Introduction

Au sein d'une société d'assurance, la **tarification** est un aspect fondamental dans le sens où elle permet de de maîtriser l'exposition face aux risques. Le marché de l'assurance automobile étant fortement concurrentiel, il est absolument essentiel pour un assureur de bien **modéliser** les coûts des sinistres qu'il sera amené à rembourser afin d'anticiper et d'incorporer ces coûts dans son tarif pour ne pas subir des changements dans la structure de son portefeuille de clients. Cette assertion est d'autant plus valable dans le cadre de la réglementation *Solvabilité 2* qui encadre le bilan prudentiel des sociétés d'assurance.
Dans ce qui suit, nous allons nous intéresser à une approche tarifaire sur un portefeuille de clients en assurance automobile au travers de la modélisation de différents aspects des caractéristiques des sinistres.

1 Structure des données

Notre jeu de données est sous la forme d'un *dataframe* ou tableau de données contenant 5 352 ménages résidant en France caractérisés par les 27 variables suivantes :

1. ***pcs*** : la catégorie socio-professionnelle (Retraités, Cadres, etc.)
2. ***RUC*** : le revenu par unité de consommation du ménage
3. ***cs*** : la catégorie sociale de revenus ou plus communément la classe (moyenne, aisée, etc.)
4. ***reves*** : le revenu estimé du ménage
5. ***crevepp*** : le quartile de revenu par personne du ménage
6. ***region*** : la région de résidence codée de 1 à 9
7. ***habi*** : le type d'habitat codé de 0 à 8
8. ***Ahabi*** : le type d'habitat (Commune rurale, Zone urbaine de 2 000 à 9 999 habitants, etc.)
9. ***Atyph*** : le statut d'occupation du logement du ménage (Locataire, Propriétaire, etc.)
10. ***agecat*** : la catégorie d'âge du chef ou de la cheffe de ménage (21-40, 41-50, etc.)
11. ***Acompm*** : la composition du ménage (Personne seule, Couple avec enfants, etc.)
12. ***nbpers*** : le nombre de personnes vivant dans le ménage
13. ***enfants*** : la présence ou non d'enfants dans le ménage
14. ***Anat*** : la nationalité du ménage (Ménage français, Au moins un étranger, etc.)

15. Bauto : la possession ou non d'un autre véhicule que celui concerné par cette assurance pour le ménage
16. ***Nbadulte*** : le nombre d'adultes vivant dans le ménage
17. ***Sinistre1*** : le montant du dommage en milliers d'€ de sinistres de type 1
18. ***Sinistre2*** : le montant du dommage en milliers d'€ de sinistres de type 2
19. ***Sinistre3*** : le montant du dommage en milliers d'€ de sinistres de type 3
20. ***Police1*** : le montant de la cotisation au titre du complément de police de type 1
21. ***Police2*** : le montant de la cotisation au titre du complément de police de type 2
22. ***Police3*** : le montant de la cotisation au titre du complément de police de type 3
23. ***durPolice1*** : la durée d'adhésion à la police 1 (en unité non précisée)
24. ***Duree*** : la durée, similaire à la variable précédente (toujours en unité non précisée mais calculée selon la formule $durPolice1 \times 100$)
25. NSin : le nombre total de sinistres pour l'ensemble du ménage
26. censure : la censure ou non des variables durPolice1 et Duree
27. Sinistre0 : le montant du dommage en milliers d'euros de sinistres de type 0

Ci-dessous, un court résumé des fonctions de ces variables :

— les 16 premières qu'on nommera variables **explicatives**, supposées **exogènes**, qui rassemblent les informations observables pour un ménage souhaitant intégrer le portefeuille de l'assureur

— *Sinistre1, Sinistre2, Sinistre3, durPolice1, Duree* et *Nsin* qu'on appellera variables **expliquées** ou **endogènes** et qui seront celles qu'on cherchera à prédire pour ajuster notre tarification puisque non observables lors de l'entrée en portefeuille du ménage ciblé

— *Police1, Police2* et *Police3* qui sont les variables dont nous chercherons à déterminer les montants par la tarification

— La variable *censure* qui est une information interne à l'assureur, non observable pour un nouveau ménage, et qui n'intervient que dans la modélisation de la durée

Nous allons énumérer quelques hypothèses faites sur les variables dans le but de faciliter les interprétations des résultats que nous produirons plus tard. Nous insistons sur le caractère arbitraire de ces hypothèses qui ne reflètent que notre compréhension :

Hypothèse 1. Les variables *RUC* et *reves* sont exprimées en €

Hypothèse 2. Les montants des 3 polices sont exprimés en milliers d'€ comme les montants des sinistres

Hypothèse 3. Les contrats d'assurance automobile étant habituellement souscrits à l'année en France, la variable *durPolice1* sera supposée en années

Hypothèse 4. Le fait qu'on ne dispose que de la durée de la police 1 peut être expliqué par le fait que celle-ci est la couverture *RC* (*Responsabilité Civile* ou, dans le jargon, au tiers) obligatoire, les deux autres polices étant optionnelles dans ce cas de figure. On peut

alors supposer que pour les clients ayant choisi cette police chez l'assureur propriétaire de cette base, on connait la durée de la souscription mais pour un client n'ayant souscrit que la complémentaire 2 ou 3 ou les deux, l'information sur la durée est connue chez un autre assureur mais pas disponible dans notre base.

2 Méthodologie & Traitement des Données

Après avoir vérifié qu'il n'y a aucune valeur manquante dans tout le jeu de données, la première remarque que l'on fait est que les variables *region* et *habi* sont sous forme de caractère. Il faut les recoder sous forme de facteur pour les rendre utilisables dans la modélisation. Nous faisons de même avec les variables *nbpers*, *Nbadulte* et *censure* puisqu'il serait plus sensé de les coder en modalités (variable discrète) que de chercher à tarifer un foyer avec 3,39 adultes par exemple (variable continue).

Ensuite, nous nous intéressons au phénomène des valeurs aberrantes (figure 3) : ces valeurs extrêmes élevées qui peuvent être source de biais trop important dans les prédictions des modèles et qui, si leur rareté le justifie, peuvent être soit éliminées soit remplacées par des valeurs plus conventionnelles pour des raisons d'homogénéité dans le portefeuille.
Dans notre cas, nous choisissons de remplacer, pour les variables numériques suivantes *RUC, reves, Sinistre1, Sinistre2, Sinistre3, Police1, Police2, Police3, durPolice1, Duree* et *NSin*, les 1% de valeurs les plus extrêmes par le quantile 99% de leur distribution empirique. Notons que ce choix peut se justifier et s'illustrer par l'exemple de $Sinistre1$: 5 298 montants de ces sinistres sont en dessous de 25 842 € et seulement 53 au-delà dont le plus gros sinistre culminant à 355 000 € , ces 53 sinistres peuvent donc être considérés comme exceptionnels par rapport à la "norme" et donc remplacés par des valeurs plus "normales".

Une autre étape dans la vérification est celle de la détection de la *colinéarité* entre les variables explicatives : la qualité de prédiction du modèle n'en pâtirait pas mais l'interprétation des coefficients de la régression pourrait s'avérer ne pas être fiable ou ces coefficients eux-même ne pas sembler significatifs quand bien même ils le seraient en réalité. Pour quantifier ces liens potentiels, nous passons par le calcul du *coefficient V de Cramer* (définition 6) qui fonctionne aussi bien pour les variables quantitatives que qualitatives et qui s'interprète de la même manière qu'un coefficient de corrélation linéaire : 0 représentant l'indépendance absolue et 1 la dépendance totale. Après analyse des figures 4 et 5, on remarque les 3 faits suivants au seuil de 70% :

- la variable *RUC* est fortement liée aux variables *cs, reves, crevpp, Acompm, nbpers, enfants* et *Nbadulte*
- les variables *habi* et *Ahabi* sont totalement dépendantes : c'est la même variable alors il y a redondance

— la variable *Acompm* est fortement dépendante des variables *nbpers* et *Nbadulte* et aussi totalement dépendante de *enfants*

Remarque 5. Certaines valeurs des auto-corrélations des variables ne sont pas exactement égales à 1, ce qui est dû au fait que la statistique du test du χ^2 d'indépendance (définition 7), sur laquelle est basée le calcul du V de Cramer, est légèrement biaisée lorsque certaines classes ont des effectifs inférieurs à 5 mais cela ne fausse en rien l'interprétation.

Pour ces raisons de dépendance évoquées, on choisira de modéliser soit avec *Acompm* soit avec *enfants*, mais pas les deux en même temps. Il en va de même pour *habi* et *Ahabi*.

Pour les mêmes raisons, on ne garde qu'une seule variable entre *Duree* et *durPolice1* qui sont aussi corrélées à 99,99 %. Étant en assurance automobile, où les contrats sont habituellement à durée annuelle, on choisit de garder la seconde en la supposant en années.

Bien que non observables à l'entrée d'un ménage, les variables *Police1, Police2* et *Police3* peuvent avoir une influence positive sur la sinistralité. En effet, on remarque parfois en automobile que le fait de se savoir mieux couvert peut avoir un effet de relâchement des conducteurs (théorie de l'*aléa moral* [2]). On va donc créer trois nouvelles variables *Pol1, Pol2* et *Pol3* qui vont synthétiser le choix de souscrire ou non à une des 3 polices et qui, elles, seront observables lors d'une nouvelle entrée en portefeuille.

Notre base de données finale comportera donc 29 variables mais seulement 26 que nous utiliserons.

Maintenant, intéressons nous à quelques statistiques sur les individus la composant. Sur les 5 352 ménages, on en compte 4 579 payant des primes pour la police 1, 5 250 pour la police 2 et 4 940 pour la police 3. Certains de ces ménages étant les mêmes, on en dénombre en tout 4 231 qui payent pour les 3 polices en même temps. 18 ménages ont uniquement souscrit la police 1, 99 uniquement la police 2 et 23 uniquement la police 3.
On a aussi noté 13 individus qui ne payent aucune prime pour aucune police, ce qui nous semble être une anomalie au vu de l'hypothèse suivante : puisqu'on est en assurance automobile, on a supposé que la police 1 pouvait représenter l'assurance au tiers, la police 2 serait une assurance complémentaire par exemple bris de glace et la police 3 une couverture complète tous risques par exemple. Sous cette hypothèse, il serait incohérent de laisser des clients dans la base qui ne souscrivent aucune de ces 3 options. On choisit donc de les enlever pour ne garder que les 5 339 restants.

La consigne pour ce rapport est d'en enlever les 3 premiers individus que nous garderons de côté pour la tarification.
Les autres serviront de base pour la connaissance des types de sinistre où la compagnie est exposée. Étant donné que les erreurs de prédiction auraient tendance à être trop optimistes si l'échantillon qui a servi à ajuster le modèle est le même que celui sur lequel il est testé, on va aussi partitionner ces 5 336 individus en deux sous-échantillons choisis de manière totalement aléatoire : un sous-échantillon d'apprentissage comprenant 5 230 ménages, soit

98 % de la population, et un sous-échantillon de test en comprenant 106, les 2 % restants. Notre base est donc enfin prête pour la modélisation.

3 Modélisation des Sinistres

La maîtrise des sinistres implique de chercher à pouvoir prédire pour un nouveau ménage entrant dans le portefeuille une ou plusieurs des caractéristiques suivantes :

— l'**occurence** du sinistre
— le **nombre** de sinistres
— les **montants** des sinistres
— la **durée** de vie du contrat

3.1 L'Occurence du Sinistre

Les modèles de régression adaptés à la prédiction de l'**occurence** ou non d'un sinistre sont des modèles dits à variable réponse *dichotomique*. Parmi ceux possibles, on a choisi le modèle ***Probit*** (définition 9) qui est un cas particulier des modèles linéaires généralisés, et le modèle ***Random Forest*** (définition 11) qui utilise en partie l'algorithme des *arbres de classification CART* (définition 10). Pour pouvoir implémenter ces deux méthodes, il nous faut une variable réponse de type binaire. On rajoute donc à la base une 27^{e} variable utilisable (30^{e} en tout) nommée *Occurence* et définie comme ceci :

$$Occurence = \begin{cases} 1 & \text{en cas de sinistre} \\ 0 & \text{sinon} \end{cases}$$

3.1.1 Le Modèle Probit

Pour notre échantillon d'apprentissage constitué précédemment, nous allons d'abord paramétrer un premier modèle en intégrant toutes les variables explicatives retenues à ce stade. Les résultats du modèle sont donnés sous forme de probabilité d'appartenir à une classe. Ce qui nous donne la règle de classification suivante : on classe en 1 pour une probabilité supérieure au seuil de 0,5 et en 0 sinon. On évaluera l'erreur en comparant les différences entre les prédictions du modèle et les vraies valeurs du sous-échantillon de test. Ici, elle vaut **24,76 %**.

Notre modèle comporte cependant beaucoup de variables explicatives. On peut l'améliorer à ce niveau en optimisant le critère de l'*AIC* (*Akaike Information Criterion*, définition

12). C'est un critère de parcimonie qui pénalise les modèles avec beaucoup de variables explicatives au profit de ceux qui en comportent peu. Pour le modèle précédent, l'AIC valait approximativement **5 597,26**. Avec une procédure *step*, ajout ou retrait de variable étape par étape, on cherche à minimiser ce critère pour arriver au résultat final d'à peu près **5 587,45**. Ce modèle retient les 13 variables explicatives suivantes : *cs, reves, region, habi, Atyph, agecat, nbpers, enfants, Bauto, Pol1, Pol2* et *Pol3*. Son erreur de prédiction vaut **24,72 %**.

Enfin, nous testons un troisième et dernier modèle qui est directement issu de celui que nous venons de construire en en éliminant les variables qui ne sont pas significatives au niveau de confiance de 95 % au test de *Wald* (définition 14). Ce dernier modèle nous conduit à retenir les 9 variables explicatives suivantes : *reves, habi, Atyph, agecat, nbpers, Bauto, Pol1, Pol2* et *Pol3* pour un AIC d'environ **5 617,6**, son erreur est calculée à **24,89 %**.

3.1.2 Le Modèle Random Forest

La méthode Random Forest base son algorithme sur la construction d'un arbre de classification. Nous avons choisi de la tester ici car son utilisation se généralise de plus en plus dans la tarification automobile en vertu de sa facilité d'implémentation et de son aptitude à fournir une vision synthétique du portefeuille de risques assurés [1]. Il se met en oeuvre de la manière suivante :

- Constitution d'un nouvel échantillon de la même taille que l'échantillon d'origine sur le principe du bootstrap : tirage aléatoire et avec remise
- Tirage aléatoire et sans remise de p variables parmi les k variables de départ (en général, on choisit $p < \sqrt{k}$)
- Choix du meilleur séparateur parmi les p variables retenues selon un critère défini (en général, l'indice de *Gini* que nous ne définirons pas ici)
- Construction du premier noeud suivant ce séparateur par un choix de l'individu de référence qui minimise l'erreur de prédiction
- On recommence à nouveau le tirage jusqu'à construction de l'arbre de classification complet
- On reprend le processus depuis le début pour construire un nouvel arbre avec un nouvel échantillon pour un total de B itérations (ici, on a choisi $B = 500$)
- Le résultat final de la prédiction est déterminé selon l'aggrégation des résultats des B arbres, dans ce cas-ci, cela équivaut à un vote à la majorité

Comme précédemment, nous avons mis cette méthode en place pour les trois modèles étudiés pour les résultats suivants : des erreurs respectives de **32,08 %, 33,02 %** et **32,08 %** pour les modèles complets (17 variables explicatives), par minimisation de l'AIC (13 variables explicatives) et enfin par conservation des variables significatives au test de Wald à 95 % (9 variables explicatives).

3.1.3 Choix du Modèle Adapté

En se basant sur l'erreur de prédiction, nous sommes clairement tentés par le choix d'un

modèle probit au détriment des Random Forest. Leur inefficacité pourrait s'expliquer par plusieurs raisons que nous choisissons de ne pas développer pour des raisons de concision. Concernant le modèle retenu, on fera le choix de garder le **probit** d'AIC minimal qui comporte les 10 variables explicatives suivantes ***cs, reves, region, habi, agecat, nbpers, Bauto, Pol1, Pol2*** et ***Pol3***.

En analysant les effets marginaux (définition 13) du modèle retenu (figure 6), on peut constater certains faits intéressants. Toutes choses égales par ailleurs, le fait d'appartenir à la catégorie d'âge 41-50 ans fait diminuer la probabilité d'occurence du sinistre de 0,04 par rapport aux 21-40 ans ; probabilité qui diminue de 0,07 pour les 51-60 ans et de 0,15 pour les 61-96 ans confirmant l'intuition que la sinistralité baisse avec le fait de prendre de l'âge. Autrement dit, les vieux conducteurs auraient une conduite plus responsable que les jeunes conducteurs.
De même, toutes choses égales par ailleurs, un habitant de l'agglomération parisienne ou d'une ville de plus de 100 000 habitants verra sa probabilité d'occurence de sinistre augmentée de 0,08 par rapport à un habitant d'une commune rurale.
Enfin, on note que le fait de souscrire à la police 2 augmente, encore sous la condition que toutes les autres caractéristiques soient les mêmes, cette probabilité de 0,16 (+0,13 pour la police 3) par rapport à un ménage ayant souscrit la police 1. Ce constat semble confirmer la théorie de l'aléa moral mentionnée plus haut.

3.2 Le Nombre de Sinistres

En assurance automobile, il est important de modéliser le nombre des sinistres dans l'objectif de réaliser la tarification des clients.
Le tableau ci-dessous présente quelques statistiques du nombre de sinistres de notre base de clients :

Minimum	1^{e} quartile	Médiane	Moyenne	3^{e} quartile	Maximum	Variance
0,00	0,00	4,00	4,25	6	16	13,79

La moyenne empirique du nombre de sinistres par client est de 4,25, la variance est de 13,79, le minimum et le 1^{e} quartile sont nuls. Le nombre maximum de sinistres est de 16 après traitement des valeurs aberrantes.
Pour ce genre de situations où on traite des valeurs d'une variable discrète, on a recours aux modèles de comptage.

3.2.1 Le Modèle de Poisson

La loi de **Poisson** de paramètre λ, notée $\mathcal{P}(\lambda)$, également appelée la loi des évènements rares, est une loi de probabilité discrète, qui se définit de la façon suivante :

Soit X une variable aléatoire, on dit que X suit une loi de Poisson de paramètre λ si :

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \text{ où k} \in \mathbb{N}$$

L'espérance et la variance d'une loi de Poisson sont : $\mathbb{E}(X) = \mathbb{V}(X) = \lambda$.

On utilise ici un modèle de Poisson pour modéliser le nombre de sinistres :

$$NSin_i \text{ le nombre de sinistres d'un client } i,\ NSin_i \sim \text{P}(\lambda_i)$$

A l'aide d'une procédure *step*, on sélectionne donc les variables significatives avec le critère AIC : *pcs*, *cs*, *reves*, *crevpp*, *region*, *habi*, *Atyph*, *agecat*, *nbpers*, *Bauto*, *Nbadulte*, *Pol1*, *Pol2* et *Pol3*. Et on effectue une régression de Poisson du nombre de sinistres *NSin* par rapport aux variables explicatives sélectionnées avec les données d'apprentissage, puis on réalise des prédictions sur ce modèle avec les données test. L'AIC de ce modèle est de **27791**. Et son erreur de prédiction est de **3,05**, ce qui signifie que notre modèle se trompe en moyenne de $\pm$ 3 sinistres sur le nombre total de sinistres que peut avoir un client.

3.2.2 Le Modèle Binomial Négatif

La loi **binomiale négative** de paramètres n et p, $BN(n, p)$, est une loi de probabilité discrète. Elle permet calculer la probabilité d'effectuer k expériences identiques et indépendantes pour obtenir n fois un évènement de probabilité p : p correspondant à un "succès" et $1 - p$ à un "échec".
Soit X une variable aléatoire, on dit que X suit une loi binomiale négative si :

$$\mathbb{P}(X = k) = C^k_{n+k-1}p^n(1-p)^k \text{ où k} \in \mathbb{N}$$

Son espérance et sa variance sont : $\mathbb{E}(X) = \frac{n(1-p)}{p}$ et $\mathbb{V}(X) = \frac{n(1-p)}{p^2}$.

On modélise donc le nombre de sinistres par un modèle binomial négatif :

$$NSin_i \text{ le nombre de sinistres d'un client } i,\ NSin_i \sim \text{BN(n, p}_i\text{)}$$

De même que pour le modèle précédent on utilise la procédure *step* et ainsi le critère AIC. On obtient les 10 variables explicatives suivantes : *pcs*, *crevpp*, *region*, *Atyph*, *agecat*, *nbpers*, *enfants*, *Pol1*, *Pol2* et *Pol3*. On fait la régression binomiale négative du nombre de sinistres *NSin* par rapport aux 10 variables explicatives précédentes avec les données d'apprentissage. On obtient un AIC égal à **25247**. L'erreur de prédiction sur l'échantillon test est de **3,03**.

3.2.3 Choix du Modèle Adapté

Le nombre de sinistres *NSin* est une variable de comptage qui permet de dénombrer les sinistres des clients. Nos modèles cherchent à estimer le nombre de sinistres qu'un client va avoir en fonction de ses caractéristiques.
En considérant le critère des erreurs de prédictions, on remarque que les erreurs pour les deux modèles sont très proches (3,05 pour le modèle de Poisson et 3,03 pour le modèle binomial négatif), même si l'erreur de prédiction du modèle binomial négatif est légèrement meilleure.
En ce qui concerne le critère AIC, le but étant de minimiser l'AIC, on choisit le modèle ayant le plus faible AIC, soit le modèle binomial négatif avec un AIC de 25247 contre 27791 pour le modèle de Poisson.
Enfin, dans le modèle de Poisson, l'espérance de la loi est égale à la variance. Il n'y a donc pas de dispersion. Pour éviter ce problème, on peut introduire le modèle de Poisson surdispersé. Cependant, afin de gérer le problème de dispersion (l'espérance du nombre de sinistres étant différente de la variance), on a opté pour le modèle binomial négatif qui est aussi approprié pour les variables de comptage ayant beaucoup de variation.
Au vu de ces éléments, nous retiendrons le modèle binomial négatif dans la suite du document pour l'estimation du nombre de sinistres. D'après ce modèle, dans les mêmes conditions de vie, un ménage locataire de son logement verrait son nombre de sinistres moyen augmenter de 0,07 par rapport à un ménage propriétaire.

3.3 Les Montants des Sinistres

3.3.1 Le Modèle Gamma

La loi **Gamma** est très souvent choisie en assurance pour modéliser des coûts de sinistre. On dit que la variable aléatoire X suit une loi de Gamma (notée Γ) si sa fonction de densité de probabilité est la suivante :

$$X \sim \Gamma(k,\theta) \iff f(x,k,\theta) = \frac{x^{k-1}e^{\frac{-x}{\theta}}}{\Gamma(k)\theta^k}$$

avec $x > 0$ et $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ la fonction *Gamma d'Euler*. L'utilisation d'un modèle Gamma suppose qu'on a déjà modélisé le fait que le client va être source de sinistralité, autrement dit que la variable *Occurence* vaut déjà 1. Pour cette partie, que nous ne choisissons d'illustrer uniquement avec le sinistre de type 3, nous travaillons avec la partie de la base qui ne concerne que les clients pour lesquels le montant de ce sinistre est strictement positif (3 565 clients). Nous les départageons en échantillon d'apprentissage de taille 3 494 et en échantillon de test de taille 71. Nous commençons par modéliser en prenant en compte toutes les variables retenues à ce stade, pour ensuite épurer le modèle avec le critère de l'AIC et

finir par ne retenir que les variables avec au moins une modalité significative suivantes : *pcs, cs, region, habi, nbpers* et bien entendu *Pol3*. Nous confrontons un modèle avec une fonction de lien inverse à un autre avec une fonction de lien identité, pour garder finalement le dernier en raison d'une meilleure erreur de prédiction sur l'échantillon de test et d'un meilleur AIC. Cette erreur absolue moyenne vaut 2 550 € sur les montants des sinistres de type 2. L'analyse des résultats de ce modèle nous montre l'importance capitale de la variable *nbpers* avec les conclusions suivantes : par rapport à un ménage avec une seule personne et tout le reste étant similaire, un ménage avec deux personnes verrait son le montant de ses sinistres de type 3 augmenté de 750 € en moyenne ; un chiffre pouvant aller jusqu'à 1 770 € pour un ménage avec 6 personnes.

3.3.2 Le Modèle Tobit

On définit le modèle **Tobit** par l'exemple suivant :

$$\begin{cases} Sinistre^* &= X'\beta + \varepsilon \\ Sinistre &= Sinistre^* \mathbb{1}_{\{Sinistre^* > 0\}} \end{cases}$$

pour une variable *Sinistre* observable seulement si sa valeur est positive ; on parle alors de censure des données. Ce modèle est adapté à un type de situation où ce sont les mêmes variables qui déterminent à la fois l'occurence et le montant du sinistre. Encore une fois, on s'intéresse aux effets marginaux afin de rendre notre modèle interprétable mais aussi aux erreurs de prédiction sur l'échantillon de test pour vérifier sa performance. Précision importante : on veut les effets marginaux sur la "vraie" variable $Sinistre^*$ et non sur sa version censurée. Ici, nous l'appliquons uniquement au sinistre 2. Sa particularité fait que nous n'avons pas besoin de filtrer les sinistres positifs contrairement au modèle Gamma.
Comme précédemment, nous cherchons les variables les plus significatives dans la régression par le Tobit. Et nous trouvons qu'elles sont *Pol1, pcs, RUC, reves, crevpp, region, habi, agecat, enfants, Anat, Nbadulte* et *NSin*. Le Tobit se trompe en moyenne de 3 400 € environ en valeur absolue par prédiction de coût de sinistre total. Le résumé de ce modèle met en lumière deux faits marquants pour nous : le fait de ne pas avoir d'enfants augmente le coût de sinistre de 930 € par rapport à un ménage où on recense un enfant, toutes choses égales par ailleurs. Sous la même condition, un ménage verra son coût de sinistre total revu à la baisse de 4 € en moyenne pour 1 € supplémentaire de revenu estimé.

L'estimateur du maximum de vraisemblance du modèle Tobit n'est pas convergent en général si les résidus sont non normaux et/ou hétéroscédastiques.

3.3.3 Le Modèle Tobit Généralisé

Dans sa définition, le modèle **Tobit généralisé** est similaire au modèle Tobit de la section précédente, à la différence près que ce ne sont pas les mêmes variables qui déclenchent

l'occurence du sinistre et qui déterminent son montant. Il offre plus de liberté de modélisation dans ce sens. On le définit comme ceci :

$$Y_i = \begin{cases} 0 & \text{si } U_1^* = Z_i\gamma + \eta_i < 0 \\ Y_i^* = X_i\beta + \varepsilon_i & \text{si } U_i^* = Z_i\gamma + \eta_i > 0 \end{cases}$$

avec U_i la variable latente non observée

La première étape consiste en une estimation d'une indicatrice qui détermine si le montant du sinistre sera nul ou non. La deuxième étape consiste en une sélection de Y uniquement pour les valeurs positives selon la régression suivante :

$$Y_i = X_i\beta + \lambda\frac{\phi(Z_i\hat{\gamma})}{1 - \Phi(Z_i\hat{\gamma})} + \eta_i$$

Cette régression est basée sur l'*inverse du ratio de Mills*.

Avant de la mettre en oeuvre, nous décidons d'abord de corriger d'éventuelles endogénéités en régressant les polices par les autres variables, en récupérant les valeurs pour une nouvelle régression sur les sinistres (*2 stage least squares*). En l'implémentant sur nos données pour le sinistre 2, nous obtenons un résumé qui indique que les variables les plus significatives pour la première haie sont notamment la région et la catégorie socio-professionnelle. On notera que par rapport à un habitant de la région Île-de-France, un habitant de la région codée 3 aura un montant de sinistre 2 inférieur en moyenne de 500 € et un habitant de la région codée 2 aura 300 € en moins, toutes choses égales par ailleurs.

3.3.4 Le Modèle Double Hurdle

Les deux modèles Tobit vus précédemment sont des cas particuliers du modèle **double hurdle** (double haie) défini comme ceci :

$$Y_{i=}\begin{cases} 0 & \text{si } U_i^* = Z_i\gamma + \eta_i < 0 \\ Y_i^* & \text{si } Y_i^* = X_i\beta + \varepsilon_i > 0 \end{cases}$$

Il est similaire dans son fonctionnement au Tobit généralisé mais se distingue par des conditions d'inversibilité de sa matrice hessienne plus complexes. Nous allons directement à son interprétation qui dénote un fait marquant : par rapport à un agriculteur, un retraité ou un ouvrier verrait son montant de sinistre baisser en moyenne de 700 € dans des conditions similaires. Nous garderons le modèle Gamma qui nous semble plus adapté aux données.

Nous retironsla variable Duree. On va étudier ici la durée d'adhésion à la police 1.
Une durée est positive et elle peut être censurée ou non :

$$\delta_i = \begin{cases} 1 \text{ si la donnée est censurée} \\ 0 \text{ sinon} \end{cases}$$

On introduit quelques notations :
— T, la durée, variable aléatoire continue, positive
— $F(t) = P[T < t]$, la fonction de répartition
— $s(t) = 1 - F(t) = P[T > t]$, la survie de la variable aléatoire T
— c_i, la variable de censure

3.4.1 Le Modèle de Cox

Le modèle de Cox ou modèle à hasard proportionnel paramétrique :

$$\lambda(t) = \lambda_0(t)exp(X'\beta) \text{ avec X} \in \mathbb{R}^k \text{ variables explicatives et } \beta \in \mathbb{R}^k$$

$$\lambda_0(t) = \lambda_0(t, \theta), \lambda_0(t) \text{ étant le hasard de base}$$

$$s(t) = exp[-exp(X'\beta) \int_0^t \lambda_0(u)du] \text{ loi de type Gumbel en X}$$

On cherche à maximiser la vraisemblance d'un échantillon de données censurées. On écrit donc la vraisemblance partielle :

$$\prod_{i \in I} \frac{exp(x_i'\beta)}{\sum_{j>i} exp(x_j'\beta)}$$

On a effectué une régression de Cox en prenant en compte la censure, sur l'échantillon d'apprentissage.
Au vu des résultats, on remarque que la composition du ménage a un impact très important sur la durée d'adhésion d'un client. Pour un couple avec enfants, le risque de quitter l'assurance augmente de 56% tandis qu'au contraire le risque de quitter l'assurance diminue de 52% pour un couple sans enfants et il diminue de 83% dans le cas d'une personne seule. La catégorie socio-professionnelle augmente le risque de quitter l'assurance de 20% pour les personnes sans activité professionnelle, de respectivement 53% et 88% pour les ouvriers et les employés etde 95% pour les retraités, les cadres et professions intellectuelles supérieures. Entre 41 et 50 ans, le risque de quitter l'assurance augmente de 12%, il est de 8% entre 51 et 60 ans mais il diminue de 3% au delà de 60 ans. Et enfin, le risque de quitter l'assurance diminue d'autant plus pour les habitants de Paris et sa banlieue que pour les habitants de villes de moins de 100 000 habitants.
On a également réalisé des prédictions sur les données test et on a obtenu une erreur de prédiction de 4 pour la durée d'adhésion à la police 1 (*durPolice1*), c'est-à-dire qu'en moyenne notre modèle peut se tromper de ± 5 ans .

3.4.2 L'Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur qui permet d'estimer une fonction de survie selon des données de durée de vie. L'objectif est d'estimer la survie $s(t)$ en ayant observé $(X_i, \delta_i)_{i=1,..,n}$, $X_i = min(T_i, c_i)$ et $\delta_i = \mathbb{1}_{\{T_i<c_i\}}$.
L'estimateur de Kaplan-Meier est défini par la formule suivante :

$$\hat{S}_n^{KM}(t) = \prod_{i=1}^{n}(1 - \frac{1-\delta_i}{n-i+1})^{\mathbb{1}_{\{X_i<t\}}}$$

$\hat{S}_n^{KM}(t) \longrightarrow s(t) sur [0,\tau_H]$ avec $\tau_H = Inf\{$ t tel que $P(X > T) = 0\}$

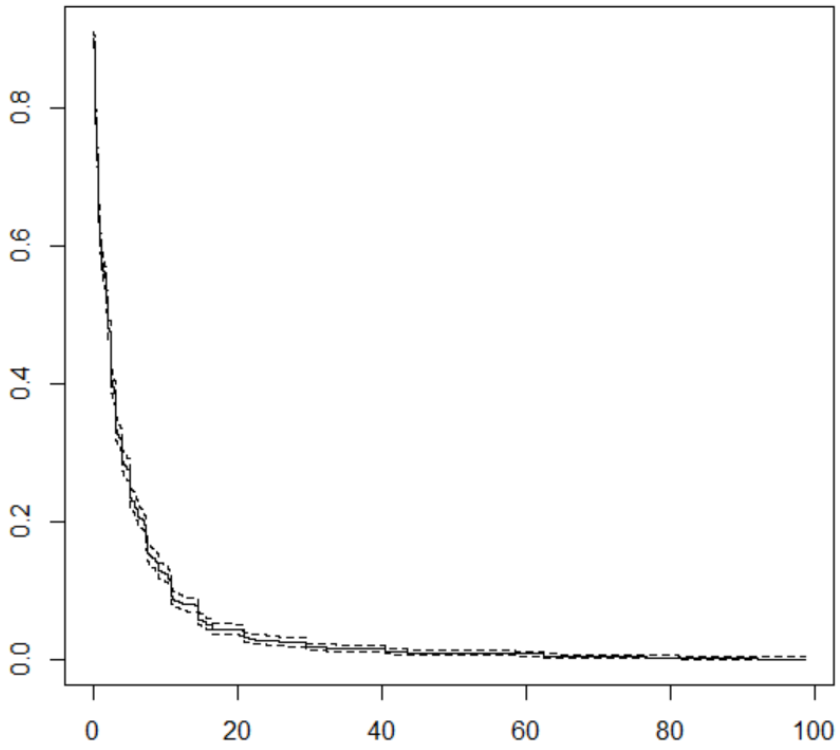

FIGURE 2 – Courbe de survie de l'estimateur de Kaplan-Meier

La figure 2 ci-dessus représente la courbe de survie pour l'estimateur de Kaplan-Meier. En abscisse, on peut observer la durée en années et en ordonnée, il y a la probabilité de survie. $\hat{S}_n^{KM}(t)$ est décroissante. En pointillés, sont représentés les intervalles de confiance à 95% pour la survie. C'est une fonction en escalier. Au vu de la courbe, on peut choisir $\tau_H = 40$.

4 Tarification

La **tarification** consiste à fixer un prix *i.e.* une prime d'assurance. Ici, nous faisons de la tarification *à priori*, c'est à dire que nous décidons de calculer la prime sur la base des informations que nous avons avant que le client ne fasse partie du portefeuille d'assurés. Le risque étant que si notre modèle de prédiction se trompe, la compagnie pourrait subir des pertes. Fort heureusement, en France comme dans la majorité des pays, la sinsitralité est révisable annuellement via un système de bonus/malus qui permet d'effectuer des corrections et d'ajuster la tarification.

Tarification rime souvent avec *segmentation* dans le monde de l'assurance. La segmentation tarifaire consiste à faire payer à chaque client (ou groupe de clients ayant les mêmes caractéristiques) le prix de son risque. Tout l'enjeu de cette méthode est de trouver les segments adéquats qui vont au-delà de la simple notion binaire de bon ou de mauvais risque : en assurance, tous les risques sont bons à prendre pour l'assureur du moment qu'ils sont bien tarifés au bon prix. Puisque ce n'est pas le sujet principal de ce mémoire, nous nous contenterons de segments arbitraires même si dans la vraie vie, cette question est sujette à des études parfois très poussées.

Au vu de nos modèles retenus, notre méthode de tarification sera la suivante :
— prédire si le client aura une occurence de sinistre ou non avec le modèle Probit : si ce n'est pas le cas, on lui fixera un tarif standard ajusté pour les clients "non sinistrables"
— si le client a été prédit "sinistrable", on prédit son nombre probable de sinistres avec le modèle binomial négatif et le coût total de ses sinistres avec le modèle Gamma
— prédire la durée de vie du contrat

Ces 4 étapes serviront à la tarification. Nous représentons ci-après les prédictions obtenues pour les 3 premières individus, représentées seulement pour le sinistre de type 3 (valeurs prédites en gras) :

	Ménage 1		Ménage 2		Ménage 3	
Occurence	**1**	0	**1**	1	**1**	1
Nombre de sinistres	**3**	0	**6**	6	**4**	4
Montant des sinistres	**2.878**	0	**3.051**	0.46	**2.112**	0.52
Durée	**0.62**	11.06	**0.92**	0.05	**0.06**	0

Nous calculons la prime pure qui vaut respectivement **959, 508** et **528 €** pour les 3 clients. Avec un taux de chargement, de gestion et de frais annexes choisi arbitrairement de 10 %, on obtient les primes d'assurances respectives de **1 055, 559** et **581 €** .

Conclusion

À travers l'étude de notre base, nous avons pu aborder différents principes de la tarification en assurance automobile. Nous avons ainsi pu voir l'importance de la modélisation via différents modèles plus ou moins adaptés aux données ou à la situation. Nous avons aussi pu nous projeter dans la prédiction des variables tarifaires et aborder de manière superficielle la segmentation.

5 Annexe 1 : Graphiques et Sorties *R*

	variable	mean	p_50	p_95	p_99	max
1	RUC	6.277521e+03	5.500000e+03	13235.2900	19117.65000	3.529412e+04
2	reves	1.487995e+04	1.125000e+04	27500.0000	40000.00000	3.416250e+06
3	Sinistre1	1.242663e+00	0.000000e+00	3.9000	25.89800	3.550000e+02
4	Sinistre2	1.615049e-01	0.000000e+00	0.6500	3.09800	3.110000e+01
5	Sinistre3	1.837128e+00	7.050000e-01	7.2990	12.33215	4.022000e+01
6	Police1	3.750700e+00	1.950000e+00	13.4545	23.08685	5.498500e+01
7	Police2	1.301746e+01	9.060000e+00	39.2764	60.74648	1.241090e+02
8	Police3	2.110487e+00	1.420000e+00	6.6358	11.69000	3.474300e+01
9	durPolice1	5.190665e+08	4.387771e-01	10.8294	40.35390	2.778043e+12
10	Duree	5.190665e+10	4.400000e+01	1083.0000	4035.00000	2.778043e+14
11	NSin	4.249253e+00	4.000000e+00	12.0000	16.00000	3.000000e+01

FIGURE 3 – Détection de potentielles valeurs aberrantes pour les variables de type numérique

	pcs	RUC	cs	reves	crevpp	region	habi	Ahabi
pcs	1	0.400449377595096	0.261733791873844	0.222262743910972	0.260498617036343	0.0882515850857117	0.117408040366573	0.149634945486635
RUC	0.400449377595096	1	0.992716877903476	0.954233021397693	0.994310943724136	0.239144447663374	0.23516335246539	0.262964188493844
cs	0.261733791873844	0.992716877903476	1	0.531758929067107	0.647030247463736	0.164128404679296	0.164411382761294	0.162603146994375
reves	0.222262743910972	0.954233021397693	0.531758929067107	1	0.405511271532949	0.112436480980777	0.102767572985719	0.132932756401464
crevpp	0.260498617036343	0.994310943724136	0.647030247463736	0.405511271532949	1	0.150786326025751	0.161831601250723	0.158569044712749
region	0.0882515850857117	0.239144447663374	0.164128404679296	0.112436480980777	0.150786326025751	1	0.367819718249296	0.475830439260804
habi	0.117408040366573	0.23516335246539	0.164411382761294	0.102767572985719	0.161831601250723	0.367819718249296	1	1
Ahabi	0.149634945486635	0.262964188493844	0.162603146994375	0.132932756401464	0.158569044712749	0.475830439260804	1	1
Atyph	0.157930126313134	0.281495903085839	0.0944287956604056	0.144822601633276	0.0664519131632025	0.0740233897860049	0.160653774261264	0.155454472750761
agecat	0.522622426545316	0.515718602933578	0.12654439173218	0.146197148469324	0.182179369932569	0.0641993619984483	0.0469002298950176	0.039219497896624
Acompm	0.327474973663456	0.924361714666352	0.251265216920886	0.221470832717138	0.332324562432139	0.0651312607893781	0.0771689296679937	0.0729192808689525
nbpers	0.209907372440808	0.929718193682969	0.268211835137381	0.135671489444991	0.407657882233796	0.0611382829494102	0.0668244737481105	0.0835121297290437
enfants	0.381749331594611	0.921407126670878	0.250287541786017	0.17254481012312	0.318941821418603	0.0665530886079691	0.0479418179950902	0.0408182396110005
Anat	0.0763604977987703	0.331539209863503	0.0991157494373669	0.0784367349680238	0.124733719419355	0.0838280460436264	0.0884956752739358	0.0851134380300591
Bauto	0.226030586948402	0.436879182954103	0.0970643065899732	0.318443945487747	0.052607303961565	0.152443581922881	0.186645589802287	0.182682975413225
Nbadulte	0.129276047669744	0.894095536610071	0.165468104484811	0.157305997211315	0.261464424408968	0.0521561884412392	0.0700584617837568	0.0852486446781775

FIGURE 4 – Détection de fortes dépendances pour les variables explicatives (seuil de 70%)

	Atyph	agecat	Acompm	nbpers	enfants	Anat	Bauto	Nbadulte
pcs	0.157930126313134	0.522622426545316	0.327474973663456	0.209907372440808	0.381749331594611	0.0763604977987703	0.226030586948402	0.129276047669744
RUC	0.281495903085839	0.515718602933578	0.924361714666352	0.929718193682969	0.921407126670878	0.331539209863503	0.436879182954103	0.894095536610071
cs	0.0944287956604056	0.12654439173218	0.251265216920886	0.268211835137381	0.250287541786017	0.0991157494373669	0.0970643065899732	0.165468104484811
reves	0.144822601633276	0.146197148469324	0.221470832717138	0.135671489444991	0.17254481012312	0.0784367349680238	0.318443945487747	0.157305997211315
crevpp	0.0664519131632025	0.182179369932569	0.332324562432139	0.407657882233796	0.318941821418603	0.124733719419355	0.052607303961565	0.261464424408968
region	0.0740233897860049	0.0641993619984483	0.0651312607893781	0.0611382829494102	0.0665530886079691	0.0838280460436264	0.152443581922881	0.0521561884412392
habi	0.160653774261264	0.0469002298950176	0.0771689296679937	0.0668244737481105	0.0479418179950902	0.0884956752739358	0.186645589802287	0.0700584617837568
Ahabi	0.155454472750761	0.039219497896624	0.0729192808689525	0.0835121297290437	0.0408182396110005	0.0851134380300591	0.182682975413225	0.0852486446781775
Atyph	1	0.207274953829999	0.121246102574861	0.0743103906273142	0.116380305314238	0.0845930800099886	0.150209578626232	0.132118581774717
agecat	0.207274953829999	1	0.442960585172643	0.330958888848026	0.595399720501039	0.0642339494899803	0.157408621022768	0.305566886524263
Acompm	0.121246102574861	0.442960585172643	1	0.784684105871826	1	0.213973132341664	0.328054871683633	0.730817078398173
nbpers	0.0743103906273142	0.330958888848026	0.784684105871826	1	0.510875468679916	0.219743252598675	0.332530346031737	0.585841617225019
enfants	0.116380305314238	0.595399720501039	1	0.510875468679916	0.999511568935127	0.0615917883415729	0.120387791310245	0.463079821084764
Anat	0.0845930800099886	0.0642339494899803	0.213973132341664	0.219743252598675	0.0615917883415729	1	0.0741213573393317	0.20286331523893
Bauto	0.150209578626232	0.157408621022768	0.328054871683633	0.332530346031737	0.120387791310245	0.0741213573393317	0.99876947809041	0.339084066813258
Nbadulte	0.132118581774717	0.305566886524263	0.730817078398173	0.585841617225019	0.463079821084764	0.20286331523893	0.339084066813258	1

FIGURE 5 – Suite de la figure 2

```
Call:
probitmfx(formula = Occurence ~ habi + agecat + nbpers + Bauto +
    Pol1 + Pol2 + Pol3, data = App[, -c(8, 11, 17:26)])

Marginal Effects:
                          dF/dx  Std. Err.       z     P>|z|
habi1                 0.0241170  0.0245672  0.9817  0.326261
habi2                 0.0584553  0.0262068  2.2305  0.025712 *
habi3                 0.0234351  0.0280489  0.8355  0.403431
habi4                 0.0648951  0.0229366  2.8293  0.004665 **
habi5                 0.0910035  0.0212183  4.2889 1.796e-05 ***
habi6                 0.0382689  0.0222417  1.7206  0.085325 .
habi7                 0.0839247  0.0161077  5.2102 1.886e-07 ***
habi8                 0.0782702  0.0182949  4.2782 1.884e-05 ***
agecat41-50          -0.0464495  0.0177002 -2.6242  0.008684 **
agecat51-60          -0.0678634  0.0215963 -3.1424  0.001676 **
agecat61-96          -0.1565444  0.0209335 -7.4782 7.537e-14 ***
nbpers2               0.1036595  0.0175136  5.9188 3.243e-09 ***
nbpers3               0.1156302  0.0182475  6.3368 2.346e-10 ***
nbpers4               0.1296623  0.0188806  6.8675 6.535e-12 ***
nbpers5               0.1571047  0.0179893  8.7332 < 2.2e-16 ***
nbpers6               0.1357742  0.0269911  5.0303 4.896e-07 ***
nbpers7               0.0636456  0.0669694  0.9504  0.341926
nbpers8               0.0369669  0.1451533  0.2547  0.798974
nbpers9               0.0318869  0.1656983  0.1924  0.847398
nbpers10              0.2443134  0.0063171 38.6746 < 2.2e-16 ***
BautoPas de vehicule -0.0499707  0.0247044 -2.0227  0.043100 *
Pol11                 0.0820411  0.0188492  4.3525 1.346e-05 ***
Pol21                 0.1539011  0.0552172  2.7872  0.005317 **
Pol31                 0.1265753  0.0261492  4.8405 1.295e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 6 – Marginales du modèle Probit retenu pour modéliser l'occurence du sinistre

.

```
Demand analysis with the Almost Ideal Demand System (AIDS)
Estimation Method: Linear Approximation (LA) with Stone Index (S)
Coefficients:
alpha
    Part1     Part2     Part3
0.4602606 0.2828485 0.2568909
beta
      Part1       Part2       Part3
-0.10891093  0.07484784  0.03406309
gamma
         Police1    Police2    Police3
Part1  0.1414394 -0.1191219 -0.0223175
Part2 -0.1191219  0.1893730 -0.0702511
Part3 -0.0223175 -0.0702511  0.0925686
```

6 Annexe 2 : Rappels & Définitions

Définition 6. **Coefficient V de Cramer** mesurant la dépendance entre les variables X et Y :

$$V = \sqrt{\frac{\chi^2}{n \times min(n_i, n_j) - 1}}$$

avec n le nombre d'individus, n_i et n_j respectivement le nombre de modalités de X et Y

Définition 7. **Statistique de test du χ^2 d'indépendance** pour une variable X à i modalités et Y à j modalités :

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

avec :
$O_{i,j}$: l'effectif observé de la classe i, j
$E_{i,j}$: l'effectif attendu de la classe i, j si les variables sont indépendantes

Définition 8. **Prime pure** : montant du sinistre moyen ou encore espérance de la perte par sinistre pour l'assureur. Pour un coût de sinistre individuel S et un nombre de sinistre N,elle correspond à :

$$\mathbb{E}(S) = \frac{1}{N} \sum_{i=1}^{N} S_i$$

Définition 9. Y est régressé par X selon un modèle dichotomique **Probit** si :

$$\mathbb{P}(Y = 1|X) = \Phi(X'\beta)$$

ou encore

$$\mathbb{P}(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

avec Y variable binaire, X vecteur des k variables explicatives précédées de la constante, β vecteur des coefficients de la régression (à estimer car inconnu à priori) et Φ la fonction de répartition de la loi normale centrée réduite

Définition 10. **Algorithme CART** (Classification And Regression Tree) : algorithme binaire d'élaboration d'un arbre de décision basé sur la construction de noeuds selon plusieurs critères (segmentation, complexité, taille minimale, etc.)

Définition 11. **Random Forest** : algorithme d'apprentissage automatique combinant la construction d'arbres de décision et le *bagging* (contraction de *bootstrap* et *aggregating*)

Définition 12.

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

avec k le nombre de paramètres et L le maximum de la fonction de vraisemblance du modèle

Définition 13. L'**effet marginal** associé à la variable k d'une régression vaut :

$$\beta_k = \frac{\partial \mathbb{E}(Y|X_k = x_k)}{\partial x_k}$$

Définition 14. Le test de **Wald** oppose les hypothèses suivantes :

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0$$

avec β_i l'effet marginal associé à la variable i dans le modèle de régression probit

7 Bibliographie

Références

[1] Rémi BELLINA. *Méthodes d'apprentissage appliquées à la tarification non-vie*. Mémoire d'actuaire, 2014. p3.

[2] Michel DENUIT Arthur CHARPENTIER. *Mathématiques de l'Assurance Non Vie*. Economica, 2005. p116.

[3] Jean-Marc ROBIN. *Économétrie des systèmes de demande*. Économie & Statistique, **324-325**, 1999. p137.

[4] Jean-Marc ROBIN. *Économétrie des systèmes de demande*. Économie & Statistique, **324-325**, 1999. p139.

8 Annexe 3 : code *R*

```
#Chargement des donnees

head(dat)

str(dat)


#sauvegarde du jeu de donnees original
jeuoriginal = dat
n = nrow(dat); n

#Verification des donnees manquantes
#install.packages("funModeling")
library(funModeling)
df_status(dat) #aucun NA

attach(dat)
```

```
#Statistiques descriptives
statdes=cbind(profiling_num(dat)[,-c(1:2,4,6:17,27)],max=c(max(RUC),
        max(reves), max(Sinistre1), max(Sinistre2), max(Sinistre3),
        max(Police1), max(Police2), max(Police3), max(durPolice1),
                                  max(Duree),max(NSin)max(Sinistre0))
statdes

#Remplacement des valeurs aberrantes au dessus du quantile 99%
for(i in 1:12){
  var = statdes[,1][i]
  dat[,var][dat[,var] > statdes[i,5]] = statdes[i,5]
}

#Verification
attach(dat)
```

```
statdes = cbind ( profiling _ num ( dat )[ , - c ( 1 : 2 ,4 ,6 : 17,27 )] , max = c ( max ( RUC ) ,
        max(reves), max(Sinistre1), max(Sinistre2), max(Sinistre3),
            max(Police1), max(Police2), max(Police3), max(durPolice1),
                          max ( Duree ) , max ( NSin )max(Sinistre0)))
statdes

#Fonction de calcul du coefficient V de Cramer
cramer = function(df, i, j) {
  s = chisq.test(df[,which(name==i)],
                 df[,which(name==j)])$statistic
  sqrt( s /
          (n*(min( nrow(table(df[,i], df[,j]))-1,
                   ncol(table(df[,i], df[,j]))-1))))
}

#Matrice des coefficients V de Cramer
matV = matrix(0, 16, 16)
name = colnames(dat)[1:16]
for(i in name) {
  for(j in name) {
    matV[which(name==i), which(name==j)] = cramer(dat, i, j)
  }
}
colnames(matV) = name; row.names(matV) = name
matV = data.frame(matV)
matV

matV1 = matV[,1:8]; matV2 = matV[,9:16]

#Mise en evidence des variables tres dependantes
#install.packages("DT")
library(DT)
seuil = 0.7
datatable(matV1) %>%
  formatStyle(columns = "pcs",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "RUC",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white")))
%>%
  formatStyle(columns = "cs",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "reves",
```

```
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white"))) %>%
  formatStyle(columns = "crevpp",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "region",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white")))
%>%
  formatStyle(columns = "habi",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "Ahabi",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white")))

datatable(matV2) %>%
  formatStyle(columns = "Atyph",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "agecat",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white")))
%>%
  formatStyle(columns = "Acompm",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "nbpers",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white"))) %>%
  formatStyle(columns = "enfants",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "Anat",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white")))
%>%
  formatStyle(columns = "Bauto",
              background = styleInterval(c(seuil,2)-1e-6,
                                         c("white", "lightblue","white"))) %>%
  formatStyle(columns = "Nbadulte",
              background = styleInterval(c(seuil, 2)-1e-6,
                                         c("white", "lightblue", "white")))

#Correlation entre durPolice1 et Duree
```

```
cor(dat$durPolice1, dat$Duree)

#Statistiques sur les clients
sum(dat$Police1 > 0) #4579
sum(dat$Police2 > 0) #5250
sum(dat$Police3 > 0) #4940
sum(dat$Police1>0 & dat$Police2>0 & dat$Police3>0)  #4231

sum(dat$Police1[dat$Police2==0 & dat$Police3==0]>0) #18
sum(dat$Police2[dat$Police1==0 & dat$Police3==0]>0) #99
sum(dat$Police3[dat$Police1==0 & dat$Police2==0]>0) #23

#13 clients qui ne payent aucune police
sum(dat$Police1==0 & dat$Police2==0 & dat$Police3==0) #13
which(dat$Police1==0 & dat$Police2==0 & dat$Police3==0)
View(dat[which(dat$Police1==0 & dat$Police2==0 & dat$Police3==0),])

#On retire les individus qui ne payent aucune police (anomalie)
dat = dat[-which(dat$Police1==0 & dat$Police2==0 & dat$Police3==0),]
str(dat)
n = nrow(dat); n

#Indice des echantillons de test et d'apprentissage
p = 0.02
x = 5; set.seed(x)
indice = sample(1:(n-3),n-3)
indiceApp = indice[(floor((n-3)*p)+1):(n-3)]; length(indiceApp)
indiceTest = indice[1:floor((n-3)*p)]; length(indiceTest)

#Rajout de la variable Occurence comme 27eme
dat$Occurence = numeric(n)
dat$Occurence[dat$NSin > 0] = 1
dat$Occurence = as.factor(dat$Occurence)
basemodel = dat[-((n-2):n),]

#Rajout des variables Pol1, Pol2 et Pol3
dat$Pol1 = numeric(n)
dat$Pol2 = numeric(n)
dat$Pol3 = numeric(n)
dat$Pol1[dat$Police1 > 0] = 1
dat$Pol2[dat$Police2 > 0] = 1
dat$Pol3[dat$Police3 > 0] = 1
dat$Pol1 = as.factor(dat$Pol1)
dat$Pol2 = as.factor(dat$Pol2)
dat$Pol3 = as.factor(dat$Pol3)
```

```
str(dat)

#Base de donnees de tarification et de modelisation
basetarif = dat[(n-2):n,]; nrow(basetarif)
basemodel = dat[-((n-2):n),]; nrow(basemodel)

#Partitionnage des echantillons
App = basemodel[indiceApp,]; dim(App)
Test = basemodel[indiceTest,]; dim(Test)
nApp = nrow(App); nApp
nTest = nrow(Test); nTest

#Fonction de calcul du risque quadratique des predictions
sqrtmse = function(modele, data, col, type="ns") {
  ifelse(type=="ns", sqrt(mean((predict(modele, data)-data[,col])**2)),
  sqrt(mean((predict(modele, data, type=type)-data[,col])**2)))
}

#Fonction de calcul de l'erreur de prediction dichotomique
epd = function(modele, data, col) {
  pred = predict.glm(modele, data, type="response")
  pred[pred > .5] = 1
  pred[pred <= .5] = 0
  (table(pred, data[,col])[1,2] +
      table(pred, data[,col])[2,1]) / nrow(data)
}

#************************
#Section 3.1 : Occurence
#************************

#Probit
OccProbit = glm(Occurence ~. , data = App[, -c(1:3,6, 9,12,15,17:28)],
                family=binomial(link="probit"))
summary(OccProbit)

OccProbitAIC = step(OccProbit, direction="both")
summary(OccProbitAIC)

summary(glm(formula = Occurence ~ cs + reves + region + habi + Atyph +
        agecat + nbpers + enfants + Bauto + Pol1 + Pol2 +
          Pol3, family = binomial(link = "probit"),
            data = App[, -c(8,11, 17:26)]))
summary(glm(formula = Occurence ~ cs + reves + habi + Atyph +
              agecat + nbpers + enfants + Bauto + Pol1 + Pol2 +
```

```
              Pol3, family = binomial(link = "probit"),
            data = App[, -c(1:3,6,9,12,15,17:28)]))
summary(glm(formula = Occurence ~ cs + reves + habi + Atyph +
              agecat + nbpers + Bauto + Pol1 + Pol2 +
              Pol3, family = binomial(link = "probit"),
            data = App[, -c(1:3,6,9,12,15,17:28)]))
summary(glm(formula = Occurence ~ reves + habi + Atyph +
              agecat + nbpers + Bauto + Pol1 + Pol2 +
              Pol3, family = binomial(link = "probit"),
            data = App[, -c(1:3,6,9,12,15,17:28)]))


OccProbitSign = glm(formula = Occurence ~ reves + habi + Atyph + agecat
                    + nbpers + Bauto + Pol1 + Pol2 + Pol3,
                    family = binomial(link = "probit"),
                    data = App[, -c(1:3,6,9,12,15,17:28)])

#Comparatif des erreurs des trois mod les
epd(OccProbit, App, 27) #24,76%
epd(OccProbitAIC, App, 27) #24,72%
epd(OccProbitSign, App, 27) #24,89%

#Marginales
#install.packages("mfx")
library(mfx)
probitmfx(formula = Occurence ~ habi + agecat + nbpers + Bauto
          + Pol1 + Pol2 + Pol3,
          data = App[, -c(1:3,6,9,12,15,17:28)])


#Random Forest
#install.packages("randomForest")
library(randomForest)
set.seed(54)
RF = randomForest(Occurence~., data = App[, -c(1:3,6,9,12,15,17:28)],
                  xtest = Test[, -c(1:3,6,9,12,15,17:28)], ytest = Test[, 28], do.trace =100,
                  importance = TRUE, ntree=500)

#Random Forest du modele AIC
set.seed(54)
RFAIC=randomForest(Occurence ~.,data =
                     App[, c(3,4,5,7,8,10,11,13,14,29,30)],
                   xtest = Test[, c(3,4,5,7,8,10,11,13,14,29,30)],
                   ytest = Test[, 28],   do.trace =100,
                   importance = TRUE, ntree=500)
```

```
#Random Forest du modele avec variables significatives 95%
set.seed(54)
RFAICSign=randomForest(Occurence ~.,data =
                         App[,c(4,7,9,10,12,15,27:30)],
                       xtest = Test[,c(4,7,9,10,12,15,28:30)],
                       ytest = Test[,27], do.trace =100,
                       importance = TRUE, ntree=500)

#-----------------------------
#Nombre de sinistres
#-----------------------------

library (MASS)
summary(App$NSin)
var(App$NSin)

#MODELE DE POISSON
mod0 = glm(NSin~1, family="poisson", data=App) #modele0

stepAIC(mod0,NSin ~ pcs + cs + reves + crevpp + region + habi + Atyph
        + agecat + nbpers + enfants + Anat + Bauto + Nbadulte,
        data=App, trace=TRUE, direction = "both")

stepAIC(glm(NSin~., data=App[,-c(1,8,11,17:25,27,28)], family=poisson))
poisson.opt = glm(formula = NSin ~ pcs + cs + reves + crevpp + region + habi +
              Atyph + agecat + nbpers + Bauto + Nbadulte + Pol1 + Pol2 +
              Pol3, family = poisson,
            data=App[,-c(1,8,11,17:25,27,28)])
summary(poisson.opt)
AIC(poisson.opt)
sqrtmse(poisson.opt, Test, 26, "response")      #3,05

summary(glm(formula = NSin ~ pcs + cs + reves + crevpp + region +
        habi + Atyph + agecat + nbpers + Bauto + Pol1 + Pol2 +
              Pol3, family = poisson,
                   data=App[,-c(1,8,11,17:25,27,28)]))
summary(glm(formula = NSin ~ pcs + reves + crevpp + region +
              habi + Atyph + agecat + nbpers + Bauto + Pol1 + Pol2 +
              Pol3, family = poisson,
            data=App[,-c(1,8,11,17:25,27,28)]))
summary(glm(formula = NSin ~ pcs + reves + crevpp + region +
              habi + Atyph + agecat + nbpers + Pol1 + Pol2 +
              Pol3, family = poisson,
            data=App[,-c(1,8,11,17:25,27,28)]))
```

```
poisson.sign = glm(formula = NSin ~ pcs + reves + crevpp + region +
                     habi + Atyph + agecat + nbpers + Pol1 + Pol2 +
                     Pol3, family = poisson,
                   data = App[, -c(1,8,11,17:25,27,28)])

sqrtmse(poisson.sign, Test, 26, "response")      #3,06

#MODELE BINOMIAL NEGATIF
mod0nb = glm.nb(NSin ~ 1, data = App)
stepAIC(glm.nb(NSin~., data = App[, -c(1,8,11,17:25,27,28)]),
        trace=TRUE, direction="both")

negBinom.opt = glm.nb(NSin ~ pcs + crevpp + region + Atyph + agecat +
        nbpers + enfants + Pol1 + Pol2 + Pol3,
                      data = App[, -c(1,8,11,17:25,27,28)],
        init.theta = 2.258693858, link = log)
summary(negBinom.opt)
AIC(negBinom.opt)
sqrtmse(negBinom.opt, Test, 26, "response")      #3,03

summary( glm.nb(NSin ~ pcs + crevpp + region + Atyph + agecat +
                  nbpers + Pol1 + Pol2 + Pol3,
                data = App[, -c(1,8,11,17:25,27,28)],
                init.theta = 2.258693858, link = log))
negBinom.sign = glm.nb(NSin ~ pcs + crevpp + region + Atyph + agecat +
                         nbpers + Pol1 + Pol2 + Pol3,
                       data = App[, -c(1,8,11,17:25,27,28)],
                       init.theta = 2.258693858, link = log)
sqrtmse(negBinom.sign, Test, 26, "response")      #3,02


#Modélisationducotsdessinistres
#=====================================
```

```
#=====================================
# Tobit avec Sinistre2
#=====================================

library(AER)
summary(tobit(Sinistre2 ~ Pol1+Pol2+Pol3+pcs+RUC+cs+reves+crevpp+region
              +habi+Atyph+agecat+enfants+Anat+Bauto+Nbadulte
              +Police1+Police2+Police3+NSin
              , data=App))
tobit2 = tobit(Sinistre2~Pol1+pcs+RUC+reves+crevpp+region+habi+agecat+
                 enfants+Anat+Nbadulte+NSin, data=App)
summary(tobit2)
sqrtmse(tobit2, Test, 18) #3,40

#=====================================
# Tobit generalise avec Sinistre2
#=====================================

library(mhurdle)

#Predictions Police2 par les autres variables
Mt2= mhurdle(Police2~RUC+region+Acompm
             |RUC+pcs+agecat+region+Atyph+Bauto+Acompm+Pol1+Pol3
             |0, dist="n", data=basemodel)
summary(Mt2)
P2pred = (1-Mt2$fitted.values[,1])*Mt2$fitted.values[,2]
head(cbind(pred=P2pred, obs=basemodel$Police2))


#Predictions Police1 par les autres variables
Mt1= mhurdle(Police1~agecat+Acompm
```

```
              |agecat+region+Acompm
              +NSin
              |0, dist="n", data=basemodel)
summary(Mt1)
P1pred = (1-Mt1$fitted.values[,1])*Mt1$fitted.values[,2]
head(cbind(pred=P1pred, obs=basemodel$Police1))


#Predictions Police3 par les autres variables
Mt3= mhurdle(Police3~agecat+Ahabi
              |Nbadulte+region+Atyph+Acompm+Bauto+Pol1
              |0, dist="n", data=basemodel)
summary(Mt3)
P3pred = (1-Mt3$fitted.values[,1])*Mt3$fitted.values[,2]

MTcorr2=mhurdle(Sinistre2~RUC+pcs+agecat+Ahabi+Atyph+Nbadulte
                |P1pred+P2pred+P3pred+Ahabi, data=basemodel)
summary(MTcorr2)
head(predict(MTcorr2,Test2))


MTcorr3=mhurdle(Sinistre3~RUC+pcs+agecat+Ahabi+Atyph+Nbadulte
                |P1pred+P2pred+P3pred+Ahabi)
summary(MTcorr3)

# avec correlation

MTcorr2=mhurdle(Sinistre2~RUC+pcs+agecat+Ahabi+Atyph+Nbadulte
                |P1pred+P2pred+P3pred+Ahabi, data=basemodel, h2=TRUE, dist="n",
summary(MTcorr2)


MTcorr3=mhurdle(Sinistre3~RUC+pcs+agecat+Ahabi+Atyph+Nbadulte
                |P1pred+P2pred+P3pred+Ahabi, h2=TRUE, dist="n", corr=TRUE)
summary(MTcorr3)

mh21=mhurdle(Sinistre2~region + pcs + Nbadulte |region  + reves+
               Police1|0,dist="n",data=basemodel)

summary(mh21)
AIC(mh21)
p=predict(mh21,Test)
sqrt(mean(((1-p[,1])*p[,2]-Test[,18])**2))

#=====================================
```

```
#Double hurdle avec Sinistre 3
#=====================================

dh3= mhurdle ( Sinistre3 ~ Acompm+agecat+Anat+Bauto+region | pcs + cs
+
                  +region+habi + Atyph+Acompm+Anat  + Bauto +
                  Nbadulte+Police3 + Police2+Police1
                , h2 = TRUE, corr=TRUE, dist = "n", data = App)
summary(dh3)
p=predict(dh3,Test)
sqrt(mean(((1-p[,1])*p[,2]-Test[,19])**2))
```

```
#==================================
#Tarification
#==================================

cbind(pred = round(predict(OccProbitSign,basetarif,type='response')),
      vrai = as.numeric(basetarif[,27])-1)
cbind(pred =round(predict(negBinom.opt,basetarif, type='response')),
      vrai = basetarif[,25])
cbind(predS3=predict(gamma3id.sign,basetarif),
      vraiS3=basetarif[,19])
cbind(pred=abs(predict(cx.app,basetarif)),
      vrai=basetarif[,23])
```