Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires

le

Par : Simon SAVOYE

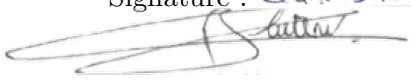
Titre : Développement et mise en application d'un algorithme de tarification sur un modèle de coût/fréquence et confrontation avec le GLM

Confidentialité : ☐*Non* ☑ *Oui* (*Durée* :☐ 1 an ☑ 2*ans*)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : ADDACTIS France
Signature : Colin BLATTNER, CEO



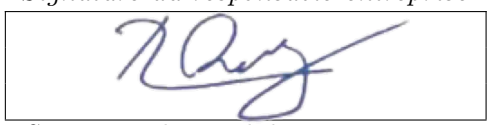
Directeur de Mémoire en entreprise :
Nom : Xuan-Quang DO
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

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

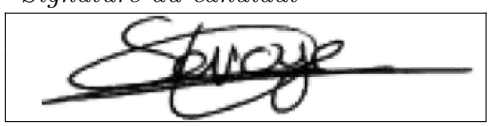
Secrétariat :

Signature du responsable entreprise



Bibliothèque :

Signature du candidat



Résumé

Les problématiques de tarification ont toujours été au coeur des travaux en actuariat. L'évaluation du risque engendré par les contrats d'assurance est, aujourd'hui plus que jamais, essentielle. La crise de la COVID-19 en est un révélateur, selon Alban de Mailly Nesle, directeur des risques et des investissements d'AXA, "elle illustre un nouveau phénomène majeur [...], l'interconnexion des risques". L'uniformisation des modes de vie et la mondialisation jouent en effet un rôle clé dans la propagation des risques. Ceux-ci se complexifient et les modèles qui les représentent doivent naturellement faire de même. L'abondance des données collectées et les améliorations des performances technologiques permettent alors de suivre cette tendance de sophistication.

Bien que les modèles linéaires généralisés, dit GLM (Generelised Linear Model), se sont depuis longtemps imposés comme la méthode standard, notamment parce que la modélisation en coût/fréquence incite à se focaliser sur l'espérance du risque, il n'en demeure pas moins qu'ils présentent des limites évidentes.
Des outils d'apprentissage plus complexes peuvent les pallier mais leur manque d'interprétabilité demeure un frein à leur application dans le monde de l'assurance.

On va donc présenter, à travers ce mémoire, un nouveau modèle plus complexe que le GLM mais qui reste interprétable. On proposera de suivre le cheminement de sa conception étape par étape.
Les premières parties de ce mémoire porteront alors sur les fondements mathématiques sur lesquels se repose notre méthode. La création de l'algorithme permettant l'implémentation de nos travaux ne sera que légèrement abordée, elle se reposera sur une méthode de descente de gradient proximale. On s'attardera d'avantage sur l'application de notre modèle à des données concrètes afin de confronter nos performances à celles du GLM, que l'on espère surpasser. Les données utilisées seront issues de la base de garantie dégât des eaux en assurance MRH (Multi-Risque habitation) d'une grande compagnie d'assurance. Les résultats seront présentés dans le dernier chapitre.

Mots-clés : *Tarification*, *Proximal Gradient Descent*, *GLM.*

Abstract

Pricing issues have always been at the heart of actuarial work. The evaluation of the risk generated by insurance contracts is, today more than ever, essential. According to Alban de Mailly Nesle, AXA's Director of Risk and Investments, the COVID-19 crisis is a revealing example of this, "it illustrates a major new phenomenon [...], the interconnection of risks". Standardization of lifestyles and globalization are playing a key role in the spread of risks. They are becoming more complex and the models that represent them must naturally do the same. The abundance of data collected and improvements in technological performance make it possible to follow this trend of sophistication.

Although generalized linear models, known as GLM (Generelised Linear Model), have long since established themselves as the standard method, in particular because cost-frequency modelling encourages a focus on risk expectation, they nevertheless have obvious limitations.
More complex learning tools can compensate for this, but their lack of interpretability remains an obstacle to their application in the insurance world.

We will therefore present, through this paper, a new model that is more complex than the GLM but which remains interpretable. We will propose to follow the development of its conception step by step.
The first parts of this paper will then focus on the mathematical foundations on which our method is based. The creation of the algorithm allowing the implementation of our work will be only slightly discussed but it relies on a proximal gradient descent method. We will focus more on the application of our model to concrete data in order to compare our performances with those of GLM, which we hope to surpass. The data used will be taken from one of the top insurance compagnie HRM (Multi-risk home insurance) water damage guarantee database. The results will be presented in the last chapter.

Keywords : *Pricing*, *Proximal Gradient Descent*, *GLM*.

Note de Synthèse

La justesse de la tarification est un enjeu crucial du monde assurantiel. C'est en effet un des marqueurs de différentiation entre les multiples acteurs sur le plan de la pérennité. Le contexte de forte tension concurrentielle ancré dans le secteur non vie exerce une friction sur les tarifs, qui est mécaniquement associée à une baisse des chargements de sécurité. Le marché est donc plus sensibles aux aléas, comme en témoigne l'impact de la crise de la COVID-19 qui, selon Standards Poor's, s'élèvera à 50 milliards de dollars pour le marché mondial de l'assurance. La maîtrise approfondie des risques est donc plus que jamais une composante majeure de stabilité.

C'est au regard de ce constat que nous allons proposer un algorithme innovant de tarification en assurance IARD qui va tenter de pallier les défauts de la méthodologie actuellement en place. Celui-ci devra être à la fois interprétable et automatique, notion à laquelle on donnera un sens plus précis par la suite.

L'approche la plus courante en pricing dommage consiste en la décomposition de la prime pure en deux parties. Ainsi, les modèles actuels cherchent à prédire séparément la fréquence d'apparition d'un sinistre et le coût moyen associé. Ces deux composantes sont, dans une très large majorité des cas, estimées via un modèle GLM (Generalised Linear Model ou Modèle Linéaire Généralisé). Il se fonde sur la représentation suivante :

$$g(\mathbb{E}[Y|X]) = X\beta,$$

où Y représente la variable cible à savoir la fréquence ou le coût, X la matrice des covariables, β l'inconnue que l'on va approcher par optimisation et g une fonction de lien. C'est cet algorithme que l'on va chercher à surpasser. On remarque que celui-ci peut uniquement modéliser des dépendances linéaires entre les features et la variable d'intérêt. Son application entraîne donc une déformation des dépendances réelles, empêchant la prise en compte de forme plus complexes. Deux solutions principales, basées sur le GLM, permettent de solutionner ce problème. La première vise à ajouter des variables avant de relancer un GLM. L'idée est d'utiliser une approximation polynomiale de la dépendance en ajoutant les variables $X_{,j}^2$, $X_{,j}^3$, ..., $X_{,j}^k$ afin d'être en mesure de trouver le polynôme d'ordre k qui s'adapte au mieux à la dépendance de la j-ième variable. Cet outil est difficilement automatisable, il nécessite soit une intervention humaine pour définir variable par variable le degré adéquat du polynôme, soit un temps de calcul suffisamment long pour tester une multitude de paramètre.
La seconde est le modèle GAM (Generalised Additive Model) qui consiste à ajuster une fonction de dépendance à chaque variable $X_{,j}$ conduisant à la modélisation

$$g(\mathbb{E}[Y|X]) = \beta_0 + f_1(X_{,1}) + ... + f_p(X_{,p}).$$

Cette méthode peut être implémentée automatiquement et corrige donc les difficultés pratiques de la précédente. Elle ne permet cependant pas de réaliser une sélection de variables, étape qui reste cruciale en assurance. En effet, bien que l'abondance de données permette une meilleure précision dans l'estimation, il sera nécessaire d'interroger l'assuré sur chacune des variables conservées. La tendance

est néanmoins à l'accélération du processus de souscription, ce qui incite à réduire drastiquement le nombre de variables nécessaires. La pénalité LASSO des méthodes GLM est donc un atout indispensable à un algorithme pertinent, permettant de réaliser automatiquement la sélection de variables. L'objectif de ce mémoire est donc de présenter une solution automatique aux limites du GLM, tout en restant interprétable, ce qui sera assuré par le fait que nos travaux reposent en grande partie sur les même concepts que ce dernier.

Nos travaux s'appuieront sur un jeu de données provenant du portefeuille d'un des plus gros assureurs français doté initialement de 89 variables et d'environ un million de lignes (donc d'individus). Celui-ci dispose d'un grand nombre de variables géospatiales mais l'on décide de travailler à la maille départementale uniquement car on sera amené à discrétiser les données par la suite, on souhaite donc garantir une exposition suffisante sur chaque modalités. On se restreindra uniquement à la garantie dégât des eaux en MRH, nos deux variables cibles seront la fréquence de sinistre ainsi que le coût moyen d'un sinistre. La première prendra en compte l'exposition comme variable offset, c'est à dire une variable dont le coefficient est fixé à 1 afin de décrire la fréquence de sinistre sur une base annuelle. La prédiction du coût moyen ne se fera que sur la base des individus ayant subi un sinistre, ce qui représente une faible proportion de la base totale (moins de 10%). On utilisera 80% des lignes comme base d'apprentissage pour entraîner notre modèle, les 20% restantes serviront à mesurer la performance de notre modélisation afin de la comparer à celles des méthodes traditionnelles. Il convient donc de décrire les spécificités de notre modélisation avant de présenter les résultats de son application sur nos données.

La limite principale des méthodes de GLM résidant dans la contrainte de linéarité imposée par le modèle, qui n'est en pratique pas toujours vérifié, notre premier objectif est de s'en affranchir. Dans ce sens, la première particularité de notre modélisation est de discrétiser l'intégralité de notre base de données, d'une manière semblable aux transformations réalisées sur les variables qualitatives. Bien entendue les variables continues, ou du moins les variables avec un nombre important de modalités, donneraient lieu à un trop grand nombre de nouvelles variables à l'issue de ce procédé. Il convient donc de se restreindre à rassembler les modalités proches (au sens de la distance euclidienne pour les variables numériques) au sein d'un même groupe. Pour éviter les disparités de représentation au sein des groupes on choisit de les rassembler par quantiles. Néanmoins si la variable est discrète et présente un faible nombre de modalités (on peut notamment penser à une variable type *Nombre d'enfants*) on créera simplement une nouvelle variable pour chacune d'entre elles. De ce fait, appliquer un modèle linéaire sur cette base associera à chaque modalité un coefficient, ce qui permettra naturellement de reproduire un spectre de forme de dépendance bien plus large qu'auparavant.

On introduit néanmoins avec ce procédé un biais de sur-apprentissage car tout coefficient $\beta_{k,j}$, associé à la j-ième modalité de la variable k sera estimé via un nombre bien plus restreint d'observations. On va donc chercher à profiter de l'information portée par les modalités proches de celle qui sert à l'estimation du coefficient $\beta_{k,j}$. Pour ce faire on va introduire une pénalisation visant à restreindre l'écart entre les coefficients associés à des modalités proches d'une unique variable. De cette façon on utilise le maximum d'information pertinente à notre disposition tout en évitant le sur-apprentissage qui pourrait conduire à un manque d'interprétabilité. Il est par exemple souhaitable que $\beta_{k,1}$ ne soit pas trop différent de $\beta_{k,2}$ pour prohiber les écarts de tarifs importants pour deux individus quasiment similaires.

Enfin, on souhaite ajouter une pénalité LASSO pour les raisons évoquées au préalable. On tra-

vaillera donc avec le problème d'optimisation suivant

$$\hat{\beta} = \arg\min_{\beta}\{-\mathcal{L}(\beta) + \lambda_1||\beta||_1 \ + \lambda_2 \sum_{k=1}^{k=m} \sum_{j=2}^{p_k-2} (\beta_{j-1}^k + \beta_{j+1}^k - 2\beta_j^k)^2\}. \quad (1)$$

$\mathcal{L}(\beta)$ est la log-vraisemblance des observations, on se place donc dans un cadre classique de maximisation de vraisemblance. On implémentera une loi de Poisson pour modéliser la fréquence ainsi qu'une loi Gamma pour le coût. On retrouve également les deux pénalités évoquées précédemment associées respectivement à un poids λ_1 et λ_2 que l'on choisira par validation croisée. Pour la seconde pénalité, m représente le nombre de variables quantitatives sujettes à la dichotomisation et p_k le nombre de modalités ou de classes associées.

La modélisation choisie est donc entièrement décrite à travers cette équation. Afin de la rendre utilisable en pratique, il est nécessaire de pouvoir calculer $\hat{\beta}$ et donc d'être en mesure de minimiser la fonction cible. On cherchera en réalité a approcher cette solution par des méthodes itératives. L'approche classique de descente de gradient n'étant pas adaptée à cause de la présence de la norme L^1, notre choix se porte sur la méthode de gradient proximal qui permet de résoudre des problèmes de minimisation pour des fonctions non différentiables. Elle est entre autres fréquemment utilisée dans la résolution d'un GLM LASSO - le lecteur pourra se référer à Parikh et Boyd, 2013 pour une explication détaillée de son fonctionnement.
Elle se repose principalement sur la fonction proximale qui se définit par

$$\text{prox}_{tf}(v) = \arg\min_{x}(f(x) + \tfrac{1}{2t}||x - v||_2^2).$$

Il s'agit du point d'arbitrage entre la minimisation de f et la proximité à l'antécédent v.

On décrit succinctement les étapes algorithmiques permettant d'obtenir une approximation de $\hat{\beta}$ qui se reposent sur les travaux présentés ici Parikh et Boyd, 2013.

- Soit à l'itération k les paramètres $\beta_k, t := t_{k-1}$ et un facteur de rétrécissement τ.
- On répète la structure suivante jusqu'à la condition d'arrêt :
  1. On pose $z := \text{prox}_{tg}(\beta_k - t\nabla f(\beta_k))$.
  2. On pose la condition d'arrêt $f(z) < \hat{f}_t(z, \beta_k)$.
  3. On diminue le pas tel que $t := \tau t$.
- On retourne $t_k := t$ et $\beta_{k+1} := z$

Ici $\hat{f}_t(x, y) = f(y) + \nabla f(y)^T(x - y) + \frac{1}{2t}||x - y||_2^2$, et $\nabla f(\beta_k)$ représente un sous-gradient de$f$enβ_k.

Il s'agit de l'algorithme de descente de gradient proximale à pas adaptatif pour lequel on peut implémenter une version accélérée nommée FISTA (Fast Iterative Soft-Thresholding Algorithm).

Une fois implémenté, il est possible d'appliquer notre modèle aux données présentées précédemment. Pour cela, on réalise une validation croisée pour optimiser les valeurs des deux hyperparamètres. On

utilise alors une validation croisée 5-folds. Une fois le couple $(\lambda_1^*, \lambda_2^*)$ déterminé pour la modélisation du coût moyen et également de la fréquence, on décide de comparer les performances de notre méthode à celle du GLM LASSO, méthode la plus courante en tarification.

On établit alors le protocole de comparabilité suivant. On sépare notre base de données en deux, une base d'apprentissage comprenant 80% des individus, et une base de test avec les individus restant. Cette séparation est identique pour tous les modèles. On évaluera les performances uniquement sur les hyperparamètres optimaux. Ceux-ci seront calculés par validation croisée sur la base d'apprentissage avec la déviance pour métrique de décision. On calculera ensuite trois indicateurs, à savoir la déviance, la RMSE (Root Mean Square Error) et la MAE (Mean Absolute Error) à la fois sur la base d'apprentissage et sur celle de test. On sera alors en mesure d'estimer l'impact du sur-apprentissage. En effet si ces métriques présentent des écarts trop importants entre le calcul réalisé sur les données d'apprentissages et celles de test, on conclura que le modèle sur-apprend trop et n'est adapté qu'aux données d'entrées.
Pour conclure, on comparera la déviance entre le GLM LASSO et notre algorithme, notamment sur la base de test, pour déterminer lequel est le plus performant. C'est l'indicateur de comparaison le plus pertinent car il s'agit d'une généralisation du calcul de la somme des carrés résiduels dans le cas où la méthode d'optimisation maximise la vraisemblance. Elle se calcule par

$$D(Y, \hat{\mu}) = -2(\mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta^*)).$$

La seconde vraisemblance est calculée en remplaçant la prédiction $g^{-1}(X\beta)$ par la vraie valeur de Y. C'est un candidat de métrique de comparaison idéal puisque l'écart de déviance entre les deux modèles suit une loi du χ^2 sous l'hypothèse que le second modèle est le vrai, ce qui permet de sélectionner le meilleur modèle.
Plus la déviance est faible, meilleur sera le modèle (par construction c'est une quantité toujours positive).

On présente alors les résultats de cette comparaison sous la forme de tableaux.

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	6695.62	2245.60	1191.93
Echantillon de test	1743.30	2213.48	1178.06

Table 1: Résultats d'un GLM LASSO sur le modèle de coût moyen

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1397.60	0.11	0.02
Echantillon de test	442.29	0.12	0.02

Table 2: Résultats d'un GLM LASSO sur le modèle de fréquence appliqué à 20 000 individus

Ces deux premiers tableaux représentent l'évaluation des performances du GLM LASSO sur les données. En ce qui concerne le modèle de fréquence, l'exposition a, dans ce modèle et au sein de notre algorithme, été implémentée en tant que variable offset. Au vu des résultats, le sur-apprentissage semble faible car les deux dernières métriques sont presque similaires d'une base à l'autre. La déviance varie plus mais c'est un effet mécanique, il ne s'agit pas d'une moyenne mais d'une somme (de log-vraisemblance), c'est donc une quantité qui augmente avec le nombre de données. Pour comparer à

taille égal on peut grossièrement multiplier la déviance issue de la base de test par 4 (ces données représentent 20% de la base totales et sont donc 4 fois moins nombreuses que celles utilisées pour la base d'apprentissage). Les ordres de grandeur ainsi obtenus laissent croire que le modèle ne tombe pas dans le sur-apprentissage.

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	6642.00	2240.91	1182.49
Echantillon de test	1743.28	2212.96	1173.01

Table 3: Résultats de notre modélisation sur le modèle de coût moyen

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1549.44	0.11	0.02
Echantillon de test	332.22	0.10	0.02

Table 4: Résultats de notre modélisation sur la fréquence appliquée à 20 000 individus

Ces deux derniers tableaux représentent les performances de notre modèle. On remarque de la même manière que l'on ne sur-apprend pas, notamment grâce à l'ajout de la seconde contrainte. Le gain de performance sur le modèle de coût est cependant très faible. Les performances des deux approches mises en concurrence sont sensiblement identiques, cependant on remarque que l'on surperforme grandement le GLM LASSO sur le modèle de fréquence.
Ces résultats sont extrêmement encourageants d'autant plus que nous disposons de plusieurs axes d'améliorations possibles pour la performance de nos travaux.

Cette modélisation permet donc d'obtenir un algorithme qui soit interprétable, automatique et plus performant que le GLM LASSO.

Synthesis note

Fair pricing is a crucial issue in the insurance world. Indeed, it is one of the markers of differentiation between the multiple actors in terms of durability. The context of strong competitive tension anchored in the non-life sector exerts a friction on tariffs, which is mechanically associated with a drop in security loads. The market is therefore more sensitive to hazards, as evidenced by the impact of the COVID-19 crisis which, according to Standards Poor's, will amount to $50 billion for the global insurance market. More than ever, in-depth risk control is a major component of stability.

It is in light of this observation that we are going to propose an innovative pricing algorithm for property and casualty insurance that will attempt to overcome the shortcomings of the methodology currently in place. This algorithm will have to be both interpretable and automatic, a notion to which we will give a more precise meaning later on.

The most common approach in damage pricing consists in breaking down the pure premium into two parts. Thus, current models seek to predict separately the frequency of occurrence of a claim and the associated average cost. These two components are, in the vast majority of cases, estimated using a GLM (Generalised Linear Model). It is based on the following representation

$$g(\mathbb{E}[Y|X]) = X\beta,$$

where Y is the target variable, i.e. frequency or cost, X the matrix of covariates, β the unknown parameter that will be approached by optimisation and g a link function. This is the algorithm that we are going to try to surpass. We notice that it can only model linear dependencies between the features and the variable of interest. Its application leads then to a distortion of the real dependencies, not allowing more complex shapes to be taken into account. Two main solutions, based on the same algorithm as the GLM, can solve this problem. The first is to add variables before re-running a GLM. The idea is to use a polynomial approximation of the dependency by adding the variables $X_{,j}^2$, $X_{,j}^3$, ..., $X_{,j}^k$ in order to be able to find the polynomial of order k which best fits the dependency of the j-th variable. This tool is difficult to automate, it requires either a human intervention defining variable by variable the adequate degree of the polynomial, or a computation time long enough to test a multitude of parameters. The second is the GAM (Generalised Additive Model) which consists in fitting with non-parametric methods a dependency function to each variable $X_{,j}$ leading to the modelling

$$g(\mathbb{E}[Y|X]) = \beta_0 + f_1(X_{,1}) + ... + f_p(X_{,p}).$$

This method can be implemented automatically and therefore provides a solution to the practical difficulties of the previous one. It does not, however, allow a selection of variables to be made, a step which remains crucial in insurance. Indeed, although the abundance of data allows for greater precision in the estimation, it will be necessary to question the insured on each of the retained variables. Nevertheless, the trend is to speed up the underwriting process, which leads to a drastic reduction in the number of variables required. The LASSO penalty of GLM methods is an indispensable asset for a relevant algorithm, allowing the selection of variables to be carried out automatically. The objective of this thesis is therefore to present an automatic solution to the GLM limits, while remaining

interpretable, which will be ensured by the fact that our work is largely based on the same concepts as the latter.

Our work will be based on a dataset from the portfolio of one of the largest French insurers with initially 89 variables and about one million lines (thus individuals). The latter has a large number of geospatial variables, but we decided to work at the departmental level only because we will have to discretize the data afterwards, so we want to guarantee sufficient exposure for each modality. We will limit ourselves to water damage coverage in HRM only, our two target variables will be the frequency of loss as well as the average cost of a loss. The first one will take into account the exposure as an offset variable, i.e. a variable whose coefficient is set to 1 in order to describe the frequency of loss on an annual basis. The prediction of the average cost will only be made on the basis of individuals who have suffered a claim, which represents a small proportion of the total base (less than 10%). 80% of the lines will be used as a learning base to train our model, the remaining 20% will be used to measure the performance of our modeling to compare it to traditional methods. It is therefore appropriate to describe the specificities of our model before presenting the results of its application on our data.

The main limitation of GLM methods lies in the linearity constraint imposed by the model, which in practice is not always verified, our first objective is to get rid of it.
In this sense, the first particularity of our modeling is to discretize our entire database, in a way similar to the transformations performed on qualitative variables. Of course, continuous variables, or at least variables with a large number of modalities, would create to too many new variables at the end of this process. It is advisable to restrict oneself to grouping together the close modalities (in the sense of the Euclidean distance for numerical variables) within the same group. In order to avoid disparities in representation within groups, it is decided to group them together by quantiles. Nevertheless, if the variable is discrete and presents a small number of modalities (we can think in particular of a classic variable *Number of children*) we will simply create a new variable for each one of them. Consequently, applying a linear model on this basis will associate a coefficient to each modality, which will naturally make it possible to reproduce a much wider spectrum of forms of dependency than before.

Nevertheless, with this procedure, an overfitting bias is introduced because any coefficient $\beta_{k,j}$, associated with the j-th modality of the variable k will be estimated from a much smaller number of observations. We will therefore try to take advantage of the information provided by the modalities close to the one used to estimate the coefficient $\beta_{k,j}$. In order to do this, we will introduce a penalty aimed at limiting the difference between the coefficients associated with modalities close to each other. In this way we use most of the relevant information at our disposal while avoiding overfitting which could lead to a lack of interpretability. For example, it is desirable that $\beta_{k,1}$ should not be too different from $\beta_{k,2}$ in order to prohibit large tariff differentials for two individuals almost similar.

Finally, we wish to add a LASSO penalty for the reasons mentioned above. We will therefore work with the following optimisation problem

$$\hat{\beta} = \arg\min_{\beta}\{-\mathcal{L}(\beta) + \lambda_1||\beta||_1 \ + \lambda_2 \sum_{k=1}^{k=m} \sum_{j=2}^{p_k-2} (\beta_{j-1}^k + \beta_{j+1}^k - 2\beta_j^k)^2\}. \tag{2}$$

$\mathcal{L}(\beta)$ is the log-likelihood of the observations, so we place ourselves in a classical likelihood maximization framework. We will implement a Poisson's law to model the frequency and a Gamma law

for the cost. We also see the two penalties mentioned above associated respectively to a weight λ_1 and λ_2 which we will choose by cross validation. For the second penalty, m represents the number of quantitative variables subject to dichotomization and p_k the number of modalities or associated classes.

The chosen modeling is therefore entirely described through this equation. In order to make it usable in practice, it is necessary to be able to compute $\hat{\beta}$ and thus be able to minimize the target function. We will actually try to approach this solution by iterative methods. As the classical approach of gradient descent is not adapted by the presence of the L^1 norm, our choice is the proximal gradient method which allows to solve minimization problems for non-differentiable functions. It is among other things frequently used in the resolution of a GLM LASSO, the reader can refer to Parikh and Boyd, 2013 for a detailed explanation of its operation.
It is mainly based on the proximal function which is defined by

$$\text{prox}_{tf}(v) = \arg\min_x (f(x) + \tfrac{1}{2t}||x - v||_2^2).$$

This is the trade-off between f minimization and proximity to the v antecedent.

The algorithmic steps to obtain an approximation of$\hat{\beta}$ are briefly described and are based once again on the work presented here Parikh and Boyd, 2013.

- Given at the iteration k the parameters $\beta_k, t := t_{k-1}$ and a shrinking factor τ.
- We repeat the following structure until the stop condition
  1. Let $z := \text{prox}_{tg}(\beta_k - t\nabla f(\beta_k))$.
  2. Let the stop condition be $f(z) < \hat{f}_t(z, \beta_k)$.
  3. Widen the step such that $t := \tau t$.
- Return $t_k := t$ and $\beta_{k+1} := z$

Here $\hat{f}_t(x, y) = f(y) + \nabla f(y)^T(x - y) + \frac{1}{2t}||x - y||_2^2$.

This is the adaptive step proximal gradient descent algorithm for which an accelerated version named FISTA (Fast Iterative Soft-Thresholding Algorithm) can be implemented.

Once implemented, it is possible to apply our model to the data presented above. To do this, we perform a cross validation to optimize the values of the two hyperparameters. A 5-fold cross-validation is then used. Once the couple $(\lambda_1^*, \lambda_2^*)$ has been determined for the modelling of the average cost and also of the frequency, we decide to compare the performances of our model to the GLM LASSO, the most common method in pricing.

We then establish the following comparability protocol. We separate our database into two folds, a learning subset comprising 80% of the individuals, and a test database with the remaining individuals. This separation is identical for all models. Performance will only be evaluated on the optimal hyperparameters. These will be calculated by cross-validation on the learning base with deviance for decision metrics. Three indicators, i.e. Deviance, RMSE (Root Mean Square Error) and MAE (Mean

Absolute Error) will then be calculated on both the learning and the test subsets. We will then be able to estimate the impact of overfitting. Indeed, if these metrics show too large discrepancies between the calculation made on the learning data and the test data, we will conclude that the model over-fits too much and is only adapted to the input data.
To conclude, we will compare the deviation between the GLM LASSO and our algorithm, with a great focus on the test subset outcome, to determine which is the most efficient. This is the most relevant comparison indicator because it is a generalisation of the calculation of the sum of residual squares in the case where the optimisation method maximises likelihood. It is calculated by

$$D(Y, \hat{\mu}) = -2(\mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta^*)).$$

The second likelihood is calculated by replacing the prediction $g^{-1}(X\beta)$ by the true value of Y.
It is an ideal comparison metric candidate since the deviation difference between the two models follows a χ^2 law under the assumption that the second model is the true one, which allows to select the best model. The smaller the deviation, the better the model (by construction it is always a positive quantity).

The results of this comparison are then presented in following tables.

Data	Deviance	RMSE	MAE
Train subset	6695.62	2245.60	1191.93
Test subset	1743.30	2213.48	1178.06

Table 5: Results of the GLM LASSO on the average cost model

Data	Deviance	RMSE	MAE
Train subset	1397.60	0.11	0.02
Test subset	442.29	0.12	0.02

Table 6: Results of the GLM LASSO on the frequency model applied to 20 000 individuals

These first two tables represent GLM LASSO's performance on our data. As far as the frequency model is concerned, exposure has, in this model and in our algorithm, been implemented as an offset variable. As we can tell from the results, overfitting does not seem to be an issue as the last two metrics are almost similar from train base to test. The deviance varies more but it is a mechanical effect, it is not an average but a sum (of log-likelihood), so it is a quantity that increases with the number of data. To make a consistent comparison, we can roughly multiply the deviance from the test subset by 4 (these data represent 20% of the total dataset and are therefore 4 times less than those used for the learning data). The orders of magnitude thus obtained suggest that the model does not overfit.

Data	Deviance	RMSE	MAE
Train subset	6642.00	2240.91	1182.49
Test subset	1743.28	2212.96	1173.01

Table 7: Results of our model on the average cost model

These last two tables represent the performances of our model. We notice in the same way that we do not overfit, notably thanks to the addition of the second constraint. The performance gain on the cost model is however very small. The performances of the two competing approaches are more or less identical, however we notice that we greatly outperform the GLM LASSO on the frequency

Data	Deviance	RMSE	MAE
Train subset	1549.44	0.11	0.02
Test subset	332.22	0.10	0.02

Table 8: Results of our model on the frequency applied to 20,000 individuals

model.
These results are extremely encouraging, all the more so as we mentioned several possible perspectives for improvement in the performance of our work.

This modeling thus makes it possible to obtain an algorithm that is interpretable, automatic and more powerful than the GLM LASSO.

Remerciements

Je tiens à adresser mes premiers remerciements à l'équipe Pricing et Consulting d'ADDACTIS France qui m'ont accueilli et intégré, me permettant de me développer tant sur le plan professionnel que personnel. Je remercie particulièrement Xuan-Quang Do qui m'a accompagné tout au long de mes travaux, répondant ainsi à mes interrogations et m'ouvrant sur de nouvelles problématiques. Les nombreux échanges que j'ai eus avec Pierre Chatelain ont également grandement participé à ma compréhension du sujet et m'ont permis d'être en mesure d'en retransmettre au mieux les subtilités et difficultés.
Je remercie également l'ensemble des membres d'ADDACTIS France pour leur sympathie à mon égard et leur grande disponibilité.

J'adresse également mes remerciements à l'ensemble du corps professoral de l'université Paris-Dauphine qui m'a permis de bénéficier d'une formation de qualité. Je remercie en particulier Christophe Dutang, responsable du master Actuariat pour ses conseils, ses enseignements et son temps. Merci à Alexandre Afgoustidis pour sa bienveillance, son écoute et sa pédagogie et qui, par son enthousiasme, a su me transmettre sa passion pour les mathématiques et leurs applications.
Je remercie également Victor-Emmanuel Brunel, professeur à l'ENSAE Paris pour les précisions techniques dont il m'a fait part en ce qui concerne l'optimisation avancée.

Pour finir je remercie Ioana Preda pour tous les échanges constructifs que j'ai eus avec elle et l'aide qu'elle m'a apportée tout au long de la réalisation de ce mémoire.

Table des matières

Introduction

La justesse de la tarification est un enjeu crucial du monde assurantiel. C'est en effet un des marqueurs de différentiation entre les multiples acteurs sur le plan de la pérennité. Le contexte de forte tension concurrentielle ancré dans le secteur non vie exerce une friction sur les tarifs, qui est mécaniquement associée à une baisse des chargements de sécurité. Le marché est donc plus sensible aux aléas, comme en témoigne l'impact de la crise de la COVID-19 qui, selon Standards Poor's, s'élèvera à 50 Milliards de dollars pour le marché mondial de l'assurance. La maîtrise approfondie des risques est donc plus que jamais essentielle pour la stabilité d'une compagnie d'assurance. Au regard de ces observations, la recherche de nouvelles approches en terme de tarification se doit d'être au coeur de toute stratégie d'entreprise assurantielle.

Ce mémoire vise donc à proposer une méthode de tarification innovante pouvant rivaliser avec celles mises en place actuellement, notamment le GLM (Generalised Linear Model). On restreindra notre champs d'étude aux produits d'assurance IARD (Incendie, Accidents et Risques Divers) et en particulier à la garantie dégât des eaux en MRH (multi-risque habitation). L'objectif est de programmer de zéro sur le logiciel R (version 4.0.3 par R Core Team, 2020) un algorithme qui soit à la fois automatique (dans un sens que l'on précisera plus tard), interprétable et qui réponde aux exigences que l'on explicitera au fur et à mesure des chapitres.

Pour rendre compte de ces travaux on commencera par exposer les grands principes de la tarification non vie avant de présenter la base de données sur laquelle portera notre étude. On proposera par la suite de détailler la démarche actuellement mise en oeuvre portant sur les méthodes de GLM afin d'en déterminer des axes d'amélioration avant de proposer la modélisation que l'on a choisi d'implémenter. Pour finir, on proposera une application de nos travaux sur un modèle de coût/fréquence que l'on confrontera aux résultats des méthodes classiques sur la même base de données.
Bien qu'une grande partie du travail effectué concerne l'implémentation algorithmique du modèle proposé, celle-ci ne sera que peu mise en avant dans ce mémoire, l'objectif étant plutôt de présenter l'impact pratique de ces travaux ainsi que les concepts mathématiques sur lesquels ils se fondent.

Chapitre 1

Notions d'assurance IARD et présentation des données

1.1 Introduction

L'objectif de ces travaux étant le développement d'un algorithme de tarification, il est essentiel dans un premier temps d'expliciter le cadre de son application. On souhaite donc dans ce chapitre expliciter les grandes étapes de la mise en place d'une tarification ainsi que décrire les données sur lesquelles cette dernière s'appuiera.

La tarification est le procédé par lequel un assureur détermine le tarif d'un produit d'assurance. Ces derniers se regroupent en deux grandes familles, les produits vie - c'est à dire ceux dont la réalisation du risque sous-jacent est lié à le durée de vie humaine - et a contrario les produits non-vie. Également appelé assurance IARD (Incendies, Accidents et Risques Divers) ou assurance dommage, c'est sur cette seconde catégorie que portera notre champ d'étude. Le mécanisme de tarification est en effet sensiblement différent d'un secteur à l'autre. Le risque vie est majoritairement basé sur les tables de mortalité, celles-ci permettant d'estimer les durées de vie résiduelle pour chaque génération et à chaque âge, elles sont la source principale de données de toute tarification. Néanmoins pour le marché IARD, un éventail bien plus large de données peut être utilisé. Le mécanisme de pricing sera détaillé par la suite mais c'est bien cette abondance de variables explicatives qui offre un choix de modélisation plus vaste dans le domaine de la tarification non-vie et qui justifie par conséquent de tels travaux.

1.2 La tarification IARD

L'assurance a pour vocation de protéger les bénéficiaires de la perte potentielle engendrée par un aléa. En ce sens, le dédommagement qu'elle sera amenée à réaliser est d'un montant inconnu et ce jusqu'à la fin de la période de couverture (voire au delà dans le cas des IBNR). Il est cependant nécessaire de connaître le prix que devra payer l'assuré pour être éligible à cette couverture. On parle alors du cycle inversé de production. La notion de tarification jouit donc en assurance d'un rôle crucial et d'autant plus complexe. La volonté de l'assureur est alors de maîtriser au mieux le risque sous-jacent au contrat afin de prévoir le plus justement les dépenses auxquelles il fera face. Il est donc important de connaître les différents produits d'assurance, leurs spécificités ainsi que leur tendance pour appréhender au

mieux un tarif. On se propose pour cela de présenter quelques produits d'assurance IARD afin de contextualiser notre étude avant de décrire les grands principes de tarification.

1.2.1 Le marché IARD et ses garanties

Le marché français non-vie pesait un poids de 93 milliards d'euros de cotisations en 2018 (contre 220 milliards pour le secteur vie). Comme décrit précédemment, il se compose d'une multitude de risques dont les plus prédominants sont portés par les garanties auto et MRH (multirisque habitation).
Le marché global de l'assurance se décompose en 26 branches d'accréditation dont la première moitié regroupe la quasi-totalité du secteur IARD. Une garantie en est prédominante, la RC (responsabilité civile). Elle correspond à l'obligation de réparer les dommages causés à autrui. On recense la RC véhicules terrestres, maritimes, aériens ainsi que la RC générale. C'est la garantie la plus coûteuse en assurance dommage, principalement à cause des sinistres dits corporels. Elle ne couvre en effet pas que les dégradations matérielles mais également physiques et est donc potentiellement vouée à prendre en charge des traitements lourds et sur une très longue période (le cas extrême souvent illustré est celui d'un accident de voiture provoquant une tétraplégie d'un tiers dans lequel la responsabilité de l'assuré est mise en cause). La garantie RC corporelle est donc extrêmement complexe à tarifer, car disposant d'une très grande variance et d'un nombre restreint de données. On présente ci-dessous l'écart temporel entre la déclaration du sinistre et le versement de l'indemnité pour une garantie liée à ce risque.

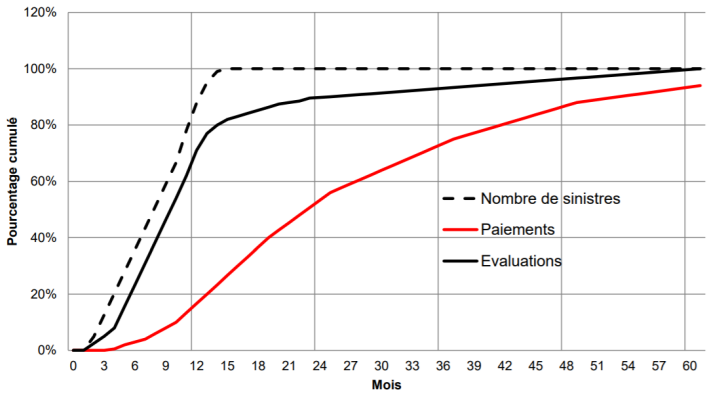


FIGURE 1.1: Garantie longue type RC corporelle, Source : Cours de Comptabilité Règlementation Assurance de M. Philippe GUYON, Université Paris-Dauphine GUYON, 2019

L'écart entre la survenance d'un sinistre et l'évaluation de son coût provient du fait qu'un individu tiers doit estimer le coût total (après décision médicale ou avis d'expert par exemple), auquel s'ajoute l'écart au versement, qui peut s'étaler sur plus de dix ans comme le révèle le graphique. Ce schéma révèle donc la complexité d'un produit RC.
Cela correspond néanmoins à une durée de versement bien supérieure à la normale, la non-vie étant en général une branche qualifiée de courte, c'est à dire dans laquelle le laps de temps entre la survenance d'un sinistre et le versement des prestations est sensiblement plus rapide que dans des contrats type épargnes. A titre d'illustration on pourra citer la garantie bris de glace, notamment auto, pour laquelle seuls quelques jours séparent la déclaration du sinistre du remboursement.

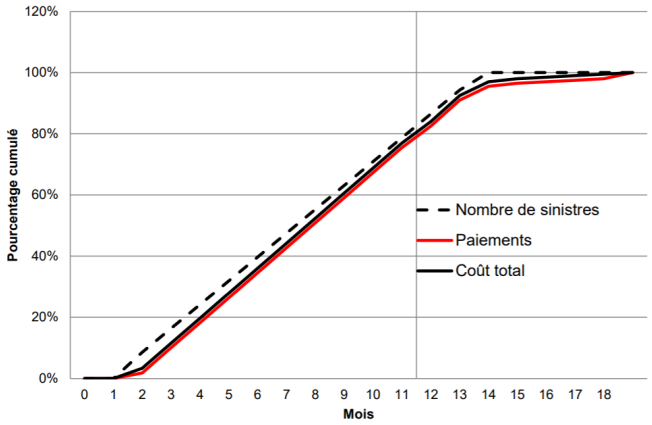


FIGURE 1.2: Garantie courte type bris de glace, , Source : Cours de Comptabilité Règlementation Assurance de M. Philippe GUYON, Université Paris-Dauphine GUYON, 2019

Au delà de cette première différence, d'autres contrastes bien plus impactants pour la tarification subsistent. En gardant les deux même garanties en titre de comparaison, il est évident que le montant d'un sinistre en RC corporelle sera en moyenne considérablement plus important que celui d'une garantie bris de glace et a contrario la fréquence suivra la tendance opposée, le nombre moyen de sinistres bris de glace est bien plus grand mais également plus stable d'une année à l'autre que celui de la RC. On ne peut donc évidemment pas les modéliser de la même façon. Il est à noter que dans la pratique un produit d'assurance se compose de différentes garanties mais on n'abordera pas ici le mécanisme par lequel la tarification les regroupe, on ne s'intéresse donc qu'à la tarification marginale.

Une des difficultés de ces travaux résidant dans le fait qu'une majeure partie du temps de travail est consacrée au développement algorithmique, il sera choisi dans la suite de ce mémoire de présenter nos résultats sur une unique garantie, le dégât des eaux en MRH. On s'attarde donc sur la description de cette dernière.

Elle a pour but de couvrir l'assuré contre des dégâts occasionnés par l'action de l'eau à des biens mobiliers ou immobiliers, y compris la capacité de jouissance d'un bien, et est une composante principale du contrat multirisque habitation. Elle est obligatoire pour les locataires et très souvent souscrite par les propriétaires, ce qui en fait un produit très bien maîtrisé par les assureurs car extrêmement représenté dans le portefeuille. On détaillera dans la partie suivante la base de données que l'on exploitera. L'abondance de données ainsi que la connaissance empirique de ce risque fait de cette garantie un très bon candidat de test. Elle est en effet relativement stable, on présente d'ailleurs l'évolution de son coût moyen et de sa fréquence au cours du temps ci-dessous.

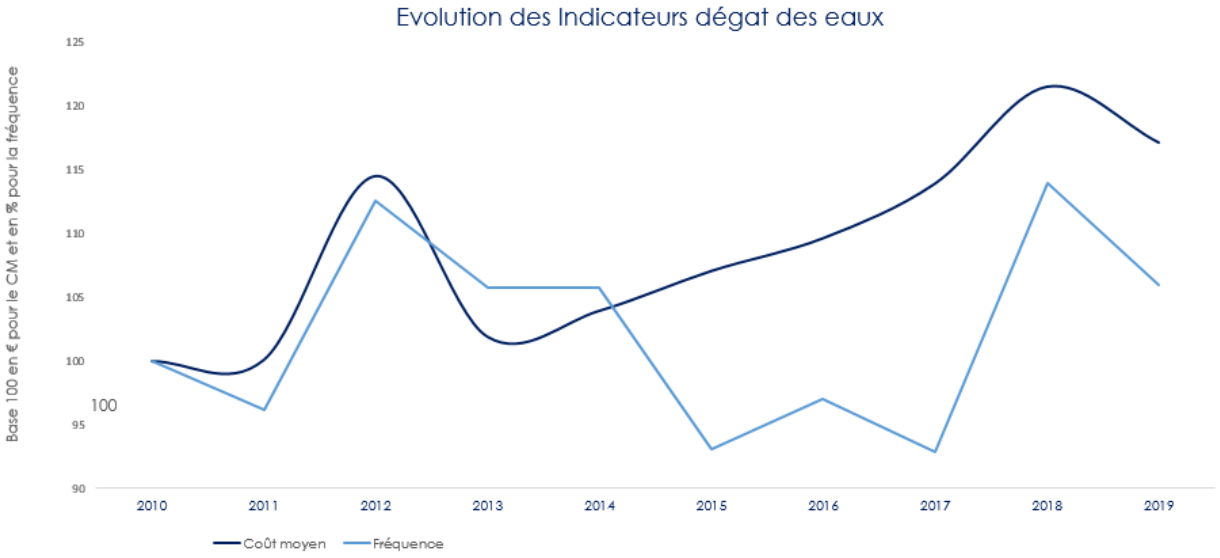


FIGURE 1.3: Evolution de la fréquence et du coût moyen de la garantie dégât des eaux sur 10 ans, Source : Données FFA

Les valeurs ont été renormalisées et sont issues des données FFA (Fédération Française de l'Assurance) 2020. La fréquence observée fait apparaître deux pics en 2011 et 2018 qui correspondent tout deux à des événements climatiques marquants, respectivement de graves inondations à l'automne 2011 et des crues abondantes en janvier 2018. Cette garantie est bien évidemment particulièrement sensible au risque climatique. Ce dernier est cependant fortement volatile et complexe à prévoir. L'ajout de variables externes types zonier dans le modèle est une première approche pour le prendre en compte, mais les observations montrent que les phénomènes impactant la garantie dégât des eaux sont relativement cycliques avec une période d'environ dix ans. On peut donc observer une certaine stabilité de la fréquence sur une moyenne mobile d'une décennie. Le coût moyen augmente quant à lui de par l'inflation, on remarque néanmoins deux pics similaires à ceux de la fréquence, il est en effet naturel de penser qu'en période d'inondations les dégâts provoqués seront bien plus importants que ceux engendrés par des dégâts des eaux classiques.

On décrit dans une seconde partie la base de données qui servira de référence à notre algorithme en présentant les différentes variables utilisées et leur impact sur la garantie dégât des eaux afin d'en peaufiner notre compréhension et notre approche. Il est au préalable nécessaire de présenter dans son ensemble le mécanisme de tarification afin de comprendre en quoi cette base de données est si essentielle.

1.2.2 Marché concurrentiel et principe de tarifications

Le processus de tarification se déroule en plusieurs étapes, la plus importante d'entre elles étant le calcul de la prime pure. C'est en réalité le montant censé représenter au mieux le risque sous-jacent au produit d'assurance, qui entraîne un gain d'espérance nulle, auquel s'ajoute diverses couches, notamment les chargements et les frais. C'est ce qui différencie la prime pure de la prime commerciale

(qui représente le tarif effectivement proposé à l'assuré). On représente sa construction par un schéma succinct.

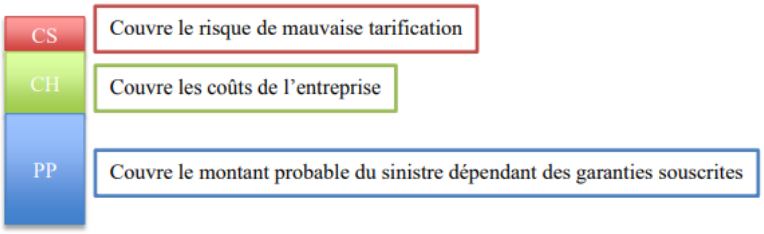


FIGURE 1.4: Décomposition de la prime commerciale

CH représente les chargements de gestion et d'acquisition et CS représente le chargement de sécurité. On s'attardera par la suite uniquement sur la prime pure.

C'est la composante principale d'un tarif. La marge provient en effet pour les contrats IARD de la justesse de l'évaluation actuarielle et des produits financiers alors qu'elle est principalement issue, pour les contrats épargne, des chargements et donc du volume de chiffre d'affaire et de PM. Le ratio combiné en non-vie ($\frac{\text{sinistres+frais}}{\text{charges}}$) est d'ailleurs très proche de 100%, conséquence d'un marché fortement concurrentiel, entraîné notamment par la possibilité de résiliation.

La maîtrise du risque est donc une composante prédominante, qui à la lumière des récents événements s'avère d'autant plus primordiale. En effet on peut citer un article de Mckinsey BECKER et al., 2020 sur l'impact de la COVID-19 sur le secteur assurantiel : "Profitable PC insurers have few similarities : they differ by region (footprints vary among Asia, Europe, and the United States) and customer or channel segment. However, they all emphasize pricing innovation and underwriting excellence and recently have made significant investments in them. Because pricing will be a primary differentiator for long-term value generation in PC, it should be an integral part of every insurance carrier's COVID-19 response strategy". L'innovation en terme de tarification est un élément commun à tous les assureurs dommages les plus rentables et est gage de pérennité face aux aléas extrêmes. C'est un véritable marqueur de différenciation, ce qui motive encore une fois ces travaux.

Plus la tarification est précise plus le chargement de sécurité vu précédemment diminue et plus la compagnie d'assurance maîtrise son risque et reste compétitive sur le marché sans avoir à sacrifier sa solvabilité.

On décrit maintenant plus en détail le processus de calcul de prime pure usuel. Celui-ci se décompose en cinq étapes conformément au document de l' PLANCHET et MISERAY, 2017 :

- La constitution d'une base de données sur laquelle s'appuiera l'apprentissage du modèle. Elle doit donc être à la fois représentative (pour la pertinence) et suffisamment volumineuse (en nombre de lignes pour la précision de l'estimation et en nombre de colonnes pour l'explicabilité).

- Un seuil de séparation des sinistres classiques, dits attritionnels et des sinistres graves que l'on modélise par la théorie des valeurs extrêmes. On s'intéresse ici uniquement aux sinistres de la première catégorie.
- Le segmentation de la population, qui permet de ressortir des groupes distincts auxquels on appliquera un tarif différent.
- Le choix de modélisation qui relie les variables à expliquer aux variables observables et observées. Les variables d'intérêt seront le coût d'un sinistre et sa fréquence (on rappel que ce procédé s'applique garantie par garantie).
- Et enfin le lissage du tarif qui permet en plus de prendre en compte les contraintes tarifaires.

Cette démarche se base donc sur une méthode de coût/fréquence. On cherche à modéliser séparément le coût moyen et la fréquence d'apparition d'un sinistre avant de les rassembler pour en extraire la prime pure.
Il est nécessaire d'introduire quelques variables pour faciliter l'écriture, en commençant par S, la charge totale d'un contrat pour une garantie spécifique.
On a alors l'écriture suivante : $S = \sum_{i=1}^{N} C_i$ où N est le nombre de sinistres ayant lieu durant la période de couverture et C_i le coût du i-ème sinistre.
Les hypothèses de modélisation sont les suivantes : $N, C_1, C_2, ..., C_N$ sont indépendants et les C_i sont identiquement distribués de loi C.
Sous ces hypothèses, il vient que $\mathbb{E}[S|X] = \mathbb{E}[N|X]\mathbb{E}[C|X]$.

L'objectif sera de prédire, en espérance, S pour un segment d'individu particulier, ce qui revient, sous chaque segment, à déterminer le coût moyen et la fréquence moyenne. On souhaite attirer l'attention du lecteur sur un point essentiel de ce raisonnement : l'intérêt de modéliser le risque moyen uniquement. En effet in fine notre objectif sera de proposer un tarif à l'assuré, utiliser l'espérance du risque souscrit pour le calcul de la prime pure peut donc sembler insuffisant. Le risque de sous tarification associé à un individu apparaît en effet comme trop important, d'autant plus que l'environnement actuariel est soumis à de nombreuses normes de solvabilité. Il est donc crucial de rappeler que le fondement même de l'assurance consiste en la diversification des risques. Un individu seul peut ne pas être en mesure de pallier financièrement la réalisation d'un sinistre tout en étant en mesure d'en couvrir le coût espéré. De ce fait, bien qu'un individu pourra, a posteriori, avoir été sous tarifé dans le sens où sa prime ne couvre pas le montant des réparations versées (ce qui reste bien l'objectif de l'assurance), la perte qu'il engendrera sur le portefeuille global sera compensée par les individus ayant subi moins de sinistres que la fréquence attendue en moyenne. C'est donc la solvabilité globale du portefeuille que l'on regarde et non pas celle de chaque assuré individuellement (bien que les principes de bonus/malus permettent de corriger des effets de l'antisélection). L'espérance est donc la mesure appropriée à notre problématique, sous l'hypothèse que les individus sont indépendants et donc que le risque se diversifie correctement. On ne détaillera pas ici le problème de corrélation des individus, typique d'un risque climatique qui affecte par essence un très grand nombre d'assurés simultanément et l'on fera par la suite l'hypothèse classique d'indépendance des variables réponses.

On se place alors naturellement dans le cadre des modèles de régression, notamment le GLM (Generalised Linear Model) qui est la méthode la plus courante en tarification. Celle-ci sera détaillée plus

amplement dans le chapitre suivant. Néanmoins cette méthode, bien que très répandue, possède des limites réelles que l'on exposera par la suite, ce qui nous pousse à nous intéresser à des structures semblables mais plus complexes tel que le modèle GAM (Generalised Additive Model). Nous allons donc développer une méthode qui s'en inspire grandement et comme nous l'avons vu, la première étape d'une telle modélisation est la constitution d'une base de données, que nous présentons dès à présent.

1.3 Descriptif de la base de données, analyse uni- et multivariée

Nous utilisons au cours de nos travaux un jeu de données provenant du portefeuille d'un des plus gros assureurs français doté initialement de 89 variables et d'environ un million de lignes (donc d'individus). Ces données étant confidentiel on les transformera avant de les exposer dans ce mémoire. Elles ne seront donc pas représentatives des données réelles mais les ordres de grandeurs seront maintenus. On détaillera dans cette partie les traitements effectués à la base de données avant de réaliser une analyse univariée puis multivariée.

1.3.1 Pré-traitement de la base

Le jeu de données avait au préalable été utilisé lors d'une mission au sein du groupe Addactis. Les modifications nécessaires à son utilisation étaient donc moins nombreuses qu'à l'accoutumée. Les données portaient néanmoins initialement sur un ensemble de garanties MRH trop grand, il était donc nécessaire de supprimer toutes les variables relatives aux garantie autres que le dégât des eaux. On supprime alors l'exposition, la fréquence et le coût relatifs à dix autres garanties. Un dictionnaire des variables permettait d'appréhender une très large majorité des caractéristiques dont nous disposions, nous avons supprimé celles qui étaient spécifiques au vocabulaire de la compagnie d'assurance, comme le numéro d'image contrat *IMAGE*, le code acte de gestion de l'image *ACTGEST*. De nombreuses variables géospatiales semblaient superflues, on disposait par exemple du département de résidence ainsi que du code INSEE et du code postal. On choisit de ne conserver que le département *DEPT* car c'est la maille qui nous semble la plus adaptée à nos travaux. En effet, on discrétisera par la suite les variables, il est donc nécessaire de travailler avec des variables disposant de suffisamment d'exposition sur chaque modalité. Il était également nécessaire de reformater des variables. Celles correspondant à des dates ont été recodées au format YYYYMMDD, de nombreuses variables ont été transtypées en facteur et l'âge, renseigné à l'origine en mois, a été modifié pour apparaître en années. Les années de naissance supérieures à 2009 étaient décalées d'un siècle, l'année 2010 étant par exemple représentée par la valeur 2110, ce que nous avons corrigé. D'autres problèmes mineurs ont également été corrigés au fur et à mesure de la prise en main des données.

A l'issue de cette première phase de transformation nous étions à même de chercher des anomalies dans le jeu de données. 67 lignes possédaient par exemple une surface habitable de $0m^2$ que nous avons supprimée. On remarque également l'existence de lignes à coût de sinistres nul mais à fréquence strictement positive, on fait alors l'hypothèse de l'existence d'une franchise qui pousse l'assuré à déclarer un sinistre qui s'avère par la suite d'un montant inférieur à la franchise. On pensera donc à ne prendre en compte que les sinistres déclarés ayant dépassé le seuil (donc à coût non nul) dans la modélisation

de la fréquence.

Le travail effectué au préalable pour l'utilisation de la base de données dans une autre mission a grandement accéléré le processus de traitement. On se propose donc maintenant de présenter analytiquement les différentes variables conservées.

1.3.2 Descriptif des variables

Nous disposons de deux variables cibles, représentant respectivement la fréquence et le coût, *NB_HS_CLIMGRAVE_DDE* et *CHG_HS_CLIMGRAVE_DDE*. Les noms des variables ont été conservé pour préserver la cohérence avec le dictionnaire des variables évoqué précédemment.

La première variable cible est discrète et sa valeur varie entre 0 et 4 sinistres. Ce n'est pas à proprement parler la variable *NB_HS_CLIMGRAVE_DDE* que l'on cherche à modéliser car celle ci n'est pas uniformément associée à une exposition d'un an. Etant donné que l'on souhaite calculer la prime pure correspondant au montant moyen annuel de la garantie dégât des eaux, il est nécessaire de prendre en compte l'exposition représentée par la variable *AA_DDEG*. On utilise alors le principe de variable offset que l'on décrira dans le chapitre suivant.
Bien que notre base de données soit composé d'environ un million d'individu, la plupart de nos applications seront réalisées sur des sous échantillon car la vitesse d'exécution de notre algorithme n'était pas la priorité. Celui-ci prends donc encore trop de temps pour être exploitable efficacement sur des quantités trop importantes de données. On présente dans le tableau suivant les principales statistiques de la variable de fréquence (modifiées pour des raisons de confidentialité) issues de l'échantillon de 200 000 individus que l'on utilisera.

Statistique	Minimum	Maximum	1er quartile	Médiane	3ème quartile	Moyenne	CV
Valeur	0	4	0	0	0	0.01342	9.542

Table 1.1: Statistiques de la variable de fréquence

La quantité CV correspond au coefficient de variation, quantité qui est comparable d'une variable à une autre, qui se calcule comme le ratio entre l'écart type et la moyenne. On remarque que la plupart des valeurs observées sont à 0 étant donné que le 3ème quantile vaut 0. En réalité seul un pourcent des assurés observés ont subi un sinistre au cours de la période d'exposition. Notre base d'apprentissage pour le modèle de coût sera alors bien plus petite que celle pour la fréquence.

La seconde variable représentant le montant des sinistres est continue, le montant maximal d'un sinistre atteint environ 50 000 €. Une grande proportion des sinistres observés ont un montant faible. On représente la densité approchée du montant des sinistres en euro ci-dessous pour visualiser la distribution globale.

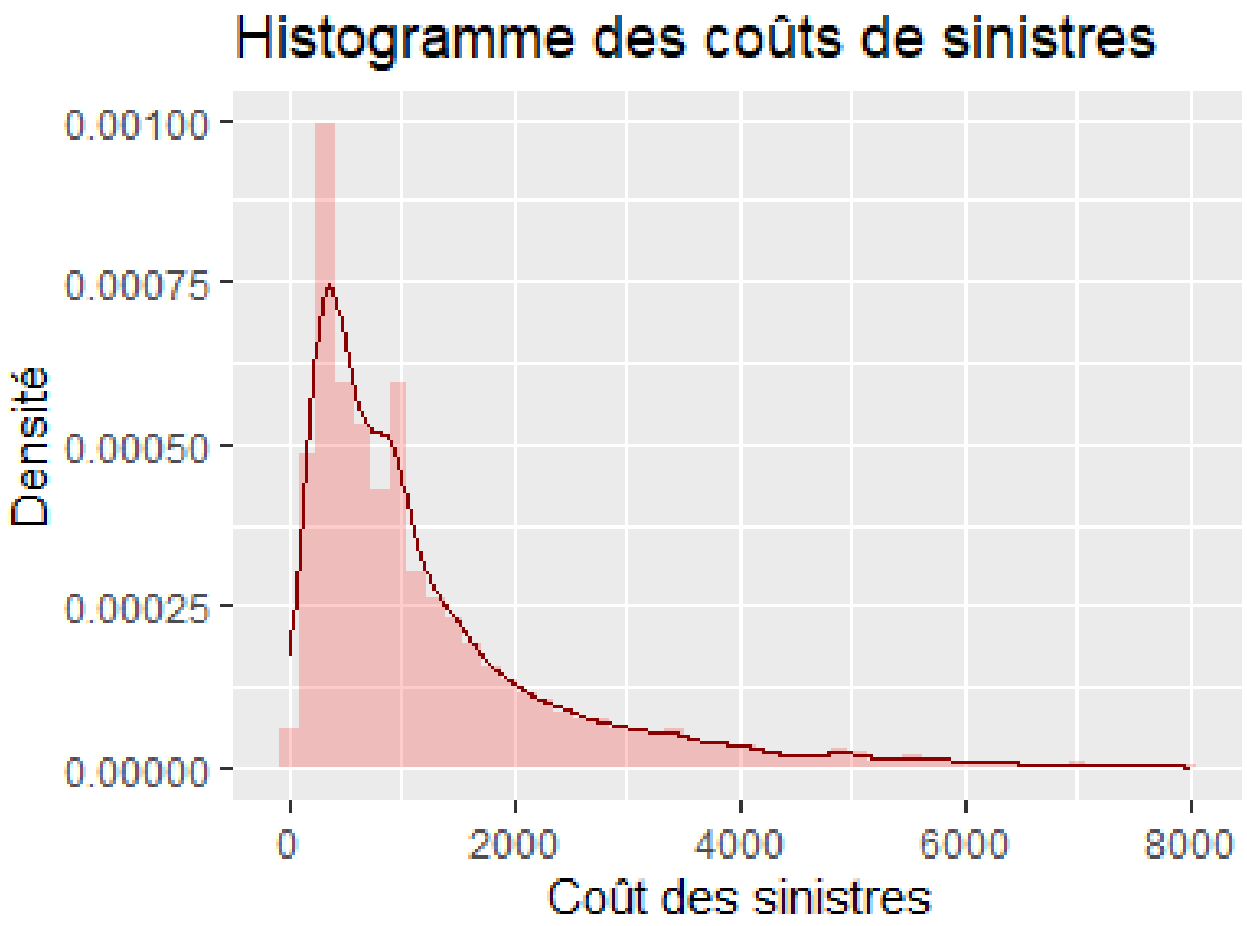


FIGURE 1.5: Histogramme des coûts de sinistres en euro

Les montants extrêmes ont été tronqués pour obtenir un graphique aux échelles lisibles. En effet les valeurs extrêmes étant trop importantes l'axe des abscisses devenait bien trop large, les montant extrêmes ont donc été supprimés avant de réalisé ce graphique.
Cette variable a du être modifiée afin d'être considéré comme la variable cible réelle. En effet pour les observations dont le nombre de sinistres observés dépasse strictement un, le montant *CHG_HS_CLIMGRAVE_DDE* correspond à la somme des sinistres reportés. Il est donc nécessaire de diviser cette variable par le nombre de sinistres *NB_HS_CLIMGRAVE_DDE* afin d'avoir une réponse représentant le coût moyen d'un sinistre. On présente ci-dessous le même tableau que précédemment afin d'obtenir les statistiques de base du coût des sinistres (notamment la moyenne que l'on cherchera à approcher au mieux).

Statistique	Minimum	Maximum	1er quartile	Médiane	3ème quartile	Moyenne	CV
Valeur	2.35	50 000	420	900	1753.04	1460.55	3.516

TABLE 1.2: Statistiques de la variable de coût

La période d'observation de nos données s'étale de 1986 à 2018 mais le premier quartile vaut 2007, ainsi la majeure partie des observations sont concentrées sur une décennie. En regardant la figure 1.3, on observe une tendance à la hausse à la fois sur les coûts et sur les fréquences au cours du temps

pour les sinistres de la garantie dégât des eaux. Cette dépendance temporelle sera modélisée car on conserve la variable *ANNEE* dans le modèle.

La garantie considérée étant le dégât des eaux en MRH, une grande partie des variables disponibles correspondent à la description du logement de l'assuré comme la superficie avec la variable *SURFACE* représentant la surface des dépendances, le nombre de pièces *NBPIECE*, le montant assuré *MTLIMITE* ou encore la zone géographique *ESPINSEE*. Une multitude d'option est disponible pour cette garantie, notamment l'option équipement, objets précieux ou encore multimédia qui sont des variables dichotomiques. Ces variables sont plus que classiques en tarification, on fait le choix de ne pas les détailler une à une afin de s'attarder spécifiquement sur celles qui présentent un intérêt particulier pour nos travaux.

Les dernières variables traitent des caractéristiques de l'assuré, telles que le nombre de sinistres qu'il a déclarés sur son contrat précédent *NBSIN*, le nombre d'enfants à sa charge *NBENFCTR*, son département *DEPT* ou son âge *AGEOCC*. On représente ci-dessous l'histogramme de l'âge des occupants des biens assurés qui jouera un rôle important par la suite.

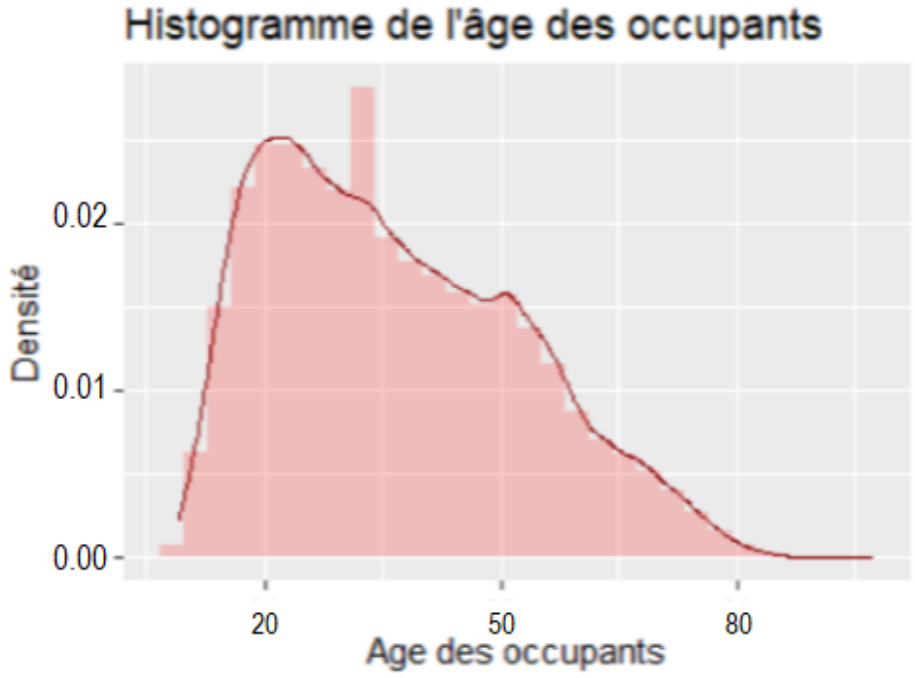


FIGURE 1.6: Histogramme de l'âge des occupants du bien assuré

L'âge le plus représenté est autour de la trentaine, le pic observé à 33 ans correspond probablement à la valeur remplie par défaut lorsque la donnée est manquante. En effet chaque variable classique dispose d'une modalité dite de référence qui approxime la valeur la plus neutre pour le risque considéré. Les pics de densité correspondent alors la plupart du temps à cette valeur par défaut.
On souhaite détailler en profondeur cette variable car on s'en servira pour illustrer certains de nos propos par la suite. On présente donc maintenant la relation entre l'âge des occupants et la sinistralité moyenne observée.

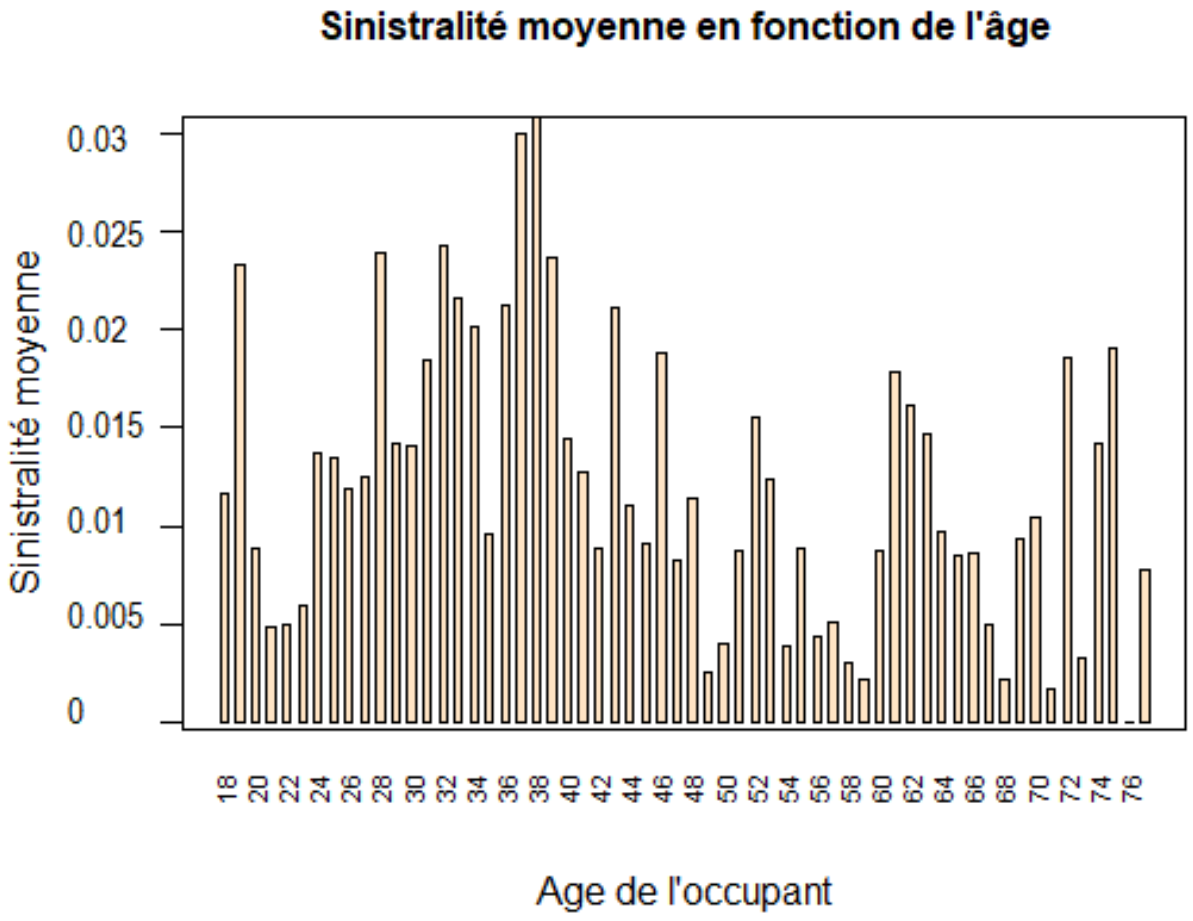


FIGURE 1.7: Diagramme en barre représentant la sinistralité moyenne observée selon l'âge des occupants

On remarque que cette dépendance n'est absolument pas linéaire. En effet on observe de nombreux pics, notamment aux alentours de 18 ans, 40 ans et après 60 ans. On interprétera ces résultats dans les chapitres suivants mais c'est bien l'observation de ce type d'indicateurs multivariés qui nous pousse à nous intéresser à de nouvelles modélisations. En effet le GLM permet uniquement de retranscrire des formes de dépendances linéaires entre les covariables et la réponse, ce qui n'est vraisemblablement pas adapté à notre jeu de données.

1.3.3 Indicateurs multvariés

On souhaite explorer d'avantage les indicateurs multivariés.

On présente donc les corrélations entre les variables à expliquer et les covariables. La matrice de corrélation représentée calcule les coefficients de corrélation de Pearson. Dans le cas du coût on obtient le graphique suivant.

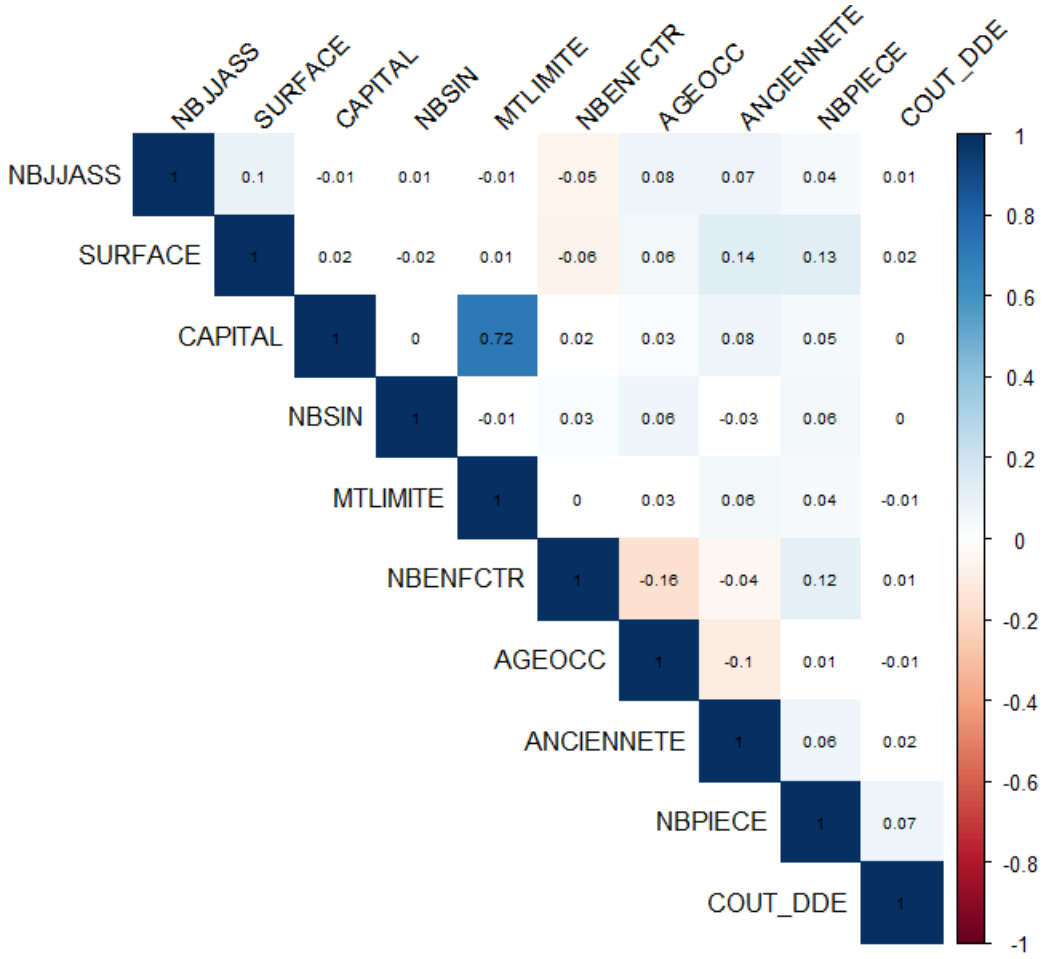


FIGURE 1.8: Matrice de corrélation entre les variables et le coût d'un sinistre

Les coefficients reflètent une relation linéaire entre deux variables continues. On observe que la variable d'intérêt *COUT_DDE* est très faiblement corrélée aux autres. Ce constat nous pousse à remettre en cause l'efficacité des méthodes linéaires comme le GLM sur ces données. Le capital est sensiblement lié au montant limite assuré ce qui semble cohérent. Le reste des variables ne sont pas corrélées. L'absence de corrélation entre la variable à expliquer et les variables explicatives nous incite à ne pas plus nous attarder sur les indicateurs multivariés dans le cadre de la modélisation du coût moyen.

Des résultats similaires apparaissent pour la variable de fréquence.

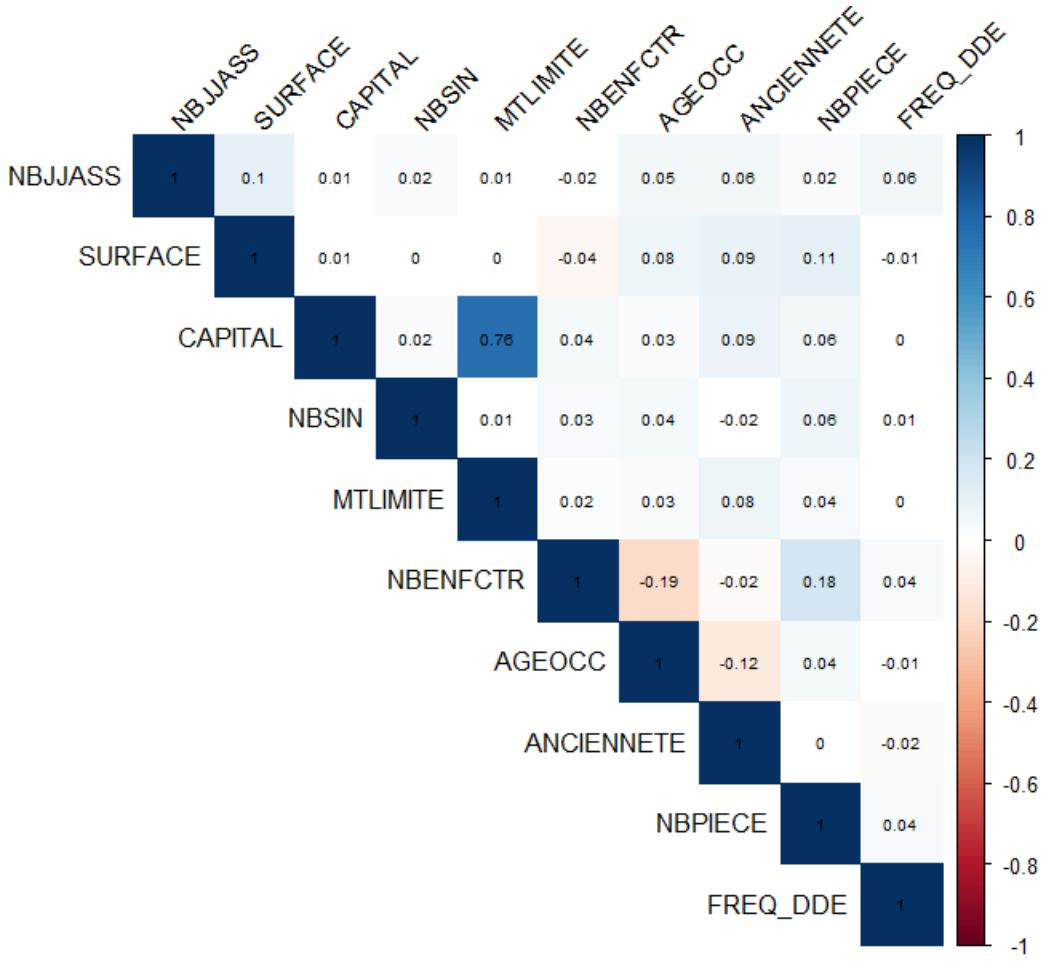


FIGURE 1.9: Matrice de corrélation entre les variables et la fréquence de sinistre

On remarque une très faible corrélation entre le nombre de jours d'assurance et la fréquence d'apparition d'un sinistre. C'est cependant la corrélation la plus importante ce qui semble instinctif. Les graphiques présentés dans la section précédente mettent cependant en exergue le fait que des dépendances de forme non linéaires peuvent apparaître entre la fréquence et une variable explicative. On cherchera alors par la suite à s'affranchir des dépendances linéaire pour améliorer nos prédictions.

Les grands principes de tarification et la base de données ayant étés présentés, on va maintenant décrire précisemment la méthode la plus couramment employé, à savoir le GLM.

Chapitre 2

Présentation du modèle actuel et de ses limites

2.1 Introduction

Dans l'objectif d'appréhender l'algorithme que nous allons implémenter il est essentiel de décrire les méthodes fondamentales desquelles il s'inspire. Nous avons vu dans le chapitre précédent que la méthodologie la plus fréquemment mise en place en tarification IARD se repose sur le GLM (Generalised Linear Model). Nous y consacrerons donc la première partie de ce chapitre avant de présenter la pénalité LASSO qui permet une régularisation ainsi que la sélection des variables. Enfin nous mettrons en exergue les limites de la modélisation actuelle, ce qui nous dirige vers des pistes d'améliorations.

2.2 Du modèle linéaire classique au modèle linéaire généralisé

Nous nous intéressons aux modèles de tarification. Ceux-ci ont une vocation prédictive, c'est à dire qu'ils sont utilisés dans le but de prédire une certaine quantité d'intérêt. Dans le cadre de la tarification actuarielle, on cherche par exemple à estimer le montant de la prime que devra payer un assuré pour s'offrir une garantie spécifique. En réalité ce n'est pas directement ce montant que l'on souhaite en sortie de notre modèle mais plutôt une estimation du coût moyen global qu'un tel contrat engendre. La prime payée dans les faits est en général supérieure à cette estimation (on se réfère au chapitre précédent et à la différenciation entre prime pure et prime commerciale). On souhaite donc trouver une méthodologie algorithmique efficace offrant en sortie une approximation la plus fine possible de la prime pure.

Pour ce faire, on commence par décrire l'espace dans lequel on travail ainsi que les objets que l'on sera amené à manipuler tout au long des parties théoriques. On dispose donc d'une matrice de données X contenant n réalisations de p variables quantitatives. Parmi celles-ci on compte des variables numériques, discrètes et continues, ainsi que des variables initialement qualitatives, que l'on a retraitées de manière classique en les discretisant modalité par modalité. On travaille également avec un vecteur Y de réponses pour ces n individus - i.e. de réalisations de notre variable aléatoire cible que l'on va par la suite chercher à prédire -.

2.2.1 Le modèle linéaire classique

Un des premiers modèles enseignés dans le cadre de la prédiction est le LM (Linear Model). Il est en effet des plus intuitifs et relativement simple dans sa résolution. Il se fonde cependant sur des hypothèses fortes, limitant ainsi son champ d'application. Il présuppose notamment que la variable Y est, à un bruit blanc prêt, une combinaison linéaire des variables explicatives. On considère donc qu'il existe β tel que $Y = X\beta + \epsilon$. Couramment, on impose également une loi sur le ϵ, c'est à dire que l'on considère $\epsilon \sim \mathcal{N}(0, \sigma^2)$. On appelle alors ce modèle le modèle linéaire Gaussien.
L'enjeu est, ici comme dans toute la suite, de trouver β, ou du moins le meilleur β possible. Cela permettra à terme de prédire Y simplement à l'aide du p-uplet de variables explicatives. On considère dès lors ϵ comme un terme d'erreur que l'on juge négligeable.
Dans un contexte assurantiel des variables typiques de la matrice X en auto sont par exemple le lieu de résidence, le modèle de la voiture assuré, l'ancienneté du permis de conduire etc... Ce sont autant de variables que l'on espère liées à notre cible Y et dont on n'aura plus qu'à renseigner les valeurs pour prédire le montant de la prime pure d'un nouvel assuré. Si ce dernier est représenté par un vecteur V de taille p de variables d'entrées, sa prédiction sera alors ${}^tV\beta$ (le symbole ${}^t.$ correspond à la transposée de la matrice).

L'élément clé de cette modélisation et de toutes celles futures réside donc dans l'évaluation de β. Une procédure couramment employée peut être la maximisation de la vraisemblance. L'objectif est de calibrer β de sorte que l'observation des données Y soit de probabilité maximale compte tenu de la matrice X. La loi de Y connaissant les données est en effet dépendante du paramètre β dans le sens où $Y|X \sim \mathcal{N}(X\beta, \Sigma)$. Il est donc possible d'écrire la densité de probabilité de cette loi en fonction de β et de l'optimiser.
On rappelle que pour une loi normale $\mathcal{N}(X\beta, \Sigma)$ on a

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} \exp\left(-\frac{1}{2}{}^t(y-\mu)\Sigma^{-1}(y-\mu)\right).$$

On souhaite maximiser cette vraisemblance selon β sachant que $\mu = X\beta$ et $\Sigma = \sigma^2 I_n$, on peut donc ne travailler qu'avec

$$\widetilde{f}_Y(y) := \exp\left(-\frac{1}{2\sigma^2}{}^t(y - X\beta)(y - X\beta)\right).$$

On passe au logarithme pour simplifier les calculs car c'est une fonction croissante qui n'impacte pas l'antécédent des extremums. On notera qu'en plus des simplifications calculatoires, cette astuce a un réel intérêt pratique car elle conduit au modèle multiplicatif. Il vient donc que

$$\hat{\beta} = \arg\max_{\beta}\{-\frac{1}{2\sigma^2}||y - X\beta||^2\}.$$

On remarque alors que raisonner par maximum de vraisemblance est, dans le cas gaussien, équivalent à optimiser β par moindres carrés ordinaires (MCO). On pose donc

$$\hat{\beta} = \arg\min_{\beta}\|Y - X\beta\|^2.$$

C'est précisément cette équation qui caractérise le modèle. Il existe en effet de nombreuses manières de définir un β cible dont la quasi totalité sont représentées par un problème de minimisation. Concernant cette équation spécifiquement, sa résolution est simple. La norme L^2 étant dérivable, on pourrait trouver β par une classique dérivation (la recherche d'optimums est en effet souvent rattachée à l'annulation de la dérivée de la fonction cible). Cependant il est plus interprétable de procéder comme suit. On remarque d'abord que $\min_{\beta}\|Y-X\beta\|^2 = \min_{\nu\in[X]}\|Y-\nu\|^2$. On utilise la notation $[X]$ pour décrire l'espace vectoriel engendré par les colonnes de X. Par définition du projeté, l'argmin du membre de droite est $\hat{\nu} = P_{[X]}(Y)$. C'est la projection orthogonale de Y sur $[X]$. Soit X^k la k-ième colonne de X, on a alors pour tout k, $< X^k, P_{[X]}(Y) - Y >= 0$ car $P_{[X]}(Y) - Y \in [X]^{\perp}$. Or, $P_{[X]}(Y) = X\hat{\beta}$, d'où $< X^k, X\hat{\beta} - Y >= 0$, soit ${}^tX^k(X\hat{\beta} - Y) = 0$. Ceci étant vrai pour tout k, on peut passer à l'écriture matricielle et il vient que ${}^tX(X\hat{\beta} - Y) = 0_{\mathbf{R}^p}$.
On trouve alors naturellement que si ${}^tXX$ est inversible, il y a unicité de la solution et

$$\hat{\beta} = ({}^tXX)^{-1}\,{}^tXY.$$

On obtient une formule fermée qui ne nécessite pas de résolution itérative évitant ainsi la lenteur d'exécution et les approximations. On reviendra par la suite sur la nécessité d'inversibilité de ${}^tXX$. On note simplement que cette condition est vérifiée si $n > p$. A cette contrainte viennent s'ajouter plusieurs hypothèses de validité du modèle qu'il convient de décrire.

L'application des résultats précédents n'est en effet valable que sous couvert de la validité de quatre postulats qui portent sur le terme d'erreur ϵ :

- $[H1]$ Les erreurs sont centrées, c'est a dire que $\mathbb{E}[\epsilon] = 0_{\mathbf{R}^p}$. Cette hypothèse est essentielle car on souhaite par la suite négliger ϵ, que l'on considère comme un terme d'erreur, donc a minima d'espérance nulle.

- $[H2]$ Les erreurs sont de variance constante, ce qui se traduit par $\forall i, \mathbb{V}[\epsilon_i] = \sigma^2$. On parle alors de modèle homoscédastique.

- $[H3]$ Les erreurs sont de plus indépendantes, ce qui entraîne l'indépendance des observations.

- $[H4]$ Enfin, les erreurs sont supposées gaussiennes. Cette hypothèse est propre au modèle linéaire gaussien mais un nombre suffisant d'observations permet de la négliger par des propriétés asymptotiques.

En pratique ces hypothèses doivent être vérifiées pour appliquer le modèle linéaire. Des outils graphiques permettent de valider les deux premières, la suivante ne relevant que du protocole expérimental de collecte des données est en général considérée vraie et la dernière pouvant être supplantée par un large nombre d'observations est également souvent négligée. On se propose donc d'étudier les graphiques associés à $[H1]$ et $[H2]$.

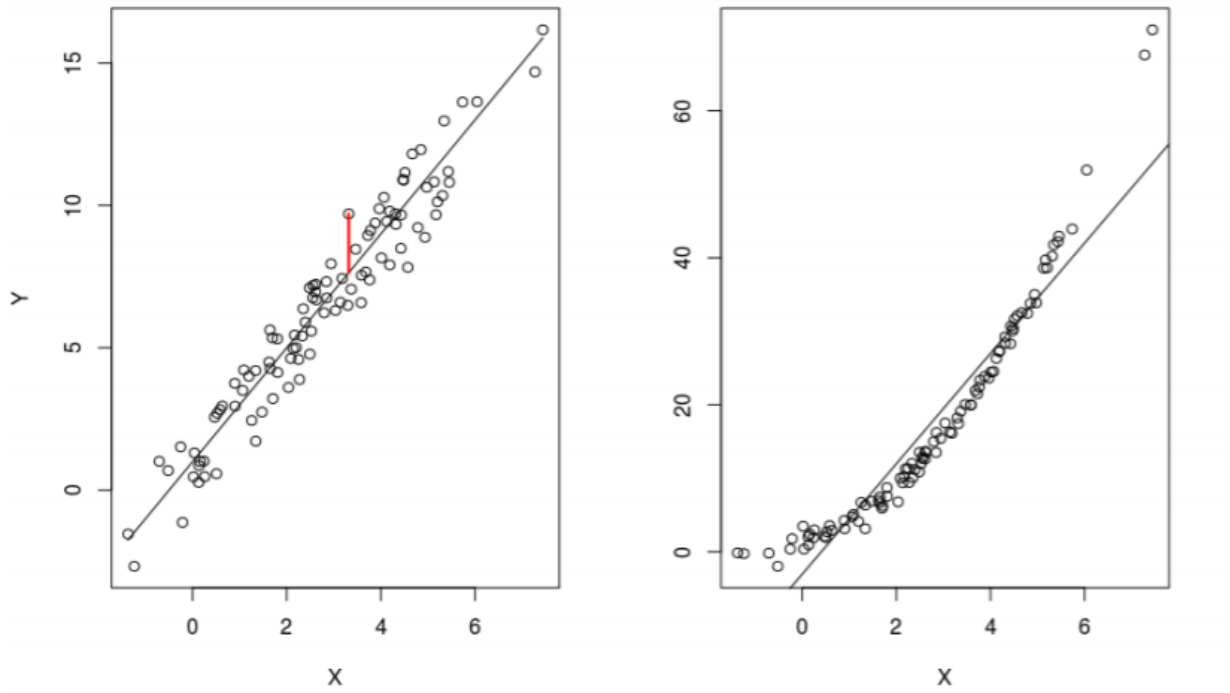


FIGURE 2.1: écart résidus/droite de régression, Source : Cours de Modèle Linéaire Mme. Sophie DONNET, Université Paris-Dauphine DONNET, 2018

On se place en une dimension pour visualiser correctement. La matrice X n'est donc composée que d'une variable. On trace alors la variable cible Y en fonction de X. On trace ensuite une droite de régression de coefficient β correspondant à notre prédiction. L'écart en rouge correspond alors au terme d'erreur d'une observation, il apparaît clairement qu'à gauche l'erreur moyenne est proche de zéro ce qui n'est pas le cas à droite. Le modèle linéaire n'est pas adapté au second jeu de données.

La seconde hypothèse concerne la constance de la variance, on s'attend donc à observer des écarts à la moyenne relativement stables peu importe la valeur de réalisation de la variable X. On présente à nouveau deux graphiques illustrant cet écart sur des jeux de données différents.

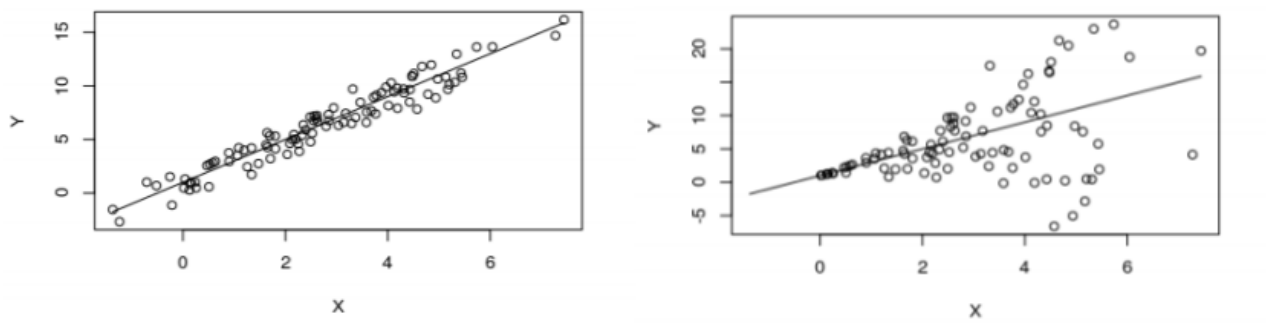


FIGURE 2.2: ϵ de variance constante à gauche et variable à droite, Source : Cours de Modèle Linéaire Mme. Sophie DONNET, Université Paris-Dauphine DONNET, 2018

Encore une fois le premier jeu de données semble plus adapté au modèle linéaire. On observe en effet que la figure de gauche présente une variance constante (l'écart à la prédiction ne semble pas dépendre de l'abscisse) alors que celle de droite est bien supérieure pour les observations élevées de X ce qui contredit l'hypothèse d'homoscédasticité.

Bien que très intuitif et relativement performant, ce premier modèle reste très limité. En effet une dernière contrainte ayant précédemment été évoquée requiert d'avoir un nombre d'observations supérieur au nombre de covariables ($n > p$). On se propose de fournir une brève démonstration en annexe pour ne pas alourdir cette partie A.1. On retient simplement que cette condition ne permet pas de traiter un nombre conséquent de variables explicatives, on se prive alors potentiellement de certaines d'entre elles. L'objectif final n'est cependant pas de toutes les conserver.
L'ajout d'une variable ne peut être que bénéfique à la prédiction car si elle n'apporte rien, son β sera de zéro. Néanmoins in fine les covariables devront être renseignées dans le modèle, on sera donc amené à questionner les futurs assurés sur chaque variable utilisée. De par le principe de concurrence, de nombreuses études sur les taux de transformations ont montré que le client est rebuté par des démarches trop longues. On souhaite donc conserver un nombre restreint de variables explicatives mais leur choix est complexe.
Il existe une méthode de sélection pour le LM mais celle-ci est longue et complexe. Il s'agit dans les faits de tester l'efficacité d'une multitude de modélisations composées chacune de variables différentes et de ne garder que celle offrant le meilleur ratio nombre de variables/précision. On y retrouve par exemple la sélection dite backward/forward dans laquelle on enlève ou ajoute successivement la variable apportant le moins ou le plus d'informations jusqu'à atteindre un critère d'arrêt précis. On peut néanmoins remédier à cette difficulté en ajoutant un terme de pénalité dans notre équation de minimisation. On reviendra sur ce point au cours de la partie suivante.

Il apparaît clairement que la limite principale du modèle linéaire est la restriction de ses hypothèses. C'est pour pallier cela que l'on présente désormais le modèle linéaire généralisé.

2.2.2 Le modèle linéaire généralisé GLM

Comme son nom l'indique, c'est en réalité une version généralisée de notre première modélisation. L'hypothèse principale est la suivante

$$g(\mathbb{E}[Y|X]) = X\beta.$$

On relâche alors grandement la contrainte de linéarité en introduisant une fonction g, appelé fonction de lien. En effet on n'impose plus directement à Y d'être linéairement dépendant de X mais simplement à son espérance conditionnelle.
On remarque rapidement l'analogie avec le modèle linéaire classique où g est l'identité. De plus la seule contrainte sur la loi de Y est désormais qu'elle appartienne à la famille exponentielle (qui englobe un grand nombre de lois, y compris la loi normale). En pratique, cette différence est essentielle. En effet, il est courant d'employer un raisonnement coût/fréquence en tarification assurantielle. On cherchera alors à modéliser indépendamment le coût moyen d'un sinistre et sa fréquence d'apparition. On est donc amené dans le premier cas à travailler avec une variable cible positive ou nulle ce qui est en contradiction avec l'hypothèse de gaussianité de modèle linéaire. On définit la famille exponentielle comme l'ensemble des lois dont la densité prend la forme

$$f_Y(y) = \exp(\tfrac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)).$$

On présente alors un résultat clé sur les lois appartenant à cette famille, on a que pour tout Y appartenant à la famille exponentielle, $\mathbb{E}[Y] = b^{'}(\theta)$.

Démonstration :
On part de $\int_{\mathcal{Y}} f_Y(y)\ \mathrm{d}y = 1$ où $\mathcal{Y}$ est le domaine de définition de f_Y.
On dérive ensuite par rapport à θ, alors, par intervention limite intégrale
$\int_{\mathcal{Y}} \frac{\partial}{\partial\theta} f_Y(y)\ \mathrm{d}y = 0 \Leftrightarrow \int_{\mathcal{Y}} \frac{y - b^{'}(\theta)}{a(\phi)} f_Y(y)\ \mathrm{d}y = 0 \Leftrightarrow \mathbb{E}[Y] = b^{'}(\theta)$

On peut en effet déduire de cette égalité que $\forall i$, $\mathbb{E}[Y_i] = b^{'}(\theta_i)$. Or, $\mathbb{E}[Y_i] = g^{-1}(X_{i,}\beta)$. On exprime ainsi θ en fonction de β de la façon suivante

$$\theta_i = (b^{'})^{-1} \circ g^{-1}(X_{i,}\beta).$$

Avec ce résultat, on possède tous les éléments nécessaire à l'optimisation de β. On va de nouveau raisonner par maximisation de vraisemblance. On cherche alors à résoudre

$$\hat{\beta} = \arg\max_{\beta}\{L(\beta)\}$$

Ou $L(\beta)$ est la vraisemblance des observations. Les individus étant supposés indépendant, leur loi de couple est égale au produit des lois. De plus, la fonction logarithme étant croissante, raisonner sur la vraisemblance ou sur le log de cette dernière est équivalent. On écrit alors la log-vraisemblance (notée $\mathcal{L}(\beta)$) du couple $(Y_1, ..., Y_n)$, puis on dérive par rapport à β pour égaliser à zéro. Enfin, on utilise

l'égalité précédemment démontrer pour faire apparaître β dans la vraisemblance de Y.

$$\begin{aligned}\mathcal{L}(\beta) &= \sum_{i=1}^{n} \log(f_Y(y_i)) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \\ &= \sum_{i=1}^{n} \frac{y_i h(X_{i,}\beta) - b \circ h(X_{i,}\beta)}{a(\phi)} + c(y_i, \phi)\end{aligned}$$

Où $h = (b^{'})^{-1} \circ g^{-1}$.
Le passage au logarithme nous a permis de simplifier l'égalité car les Y appartiennent à la famille exponentielle.

Une attention particulière doit être portée à la modélisation lorsque Y représente la fréquence du risque. L'exposition y joue un rôle particulier et doit être traitée en variable offset, c'est à dire une variable dont le coefficient est fixé à un. On ne s'intéresse en réalité pas à la fréquence en elle-même mais plutôt au taux de sinistralité par unité de temps. Notre objectif étant de réussir à calculer une prime sur une base annuelle, nos prédictions doivent représenter le risque sur une période d'un an. Malheureusement les observations sont réalisées sur des individus qui ne vérifient pas nécessairement cette propriété. On doit donc modifier légèrement la modélisation pour le prendre en compte. On travaille précisément avec l'équation suivante

$$g(\mathbb{E}[\tfrac{Y}{S}|X]) = X\beta.$$

On se place dans le cadre de la loi Poisson puisque c'est celle-ci que l'on utilisera pour modéliser la fréquence. Ici S représente l'exposition, mais c'est bien $Y|X$ qui suit une loi de Poisson et non pas le ratio. Etant donné que l'on cherche à prédire Y, qui est la loi de la fréquence de sinistres sur une année, on réécrit l'égalité comme suit, $log(\mathbb{E}[Y|X]) = X\beta + log(S)$ (ici g est la fonction de lien logarithme). La meilleur prédiction, une fois l'optimisation sur β réalisée sera donc $exp(X\beta + log(S))$. On remarque bien que notre modélisation est équivalente à un modèle linéaire dans lequel la variable exposition est entrée en logarithme et est associée à un coefficient de 1. La log-vraisemblance dans ce cas particulier est

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i X_{i,}\beta - s_i e^{X_{i,}\beta} + y_i \log(s_i) - \log(y_i!).$$

L'objectif restera de maximiser la log-vraisemblance pour trouver le β optimal associé à toutes les autres variables, auquel viendra s'ajouter en dernière composante le coefficient 1 associé à l'exposition.

Dans le cadre général, afin de maximiser la log-vraisemblance, il suffit de dériver selon β_i et égaliser à 0. Malheureusement la solution n'est pas explicite, on va donc utiliser une résolution numérique par

un algorithme de type Newton-Raphson. Celui-ci est présenté en annexe, d'abord en une dimension pour visualiser son application puis dans sa forme généralisée.A.3

On est donc en mesure de résoudre le problème d'optimisation associé au GLM. Néanmoins, il subsiste encore une limite à son application dans le sens où la sélection de variables est tout autant contraignante que dans le modèle linéaire. La section suivante introduit alors la pénalisation LASSO qui représente une solution efficace à ce problème.

2.3 Pénalisation LASSO

Le LASSO (pour Least Absolute Shrinkage and Selection Operator) correspond concrètement à l'application d'une pénalité à un modèle déjà existant. On peut donc parler de modèle linéaire LASSO, de GLM LASSO ainsi que, plus tard, de Spline LASSO.

2.3.1 Présentation de la contrainte LASSO

L'implémentation de cette méthode consiste en la modification du problème d'optimisation initial. Prenons le cas du modèle linéaire classique. On s'intéresse à la résolution de l'équation suivante

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 \tag{2.1}$$

.

On va simplement y ajouter un terme qui aura pour but de pénaliser les β_j. On se retrouve alors avec cette équation

$$\hat{\beta} = \arg\min_{\beta} \{\|Y - X\beta\|^2 + \lambda||\beta||_1\}. \tag{2.2}$$

Ce raisonnement se base sur le fait que les variables associées à des β_j de faible valeur ont finalement peu d'impact sur la prédiction de Y (en supposant que la matrice X ait été centrée et réduite).
Prenons par exemple β^* l'optimum de l'équation (2.1) et posons j et $\widetilde{\beta}$, tel que

$$\widetilde{\beta}_k = \beta^*_k \ \forall k \neq j \text{ et } \widetilde{\beta}_j = 0.$$

Si $\lambda|\beta^*_j| > ||Y - X\widetilde{\beta}||^2 - ||Y - X\beta^*||^2$ alors β^* n'est pas l'optimum de (2.2).

On remarque de plus que quelle que soit la valeur de β^*_j on peut toujours trouver un λ suffisamment grand tel que la condition est vérifiée. On comprend donc intuitivement que cette pénalité à pour

conséquence de forcer les β les moins explicatifs à zéro. En effet si l'on suppose la matrice X centrée réduite, les variables associées à des β_k de faibles valeurs ont un faible impact dans la minimisation de $||Y-X\beta||$ et seront donc plus enclins à vérifier la condition de non optimalité. Ce nouveau problème de minimisation permet donc une sélection de variables. Il n'est cependant pas facile à implémenter. En effet rares sont les problèmes d'optimisation solutionnés par des formules fermées, la démarche classique étant une descente de gradient. La norme L^1 n'étant pas dérivable cette opération est bien plus complexe. On présentera dans ce chapitre une démarche simpliste de résolution que l'on complétera dans par la suite avec celle mise en pratique dans notre algorithme et qui requiert des connaissances bien plus théoriques.

2.3.2 Existence d'une solution

Comme décrit précédemment, la contrainte LASSO peut s'appliquer à la plupart des problèmes d'optimisation. Pour se placer dans un cadre général, on pose alors l'équation suivante

$$\hat{\beta} \in \arg\min_{\beta}\{f(\beta) + \lambda\|\beta\|_1\}.$$

On impose la continuité de f ainsi que sa coercivité. Par propriété d'additivité, la fonction $g_\lambda : \beta \mapsto f(\beta)+\lambda\|\beta\|_1$ est également continue et coercive. Ainsi g admet nécessairement un minimum. Afin de faciliter les calculs, on développera dans cette partie le raisonnement sur le modèle linéaire LASSO uniquement. On reviendra dans le chapitre suivant sur la résolution spécifique à notre problème. On cherche donc à résoudre

$$\hat{\beta} \in \arg\min_{\beta}\{||Y - X\beta||_2^2 + \lambda\|\beta\|_1\} \tag{2.3}$$

2.3.3 "unicité" du minimum

On cherche maintenant à montrer qu'il n'existe qu'une "unique" solution. Cette unicité n'est pas à entendre au sens classique, on démontrera simplement que toute solution du problème possède une unique image par X. En effet d'un point de vue pratique on ne s'intéresse à β que dans le but de calculer $X\beta$ pour nos prédictions. Cette définition de l'unicité est donc amplement suffisante.

On propose un raisonnement par l'absurde. Supposons donc β_1 et β_2 deux solutions de (2.3). On va montrer que $X\beta_1 = X\beta_2$, c'est à dire que leur image par X coïncide.

On suppose que $X\beta_1 \neq X\beta_2$. On pose alors $\beta_3 := \frac{1}{2}(\beta_1 + \beta_2)$. Il vient que

$$g_\lambda(\beta_3) = \sum_{i=1}^{n}(Y_i - \tfrac{1}{2}((X\beta_1)_i + (X\beta_2)_i))^2 + \lambda\sum_{j=1}^{p}\tfrac{1}{2}|\beta_{1j} + \beta_{2j}|.$$

Or, si $x \leq y$ on a que $(\frac{x+y}{2})^2 < \frac{x^2+y^2}{2}$ (une démonstration de ce résultat est disponible en annexe A.2). En combinant ce résultat pour la partie de gauche et l'inégalité triangulaire pour celle de droite, il

vient que $g_\lambda(\beta_3) < \frac{g_\lambda(\beta_1)+g_\lambda(\beta_2)}{2}$. β_1 et β_2 étant solutions de (2.3), ils minimisent g_λ d'où $\frac{g_\lambda(\beta_1)+g_\lambda(\beta_2)}{2} = g_\lambda(\beta_1)$.
On a alors $g_\lambda(\beta_3) < g_\lambda(\beta_1)$ ce qui est absurde car β_1 est un minimum de g_λ.
On a donc prouvé l'unicité du minimum.

La difficulté de la résolution des problèmes LASSO réside dans la non dérivabilité de la norme L^1. En effet, la fonction $h : \beta \mapsto ||\beta||_1$ n'est pas dérivable bien qu'elle vérifie les hypothèses permettant une généralisation de la dérivée. On parlera par exemple dans la sous-section suivante de sous différentiel.

2.3.4 Résolution théorique

On expose ici le théorème permettant de résoudre l'équation. En effet, au vu des difficultés listées plus haut, il est impossible de trouver β simplement en annulant la dérivé de la fonction à optimiser. On a donc recours au résultat suivant

Soit f une fonction convexe différentiable,

$$\hat{\beta} \in \arg\min_{\beta}\{f(\beta) + \lambda\|\beta\|_1\} \iff \exists \delta = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{pmatrix} \in \partial\mathcal{N}_1(\hat{\beta}),\ \frac{\partial f}{\partial \beta_j}(\hat{\beta}) + \lambda\partial_j = 0\ \forall j = 1, .., p.$$

Il s'agit d'une extension de la méthode d'optimisation par dérivation.
On pose $\partial\mathcal{N}_1(\hat{\beta}) := \{x \in \mathbf{R}^p, x_j = sign(\hat{\beta}_j)$ si $\hat{\beta}_j \neq 0, x_j \in [-1, 1]$ sinon$\}$, définit comme le sous différentiel de la norme L^1. Si l'on connaît les δ_i, on peut trouver β en annulant le sous différentiel coordonnée par coordonnée.

Le problème d'optimisation du LASSO n'est pas résolvable par formule fermée, on a donc recours à un algorithme d'approximation que l'on détaille maintenant.

2.3.5 Algorithme d'approximation de la solution LASSO

On va réaliser une optimisation composante par composante et itérer jusqu'à convergence. On commence par centrer et réduire les données pour des facilités de calcul et uniformisation. On travaille donc sur la matrice $\widetilde{X}$ définie comme suit : $\widetilde{X}_{i,j} := \frac{X_{i,j} - \bar{X}_{.,j}}{sd(X_{.,j})}$. A chaque itération on va résoudre pour chaque coordonnée j de β

$$\hat{\beta}_j = \arg\min_{\beta_j}\{\sum_{i=1}^{n}(Y_i - (\widetilde{X}\beta)_i)^2 + \lambda\sum_{i=1}^{p}|\beta_i|\} \text{ avec } \beta_k \text{ fixé pour } k \neq j.$$

On définit donc $u_j : \beta_j \mapsto \sum_{i=1}^{n}(Y_i - (\widetilde{X}\beta)_i)^2 + \lambda\sum_{i=1}^{p}|\beta_i|$. Cette fonction est optimisable sur $\mathbf{R}^{+*}$ et sur $\mathbf{R}^{-*}$. On réalise son optimisation sur $\mathbf{R}^{+*}$ à titre d'exemple :

$$\forall \beta_j \in \mathbf{R}^{+*},\ u_j'(\beta_j) = -2\sum_{i=1}^{n} \widetilde{X}_{i,j}(Y_i - (\widetilde{X}\beta)_i) + \lambda$$
$$= -2nR_j + 2n\beta_j + \lambda,$$

$$\text{avec } R_j = \tfrac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{i,j}(Y_i - \sum_{l=1, l\neq j}^{p} \widetilde{X}_{i,l}\beta_l).$$

En effet, $\sum_{i=1}^{n} \widetilde{X}_{i,j}^2 = n$ par construction. On a donc

$$\hat{\beta}_j = R_j - \tfrac{\lambda}{2n} \text{ sur } \mathbf{R}^{+*}, \text{ i-e lorsque } R_j > \tfrac{\lambda}{2n}.$$

On applique le même raisonnement sur $\mathbf{R}^{-*}$ et dans le dernier cas, le minimum est atteint en $\beta_j = 0$.

On a donc l'algorithme suivant :

Initialisation : On choisit $\hat{\beta}_{init} \in \mathbf{R}^p$ arbitrairement On centre et on réduit les données X tel que $\widetilde{X}_{i,j} = \frac{X_{i,j} - \bar{X}_{.,j}}{sd(X_{.,j})}$.

Récurrence : On effectue les étapes suivantes jusqu'à convergence (on définit un critère d'arrêt) ou jusqu'à un nombre d'itérations maximal fixé. A l'itération k on transforme la j-ième composante comme suit :
$\hat{\beta}_j^{(k)} = sign(R_j^{(k)})(|R_j^{(k)}| - \frac{\lambda}{2n})_+$, avec $R_j^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{i,j}(Y_i - \sum_{l=1, l\neq j}^{p} \widetilde{X}_{i,l}\hat{\beta}_l^{(k-1)})$.

Cette récurrence converge bien vers le β optimal. Nous sommes désormais en mesure d'appliquer une pénalité LASSO dans le cadre de la régression linéaire classique. Il existe des méthodes de résolution plus complexes dans le cadre général dont on détaillera les calculs par la suite. On remarque simplement à cette étape que la démarche actuelle demande une itération composante par composante ce qui laisse présager une lenteur d'exécution.

Outre la pénalité LASSO, il existe une multitude de contraintes, notamment Ridge, Elasticnet etc. Toutes ont vocation à régulariser les coefficients de β. Dans le chapitre suivant on introduira une seconde pénalité que l'on utilisera au sein de notre algorithme. L'avantage majeur de la pénalité LASSO reste la sélection automatique des variables. Dans les faits on utilise d'ailleurs actuellement majoritairement la modélisation GLM LASSO pour la tarification en assurance. Il était donc essentiel de comprendre son fonctionnement. C'est également cette méthode que l'on va tenter de challenger avec notre algorithme. On va alors chercher à exposer les limites de ce modèle pour en trouver des axes d'améliorations possibles.

2.4 Les limites de la modélisation actuelle

Malgré ses performances, on peut aisément identifier des faiblesses au GLM. Celui-ci requiert en effet une forme de dépendance linéaire car il se base sur l'hypothèse suivante

$$g(\mathbb{E}[Y|X]) = X\beta.$$

Bien que plus laxiste que celle du modèle linéaire, il est évident que cette contrainte n'est pas toujours vérifiée. L'application d'un GLM dans un tel cas ne donnerait alors pas les meilleurs résultats possibles. On souhaite montrer en quoi cette hypothèse de linéarité représente une réelle limite pratique. Il peut sembler au premier abord que la fonction de lien g permet de s'affranchir d'une dépendance linéaire, cependant c'est un paramètre relativement fixe dont la valeur dépend simplement de la loi donnée à Y. On récapitule ici les principales fonctions lien.

log	$g(\mu_i) = \log \mu_i$
logit	$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
probit	$g(\mu_i) = \Phi^{-1}(\mu_i)$, où $\Phi(.)$ est la fdc $\mathcal{N}(0,1)$
complementary log-log	$g(\mu_i) = \log\left(-\log(1-\mu_i)\right)$
log-log	$g(\mu_i) = \log(-\log(\mu_i))$

FIGURE 2.3: Les principales fonctions de lien, Source : THOMAS, 2016

Elles sont pour la plupart relativement simples et ne permettent pas de s'adapter aux formes de dépendances entre les covariables et la variable d'intérêt. En effet la plupart sont monotones et sont de plus généralement fixées à l'avance et ne dépendent que du Y considéré, ce qui exclut tout ajustement aux données. D'autre part, les liens entre les variables explicatives et Y n'ont aucune raison d'être identiques d'un $X_{.,j}$ à l'autre alors que la fonction de lien est unique.

On présente ci-dessous la forme de dépendance de la fréquence d'un risque en fonction de l'âge. Les données automobiles sont parfaitement adaptées à notre propos car l'accidentalité est très importante sur deux tranches d'âge distinctes ce qui est en contradiction avec un lien linéaire. Cet exemple fait écho à la figure 1.7 mais se base sur des données plus conventionnels rendant son interprétation plus intuitive pour un regard extérieur au monde de l'assurance.

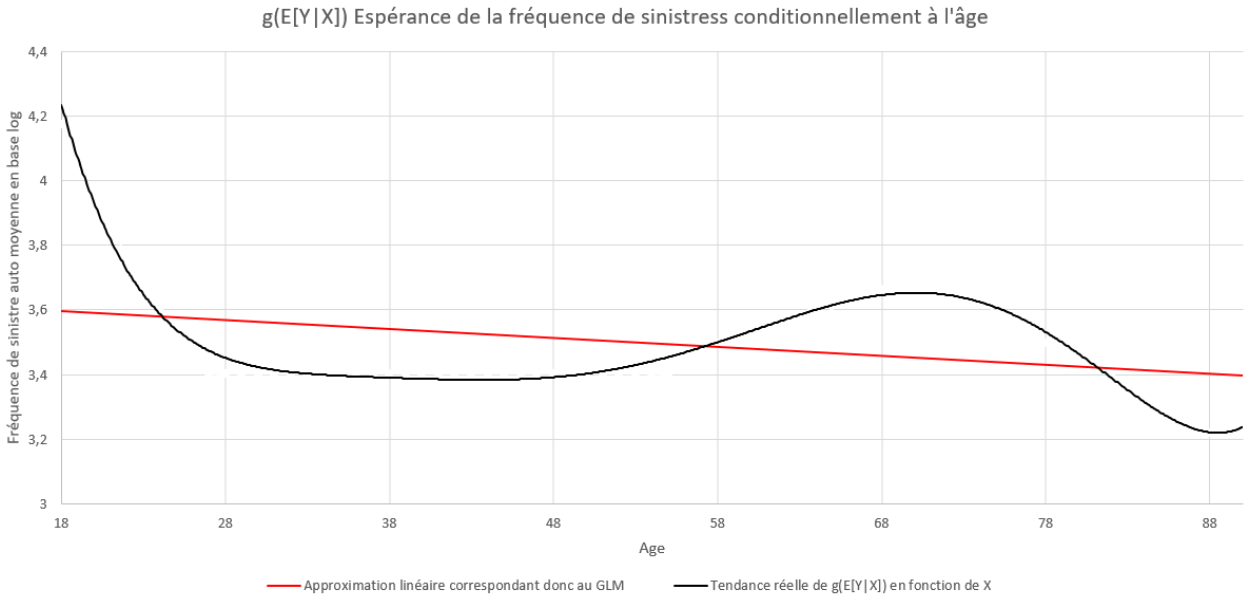


FIGURE 2.4: Relation de dépendance entre l'âge et le log de la fréquence moyenne conditionnelle

La courbe en noir correspond à la dépendance réelle entre la variable âge et $g(E[Y|X])$. On observe que les jeunes conducteurs sont plus sujets à des accidents, au même titre que les personnes plus âgées bien que dans une moindre mesure. Ce phénomène est intuitif et ne peut pas être correctement capté par un GLM. En effet on a tracé en rouge l'interpolation linéaire de la dépendance qui représente la meilleure approximation linéaire de l'effet observé. C'est donc par définition l'approximation faite par le GLM, le coefficient directeur correspondant alors au β associé à l'âge.
L'écart entre prédiction et réalisation sera conséquent de par la violation de l'hypothèse de linéarité.

Malgré tout, le GLM reste le modèle majoritairement utilisé en tarification IARD. Au regard de cet exemple, il semble possible de le sur-performer en s'astreignant de cette contrainte de linéarité. On sait d'ailleurs que certains modèles de machine learning non paramétriques peuvent présenter une MSE (Mean Square Error) plus faible, notamment la random forest. Ces méthodes restent cependant peu interprétables, ce qui pousse à les délaisser. On souhaite en effet être en mesure d'avoir une tarification interprétrable afin de la justifier auprès des assurés notamment.

On a déjà évoqué dans le chapitre précèdent un autre modèle plus complexe que le GLM et qui représente un bon compromis entre interprétabilité et complexité, le modèle GAM (Generalised Additive Model). Celui-ci se repose sur l'hypothèse suivante

$$g(\mathbb{E}[Y|X]) = \beta_0 + f_1(X_{,1}) + ... + f_p(X_{,p}).$$

Les fonctions f_1, f_2, ..., f_p sont des fonctions de dépendances qui sont ajustées sur les données. Cette méthode peut être implémentée automatiquement et corrige en partie les limites du GLM. Les limites de cette méthode résident néanmoins dans le temps de calcul nécessaire à l'ajustement des fonctions f_i pour chaque variables, un potentiel sur-apprentissage et une complexité importante pour ajouter une pénalité permettant la sélection de variables.

La pénalité LASSO des méthodes GLM est en effet un atout indispensable à un algorithme pertinent, permettant de réaliser automatiquement la sélection de variables. On cherchera donc à produire un modèle semblable au GAM mais auquel on peut ajouter cette contrainte tout en corrigeant un éventuel sur-apprentissage.

On présente alors dans le chapitre suivant la modélisation que l'on choisit de mettre en oeuvre.

Chapitre 3

L'algorithme du Spline LASSO

3.1 Introduction

On va dans ce chapitre présenter en détail notre algorithme, de l'idée de sa conception à son implémentation tout en démontrant sa convergence.
Celui-ci se base sur l'observation des limites du GLM évoquées précédemment ainsi que sur les défauts des méthodes plus complexes. Une grande attention sera portée aux concepts mathématiques sous-jacents à notre méthode.

3.2 Présentation générale du modèle

Du fait de la prépondérance du modèle GLM en assurance, il est essentiel, dans l'objectif de produire un algorithme plus performant, d'en corriger les défauts énoncés précédemment. L'illustration de la dépendance non linéaire entre l'âge et la fréquence de la sinistralité automobile représente un axe d'amélioration crucial que nous allons exploiter en discrétisant l'intégralité des variables de notre base de données.

3.2.1 La discrétisation des variables

En effet, avoir une variable par modalité d'âge nous permettrait de capter avec précision la forme de dépendance à la fréquence. Il s'agit donc du premier point clé de notre modélisation : nous allons discrétiser toutes les variables. Il convient alors de définir le mécanisme de ce formatage.

Ce processus est déjà implémenté en pratique pour les variables qualitatives. En effet les méthodes classiques, notamment le GLM, doivent uniquement traiter des bases de données numériques. Les variables non numériques sont donc reformatées de la manière suivante : on crée une colonne par modalité présente dans la base de données, dans lesquelles les individus sont classifiés de manière binaire pour indiquer l'appartenance à cette modalité (1 dans la colonne de la modalité de l'individu et 0 ailleurs).

Néanmoins la matrice de données X ne doit pas être singulière, on souhaite par exemple qu'elle soit de rang plein. Dans le modèle linéaire classique, cette condition est nécessaire à l'inversibilité de la matrice ${}^tXX$, plus généralement elle est en réalité indispensable à l'unicité de la solution. En effet dans tous les problèmes d'optimisation présentés jusqu'à présent apparaissait le terme $X\beta$, on remarque rapidement que si X est singulière, pour tout β^* solution de l'équation il existe un $\tilde{\beta}^*$ différent également solution.

Si par exemple la troisième colonne est le double de la seconde il vient que $\tilde{\beta}_i^* := \begin{cases} \beta^* - 2 & \text{si} \quad i = 2 \\ \beta^* - 1 & \text{si} \quad i = 3 \\ \beta^* & \text{sinon} \end{cases}$

est aussi un optimum.

Conceptuellement cette unicité n'est pas nécessaire mais en terme de significativité des variables, d'interprétabilité du modèle ainsi que pour la vitesse de convergence de l'algorithme et l'initialisation elle s'avère être fondamentale.

Afin d'éviter les dépendances entre variables, il est donc nécessaire par la suite de supprimer une colonne. En effet la matrice issue d'une variable qualitative est dite disjonctive complète, c'est à dire que chaque ligne dispose d'un et unique 1. De ce fait la somme de toutes les colonnes issues d'une même variable originelle sera égale à l'intercept. La matrice ne sera alors pas de rang plein. Pour illustrer ces propos on présente une matrice de design fictive avant discrétisation,

$$\begin{pmatrix} \text{Intercept} & \text{Classe de logement} \\ 1 & a \\ 1 & a \\ 1 & b \\ 1 & c \\ 1 & a \\ 1 & b \end{pmatrix},$$

que l'on discrétise de la manière suivante,

$$\begin{pmatrix} \text{Intercept} & \text{Classe de logement a} & \text{Classe de logement b} & \text{Classe de logement c} \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

On remarque alors que la somme des trois dernières colonnes est égale à l'intercept, il est donc nécessaire d'en supprimer une.

On va appliquer un raisonnement très similaire pour discrétiser les variables numériques. Il faut cependant être prudent quant au nombre de valeurs différentes prises par les variables. On ne peut évidemment pas créer une variable par modalité lorsque l'on s'attarde sur le salaire ou sur toute autre

variable continue. On distinguera alors les variables numériques discrètes avec un nombre restreint de modalités des variables numériques (discrètes ou continues) ayant un trop grand nombre de modalités.

Dans le premier cas on appliquera à l'identique la dichotomisation induite par celle des variables qualitatives. Dans le second, on regroupera les modalités en classe selon une des méthodes proposées ci-dessous.

Classes à pas fixe : La première est sûrement la plus intuitive. Elle consiste simplement à se fixer un nombre k de classes et diviser l'espace des valeurs en k groupes uniformes. Si l'on souhaite par exemple appliquer ce raisonnement sur l'âge en considérant une plage de 15 à 120 ans que l'on divise en 8 groupes, on obtient les classes suivantes : $[15; 30[, [30; 45[, ..., [105; 120]$. Un défaut majeur de cette méthode est l'exposition. Il semble naturel sur cet exemple que les groupes extrêmes seront bien moins représentés que les groupes centraux, ainsi, la variance des β associés aux premiers et derniers groupes sera très élevée et favorisera le sur-apprentissage.

Classes à même effectif : Pour pallier ce problème une seconde idée pourrait être de séparer les classes selon les quantiles et non plus de manière uniforme. La mise en place d'une telle méthode reste simple et permet efficacement de corriger le défaut de la précédente. Il est néanmoins nécessaire de faire attention aux variables que l'on manipule. Par exemple cette méthode sera inefficace dans le cadre de variables mixtes - c'est à dire a priori continue mais avec une Dirac. La variable *SURFACE_DEP* représentant la surface de dépendances de notre base de données en est une parfaite illustration. La majorité des individus seront représentés par un zéro, mais la variable semblera continue au sein de la population possédant une dépendance. Le quantile d'ordre 1/10 sera alors égal au quantile d'ordre 2/10 (valant 0) si plus de 20% de l'échantillon n'en possède pas. Pour adapter cette méthode dans la pratique il faut donc repérer les Dirac et les traiter comme un quantile à part entière avant de refaire une subdivision en dehors de la valeur des potentielles Dirac.

Classes induites par CART : Une dernière méthode basée sur les CART (Classification And Rgression Trees) peut également être envisagée. On propose en annexe une brève présentation des arbres de décision A.4. Il est ici simplement nécessaire de rappeler qu'une telle méthode segmente l'espace des covariables de sorte à ce que chaque sous-groupe soit le plus identifiable possible vis-à-vis de la variable cible Y. On réalise alors un CART avec une unique variable ce qui aura pour conséquence de segmenter l'espace de cette dernière en maximisant le pouvoir explicatif de chaque région. Le découpage retenu aura alors un réel sens car les β de chaque modalité d'une même région seront par essence proches (minimisation de l'inertie intra-classe) alors que les β de modalités de régions différentes seront par nature plus éloignés (maximisation de l'inertie inter-classes). Néanmoins l'exposition reste dans cette méthode un problème majeure.

Notre choix se portera alors sur la seconde option qui offre des classes homogènes en terme d'exposition. L'application des CART ayant néanmoins une interprétabilité forte pourra être explorée dans le futur.

Une fois la méthode choisie, il est nécessaire, avant de pouvoir l'appliquer, de définir le nombre de classes que l'on souhaite créer pour les variables ayant trop de modalités. Par simplicité de modélisation, on souhaite déterminer un nombre unique et indépendant de la variable à discrétiser.

Afin d'avoir un ratio d'explicabilité/rapidité de calcul décent, on fixe le nombre de groupes par variable à 10. Ce choix pourrait faire l'objet d'une étude approfondie, mais étant donné le peu de documentation sur le sujet ainsi que les restrictions de temps, on choisit de ne pas y consacrer trop de ressources. Une démarche envisagée pour une potentielle amélioration serait de calibrer ce paramètre par validation croisée.

Ce mécanisme de discrétisation permet alors de retranscrire un spectre bien plus large de dépendances. A ce stade on peut cependant faire deux remarques.
La première étant qu'il est déjà possible de contourner la contrainte de linéarité à l'aide de méthode Spline (par le modèle GAM par exemple). En effet, il est possible de déformer le modèle pour forcer des dépendances polynomiales en ajoutant pour une variable $X_{,i}$ plusieurs variables de types $X_{,i}^k$. De cette manière on résout

$$g(\mathbb{E}[Y|X]) = \beta_0 + X_{,1}\beta_1 + X_{,1}^2\beta_2 + ... + X_{,1}^k\beta_k$$

ce qui nous permet de modéliser des dépendances polynomiales de n'importe quel degré.
L'ajout de ces nouvelles variables est possible sous R à l'aide du package Spline mais non utilisé dans ce projet. Néanmoins cette implémentation nécessite de définir pour chaque variable du modèle le degré du polynôme qui lui sera associé, ce qui peut se faire manuellement après étude de la forme de la dépendance, ou bien en un temps de calcul relativement long avec des méthodes automatiques. Notre méthodologie a l'avantage considérable d'accélérer ce procédé et de ne pas modéliser simplement des formes polynomiales (ou qui demanderaient un degré beaucoup trop élevé pour une approximation suffisante).

La seconde remarque que l'on peut apporter est que pour le moment il n'y a pas de nécessité de développer un quelconque algorithme, il suffit d'appliquer un GLM à la base de données nouvellement discrétisée ou bien un modèle GAM en acceptant un temps de calcul potentiellement long. En réalité la méthodologie que l'on propose est sensiblement plus rapide pour approximer les formes de dépendances car elle correspond à un GLM sur la base discrétisée. De plus en la développant nous même, on peut y ajouter un grand nombre de spécificité, comme la pénalité LASSO et une seconde contrainte permettant de réduire le sur-apprentissage. En effet les méthodes comme le GAM y sont bien plus sensibles car elles retranscrivent le lien entre les covariables et Y en se basant sur les données. C'est donc pour pallier ce défaut que nous introduisons la seconde spécificité de notre modèle, la pénalité de lissage.

3.2.2 Le lissage des coefficients

L'objectif de notre algorithme étant le pricing de produit d'assurance, il est crucial de rendre nos résultats de sortie interprétables. Pour ce faire, il est souhaitable que les coefficients β issues de modalités proches (par exemple deux âges qui diffèrent d'une unique année) ne soient pas trop éloignés. Si le β_{age_22} diffère grandement du β_{age_23}, un même individu pourra se voir proposer deux tarifs sensiblement différents si ceux-ci sont réalisés à quelques mois d'intervalle.

En pratique, le portefeuille de l'assureur étant découpé en segments au sein desquels les individus sont supposé similaires, ceux qui se "ressemblent" doivent se voir offrir un tarif équivalent. Malheureusement il est plus probable d'obtenir des coefficients volatiles avec notre méthode qu'avec un GLM classique, on parle alors de sur-apprentissage, ce qui augmente drastiquement la variance de notre

estimation.

Une forte exposition peut néanmoins réduire cette variance, cependant notre modélisation discrétise les variables, il est donc désormais nécessaire d'avoir une exposition suffisante sur toutes les modalités et non plus simplement un nombre d'observations globales suffisant. Le nombre d'enfants est une parfaite illustration de ce problème. il n'est pas improbable de travailler sur une base de données dans laquelle une et une seule famille possède 6 enfants. Le β associé à la variable nouvellement créée *nombre_d'enfants_6* sera alors déterminé par une unique observation et ne sera donc pas robuste. Étant plus sensible au sur-apprentissage, notre méthode aura alors tendance à générer des coefficients plus erratiques d'une modalité à l'autre. Pour éviter ces écarts trop importants il est nécessaire d'appliquer une contrainte de lissage.

L'objectif sera alors de forcer les β voisins à ne pas être trop disparates. On corrige alors le sur-apprentissage tout en rendant notre modèle plus interprétable. La contrainte mise en place est quadratique et prend la forme suivante

$$\lambda_2 \sum_{k=1}^{k=m} \sum_{j=2}^{p_k-2} (\beta_{j-1}^k + \beta_{j+1}^k - 2\beta_j^k)^2. \tag{3.1}$$

où p_k est le nombre de modalités de la nouvelle variable discrétisée à partir de la variable k et m le nombre de variables numériques ayant été discrétisées.
Cette contrainte s'applique évidemment uniquement sur des variables numériques car il n'y a pas de notion de proximité sur les modalités d'une variable qualitative, forcer une similarité entre les coefficients n'aurait donc pas de sens. De plus, il faut que les valeurs numériques aient une réelle signification, c'est à dire qu'elles soient associées à une métrique. Si l'on prend par exemple le cas de la variable Département *DEPT* de notre base, il est possible qu'elle soit codée au format numérique. Cependant aucune distance ne fait sens dans ce cas précis, les départements étant numérotés par ordre alphabétique et non par proximité géographique. Ils sont donc à exclure de cette seconde contrainte.

Pour que l'équation (3.1) fasse sens, il est nécessaire que les colonnes nouvellement créées aient été triées (i-e que la variable *age_18* soit positionnée juste avant *age_19*). On notera alors que si des modalités viennent à manquer dans notre jeu de données, la formulation (3.1) forcera un rapprochement entre deux modalités qui n'ont pas nécessairement de raison d'être proches. On prendra pour exemple extrême une base d'individus dont l'âge est soit égal à 18 soit à 80, il n'y a alors aucune raison pour que les deux coefficients liés à l'âge soient similaires mais notre modélisation ne pourra pas prendre ce cas de figure en compte. Dans la pratique il est très rare que l'on y soit confronté mais il est important de garder en mémoire ces particularités pour potentiellement analyser des résultats incohérents au premier abord. D'autant plus que si les modalités ont été regroupées en quantiles, l'étendue de ces derniers peut varier et biaiser la pertinence de notre contrainte. On pourrait alors pondérer la contrainte selon l'écart entre les modalités qui rentrent en jeu. Au vu des résultats de notre discrétisation, la prise en compte de cette difficulté ne semble pas nécessaire, on pourra cependant y trouver un axe d'amélioration futur.

La mise en place de cette contrainte permet donc de pallier nettement le problème de sur-apprentissage. Elle résout donc un problème majeur des modélisations complexes de type GAM et

justifie la nécessité de créer un algorithme de zéro pour le rendre plus flexible aux difficultés pratiques.

Notre modélisation présente donc des avantages théoriques considérables, dont notamment la capacité à reproduire une large variété de dépendances, complètement automatiquement, tout en restant interprétable et cohérent de par l'ajout d'une nouvelle contrainte. Enfin, l'ajout d'une pénalité de type LASSO permet de procéder à une sélection automatique des variables au cours de notre construction de tarif.

Notre modèle semble donc être un candidat pertinent pour répondre aux défauts de la méthode classique de tarification par GLM. Néanmoins, il est important de garder à l'esprit que la discrétisation des variables ainsi réalisée entraîne naturellement une perte d'information. En effet, dans le cas où le nombre de modalités est trop important, on ne gardera l'information que sur le quantile et non pas la valeur exacte. On prévoit alors que cette perte d'information soit compensée par la plus-value apportée par une telle modélisation. Il apparaît alors évident que sur une application où la dépendance est réellement linéaire il sera difficile de sur-performer le GLM. On réalisera donc une comparaison sur des données ne présentant pas cette particularité afin d'estimer réellement l'impact de notre modèle.

3.3 Résolution mathématique du problème de minimisation

On s'intéresse maintenant à l'aspect théorique de notre modélisation. Il est nécessaire dans un premier temps d'en présenter l'équation de minimisation afin d'en détacher une approche de résolution. On exposera ensuite la démarche d'approximation itérative mise en place.

3.3.1 Problème de minimisation et approche théorique

On procède classiquement par maximisation de la vraisemblance, à laquelle on vient ajouter une contrainte LASSO pour la sélection des variables ainsi que la contrainte de lissage évoquée précédemment. On travaille donc avec l'équation finale suivante

$$\hat{\beta} = \arg\min_{\beta}\{-\mathcal{L}(\beta) + \lambda_1||\beta||_1 \ + \lambda_2 \sum_{k=1}^{k=m} \sum_{j=2}^{p_k-2} (\beta_{j-1}^k + \beta_{j+1}^k - 2\beta_j^k)^2\}. \tag{3.2}$$

où $\mathcal{L}(\beta)$ représente la log-vraisemblance des observations.

Nous avons expliqué dans le chapitre précédent que les problèmes de minimisation pouvaient se résoudre par l'annulation de la dérivée de la fonction cible. Dans le cas de l'équation (3.2) la norme L^1 rend le calcul bien moins naturel. Il est nécessaire d'avoir recours à des méthodes plus complexes qui s'inspirent du sous-différentiel décrit dans le chapitre précédent. On présente donc ici la Proximal Gradient Descent qui est un outil d'optimisation pour des fonctions non nécessairement régulières et de grandes dimensions (Une grande partie des notions abordées dans cette partie sont fondées sur les deux cours suivants Parikh et Boyd, 2013, Tibshirani, 2013). Il s'agit d'une méthode plus abstraite

que les descentes de gradients classiques mais permettant d'optimiser un spectre plus large de fonctions, notamment non dérivables, et de traiter des données bien plus volumineuses.

On décrit alors dans cette partie les différents aspects de cette méthode. Elle repose sur une fonction appelée *proximal* qui se définit comme suit

$$\text{prox}_f(v) = \arg\min_x (f(x) + \frac{1}{2}||x - v||_2^2). \tag{3.3}$$

Il s'agit intuitivement du point qui se rapproche du minimum de f sans être trop éloigné de l'antécédent v.
La plupart du temps on sera amené à utiliser l'opérateur proximal de la fonction tf, pour $t \neq 0$, qui se réécrit $\text{prox}_{tf}(v) = \arg\min_x (tf(x) + \frac{1}{2}||x - v||_2^2) = \arg\min_x (f(x) + \frac{1}{2t}||x - v||_2^2)$.

On illustre son action dans le graphique suivant. Les points bleus correspondent aux antécédents, les points rouges aux images par la fonction proximal et le trait en gras représente la frontière du domaine de définition de f.

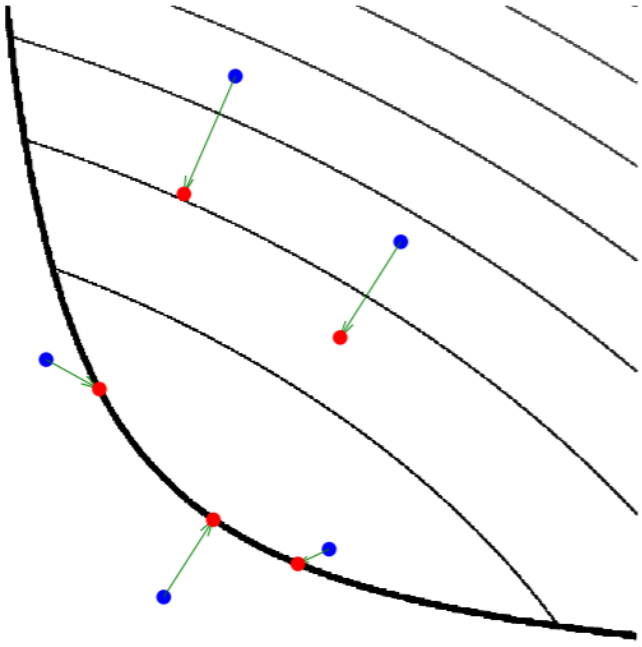


FIGURE 3.1: Application de la fonction proximal

Au plus t est grand au plus vite on se rapproche du minimum de f dans le sens où $\frac{1}{2t}$ pondère le poids du facteur $||x - v||_2^2$ dans la minimisation.

Cette fonction est intrinsèquement liée aux problèmes de minimisation, elle est en effet semblable à une étape de descente de gradient puisque sous certaines conditions on a, lorsque t est petit, l'équivalence suivante

$$\text{prox}_{tf}(v) \approx v - t\nabla f(v). \tag{3.4}$$

De plus, on va démontrer un résultat clé duquel découlent tous les algorithmes de type proximal,

Soit f sous différentiable sur sous domaine E et $v^\star \in E$, alors

$$prox_{tf}(v^\star) = v^\star \Leftrightarrow v^\star = \arg\min_{v \in E} f(v). \tag{3.5}$$

Démonstration :
$\Leftarrow$
Pour simplifier les calculs on suppose que f est sous différentiable sur son domaine de définition.
Soit $v^\star$ le minimisateur f :
$\forall v, f(v) > f(v^\star) \Rightarrow f(v) + \frac{1}{2}||v - v^\star||_2^2 > f(v^\star) = f(v^\star) + \frac{1}{2}||v^\star - v^\star||_2^2$
$v^\star$ minimise donc $f(v) + \frac{1}{2}||v - v^\star||_2^2$.
On a alors par définition que $v^\star = \text{prox}_f(v^\star)$.
$\Rightarrow$
On utilise la caractérisation du minimum d'une fonction convexe par le sous-différentiel :
$\widetilde{v} = \arg\min_v(f(v) + \frac{1}{2}||v - v^\star||_2^2) \Leftrightarrow 0 \in \partial f(\widetilde{v}) + (\widetilde{v} - v^\star)$.
On a ici $\widetilde{v} = \text{prox}_f(v^\star) = v^\star$ par hypothèse. On a donc $0 \in \partial f(\widetilde{v^\star})$, donc $v^\star$ minimise f par caractérisation du sous-différentiel.
Le sous-différentiel de f en x noté $\partial f(x)$ désigne l'ensemble des pentes de toutes les minorantes affines de f qui passent par le point $(x, f(x))$.

En somme, les points fixes de la fonction proximal sont les minimisateurs de f. On peut donc se restreindre à la recherche de points fixes de prox_{tf}. Si cette fonction était contractante (Lipschitzienne de constante strictement inférieur à 1), le théorème du point fixe de Banach aurait garanti que la suite définie par $x_{n+1} = \text{prox}_{tf}(x_n)$ converge vers son unique point fixe. Cette propriété est néanmoins trop restrictive, on en présente donc une plus souple, vérifiée par prox_{tf} qui conduira naturellement à un algorithme de minimisation

$$||\text{prox}_{\lambda f}(x) - \text{prox}_{\lambda f}(y)||_2^2 \leq (x - y)^T(\text{prox}_{\lambda f}(x) - \text{prox}_{\lambda f}(x)) \tag{3.6}$$

On ne prouvera pas cette propriété, appelée non expansivité forte mais on exploitera le fait que pour toute fonction f la vérifiant et admettant un point fixe x^*, la suite $x_{n+1} := f(x_n)$ converge vers x^*.
Tous les algorithmes de type Proximal Gradient Descent reposent donc sur le stratégie induite par cette propriété.

On va désormais approfondir la méthode implémentée. Elle diffère très légèrement de celle présentée précédemment mais elle repose néanmoins sur les mêmes concepts mathématiques.

3.3.2 Algorithme de descente de gradient proximal accélérée à pas adaptatif

On rappelle notre problème de minimisation (3.2) dans lequel $\beta^* = \arg\min_{\beta}(h(\beta))$ avec,

h défini par $h(\beta) = -\mathcal{L}(\beta) + \lambda_1||\beta||_1 \ + \lambda_2 \sum_{k=1}^{k=m} \sum_{j=2}^{p_k-2} (\beta_{j-1}^k + \beta_{j+1}^k - 2\beta_{j1}^k)^2$.
On décompose ainsi :

$$\begin{gathered} h(\beta) = f(\beta) + g(\beta), \\ \text{où } f(\beta) = -\mathcal{L}(\beta) + \lambda_2 \sum_{k=1}^{k=m} \sum_{j=2}^{p_k-2} (\beta_{j-1}^k + \beta_{j+1}^k - 2\beta_{j1}^k)^2 \\ \text{et } g(\beta) = \lambda_1||\beta||_1. \end{gathered}$$

La fonction f est différentiable et g est convexe.

La méthode de descente de gradient proximal que l'on choisit d'implémenter consiste à itérer jusqu'à convergence de la suite

$$\beta_{k+1} := \text{prox}_{t_k g}(\beta_k - t_k \nabla f(\beta_k)). \tag{3.7}$$

On y tire avantage de la partie différentiable pour accélérer la convergence. On applique simplement une descente de gradient sur f, puis on cherche un point proche de l'itération de la descente de gradient qui se rapproche du minimum de g.
On se sert de la caractérisation du sous-gradient pour démontrer la convergence de cette méthode.

$$\begin{aligned} x^* \in \arg\min_x(f(x) + g(x)) &\iff 0 \in \nabla f(x^*) + \partial g(x^*) \\ &\iff 0 \in t\nabla f(x^*) + t\partial g(x^*) + x^* - x^* \\ &\iff (Id + t\partial g)(x^*) \ni (Id - t\nabla f)(x^*) \\ &\iff x^* = (Id + t\partial g)^{-1}(Id - t\nabla f)(x^*) \\ &\iff x^* = \text{prox}_{tg}(x^* - t\nabla f(x^*)) \end{aligned}$$

La dernière équivalence vient du fait que $\text{prox}_{tg} = (Id + t\partial g)^{-1}$. En effet, en supposant g sous-différentiable sur son domaine par convenance

$$\begin{gathered} z \in (Id + t\partial g)^{-1}(x) \iff x \in (Id + t\partial g)(z) = z + t\partial g(z) \\ \iff 0 \in \partial g(z) + \tfrac{1}{t}(z - x) \iff 0 \in \partial_z(g(z) + \tfrac{1}{2t}||z - x||_2^2 \\ \iff z = \arg\min_u(g(u) + \tfrac{1}{2t}||u - x||_2^2) \end{gathered}$$

On a bien $z \in (Id + t\partial g)^{-1}(x) \iff z = \text{prox}_{tg}(x)$ ce qui en particulier montre que $(Id + t\partial g)^{-1}$ est à valeur unique justifiant ainsi le passage de l'appartenance à l'égalité au sein des deux démonstrations précédentes.
Il y a bien équivalence entre la recherche du minimisateur β^* et celle d'un point fixe de
$\text{prox}_{tg}(\beta^* - t\nabla f(\beta^*))$. On approche alors simplement ce point fixe par l'application récursive définit

en (3.7).

Bien que l'on ait démontrer que cette récurrence converge vers β^*, il peut sembler que l'on a artificiellement résolu le problème d'optimisation (3.2) en dissimulant la recherche du minimum au sein de la fonction *prox*. En réalité celle-ci a une forme explicite pour un grand nombre de fonctions, notamment pour la norme L^1. En effet, par un calcul succinct, il vient que

$$\text{prox}_{t\lambda||.||_1}(v) = \arg\min_x(\lambda||x||_1 + \tfrac{1}{2t}||x-v||_2^2) = S_{\lambda t}(v)$$
$$\text{tel que } [S_{\lambda t}(v)]_i = \begin{cases} v_i - \lambda & \text{si } v_i > \lambda \\ v_i + \lambda & \text{si } v_i < -\lambda \\ 0 & \text{sinon.} \end{cases}$$

$S_{\lambda t}(v)$ est l'opérateur dit de seuillage doux (ou soft-thresholding), pour lequel on remarque que lorsque l'itération de descente de gradient qui porte sur f a une composante inférieure à λ en valeur absolue, celle-ci sera fixée à zéro par l'opérateur proximal, d'où la sélection de variables.

Il est maintenant nécessaire de s'attarder sur la présence de t_k dans notre formule d'itération (3.7), celui-ci joue le rôle du pas dans la descente de gradient et également de poids dans la pondération de l'opérateur proximal. Il est indicé par k car la méthode implémentée présente un pas adaptatif. En effet le pas est un paramètre critique dans les problèmes d'optimisation algorithmique pouvant parfois empêcher la convergence vers le β^*. Afin d'améliorer notre algorithme et d'accélérer la convergence, on va chercher le pas "optimal" à chaque itération. Cette recherche ayant un coût, il est important de veiller à ce que celui-ci ne surpasse pas le gain de vitesse dû à l'obtention d'un pas plus performant. De ce fait, on autorise uniquement notre pas à décroître, l'idée étant d'en prendre un élevé au début pour vite se rapprocher du minimum et de le diminuer lorsque l'on se rapproche de β^*.
La méthode se définit de la manière suivante :

- Soit à l'itération k les paramètres $\beta_k, t := t_{k-1}$ et un facteur de rétrécissement τ.
- On répète la structure suivante jusqu'à la condition d'arrêt :
  1. On pose $z := \text{prox}_{tg}(\beta_k - t\nabla f(\beta_k))$.
  2. On pose la condition d'arrêt $f(z) < \hat{f}_t(z, \beta_k)$.
  3. On diminue le pas tel que $t := \tau t$.
- On retourne $t_k := t$ et $\beta_{k+1} := z$

Ici $\hat{f}_t(x,y) = f(y) + \nabla f(y)^T(x-y) + \frac{1}{2t}||x-y||_2^2$.

Enfin, dans l'optique d'accélérer la convergence tout en minimisant les calculs lourds, nous utilisons la version accélérée de cet algorithme. Elle consiste à diminuer le nombre d'itérations nécessaires à la convergence en laissant dans la composante β_k une trace des itérations passées en posant
$v_k := \beta_k + \frac{k}{k+3}(\beta_k - \beta_{k-1})$.

Le coefficient de poids $\frac{k}{k+3}$ est issu des travaux de Nesterov sur l'algorithme FISTA (Fast Iterative Soft-thresholding Algorithm).
Aux premières occurrences, le paramètre d'intérêt β_k varie grandement d'une itération à l'autre, le facteur $\beta_k - \beta_{k-1}$ correspondant à cette évolution sera alors très grand. Son ajout entraînera β_k vers la direction de l'évolution du passé, accentuant alors celle-ci. On se rapprochera donc plus rapidement du minimum. Plus tard, les itérations successives se rapprocheront de par la convergence, le facteur $\frac{k}{k+3}$ tendant vers 1, l'expression $\beta_k + \frac{k}{k+3}(\beta_k - \beta_{k-1})$ sera donc proche de β_k.
La relation d'itération, en considérant t_k le pas optimal calculé au préalable, est

$$\beta_{k+1} := \text{prox}_{t_k g}(v_k - t_k \nabla f(v_k)). \tag{3.8}$$

On représente ci-dessous l'impact d'une telle modification pour une itération fixée.

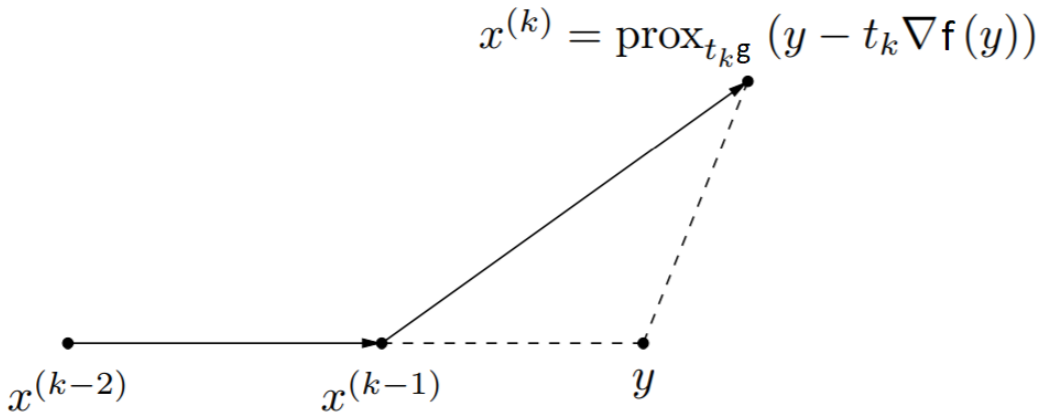


FIGURE 3.2: Elan porté par l'itération passée dans la minimisation

3.3.3 Implémentation en pratique et difficultés rencontrées

On a jusqu'ici présenté les outils nécessaires à la création de l'algorithme et la démarche mise en oeuvre. On souhaite dans cette partie décrire rapidement la structure de l'algorithme que nous avons implémenté afin d'en extraire des difficultés pratiques auxquelles nous avons fait face.

La première partie de notre code consiste à décrire les différentes lois prises en compte, en détaillant les sous-fonctions apparaissant dans l'écriture famille exponentielle. On récapitule les deux lois principales, que sont la loi de Poisson et la loi Gamma, dans le tableau suivant.

Loi	θ	$b(\theta)$	ϕ	$c(y,\phi)$	$a(\phi)$
$Gamma(a,b)$	$\frac{b}{a}$	$\log(\theta)$	$\frac{1}{a}$	$\frac{-\log(\phi)}{\phi} + (\frac{1}{\phi} - 1)\log(y) - \log(gamma(\frac{1}{\phi}))$	$-\phi$
$Poisson(\lambda)$	λ	e^{θ}	1	$-\log(y!)$	ϕ

TABLE 3.1: Récapitulatif des lois

Par la suite, l'algorithme sera appliqué aux fonctions $b(.), c(y,\phi), ...$ fixée.
On rappelle que notre approximation du problème (3.2) traite séparément la partie différentiable f de celle non différentiable g. Notre première étape consiste à calculer d'une part l'opérateur proximale de la norme L^1 et d'autre part le gradient de la partie différentiable, correspondant à l'opposé de la log-vraisemblance et à la contrainte associée à λ_2.
Le gradient de la seconde contrainte est assez complexe de par la multitude de cas à distinguer, par exemple si $p > 3$ il vaut

$$\frac{\partial}{\partial\beta_j}\sum_{i=2}^{p-2}(\beta_{i-1}+\beta_{i+1}-2\beta_i)^2 = \begin{cases} 2(\beta_1+\beta_3-2\beta_2) \text{ si } j=1 \\ -4(\beta_1+\beta_3-2\beta_2)+2(\beta_2+\beta_4-2\beta_3) \text{ si } j=2 \\ -4(\beta_{p-2}+\beta_p-2\beta_{p-1})+2(\beta_{p-3}+\beta_{p-1}-2\beta_{p-2}) \text{ si } j=p-1 \\ 2(\beta_{p-2}+\beta_p-2\beta_{p-1}) \text{ si } j=p \\ 2(\beta_j+\beta_{j+2}-2\beta_{j+1})+2(\beta_{j-2}+\beta_j-2\beta_{j-1})-4(\beta_{j-1}+\beta_{j+1}-2\beta_j) \text{ sinon} \end{cases}$$

Une fois ces fonctions codées, nous disposons de tous les outils nécessaires à la résolution de notre minimisation. Il est nécessaire, avant de faire appel à ces fonctions, d'adapter la base de données en la dichotomisant comme décrit précédemment. Cette discrétisation des variables entraîne alors des difficultés pratiques. Par exemple, lorsque l'on réalisera une validation croisée pour trouver les λ optimaux ou mesurer la performance de nos travaux il arrivera que notre division en folds laisse ressortir un fold pour lequel certaines variables seront identiquement nulles de par la sparsité de X. De ce fait, l'estimation du β_j associé aux modalités en questions sera impossible à optimiser. On fait alors le choix, quand ce problème se présente, de supprimer les colonnes problématique de notre modélisation le temps de l'optimisation du fold concerné. L'erreur globale sera peu affectée car cette situation ne concerne que très peu de modalités en pratique, les valeurs optimales des λ ne seront donc vraisemblablement pas significativement modifiées. De plus, si les performances du modèle semblent satisfaisantes, il sera possible de recalculer β^* à l'aide de la totalité de la base ce qui palliera ce problème.

Une difficulté supplémentaire en terme d'implémentation concerne les lois multi-paramétrées comme la loi Gamma. En effet, notre problème d'optimisation porte sur β (et donc sur θ), mais celui-ci ne contient pas l'intégralité de l'information nécessaire pour retrouver les deux paramètres de la loi. On remarque par exemple dans le tableau 3.1 que le paramètre a apparaît en dehors de θ, il serait donc nécessaire de le connaître pour calculer le gradient de la partie différentiable. Il est possible de le rajouter dans le problème d'optimisation mais la démarche couramment utilisée consiste plutôt à l'approcher par l'estimateur de Pearson (voir ALTMAN, 2019).

Pour finir, l'implémentation d'une variable offset dans la modélisation de la fréquence de sinistre fut relativement complexe. La documentation des packages existants omet en effet souvent les détails associés à cette spécificité que nous avons du implémenter manuellement, nous sommes tout de même parvenus à la modéliser correctement. L'exposition est donc ajoutée classiquement dans le modèle, avec un paramètre β forcé à un.

Chapitre 4

Mise en oeuvre de notre modèle et comparaison avec GLM et GAM

4.1 Introduction

Ce chapitre présente les résultats de notre algorithme sur la base décrite précédemment. On cherche à modéliser la prime pure par coût/fréquence et l'on comparera les résultats de notre modèle à ceux d'un GLM LASSO et d'un modèle GAM.

Afin de rendre cette comparaison cohérente il est nécessaire d'établir un protocole de comparabilité.
Tous les modèles ayant des hyperparamètres feront l'objet d'une validation croisée en amont permettant de les calibrer au mieux. C'est le cas notamment du GLM LASSO qui contient un hyperparamètre et de notre modélisation qui en contient deux. On ne s'attardera sur leur optimisation que dans le cas de nos travaux, celle concernant le GLM LASSO étant directement réalisée par les packages utilisés.
Cette optimisation se fera sur 80% de la base afin de pouvoir mesurer par la suite le sur-apprentissage. Les 20% restants n'ayant pas servi à l'apprentissage seront utilisés pour calculer trois métriques distinctes qui serviront de points de comparaison entre les différents modèles. Les bases de test et d'apprentissage seront identiques d'un modèle à l'autre grâce à l'utilisation d'une graine en début de code.

Les métriques choisies permettent de rendre compte au mieux de la pertinence du modèle. On calculera ainsi

- **La déviance :** elle s'exprime comme la différence entre la vraisemblance du modèle estimé et celle du modèle saturé. C'est une généralisation du calcul de la somme des carrés résiduels dans le cas où la méthode d'optimisation maximise la vraisemblance. Elle se calcule par

$$D(Y, \hat{\mu}) = -2(\mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta^*)).$$

  La seconde vraisemblance est calculée en remplaçant la prédiction $g^{-1}(X\beta)$ par la vraie valeur de Y.

C'est un candidat de métrique de comparaison parfait puisque l'écart de déviance entre les deux modèles suit une loi du χ^2 sous l'hypothèse que le second modèle est le vrai, ce qui permet de sélectionner le meilleur modèle par tests statistiques.

Plus la déviance est faible, meilleur sera le modèle (par construction c'est une quantité toujours positive).

- **La RMSE (Root-Mean-Square Error) :** qui correspond à la racine de la moyenne des erreurs quadratiques. Elle permet de valider la justesse des prédictions, néanmoins dans le cadre d'un modèle où l'on cherche à estimer la moyenne elle sera naturellement élevée. Elle se calcule de la manière suivante

$$RMSE = \sqrt{\frac{1}{n}\sum(Y_i - \hat{Y}_i)^2}.$$

- **La MAE (Mean-Absolute Error) :** qui est semblable à la RMSE mais se calcule par

$$RMSE = \frac{1}{n}\sum|Y_i - \hat{Y}_i|.$$

On présente donc dans ce chapitre les résultats de ces métriques à la fois pour le modèle de coût et de fréquence pour le modèle GLM LASSO, GAM ainsi que pour notre modélisation.

4.2 Application du modèle GLM LASSO

On commence par exposer les résultats du modèle de coût pour le modèle GLM LASSO. Celui-ci sera calculé à l'aide du package *h2o* (LeDell, 2016) car bien que l'utilisation du package *glmnet* (Friedman et al., 2010) soit plus courante, elle ne permet pas de modéliser une loi Gamma. On optimise alors λ par validation croisé de manière automatique. Sa valeur optimale est de 0.00095. Le processus permettant cette optimisation sera plus détaillé pour le modèle de fréquence car les fonctions utilisées disposent de sorties graphiques claires. Le tableau ci-après montre les performances du modèle pour ce λ optimal.

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	6695.62	2245.60	1191.93
Echantillon de test	1743.30	2213.48	1178.06

Table 4.1: Résultats d'un GLM LASSO sur le modèle de coût moyen

Les deux dernières métriques affichent des valeurs élevées, elles correspondent à un calcul des erreurs de prédiction qui, dans le cas d'un GLM, sont extrêmement grandes. En effet la seule prédiction que ce type de modèle peut faire est l'espérance qui pour une variable avec un écart type d'environ 2000 est souvent très éloignée de la valeur réelle. On rappelle que cette modélisation fait cependant sens car l'on s'intéresse à l'espérance du risque.
Le sur-apprentissage semble assez faible, en effet la RMSE et la MAE sont très stables entre les échantillons de test et d'apprentissage. C'est bien l'écart d'une même métrique calculée à la fois sur la base d'apprentissage et sur celle de test qui permet de mesurer le sur-apprenstissage. Les métriques calculées sur la base d'apprentissage étant par construction d'une valeur assez faible, car le modèle à

été calibrer pour minimiser l'une d'entre elles, si l'ordre de grandeur des valeurs issues de la base de test est similaire on en conclut que notre modèle se généralise bien à de nouvelles données.
La déviance est une mesure qui dépend du nombre de données, celle de la base de test sera donc naturellement plus faible car calculée sur moins d'individus. La base d'apprentissage comportant 80% des lignes, elle est 4 fois plus grande ; une simple remise à l'échelle indicative permet de voir qu'il y a un écart de 277 ($= 4 \times 1743 - 6695$) entre les deux valeurs à nombre d'individus égal. Ce sont bien les données ayant servi à l'apprentissage qui présentent la déviance la plus faible.

On analyse maintenant les résultats issus de la modélisation de la fréquence par GLM LASSO. On rappel que dans la modélisation de la fréquence, la variable exposition doit être considérée comme une variable offset, ce qui est rendu directement possible via les package utilisés et qui a été implémenté également dans le cas de notre méthode. L'optimisation du paramètre λ a ici été réalisée par la fonction *cv.glmnet*. Son utilisation est plus courante que la méthode analogue proposée dans le package *h2o* et prend cette fois en compte la loi Poisson. On recherche comme indiqué précédemment le meilleur hyperparamètre par validation croisée, résultat que l'on illustre par le graphique suivant

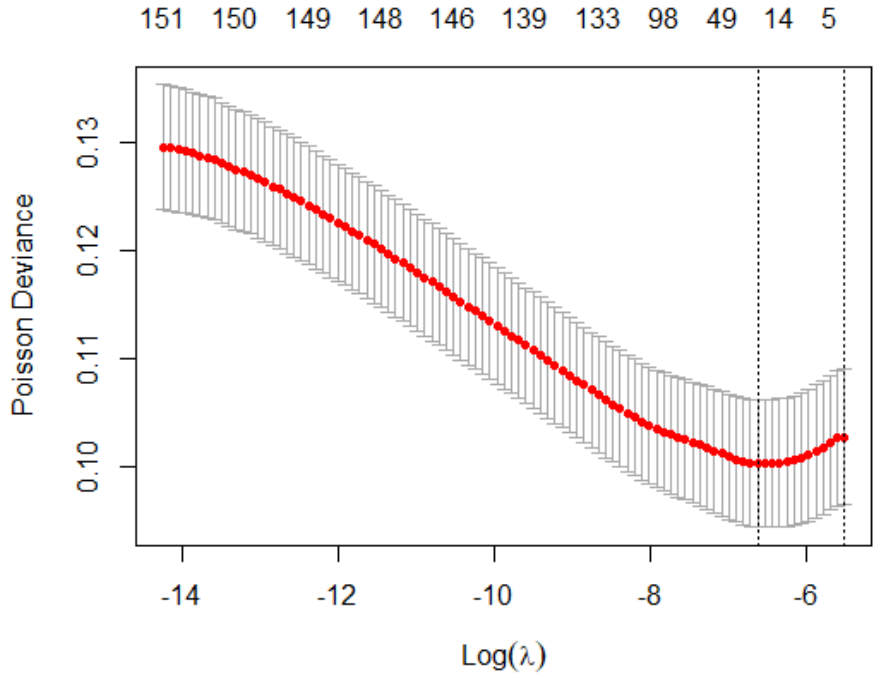


FIGURE 4.1: Recherche du lambda optimal par validation croisée

La métrique de sélection utilisée est la déviance, ce sera le même critère que l'on choisira pour notre propre recherche d'hyperparamètres. La forme en **U** est caractéristique de ces applications, on cherchera alors une plage suffisamment large de valeurs de λ pour que la fonction objectif ait le temps de décroître puis de recroître tout en choisissant un pas suffisamment fin pour approcher au mieux la valeur optimale.

Pour finir, on présente le tableau récapitulatif des performances du GLM LASSO. Celui-ci servira de point d'appui aux comparaisons avec les autres modèles.

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1397.60	0.11	0.02
Echantillon de test	442.29	0.12	0.02

TABLE 4.2: Résultats d'un GLM LASSO sur le modèle de fréquence appliqué à 20 000 individus

Le sur-apprentissage semble encore une fois assez faible pour les même raisons que précédemment.
On va maintenant présenter les résultats issus du modèle GAM.

4.3 Application du modèle GAM

Cette section permettra de mettre en exergue l'intérêt de notre modélisation en comparaison au GLM LASSO.

On rappelle que le modèle GAM (General Additiv Model) est une extension du GLM dans lequel la dépendance entre les covariables et la réponse Y n'est plus nécessairement modélisée linéairement. Ainsi, on part de l'hypothèse suivante

$$g(\mathbb{E}[Y|X]) = \beta_0 + f_1(X_{,1}) + ... + f_p(X_{,p}). \tag{4.1}$$

On réalise alors à l'aide de la fonction $gam()$ cette modélisation sur notre base de donnée avant discrétisation.
Les sorties R du package *gam* (HASTIE, 2019) permettent de voir l'allure des fonctions f_i approximées pour chaque variable. De cette manière, si celles-ci n'apparaissent pas comme linéaires, il sera probable qu'une modélisation de type GLM ne soit pas la plus adaptée. On va donc exposer les résultats de cette modélisation, à la fois pour le modèle de coût et pour le modèle de fréquence.

4.3.1 Modèle de coût moyen

La performance du modèle est calculée avec la métrique décrite précédemment.

On s'intéresse cependant ici surtout aux sorties graphiques. On présente dans un premier temps deux cas extrêmes.
Le premier concerne la variable *ANCIENNETE* dont la forme de la fonction f approché est quasi-linéaire.

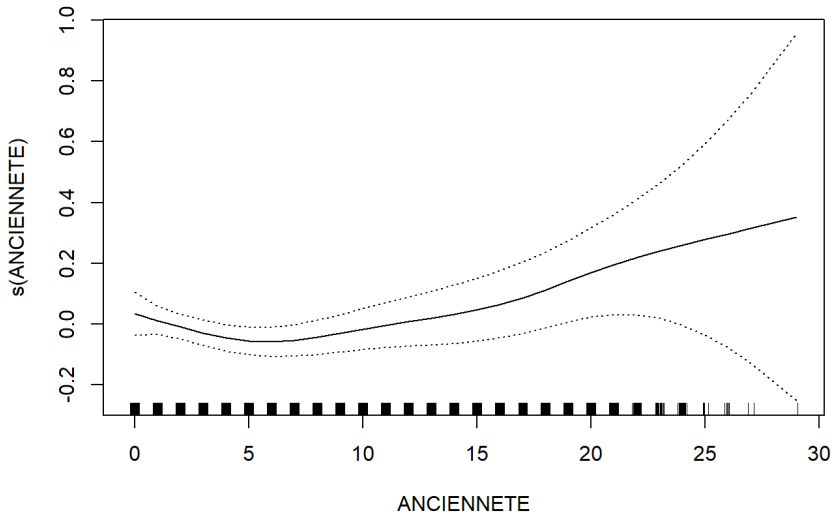


FIGURE 4.2: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et l'ancienneté

On peut donc aisément supposer que cette variable serait parfaitement adaptée à une modélisation GLM. Les coefficients associés à ce type de variables sont sensiblement les mêmes que ceux produits par le GLM.
Le second concerne la variable *AnneeNaiss* pour laquelle la dépendance n'est absolument pas linéaire.

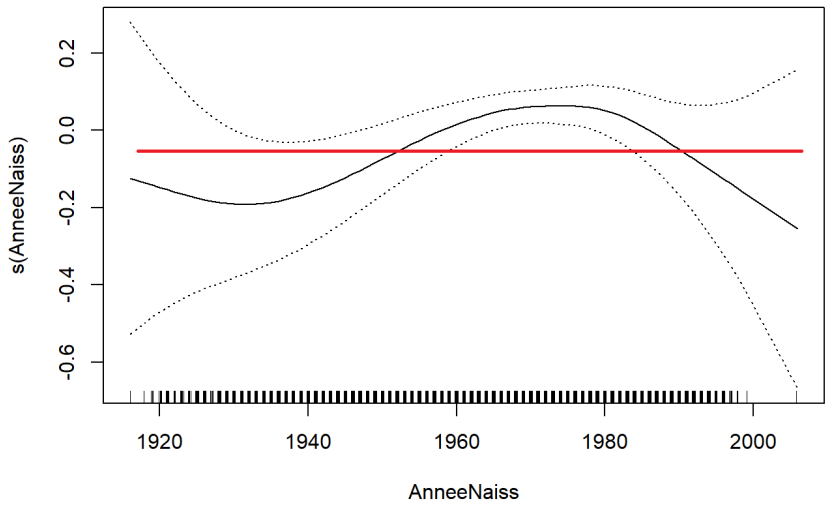


FIGURE 4.3: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et l'année de naissance de l'assuré

On y a ajouté la droite d'approximation affine dont la pente est supposée correspondre au coefficient du GLM. On remarque alors que malgré le fait que cette variable soit significative au seuil de 5%, elle était associée à un coefficient nul et donc jugée non significative dans la modélisation classique GLM.

On fournit en annexe A.5 la plupart des graphiques de sortie du modèle GAM pour le modèle de coût . On y observe que pour une grande partie des variables la dépendance n'est pas linéaire. Malgré tout son approximation par une droite reste dans quelques cas assez satisfaisante. Le capital total assuré est cependant assez éloigné d'une forme linéaire et est une variable explicative sur le modèle de coût ce qui justifie une nouvelle fois la recherche de modèles plus complexes.

Pour conclure sur l'application du modèle GAM pour le coût moyen, on s'attarde sur l'avant dernier graphique représentant la variable *NBSIN*.

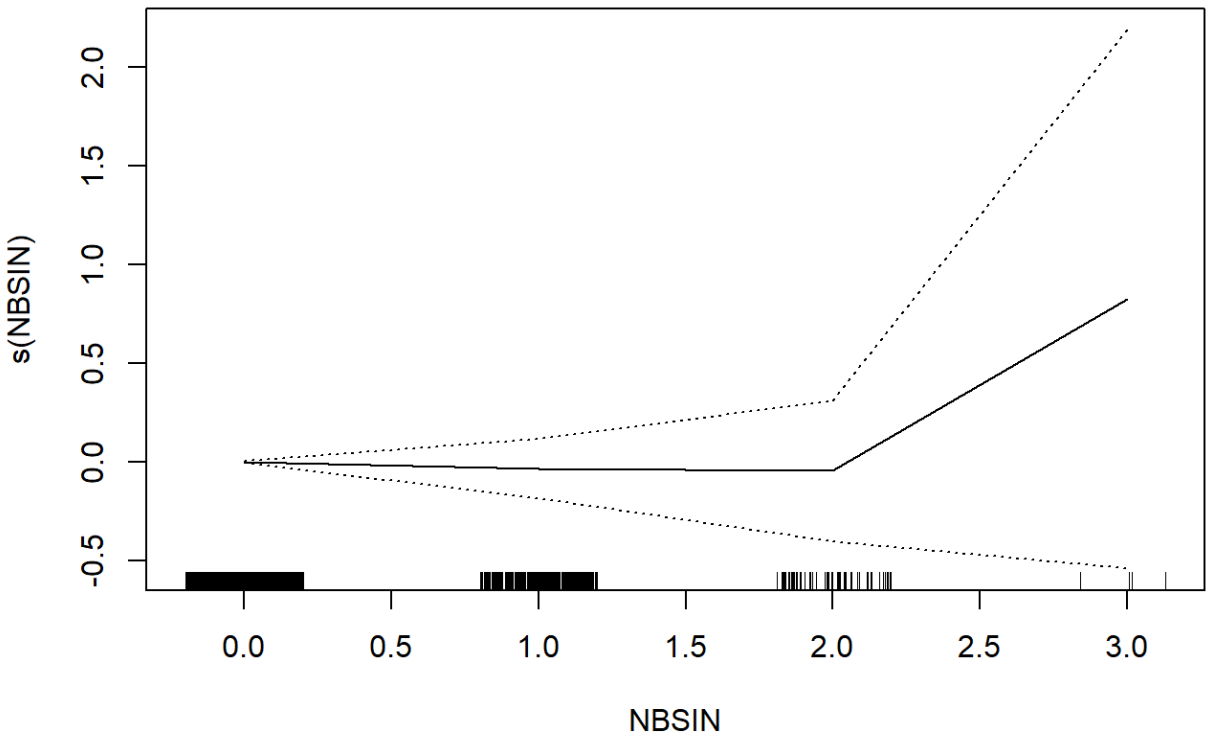


FIGURE 4.4: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et le nombre de sinistres déclaré sur le contrat antérieur

Celle-ci présente un décrochage très net à la valeur 2. La tendance semble dans un premier temps presque nulle puis augmente drastiquement pour les individus ayant enregistré 3 sinistres sur leur contrat précédent. Si l'on s'attarde sur la fréquence des modalités, représentée par l'épaisseur des traits sur l'axe des abscisses, on remarque qu'il y a très peu d'exposition au delà de la valeur 2 com-

parativement au reste. Il n'y a en réalité que 4 individus qui ont un nombre de sinistres antérieurs supérieur ou égal à 3. Si un seul d'entre eux a subi un sinistre lourd, ce modèle va sur-apprendre des données et présenter, comme c'est le cas ici, des résultats peu convaincants et peu interprétables. En effet, dans notre base, un de ces 4 individus a subi un sinistre d'un montant de 10 000, ce qui est plus de six fois supérieur à la moyenne expliquant ainsi l'allure du graphique. Cette remarque met alors en exergue l'utilité concrète de notre seconde pénalité dans la modélisation de nos données.

On présente à titre indicatif le tableau de performance de cette modélisation.

	Déviance	RMSE	MAE
Echantillon d'apprentissage	6536.81	2229.53	1191.60
Echantillon de test	1777.76	2226.92	1190.73

TABLE 4.3: Résultat sur le modèle de coût pour le modèle GAM

4.3.2 Modèle de fréquence

De la même manière on va présenter en détail deux graphiques issus du modèle de fréquence avant d'exposer les résultats de façon plus générale.

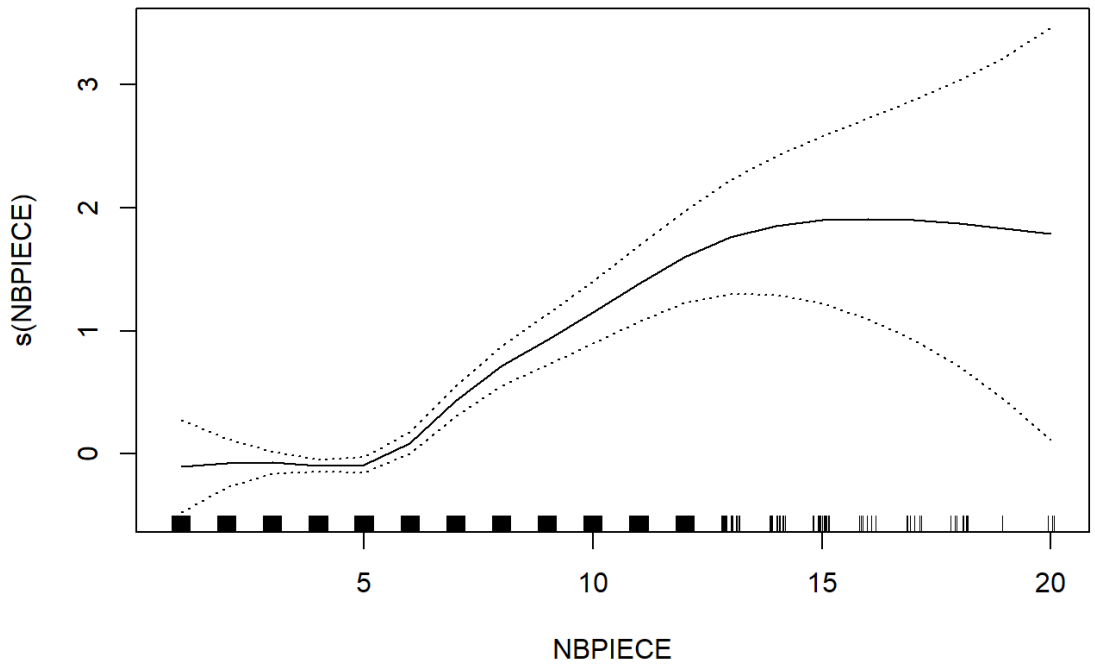


FIGURE 4.5: Fonction de dépendance approché par le GAM pour le lien entre la fréquence de sinistralité et le nombre de pièces du bien

Il apparaît clairement que cette dépendance n'est pas linéaire, elle est plus semblable à une fonction en escalier. De 1 à 5 pièces la fréquence semble stable, mais à l'ajout d'une pièce supplémentaire on remarque un décrochage net. Ce phénomène est très facilement interprétable. En effet ce sont les salles d'eaux qui sont la plupart du temps la cause d'un sinistre en garantie dégât des eaux. Néanmoins il est évident que lorsqu'un bien dispose déjà d'une salle de bain, la transition de 3 à 4 pièces sera bien plus probablement due à l'ajout d'une chambre qu'à celui d'une pièce d'eau.

A contrario, la variable *AnneeEff* représentant l'année du contrat semble jouer un rôle sensiblement linéaire dans la prédiction de la fréquence.

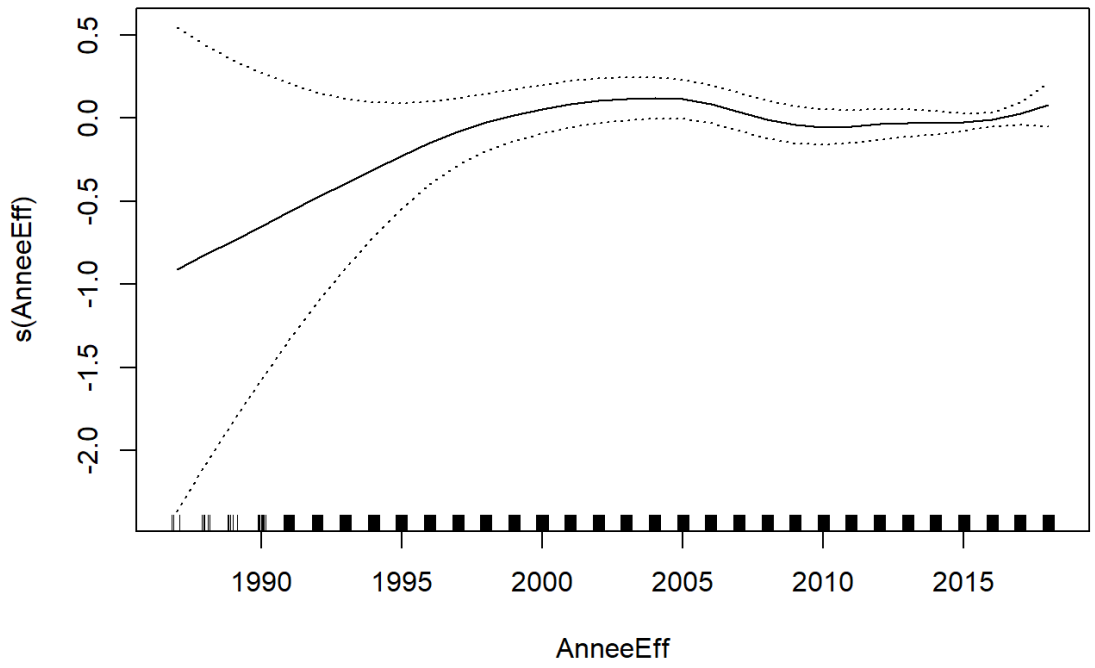


FIGURE 4.6: Fonction de dépendance approché par le GAM pour le lien entre la fréquence de sinistralité et l'année effective du contrat

Le léger pic observé aux années 2002-2003 correspond probablement aux inondations les plus marquantes de la décennie à savoir celle du Gard en septembre 2002 et celles de décembre 2003 dues à un débordement du Rhône et de la Loire.
Une fois de plus la totalité des graphiques est disponible en annexe A.5.2.

Pour finir, on présente le tableau des performances du modèle GAM à travers les trois métriques exposées précédemment.

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1222.11	0.11	0.02
Echantillon de test	563.45	0.12	0.02

TABLE 4.4: Résultat sur le modèle de fréquence avec variable offset sur 20 000 individus pour le modèle GAM

On remarque que le sur-apprentissage est assez faible car la MAE et la RMSE sont assez similaires sur les deux jeux de données. Néanmoins le GLM LASSO semble être plus performant que le GAM sur le modèle de fréquence étant donné que la déviance de ce dernier est plus élevée.
Une fois de plus la présentation de ce modèle sert principalement à appuyer l'intérêt de nos travaux en montrant les limites concrètes de la modélisation classique. les performances de cette dernière étant similaires ou supérieures à celles du GAM, on espère pouvoir tirer partie de la plus grande paramétrisation de notre modélisation pour surpasser le GLM LASSO.

4.4 Application de la méthode développée

4.4.1 Application au modèle de coût

L'objectif est ici d'appliquer notre algorithme en prenant comme variable Y le coût moyen des sinistres $COUT\_DDE$. Pour ce faire dès qu'une ligne est associée à plusieurs sinistres on modifie la variable d'intérêt de telle sorte qu'elle reflète le coût d'un unique sinistre. Ainsi, $\widetilde{Cout\_DDE} = \frac{COUT\_DDE}{FREQ\_DDE}$.

Notre modélisation présentant deux contraintes, il convient de déterminer les valeurs optimales des λ associées. Pour ce faire on réalise une validation croisée à 5 folds. Afin de rendre nos résultats comparables entre les différentes valeurs de λ ainsi qu'entre les différents modèles, on utilise la déviance standardisée comme métrique. On présente dans un premier temps les résultats de notre validation croisée.

lambda1	lambda2	deviance	MAE	RMSE
0.000001	0.00050	1485.25578	1240.15715	2284.56162
0.000001	0.00100	1515.66546	1264.00342	2300.87744
		...		
0.000001	0.00950	1480.09950	1237.20520	2283.91400
0.000001	0.01000	1479.99139	1237.34260	2283.94194
0.00011	0.00050	1465.39889	1230.26917	2277.73524
0.00011	0.00100	1483.68106	1243.55237	2285.49118
		...		
0.00011	0.00950	1462.63984	1228.89252	2277.33716
0.00011	0.01000	1470.54400	1226.50981	2278.60000
		...		
		...		
0.00089	0.00050	1415.51665	1204.78875	2260.62796
0.00089	0.00100	1423.49685	1214.33370	2264.85722
		...		
0.00089	0.00950	1422.74938	1201.55459	2261.61131
0.00089	0.01000	1417.21999	1198.21526	2259.22342
0.00095	0.00050	1412.71456	1203.07529	2259.34763
0.00095	0.00100	1416.70875	1209.70784	2261.50811
		...		
0.00095	0.00950	1418.43457	1202.31295	2260.73016
0.00095	0.01000	1417.54683	1200.79168	2260.42203
0.00100	0.00050	1413.67181	1201.35349	2259.29319
0.00100	0.00100	1421.80328	1210.56504	2262.95914
		...		
0.00100	0.00900	1428.03419	1195.29206	2263.86451
0.00100	0.00950	1425.04430	1193.78674	2262.17266

TABLE 4.5: Tableau restreint de la validation croisée du modèle de coût sur le couple (λ_1, λ_2)

On a simplement supprimé des lignes pour rendre le tableau plus lisible mais sa forme complète est disponible en annexe (A.1). La déviance minimale est atteinte en $(\lambda_1^* = 0.00095, \lambda_2^* = 0.00050)$ et vaut 1412.71456. On a par la suite recalculé pour des valeurs plus petites de λ_2 la valeur de la déviance pour s'assurer que notre valeur était le minimisateur, ce qui est le cas.

On va donc faire apprendre notre modèle sur la base d'apprentissage avec ce couple $(\lambda_1^*, \lambda_2^*)$ et calculer les métriques sur la base d'apprentissage puis sur celle de test. On présente les résultats dans le tableau suivant

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	6642.00	2240.91	1182.49
Echantillon de test	1743.28	2212.96	1173.01

TABLE 4.6: Résultats de notre modélisation sur le modèle de coût moyen

Le sur-apprentissage est assez faible. On remarque que l'on surpasse le modèle GAM. En ce qui concerne la confrontation avec le modèle GLM, notre déviance sur la base de test est quasiment identique (on ne la bat que de 0.02). Néanmoins les autres métriques, comme la MAE semblent être légèrement meilleures avec notre modélisaton. On remarquera des résultats sensiblement plus convaincants sur la modélisation de la fréquence.

Afin de visualiser en pratique l'impact de nos choix de modélisation, on expose ci-après deux graphiques confrontant le coût moyen observé au coût moyen prédit par notre modèle.

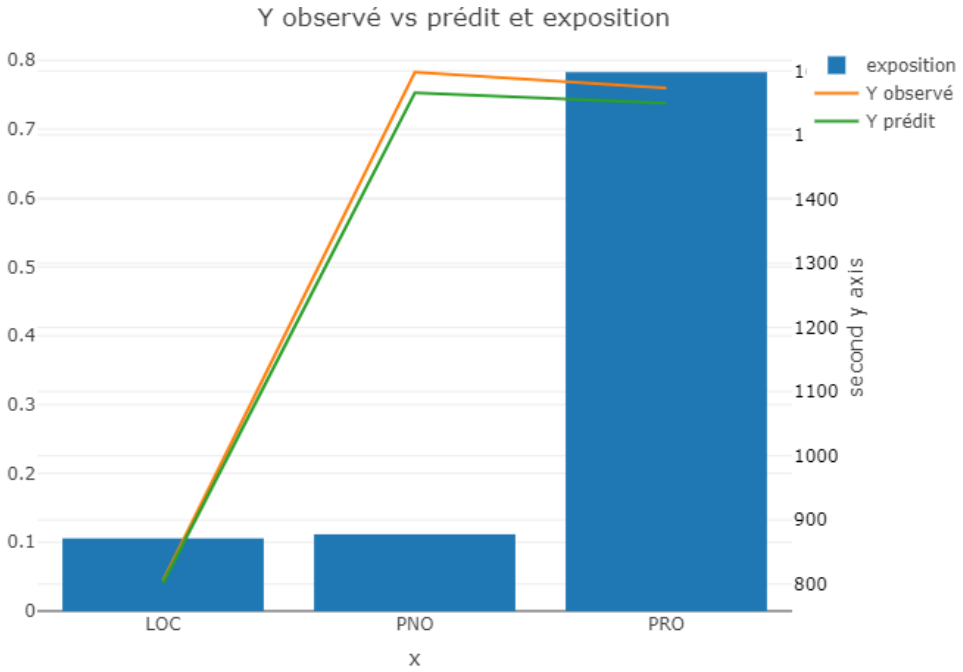


FIGURE 4.7: Coût moyen prédit contre observé pour la variables QUALIT

On commence par présenter ces résultats pour la variable *Qualit* qui indique si le bénéficiaire est propriétaire ou locataire. C'est une des variables les plus discriminantes habituellement en MRH ce qui nous pousse à nous y intéresser. Le diagramme en barre indique la proportion d'individus appartenant aux classes décrites en abscisses. Celle-ci est associée à l'ordonnée principale à gauche. Les courbes orange et verte représentent respectivement la moyenne observé et la moyenne prédite pour la variable d'intérêt modalité par modalité. Les deux courbes sont suffisamment proches pour que l'on conclue en un bon ajustement de notre modèle.

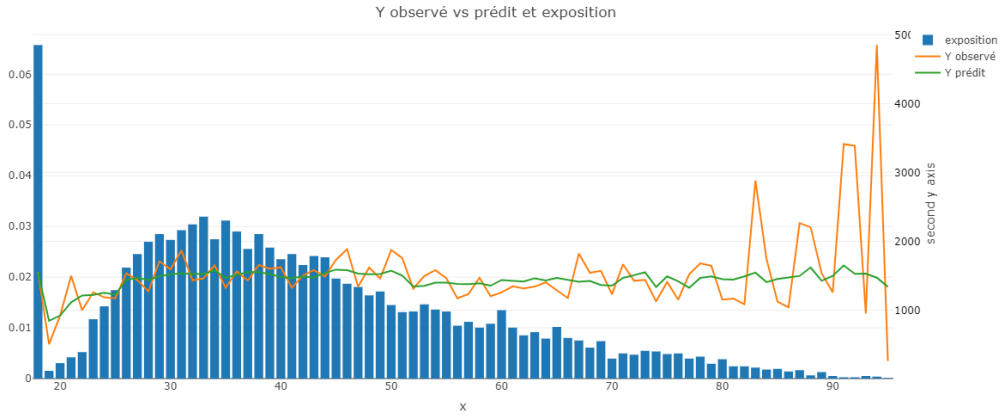


FIGURE 4.8: Coût moyen prédit contre observé pour la variables AGEOCC

Ce second graphique nous permet de visualiser l'impact de la seconde pénalité ajoutée dans notre modèle. On y remarque clairement que la courbe des prédictions est une version lissée de la courbe des observés. Cela vient du fait que les coefficients proches ont été lissés pour ne pas présenter des valeurs trop distinctes (soit par la seconde contrainte soit par le fait d'avoir le même coefficient pour plusieurs modalités au sein d'un même quantile.
De plus on s'aperçoit que plus une modalité est représentée, plus sa prédiction sera proche de la réalité. En effet l'écart entre les deux courbes est plus grand pour les valeurs extrêmes d'âge qui sont, comme l'indique le diagramme en barre, très peu représentées. La volatilité de la courbe des observées s'explique par ce manque d'effectif, problème que notre modèle corrige en lissant les effets.

On applique maintenant notre méthode à la variable de fréquence.

4.4.2 Application au modèle de fréquence

De la même manière que précédemment, on applique ici notre algorithme afin de prédire la fréquence d'apparition d'un sinistre. On rappelle que la variable exposition est traitée comme une variable offset. La validation croisée a été effectuée après de longues recherches en amont sur une plage optimale de recherche afin de faire moins de calculs. Le tableau complet est également disponible en annexe (A.2).

lambda1	lambda2	deviance	MAE	RMSE
0.000001	0.000500	391.641033	0.022155	0.113762
0.000001	0.001625	369.473420	0.023164	0.113467
0.000001	0.002750	371.304258	0.023161	0.113505
		...		
		...		
0.000100	0.000500	301.994307	0.022102	0.112003
0.000100	0.001625	301.890727	0.023179	0.112009
0.000100	0.002750	301.974331	0.023106	0.111993
0.000500	0.000500	307.465660	0.023171	0.112076
0.000500	0.001625	307.417767	0.023212	0.112071
0.000500	0.002750	306.850662	0.023106	0.112058
0.001556	0.000500	312.242608	0.023233	0.112181
0.001556	0.001625	312.242608	0.023233	0.112181
0.001556	0.002750	312.242608	0.023233	0.112181
0.002611	0.000500	313.100603	0.023268	0.112200
0.002611	0.001625	313.100603	0.023268	0.112200
0.002611	0.002750	313.100603	0.023268	0.112200
		...		
		...		

TABLE 4.7: Tableau restreint de la validation croisée du modèle de fréquence sur le couple (λ_1, λ_2)

Une fois ces valeurs trouvées on apprend de nouveau notre modèle sur la base d'apprentissage puis on calcule les métriques.

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1549.44	0.11	0.02
Echantillon de test	332.22	0.10	0.02

TABLE 4.8: Résultats de notre modélisation sur la fréquence appliquée à 20 000 individus

On remarque que le sur-apprentissage est très faible, voire inexistant. En effet la déviance calculée sur l'échantillon de test et remise à l'échelle en multipliant par 4 est inférieure à celle calculée sur la base d'apprentissage. Ce calcul est indicatif mais permet néanmoins de déterminer l'importance du sur-apprentissage. On remarque surtout à travers ces résultats que notre modélisation a une déviance plus faible sur l'échantillon de test que le GLM LASSO. C'est bien cette déviance qu'il faut prendre en compte pour comparer les performances des modèles, on peut donc conclure que nos travaux permettent un gain sensible de performance.

On présente pour finir des graphiques similaires à ceux évoqués dans la modélisation du coût moyen.

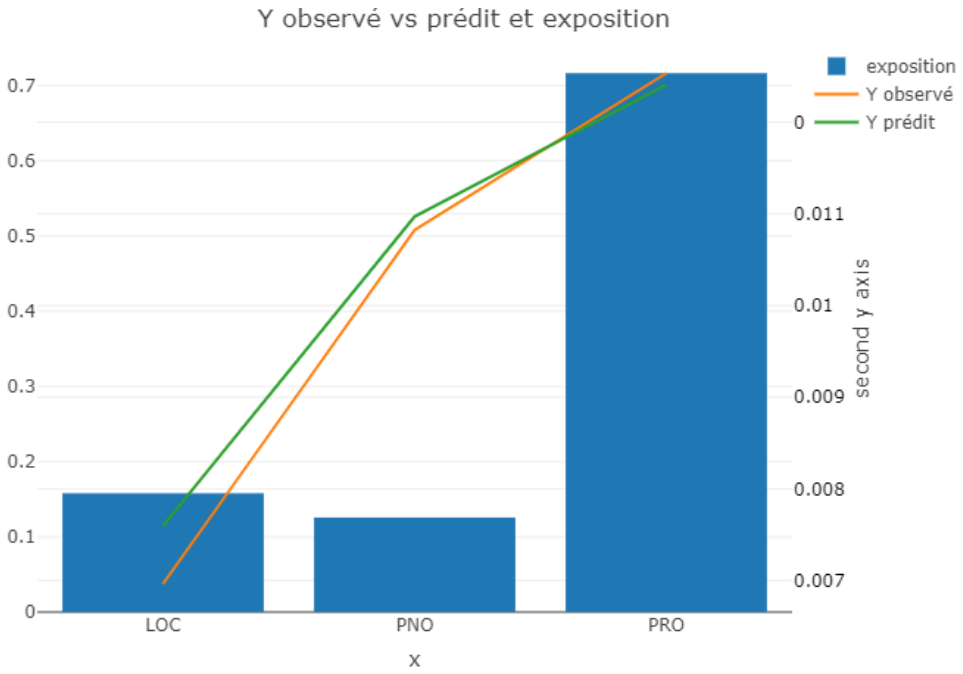


FIGURE 4.9: Fréquence prédite contre observée pour la variables QUALIT

Une fois de plus la prédiction pour une des variables les plus couramment explicative en MRH semble plus que satisfaisante.

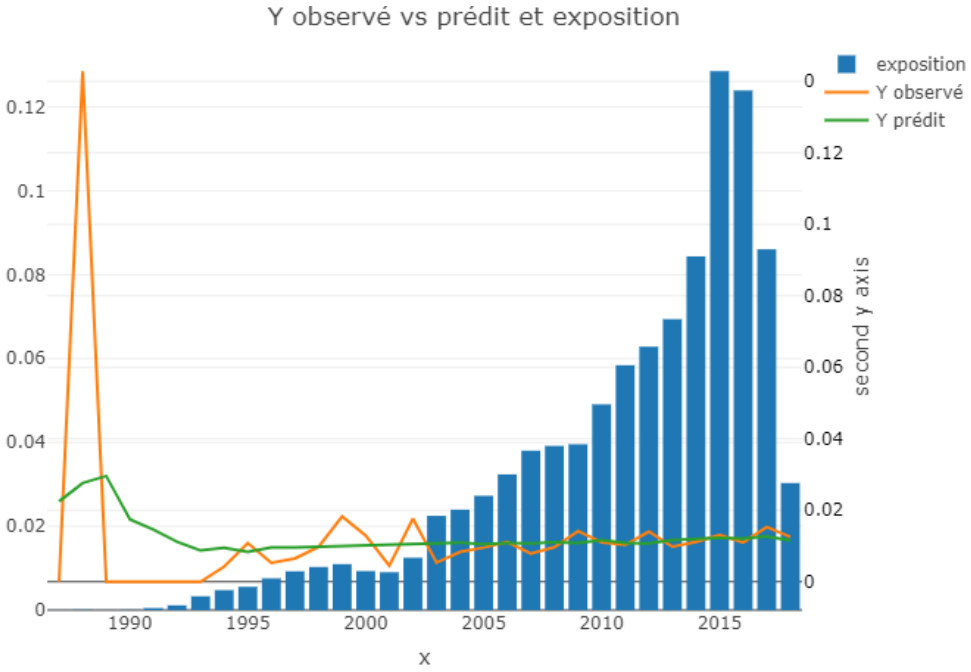


FIGURE 4.10: Fréquence prédite contre observée pour la variables AnneeEff

On observe encore le lissage effectué par notre modèle qui s'interprète d'autant mieux pour les modalités peu représentées. On retrouve également, comme remarqué dans le modèle GAM, l'aspect quasi-linéaire de cette variable ce qui appuie la cohérence de nos travaux.

4.5 Récapitulatif des différents résultats

Pour récapituler on a obtenu, dans le modèle de coût moyen, les performances suivantes.

- Pour le GLM

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	6695.62	2245.60	1191.93
Echantillon de test	1743.30	2213.48	1178.06

TABLE 4.1: Résultats d'un GLM LASSO sur le modèle de coût moyen

- Pour le GAM

	Déviance	RMSE	MAE
Echantillon d'apprentissage	6536.81	2229.53	1191.60
Echantillon de test	1777.76	2226.92	1190.73

TABLE 4.3: Résultat sur le modèle de coût pour le modèle GAM

- Pour notre modélisation

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	6642.00	2240.91	1182.49
Echantillon de test	1743.28	2212.96	1173.01

TABLE 4.5: Résultats de notre modélisation sur le modèle de coût moyen

Le problème de sur-apprentissage ne semble jamais se poser. Le GAM sous-performe les deux autres modélisations. La modélisation présentée tout au long de ce mémoire semble aussi performante que le GLM.
Il semble difficile de parvenir à comparer les coefficients prédits entre les modèles étant donné que nous avons discrétisés l'intégralité de la base de données d'autant plus que le GAM ne présente pas la même structure que les autres modèles.
On peut néanmoins s'attarder sur quelques coefficients. Celui de la variable *QUALIT* semble être un candidat intéressant de par son explicabilité. Etant une variable discrète, elle a de plus été discrétisée de la même manière dans le GLM et dans notre algorithme. La modalité de référence est dans les deux cas *LOC* correspondant à un bénéficiaire locataire. Pour le GLM, la modalité Propriétaire est associée à un coefficient de -0.62 et de 0.20 pour la modalité *PNO*.
On présente dans le graphique suivant la valeur des coefficients approchés par notre algorithme.

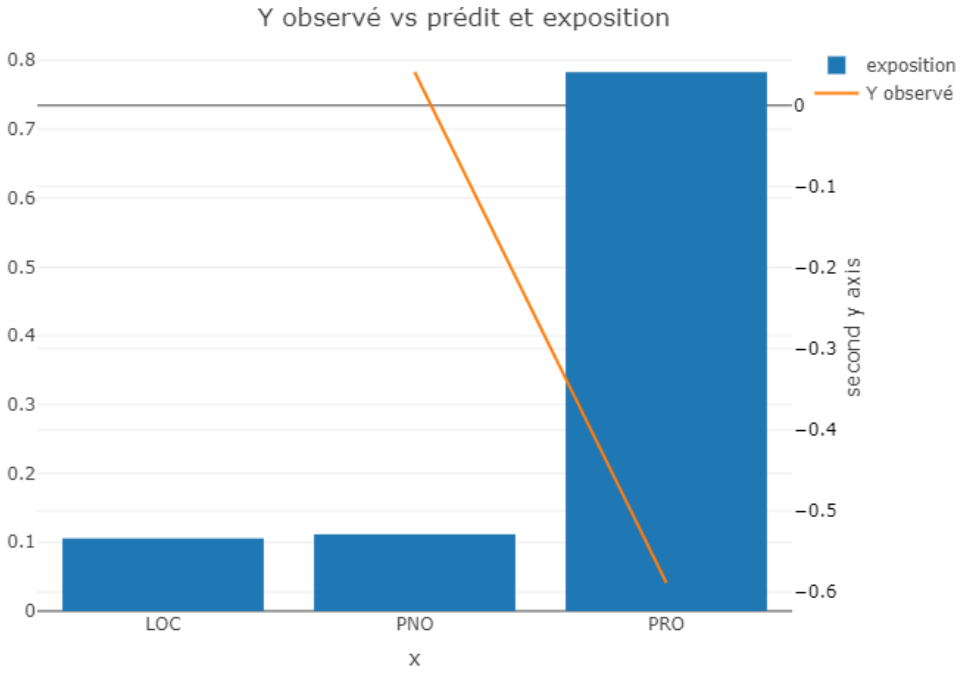


FIGURE 4.11: Coefficients estimés pour la modélisation du coût moyen pour la variable QUALIT

Le coefficient associé à *PRO* est sensiblement similaire dans les deux approches. Le fait qu'il soit négatif est classique. L'écart sur l'autre modalité vient probablement du manque d'exposition de cette dernière. Plus une valeur est présente dans nos données, plus l'estimation de son coefficient se fera avec justesse.

Pour le modèle de fréquence on regroupe à nouveau les tableaux de performance.

- Pour le GLM

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1397.60	0.11	0.02
Echantillon de test	442.29	0.12	0.02

TABLE 4.2: Résultats d'un GLM LASSO sur le modèle de fréquence appliqué à 20 000 individus

- Pour le GAM

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1222.11	0.11	0.02
Echantillon de test	563.45	0.12	0.02

TABLE 4.4: Résultat sur le modèle de fréquence avec variable offset sur 20 000 individus pour le modèle GAM

- Pour notre modélisation

Données utilisées	Déviance	RMSE	MAE
Echantillon d'apprentissage	1549.44	0.11	0.02
Echantillon de test	332.22	0.10	0.02

TABLE 4.6: Résultats de notre modélisation sur la fréquence appliquée à 20 000 individus

Notre méthode sur-performe très largement le GLM (et le GAM qui sous-performe). Cette amélioration est d'autant plus importante au vu des ordres de grandeur de la déviance.
On expose une dernière fois une analyse sur les différences de coefficients avec le GLM.

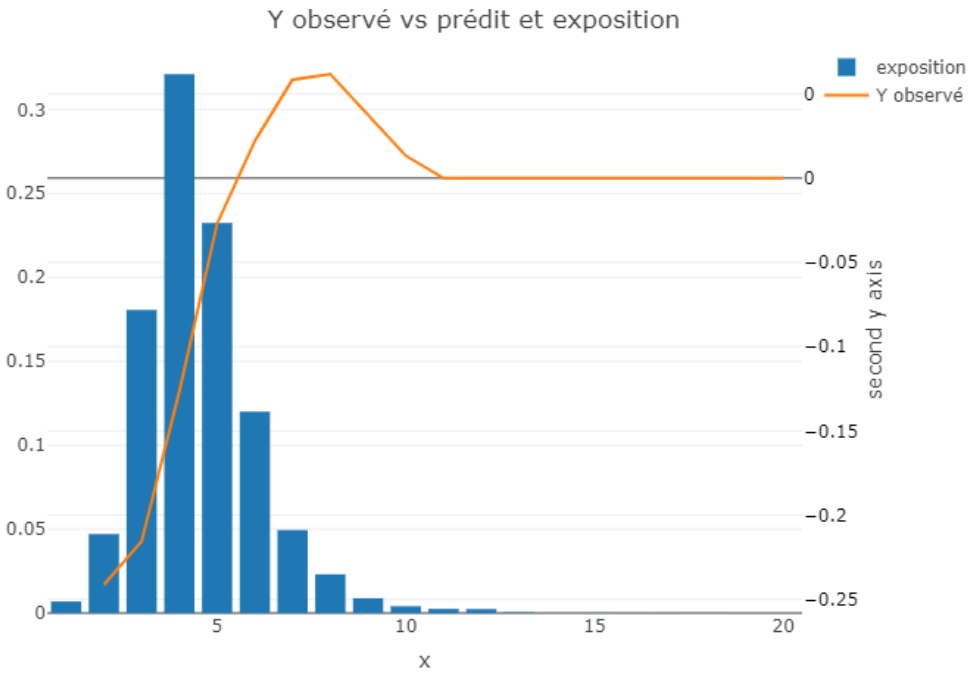


FIGURE 4.12: Coefficients estimés pour la modélisation de la fréquence pour la variable NBPIECE

Le coefficient associé à cette variable dans le GLM est de 0.09. On remarque dans notre cas que la discrétisation apporte une réelle plus value en s'affranchissant de la contrainte de linéarité. La moyenne de nos différents coefficients correspond à celui du GLM mais retranscrit le caractère non linéaire de l'impact de l'ajout d'une nouvelle pièce dans un bien sur la fréquence des sinistres.

Pour finir, à la fois sur la prédiction du coût moyen et de la fréquence, de nombreux axes d'amélioration sont déjà envisagés. On prendra pour exemple l'augmentation du nombre de modalités maximales avant regroupement par quantile ou encore l'augmentation du nombre de quantiles qui étaient restreints pour des raisons de temps de calcul. Ce gain sensible de performance au global nous confirme que, sur les données utilisées, nous avons déterminé un modèle

Conclusion

Au regard de la brutalité de l'impact de la crise de la COVID-19 sur le marché de l'assurance nous avons déterminé que la justesse de tarification était la clé de voûte de la pérennité des compagnies. La diversification des risques les rendant plus complexes à analyser et maîtriser, il nous a semblé important de chercher à complexifier les méthodes qui les modélisaient.

Pour ce faire il était essentiel de s'attarder sur les limites de la modélisation la plus courante, le GLM. Nous avons donc cherché dans un premier temps à créer un modèle capable de s'affranchir des contraintes de linéarité. La discrétisation des variables semblait une solution naturelle qui permettait de dépeindre un large spectre de forme de dépendances et ce, automatiquement. Cette astuce se rapprochait sensiblement des méthodes GAM nous incitant ainsi à trouver une solution au sur-apprentissage dont ces dernières souffraient. L'ajout de la seconde contrainte a donc permis de proposer un modèle plus robuste que le GAM tout en en conservant les avantages.

Les grands principes de notre modèle ayant été posés, l'objectif était alors de mettre en oeuvre un algorithme permettant de solutionner le problème de minimisation induit. Les méthodes de descente de gradient proximales étaient les plus adaptées et nous avons cherché à démontrer la validité de leur application à nos données.

Pour finir, nous avons pu appliquer notre modèle au jeu de données d'une des compagnies d'assurance possédant des bases de données parmi les plus complètes de la garantie dégât des eaux en MRH
.
Nous avons ainsi pu constater que notre modèle sur-performe le GLM dans le cas de l'estimation de la fréquence d'apparition de sinistres. Le gain en modélisation du coût moyen est cependant négligeable.

Bien que les premiers résultats soient déjà satisfaisants, il est encore possible d'améliorer nos travaux. En effet, il serait plus légitime d'employer une pénalité GROUP LASSO afin de rendre un sens plus concret à notre sélection de variables, celle-ci permettant de sélectionner toutes ou aucunes des modalitées issues d'une même variable originelle. Les méthodes de discrétisation des données peuvent aussi faire l'objet d'une étude comparative visant à déterminer plus précisément laquelle serait la plus efficace. Enfin, la vitesse d'exécution de l'algorithme pourrait être améliorée pour permettre une application à des données plus larges et l'objectif à terme serait de créer un package mettant en oeuvre ce modèle pour en permettre une utilisation plus globale.

Bibliographie

Altman, R. (2019). Estimateur du paramètre de dispersion. *Simon Fraser University*.

Becker, G., Klotzki, U., McElhaney, D. et Srivastava, A. (2020). The post-COVID-19 pricing imperative for PC insurers. url : https://www.mckinsey.com/industries/financial-services/our-insights/the-post-covid-19-pricing-imperative-for-p-and-c-insurers.

Donnet, S. (2018). Cours de Modèle Linéaire. *Université Paris-Dauphine*.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33.1, p. 1-22.

Guyon, P. (2019). Cours de Comptabilité Réglementation Assurance. *Université Paris-Dauphine*.

Hastie, T. (2019). gam: Generalized Additive Models. url : https://CRAN.R-project.org/package=gam.

LeDell, E. (2016). h2o: R Interface for H2O. url : http://www.h2o.ai.

Parikh, N. et Boyd, S. (2013). Proximal Algorithms. *Stanford University*.

Planchet, F. et Miseray, A. (2017). Tarification IARD, Introduction aux techniques avancées.

R Core Team (2020). R: A Language and Environment for Statistical Computing. url : http://www.R-project.org/.

Thomas, J. (2016). Modèles linéaires GLMs analyse logit Régression de poisson. *Working paper ISFA, Lyon*.

Tibshirani, R. (2013). Proximal Gradient Descent. *Carnegie Mellon University*.

Witten, D., Hastie, T., James, G. et Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Annexe A

A.1 Modèle linéaire et fléau de la grande dimension

L'objectif est de démontrer que la validité des résultats exposés pour le modèle linéaire sont conditionnels au fait que $n > p$. La solution $\hat{\beta}$ fait intervenir le terme $({}^tXX)^{-1}$, or pour que la matrice ${}^tXX$ soit inversible, il faut qu'elle soit de rang plein.

D'après le théorème du rang pour une matrice X de rang p, $p = dim(Ker(X)) + rg(X)$. Par conséquent, X est de rang plein si et seulement si X est injective (par abus de notation X représente à la fois la matrice et l'endomorphisme associé).
On va maintenant montrer que $rg(X) = p \iff {}^t XX$ inversible.
$\Leftarrow$
Soit ${}^tXX$ inversible et soit $u \in R^p$ tel que $Xu = 0_{R^n}$. Alors ${}^tXXu = 0_{R^p}$, d'où $u = 0_{R^p}$. Donc X est injective, soit de rang plein.
$\Rightarrow$
Soit $u \in R^p$ tel que ${}^tXXu = 0_{R^p}$, alors $u{}^tXXu = 0$, soit $||Xu|| = 0$. Donc $Xu = 0$, or X est de rang plein donc $u = 0_{R^p}$

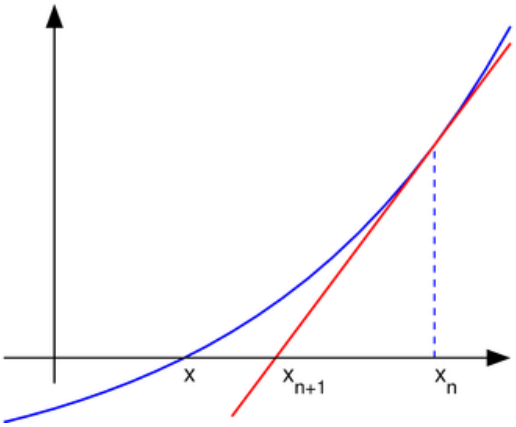
A.2 Démonstration de l'égalité $(\frac{x+y}{2})^2 \leq \frac{x^2+y^2}{2}$

$(\frac{x+y}{2})^2 \leq \frac{x^2+y^2}{2} \iff x^2 + y^2 + 2xy \leq 2(x^2 + y^2)$.
Or, on a que
$0 \leq (x-y)^2 \iff 0 \leq x^2 + y^2 - 2xy \iff 2xy \leq x^2 + y^2$.
On a donc bien que
$x^2 + y^2 + 2xy \leq 2(x^2 + y^2)$

A.3 Algorithme de Newton

Il s'agit d'un algorithme classique pour trouver le zéro d'une fonction de manière itérative. On commence par l'exposer en dimension une.
On part d'un point x_0 idéalement "proche" du zéro de la fonction f que l'on cherche à approcher. On va alors identifier f à sa tangente en x_0. On trouve le zéro de cette tangente, puis on vient lire son image par f. On réitère ensuite en partant de ce nouveau point.

FIGURE A.1: Illustration d'une itération de Newton-Raphson en dimension une



On rappel que l'équation de la tangente d'une fonction f en x_0 est : $x \mapsto f(x_0) + (x - x_0)f^{'}(x_0)$. On définit donc l'algorithme qui suit. On part de x_0 puis on itère jusqu'à convergence la suite suivante :

$$x_{k+1} = x_k - \frac{f(x_k)}{f^{'}(x_k)}.$$

On convergera alors vers un zéro de la fonction sous certaines conditions. Cette méthode dite de Newton-Raphson permet donc de trouver des extremums locaux en annulant la dérivée de la fonction. On va alors la généraliser à la dimension p pour l'adapter à notre problématique.On obtient l'algorithme suivant :

1. Choisir un point de départ arbitraire β_0
2. Itérer comme suit : $\beta^{k+1} = \beta^k - [\mathbb{E}[\mathcal{H}(\mathcal{L})(\beta^k)]]^{-1}\nabla\mathcal{L}(\beta^k)$
3. On s'arrête dès lors que $\beta^{k+1} \approx \beta^k$.

La fonction à annuler est la dérivée de la log-vraisemblance $\mathcal{L}(\beta)$, sa propre dérivée est donc bien la matrice Hessienne $\mathcal{H}(\mathcal{L})(\beta)$.
On va maintenant calculer la dérivée de la log-vraisemblance ainsi que la Hessienne afin d'implémenter notre méthode. Pour faciliter les calculs, on va prendre comme fonction de lien la fonction canonique i-e $g = (b^{'})^{-1}$. On ne détaillera que succinctement les calculs.

$$\frac{\partial\mathcal{L}(\beta)}{\partial\beta_j} = \sum_{i=1}^{n} \frac{y_i X_{i,j} - X_{i,j} b^{'}(X_{i,\beta})}{a(\phi)}.$$

On note alors $\mu(\beta) := \begin{pmatrix} b^{'}(X_1,\beta) \\ . \\ . \\ b^{'}(X_n,\beta) \end{pmatrix}$, il vient que $\nabla\mathcal{L}(\beta) = \frac{1}{a(\phi)} X^{'}(Y - \mu(\beta))$.
Il convient ensuite d'étudier la matrice Hessienne du log de la vraisemblance.

$$\begin{aligned}[\mathcal{H}(\mathcal{L})]_{i,j} :&= \frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_j\partial\beta_k} \\ &= -\frac{1}{a(\phi)} \sum_{i=1}^{n} b^{''}(X_i,\beta) X_{i,j} X_{i,k}.\end{aligned}$$

On définit alors $\mathcal{W}(\beta) := diag(b^{''}(X_{i,}\beta)) \in \mathcal{M}_{n\times n}$. On peut donc réécrire $\mathcal{H}(\mathcal{L}) = -\frac{1}{a(\phi)}X^{'}\mathcal{W}(\beta)X$.
On injecte ces résultats dans l'équation de la suite des β_k, il suffit donc d'itérer l'équation suivante

$$\begin{aligned}\beta^{k+1} &= \beta^k - [\mathcal{H}(\mathcal{L})(\beta^k)]^{-1}\nabla\mathcal{L}(\beta^k) \\ &= \beta^k + [X^{'}\mathcal{W}(\beta)X]^{-1}X^{'}(Y-\mu(\beta^k))\end{aligned}$$

A.4 CART (Classification And Regression Tree)

Ce sont des modèles de classification ou de régression que l'on présente succinctement, pour plus d'informations le lecteur pourra se renseigner avec le livre suivant Witten et al., 2013 ;
Ils se représentent de la manière suivante :
Chaque noeud correspond à un seuil associé à une des variables X_j de la base de données. Les noeuds

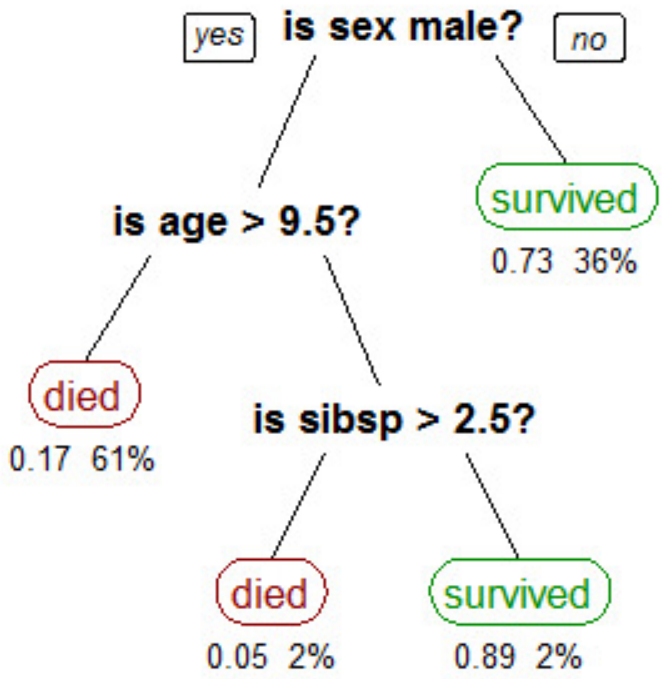


Figure A.2: Arbre de décision sur le jeu de données du Titanic, Source : https ://en.wikipedia.org

fils gauche et droite correspondent aux individus du noeud parent respectivement inférieur et supérieur au seuil séparateur. A chaque étape on sélectionne la variable et le seuil qui minimisent la déviance de la nouvelle partition.
Ce processus se poursuit jusqu'à ce que les noeuds finaux contiennent moins d'individus que le critère d'arrêt fixé au préalable. Ces noeuds finaux sont appelés feuilles et servent de prédicteurs. Dans le

cadre de la classification on attribuera en effet à chaque feuille une réponse calculée comme la classe majoritaire des individus du noeud, la prédiction en régression sera quant à elle calculée comme la moyenne des réponses des individus. Du fait que l'arbre décrive une partition de l'espace tout entier, tout nouvel individu appartiendra à une et unique feuille permettant ainsi la prédiction.

Afin de pallier le sur-apprentissage il est possible de créer plusieurs arbres en ne présentant à chaque étape que quelques variables possibles pour le choix du partitionnement. L'agrégation de ces multiples arbres permet alors une prédiction plus robuste. Ce sont les méthodes de Random Forest.

A.5 Sorties graphiques du modèle GAM

A.5.1 Modèle de coût

On présente alors ces mêmes graphiques pour la plupart des variables numériques de notre base de données.

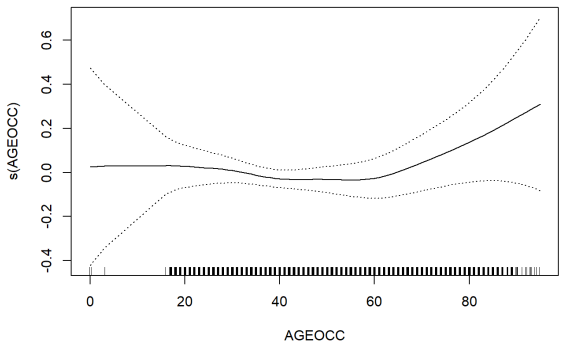


FIGURE A.3: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et l'âge de l'occupant

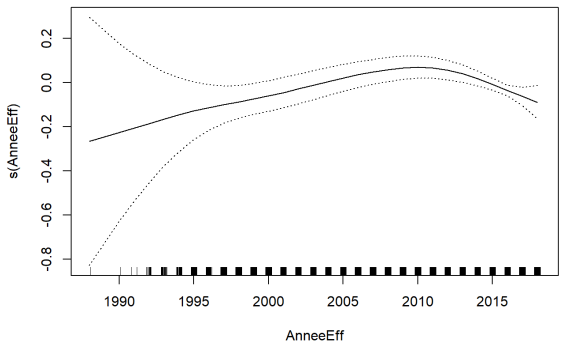


FIGURE A.4: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et l'année effective du contrat

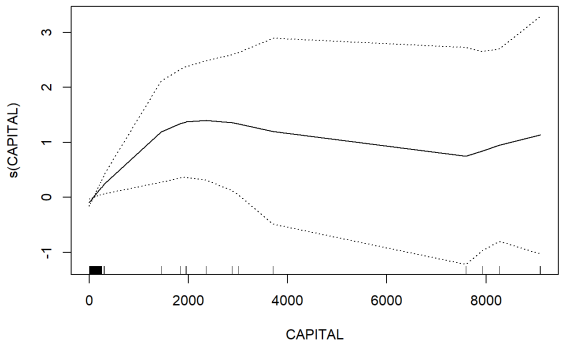


FIGURE A.5: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et le capital total assuré

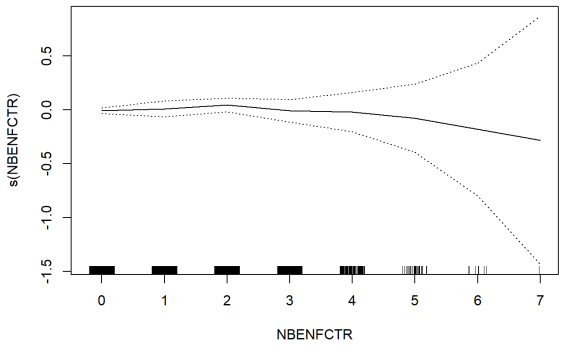


FIGURE A.6: Fonction de dépendance approcheé par le GAM pour le lien entre le coût moyen des sinistres et le nombre d'enfants à charge de l'assuré

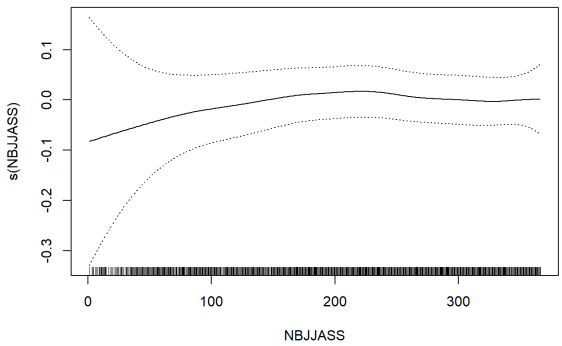


FIGURE A.7: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et la période cumulée d'assurance en jour

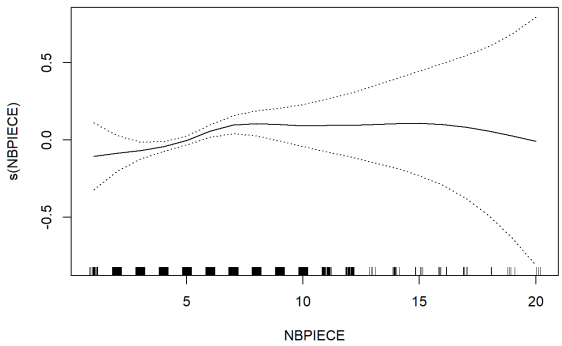


FIGURE A.8: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et le nombre de pièces du bien assuré

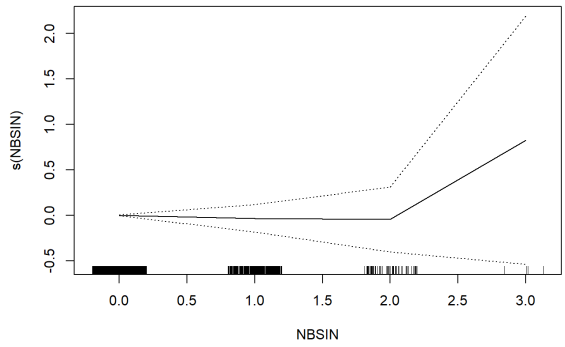


FIGURE A.9: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et le nombre de sinistres déclaré sur le contrat antérieur

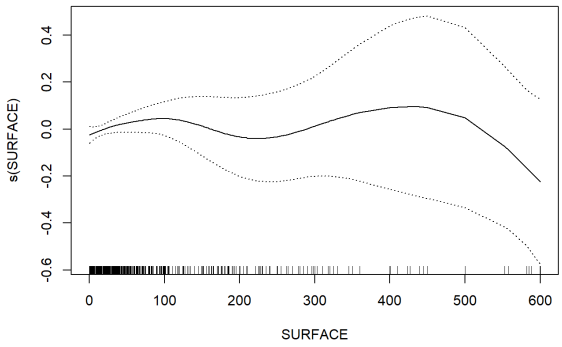


FIGURE A.10: Fonction de dépendance approchée par le GAM pour le lien entre le coût moyen des sinistres et la surface des dépendances

A.5.2 Modèle de fréquence

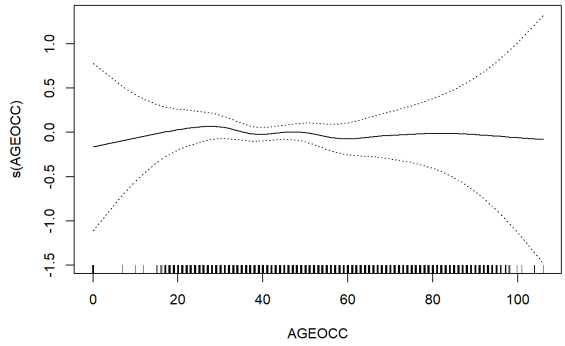


FIGURE A.11: Fonction de dépendance approchée par le GAM pour le lien entre la fréquence de sinistralité et l'âge de l'occupant

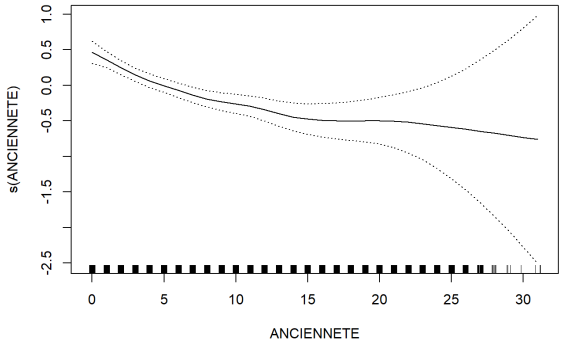


FIGURE A.12: Fonction de dépendance approchée par le GAM pour le lien entre la fréquence de sinistralité et l'ancienneté

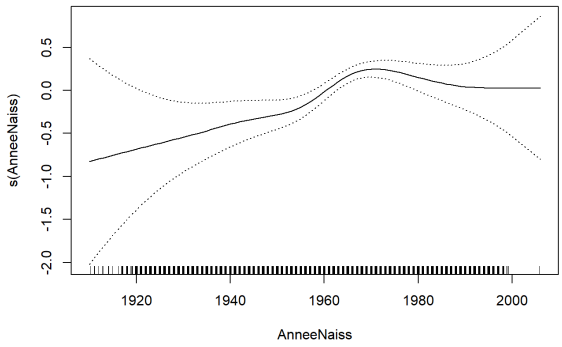


FIGURE A.13: Fonction de dépendance approchée par le GAM pour le lien entre la fréquence de sinistralité et l'année de naissance de l'assuré

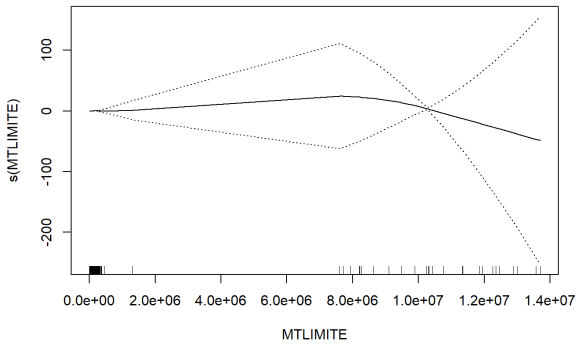


FIGURE A.14: Fonction de dépendance approchée par le GAM pour le lien entre la fréquence de sinistralité et le montant limite assuré

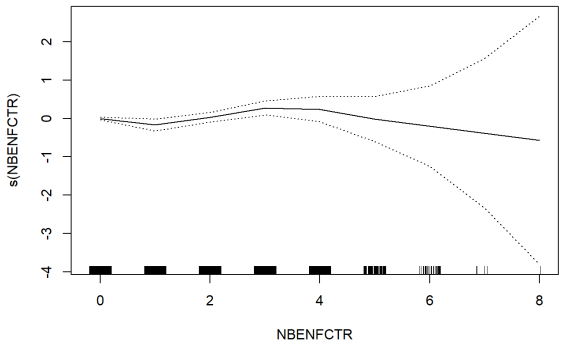


FIGURE A.15: Fonction de dépendance approchée par le GAM pour le lien entre la fréquence de sinistralité et le nombre d'enfants à charge

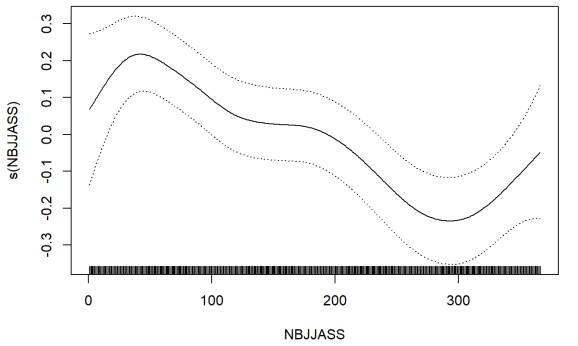


FIGURE A.16: Fonction de dépendance approchée par le GAM pour le lien entre la fréquence de sinistralité et le nombre de jours d'assurance

A.6 Tableau complet des validations croisées de notre algorithme

On présente le tableau complet associé au tableau 4.5 sur le modèle de coût.

TABLE A.1: Tableau complet de la validation croisée du modèle de coût sur le couple (λ_1, λ_2)

lambda1	lambda2	deviance	MAE	RMSE
0.000001	0.00050	1485.25578	1240.15715	2284.56162
0.000001	0.00100	1515.66546	1264.00342	2300.87744
0.000001	0.00150	1491.73945	1242.08212	2286.51681
0.000001	0.00200	1529.71174	1209.93866	2287.11512
0.000001	0.00250	1487.46642	1229.53049	2283.42908
0.000001	0.00300	1521.19374	1225.93839	2290.60497
0.000001	0.00350	1528.93705	1241.61518	2297.44529
0.000001	0.00400	1513.88433	1248.33210	2294.93916
0.000001	0.00450	1495.40892	1238.04609	2285.09036
0.000001	0.00500	1490.24004	1239.26315	2285.84617
0.000001	0.00550	1526.68682	1223.73706	2291.10398
0.000001	0.00600	1508.22552	1236.45077	2291.20050
0.000001	0.00650	1514.75358	1238.18970	2293.22181
0.000001	0.00700	1479.83306	1237.16680	2283.89097
0.000001	0.00750	1480.19100	1237.60513	2284.10671
0.000001	0.00800	1480.40145	1237.21186	2284.33013
0.000001	0.00850	1507.29767	1236.13190	2291.33804
0.000001	0.00900	1488.68782	1235.36698	2285.91210
0.000001	0.00950	1480.09950	1237.20520	2283.91400
0.000001	0.01000	1479.99139	1237.34260	2283.94194
0.00011	0.00050	1465.39889	1230.26917	2277.73524
0.00011	0.00100	1483.68106	1243.55237	2285.49118
0.00011	0.00150	1474.81227	1237.69931	2282.14311
0.00011	0.00200	1498.81873	1210.86489	2282.88460
0.00011	0.00250	1483.68790	1242.63386	2283.21169
0.00011	0.00300	1495.25109	1207.84230	2280.63054
0.00011	0.00350	1519.45344	1217.89230	2290.01456
0.00011	0.00400	1517.01862	1212.30730	2288.30830
0.00011	0.00450	1473.46231	1224.49961	2277.04128
0.00011	0.00500	1468.62337	1225.36011	2277.81053
0.00011	0.00550	1512.18662	1207.31559	2284.63046
0.00011	0.00600	1486.79943	1221.45189	2281.48244
0.00011	0.00650	1492.52353	1223.77750	2283.17437
0.00011	0.00700	1500.55827	1227.38361	2285.83242
0.00011	0.00750	1462.86327	1229.30621	2277.53563
0.00011	0.00800	1463.13681	1229.78367	2277.72262
0.00011	0.00850	1490.87013	1212.06788	2281.67974
0.00011	0.00900	1490.91085	1224.67159	2283.02016
0.00011	0.00950	1462.63984	1228.89252	2277.33716
0.00011	0.01000	1470.54400	1226.50981	2278.60000

lambda1	lambda2	deviance	MAE	RMSE
0.00016	0.00050	1455.52995	1225.85876	2274.29039
0.00016	0.00100	1478.87621	1251.78193	2287.86174
0.00016	0.00150	1468.49556	1231.24846	2279.63801
0.00016	0.00200	1481.45497	1221.53400	2280.75217
0.00016	0.00250	1492.10553	1224.63071	2285.04497
0.00016	0.00300	1482.50415	1211.22948	2278.98805
0.00016	0.00350	1515.16867	1219.13946	2291.31095
0.00016	0.00400	1495.97041	1223.03735	2285.66478
0.00016	0.00450	1494.03977	1221.27334	2283.14651
0.00016	0.00500	1478.02161	1222.26394	2278.88897
0.00016	0.00550	1505.87004	1206.75613	2284.31205
0.00016	0.00600	1501.82756	1216.65096	2286.39632
0.00016	0.00650	1486.37641	1218.81114	2281.01413
0.00016	0.00700	1494.33924	1222.33681	2283.52238
0.00016	0.00750	1453.32601	1221.40535	2272.93282
0.00016	0.00800	1465.06534	1218.21206	2275.14416
0.00016	0.00850	1481.38119	1209.34951	2279.09702
0.00016	0.00900	1482.58306	1216.03241	2279.79087
0.00016	0.00950	1455.61853	1225.73988	2274.78346
0.00016	0.01000	1462.50944	1223.27461	2275.96995
0.00021	0.00050	1447.65684	1222.13873	2271.55971
0.00021	0.00100	1472.04882	1243.32042	2284.52361
0.00021	0.00150	1458.04315	1228.51375	2275.66278
0.00021	0.00200	1474.81709	1214.47240	2278.15155
0.00021	0.00250	1495.82641	1208.54479	2282.44201
0.00021	0.00300	1462.29642	1212.26236	2273.41868
0.00021	0.00350	1498.21846	1211.17736	2284.53875
0.00021	0.00400	1514.92875	1221.65074	2292.07544
0.00021	0.00450	1495.70593	1225.08646	2286.37793
0.00021	0.00500	1468.57089	1216.58502	2273.55735
0.00021	0.00550	1485.25641	1210.56999	2280.68632
0.00021	0.00600	1504.62818	1209.23574	2284.87504
0.00021	0.00650	1494.26512	1211.64086	2283.17169
0.00021	0.00700	1486.56834	1219.46912	2281.28491
0.00021	0.00750	1446.45273	1219.80832	2270.26975
0.00021	0.00800	1460.28875	1210.40766	2272.56247
0.00021	0.00850	1478.28572	1202.81507	2277.21118
0.00021	0.00900	1478.06623	1212.25154	2278.30393
0.00021	0.00950	1444.69699	1223.39232	2270.29730
0.00021	0.01000	1448.71506	1223.19618	2272.00900
0.00026	0.00050	1444.80603	1219.98451	2270.66349

lambda1	lambda2	deviance	MAE	RMSE
0.00026	0.00100	1462.50140	1239.36127	2280.07412
0.00026	0.00150	1455.86786	1224.75789	2274.76651
0.00026	0.00200	1472.15254	1216.37469	2278.93211
0.00026	0.00250	1493.46785	1205.48741	2280.89513
0.00026	0.00300	1457.65708	1210.12610	2271.84115
0.00026	0.00350	1485.18769	1198.83976	2277.26867
0.00026	0.00400	1493.44566	1221.67710	2284.06450
0.00026	0.00450	1497.28255	1230.44457	2285.64401
0.00026	0.00500	1491.64208	1218.99864	2281.43027
0.00026	0.00550	1473.59747	1212.73628	2277.69271
0.00026	0.00600	1476.83237	1206.77298	2276.63766
0.00026	0.00650	1466.44057	1209.46137	2273.22156
0.00026	0.00700	1470.18616	1208.95347	2273.59347
0.00026	0.00750	1454.30832	1208.36953	2270.52948
0.00026	0.00800	1460.17562	1209.74802	2272.93342
0.00026	0.00850	1469.88492	1204.85717	2274.85434
0.00026	0.00900	1473.39659	1209.60044	2276.54677
0.00026	0.00950	1456.15081	1219.57832	2272.51056
0.00026	0.01000	1445.14310	1221.01487	2270.60456
0.00032	0.00050	1442.41071	1218.48109	2270.69628
0.00032	0.00100	1462.95424	1237.21471	2280.87069
0.00032	0.00150	1451.28723	1229.18851	2273.97882
0.00032	0.00200	1462.74115	1229.10699	2279.06915
0.00032	0.00250	1475.99682	1200.34567	2275.97536
0.00032	0.00300	1452.10794	1223.76070	2272.73766
0.00032	0.00350	1475.93341	1193.23403	2273.58099
0.00032	0.00400	1495.60249	1184.77726	2277.06700
0.00032	0.00450	2771.54000	1199.66845	2338.72370
0.00032	0.00500	2261.55251	1182.70346	2331.39637
0.00032	0.00550	1469.12805	1211.34707	2275.85687
0.00032	0.00600	1473.29489	1200.56003	2274.41469
0.00032	0.00650	1479.90884	1200.57502	2276.04275
0.00032	0.00700	1467.91890	1198.26771	2271.27931
0.00032	0.00750	1472.99551	1197.79823	2272.08614
0.00032	0.00800	1460.98964	1200.10905	2271.37787
0.00032	0.00850	1468.24181	1198.65945	2272.98729
0.00032	0.00900	1472.97971	1198.16866	2274.29814
0.00032	0.00950	1445.68577	1222.82565	2270.28654
0.00032	0.01000	1440.98822	1218.51433	2268.73747
0.00037	0.00050	1438.32969	1215.30265	2267.74140
0.00037	0.00100	1457.51411	1225.49732	2275.83601
0.00037	0.00150	1446.64668	1229.06623	2271.90195
0.00037	0.00200	1472.10112	1218.38291	2279.33395

lambda1	lambda2	deviance	MAE	RMSE
0.00037	0.00250	1492.55427	1201.70931	2281.41541
0.00037	0.00300	1458.41863	1214.42581	2273.74488
0.00037	0.00350	1461.44265	1202.09876	2271.92611
0.00037	0.00400	1500.56515	1188.09898	2280.19147
0.00037	0.00450	1517.10258	1213.80438	2288.42619
0.00037	0.00500	2728.16540	1225.56196	2326.47077
0.00037	0.00550	1454.78474	1215.54720	2272.62575
0.00037	0.00600	1466.05984	1205.32307	2273.51798
0.00037	0.00650	1469.83108	1203.05558	2273.79148
0.00037	0.00700	1465.50087	1197.81150	2270.79162
0.00037	0.00750	1448.51382	1202.65882	2267.90012
0.00037	0.00800	1460.83871	1199.29367	2271.53344
0.00037	0.00850	1464.22638	1199.20689	2272.58766
0.00037	0.00900	1467.69271	1198.98268	2273.51233
0.00037	0.00950	1464.13212	1212.57092	2273.61759
0.00037	0.01000	1437.80490	1216.85421	2267.62192
0.00042	0.00050	1432.03583	1213.87433	2265.94755
0.00042	0.00100	1446.69997	1226.32253	2273.58417
0.00042	0.00150	1443.43176	1229.75728	2271.85101
0.00042	0.00200	1449.76145	1225.57649	2274.47797
0.00042	0.00250	1473.84894	1215.05006	2278.31648
0.00042	0.00300	1445.62773	1211.36086	2269.17216
0.00042	0.00350	1455.92192	1201.62007	2270.42597
0.00042	0.00400	1471.26402	1188.61628	2273.19876
0.00042	0.00450	1567.96140	1162.44745	2288.86199
0.00042	0.00500	1493.02583	1213.35629	2280.63761
0.00042	0.00550	1451.77275	1210.63229	2271.60264
0.00042	0.00600	1460.00627	1199.83104	2271.89013
0.00042	0.00650	1465.73165	1196.47089	2273.04829
0.00042	0.00700	1453.22484	1193.61832	2268.24732
0.00042	0.00750	1454.96720	1192.54438	2268.55223
0.00042	0.00800	1460.97139	1193.17258	2270.82348
0.00042	0.00850	1463.81123	1193.67539	2271.58253
0.00042	0.00900	1455.43676	1203.40233	2271.95664
0.00042	0.00950	1446.31001	1206.67581	2269.03924
0.00042	0.01000	1441.03303	1221.91015	2270.02642
0.00047	0.00050	1429.82759	1212.37637	2265.13945
0.00047	0.00100	1451.79099	1231.64172	2275.39477
0.00047	0.00150	1440.69505	1216.38371	2268.12353
0.00047	0.00200	1455.75759	1222.80625	2275.52954
0.00047	0.00250	1476.51156	1208.07383	2278.40832
0.00047	0.00300	1452.71773	1216.50520	2272.63875
0.00047	0.00350	1458.51533	1204.71611	2272.03748

lambda1	lambda2	deviance	MAE	RMSE
0.00047	0.00400	1476.16395	1181.75269	2274.23764
0.00047	0.00450	1512.37727	1191.07741	2286.33517
0.00047	0.00500	1937.91759	1181.98291	2311.13498
0.00047	0.00550	1450.55031	1208.80835	2270.57272
0.00047	0.00600	1455.00913	1198.67365	2270.41122
0.00047	0.00650	1458.39691	1196.58082	2271.17321
0.00047	0.00700	1450.46794	1191.02080	2267.37352
0.00047	0.00750	1451.21116	1190.54244	2267.55398
0.00047	0.00800	1456.76278	1191.08587	2269.73482
0.00047	0.00850	1459.86773	1191.39651	2270.74851
0.00047	0.00900	1463.78493	1191.77250	2271.96530
0.00047	0.00950	1462.03377	1199.12937	2271.92258
0.00047	0.01000	1429.88664	1206.92561	2263.78485
0.00053	0.00050	1429.04962	1212.95062	2265.35261
0.00053	0.00100	1444.04247	1228.24247	2273.57921
0.00053	0.00150	1436.77004	1217.26152	2267.22035
0.00053	0.00200	1443.94803	1227.29613	2272.06473
0.00053	0.00250	1461.81188	1208.84414	2275.63788
0.00053	0.00300	1437.61479	1212.04970	2266.90624
0.00053	0.00350	1441.81274	1202.47938	2266.43388
0.00053	0.00400	1458.35394	1175.10432	2268.25933
0.00053	0.00450	1564.63412	1163.44582	2290.01356
0.00053	0.00500	1519.47902	1182.55039	2280.44540
0.00053	0.00550	1435.46683	1207.16696	2266.07668
0.00053	0.00600	1448.78350	1187.94083	2267.09895
0.00053	0.00650	1451.06825	1186.49531	2267.56807
0.00053	0.00700	1453.69348	1184.50935	2268.08368
0.00053	0.00750	1455.90463	1183.81373	2268.63458
0.00053	0.00800	1459.03750	1183.26232	2269.42817
0.00053	0.00850	1462.81925	1183.37603	2270.50003
0.00053	0.00900	1467.30642	1183.08119	2271.76936
0.00053	0.00950	1452.16125	1200.61484	2270.57242
0.00053	0.01000	1446.95424	1212.32162	2270.15743
0.00058	0.00050	1427.62963	1210.22810	2264.30710
0.00058	0.00100	1441.40782	1225.39465	2270.46806
0.00058	0.00150	1436.53474	1220.52010	2267.90020
0.00058	0.00200	1440.72242	1218.57936	2268.23288
0.00058	0.00250	1457.21106	1204.84996	2272.92964
0.00058	0.00300	1438.04223	1204.89566	2266.23002
0.00058	0.00350	1443.92921	1203.12405	2268.03409
0.00058	0.00400	1460.64083	1184.80178	2270.48550
0.00058	0.00450	1467.85670	1181.61585	2272.37299
0.00058	0.00500	1744.50272	1162.46606	2304.89640

lambda1	lambda2	deviance	MAE	RMSE
0.00058	0.00550	1429.36948	1210.24861	2264.91593
0.00058	0.00600	1444.59415	1188.63232	2266.00449
0.00058	0.00650	1457.08770	1187.12108	2269.67675
0.00058	0.00700	1461.47429	1187.50192	2271.00723
0.00058	0.00750	1450.71112	1186.03552	2267.59899
0.00058	0.00800	1452.98101	1184.82978	2268.25323
0.00058	0.00850	1456.13970	1185.21464	2269.26844
0.00058	0.00900	1459.93221	1185.19551	2270.46515
0.00058	0.00950	1453.55059	1199.31262	2269.72473
0.00058	0.01000	1438.12335	1205.06866	2265.99659
0.00063	0.00050	1424.55880	1210.26642	2264.12755
0.00063	0.00100	1438.09489	1224.01389	2270.65316
0.00063	0.00150	1433.60033	1221.70264	2267.14870
0.00063	0.00200	1441.34989	1224.92798	2270.61842
0.00063	0.00250	1457.24333	1205.53886	2273.43770
0.00063	0.00300	1435.05411	1198.76691	2264.29578
0.00063	0.00350	1441.05223	1195.85074	2265.80968
0.00063	0.00400	1449.41029	1188.68573	2267.66765
0.00063	0.00450	1551.25979	1170.58023	2285.22076
0.00063	0.00500	1492.61167	1176.66651	2271.69093
0.00063	0.00550	1444.17690	1199.26528	2266.20742
0.00063	0.00600	1434.88528	1196.13560	2264.10671
0.00063	0.00650	1451.68451	1189.27658	2268.45963
0.00063	0.00700	1457.02786	1188.23709	2270.02576
0.00063	0.00750	1442.34976	1186.83201	2264.82947
0.00063	0.00800	1449.56388	1184.12282	2267.28674
0.00063	0.00850	1453.24238	1183.47147	2268.31384
0.00063	0.00900	1453.83527	1188.37864	2268.80112
0.00063	0.00950	1439.56469	1205.33094	2267.43175
0.00063	0.01000	1422.15062	1203.72534	2261.36629
0.00068	0.00050	1421.52175	1207.74223	2261.43462
0.00068	0.00100	1430.92952	1220.31594	2267.10575
0.00068	0.00150	1431.35891	1218.11098	2266.22453
0.00068	0.00200	1441.75109	1219.14924	2270.43075
0.00068	0.00250	1456.11663	1199.46388	2271.67886
0.00068	0.00300	1441.21378	1203.33656	2267.45092
0.00068	0.00350	1439.47770	1202.75457	2266.23871
0.00068	0.00400	1458.36384	1192.88765	2270.97341
0.00068	0.00450	1666.71567	1144.73723	2302.43364
0.00068	0.00500	1629.25463	1159.16592	2303.30476
0.00068	0.00550	1637.13557	1196.76092	2293.92493
0.00068	0.00600	1446.66158	1193.46748	2267.96134
0.00068	0.00650	1454.68015	1186.09786	2269.27259

lambda1	lambda2	deviance	MAE	RMSE
0.00068	0.00700	1444.29669	1183.77394	2265.14974
0.00068	0.00750	1445.40265	1183.03936	2265.37786
0.00068	0.00800	1450.02238	1183.62117	2267.27366
0.00068	0.00850	1453.18917	1183.58520	2268.21025
0.00068	0.00900	1457.11148	1183.89835	2269.41042
0.00068	0.00950	1455.37188	1190.58104	2269.13633
0.00068	0.01000	1421.68035	1202.48239	2261.00641
0.00074	0.00050	1419.55357	1208.02179	2261.97500
0.00074	0.00100	1435.70917	1219.96484	2269.33884
0.00074	0.00150	1435.92629	1224.10682	2268.06911
0.00074	0.00200	1439.81792	1226.31089	2270.10276
0.00074	0.00250	1469.33546	1212.17042	2278.60003
0.00074	0.00300	1438.61463	1202.04647	2266.31115
0.00074	0.00350	1445.25812	1204.85945	2268.65347
0.00074	0.00400	1452.91623	1184.06827	2268.53181
0.00074	0.00450	1476.52102	1184.44819	2274.97612
0.00074	0.00500	1496.57306	1184.55883	2282.26773
0.00074	0.00550	1474.35628	1198.67068	2275.60247
0.00074	0.00600	1435.83685	1198.97615	2264.95953
0.00074	0.00650	1444.21711	1191.91627	2266.59532
0.00074	0.00700	1448.45007	1190.50526	2267.94800
0.00074	0.00750	1452.10974	1191.13076	2269.34011
0.00074	0.00800	1445.28258	1184.56386	2265.81214
0.00074	0.00850	1448.32559	1184.66086	2266.70719
0.00074	0.00900	1439.29084	1195.43138	2266.62104
0.00074	0.00950	1436.51932	1202.25541	2266.25308
0.00074	0.01000	1425.97010	1214.91862	2264.20731
0.00079	0.00050	1418.91380	1206.53747	2261.74753
0.00079	0.00100	1433.37530	1217.27179	2268.06707
0.00079	0.00150	1428.78420	1219.92802	2264.91024
0.00079	0.00200	1451.97157	1234.61478	2275.94559
0.00079	0.00250	1463.13793	1212.16309	2276.59521
0.00079	0.00300	1437.14880	1201.88403	2265.88198
0.00079	0.00350	1439.45180	1200.89699	2266.11200
0.00079	0.00400	1449.73219	1185.46679	2267.83437
0.00079	0.00450	1585.23380	1157.26321	2289.24244
0.00079	0.00500	1556.81005	1189.71205	2290.53701
0.00079	0.00550	1522.10070	1191.79941	2282.30020
0.00079	0.00600	1431.81725	1199.95014	2263.75639
0.00079	0.00650	1438.22117	1190.65725	2264.34807
0.00079	0.00700	1436.31073	1187.58146	2263.11427
0.00079	0.00750	1437.96270	1184.81528	2263.30845
0.00079	0.00800	1441.92444	1185.98545	2265.03473

lambda1	lambda2	deviance	MAE	RMSE
0.00079	0.00850	1444.40484	1186.43975	2265.86518
0.00079	0.00900	1447.85027	1186.50723	2267.00930
0.00079	0.00950	1436.68673	1196.28292	2263.76762
0.00079	0.01000	1419.09788	1207.04651	2260.75618
0.00084	0.00050	1414.81021	1205.02961	2260.10695
0.00084	0.00100	1429.90120	1215.10560	2265.65448
0.00084	0.00150	1438.18335	1221.42018	2268.72113
0.00084	0.00200	1437.98645	1221.40207	2269.05365
0.00084	0.00250	1459.69025	1211.10662	2275.18485
0.00084	0.00300	1449.43436	1199.27849	2269.09599
0.00084	0.00350	1439.25432	1206.63525	2267.22145
0.00084	0.00400	1450.66287	1191.88088	2269.18608
0.00084	0.00450	1477.97230	1178.40397	2275.38740
0.00084	0.00500	1479.42932	1213.26981	2282.64970
0.00084	0.00550	1450.25815	1218.69322	2271.88018
0.00084	0.00600	1432.05717	1199.15864	2263.83056
0.00084	0.00650	1433.99735	1191.48425	2263.53288
0.00084	0.00700	1431.84769	1188.95964	2262.31733
0.00084	0.00750	1432.56664	1188.16293	2262.41481
0.00084	0.00800	1437.59558	1186.55920	2264.29896
0.00084	0.00850	1439.21642	1187.21289	2265.00919
0.00084	0.00900	1437.16540	1192.76207	2264.86718
0.00084	0.00950	1430.38596	1199.17527	2262.78916
0.00084	0.01000	1431.90312	1198.58557	2263.16747
0.00089	0.00050	1415.51665	1204.78875	2260.62796
0.00089	0.00100	1423.49685	1214.33370	2264.85722
0.00089	0.00150	1431.34517	1218.08035	2265.98251
0.00089	0.00200	1437.00568	1219.70920	2268.00474
0.00089	0.00250	1449.71287	1214.08411	2272.10925
0.00089	0.00300	1430.88516	1210.98047	2264.55525
0.00089	0.00350	1435.74507	1201.73153	2265.45917
0.00089	0.00400	1442.07574	1192.17498	2266.19053
0.00089	0.00450	1576.27681	1161.24144	2290.18077
0.00089	0.00500	1472.36030	1195.24197	2277.40112
0.00089	0.00550	1468.79432	1191.65202	2273.37201
0.00089	0.00600	1426.08761	1203.76085	2262.26159
0.00089	0.00650	1429.46438	1188.31830	2261.89640
0.00089	0.00700	1436.89129	1180.01829	2262.64088
0.00089	0.00750	1564.52105	1163.30746	2277.38132
0.00089	0.00800	1555.48723	1162.22587	2277.72923
0.00089	0.00850	1430.22198	1190.28834	2263.26097
0.00089	0.00900	1426.19551	1195.48859	2262.34719
0.00089	0.00950	1422.74938	1201.55459	2261.61131

lambda1	lambda2	deviance	MAE	RMSE
0.00089	0.01000	1417.21999	1198.21526	2259.22342
0.00095	0.00050	1412.71456	1203.07529	2259.34763
0.00095	0.00100	1416.70875	1209.70784	2261.50811
0.00095	0.00150	1427.30516	1217.48439	2266.24013
0.00095	0.00200	1434.42043	1221.02118	2267.12728
0.00095	0.00250	1446.70889	1222.21395	2271.86048
0.00095	0.00300	1440.73906	1209.22723	2266.64486
0.00095	0.00350	1435.07842	1204.23702	2266.13291
0.00095	0.00400	1442.94799	1188.90825	2266.52368
0.00095	0.00450	1457.20417	1174.61078	2270.11219
0.00095	0.00500	1463.20733	1193.64491	2274.98379
0.00095	0.00550	1453.77945	1190.76147	2270.47900
0.00095	0.00600	1425.14070	1197.93263	2261.91044
0.00095	0.00650	1431.85833	1177.08479	2262.23615
0.00095	0.00700	1435.74029	1177.60772	2262.66677
0.00095	0.00750	1436.77265	1176.02457	2262.82395
0.00095	0.00800	1427.56012	1188.45527	2262.68486
0.00095	0.00850	1427.84419	1188.01128	2262.74242
0.00095	0.00900	1428.81712	1189.24027	2263.21298
0.00095	0.00950	1418.43457	1202.31295	2260.73016
0.00095	0.01000	1417.54683	1200.79168	2260.42203
0.00100	0.00050	1413.67181	1201.35349	2259.29319
0.00100	0.00100	1421.80328	1210.56504	2262.95914
0.00100	0.00150	1430.09646	1214.33506	2266.35269
0.00100	0.00200	1431.33572	1218.87474	2265.60044
0.00100	0.00250	1449.99507	1219.64664	2272.54459
0.00100	0.00300	1444.31372	1205.62906	2267.68629
0.00100	0.00350	1431.64135	1203.63860	2264.10729
0.00100	0.00400	1442.17542	1193.41203	2266.84311
0.00100	0.00450	1467.09671	1167.38834	2272.08157
0.00100	0.00500	1482.58104	1168.60151	2277.26269
0.00100	0.00550	1453.90497	1195.98383	2269.35062
0.00100	0.00600	1426.23037	1199.13733	2262.33563
0.00100	0.00650	1428.96620	1184.77663	2261.90650
0.00100	0.00700	1523.67907	1161.83812	2274.02516
0.00100	0.00750	1514.39298	1161.17905	2272.89968
0.00100	0.00800	1514.53625	1161.17218	2273.82867
0.00100	0.00850	1508.58335	1162.12487	2273.50282
0.00100	0.00900	1428.03419	1195.29206	2263.86451
0.00100	0.00950	1425.04430	1193.78674	2262.17266

On présente le tableau complet associé au tableau 4.7 sur le modèle de fréquence.

TABLE A.2: Tableau complet de la validation croisée du modèle de fréquence sur le couple (λ_1, λ_2)

lambda1	**lambda2**	**deviance**	**MAE**	**RMSE**
0.000001	0.000500	391.641030	0.022154	0.113762
0.000001	0.001625	369.473420	0.023164	0.113467
0.000001	0.002750	371.304258	0.023161	0.113505
0.000012	0.000500	495.146667	0.054183	0.220401
0.000012	0.001625	330.199700	0.023157	0.112984
0.000012	0.002750	330.625063	0.023132	0.112959
0.000023	0.000500	323.264137	0.022140	0.112750
0.000023	0.001625	321.712494	0.022987	0.112611
0.000023	0.002750	321.381415	0.023159	0.112688
0.000034	0.000500	321.774687	0.021125	0.112569
0.000034	0.001625	316.124643	0.023230	0.112535
0.000034	0.002750	316.125673	0.023189	0.112522
0.000045	0.000500	314.909860	0.022140	0.112358
0.000045	0.001625	312.898455	0.022771	0.112265
0.000045	0.002750	312.468671	0.023208	0.112378
0.000056	0.000500	309.520225	0.022149	0.112256
0.000056	0.001625	307.347400	0.022915	0.112156
0.000056	0.002750	307.575604	0.023160	0.112203
0.000067	0.000500	305.186319	0.022469	0.112120
0.000067	0.001625	305.097733	0.022927	0.112064
0.000067	0.002750	305.476691	0.023146	0.112116
0.000078	0.000500	303.323725	0.021797	0.112021
0.000078	0.001625	303.264221	0.023366	0.112050
0.000078	0.002750	303.054737	0.023113	0.112045
0.000089	0.000500	303.159314	0.022070	0.112039
0.000089	0.001625	302.633380	0.023089	0.112034
0.000089	0.002750	302.670781	0.023106	0.112015

lambda1	lambda2	deviance	MAE	RMSE
0.000100	0.000500	301.994307	0.022102	0.112003
0.000100	0.001625	301.890727	0.023179	0.112009
0.000100	0.002750	301.974331	0.023106	0.111993
0.000500	0.000500	307.465660	0.023171	0.112076
0.000500	0.001625	307.417767	0.023212	0.112071
0.000500	0.002750	306.850662	0.023106	0.112058
0.001556	0.000500	312.242608	0.023233	0.112181
0.001556	0.001625	312.242608	0.023233	0.112181
0.001556	0.002750	312.242608	0.023233	0.112181
0.002611	0.000500	313.100603	0.023268	0.112200
0.002611	0.001625	313.100603	0.023268	0.112200
0.002611	0.002750	313.100603	0.023268	0.112200
0.003667	0.000500	313.100603	0.023268	0.112200
0.003667	0.001625	313.100603	0.023268	0.112200
0.003667	0.002750	313.100603	0.023268	0.112200
0.004722	0.000500	313.100603	0.023268	0.112200
0.004722	0.001625	313.100603	0.023268	0.112200
0.004722	0.002750	313.100603	0.023268	0.112200