	22,13
F	B	SM	15	0	0	25	115	1	Type2	1	...	23,55	13,90	10,02	11,40	8,75

Figure 8.14: Residual Life expectancy at age 65 predictions per model

Appendix I : A/E of several comorbidities

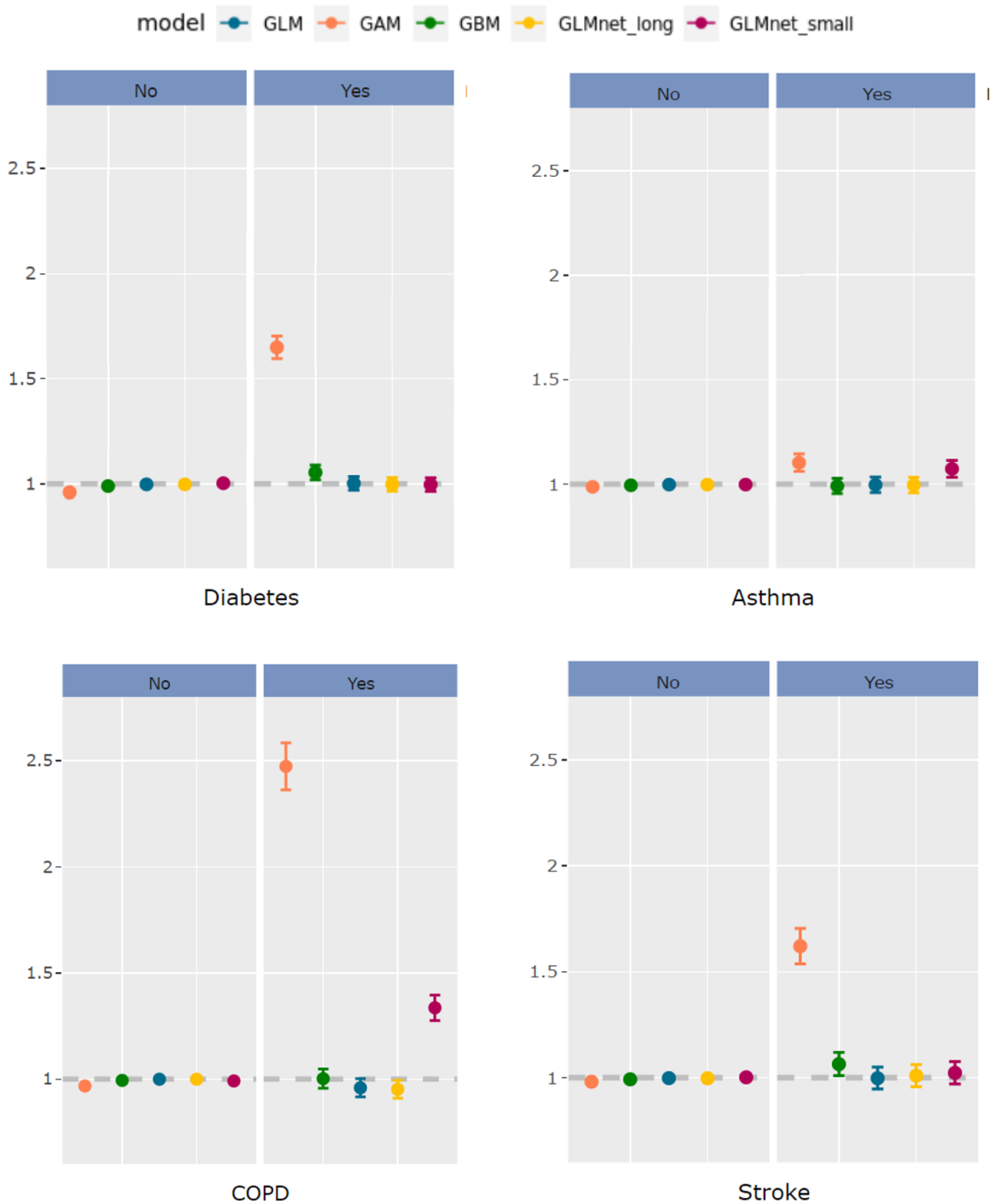


Figure 8.15: A/E of four comorbidities

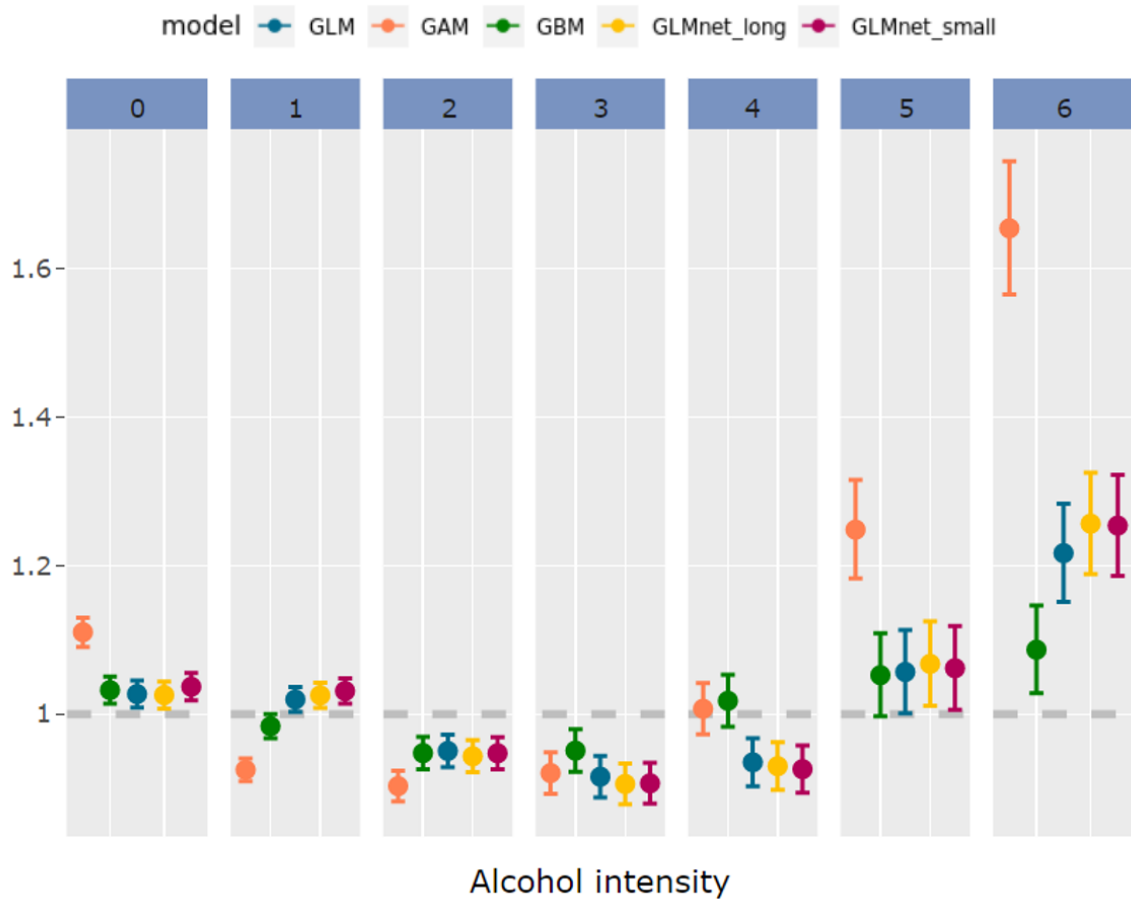


Figure 8.16: A/E of alcohol intensity per group

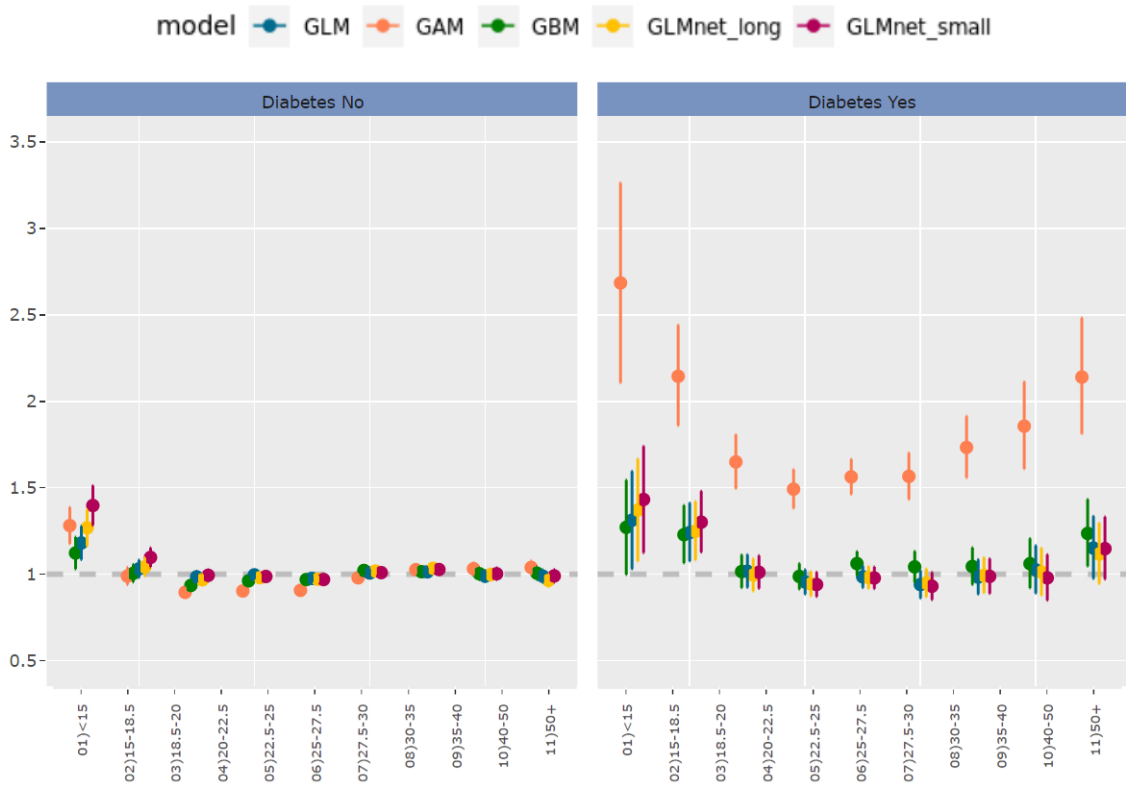


Figure 8.17: A/E of Systolic blood pressure depending on diabetes status