UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÉLISATION DE LA RÉSERVE AVEC MODÈLES LINÉAIRES GÉNÉRALISÉS PONDÉRÉS PAR LA RÉCIPROQUE DE LA PROBABILITÉ DE CENSURE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ALEXANDRE LEBLANC

JUIN 2023

REMERCIEMENTS

J'aimerais remercier mon directeur de recherche M. Mathieu Pigeon pour ses conseils pendant mes recherches. J'aimerais également remercier la Chaire Co-operators en analyse des risques actuariels pour son soutien financier lors de mes études supérieures ainsi que la compagnie d'assurance générale Co-operators pour l'opportunité de stage en recherche et innovation lors de la pandémie et qui a fourni les données réelles utilisées pour mon projet. Enfin, mais pas le moindre, je voudrais remercier ma famille qui m'a soutenu tout au long de mes études.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Le provisionnement est une des fonctions principales d'un assureur et la réserve qui en résulte constitue une dette importante pour celui-ci. Afin d'assurer la solvabilité de la compagnie, l'assureur emploie des actuaires pour calculer cette réserve. Selon la granularité des données, les méthodes de calcul de la réserve se divisent en deux approches : les modèles collectifs et les modèles individuels. Les modèles individuels sont rarement utilisés et de manière générale ne tiennent pas compte du mécanisme de censure observé parmi les réclamations ouvertes au moment du calcul de la réserve. Dans une série d'articles (Lopez *et al.*, 2016), (Lopez *et al.*, 2019) et (Lopez et Milhaud, 2021), des modèles individuels qui utilisent des arbres de régression sont proposés afin de prédire la réserve individuelle pour chaque réclamation ouverte. Des modèles linéaires généralisés pondérés qui prennent en compte le mécanisme de censure sont proposés dans ce mémoire. Les fonctions de vraisemblance sont calculées afin d'assurer des propriétés asymptotiques désirables. Les modèles collectifs et les modèles individuels proposés sont analysés avec des données réelles fournies par un grand assureur canadien.

Mots-clés — Réserves individuelles, Réserve RBNS, Modèles linéaires généralisés, Modèles linéaires généralisés pondérés, Arbres de régression, Arbres de régression pondérés, IPCW, Kaplan-Meier, Modèles individuels, Assurance IARD

INTRODUCTION

Un actuaire travaillant pour un assureur IARD (incendie, accidents et risques divers) a le mandat financier d'établir une *réserve* selon les régulations portant sur la suffisance du capital. La réserve est calculée afin de déterminer un montant pour couvrir l'impact des paiements futurs relatifs à une *date d'évaluation*. Plus précisément, la réserve représente une charge qui comprend plusieurs éléments, dont certains sont fixes et certains qui doivent être estimés. Par exemple, les dépenses pour des ajustements aux réclamations sont fixes et les provisions nécessaires à la date d'évaluation pour absorber les paiements futurs sont inconnues.

Un des objectifs poursuivis par la science actuarielle est de mieux capturer le développement d'une réclamation et de construire un modèle ayant des propriétés statistiques intéressantes. En parallèle, les organismes réglementaires tentent de mieux refléter divers risques. Par exemple, le BSIF (bureau du surintendant des institutions financières du Canada) requiert que les assureurs IARD complètent une évaluation interne des risques et de la solvabilité qui peut comprendre des analyses de différents scénarios financiers et de la dépendance entre certains risques ou certaines lignes d'affaires.

Par le fait que l'actuaire calcule la réserve en fonction de la date d'évaluation, les paiements futurs doivent être prédits. À cause de cette censure à droite du temps, la réserve présente un biais négatif, c'est-à-dire que la réserve individuelle ultime est sous-estimée. Ce mémoire propose un modèle *individuel* de régression linéaire généralisé qui corrige la censure de la durée de la réclamation. Ce dernier sera comparé au modèle proposé par (Lopez *et al.*, 2016) qui corrige le biais de la valeur estimée de la réserve individuelle par un arbre de régression. En particulier, la réserve prédite par ce travail sera celle issue des réclamations déclarées à la date de calcul de la réserve et non celles qui seront déclarées après la date de calcul de la réserve.

Pour les modèles linéaires généralisés et les arbres de régression, la réciproque de la probabilité de censure de la durée sert de poids lors de l'algorithme d'estimation de chaque modèle. Ces poids seront dorénavant appelés par l'abréviation IPCW (qui réfère à leur nom en anglais *inverse probability censoring weights*). Selon (Vock *et al.*, 2016), ces poids ont été employés pour plusieurs modèles statistiques afin d'expliquer la durée jusqu'à un évènement (*time-to-event data*) tout en permettant de conserver la propriété d'absence de biais et la convergence des estimateurs des paramètres.

Au premier chapitre, la réserve sera définie en générale ainsi que la réserve modélisée au présent mémoire. De plus, le processus de développement d'une réclamation en assurance IARD sera présenté en détail.

Au deuxième chapitre, suivant la description de la réserve visée, une revue des modèles collectifs sera présentée. En particulier, le modèle non paramétrique de Mack (Mack, 1993) et le rééchantillonnage de celui-ci sera décrit. Un autre modèle collectif d'importance historique, le modèle linéaire généralisé avec variable réponse appartenant à la famille d'*exponential dispersion* (ED) sera également décrit. Un exemple illustratif sera fourni pour le modèle de Mack et pour le modèle linéaire généralisé.

Au troisième chapitre, un historique des modèles individuels sera présenté afin de justifier l'inclusion des variables explicatives relatives aux dossiers des assurés. Il y aura une introduction aux notions nécessaires de l'apprentissage statistique et aux mécanismes de censure. Il y aura une introduction au calcul des poids IPCW et aux arbres de régression pondérés. L'algorithme wCART qui permet l'ajustement des arbres de régression pondérés sera décrit et illustré par un exemple.

Au quatrième chapitre, les stratégies de prédiction des arbres de régression et l'élagage de ceux-ci seront introduits. Des exemples illustratifs des arbres de régression pondérés permettront d'éclairer les stratégies de prédiction de la réserve individuelle. De plus, une section sera consacrée à la construction de la base de données individuelles.

Au cinquième chapitre, la contribution principale de ce mémoire sera développée. Les

propriétés de l'estimateur du maximum de vraisemblance seront énumérées. La fonction de vraisemblance en présence de données censurées sera développée. Il sera démontré que l'estimateur calculé en maximisant la fonction de vraisemblance pondérée par les IPCW respecte les mêmes propriétés que l'estimateur du maximum de vraisemblance. Deux modèles linéaires généralisés pondérés par les IPCW comparables aux arbres de régression seront proposés. Également, des exemples des GLM pondérés par les IPCW permettront d'illustrer la méthodologie proposée.

Finalement, au dernier chapitre, une analyse numérique complète sera présentée. Elle sera basée sur un jeu de données réelles fourni par un grand assureur canadien. Elle se focalisera sur la modélisation de la réserve globale, autant pour les modèles collectifs que pour les modèles individuels. On s'intéressera surtout au comportement de la distribution de la réserve modélisée et les moments de cette dernière ainsi qu'au comportement longitudinal des modèles individuels sur base annuelle. À partir des constats à propos de l'adéquation des modèles ci-haut, des améliorations futures seront proposées.

CHAPITRE I

LE PROVISIONNEMENT EN ASSURANCE IARD

1.1 Introduction

Une police d'assurance est un contrat entre un assureur et un assuré. Moyennant une prime versée à l'assureur, l'assuré peut recevoir un montant à la suite d'une réclamation, dont ledit montant peut être divisé en plusieurs paiements en fonction du développement de la réclamation ainsi que de l'administration de la réclamation par l'assureur et il peut y avoir un délai entre la survenance d'un sinistre et la déclaration de celui-ci à l'assureur. L'événement causant la perte peut prendre plusieurs mois avant d'engendrer des dommages. En assurance automobile, plusieurs assureurs offrent un service de déclaration automatique ou encore la police peut inclure une clause limitant la durée après laquelle l'assuré peut soumettre une réclamation. De plus, il y a un délai entre la déclaration d'un sinistre et la fermeture du dossier qui dépend de la complexité de la réclamation et des processus administratifs de l'assureur. Enfin, il est possible que le dossier lié à une réclamation se réouvre : la date de réouverture pourrait alors être considérée comme une deuxième date de déclaration.

Formellement, pour une réclamation avec identifiant k, on note $T_1^{(k)}$ la date de survenance du sinistre, $T_2^{(k)}$ la date de la déclaration du sinistre à l'assureur du sinistre survenu au temps $T_1^{(k)}$ et $T_4^{(k)}$ la date de fermeture. Pour un nombre entier de paiements M, on note $P_m^{(k)} \in \mathbb{R}$, $m = 1, \ldots, M$, le montant de chacun des paiements versés et $T_{3,m}^{(k)}$ la date du m^{e} paiement. Il est à noter que les montants $P_m^{(k)}$ ne sont pas forcément positifs. La

réouverture du dossier d'une réclamation peut survenir si un deuxième sinistre découlant du premier se manifeste dont la date de cette dernière est notée par $T_1^{(k')}$. Un schéma du processus de développement d'une réclamation est illustré à la Figure 1.1.

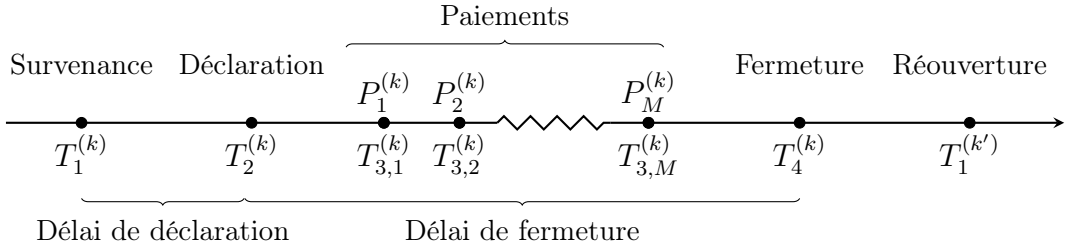


Figure 1.1: Développement d'une réclamation quelconque avec identifiant k.

La réserve actuarielle étant à la base une notion comptable, il est important de traduire celle-ci en une notion statistique. La définition comptable définie par (Sharara *et al.*, 2010) est la charge composée de la somme des charges liées aux dépenses (*allocated loss adjustment expenses*, ou ALAE) et la somme des charges liées aux réclamations. Les frais peuvent inclure des frais de justice, des frais médicaux, des frais d'ingénierie et d'autres frais payés par l'assureur. La provision peut être divisée en deux morceaux : les *case reserves* (les montants réservés par l'assureur afin d'indemniser les assurés), qui peuvent être parfois appelées les provisions projetées relativement à la date d'évaluation pour réclamations déclarées et ouvertes (*reported but not settled expenses*, ou RBNS) et les provisions projetées relativement à la date d'évaluation pour sinistres survenus, mais non déclarés (*incurred but not reported expenses*, ou IBNR).

Il y a plusieurs subtilités de la définition ci-dessus qui seront ignorées pour le calcul de la réserve dans le présent mémoire. Par exemple, les charges liées aux polices doivent être actualisées au taux d'intérêt du portefeuille des actifs de l'assureur. La prime non acquise contribue aussi à la réserve puisqu'elle est une dette créée à l'évaluation de la réserve si les primes sont payées par l'assuré à l'avance. Il y a également certaines conséquences comptables qui découlent de la prédiction de la réserve RBNS. Au fil du temps, si la réserve augmente, les impôts seront réduits en conséquence puisque l'augmentation en

question représente une dépense.

L'objectif principal de ce mémoire est de développer une méthodologie d'imputation de la réserve finale pour chaque réclamation ouverte à la date d'évaluation de la réserve. En conséquence, l'accent sera mis sur l'estimation de la réserve RBNS et non sur les notions comptables mentionnées, ni sur l'estimation de la réserve IBNR.

1.2 La réserve en temps continu

Soit t_e la date d'évaluation mesurée en périodes d'une certaine durée (jours, trimestres, années, etc.) relativement à une date d'origine ad hoc t_o, la première date d'accident observée au portefeuille.

Alors, pour une réclamation avec identifiant k, si

1. $T_1^{(k)} < t_e < T_2^{(k)}$, un sinistre est de type IBNR puisqu'il est survenu après la date d'évaluation, mais n'a pas été déclaré à l'assureur avant la date d'évaluation. L'assureur possède des informations au sujet de l'assuré, mais n'a aucune information relative à la réclamation ;
2. $T_2^{(k)} < t_e < T_4^{(k)}$, un sinistre est de type RBNS puisqu'il est survenu et a été déclaré à l'assureur avant la date d'évaluation. L'assureur possède des informations au sujet de l'assuré et des informations recueillies au sujet de la réclamation k avant la date d'évaluation. Le dossier de la réclamation k est dit *ouvert* ;
3. $t_e > T_4^{(k)}$, le dossier de la réclamation k est dit *fermé*. Pour le présent mémoire, on exclut la possibilité qu'une réclamation puisse réouvrir. Alors, pour toute réclamation fermée avant la date d'évaluation, on suppose que la totalité de l'information relevant de cette dernière est à la disposition de l'assureur.

Puisque seulement la réserve RBNS sera d'intérêt, le montant cumulatif payé par l'assureur

pour la réclamation k au temps t est défini comme

$$C_t^{(k)} = \mathbb{1}(T_2^{(k)} \leq t) \sum_{\{m \in \mathbb{N}: T_{3,m}^{(k)} \leq t\}} P_m^{(k)},$$

où

$$\mathbb{1}(T_2^{(k)} \leq t) = \begin{cases} 1, & \text{la réclamation } k \text{ est déclarée à l'assureur avant } t, \\ 0, & \text{sinon.} \end{cases}$$

Naturellement, on s'intéresse à modéliser la réserve qui sera connue seulement quand $t = T_4$. Alors, le montant total payé (*total amount paid*) pour la réclamation k est $C_{T_4}^{(k)} = C_{T_4^{(k)}}^{(k)}$ et la réserve pour la réclamation k est

$$\widehat{R}_{t_e}^{(k)} = \widehat{C}_{T_4}^{(k)} - C_{t_e}^{(k)}, \tag{1.1}$$

où $\widehat{C}_{T_4}^{(k)}$ une projection réalisée à l'aide des informations disponibles à la date d'évaluation.

La réserve totale est définie alors comme

$$\begin{aligned} \widehat{R}_{t_e} &= \sum_{k=1}^{N} \widehat{R}_{t_e}^{(k)} \\ &= \sum_{k=1}^{N} \widehat{R}_{t_e}^{(k)} \{\mathbb{1}(T_1^{(k)} < t_e \leq T_2^{(k)}) + \mathbb{1}(T_2^{(k)} < t_e < T_4^{(k)})\} \\ &= \sum_{k=1}^{N} \widehat{R}_{t_e}^{(k)} \mathbb{1}(T_1^{(k)} < t_e \leq T_2^{(k)}) + \sum_{k=1}^{N} \widehat{R}_{t_e}^{(k)} \mathbb{1}(T_2^{(k)} < t_e < T_4^{(k)}) \\ &= \widehat{R}_{t_e}^{\text{IBNR}} + \widehat{R}_{t_e}^{\text{RBNS}}, \end{aligned} \tag{1.2}$$

où N est le nombre total de réclamations au portefeuille. On note que pour $t_e \geq T_4^{(k)}$ ou $t_e < T_1^{(k)}$, $\widehat{R}_{t_e}^{(k)} = 0$. En particulier, le premier terme de l'Équation (1.2) (la réserve IBNR) sera $C_{t_e}^{(k)} = 0$ pour tout $k = 1, 2, \ldots, N$.

1.3 La réserve en temps discret

Au sein de cette section, toutes durées définies dans la section précédente seront substituées par des durées discrètes où N (muni d'indices et d'exposants) réfère à la durée T (muni

d'indices et d'exposants) mesurée en unités discrètes de temps. Par exemple, la date d'évaluation mesurée en temps continu est t_e et est n_e en temps discret, $N_4^{(k)}$ est la durée à la fermeture de la k^{e} réclamation en temps discret tandis que $T_4^{(k)}$ est la durée de fermeture en temps continu, etc.

Au prochain chapitre, les modèles collectifs agrègent les paiements par période de survenance de sinistre et par période de développement. On peut maintenant développer l'analogue discret de l'Équation (1.2).

Soit $i = 1, \ldots, I$ et $j = 1, \ldots, J$ la période de survenance et la période de développement, respectivement. Pour une réclamation k avec période de survenance i, $i - 1 < T_1^{(k)} \leq i$ et les paiements versés après j périodes de développement doivent respecter $j - 1 < T_{3,m}^{(k)} - T_1^{(k)} \leq j$.

Soit le montant payé après j périodes de développement pour la réclamation k ayant une période de survenance i, on a alors

$$Y_{i,j}^{(k)} = \sum_{m \in \mathcal{M}_{i,j}^{(k)}} P_m^{(k)}, \tag{1.3}$$

où

$$\mathcal{M}_{i,j}^{(k)} = \{m \in \mathbb{N} : (i,j) \in \{1, \ldots, I\} \times \{1, \ldots, J\}, T_1^{(k)} \in (i-1, i], T_{3,m}^{(k)} - T_1^{(k)} \in (j-1, j]\}$$

est le nombre de paiements dans le développement de la réclamation k.

Il s'ensuit que le montant total payé pour sinistres survenus à la période de survenance i après j périodes de développement est

$$Y_{i,j} = \sum_{k=1}^{N} Y_{i,j}^{(k)}. \tag{1.4}$$

En temps discret, on peut définir l'analogue de l'Équation (1.1). En effet, si l'on pose $C_{i,j}^{(k)} = \sum_{s=1}^{j} Y_{i,s}^{(k)}$ et $C_{i,j} = \sum_{s=1}^{j} Y_{i,s}$,

$$\widehat{R}_{n_e}^{(k)} = \sum_{i=1}^{I} \widehat{C}_{i,J}^{(k)} - \sum_{(i,j) \in \Delta} C_{i,j}^{(k)}$$

où $\Delta = \{(i,j) \in \mathbb{N}^2 : i + j - 1 \leq n_e\}$.

Finalement, la réserve totale, évaluée à la période n_e, est

$$\begin{aligned}
\widehat{R}_{n_e} &= \sum_{k=1}^{N} \widehat{R}_{n_e}^{(k)} \\
&= \sum_{k=1}^{N} \widehat{R}_{n_e}^{(k)} \{\mathbb{1}(N_1^{(k)} < n_e < N_2^{(k)}) + \mathbb{1}(N_2^{(k)} \leq n_e < N_4^{(k)})\} \\
&= \sum_{k=1}^{N} \widehat{R}_{n_e}^{(k)} \mathbb{1}(N_1^{(k)} < n_e < N_2^{(k)}) + \sum_{k=1}^{N} \widehat{R}_{n_e}^{(k)} \mathbb{1}(N_2^{(k)} \leq n_e < N_4^{(k)}) \\
&= \widehat{R}_{n_e}^{\text{IBNR}} + \widehat{R}_{n_e}^{\text{RBNS}}.
\end{aligned}$$

Dans le cadre des modèles collectifs, les paiements avec symboles C sont dits *cumulatifs* et les paiements avec symboles Y sont dits *incrémentaux*. On peut également définir la réserve en fonction des paiements incrémentaux. Utilisant l'Équation (1.3), on a

$$\widehat{R}_{n_e}^{(k)} = \sum_{(i,j) \in \overline{\Delta}} \widehat{Y}_{i,j}^{(k)}$$

et

$$\widehat{R}_{n_e} = \sum_{(i,j) \in \overline{\Delta}} \widehat{Y}_{i,j},$$

où $\overline{\Delta} = \{1, \ldots, I\} \times \{1, \ldots, J\} \setminus \Delta$.

Les prochains chapitres traitent les modèles collectifs et les modèles individuels de provisionnement. Les modèles collectifs visent à modéliser la réserve $\widehat{R}_{n_e}$ en temps discret (typiquement sur base annuelle) en agrégeant les informations individuelles de chaque réclamation. Pour cette raison, les modèles collectifs peuvent être surparamétrés, mais certains modèles collectifs proposés dans les dernières années (p. ex. le modèle de (Wahl *et al.*, 2019)) proposent une modélisation individuelle du processus de paiements ainsi que le statut (ouvert ou fermé) de chaque réclamation et d'agréger les prédictions de ces derniers afin d'obtenir une prédiction de la réserve RBNS pour une période future. La modélisation individuelle des réclamations permet une modélisation plus flexible puisque chaque réclamation ouverte peut avoir son propre modèle et de capturer la

variabilité induite par les paiements pertinents ou par des dépendances entre des données individuelles (p. ex. la dépendance entre la durée à la fermeture et le montant total à payer) de la réclamation en question. Par contre, avec des données longitudinales, les modèles individuels demandent des ressources informatiques plus performantes que les modèles collectifs qui ne conservent pas des données agrégées par période d'accident et par période de développement et sont normalement plus complexes à mettre à jour. Les modèles proposés dans ce mémoire permettent de mettre à jour les réserves prédites à l'aide d'un seul modèle ou de mettre à jour la réserve d'une seule réclamation qui possède son propre modèle.

CHAPITRE II

MODÈLES COLLECTIFS DE PROVISIONNEMENT EN ASSURANCE IARD

2.1 Introduction

La littérature académique des modèles collectifs est en évolution depuis plus d'un demi-siècle. Les principaux modèles collectifs sont des dérivés de la technique *Chain Ladder* (CL) (Mack, 1993) et des *modèles linéaires généralisés* (GLM) développés par (McCullagh et Nelder, 1989). Ces modèles demeurent les plus souvent utilisés en pratique et plusieurs modifications ont été proposées afin de faciliter leurs mises en oeuvre. Un survol des propriétés des modèles CL et des GLM sera présenté, mais les variantes ne seront pas abordées. Une étude approfondie des modèles collectifs est faite dans (Wüthrich et Merz, 2008).

Ces modèles permettent une estimation des deux premiers moments (espérance et variance) de la distribution de la réserve. Souvent, on s'intéresse à des moments plus élevés ou encore à des quantiles qui permettent d'évaluer la forme de la distribution et de calculer la probabilité des montants extrêmes. À ces fins, le *bootstrap* (Efron et Tibshirani, 1993) est une technique de rééchantillonage qui permet d'estimer la distribution d'une statistique à partir d'un seul échantillon. Cette technique s'applique au présent mémoire pour produire la distribution de l'estimateur de l'espérance de la réserve et d'estimer ses quantiles.

2.1.1 Triangles de développement et notation

Les montants incrémentaux et collectifs sont souvent illustrés à l'aide de *triangles de développement* (on *run-off triangles* en anglais). Un triangle de développement en scidant en deux une table de dimensions $I \times J$ relativement à la diagonale qui commence dans le coin inférieur gauche. Les triangles de développement qui affiche le développement des réclamations à travers le temps. Les entrées de Les triangles peuvent être de type *cumulatif* si ses entrées sont $C_{i,j}$ ou de type *incrémental* si ses entrées sont $Y_{i,j}$.

		Année de développement			
		1	$\cdots$		J
Année d'accident	1	$Y_{1,1}$	$Y_{1,2}$	$\cdots$	$Y_{1,J}$
	2	$Y_{2,1}$	$Y_{2,2}$	$\cdots$	$Y_{2,J}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	I	$Y_{I,1}$	$Y_{I,2}$	$\cdots$	$Y_{I,J}$

Tableau 2.1: Exemple de triangle de développement incrémental

À titre d'exemple, le Tableau 2.1 illustre un triangle de développement incrémental avec I années d'accident et J années de développement où les paiements enregistrés avant le calcul de la réservé se retrouvent à gauche de l'escalier. Pour obtenir le triangle de paiements cumulatifs, il suffit de remplacer $Y_{i,j}$ par $C_{i,j} = \sum_{k=1}^{j} Y_{i,k}$.

2.2 Méthode Chain Ladder

Les origines de la méthode Chain Ladder (CL) ne sont pas bien documentées, mais la méthode CL est une heuristique utilisée en provisionnement depuis au moins un demi-siècle. La méthode Chain Ladder est simple à comprendre et à mettre en oeuvre, mais son estimation de la réserve est déterministe. Pour les modèles collectifs, une hypothèse

classique est de forcer $n_e = I = J$ (triangle de développement « carré »). La méthode CL impose deux hypothèses sur les entrées $C_{i,j}$:

1. les vecteurs $\{C_{i,j}\}_{j=1}^{I}$ et $\{C_{i',j}\}_{j=1}^{I}$ sont indépendants pour $i \neq i'$;
2. les paiements cumulatifs pour la période de survenance i et de la période de développement $j-1$ à j sont liés par un facteur de développement λ_{j-1} tel que $C_{i,j} = \lambda_{j-1} C_{i,j-1},\ i = 1, \ldots, I,\ j = 2, \ldots, I.$

L'estimateur CL du facteur de développement λ_j est

$$\widehat{\lambda}_j = \frac{\sum_{i=1}^{I-j-1} C_{i,j+1}}{\sum_{i=1}^{I-j-1} C_{i,j}}, \quad j = 1, \ldots, I-1. \tag{2.1}$$

Le montant cumulatif projeté à la période de développement I pour une réclamation survenue à la période d'accident i est

$$\widehat{C}_{i,I} = \left[\prod_{j=I-i+1}^{I-1} \widehat{\lambda}_j \right] C_{i,I-i+1}, \quad i = 2, \ldots, I. \tag{2.2}$$

Il s'ensuit que la réserve évaluée à la période de développement $I-i+1$ pour les sinistres survenus à la période i est

$$\widehat{R}_{i,I} = \widehat{C}_{i,I} - C_{i,I-i+1}$$

et que la réserve totale calculée est

$$\widehat{R}_I = \sum_{i=2}^{I} \widehat{R}_{i,I}.$$

Puisque la méthode CL est déterministe, elle n'admet aucune inférence statistique ni aucune évaluation de la distribution prédictive de la réserve. Le modèle de Mack (Mack, 1993) est une version stochastique du modèle CL où les $C_{i,j}$ sont des variables aléatoires avec des hypothèses sur l'espérance conditionnelle et sur la variance conditionnelle des paiements cumulatifs :

1. les vecteurs aléatoires $(C_{i,1}, \ldots, C_{i,I})$ et $(C_{i',1}, \ldots, C_{i',I})$ sont indépendants pour chaque paire d'entiers $(i, i') \in \{1, \ldots, I\}^2$ où $i \neq i'$;
2. $E\left[C_{i,j} \mid C_{i,1}, \ldots, C_{i,j-1}\right] = \lambda_{j-1} C_{i,j-1}$, $1 \leq i \leq I$ et $2 \leq j \leq I$;
3. $\mathrm{Var}\left[C_{i,j} \mid C_{i,1}, \ldots, C_{i,j-1}\right] = \sigma_{j-1}^2 C_{i,j-1}$, $1 \leq i \leq I$ et $2 \leq j \leq I$.

Avec les hypothèses ci-dessus, il est simple de démontrer que les estimateurs $\widehat{\lambda}_j$ du modèle de Mack sont sans biais et non corrélés. En effet,

$$\begin{aligned}
E\left[\widehat{\lambda}_j\right] &= E\left[E\left[\widehat{\lambda}_j \mid \mathcal{B}_j\right]\right], \quad \mathcal{B}_j = \{C_{i,k} : i + j - 1 \leq I, k \leq j\} \\
&= E\left[\frac{\sum_{i=1}^{I-j-1} E[C_{i,j+1} \mid \mathcal{B}_j]}{\sum_{i=1}^{I-j-1} C_{i,j}}\right] \\
&= E\left[\frac{\sum_{i=1}^{I-j-1} \lambda_j C_{i,j}}{\sum_{i=1}^{I-j-1} C_{i,j}}\right] \\
&= E\left[\lambda_j\right] \\
&= \lambda_j
\end{aligned}$$

et

$$\begin{aligned}
E\left[\prod_{j=1}^{I-1} \widehat{\lambda}_j\right] &= E\left[E\left[\prod_{j=1}^{I-1} \widehat{\lambda}_j \mid \mathcal{B}_{I-1}\right]\right] \\
&= E\left[\prod_{j=1}^{I-2} \widehat{\lambda}_j E\left[\widehat{\lambda}_{I-1} \mid \mathcal{B}_{I-1}\right]\right] \\
&= E\left[\prod_{j=1}^{I-2} \widehat{\lambda}_j\right] \lambda_{I-1} \\
&= \ldots \\
&= \prod_{j=1}^{I-1} \lambda_j
\end{aligned}$$

si l'on répète la même démarche $I-2$ fois en utilisant le théorème de l'espérance totale successivement sur les ensembles $\mathcal{B}_{I-2}, \ldots, \mathcal{B}_1$. Puisque l'espérance du produit des facteurs de développement est le produit des espérances, la covariance des $\widehat{\lambda}_j$ est nulle. En conséquence, l'estimateur $\widehat{C}_{i,I}$ est sans biais pour tout $i = 1, \ldots, I$.

Il est démontré dans le livre (Wüthrich et Merz, 2008) que l'estimateur

$$\widehat{\sigma}_j^2 = \frac{1}{I-j-1} \sum_{i=1}^{I-j} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - \widehat{\lambda}_j \right)^2, \quad j = 1, \ldots, I-2$$

est également sans biais. S'il y a des réclamations donc les périodes développement dépassent la période I, il y a nécessairement des réclamations ouvertes à la période d'évaluation et il faut extrapoler le montant cumulé en fonction des périodes de survenance et de développement. (Mack, 1993) suppose une décroissance exponentielle de l'estimateur $\widehat{\sigma}_j^2$ pour $j \geq I-1$. Si l'on observe que $\widehat{\sigma}_{I-2}^2 > \widehat{\sigma}_{I-3}^2$, l'extrapolation pour $\widehat{\sigma}_{I-1}^2$ se fait à l'aide de

$$\frac{\widehat{\sigma}_{I-3}^2}{\widehat{\sigma}_{I-2}^2} = \frac{\widehat{\sigma}_{I-2}^2}{\widehat{\sigma}_{I-1}^2}.$$

En général,

$$\widehat{\sigma}_{I-1}^2 = \min \left\{ \frac{\widehat{\sigma}_{I-2}^4}{\widehat{\sigma}_{I-3}^2}, \min\{\widehat{\sigma}_{I-3}^2, \widehat{\sigma}_{I-2}^2\} \right\}.$$

Dans le livre (Wüthrich et Merz, 2008), on dérive l'estimateur de l'erreur quadratique moyenne de l'estimateur (*mean square error*, ou MSE) de la réserve pour les sinistres survenus à la période i où

$$\widehat{\mathrm{MSE}}(\widehat{R}_{i,I}) = \widehat{C}_{i,I} \sum_{k=I+1-i}^{I-1} \frac{\widehat{\sigma}_k^2}{\widehat{\lambda}_k^2} \left(\frac{1}{\widehat{C}_{i,k}} + \frac{1}{\sum_{j=1}^{I-k} C_{j,k}} \right), \quad i = 1, \ldots, I$$

et la variance de l'estimateur du total de la réserve

$$\widehat{\mathrm{MSE}}(\widehat{R}_I) = \sum_{i=2}^{I} \left[\widehat{\mathrm{Var}} \left[\widehat{R}_{i,I} \right] + \widehat{C}_{i,I} \left(\sum_{m=i+1}^{I} C_{m,I} \right) \sum_{k=I+1-i}^{I-1} \frac{2\widehat{\sigma}_k^2 / \widehat{\lambda}_k^2}{\sum_{m=1}^{I-k} C_{m,k}} \right],$$

où $\widehat{R}_I = \sum_{i=2}^{I} \widehat{R}_{i,I}$.

Enfin, on note que le modèle de Mack présente plusieurs faiblesses dont les principales sont :

— pour chaque période de survenance, le développement des paiements cumulatifs est complètement déterminé par les facteurs $\widehat{\lambda}_j,\ j = 1, \ldots, I-1$;
— les changements à la jurisprudence, aux processus de souscription et à la gestion de risque ne sont pas reflétés par l'heuristique CL ;
— l'estimateur $\widehat{R}_{i,I}$ est de plus en plus variable lorsque i approche I ;
— l'inférence statistique de l'adéquation des facteurs de développement est impossible.

2.3 Modèles linéaires généralisés

2.3.1 Famille de distributions *exponential dispersion*

La famille de distributions *exponential dispersion* (ED) définie dans l'article (Jørgensen, 1987) est l'ensemble des fonctions de densité de probabilité dont la forme est donnée par

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

pour des fonctions a, b, c et le vecteur de paramètres (θ, ϕ). La fonction c ne peut dépendre de paramètre θ, le support de f_Y ne peut dépendre de (θ, ϕ) et $a(\phi) > 0$ pour tout ϕ.

Dans la famille ED, la variance $\mathrm{Var}[Y]$ est une fonction de l'espérance $E[Y]$:

$$E[Y] = \frac{\partial}{\partial\theta} b(\theta) = \mu$$
$$\mathrm{Var}[Y] = a(\phi)\frac{\partial^2}{\partial\theta^2} b(\theta) = a(\phi)V(\mu)$$

pour une fonction V de l'espérance de Y.

2.3.2 Modèle de régression linéaire généralisé et ses propriétés

Soit g, une fonction bijective et dérivable appelée *fonction de lien*. Pour un vecteur aléatoire $(Y_1, \ldots, Y_n)$ composé de variables aléatoires indépendantes, mais non identiquement distribuées appartenant à la famille ED, les GLM tentent d'estimer l'espérance

conditionnelle $E[Y_i \mid \mathbf{X}_i]$ par une matrice non stochastique $\mathbf{X} = [\mathbf{X}_1 \mid \cdots \mid \mathbf{X}_n]^T$, où $\mathbf{X}_i$ correspond à un vecteur colonne de taille $p+1$ comprenant les informations explicatives relatives à l'observation i. En particulier, le GLM pose

$$g(E[Y_i \mid \mathbf{X}_i]) = \mathbf{X}_i^T \boldsymbol{\beta},$$

pour $\boldsymbol{\beta}$ un vecteur colonne d'inconnus de longueur $p+1$. Alors, $\mu_i = E[Y_i \mid \mathbf{X}_i] = g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta})$.

L'estimateur du vecteur $\boldsymbol{\beta}$ est généralement trouvé par la méthode du *maximum de vraisemblance* (MV). L'estimateur du MV, $\widehat{\boldsymbol{\beta}}$, est convergent si l'espérance conditionnelle est correctement spécifiée relativement à la famille ED. Concrètement, l'estimateur par le MV est obtenu en maximisant la fonction de vraisemblance $\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{X}) = \prod_{i=1}^n f_{Y_i}(y_i; \theta_i, \phi_i)$, où θ_i est une fonction de $\mathbf{X}_i$ et $\boldsymbol{\beta}$.

Puisque la fonction logarithmique est monotone croissante, on peut maximiser la fonction de log-vraisemblance $\ell(\boldsymbol{\beta} \mid \mathbf{X}) = \log\{\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{X})\}$. Donc, pour un vecteur aléatoire $(Y_1, \ldots, Y_n)$, la fonction de log-vraisemblance est

$$\begin{aligned} \ell(\boldsymbol{\beta} \mid \mathbf{X}) &= \sum_{i=1}^n \ell_i(\boldsymbol{\beta} \mid \mathbf{X}_i) \\ &= \sum_{i=1}^n \log\{f_{Y_i}(y_i; \theta_i, \phi_i)\}. \end{aligned} \tag{2.3}$$

Si on calcule le gradient de l'Équation (2.3), on obtient la fonction de *score* $\mathcal{U}$ telle que

$$\begin{aligned} \mathcal{U}(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta} \mid \mathbf{X}) &= \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0}, \cdots, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p} \right) \\ &= \left(\sum_{i=1}^n \frac{(y_i - \mu_i)}{g'(\mu_i)\mathrm{Var}[Y_i]}, \sum_{i=1}^n \frac{(y_i - \mu_i)X_{i,1}}{g'(\mu_i)\mathrm{Var}[Y_i]}, \cdots, \sum_{i=1}^n \frac{(y_i - \mu_i)X_{i,p}}{g'(\mu_i)\mathrm{Var}[Y_i]} \right) \\ &= \mathbf{X}^T \mathbf{D}(\mathbf{y} - \boldsymbol{\mu}), \end{aligned} \tag{2.4}$$

où $\mathbf{D}$ est une matrice diagonale telle que $\mathbf{D}_i = [g'(\mu_i)\mathrm{Var}[Y_i]]^{-1}$.

Posant $\mathcal{U}(\boldsymbol{\beta}) = \mathbf{0}$, on peut obtenir un vecteur $\widehat{\boldsymbol{\beta}}$ qui maximise ℓ selon certaines hypothèses.

Quant à la variance de l'estimateur $\boldsymbol{\beta}$, la matrice d'information Fisher notée par $\mathcal{I}(\boldsymbol{\beta})$ est définie par

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}) &= -E\left[\nabla^2 \ell(\boldsymbol{\beta} \mid \mathbf{X})\right] \\ &= \mathbf{X}^T \mathbf{W} \mathbf{X},\end{aligned}$$

où $\mathbf{W}$ est une matrice diagonale telle que $\mathbf{W}_i = [g'(\mu_i)^2 \mathrm{Var}[Y_i]]^{-1}$.

Si $n \to \infty$, $\mathcal{I}(\boldsymbol{\beta}) = E[\mathcal{U}(\boldsymbol{\beta})^T \mathcal{U}(\boldsymbol{\beta})]$. Par l'inégalité Cramer-Rao, si $\widehat{\boldsymbol{\beta}}$ est sans biais, $\mathrm{Var}[\widehat{\boldsymbol{\beta}}] = \mathcal{I}(\boldsymbol{\beta})^{-1}$.

L'unicité de l'estimateur $\widehat{\boldsymbol{\beta}}$ est déterminée à l'aide de certaines hypothèses selon (Wedderburn, 1976) qui dépendent du choix de la fonction de densité de probabilité de la variable aléatoire réponse.

L'estimateur $\widehat{\boldsymbol{\beta}}$ obtenu par maximum de vraisemblance est un estimateur MVUE (*minimum variance unbiased estimator*) asymptotiquement et suit une loi normale (McCullagh et Nelder, 1989). Ainsi, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1})$ lorsque $n \to \infty$.

Les GLM ont été appliqués dans plusieurs domaines même avant leur définition formelle dans l'ouvrage (McCullagh et Nelder, 1989). En actuariat, un survol des modèles orientés vers la tarification ou le provisionnement (dont plusieurs utilisent des GLM) a été rédigé par (Blier-Wong *et al.*, 2021).

Pour modéliser la réserve, il est typique de considérer une classe de GLM nommée les *modèles log-linéaires*. Ces derniers supposent typiquement que la variable réponse $Y_{i,j}$ est une cellule du triangle de développement incrémental de taille $I \times I$. Par exemple, le modèle log-linéaire (Renshaw et Verrall, 1998) où $Y_{i,j} \sim \text{Poisson}(\mu_{i,j} = \exp\{\beta_0 + \alpha_i + \beta_j\})$ où $\alpha_1 = \beta_1 = 0$ afin que le modèle soit identifiable et qu'il parvienne à la même réserve que la méthode CL. Alors, la prédiction du montant incrémental $Y_{i,j}$ est $\widehat{\mu}_{i,j} = \exp\{\widehat{\beta}_0 + \mathbb{1}(i \geq 1)\widehat{\alpha}_i + \mathbb{1}(j \geq 1)\widehat{\beta}_j\}$, $(i,j) \in \{1, \ldots, I\}^2$.

On peut augmenter le modèle ci-haut en modifiant la variance conditionnelle pour inclure

un facteur multiplicatif ϕ. Par contre, puisqu'on spécifie la forme de la variance par $\text{Var}[Y_{i,j}] = \phi V(E[Y_{i,j}])$, la variable $Y_{i,j}$ est libre de fonction de distribution. Au lieu d'estimer les paramètres par le MV, ils sont estimés par des méthodes de *quasi-vraisemblance*. Les estimateurs du *quasi maximum de vraisemblance* sont toujours convergents et suivent asymptotiquement une loi normale, mais ne sont plus efficaces. La fonction de score donnée par l'Équation (2.4) à optimiser dans le cas de la quasi-vraisemblance reste la même et le paramètre ϕ est souvent estimé par la méthode des moments (MOM). L'estimateur MOM de ϕ est

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}, \tag{2.5}$$

où p est le nombre de paramètres (excluant l'ordonnée à l'origine) du modèle.

Contrairement au modèle de Mack, les GLM offrent plusieurs avantages tels que

- ils permettent l'inclusion de covariables (même celles du dossier) ;
- ils permettent, pour la régression de Poisson, le calcul d'une réserve identique à celle de la méthode CL ;
- ils permettent d'obtenir des intervalles de confiance pour les paramètres ainsi que des intervalles de prédiction ;
- la nature semi-paramétrique des GLM permet plusieurs méthodes de rééchantillonage.

2.3.3 Familles de densités utilisées et choix de la fonction de lien

Selon l'Équation (2.4), la fonction de lien g est importante pour la convergence du GLM, mais également pour son interprétation. Le choix naturel de fonction de lien est dit *canonique* si $\theta = g(\mu)$. Si la fonction de lien canonique est utilisée,

- la covariance échantillonnale entre $\mathbf{y}$ et la j^{e} colonne de la matrice $\mathbf{X}$ (à une somme près) est une statistique exhaustive pour β_j ;
- l'espérance de l'information de Fisher est identique à l'information de Fisher observée.

En assurance IARD, la fonction de canonique est parfois délaissée en faveur de la fonction de lien logarithmique. Le lien logarithmique force l'espérance μ à être d'une forme multiplicative. En effet,

$$\log(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta} \implies \mu_i = \exp\{\mathbf{X}_i \boldsymbol{\beta}\} = P_0 r_{i,1} \cdots r_{i,p},$$

où $P_0 = e^{\beta_0}$ et $r_{i,j} = e^{X_{i,j}\beta_j}$, $j = 1, \ldots, p$.

En pratique, l'utilisation des *relativités* $r_{i,j}$ est fréquente en tarification. Le Tableau 2.2 présente les distributions utilisées dans le présent mémoire ainsi que les fonctions de lien choisies.

Distribution	Support	Espérance	Variance	Fonction de lien	
				Canonique	Choisie
ODP(μ, ϕ)	$\mathbb{N}_0$	μ	$\phi\mu$	$\log(\mu)$	$\log(\mu)$
Bernoulli(n, p)	$\{0, 1\}$	p	$p(1-p)$	$\log\left(\frac{p}{1-p}\right)$	$\log\left(\frac{p}{1-p}\right)$
Tweedie(μ, p, ϕ)	$\mathbb{R}_{\geq 0}$	μ	$\phi\mu^p$	$\begin{cases} \frac{\mu^{1-p}}{1-p}, & p \neq 1 \\ \log(\mu), & p = 1 \end{cases}$	$\log(\mu)$

Tableau 2.2: Résumé des GLM utilisés et leurs fonctions de lien.

Formellement, il est possible d'évaluer l'ajustement de la fonction de lien par un test de score décrit dans l'article (Breslow, 2007). Il suffit de créer une fonction de lien plus générale qui a dans son image les fonctions de liens usuelles (p. ex. identité, logit, logarithmique, etc.) et de tester plusieurs de ces dernières comme hypothèses nulles.

2.4 Exemples

2.4.1 Méthode de Mack

Afin d'illustrer la méthode CL et le modèle linéaire généralisé appliqué au calcul de la réserve globale, le triangle cumulatif `usautotri9504` provenant de la librairie `CASdatasets` de `R` sera utilisé. Pour les fins des exemples suivants, les années d'accident 2000 jusqu'à 2004 sont choisies.

		Année de développement				
		1	2	3	4	5
Année d'accident	1	22 327	39 312	46 848	51 065	53 242
	2	23 141	40 527	48 284	52 661	·
	3	24 301	42 168	50 356	·	·
	4	24 210	41 640	·	·	·
	5	24 468	·	·	·	·

Tableau 2.3: Triangle de développement cumulatif.

Les estimateurs des facteurs de développement de la méthode CL sont définis par l'Équation (2.1). Pour le premier, on obtient

$$\widehat{\lambda}_1 = \frac{39\,312 + 40\,527 + 42\,168 + 41\,640}{22\,327 + 23\,141 + 24\,301 + 24\,210} = 1{,}7413.$$

En répétant la démarche pour chaque année de développement, on obtient les résultats présentés dans le Tableau 2.4.

Année de développement (j)	1	2	3	4	5
$\widehat{\lambda}_j$	1,7413	1,1925	1,0903	1,0426	1,0000

Tableau 2.4: Facteurs de développement de la méthode CL.

À l'aide des facteurs de développement et de l'Équation (2.2), on obtient une prédiction du montant cumulatif pour les accidents survenus pendant la deuxième année de $\widehat{C}_{2,5} = \widehat{\lambda}_4 C_{2,4} = 1{,}042632(52\,661) = 54\,906{,}04$. Alors, la réserve pour les accidents survenus lors de la deuxième année est $\widehat{R}_{2,5} = \widehat{C}_{2,5} - C_{2,4} = 54\,906{,}04 - 52\,661 = 2\,245{,}04$.

Les prédictions peuvent ensuite être calculées pour les espaces vierges du Tableau 2.3. On obtient ainsi les résultats présentés dans le Tableau 2.5.

		Année de développement				
		1	2	3	4	5
Année d'accident	1	22 327	39 312,00	46 848,00	51 065,00	53 242,00
	2	23 141	40 527,00	48 284,00	52 661,00	**54 906,04**
	3	24 301	42 168,00	50 356,00	**54 905,04**	**57 245,75**
	4	24 210	41 640,00	**49 653,87**	**54 139,49**	**56 447,56**
	5	24 468	**42 606,48**	**50 806,37**	**55 396,09**	**57 757,74**

Tableau 2.5: Prédictions de la méthode CL.

Alors, les réserves par année d'accident sont données par $\widehat{R}_{i,I} = \widehat{C}_{i,5} - C_{i,5-i+1},\ i = 2, 3, 4, 5$ et présentées dans le Tableau 2.6.

Année d'accident (i)	2	3	4	5
$\widehat{R}_i$	2 245,04	6 889,75	14 807,56	33 289,74

Tableau 2.6: Réserve CL prédite par année d'accident.

Enfin, la réserve totale prédite par la méthode CL est $\widehat{R} = \sum_{i=2}^{5} \widehat{R}_{i,5} = 57\,232,09$.

2.4.2 Modèles linéaires généralisés

Les modèles linéaires généralisés sont souvent réalisés en utilisant les montants incrémentaux. Ces derniers sont calculés dans le Tableau 2.7.

		Année de développement				
		1	2	3	4	5
Année d'accident	1	22 327	16 985	7 536	4 217	2 177
	2	23 141	17 386	7 757	4 377	·
	3	24 301	17 867	8 188	·	·
	4	24 210	17 430	·	·	·
	5	24 468	·	·	·	·

Tableau 2.7: Triangle de développement incrémental.

On rappelle que dans un modèle log-linéaire pour l'espérance conditionnelle du montant incrémental en fonction de l'année d'accident (i) et de l'année de développement (j), on a

$$\log E[Y_{i,j} \mid i, j] = \beta_0 + \alpha_i + \beta_j, \quad \alpha_1 = \beta_1 = 0.$$

En utilisant la fonction `glm`, les paramètres $\alpha_i,\ i = 2, \ldots, 5$ et $\beta_j,\ j = 0, 2, \ldots, 5$ sont estimés à l'aide d'une hypothèse de distribution du montant incrémental $Y_{i,j}$ ou encore

hypothèses sur les moments de $Y_{i,j}$. Un choix populaire est $Y_{i,j} \sim \text{ODP}(\mu_{i,j}, \phi)$ qui entraîne les relations $E[Y_{i,j} \mid i,j] = \mu_{i,j}$ et $\text{Var}[Y_{i,j} \mid i,j] = \phi\mu_{i,j}$ pour l'espérance et la variance, respectivement.

Les résultats d'estimation sont donnés au Tableau 2.8.

	$\widehat{\beta}_0$	10,023712	
$\widehat{\beta}_2$	-0,299330	$\widehat{\alpha}_2$	0,030776
$\widehat{\beta}_3$	-1,093247	$\widehat{\alpha}_3$	0,072506
$\widehat{\beta}_4$	-1,673546	$\widehat{\alpha}_4$	0,058465
$\widehat{\beta}_5$	-2,338009	$\widehat{\alpha}_5$	0,081410
	$\widehat{\phi}$	3,089769	

Tableau 2.8: Valeurs estimées pour les paramètres d'une régression de Poisson sur-dispersée.

Par exemple, le montant incrémental prédit à l'année d'accident 2 et l'année de développement 5 est $\widehat{\mu}_{2,5} = \exp\{\widehat{\beta}_0 + \widehat{\alpha}_2 + \widehat{\beta}_5\} = \exp\{\widehat{\beta}_0 + \widehat{\alpha}_2 + \widehat{\beta}_5\} = \exp\{7{,}716479\} = 2\ 245{,}04$. Tous les résultats sont présentés dans le Tableau 2.9.

		Année de développement				
		1	2	3	4	5
Année d'accident	1	22 327	16 985	7 536	4 217	2 177
	2	23 141	17 386	7 757	4 377	**2 245,04**
	3	24 301	17 867	8 188	**4 549,04**	**2 340,71**
	4	24 210	17 430	**8 013,88**	**4 485,61**	**2 308,07**
	5	24 468	**18 138,48**	**8 199,88**	**4 589,73**	**2 361,64**

Tableau 2.9: Prédictions de la régression de Poisson sur-dispersée.

La réserve prédite est alors $\widehat{R} = \sum_{(i,j)\in\overline{\Delta}} \widehat{\mu}_{i,j} = 57\ 232{,}09$. Tel que précité, la réserve prédite par le GLM de Poisson est la même que celle prédite par la méthode CL.

2.5 Techniques de rééchantillonage

Le bootstrap est une technique de rééchantillonage permet d'estimer une statistique qui découle d'une distribution sous-jacente ainsi que la dispersion de cette dernière. Cette technique est également souvent utilisée dans le contexte d'apprentissage statistique (à définir dans les prochains chapitres) afin de mitiger le risque le surajustement et de mesurer l'erreur d'estimation d'un modèle de régression supervisé. En actuariat IARD afin d'estimer les moments et les quantiles de la distribution prédictive de la réserve lorsque ces derniers ne peuvent pas être obtenus de façon analytique. Deux références excellentes pour l'utilisation du bootstrap appliqué à la modélisation de la réserve sont (England et Verrall, 2002) et (Taylor et McGuire, 2016).

Pour illustrer le bootstrap d'un modèle de régression, on considère l'ensemble

$$\mathcal{D}_n = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$$

de n paires de vecteurs de covariables $\mathbf{X}_i$ et de réponses y_i. La prédiction de y_i est définie par $\widehat{y}_i = f(\mathbf{X}_i; \boldsymbol{\upsilon}_i)$ où f est une fonction de régression qui prend comme arguments le vecteur de covariables $\mathbf{X}_i$ et un vecteur de paramètres $\boldsymbol{\upsilon}_i$, et retourne une prédiction $\widehat{y}_i$. En fonction du modèle de régression, $\boldsymbol{\upsilon}_i$ peut comprendre des inconnus qui devront être estimés ou encore des hyperparamètres qui devront être choisis ou validés par l'utilisateur.

En général, le bootstrap possède trois variantes : paramétrique, semi-paramétrique et non paramétrique. Les deux premières seront d'intérêt pour le présent mémoire, alors l'exposition sera limitée à celles-ci.

Algorithme 1 : L'algorithme du bootstrap des résidus.

Input : Vecteurs $\widehat{\boldsymbol{\varepsilon}}$, $\widehat{\boldsymbol{\upsilon}}$ et fonction de régression f

Output : Pseudo-paramètres estimés $\widehat{\boldsymbol{\upsilon}}^*$

Data : Base de données $\mathcal{D}_n$

1. Piger un échantillon aléatoire simple avec remise de taille n des résidus de Pearson $\widehat{\boldsymbol{\varepsilon}}^* = \dfrac{y_i - \widehat{y}_i}{\sqrt{\widehat{\mathrm{Var}}[Y_i]}}$;
2. Calculer les pseudo-réponses $\widehat{y}_i^* = f(\mathbf{X}_i; \widehat{\boldsymbol{\upsilon}}) + \widehat{\varepsilon}_i^*$, $i = 1, \ldots, n$;
3. Réajuster les paramètres du modèle avec $y_i \to \widehat{y}_i^*$ et les variables explicatives $\mathbf{X}_i$ pour $i = 1, \ldots, n$;
4. Extraire $\widehat{\boldsymbol{\upsilon}}^*$.

Figure 2.1: Algorithme du bootstrap des résidus de Pearson.

2.5.1 Le bootstrap semi-paramétrique

Le bootstrap semi-paramétrique est souvent nommé « rééchantillonnage des résidus ». Un résidu est le résultat d'une fonction de distance $\widehat{\varepsilon}_i = h(y_i, \widehat{y}_i)$ qui tente de mesurer l'écart entre la réponse y_i et la prédiction $\widehat{y}_i$. L'algorithme du bootstrap des résidus est décrit à la Figure 2.1.

Une hypothèse classique qui justifie l'utilisation du bootstrap semi-paramétrique est que les résidus sont indépendants et identiquement distribués. Ceci implique que les résidus $\widehat{\boldsymbol{\varepsilon}}$ et les pseudo résidus $\widehat{\boldsymbol{\varepsilon}}^*$ possèdent la même distribution. Pour cette raison, les résidus de Pearson sont fréquemment utilisés (voir l'étape 1 de l'algorithme de la Figure 2.1). Les fonctions `BootChainLadder` et `glmReserve` permettent d'effectuer le bootstrap semi-paramétrique pour les modèles GLM collectifs et le modèle de Mack, respectivement.

2.5.2 Le bootstrap paramétrique

Le bootstrap paramétrique, parfois appelé simplement « simulation », demande que la distribution de la variable réponse soit connue. En particulier, si la i^{e} réponse y_i est une réalisation de la variable aléatoire $Y_i \sim F(\boldsymbol{\theta}_i)$, où $\boldsymbol{\theta}$ est un vecteur de paramètres et F une fonction de distribution.

Plusieurs fonctions en `R` permettent la simulation de variables aléatoires, dont `rtweedie` et `rnbinom` pour les variables aléatoires Tweedie, binomiale-négative et Poisson surdispersées, respectivement. Une référence classique pour la simulation est (Devroye, 1986).

La simulation est nécessairement plus rapide que le bootstrap semi-paramétrique puisqu'elle ne nécessite pas de réajustement du modèle relativement à des pseudo réponses. Puisque les arbres de régression sont des modèles de régression non paramétriques, le bootstrap non paramétrique est naturellement approprié pour celui-ci. Par contre, le grand nombre de réajustements peut être prohibitif. Alors, un rééchantillonage des prédictions sera favorisé pour évaluer la distribution prédictive de la réserve. Les arbres de régression seront présentés au prochain chapitre ainsi que l'introduction aux mécanismes de censure.

CHAPITRE III

INTRODUCTION AUX MODÈLES INDIVIDUELS DE PROVISIONNEMENT ET AUX ARBRES DE RÉGRESSION

3.1 Introduction

Le présent chapitre sera divisé en trois sections. En premier lieu, un survol des modèles individuels de réserve sera présenté afin de clarifier les différences entre ces derniers et les modèles collectifs, les modèles granulaires et les modèles de provisionnement par imputation. Il y aura une comparaison de haut niveau des avantages et désavantages de ces modèles qui motivera l'utilisation des modèles de provisionnement par imputation.

Deuxièmement, les préliminaires théoriques pertinents à ce mémoire pour les modèles par imputation seront présentés. Il y aura une introduction aux mécanismes de censure (*dropout mechanisms*) et à l'impact que ceux-ci peuvent avoir sur les hypothèses des modèles de régression. Les IPCW seront définis en fonction de la réciproque de la probabilité que le délai de fermeture soit censuré à la d'évaluation du portefeuille. Cette probabilité sera calculée à l'aide de l'estimateur Kaplan-Meier ainsi qu'à l'aide d'une régression logistique.

Finalement, les arbres de régression seront présentés. Ils permettront le développement du modèle de provisionnement par imputation par arbres de régression pondérés basés sur (Lopez *et al.*, 2016), (Lopez *et al.*, 2019), (Lopez et Milhaud, 2021) au prochain chapitre.

3.2 Survol des modèles individuels de provisionnement

À la suite des augmentations simultanées du volume et de la richesse des informations disponibles à un assureur, les modèles de provisionnement tentent maintenant d'inclure des informations individuelles afin de prédire la réserve globale d'un portefeuille d'assurés avec une précision supérieure. C'est ce qui distingue principalement les approches individuelles des approches collectives : la possibilité de modéliser précisément chacune des composantes du développement d'une réclamation pour chacun des assurés d'un portefeuille.

Lors des dernières vingtaines années, on a assisté à une augmentation rapide du nombre de modèles de type individuel pour la réserve actuarielle en assurance IARD. Cette classe de modèles peut porter plusieurs noms selon la source consultée. Selon (Taylor, 2019), les modèles individuels sont divisés en deux sous-groupes : les modèles granulaires et les modèles par apprentissage statistique. Les modèles granulaires tentent de modéliser la réserve par module où chaque module est un modèle (ou encore un ensemble de modèles) qui tente d'expliquer une étape du processus de développement d'une réclamation. Par exemple, (Antonio et Plat, 2014) modélise la réserve RBNS par un processus de simulation pour chaque réclamation ouverte. Le temps jusqu'à la survenance du prochain événement relatif au développement de la réclamation (un paiement, une fermeture avec paiement ou une fermeture sans paiement) demande la simulation d'une réalisation d'une variable aléatoire uniforme sur l'intervalle $(0, 1)$. Ensuite, le type d'événement est simulé ainsi que, le cas échéant, le montant du paiement. En fonction du nombre de lignes d'affaires présentes au portefeuille d'intérêt, ce type de modèle repose sur plusieurs regroupements d'hypothèses et peut être long à mettre en oeuvre puisque le problème de prédiction du Chapitre 2 est remplacé par un problème de simulation.

La classe de modèles par apprentissage statistique est une méthode plus directe, car la variable réponse est le paiement incrémental ou cumulatif à une période donnée. Les modèles de cette classe peuvent prendre plusieurs formes et, dans ce mémoire, on s'intéressera particulièrement aux arbres de régression et aux GLM. En général, les

modèles par apprentissage statistique reposent sur une méthode statistique qui demande moins d'intervention de la part de l'utilisateur que les modèles granulaires. À titre d'exemple, (McGuire *et al.*, 2018) utilise la pénalisation LASSO pour automatiser la sélection de variables dans un modèle similaire aux modèles GLM susmentionnés, (Wüthrich, 2016) utilise les arbres de régression pour calculer la probabilité qu'une réclamation ait un paiement à une certaine période et (Gabrielli, 2021) utilise un réseau de neurones afin de calculer la réserve RBNS pour chaque réclamation qui possède une série incomplète de paiements. L'article (Blier-Wong *et al.*, 2021) présente une liste plus exhaustive de modèles de provisionnement granulaires et d'apprentissage statistique. Les modèles granulaires peuvent éclairer le processus de développement d'une réclamation, mais peuvent aussi être difficiles à formuler. Les modèles par apprentissage statistique, même s'ils sont basés sur une construction simple, peuvent être opaques et difficiles à interpréter. Le modèle proposé par ce travail au prochain chapitre est une reformulation du modèle de (Lopez *et al.*, 2016) à l'aide de GLM qui conserve plusieurs propriétés statistiques désirables.

Les modèles granulaires permettent une modélisation plus précise du processus de réclamation, mais les modules sont souvent structurés en cascade (ou de façon hiérarchique) : un module dépend des sorties d'autres modules. La structure de dépendance peut ainsi s'avérer très complexe. De plus, lorsque le nombre de modules augmente, le nombre total de paramètres augmente ce qui peut conduire à une surparamétrisation. Même s'ils ne tiennent généralement pas compte de la censure, ils permettent souvent une amélioration de la précision des prédictions comparativement aux modèles collectifs. Cependant, il peut être difficile d'identifier précisément quel(s) module(s) contribue(nt) à cette amélioration.

Les modèles par apprentissage statistique varient énormément dans leurs constructions ce qui fait que certains sont plus simples à interpréter que d'autres. Les modèles basés sur le maximum de vraisemblance pénalisé (LASSO ou autre) sont aussi simples que les GLM en pratique. Les réseaux de neurones, surtout quand ils sont appliqués au provisionnement, peuvent être difficiles à interpréter sans une structure de modèle spéciale.

D'un point de vue réglementaire, (Charpentier et Pigeon, 2016) note que les modèles individuels permettent une plus grande flexibilité afin de satisfaire les balises financières comparativement aux modèles collectifs. Les modèles individuels sont entraînés à l'aide de beaucoup plus de données, supposent moins de contraintes structurelles sur celles-ci et offrent l'avantage de permettre la modélisation précise de la dépendance.

3.3 Apprentissage statistique

En général, l'apprentissage statistique réfère à la modélisation statistique d'un certain processus défini qui prend comme entrée un ensemble de données et des hyperparamètres décidés par l'utilisateur. Si le processus en question retourne une sortie numérique ou catégorielle, l'apprentissage est appelé *supervisé*. Par exemple, l'apprentissage supervisé englobe tous les modèles de régression.

3.3.1 La séparation de données d'entraînement et de validation

Pour l'apprentissage supervisé, il est typique de diviser une base de données $\mathcal{D}_n$ de taille n en bases de données d'entraînement et de validation pour l'ajustement du modèle et de validation pour calculer des mesures d'erreur, respectivement. Dans la modélisation de la réserve, les données d'entraînement et de validation sont naturellement tranchées en fonction de la base de données $\mathcal{D}_n$ et de la date d'évaluation de la réserve. Soit $\mathcal{D}_{\mathcal{T}}$, la base de données d'entraînement composée de réclamations déclarées avant l'évaluation du portefeuille et $\mathcal{D}_{\mathcal{V}}$ la base de données de validation composée des réclamations déclarées après la date d'évaluation. En temps discret, pour une base de données $\mathcal{D}_n$ de taille n comprenant N réclamations,

$$\mathcal{D}_{\mathcal{T}} = \{(\mathbf{X}_i, y_i) \in \mathcal{D}_n \mid i^{(k)} + j \leq n_e, 1 \leq j \leq J, k = 1, \ldots, N\}$$

et $\mathcal{D}_{\mathcal{V}} = \mathcal{D}_n \setminus \mathcal{D}_{\mathcal{T}}$.

Quant aux modèles individuels proposés dans ce mémoire, l'ensemble d'entraînement $\mathcal{D}_{\mathcal{T}}$ sera construit selon une méthode décrite dans la Section 4.4. Ensuite, chaque réclamation

ouverte présente dans la base de données d'entraînement aura une base de données propre qui sera un sous-ensemble de $\mathcal{D}_{\mathcal{T}}$.

3.3.2 Validation croisée

La validation croisée est une technique d'échantillonnage permettant de déterminer la fiabilité d'un modèle. En général, la validation croisée de k blocs demande de séparer un jeu de données en k sous-ensembles de même taille. L'ajustement d'un modèle est ensuite effectué k fois, une fois pour chaque regroupement de blocs de taille $k-1$. Des prédictions sont ensuite calculées pour chaque modèle sur le bloc omis lors de l'ajustement qui sera nommé le bloc de validation.

Typiquement, pour un modèle avec une seule variable réponse où le cout de l'erreur de la prédiction est constant, le MSE ou encore la racine carrée du MSE est souvent utilisé. Afin de simplifier l'ajustement des modèles qui suivent, le RMSE est utilisé, malgré que le cout de la surestimation de la réserve soit inférieur au cout de la sous-estimation de la réserve. La validation croisée sera appliquée à la sélection des hyperparamètres des modèles non paramétriques qui suivent.

3.4 Introduction aux mécanismes de censure

Dans un contexte de provisionnement, la date d'évaluation de la réserve induit nécessairement un mécanisme de censure et la création de données manquantes (*dropout mecanism*). Il est important de définir le type de mécanisme qui cause le manque de données (*missingness*). (Little et Rubin, 1986) définit trois types de données manquantes selon le mécanisme sous-jacent.

Avant tout, on pose que

- $\mathbf{Z} = [\mathbf{y} \mid \mathbf{X}] = [\mathbf{z}_1 | \ldots | \mathbf{z}_n]^T$ est une matrice comprenant le vecteur réponse et la matrice des covariables ;
- $\boldsymbol{\delta}$ est un vecteur dont $\delta_i = \mathbb{1}(i^{\text{e}}$ ligne de $\mathbf{Z}$ possède une valeur manquante$)$;

— le mécanisme de censure est décrit par une densité conditionnelle $f_{\boldsymbol{\Delta}|\mathbf{Z}}$;

— $\boldsymbol{\gamma}$ est un vecteur de paramètres non observables.

Les types de données manquantes sont :

1. *Missing completely at random* (MCAR) : la densité $f_{\boldsymbol{\Delta}|\mathbf{Z}}(\delta_i \mid \mathbf{z}_i; \boldsymbol{\gamma}) = f_{\boldsymbol{\Delta}}(\delta_i)$, c'est-à-dire que la densité f ne peut pas dépendre des observations $\mathbf{z}_i$ ni des variables non observables $\boldsymbol{\gamma}$. Alors, $\Delta_i \perp\!\!\!\perp \mathbf{Z}_i$.
2. *Missing at random* (MAR) : la densité $f_{\boldsymbol{\Delta}|\mathbf{Z}}(\delta_i \mid \mathbf{z}_i; \boldsymbol{\gamma}) = f_{\boldsymbol{\Delta}|\mathbf{X}}(\delta_i \mid \mathbf{X}_i)$, c'est-à-dire que la densité f peut dépendre des observations $\mathbf{X}_i$, mais pas des paramètres non observables $\boldsymbol{\gamma}$. Alors, $\Delta_i \perp\!\!\!\perp \mathbf{Z}_i \mid \mathbf{X}_i$.
3. *Missing not at random* (MNAR) : la densité $f_{\boldsymbol{\Delta}|\mathbf{Z}}(\delta_i \mid \mathbf{z}_i; \boldsymbol{\gamma})$ dépend du vecteur $\mathbf{z_i}$.

Le test de Little (Enders, 2010) permet d'effectuer un test d'hypothèse pour juger si le mécanisme d'abandon est de type MCAR. Il est possible d'effectuer une régression logistique avec δ_i comme variable réponse et $\mathbf{z}_i$ comme covariables pour évaluer l'importance des covariables $\mathbf{X}_i$ et de la variable réponse y_i sur la probabilité d'abandon. En général, il est difficile, voire souvent impossible de détecter la structure d'abandon et le type de mécanisme. Ainsi, ce dernier est généralement posé en hypothèse. En provisionnement, il y a plusieurs variables qui peuvent avoir un impact sur la probabilité d'abandon (p. ex. le délai de déclaration). De plus, parce que la date d'évaluation de la réserve est déterminée par l'utilisateur, il sera supposé que le mécanisme d'abandon est de type MAR.

Étant donné des informations censurées (ou manquantes), l'imputation de celles-ci peut être nécessaire avant de procéder à l'ajustement d'un modèle. Une alternative à l'imputation est l'utilisation des IPCW telle que proposée par (Lopez *et al.*, 2016), (Lopez *et al.*, 2019), et (Lopez et Milhaud, 2021).

3.4.1 Calcul des IPCW

Les mécanismes de censure décrits ci-dessus proviennent de l'analyse de survie où les données manquantes surviennent souvent à cause d'une erreur imprévisible ou de manière aléatoire. Dans le contexte du provisionnement individuel, la censure est induite à cause

de l'évaluation de la réserve et le statut de chaque réclamation (ouvert ou fermé) est aléatoire.

Utilisant les définitions de la référence (Tsiatis, 2007), les données complètes (*full data*) sont toutes les données qui peuvent être recueillies à propos d'un ensemble de réclamations dans un horizon de temps prédéterminé. Les données observées (*observed data*) sont les données qui sont réellement recueillies, dont une certaine proportion est manquante.

Supposons que les mécanismes d'abandon sont de type MCAR. Si on suppose que les variables aléatoires Δ_i (dont la réalisation est notée δ_i) et Y_i (dont la réalisation est notée y_i) sont indépendamment et identiquement distribuées (i.i.d.), il s'ensuit que la moyenne des données complètes converge en probabilité vers l'espérance

$$\begin{aligned}\frac{\sum_{i=1}^{n}\delta_i y_i}{\sum_{i=1}^{n}\delta_i} &\xrightarrow{P} \frac{E[\Delta Y]}{E[\Delta]}\\ &= \frac{E[\Delta]E[Y]}{E[\Delta]}\\ &= E[Y],\end{aligned}$$

par indépendance de Δ_i et Y_i.

Par contre, si le mécanisme d'abandon est de type MAR, par les mêmes hypothèses sur Δ_i et Y_i, et si on pose la probabilité $P(\Delta_i = 1 \mid \mathbf{X}_i) = \pi(\mathbf{X}_i)$, alors l'*inverse probability censoring weight* (IPCW) est la réciproque de la probabilité de censure (ou d'abandon). On peut vérifier que

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\delta_i y_i \pi(\mathbf{X}_i)^{-1} &\xrightarrow{P} E[\Delta Y \pi(\mathbf{X})^{-1}] \qquad (3.1)\\ &= E[E[\Delta Y \pi(\mathbf{X})^{-1} \mid Y, \mathbf{X}]]\\ &= E[Y E[\Delta \mid Y, \mathbf{X}]]\pi(\mathbf{X})^{-1}\\ &= E[Y\pi(\mathbf{X})]\pi(\mathbf{X})^{-1}\\ &= E[Y].\end{aligned}$$

Afin d'estimer la fonction de probabilité $\pi(\mathbf{X}_i)$, deux estimateurs seront proposés :

1. Estimateur par MV par la régression logistique (IPCW GLM) ;
2. Estimateur non paramétrique de Kaplan-Meier (IPCW K-M).

Par la régression logistique, il est possible d'incorporer des covariables. Précisément, on pose

$$\log\left(\frac{\pi(\mathbf{X}_i)}{1-\pi(\mathbf{X}_i)}\right) = \mathbf{X}_i^T\boldsymbol{\beta}.$$

Les propriétés de l'estimateur $\widehat{\boldsymbol{\beta}}$ sont décrites à la Section 2.3 pour les modèles linéaires généralisés.

L'estimateur Kaplan-Meier (K-M) est un estimateur non paramétrique de la fonction de survie de la durée à la survenance d'un événement. Dans le cas de du provisionnement en assurance IARD, la durée d'intérêt est le nombre de nombre de jours avant la fermeture de la réclamation relativement à la date d'accident. L'estimateur de la fonction de survie de K-M est

$$\widehat{S}_T(t) = \prod_{\{j|t_j\leq t\}}\left(1-\frac{d_j}{n_j}\right), \tag{3.2}$$

où t_j est la j^{e} durée observée jusqu'à la fermeture (en ordre croissant), d_j est le nombre d'observations ayant la durée t_j et n_j le nombre d'observations des réclamations qui sont ouvertes après t jours. Les observations ayant $t = t_j$ sont comptées par les n_j. La fonction `prodlim` permet de calculer l'estimateur K-M de la fonction de survie ainsi que l'estimateur *inversée* (où l'argument n'est plus le délai de fermeture, mais plutôt le délai jusqu'à la date d'évaluation).

La fonction de survie inversée (pour données censurées) sera utilisée au prochain chapitre pour illustrer les arbres de régression afin d'estimer le montant total payé.

3.5 Arbres de régression

Les modèles de régression présentés dans les sections précédentes sont de type paramétrique. En contrepartie, les arbres de régression sont de type non paramétrique. Les arbres de régression divisent le produit cartésien de l'étendue de chaque covariable en sous-ensembles et une prédiction de la variable réponse est attribuée à chaque sous-ensemble.

En général, soit $\mathbf{X}$ une matrice de covariables telle que définie à la Section 2.3. Soit $\mathbf{X}_j^r$ l'étendue de la j^{e} colonne de la matrice $\mathbf{X}$, notée $\mathbf{X}_j$. Par exemple, si $\mathbf{X}_j$ est une covariable numérique, $\mathbf{X}_j^r$ est un intervalle fermé, si $\mathbf{X}_j$ est une covariable catégorielle, $\mathbf{X}_j^r$ est un ensemble discret.

Un arbre de régression est une fonction $f(\cdot\,; \boldsymbol{\upsilon}, \mathbf{w})$ où

$$f(\mathbf{x}_0; \boldsymbol{\upsilon}, \mathbf{w}) = \overline{y}_l^{\mathbf{w}} = \frac{\sum_{i=1}^n w_i y_i \mathbb{1}(\mathbf{X}_i \in \mathcal{R}_l)}{\sum_{i=1}^n w_i \mathbb{1}(\mathbf{X}_i \in \mathcal{R}_l)}, \quad \text{pour } \mathbf{x}_0 \in \mathcal{R}_l, \tag{3.3}$$

$l = 1, \ldots, L$ pour un arbre ayant L *feuilles* qui sont les sous-ensembles $\mathcal{R}_1, \ldots, \mathcal{R}_L$ et le vecteur d'hyperparamètres est noté par $\boldsymbol{\upsilon}$. Pour prédire $\overline{y}_l$ il faut qu'un vecteur de covariables $\mathbf{x}_0 \in \mathcal{R}_l$ où $\mathcal{R}_l \subset \mathbf{X}_1^r \times \cdots \times \mathbf{X}_p^r$. On note que $\bigcup_{l=1}^L \mathcal{R}_l = \mathbf{X}_1^r \times \cdots \times \mathbf{X}_p^r$ afin de s'assurer que les $\mathcal{R}_l$ partitionnent l'espace des covariables.

On peut réécrire l'Équation (3.3) à l'aide de fonctions indicatrices :

$$f(\mathbf{x}_0; \boldsymbol{\upsilon}, \mathbf{w}) = \sum_{s=1}^{L} \overline{y}_s^{\mathbf{w}} \mathbb{1}(\mathbf{x}_0 \in \mathcal{R}_s). \tag{3.4}$$

L'algorithme CART (*classification and regression trees*) développé par (Breiman *et al.*, 1984) permet de déterminer les noeuds optimaux de l'arbre qui minimise l'écart quadratique moyen de la variable réponse relativement à la moyenne arithmétique calculée sur un sous-ensemble de $\mathbf{X}_1^r \times \cdots \times \mathbf{X}_p^r$. L'algorithme CART correspond exactement à l'algorithme wCART avec $w_i = 1$. L'algorithme wCART est présenté à la Figure 3.1.

On note que cet algorithme permet de prendre en compte les interactions entre les covariables. Afin de limiter le sur-apprentissage (le biais de la prédiction est petit

Algorithme 2 : Construction d'un arbre de régression par la méthode wCART

Input : L'ensemble d'hyperparamètres $\mathcal{P}_{\boldsymbol{\upsilon}}$ et les poids $\mathbf{w}$

Output : Arbre de régression f

Data : Base de données $\mathcal{D}_{\mathcal{T}}$ (avec p covariables numériques)

1. Étant donné les hyperparamètres $\boldsymbol{\upsilon}$, soit $\mathcal{P}_{\boldsymbol{\upsilon}}$ l'ensemble des valeurs acceptables.

2. Soit $\mathcal{C}$ la partition de $\mathbf{X}_1^r \times \cdots \times \mathbf{X}_p^r$. Initialement, $\mathcal{C} = \mathbf{X}_1^r \times \cdots \times \mathbf{X}_p^r$.

3. Créer une vecteur d'hyperparamètres temporaire $\boldsymbol{\upsilon}^* = \boldsymbol{\upsilon}$.

4. **while** $\boldsymbol{\upsilon}^* \in \mathcal{P}_{\boldsymbol{\upsilon}}$ **do**

   foreach $j \in \{1, \ldots, p\}$ *et chaque sous-ensemble* $\mathcal{R}$ *de la partition* $\mathcal{C}$ **do**

   — Pour un point de séparation s,

$$\mathcal{R}_1(j,s) = \{\text{Élements de } \mathbf{X}_j^r \text{ inférieurs à } s\} \quad \text{et}$$
$$\mathcal{R}_2(j,s) = \{\text{Élements de } \mathbf{X}_j^r \text{ supérieurs ou égaux à } s\}$$

   Chercher la covariable avec indice j et le point s qui minimise

$$\min_{j,s} \left\{ \min_{c_1} \sum_{X_{i,j} \in \mathcal{R}_1(j,s)} w_i (y_i - c_1)^2 + \min_{c_2} \sum_{X_{i,j} \in \mathcal{R}_2(j,s)} w_i (y_i - c_2)^2 \right\} \quad \text{où}$$

$$c_1 = \frac{\sum_{i=1}^n w_i y_i \mathbb{1}(X_{i,j} \in \mathbf{X}_j \cap \mathcal{R}_1(j,s))}{\sum_{i=1}^n w_i \mathbb{1}(X_{i,j} \in \mathbf{X}_j \cap \mathcal{R}_1(j,s))} \quad \text{et}$$
$$c_2 = \frac{\sum_{i=1}^n w_i y_i \mathbb{1}(X_{i,j} \in \mathbf{X}_j \cap \mathcal{R}_2(j,s))}{\sum_{i=1}^n w_i \mathbb{1}(X_{i,j} \in \mathbf{X}_j \cap \mathcal{R}_2(j,s))};$$

   — Réévaluer $\boldsymbol{\upsilon}^*$ avec les informations du noeud créé ci-haut (p. ex. le nombre minimal d'observations dans un noeud est $\min\{|\mathbf{X}_j \cap \mathcal{R}_1(j,s)|, |\mathbf{X}_j \cap \mathcal{R}_2(j,s)|\}$ et doit respecter $\boldsymbol{\upsilon}$). On obtient une nouvelle partition $\mathcal{C} \to (\mathcal{C} \setminus \mathcal{R}) \cup (\mathcal{R}_1 \cup \mathcal{R}_2)$.

5. Retourner la fonction f définie ci-haut sur la partition $\mathcal{C}$ de $\mathbf{X}_1^r \times \cdots \times \mathbf{X}_p^r$. Alors, f est définie selon (3.4), supposant que les $\mathcal{R}_l$ sont libellés uniquement.

Figure 3.1: Étapes de l'algorithme wCART.

relativement à la variance de la prédiction), on effectue une pénalisation des noeuds de l'arbre afin de faire un élagage de l'arbre. La convergence de l'estimateur $\overline{y}_l$ pour chaque feuille ainsi que la convergence de la prédiction $\widehat{f}$ sont démontrées dans (Lopez *et al.*, 2016).

Soit T un arbre de régression quelconque et $|T| = L$ le nombre de feuilles de celui-ci. On note $N_l = |\mathcal{R}_l|,\ l = 1, \ldots, L$ et $\widehat{f}(\mathbf{x}_0; \boldsymbol{\upsilon}, \mathbf{w}) = \overline{y}_l^{\mathbf{w}},\ \mathbf{x}_0 \in \mathcal{R}_l$ pour l'arbre T. Une fois l'arbre estimé par l'algorithme wCART, l'élagage peut aussi être fait avec les poids w_i.

$$C_\alpha(T) = \sum_{l=1}^{L} N_l Q_l(T) + \alpha|T| \tag{3.5}$$

pour l'arbre T avec $|T|$ feuilles et

$$Q_l(T) = \frac{1}{N_l} \sum_{i=1}^{n} w_i \mathbb{1}(\mathbf{X}_i \in \mathcal{R}_l)[y_i - \widehat{f}(\mathbf{X}_i; \boldsymbol{\upsilon}, \mathbf{w})]^2.$$

Pour chaque valeur de $\alpha \geq 0$, on peut démontrer qu'il existe un sous-arbre qui minimise la valeur de $C_\alpha(T)$ (Breiman *et al.*, 1984) pour $w_i = 1$. Alors, pour un α donné, on élimine successivement les noeuds qui produisent la plus petite augmentation de la somme $\sum_{l=1}^{L} N_l Q_l(T)$. À l'extrême, on obtient un arbre trivial sans noeud. Afin de mitiger le risque de surajustement, pour chaque valeur de α choisie, la validation croisée est utilisée afin de minimiser l'erreur quadratique de la prédiction du bloc de validation après que l'élagage est effectué sur les blocs d'entraînement. Typiquement, le nombre de blocs est $k = 5$ ou $k = 10$. La convergence de la stratégie d'élagage de l'algorithme wCART a été démontrée dans (Lopez *et al.*, 2016).

Les arbres de régression, selon l'algorithme CART, présentent plusieurs avantages et inconvénients :

— Les variables catégorielles sont absentes jusqu'à présent. Pour une covariable comprenant q catégories, il y a $2^{q-1} - 1$ façons de partitionner celle-ci comme à l'algorithme présenté à la Figure 3.1.

— Les arbres peuvent avoir une grande variance de prédiction puisqu'un changement

dans le jeu de données peut entraîner une série de noeuds différents. Ceci survient à cause de la nature hiérarchique de l'arbre.

— La fonction f n'est pas continue.

— Si la relation entre la variable réponse et les covariables est additive, l'algorithme CART ne capturera pas forcément cette dernière.

3.6 Exemple

Cette section présente un exemple simple d'un arbre de régression pondéré avec deux covariables continues avec un ensemble d'hyperparamètres donné. De plus, les poids seront fixes puisqu'un exemple de calcul des poids IPCW pour l'application à la modélisation de la réserve sera abordé au prochain chapitre.

Avant de décrire l'ensemble de données d'entraînement fictif, il est nécessaire de décrire les hyperparamètres d'un arbre de régression. Les hyperparamètres des arbres sont les suivants : le nombre minimal d'observations par feuille (noeud terminal), le nombre minimal d'observations par séparation, la profondeur maximale de l'arbre (le nombre de couches de noeuds) et le paramètre de complexité, noté par α dans l'Équation (3.5) pour effectuer l'élagage. La fonction `rpart` prend les mêmes hyperparamètres nommés `minbucket`, `minsplit`, `maxdepth` et `cp`, respectivement.

Les données d'entraînement sont présentées dans le Tableau 3.1.

i	$X_{i,1}$	$X_{i,2}$	y_i	w_i
1	0,5	6	50	1,25
2	1	3,5	10	0
3	3	2	100	1,1
4	4	1	30	1,05

Tableau 3.1: Exemple de données fictives.

L'espace des covariables est alors $\mathbf{X}_1^r \times \mathbf{X}_2^r = [0{,}5\,,4] \times [1,6]$. Supposons que les hyperparamètres sont `minbucket` $= 1$, `minsplit` $= 3$, `maxdepth` $= 5$ et `cp` $= 0$.

Si l'on choisit la variable $\mathbf{X}_1$ en premier, les candidats initiaux pour un point de séparation s sont $(0{,}75\,,2\,,3{,}5)$ (on suppose que l'on choisit, à chaque fois, le point milieu entre deux valeurs successives de $\mathbf{X}_1$).

1. Pour $s = 0{,}75$, les moyennes arithmétiques par partition sont

$$c_1 = \frac{1{,}25(50)}{1{,}25} = 50 \text{ et } c_2 = \frac{0(10) + 1{,}1(100) + 1{,}05(30)}{0 + 1{,}1 + 1{,}05} = 65{,}81$$

   et l'erreur quadratique moyenne est

$$1{,}25(50-50)^2 + 0(10-65{,}81)^2 + 1{,}1(100-65{,}81)^2 + 1{,}05(30-65{,}81)^2 = 2632{,}326;$$

2. Pour $s = 2$, les moyennes arithmétiques par partition sont

$$c_1 = \frac{1{,}25(50) + 0(10)}{1{,}25 + 0} = 50 \text{ et } c_2 = \frac{1{,}1(100) + 1{,}05(30)}{1{,}1 + 1{,}05} = 65{,}81$$

   et l'erreur quadratique moyenne est

$$1{,}25(50-50)^2 + 0(10-50)^2 + 1{,}1(100-65{,}81)^2 + 1{,}05(30-65{,}81)^2 = 2632{,}326;$$

3. Pour $s = 3{,}5$, les moyennes arithmétiques par partition sont

$$c_1 = \frac{1{,}25(50) + 0(10) + 1{,}1(100)}{1{,}25 + 0 + 1{,}1} = 73{,}40 \text{ et } c_2 = \frac{1{,}05(30)}{1{,}05} = 30$$

   et l'erreur quadratique moyenne est

$$1{,}25(50-73{,}40)^2 + 0(10-73{,}40)^2 + 1{,}1(100-73{,}40)^2 + 1{,}05(30-30)^2 = 1462{,}766.$$

Alors, la partition de l'espace des covariables est $\mathcal{C} = \{[0{,}5\,,3{,}5) \times [1,6]\} \cup \{[3{,}5\,,4] \times [1,6]\}$.

Ensuite, on peut partitionner $\mathcal{C}$ davantage selon $\mathbf{X}_1$ ou $\mathbf{X}_2$. D'une manière qui est hors de la portée du présent mémoire, l'algorithme wCART détermine que le prochain point de séparation est $X_{i,2} = 4$. Finalement, l'algorithme wCART retourne la partition

$\mathcal{C} = \{[0{,}5\,,3{,}5) \times [1,4)\} \cup \{[0{,}5\,,3{,}5) \times [4,6]\} \cup \{[3{,}5\,,4] \times [1,6]\}$. C'est à cette étape que l'algorithme wCART termine à cause des hyperparamètres.

Pour réaliser l'élagage de l'arbre illustré à la Figure 3.2, il suffit de comprendre que l'objectif est de réduire la variance des prédictions au détriment d'une augmentation du biais des prédictions en éliminant les branches de l'arbre qui diminuent l'erreur quadratique de peu relativement à un facteur de pénalisation α. Par défaut, la fonction `rpart` choisit $k = 10$ comme nombre de blocs pour la validation croisée. Lorsque le nombre de blocs de la validation croisée augmente, le nombre de données dans chaque bloc de validation diminue. Il s'ensuit que l'erreur de validation augmente pour les sous-arbres de la séquence déterminée par le nombre de noeuds terminaux. À l'inverse, diminuer le nombre de blocs cause davantage de données à se retrouver dans le bloc de validation qui diminue l'erreur de validation. On s'attend à ce qu'un bias se manifeste par contre puisque la taille du bloc de validation est inférieure à l'ensemble de données d'entraînement. Les noeuds de l'arbre contiennent la prédiction du montant total payé et le nombre d'observations de l'ensemble d'entraînement (brut et en pourcentage).

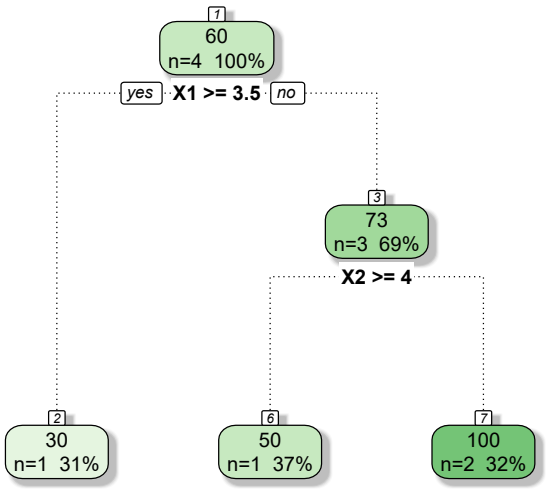


Figure 3.2: Arbre de régression non élagué de l'exemple illustratif.

CHAPITRE IV

ARBRES DE RÉGRESSION PONDÉRÉS POUR LA RÉSERVE INDIVIDUELLE

4.1 Introduction

Le chapitre précédent a résumé les principes théoriques nécessaires pour le présent chapitre qui présente les modèles utilisés pour l'estimation du montant total payé pour chaque réclamation ouverte à la date d'évaluation. Les modèles proposés dans ce chapitre sont issus de l'article (Lopez et Milhaud, 2021) qui tente de modéliser le montant total payé directement ou par l'entremise de l'estimation du délai de fermeture.

La notation et les hypothèses de ces modèles seront présentées afin d'apporter des précisions relevant de l'application à la modélisation de la réserve individuelle. L'article précité propose que la modélisation du montant total payé soit conditionnelle au délai de fermeture. (Lopez et Milhaud, 2021) propose deux façons de traiter la censure qui donne lieu à deux familles de modèles : une famille qui nécessite une seule étape de modélisation et une famille qui nécessite deux étapes de modélisation. Forcément, il y a des considérations statistiques différentes pour chaque famille qui seront traitées dans ce chapitre et dans le prochain. Des exemples illustratifs des arbres de régressions pour l'estimation de la réserve seront fournis.

4.2 Notation et hypothèses

La notation des modèles d'arbres de régression nécessite des informations individuelles pour chaque réclamation d'un portefeuille. La notation utilisée est identique à celle des articles (Lopez *et al.*, 2016), (Lopez *et al.*, 2019) et (Lopez et Milhaud, 2021).

Le schéma présenté à la Figure 4.1 résume la notation pertinente pour une réclamation quelconque.

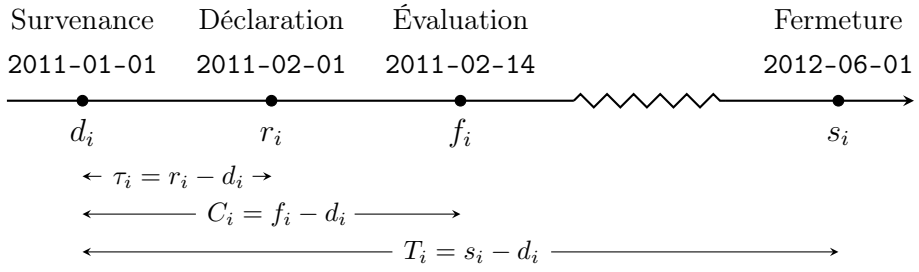


Figure 4.1: Exemple du développement d'une réclamation (avec indice i) ouverte à la date d'évaluation.

Le délai de fermeture de la i^{e} réclamation sera noté par T_i et censuré par le délai C_i dans le cas où s_i, la date de fermeture, n'est pas encore observée. La durée observée à la date d'évaluation de la réserve est définie par $Y_i = \inf(T_i, C_i)$.

Soit $\delta_i = \mathbb{1}(T_i \leq C_i)$ une indicatrice de la fermeture et M_i le paiement cumulatif à la fermeture. Si $\delta_i = 0$ (la réclamation i est ouverte à la date d'évaluation de la réserve), M_i est inconnu. En contrepartie, si $\delta_i = 1$ (la réclamation i est fermée à la date d'évaluation de la réserve), le montant M_i est connu et contribue à l'entraînement du modèle de régression pondéré. Certaines variables sont définies en fonction des données disponibles au moment du calcul de la réserve. Si les données ne sont pas de nature longitudinale ou si les paiements passés pour chaque réclamation ne sont pas disponibles, (Lopez *et al.*, 2016) propose que le montant cumulatif observé est $N_i = \delta_i M_i$. Si des paiements partiels ont été observés au moment du calcul de la réserve, (Lopez et Milhaud, 2021) propose

que N_i soit la somme des paiements partiels et que cette variable soit incluse parmi les covariables pour la régression de M_i. Ceci pose des difficultés puisque $M_i = N_i$ si $\delta_i = 1$, c'est-à-dire si la variable réponse est incluse comme variable explicative.

On rappelle que le mécanisme de censure impose des hypothèses sur la dépendance entre (T, C) et $(M, T, \mathbf{X})$. Afin que le modèle d'arbre de régression soit identifiable, (Lopez *et al.*, 2016) pose les hypothèses suivantes :

1. C_i est indépendant de (M_i, T_i) ;
2. $P(T_i \leq C_i \mid M_i, T_i, \mathbf{X}_i) = P(T_i \leq C_i \mid T_i)$.

Les auteurs supposent que la censure ne dépend pas de la variable réponse M, ni des covariables $\mathbf{X}$. Alors, les auteurs supposent que le mécanisme de censure est de type MCAR. L'estimateur Kaplan-Meier sans covariable de la fonction de survie de la durée jusqu'à la fermeture respecte les hypothèses proposées.

Le mécanisme de censure MAR demande que la variable indicatrice de censure δ_i dépende des covariables $\mathbf{X}_i$. Alors, la fonction de probabilité conditionnelle de censure devient $P(T_i \leq C_i \mid M_i, T_i, \mathbf{X}_i) = P(T_i \leq C_i \mid T_i, \mathbf{X}_i)$.

Une dernière hypothèse importante est l'ensemble de données avec lequel les IPCW sont calculés. (Lopez *et al.*, 2016) propose de calculer l'estimateur Kaplan-Meier en utilisant toutes les observations, non seulement les données d'entraînement. Les auteurs remarquent que la prédiction de la réserve et la prédiction de l'état (ouvert ou fermé) d'une réclamation sont deux problèmes disjoints et que le second peut s'éprouver instable si la taille de l'ensemble de données est petite.

4.3 Formulation des arbres de régression individuels

L'arbre de régression pondéré a été décrit au chapitre précédent sans avoir défini les poids w_i. La définition de ces derniers a été retardée jusqu'à ce chapitre. Puisqu'on suppose que le mécanisme de censure est de type MAR, l'estimation de l'espérance avec données complètes retourne un résultat biaisé. Les poids utilisés pour mettre en oeuvre

l'algorithme wCART sont $w_i = \dfrac{\delta_i}{\pi(\mathbf{X}_i)}$ où $\pi(\mathbf{X}_i) = P(T_i \leq C_i \mid \mathbf{X}_i)$ et $\delta_i = \mathbb{1}(T_i \leq C_i)$. Le poids w_i est positif si la réclamation i est fermée à la date d'évaluation et nul sinon. Le poids w_i est croissant en T_i. Rappelant la fonction de prédiction de l'algorithme de l'arbre de décision (3.4), toutes les réclamations contribuent à la prédiction par l'intermédiaire des poids, mais uniquement les montants des réclamations fermées sont inclus dans le calcul des moyennes.

Comme observé à la Section 3.4.1, la probabilité $\pi(\mathbf{X}_i)$ peut être estimée à l'aide de la régression logistique où

$$\widehat{\pi}_1(\mathbf{X}_i) = P(\Delta_i = 1 \mid \mathbf{X}_i) = (1 + \exp\{-\mathbf{X}_i\widehat{\boldsymbol{\beta}}\})^{-1} \tag{4.1}$$

ou encore par l'estimateur K-M de la fonction de survie du délai C_i. Utilisant l'estimateur K-M (sans covariables), la probabilité $\pi(\mathbf{X}_i)$ est estimée par

$$\widehat{\pi}_2(\mathbf{X}_i) = \widehat{S}_C(Y_i). \tag{4.2}$$

L'estimateur $\widehat{S}_C$ est la fonction de survie Kaplan-Meier de C et se calcule par

$$\widehat{S}_C(t) = \prod_{\{j|t_j \leq t\}} \left(1 - \frac{c_j}{n_j}\right) \tag{4.3}$$

avec c_j étant le nombre de censures observées après une durée t_j.

L'estimateur Kaplan-Meier est l'estimateur de l'espérance $E[\Delta_i(t)] = E[\mathbb{1}(C_i \geq t)] = P(C_i \geq t) = S_{C_i}(t)$ si C_i est une variable aléatoire continue.

À la prochaine section, chaque réclamation ouverte ($w_i = 0$) possède un arbre de régression estimé par l'algorithme wCART qui se prévaut des réclamations réclamations ouvertes et les réclamations fermées, mais les premiers sont seulement reconnus par les poids et non pour l'optimisation des noeuds des arbres.

Également, comme le montre (Lopez *et al.*, 2016), la validation croisée inclut également les poids w_i. En effet, le MSE du bloc de validation pour un facteur de pénalisation α est

$$\mathcal{V}(\alpha) = \sum_{i \in F_k} \frac{\delta_i}{\widehat{\pi}(\mathbf{X}_i)} [M_i - \widehat{f}^{K_\alpha}(\mathbf{X}_i; \boldsymbol{\upsilon}, \mathbf{w})]^2,$$

où F_k est l'ensemble d'indices du $k$ème bloc pour la validation croisée et $\widehat{f}^{K_\alpha}$ est l'arbre pénalisé avec K_α feuilles.

4.4 Construction des données individuelles

Les triangles de développement sont la structure de données qui permet de mettre en oeuvre les modèles collectifs. Pour les modèles individuels, la structure de données n'a pas encore été introduite. Elle est nécessaire avant l'estimation de la réserve individuelle, car la gestion des données d'un assureur peut avoir un grand impact sur la modélisation avec modèles granulaires et avec modèles individuels. Typiquement, un assureur collecte des informations explicatives pour chaque réclamation relative à la transaction, c'est-à-dire que pour chaque transaction, il y a une ligne qui contient les informations explicatives de l'assuré, de la police et de l'accident. Le Tableau 4.1 présente un exemple fictif de transactions.

Police	Réclamation	Réclamant	Accident	Déclaration	Transaction	Fermeture	Sous-couverture	Paiement
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-04-10	-	Medical	256,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-04-12	-	Medical	21,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-05-01	-	Expense	16,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-06-13	-	Expense	−16,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-07-30	-	Medical	0,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-07-30	-	Expense	−16,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-07-30	2010-07-30	Medical	−21,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2010-07-31	2010-07-31	Medical	−256,00

Tableau 4.1: Exemple de données transactionnelles.

Certaines transactions peuvent paraître superflues puisque les paiements de celles-ci peuvent être nuls ou même négatifs. Les transactions peuvent posséder des paiements nuls puisqu'ils reflètent le changement aux caractéristiques de la réclamation ou de l'assuré (qui ne sont pas affichées au Tableau 4.1). En contrepartie, les montants positifs sont des sommes qui doivent être versées par l'assureur et les montants négatifs reflètent des

sommes qui ont été versés par l'assureur. Puisqu'on s'intéresse au montant total à payer sur base individuelle, le montant M_i comprend la somme des paiements positifs et N_i la somme des paiements positifs observés avant l'évaluation de la réserve. Les données transactionnelles ne sont typiquement pas utilisées pour l'ajustement des modèles de provisionnement. Les données photographiques sont préférées. Celles-ci sont des données transactionnelles agrégées par période (p. ex. par mois, par trimestre, par semestre, ou par année) qui permettent de mitiger l'effet des changements aux caractéristiques du risque ainsi que les changements à l'administration des réclamations.

Par exemple, si la réclamation illustrée au Tableau 4.1 est agrégée par mois, pour chaque sous-couverture, on obtient une nouvelle base de données présentée dans le Tableau 4.2.

Police	Réclamation	Réclamant	Accident	Déclaration	Mois de Dév.	Fermeture	Sous-couverture	Paiement
P0001	R0001	C0001	2010-01-01	2010-02-06	4	-	Medical	277,00
P0001	R0001	C0001	2010-01-01	2010-02-06	5	-	Expense	16,00
P0001	R0001	C0001	2010-01-01	2010-02-06	6	-	Expense	0,00
P0001	R0001	C0001	2010-01-01	2010-02-06	7	2010-07-31	Medical	0,00

Tableau 4.2: Exemple de données photographiques (mensuelles) pour les paiements positifs.

Le mois de développement est calculé relativement à la date d'accident. On remarque qu'il y a des périodes manquantes pendant lesquelles il est évident que la réclamation est ouverte ou encore qu'il y a des changements aux covariables de la réclamation ou le réclamant. De plus, si l'objectif est de modéliser la réserve finale directement en utilisant un seul modèle par réclamation ouverte, il faut finalement regrouper les données selon la sous-couverture. Quant aux covariables qui ne sont pas illustrées dans le Tableau 4.2, les covariables observées à la transaction la plus récente pour chaque période de développement sont choisies.

Si l'on insère les mois de développement pendant lesquels il n'y a pas de transaction et si on impute les covariables manquantes de ces périodes, on obtient une troisième base de

données présentée dans le Tableau 4.3.

Police	Réclamation	Réclamant	Accident	Déclaration	Mois de Dév.	Fermeture	Paiement	Cumul
P0001	R0001	C0001	2010-01-01	2010-02-06	1	-	0,00	0,00
P0001	R0001	C0001	2010-01-01	2010-02-06	2	-	0,00	0,00
P0001	R0001	C0001	2010-01-01	2010-02-06	3	-	0,00	0,00
P0001	R0001	C0001	2010-01-01	2010-02-06	4	-	277,00	277,00
P0001	R0001	C0001	2010-01-01	2010-02-06	5	-	16,00	293,00
P0001	R0001	C0001	2010-01-01	2010-02-06	6	-	0, 00	293,00
P0001	R0001	C0001	2010-01-01	2010-02-06	7	2010-07-31	0,00	293,00

Tableau 4.3: Exemple de données photographiques (mensuelles) avec périodes manquantes insérées et paiements cumulatifs calculés.

Il est possible qu'il y ait plus d'un réclamant associé à une même réclamation (contrairement à l'exemple fictif ci-haut). La base de données $\mathcal{D}_{\mathcal{T}}$ est formée en fonction du nombre de périodes désirées, la durée d'une période et la date d'évaluation de la réserve. De plus, puisque les réclamations ouvertes au moment de l'évaluation de réserve n'auront aucun impact sur la prédiction du montant total payé individuel, notons $\mathcal{D}_{\mathcal{T}}^{j} = \{(\mathbf{X}_i, y_i) \in \mathcal{D}_{\mathcal{T}} \mid \delta_i = j\}$ comme le sous-ensemble de données de $\mathcal{D}_{\mathcal{T}}$ qui comprend seulement les réclamations fermées ($j = 1$) ou ouvertes ($j = 0$). Les données les plus récentes relativement à la date d'évaluation sont choisies et l'indicatrice δ_i de fermeture est calculée. Par exemple, si f_i = 2010-03-01, l'information explicative disponible à l'évaluation de la réserve sont celles du troisième mois de développement du tableau 4.3 et $\delta_i = 0$. Par contre, si f_i = 2010-07-01, les informations explicatives du septième mois sont sélectionnées et $\delta_i = 1$.

4.5 Modèles individuels à une étape

Jusqu'à présent, la censure du délai de fermeture a été prise en compte par l'inclusion des poids qui dépendent de la probabilité $P(T_i \leq C_i)$, mais il n'a pas encore été expliqué comment utiliser l'algorithme wCART et les poids IPCW afin de formuler un modèle

individuel pour la réserve. Les articles (Lopez *et al.*, 2016), (Lopez *et al.*, 2019), (Lopez et Milhaud, 2021) proposent plusieurs modèles et les modèles sélectionnés pour le présent mémoire reflètent les deux derniers articles. Les régressions proposées peuvent être de plusieurs types : des régressions ordinaires (1), des régressions avec biais de sélection (2) ou des régressions tronquées (3). La régression ordinaire est la modélisation de l'espérance (ou d'un quantile) conditionnellement au vecteur de covariables. La régression avec biais de sélection est semblable à la régression ordinaire, mais où l'ensemble de données d'entraînement est assujetti à une certaine condition qui ne dépend pas de la variable réponse. Finalement, la régression tronquée est semblable à la régression avec biais de sélection, mais où la condition d'inclusion dépend de la variable réponse.

Les modèles à une étape sont des modèles de régression avec biais de sélection. Pour chaque réclamation ouverte ($\delta_i = 0$), les modèles à une étape tentent de modéliser l'espérance $E[M_i \mid \mathbf{X}_i, T_i > y]$ où y est la durée observée de la réclamation i. L'article (Lopez et Milhaud, 2021) propose d'exprimer l'espérance conditionnelle $E[M_i \mid \mathbf{X}_i, T_i > y]$ comme

$$E[M_i \mid \mathbf{X}_i, T_i > y] = \frac{E[M_i \mathbb{1}(T_i > y) \mid \mathbf{X}_i]}{E[\mathbb{1}(T_i > y) \mid \mathbf{X}_i]}.$$

Cette reformulation demande d'exécuter l'algorithme wCART une fois pour estimer le numérateur et une deuxième fois pour estimer le dénominateur. L'espérance de gauche (stratégie **A1**) est plus simple et efficace à modéliser comme remarque (Lopez *et al.*, 2019). Il y a une grande instabilité dans le rapport d'espérances de droite (stratégie **A2**), car il dépend du nombre des réclamations fermées qui respectent la condition $T_i > y$.

Soit $\mathcal{D}_{\mathcal{T}}(y) = \{(\mathbf{X}_i, y_i) \in \mathcal{D}_{\mathcal{T}} \mid T_i > y\}$ l'ensemble d'entraînement pour la réclamation ouverte ayant la durée observée y. On peut partitionner l'ensemble $\mathcal{D}_{\mathcal{T}}(y) = \mathcal{D}^0_{\mathcal{T}}(y) \cup \mathcal{D}^1_{\mathcal{T}}(y)$ où $\mathcal{D}^j_{\mathcal{T}}(y) = \{(\mathbf{X}_i, y_i) \in \mathcal{D}_{\mathcal{T}} \mid T_i > y, \delta_i = j\}$, $j = 0, 1$. Une augmentation en y cause une diminution de la taille de $\mathcal{D}_{\mathcal{T}}(y)$ et si $|\mathcal{D}^1_{\mathcal{T}}(y)| = 0$, on remarque que la fonction de prédiction donnée par l'Équation (3.3) est indéterminée. De plus, si $|\mathcal{D}^0_{\mathcal{T}}(y)| = 0$, le dénominateur de l'Équation (3.3) est trivial. Puisque chaque réclamation ouverte est attribuée un ensemble d'entraînement propre, il y a un intérêt à optimiser ses

hyperparamètres. Il possible d'optimiser certains hyperparamètres des arbres avec la fonction `train` de la librairie `caret`. Afin d'assurer la convergence des modèles individuels pour $|\mathcal{D}_{\mathcal{T}}^1(y)|$ suffisamment grand, on élimine les réclamations ouvertes dont le nombre de réclamations fermées ($w_i > 0$) est inférieur à un seuil entier c. À l'extrême, si $|\mathcal{D}_{\mathcal{T}}^1(y)| = 1$, la fonction `rpart` retourne un arbre de régression trivial selon la stratégie **A1**. Également, si $\mathcal{D}_{\mathcal{T}}(y) = \mathcal{D}_{\mathcal{T}}^1(y)$ ou $\mathcal{D}_{\mathcal{T}}(y) = \mathcal{D}_{\mathcal{T}}^0(y)$, le dénominateur de **A2** est trivial et lance une erreur vis-à-vis la fonction `rpart`. Les algorithmes d'ajustement des modèles à une étape (stratégies **A1** et **A2**) sont résumés par les Figures 4.2 et 4.3.

Il est important de remarquer que la définition de l'ensemble d'entraînement $\mathcal{D}_{\mathcal{T}}(y)$ est modifiable. En particulier, il est possible remplacer la condition stricte ($T_i > y$) par une condition inclusive, c'est-à-dire ($T_i \geq y$). L'article (Lopez *et al.*, 2016) propose une censure bivariée en proposant la condition ($M_i > m, T_i > y$) mais demande l'utilisation d'une fonction de survie bivariée en conséquence pour le calcul des poids.

4.6 Modèles individuels à deux étapes

Les modèles individuels à deux étapes sont des modèles d'arbre de régression pour l'espérance $E[M_i \mid \mathbf{X}_i, T_i]$. On note que les modèles de cette classe demeurent individuels, mais pour lequel il y a un seul modèle pour le montant final. Ce modèle peut être directement ajusté à l'aide de l'algorithme wCART présenté à la Figure 3.1 avec les réclamations fermées. Puisque T_i est censuré, il reste à estimer celui-ci pour chaque réclamation ouverte. (Lopez et Milhaud, 2021) propose d'ajuster des modèles à une étape pour le délai de fermeture avec arbres de régression individuels (1) et ensuite d'estimer le montant total payé avec un arbre qui prend comme covariable le délai de fermeture (2). On obtient la prédiction $\widehat{M_i}$ en insérant la prédiction $\widehat{T_i}$. Les modèles à deux étapes seront englobés par la stratégie appelée **B**.

En réalité, malgré qu'il est naturel d'appliquer les arbres de régression pour prédire le délai de fermeture, (Lopez et Milhaud, 2021) propose plusieurs modèles alternatifs. Seulement les arbres régression seront traités dans ce mémoire.

Algorithme 3 : Estimation de la réserve avec modèles à une étape (**A1**)

Input : Ensemble d'hyperparamètres $\mathcal{P}_{\boldsymbol{\upsilon}}$, poids $\mathbf{w}$ et seuil du nombre de réclamations fermées c

Output : Arbres de régression $\widehat{f}_1, \ldots, \widehat{f}_m$ (une pour chaque réclamation ouverte) et la réserve totale $\widehat{R}$

Data : Base de données $\mathcal{D}_{\mathcal{T}}$

1. Éliminer les réclamations ouvertes dont $|\mathcal{D}^1_{\mathcal{T}}(y_j)| < c$ pour chaque réclamation ouverte $j = 1, \ldots, m$.

2. Retourner les durées observées $y_1, \ldots, y_m$ pour les réclamations ouvertes restantes.

3. **foreach** $j \in \{1, \ldots, m\}$ *pour chaque indice de réclamation ouverte* **do**
   - Optimiser l'arbre avec hyperparamètres $\boldsymbol{\upsilon}^* \in \mathcal{P}_{\boldsymbol{\upsilon}}$ et retourne les hyperparamètres optimaux $\boldsymbol{\upsilon}_j$;
   - Appliquer l'algorithme wCART avec réponse M_i et poids $w_i = \dfrac{\delta_i}{\widehat{\pi}(\mathbf{X}_i)}$ l'aide de l'ensemble d'entraînement $\mathcal{D}^1_{\mathcal{T}}(y_j)$;
   - Élaguer l'abre de la réclamation avec indice j avec un paramètre de complexité sélectionné par l'utilisateur afin d'obtenir $\widehat{f}_j(\cdot\,; \boldsymbol{\upsilon}_j, \mathbf{w})$.

4. Retourner les fonctions $\widehat{f}_1(\cdot\,; \boldsymbol{\upsilon}_1, \mathbf{w}), \ldots, \widehat{f}_m(\cdot\,; \boldsymbol{\upsilon}_m, \mathbf{w})$ estimées.

5. Calculer la réserve globale

$$\widehat{R} = \sum_{j=1}^{m} \widehat{R}_j = \begin{cases} \sum_{j=1}^{m} \widehat{f}_j(\mathbf{X}_j; \boldsymbol{\upsilon}_j, \mathbf{w}), \text{ si les réclamations n'ont pas de paiements partiels}\,; \\ \sum_{j=1}^{m} [\widehat{f}_j(\mathbf{X}_j; \boldsymbol{\upsilon}_j, \mathbf{w}) - N_j], \text{ sinon.} \end{cases}$$

Figure 4.2: Étapes du calcul de la réserve avec modèles individuels à une étape (**A1**).

Algorithme 4 : Estimation de la réserve avec modèles à une étape (**A2**)

Input : L'ensemble d'hyperparamètres $\mathcal{P}_{\boldsymbol{\upsilon}}$ pour les arbres du numérateur,
L'ensemble d'hyperparamètres $\mathcal{P}_{\boldsymbol{\xi}}$ pour les arbres du dénominateur,
les poids $\mathbf{w}$ et seuil du nombre de réclamations fermées c

Output : Arbres de régression $\widehat{f}_1, \ldots, \widehat{f}_m$ pour le numérateur, $\widehat{g}_1, \ldots, \widehat{g}_m$ pour le dénominateur et la réserve totale $\widehat{R}$

Data : Base de données $\mathcal{D}_{\mathcal{T}}$

1. Éliminer les réclamations ouvertes dont $|\mathcal{D}^1_{\mathcal{T}}(y_j)| < c$ pour chaque réclamation ouverte $j = 1, \ldots, m$.
2. Retourner les durées observées $y_1, \ldots, y_m$ pour les réclamations ouvertes restantes.
3. **foreach** $j \in \{1, \ldots, m\}$ *pour chaque indice de réclamation ouverte* **do**
   - Optimiser l'arbre du numérateur avec hyperparamètres $\boldsymbol{\upsilon}^* \in \mathcal{P}_{\boldsymbol{\upsilon}}$ et retourne les hyperparamètres optimaux $\boldsymbol{\upsilon}_j$;
   - Optimiser l'arbre du dénominateur avec hyperparamètres $\boldsymbol{\xi}^* \in \mathcal{P}_{\boldsymbol{\xi}}$ et retourne les hyperparamètres optimaux $\boldsymbol{\xi}_j$;
   - Appliquer l'algorithme wCART avec poids $w_i = \dfrac{\delta_i}{\widehat{\pi}(\mathbf{X}_i)}$ au numémateur $M_i \mathbb{1}(T_i > y_j)$ l'aide de l'ensemble d'entraînement $\mathcal{D}^1_{\mathcal{T}}$;
   - Appliquer l'algorithme wCART avec poids $w_i = \dfrac{\delta_i}{\widehat{\pi}(\mathbf{X}_i)}$ au dénominateur $\mathbb{1}(T_i > y_j)$ l'aide de l'ensemble d'entraînement $\mathcal{D}^1_{\mathcal{T}}$;
   - Élaguer les arbres du numérateur et du dénominateur avec indices j avec un paramètre de complexité sélectionné par l'utilisateur.
4. Retourner les fonctions estimées $\widehat{f}_1(\cdot\,; \boldsymbol{\upsilon}_1, \mathbf{w}), \ldots, \widehat{f}_m(\cdot\,; \boldsymbol{\upsilon}_m, \mathbf{w})$ et $\widehat{g}_1(\cdot\,; \boldsymbol{\xi}_1, \mathbf{w}), \ldots, \widehat{g}_m(\cdot\,; \boldsymbol{\xi}_m, \mathbf{w})$.
5. Calculer la réserve globale

$$\widehat{R} = \sum_{j=1}^{m} \widehat{R}_j = \begin{cases} \sum_{j=1}^{m} \frac{\widehat{f}_j(\mathbf{X}_j; \boldsymbol{\upsilon}_j, \mathbf{w})}{\widehat{g}_j(\mathbf{X}_j; \boldsymbol{\xi}_j, \mathbf{w})}, \text{ si les réclamations n'ont pas de paiements partiels ;} \\ \sum_{j=1}^{m} \left[\frac{\widehat{f}_j(\mathbf{X}_j; \boldsymbol{\upsilon}_j, \mathbf{w})}{\widehat{g}_j(\mathbf{X}_j; \boldsymbol{\xi}_j, \mathbf{w})} - N_j \right], \text{ sinon.} \end{cases}$$

Figure 4.3: Étapes du calcul de la réserve avec modèles individuels à une étape (**A2**).

4.6.1 Comparaison des modèles à une étape et des modèles à deux étapes

Les modèles de la stratégie **A1**, **A2** et **B** ont l'objectif commun de prédire le montant total payé pour chaque réclamation ouverte au moment du calcul de la réserve mais ont différentes interprétations.

La distinction entre les stratégies **A1** et **A2** repose sur deux différents traitements des données en présence du biais de sélection induit par la condition $T_i > y$. La stratégie **A1** tente de prédire l'espérance où seulement les réclamations qui ont le délai de fermeture est supérieures au délai observé (y) sont conservés. La stratégie **A2** alors tente de prédire la même espérance où les réclamations qui ont des délais de fermeture inférieurs au délai observé sont censurées et au lieu d'éliminer les données qui ne respectent la condition de sélection, on suppose que les montants ne sont pas encore réalisés et que ces montants totaux payés sont posés à zéro. Dit autrement, le biais de sélection de la stratégie **A1** est exprimé par l'entremise d'une censure qui doit être corrigée par la modélisation de la probabilité de la même censure. Comme illustration arithmétique,

$$\frac{1}{\sum_{i=1}^{n} w_i \mathbb{1}(T_i > y)} \sum_{i=1}^{n} w_i m_i \mathbb{1}(T_i > y) = \left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i \mathbb{1}(T_i > y)} \right) \left[\frac{1}{\sum_{i=1} w_i} \sum_{i=1}^{n} w_i m_i \mathbb{1}(T_i > y) \right]. \tag{4.4}$$

Il est évident par la formule ci-haut que la probabilité de $T_i > y$ a un grand impact sur la prédiction du montant total payé qui peut cause une instabilité dans la prédiction, mais l'élagage des arbres peuvent révéler des différences entre les deux approches puisqu'elles capturent la variabilité de la variable réponse selon deux interprétations du même phénomène. Puisque la stratégie **A1** ne pose pas les montants totaux payés à zéro, il est moins probable que cette stratégie supprime des branches. Par contre, le numérateur et le dénominateur de la stratégie **A2** risquent d'avoir beaucoup de branches retirées s'il y a plusieurs réclamations qui ne respectent pas la condition $T_i > y$.

La stratégie **B** est la plus a comprendre puisqu'elle repose sur un seul arbre avec une variable explicative censurée pour lequel la censure est corrigée par les poids IPCW. Elle suppose que la relation entre les covariables et le montant total payé est la même pour

Réclamation	Accident	Déclaration	Femeture	Legal	f_i	τ_i	C_i	T_i	Y_i	δ_i	M_i
CLM-1	1993-08-01	1993-09-01	1993-10-01	No	1996-06-01	32	1080	62	62	1	87,75
CLM-2	1993-12-01	1994-01-01	1994-02-01	No	1996-06-01	32	958	63	63	1	353,62
CLM-3	1994-01-01	1994-01-01	1994-02-01	Yes	1996-06-01	1	927	32	32	1	688,83
CLM-4	1994-04-01	1994-04-01	1994-05-01	No	1996-06-01	1	837	31	31	1	172,80
CLM-5	1994-08-01	1994-09-01	1994-09-01	No	1996-06-01	32	715	32	32	1	43,29
CLM-6	1994-12-01	1995-01-01	1995-01-01	Yes	1996-06-01	32	593	32	32	1	2 915,43
CLM-7	1995-02-01	1995-02-01	1995-03-01	No	1996-06-01	1	531	29	29	1	496,16
CLM-8	1995-05-01	1995-05-01	1995-07-01	Yes	1996-06-01	1	442	62	62	1	1 593,20
CLM-9	1995-08-01	1995-08-01	1995-09-01	No	1996-06-01	1	350	32	32	1	122,87
CLM-10	1995-11-01	1995-11-01	1995-12-01	No	1996-06-01	1	258	31	31	1	1 943,10
CLM-11	1996-01-01	1996-03-01	1996-03-01	No	1996-06-01	61	197	61	61	1	92,58
CLM-12	1996-06-01	1996-07-01	1996-08-01	No	1996-06-01	31	45	62	45	0	397,49

Tableau 4.4: Données individuelles pour la démonstration de l'ajustement des arbres de régression.

toutes réclamations ouvertes et que l'impact que le réclamation soit ouverte est totalement absorbé par la covariable censurée T_i. J'estime que les prédictions de la stratégie **B** sous-estimeront la réserve globale par le fait qu'il peut exister des dépendances entre T_i et les autres covariables dont l'arbre ne prendra pas compte.

4.7 Exemples

Dans cette section, on présente des exemples pour les modèles à une étape (**A1** et **A2**) et à deux étapes avec des données individuelles. Les données en question sont un sous-ensemble de 12 réclamations de l'ensemble `ausautoBI8999` provenant d'un assureur automobile australien. La base de données est encore une fois issue de la librairie `CASdatasets`. La Figure 4.4 illustre les réclamations choisies. Certaines informations quant aux 12 réclamations sont modifiées afin d'assurer l'existence des prédictions.

Il y a une réclamation ouverte à la date d'évaluation de la réserve (qui normalement est la même pour chaque ligne). Les variables explicatives qui seront utilisées sont la

présence d'un avocat au dossier (Legal) et le délai de déclaration (τ_i). Les poids IPCW seront calculés à l'aide de l'estimateur K-M de la fonction de survie à l'aide de la fonction `prodlim` de la librairie `survival`. En utilisant l'Équation (4.3), on peut calculer l'estimateur K-M de la fonction de survie de C.

Y_j	n_j	d_j	c_j	$\widehat{S}_C(Y_j)$
29	12	1	0	$1 - \frac{0}{12} = 1$
31	11	2	0	$1(1 - \frac{0}{11}) = 1$
32	9	4	0	$1^2(1 - \frac{0}{9}) = 1$
45	5	0	1	$1^3(1 - \frac{1}{5}) = \frac{4}{5}$
61	4	1	0	$1^3(\frac{4}{5})(1 - \frac{0}{4}) = \frac{4}{5}$
62	3	2	0	$1^4(\frac{4}{5})(1 - \frac{0}{3}) = \frac{4}{5}$
63	1	1	0	$1^5(\frac{4}{5})(1 - \frac{0}{1}) = \frac{4}{5}$

Tableau 4.5: Données de survie de C_i pour les réclamations illustratives.

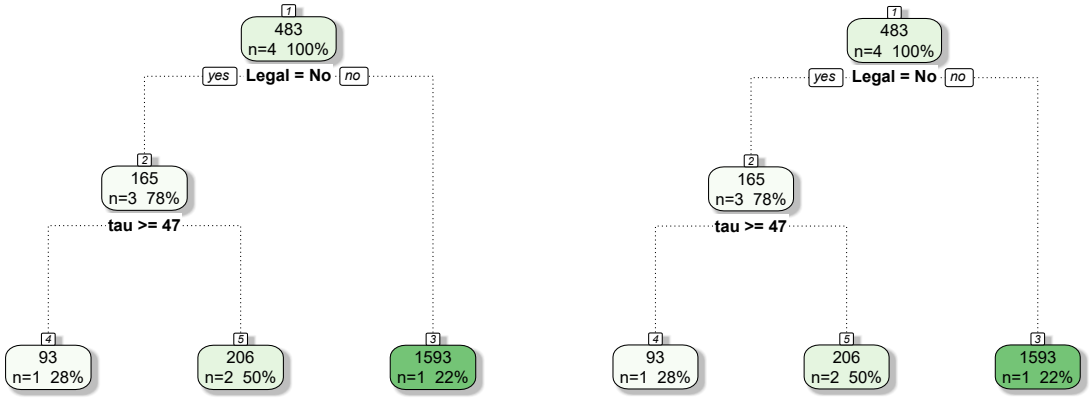
Les principaux éléments sont repris dans le Tableau 4.5. Dans ce dernier, les poids sont

$$w_i = \begin{cases} 1, & i = \texttt{CLM-3}, \texttt{CLM-4}, \texttt{CLM-5}, \texttt{CLM-6}, \texttt{CLM-7}, \texttt{CLM-9}, \texttt{CLM-10}; \\ \frac{5}{4}, & i = \texttt{CLM-1}, \texttt{CLM-2}, \texttt{CLM-8}, \texttt{CLM-11}; \\ 0, & i = \texttt{CLM-12}. \end{cases}$$

Les hyperparamètres des arbres de régression sont le nombre minimal d'observations par feuille, le nombre minimal d'observations pour une séparation, la profondeur maximale de l'arbre et le paramètre de complexité (α) pour l'élagage de l'arbre. Utilisant la fonction `rpart` de la librairie du même nom, ces derniers sont nommés `minbucket`, `minsplit`, `maxdepth` et `cp` respectivement. Pour les exemples suivants, `minbucket` = 1, `minsplit` = 3, `maxdepth` = 5 et `cp` = 0.

Les données d'entraînement $\mathcal{D}^1_{\mathcal{T}}(45)$ comprennent les réclamations `CLM-1`, `CLM-2`, `CLM-8` et `CLM-11` par la définition de $\mathcal{D}^1_{\mathcal{T}}(y)$ et le fait que $\tau_i < C_i$ pour $i = 1, \ldots, 11$. Si on

applique l'algorithme wCART sur $\mathcal{D}_{\mathcal{T}}^{1}(45)$ avec les hyperparamètres définis au paragraphe précédent, on obtient les arbres non élagué et élagué présentés à la Figure 4.4.

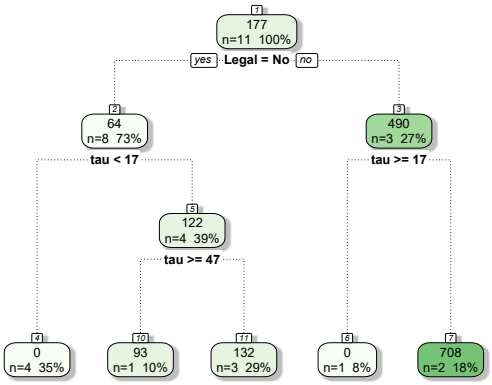


(a) Arbre non élagué

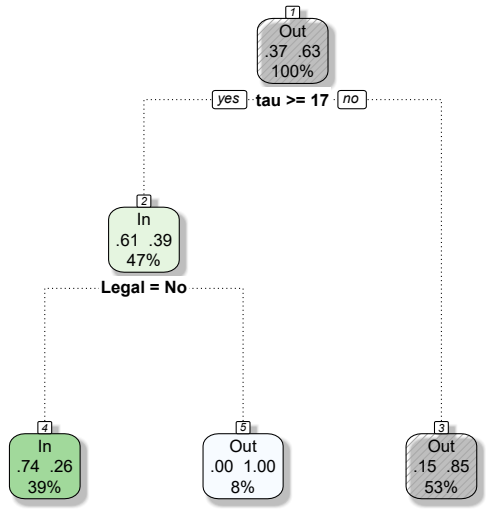
(b) Arbre élagué

Figure 4.4: Exemples d'arbres de régression pour $\widehat{M}_{12}$ de la stratégie **A1**.

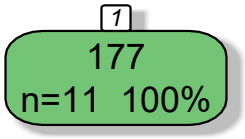
Puisqu'il n'y a pas de paiements partiels qui s'inscrivent dans les données, le montant total prédit pour la réclamation `CLM-12` est $\widehat{M}_{12} = 205{,}91$ selon la stratégie **A1**. Ensuite, pour la stratégie **A2**, il y aura 2 arbres non élagués et 2 arbres élagués pour le numérateur et le dénominateur respectivement. Ces derniers sont présentés à la Figure 4.5.



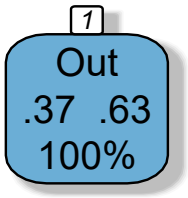
(a) Arbre non élagué du numérateur



(b) Arbre non élagué du dénominateur



(c) Arbre élagué du numérateur



(d) Arbre élagué du dénominateur

Figure 4.5: Exemples d'arbres de régression pour $\widehat{M}_{12}$ de la stratégie **A2** .

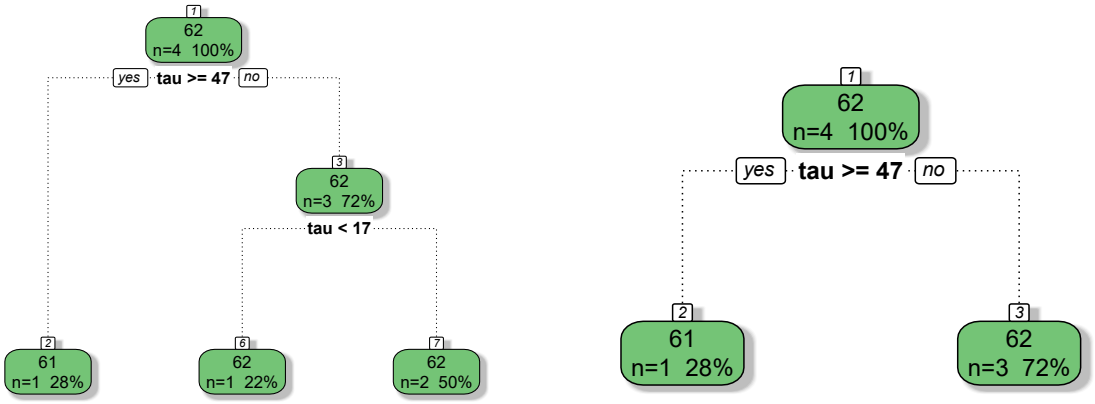
Ainsi, la prédiction du montant total pour la réclamation `CLM-12` est $\widehat{M}_{12} = \dfrac{177{,}3251}{0{,}3673} = 482{,}71$. Selon la Formule 4.4, il s'ensuit qu'en effet les deux stratégies sont équivalentes. En effet,

$$\frac{1}{\sum_{i=1} w_i \mathbb{1}(T_i > 45)} \sum_{i=1} w_i m_i \mathbb{1}(T_i > 45) = \frac{87{,}75(1{,}25) + 353{,}62 + 1\ 593{,}20 + 92{,}58(1{,}25)}{1{,}25(2) + 2(1)} = 482{,}71.$$

Par contre, l'élagage élimine toutes les branches des arbres du numérateur et du dénomi-

nateur.

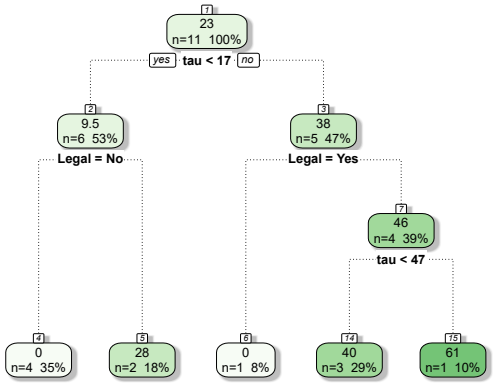
Pour les modèles à deux étapes, il faut estimer les modèles à une étape pour le délai de fermeture T_i. Ces arbres sont présentés à la Figure 4.6.



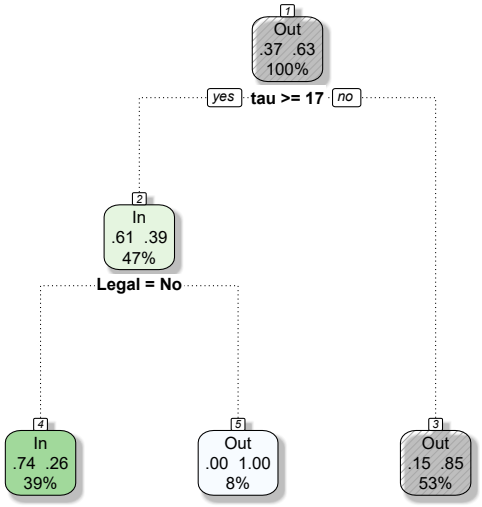
(a) Arbre non élagué (b) Arbre élagué

Figure 4.6: Exemples d'arbres de régression pour $\widehat{T}_{12}$ de la stratégie **A1**.

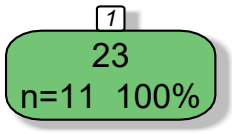
D'après la stratégie **A1**, la prédiction du délai de fermeture de la réclamation `CLM-12` est $\widehat{T}_{12} = 62{,}31$.



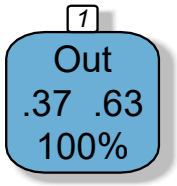
(a) Arbre non élagué du numérateur



(b) Arbre non élagué du dénominateur



(c) Arbre élagué du numérateur



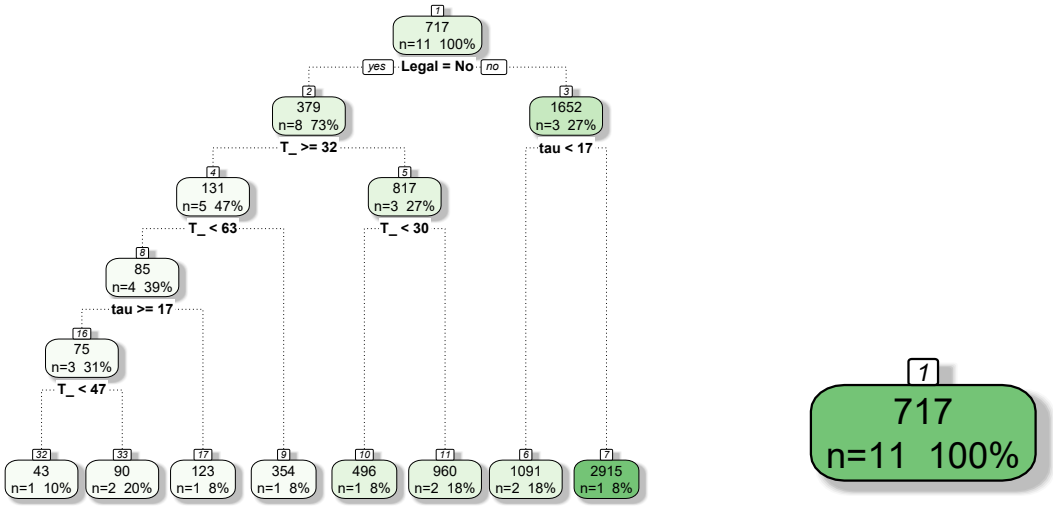
(d) Arbre élagué du dénominateur

Figure 4.7: Exemples d'arbres de régression pour $\widehat{T}_{12}$ de la stratégie **A2**.

Suivant la stratégie **A2**, la prédiction du délai de fermeture de la réclamation `CLM-12` est également $\widehat{T}_{12} = \dfrac{22{,}7551}{0{,}3673} = 61{,}9\overline{4}$. Les arbres pour y parvenir sont présentés à la Figure 4.7. La variable T_i est moins sur-dispersée que la variable M_i, alors il en découle que les arbres pour les délais de fermeture seront moins profonds. Rappelant l'équivalence de la Formule 4.4, si les stratégies **A1** et **A2** donnent des arbres à une seule feuille, les prédictions de stratégies précitées sont identiques. Comme mentionné à la Section 4.6.1,

puisque la stratégie **A2** tente de séparer la modélisation de la variable réponse avec avec biais de sélection (stratégie **A1**) en deux modèles, un traitant la probabilité d'inclusion dans l'ensemble d'entraînement et le deuxième deuxième la modélisation de la variable réponse assujettie à une censure, il est plus probable que l'élagage ait un effet plus prononcé pour la stratégie **A2** comparativement à la stratégie **A1**.

Enfin, pour obtenir la prédiction $\widehat{M}_{12}$, il suffit d'évaluer l'espérance $E[M_{12} \mid \mathbf{X}_{12}, T_{12}]$ avec la prédiction $\widehat{T}_{12}$.



(a) Arbre non élagué

(b) Arbre élagué

Figure 4.8: Exemples d'arbres de régression pour $\widehat{M}_{12}$ selon la stratégie **B**.

Les arbres de régression pour $\widehat{M}_{12}$ selon la stratégie **B** sont présentés à la Figure 4.8. Ainsi, peu importe la prédiction $\widehat{T}_{12}$, la prédiction du montant total payé est $\widehat{M}_{12} = 716{,}81$.

CHAPITRE V

GLM PONDÉRÉS DE LA RÉSERVE INDIVIDUELLE

5.1 Introduction

Dans le présent chapitre, on fera une introduction aux propriétés de l'estimateur MV et de la fonction de vraisemblance pour données assujetties à un biais de sélection. En introduisant les poids IPCW, il est possible de retrouver la vraisemblance originale. Suivant la même idée que le modèle individuel à une étape de la Section 4.5, il est possible de développer un modèle équivalent à l'aide d'un GLM pondéré par les poids IPCW (wGLM). De plus, un modèle à deux étapes peut être mis en oeuvre par l'application du modèle de régression proposé par (Matsouaka et Atem, 2020) qui permet l'ajustement d'un GLM qui comprend une covariable censurée. Ce modèle est une régression logistique qui a été appliquée en analyse de survie, mais les démarches de l'article susmentionné ne sont pas uniques et peuvent s'adapter à la régression de Poisson sur-dispersée pour la prédiction de la réserve individuelle.

Cette approche présente deux avantages importants pour la modélisation individuelle de la réserve :

- les modèles GLM sont plus simples d'interprétation que les arbres de régression ;
- les IPCW sont suffisamment flexibles pour accommoder plusieurs types de biais de sélection et peuvent être utilisés au sein plusieurs modèles de régression.

Les modèles GLM peuvent s'ajouter aux approches qui permettent de modéliser la réserve

individuelle à l'aide des IPCW. L'estimation à l'aide des réclamations fermées seulement donne lieu à des erreurs d'inférence.

Comme dans les précédents chapitres, il y aura des exemples illustratifs afin d'illustrer les modèles. Afin d'évaluer l'impact de l'utilisation des poids IPCW, on va comparer les régressions pondérées avec les régressions naïves utilisant seulement les réclamations fermées à partir de bases de données simulées qui varient en taille et dans la proportion de réclamations ouvertes au moment du calcul de la réserve.

5.2 Précisions sur les propriétés du maximum de vraisemblance

Dans cette section, les propriétés de l'estimateur du maximum de vraisemblance sont revisées pour la fonction de vraisemblance pondérée par les IPCW. On verra que l'inclusion des IPCW permet le même comportement asymptotique de l'estimateur du MV classique. La convergence de l'estimateur par MV_{IPCW} et l'efficacité de l'estimateur par MV_{IPCW} seront vérifiées.

Enfin, dans le présent mémoire, la régression de Poisson sur dispersée est utilisée pour l'estimation de la réserve individuelle pour chaque réclamation ouverte.

5.2.1 Définitions

La fonction de vraisemblance pondérée est

$$\mathcal{L}^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}^{w_i}(y_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \boldsymbol{\theta})^{w_i}.$$

Le vecteur $\mathbf{w} = (w_1, \ldots, w_n)$ comprend des poids non négatifs. On obtient directement la fonction de log-vraisemblance pondérée :

$$\ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})^{w_i}\} = \sum_{i=1}^{n} w_i \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}.$$

Les hypothèses de la fonction de vraisemblance pondérée sont données dans l'Annexe A.

5.2.2 Propriétés

Il peut être démontré que l'estimateur MV_{IPCW} est convergent et asymptotiquement normal. Cette démonstration est tirée de (Matsouaka et Atem, 2020) qui montre que l'espérance de la fonction de score pondérée par les IPCW est égale à $\mathbf{0}$. Par raisonnement semblable, la matrice d'information de Fisher pondérée par les IPCW est identique (en espérance) à celle non pondérée. Puisque les IPCW sont définis par $w_i = \dfrac{\Delta_i}{\pi_i(\mathbf{X}_i)}$ pour $i = 1, \ldots, n$ et $E[\Delta_i \mid \mathbf{X}_i] = \pi_i(\mathbf{X}_i)$. L'espérance de la fonction de score $\mathcal{U}^{\mathbf{w}}(\boldsymbol{\theta}) = \nabla \ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y})$ est

$$
\begin{aligned}
E[\mathcal{U}^{\mathbf{w}}(\boldsymbol{\theta})] &= E[\nabla \ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y})] \\
&= E\left[\left(\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} w_i \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right)_{j=1,\ldots,p}\right] \\
&= \left(E\left[\sum_{i=1}^{n} w_i \frac{\partial}{\partial \theta_j} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]\right)_{j=1,\ldots,p} \\
&= \left(\sum_{i=1}^{n} E\left[w_i \frac{\partial}{\partial \theta_j} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]\right)_{j=1,\ldots,p} \\
&= \left(\sum_{i=1}^{n} E\left[E[w_i \mid \mathbf{X}_i] \frac{\partial}{\partial \theta_j} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]\right)_{j=1,\ldots,p} \\
&= \left(\sum_{i=1}^{n} E\left[E[\Delta_i \mid \mathbf{X}_i] \pi_i(\mathbf{X}_i)^{-1} \frac{\partial}{\partial \theta_j} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]\right)_{j=1,\ldots,p} \\
&= \left(\sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta_j} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]\right)_{j=1,\ldots,p} \\
&= \left(\sum_{i=1}^{n} E\left[\frac{\partial}{\partial \theta_j} \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]\right)_{j=1,\ldots,p} \\
&= \mathbf{0}.
\end{aligned}
$$

Ensuite, la matrice d'information de Fisher est

$$
\mathcal{I}^{\mathbf{w}}(\boldsymbol{\theta})_{(j,k)} = -E\left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \sum_{i=1}^{n} w_i \log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right]
$$

$$
\begin{aligned}
&= -E\left[\sum_{i=1}^{n} w_i \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right] \\
&= -E\left[\sum_{i=1}^{n} E\left[w_i \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\} \mid \mathbf{X}_i\right]\right] \\
&= -E\left[\sum_{i=1}^{n} E\left[w_i \mid \mathbf{X}_i\right] \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right] \\
&= -E\left[\sum_{i=1}^{n} E\left[\Delta_i \mid \mathbf{X}_i\right] \pi_i(\mathbf{X}_i)^{-1} \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right] \\
&= -E\left[\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log\{f_{Y_i}(y_i \mid \boldsymbol{\theta})\}\right] \\
&= \mathcal{I}(\boldsymbol{\theta})_{(j,k)}.
\end{aligned}
$$

Il s'ensuit que $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1})$ et que $\widehat{\boldsymbol{\theta}}$ est convergent.

5.3 Théorie de la vraisemblance avec biais de sélection

Soit la $f_{M,\mathbf{X}}(m, \mathbf{x})$ la vraisemblance d'une observation du montant total payé et ses covariables. La vraisemblance d'une observation du montant total (censuré) et ses covariables est $f_{\Delta M,\mathbf{X},\Delta}(\delta m, \mathbf{x}, \delta)$.

Développant la fonction de vraisemblance,

$$
\begin{aligned}
f_{\Delta M,\mathbf{X},\Delta}(\delta m, \mathbf{x}, \delta) &= f_{M,\mathbf{X},\Delta}(m, \mathbf{x}, \delta = 1)^{\delta} f_{\mathbf{X},\Delta}(\mathbf{x}, \delta = 0)^{1-\delta} \\
&= [f_{M|\mathbf{X},\Delta}(m \mid \mathbf{x}, \delta = 1) f_{\Delta|\mathbf{x}}(\delta = 1 \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})]^{\delta} [f_{\Delta|\mathbf{X}}(\delta = 0 \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})]^{1-\delta} \\
&= [f_{M|\mathbf{X}}(m \mid \mathbf{x}) f_{\Delta|\mathbf{x}}(\delta = 1 \mid \mathbf{x})]^{\delta} [f_{\Delta|\mathbf{X}}(\delta = 0 \mid \mathbf{x})]^{1-\delta} f_{\mathbf{X}}(\mathbf{x}) \\
&\propto [f_{M|\mathbf{X}}(m \mid \mathbf{x}) f_{\Delta|\mathbf{x}}(\delta = 1 \mid \mathbf{x})]^{\delta} [f_{\Delta|\mathbf{X}}(\delta = 0 \mid \mathbf{x})]^{1-\delta}
\end{aligned}
$$

par l'hypothèse MAR qui suppose que $\Delta \perp\!\!\!\perp M \mid \mathbf{X}$. Alors, $f_{M|\mathbf{X}}(m \mid \mathbf{x}) = \mathcal{L}(\boldsymbol{\beta} \mid \mathbf{x})$ et $f_{\Delta|\mathbf{X}}(\delta = 1 \mid \mathbf{x}) = \pi(\boldsymbol{\eta} \mid \mathbf{x})$ sont des fonctions de vraisemblance pour le montant total et l'indicatrice de censure respectivement.

Puisque les covariables sont non stochastiques, la vraisemblance $f_{\Delta M,\mathbf{X},\Delta}(\delta m, \mathbf{x}, \delta)$ est

$$\mathcal{L}_1(\boldsymbol{\beta} \mid \mathbf{x}, \delta)\mathcal{L}_2(\boldsymbol{\eta} \mid \mathbf{x}) = \{\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{x})^{\delta}\}\{\pi(\boldsymbol{\eta} \mid \mathbf{x})^{\delta}[1 - \pi(\boldsymbol{\eta} \mid \mathbf{x})]^{1-\delta}\}. \tag{5.1}$$

La fonction de vraisemblance globale est un produit de deux fonctions de vraisemblance. La fonction de vraisemblance $\mathcal{L}_2$ est la vraisemblance d'une famille binomiale avec toutes les réclamations et $\mathcal{L}_1$ est la vraisemblance qui contribue seulement pour les réclamations fermées. Afin de pallier la perte d'efficacité par le biais de sélection, on peut modifier la vraisemblance $f_{\Delta M,\mathbf{X},\Delta}(\delta m, \mathbf{x}, \delta)$ par

$$\mathcal{L}_1^w(\boldsymbol{\beta} \mid \mathbf{x})\mathcal{L}_2(\boldsymbol{\eta} \mid \mathbf{x}) = \{\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{x})^{w}\}\{\pi(\boldsymbol{\eta} \mid \mathbf{x})^{\delta}[1 - \pi(\boldsymbol{\eta} \mid \mathbf{x})]^{1-\delta}\} \tag{5.2}$$

où $w = \dfrac{\delta}{\pi(\mathbf{x} \mid \boldsymbol{\eta})}$.

Par le même raisonnement qu'à la Section 5.2.2, il s'ensuit que l'estimateur du MV_{IPCW} de $\boldsymbol{\beta}$ est convergent, efficace et sans biais. Malgré la présence de données manquantes (à cause de la censure), il est possible de faire l'inférence statistique basée sur les données observées.

5.4 Modèles individuels à une étape

De façon similaire au modèle présenté à la Section 4.5 pour chaque réclamation ouverte ($\delta_i = 0$), les modèles à une étape tentent de modéliser l'espérance $E[M_i \mid \mathbf{X}_i, T_i > y]$ où y est encore la durée observée de la réclamation i. Au lieu de modéliser l'espérance conditionnelle avec un arbre de régression, elle sera modélisée par un wGLM. En particulier, une Poisson sur-dispersée sera la famille sélectionnée pour ce mémoire. Alors, on suppose que pour chaque réclamation ouverte, $M_i \mid \mathbf{X}_i, T_i > y$ est modélisé par un wGLM de la famille $\text{ODP}(\mu_i = \exp\{\mathbf{X}_i^T\boldsymbol{\beta}\}, \phi)$.

Posant $\Delta_1 = \mathbb{1}(T_i \leq C_i)$ et $\Delta_2 = \mathbb{1}(T_i > y)$, où y est une réalisation quelconque de Y_i. Le montant final cumulé M_i est seulement observé si $\Delta_1\Delta_2 = 1$. Sinon, le montant total

cumulé est inconnu. La vraisemblance est alors

$$\begin{aligned}
f_{\Delta_1\Delta_2 M,\mathbf{X},\Delta_1,\Delta_2}(\delta_1\delta_2 m,\mathbf{x},\delta_1,\delta_2) &= f_{M,\mathbf{X},\Delta_1,\Delta_2}(m,\mathbf{x},\delta_1=1,\delta_2=1)^{\delta_1\delta_2} \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=1,\delta_2=0)^{\delta_1(1-\delta_2)} \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=0,\delta_2=1)^{(1-\delta_1)\delta_2} \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=0,\delta_2=0)^{(1-\delta_1)(1-\delta_2)} \\
&= [f_{M|\mathbf{X},\Delta_1,\Delta_2}(m \mid \mathbf{x},\delta_1=1,\delta_2=1) \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=1,\delta_2=1)]^{\delta_1\delta_2} \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=1,\delta_2=0)^{\delta_1(1-\delta_2)} \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=0,\delta_2=1)^{(1-\delta_1)\delta_2} \\
&\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x},\delta_1=0,\delta_2=0)^{(1-\delta_1)(1-\delta_2)} \\
&\propto f_{M|\mathbf{X},\Delta_1,\Delta_2}(m \mid \mathbf{x},\delta_1=1,\delta_2=1)^{\delta_1\delta_2} \\
&\quad f_{\Delta_1,\Delta_2|\mathbf{X}}(\delta_1=1,\delta_2=1 \mid \mathbf{x})^{\delta_1\delta_2} \\
&\quad f_{\Delta_1,\Delta_2|\mathbf{X}}(\delta_1=1,\delta_2=0 \mid \mathbf{x})^{\delta_1(1-\delta_2)} \\
&\quad f_{\Delta_1,\Delta_2|\mathbf{X}}(\delta_1=0,\delta_2=1 \mid \mathbf{x})^{(1-\delta_1)\delta_2} \\
&\quad f_{\Delta_1,\Delta_2|\mathbf{X}}(\delta_1=0,\delta_2=0 \mid \mathbf{x})^{(1-\delta_1)(1-\delta_2)}.
\end{aligned}$$

Il s'ensuit que

$$\begin{aligned}
f_{\Delta_1\Delta_2 M,\mathbf{X},\Delta_1,\Delta_2}(\delta_1\delta_2 m,\mathbf{x},\delta_1,\delta_2) &\propto f_{M|\mathbf{X}}(m \mid \mathbf{x})^{\delta_1\delta_2} \\
&\quad \prod_{(i,j)\in\{0,1\}^2} f_{\Delta_1,\Delta_2|\mathbf{X}}(\delta_1=i,\delta_2=j \mid \mathbf{x})^{ij} \quad (5.3)
\end{aligned}$$

par l'hypothèse MAR qui suppose que $(\Delta_1,\Delta_2) \perp\!\!\!\perp M \mid \mathbf{X}$. Pour chaque réclamation ouverte (c'est-à-dire pour chaque valeur y), il faut calculer les poids $w_i = \dfrac{\delta_{1,i}\delta_{2,i}}{S_{C,T}(Y_i,y)}$ et pondérer la vraisemblance $f_{M_i|\mathbf{X}_i}(m_i \mid \mathbf{x}_i)$ par w_i. Puisque $T_i \perp\!\!\!\perp C_i$, $S_{C,T}(Y_i,y) = S_C(Y_i)S_T(y)$. La fonction de survie bivariée a lieu à cause du double biais de sélection : une fois pour le statut des réclamations et une fois pour la borne du délai de fermeture. Pareillement au développement de la vraisemblance donnée par l'Équation (5.2), l'expo-

sant de la vraisemblance du montant total payé est remplacé par w_i. L'algorithme de la stratégie **A1** est présenté à la Figure 5.1.

De façon similaire à l'Équation (4.5), la stratégie **A2** demande la modélisation du numérateur $E[M_i \mathbb{1}(T_i > y) \mid \mathbf{X}_i]$ et du dénominateur $E[\mathbb{1}(T_i > y) \mid \mathbf{X}_i]$ pour une réclamation ouverte qui a une durée observée y. La variable aléatoire au numérateur est une variable aléatoire qui est censurée selon la durée observée y. Typiquement, la condition de censure d'une variable dépend uniquement de la variable elle-même, mais dans ce cas elle dépend d'une autre variable, la durée observée de la réclamation.

Le numérateur est exprimé par la vraisemblance

$$f_{\Delta_1\Delta_2 M,\mathbf{X},\Delta_1,\Delta_2}(\delta_1\delta_2 m, \mathbf{x}, \delta_1, \delta_2) = f_{M\Delta_2,\mathbf{X},\Delta_1,\Delta_2}(m\delta_2, \mathbf{x}, \delta_1 = 1, \delta_2)^{\delta_1} \\ f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x}, \delta_1 = 0, \delta_2)^{1-\delta_1}.$$

Si l'indicateur $\mathbb{1}(T_i > y)$ est traité comme non stochastique pour chaque réclamation ouverte, on obtient exactement la vraisemblance donnée par l'Équation (5.1) où

$$\begin{aligned} f_{\Delta_1\Delta_2 M,\mathbf{X},\Delta_1}(\delta_1\delta_2 m, \mathbf{x}, \delta_1, \delta_2) &= f_{M\Delta_2,\mathbf{X},\Delta_1,\Delta_2}(m\delta_2, \mathbf{x}, \delta_1 = 1, \delta_2)^{\delta_1} \\ &\quad f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x}, \delta_1 = 0, \delta_2)^{1-\delta_1} \\ &\propto f_{M\Delta_2|\mathbf{X},\Delta_1,\Delta_2}(m\delta_2 \mid \mathbf{x}, \delta_1 = 1, \delta_2)^{\delta_1} \\ &\quad f_{\Delta_1|\mathbf{X},\Delta_2}(\delta_1 = 1 \mid \mathbf{x}, \delta_2)^{\delta_1} f_{\Delta_1|\mathbf{X},\Delta_2}(\delta_1 = 0 \mid \mathbf{x}, \delta_2)^{1-\delta_1} \\ &= f_{M\Delta_2|\mathbf{X},\Delta_1}(m\delta_2 \mid \mathbf{x})^{\delta_1} \\ &\quad f_{\Delta_1|\mathbf{X}}(\delta_1 = 1 \mid \mathbf{x})^{\delta_1} f_{\Delta_1|\mathbf{X}}(\delta_1 = 0 \mid \mathbf{x})^{1-\delta_1} \end{aligned}$$

ce qui implique que le numérateur sera estimé par un GLM pondéré avec poids $w_i = \dfrac{\delta_{1,i}}{\pi(\mathbf{X}_i)}$. Encore une fois, par l'hypothèse MAR, on suppose que $\Delta_1 \perp\!\!\!\perp M\Delta_2 \mid \mathbf{X}$. On supposera que $M_i \mathbb{1}(T_i > y) \mid \mathbf{X}_i \sim \text{ODP}(\mu_i = \exp\{\mathbf{X}_i^T \boldsymbol{\beta}\}, \phi_i)$ pour l'application de la stratégie **A2**.

Quant au dénominateur, sa vraisemblance est exprimée par

$$f_{\Delta_1\Delta_2,\mathbf{X},\Delta_1,\Delta_2}(\delta_1\delta_2, \mathbf{x}, \delta_1, \delta_2) = f_{\mathbf{X},\Delta_1,\Delta_2}(\mathbf{x}, \delta_1 = 1, \delta_2)^{\delta_1} f_{\mathbf{X},\Delta_1}(\mathbf{x}, \delta_1 = 0)^{1-\delta_1}$$

$$\propto f_{\Delta_2|\mathbf{X},\Delta_1}(\delta_2 \mid \mathbf{x}, \delta_1 = 1)^{\delta_1} f_{\Delta_1|\mathbf{X}}(\delta_1 = 1 \mid \mathbf{x})^{\delta_1} f_{\Delta_1|\mathbf{X}}(\delta_1 = 0 \mid \mathbf{x})^{1-\delta_1}$$
$$= f_{\Delta_2|\mathbf{X}}(\delta_2 \mid \mathbf{x})^{\delta_1} f_{\Delta_1|\mathbf{X}}(\delta_1 = 1 \mid \mathbf{x})^{\delta_1} f_{\Delta_1|\mathbf{X}}(\delta_1 = 0 \mid \mathbf{x})^{1-\delta_1}$$

ce qui implique que le dénominateur sera estimé par un GLM pondéré avec poids $w_i = \dfrac{\delta_{1,i}}{\pi(\mathbf{X}_i)}$. On supposera que $\mathbb{1}(T_i > y) \mid \mathbf{X}_i \sim \text{Bernoulli}(p_i = [1 + \exp\{-\mathbf{X}_i^T \boldsymbol{\zeta}\}]^{-1})$ pour l'application de la stratégie **A2**. L'algorithme de la stratégie **A2** est présenté à la Figure 5.2.

5.5 Modèles individuels à deux étapes

Pour les modèles à deux étapes, il faut encore une fois modéliser l'espérance conditionnelle $E[T_i \mid \mathbf{X}_i, T_i > y]$. Pour les fins de ce mémoire qui tente de comparer les modèles de réserve, il n'y a pas d'intérêt de comparer plusieurs modèles de régression pour le délai de fermeture. Le seul modèle qui sera considéré est la stratégie **A1** avec les poids IPCW K-M qui modélise l'espérance conditionnelle $E[T_i \mid \mathbf{X}_i, T_i]$ dont la prédiction $\widehat{T}_i$ est un argument de la fonction d'espérance conditionnelle $E[M_i \mid \mathbf{Z}_i] = \exp\{\mathbf{Z}_i^T \boldsymbol{\eta}\}$ où $\mathbf{Z}_i = (\mathbf{X}_i^T, T_i)^T$ est un vecteur colonne de covariables $\mathbf{X}_i$ auquel T_i est ajouté et $\boldsymbol{\eta}$ est un vecteur de paramètres de longueur $p + 1$.

Il n'est pas nécessaire de décrire l'algorithme pour les modèles à deux étapes parce qu'il demande un appel de l'algorithme 4.2 pour l'espérance de T_i et un appel de la fonction `glm` avec poids IPCW $\mathbf{w}$ avec matrice de covariables $\mathbf{Z} = [\mathbf{Z}_1 \mid \cdots \mid \mathbf{Z}_n]$. Comme les modèles à une étape, le montant total payé des modèles à deux étapes sera modélisé par une famille ODP.

Pour le modèle wGLM, l'estimateur MOM du paramètre de dispersion ϕ est

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} \tag{5.4}$$

pour les modèles à une étape. Pour les modèles à deux étapes, il suffit de réduire les degrés de liberté d'un à cause de l'ajout de la covariable T_i.

Algorithme 5 : Estimation de la réserve avec modèles à une étape (**A1**)

Input : Les poids $\mathbf{w}$ et seuil du nombre de réclamations fermées c

Output : Vecteurs d'estimations $(\widehat{\boldsymbol{\beta}}_1, \widehat{\phi}_1), \ldots, (\widehat{\boldsymbol{\beta}}_m, \widehat{\phi}_m)$ (un pour chaque réclamation ouverte) et la réserve totale $\widehat{R}$

Data : Base de données $\mathcal{D}_{\mathcal{T}}$

1. Éliminer les réclamations ouvertes dont $|\mathcal{D}^1_{\mathcal{T}}(y_i)| < c$ pour chaque réclamation ouverte $i = 1, \ldots, m$.
2. Calculer l'estimateur K-M de la fonction de survie $\widehat{S}_T(y)$ à l'aide de l'ensemble $\mathcal{D}_{\mathcal{T}}$.
3. Retourner les durées observées $y_1, \ldots, y_m$ pour les réclamations ouvertes restantes.
4. **foreach** $j \in \{1, \ldots, m\}$ *pour chaque indice de réclamation ouverte* **do**
   - Appliquer l'algorithme wGLM avec poids $w_i = \dfrac{\delta_{1,i}\delta_{2,i}}{\widehat{S}_C(Y_i)\widehat{S}_T(y_j)}$ à l'aide de l'ensemble d'entraînement $\mathcal{D}^1_{\mathcal{T}}(y_j)$.
5. Retourner les vecteurs d'estimations $(\widehat{\boldsymbol{\beta}}_1, \widehat{\phi}_1), \ldots, (\widehat{\boldsymbol{\beta}}_m, \widehat{\phi}_m)$.
6. Calculer la réserve globale

$$\widehat{R} = \sum_{j=1}^{m} \widehat{R}_j = \begin{cases} \sum_{j=1}^{m} \exp\{\mathbf{X}_j^T \widehat{\boldsymbol{\beta}}_j\}, \text{ si les réclamations n'ont pas de paiements partiels ;} \\ \sum_{j=1}^{m} [\exp\{\mathbf{X}_j^T \widehat{\boldsymbol{\beta}}_j\} - N_j], \text{ sinon.} \end{cases}$$

Figure 5.1: Étapes du calcul de la réserve avec modèles individuels à une étape (**A1**).

Algorithme 6 : Estimation de la réserve avec modèles à une étape (**A2**)

Input : Les poids **w** et seuil du nombre de réclamations fermées c

Output : Les vecteurs d'estimations $(\widehat{\boldsymbol{\beta}}_1, \widehat{\phi}_1), \ldots, (\widehat{\boldsymbol{\beta}}_m, \widehat{\phi}_m)$ pour le numérateur, les vecteurs d'estimations $\widehat{\boldsymbol{\zeta}}_1, \ldots, \widehat{\boldsymbol{\zeta}}_m$ pour le dénominateur et la réserve totale $\widehat{R}$

Data : Base de données $\mathcal{D}_{\mathcal{T}}$

1. Éliminer les réclamations ouvertes dont $|\mathcal{D}^1_{\mathcal{T}}(y_j)| < c$ pour chaque réclamation ouverte $j = 1, \ldots, m$.

3. Retourner les durées observées $y_1, \ldots, y_m$ pour les réclamations ouvertes restantes.

4. **foreach** $j \in \{1, \ldots, m\}$ *pour chaque indice de réclamation ouverte* **do**
   - Appliquer l'algorithme wGLM avec poids $w_i = \dfrac{\delta_i}{\widehat{\pi}(\mathbf{X}_i)}$ au numémateur $M_i \mathbb{1}(T_i > y_j)$ l'aide de l'ensemble d'entraînement $\mathcal{D}^1_{\mathcal{T}}$;
   - Appliquer l'algorithme wGLM avec poids $w_i = \dfrac{\delta_i}{\widehat{\pi}(\mathbf{X}_i)}$ au dénominateur $\mathbb{1}(T_i > y_j)$ l'aide de l'ensemble d'entraînement $\mathcal{D}^1_{\mathcal{T}}$.

5. Retourner les vecteurs d'estimations $(\widehat{\boldsymbol{\beta}}_1, \widehat{\phi}_1), \ldots, (\widehat{\boldsymbol{\beta}}_m, \widehat{\phi}_m)$ des paramètres du numérateur et les estimations $\widehat{\boldsymbol{\zeta}}_1, \ldots, \widehat{\boldsymbol{\zeta}}_m$ des paramètres du dénominateur.

6. Calculer la réserve globale

$$\widehat{R} = \sum_{j=1}^{m} \widehat{R}_j = \begin{cases} \sum_{j=1}^{m} \frac{\exp\{\mathbf{X}_j^T \widehat{\boldsymbol{\beta}}_j\}}{(1+\exp\{-\mathbf{X}_j^T \widehat{\boldsymbol{\zeta}}_j\})^{-1}}, & \text{si les réclamations n'ont pas de paiements partiels ;} \\ \sum_{j=1}^{m} \left[\frac{\exp\{\mathbf{X}_j^T \widehat{\boldsymbol{\beta}}_j\}}{(1+\exp\{-\mathbf{X}_j^T \widehat{\boldsymbol{\zeta}}_j\})^{-1}} - N_j \right], & \text{sinon.} \end{cases}$$

Figure 5.2: Étapes du calcul de la réserve avec modèles individuels à une étape (**A2**).

Y_j	n_j	d_j	$\widehat{S}_T(Y_j)$
29	12	0	$1 - \frac{0}{12} = 1$
31	11	0	$1(1 - \frac{0}{11}) = 1$
32	9	0	$1^2(1 - \frac{0}{9}) = 1$
45	5	1	$1^3(1 - \frac{1}{5}) = \frac{4}{5}$
61	4	0	$1^3(\frac{4}{5})(1 - \frac{0}{4}) = \frac{4}{5}$
62	3	0	$1^4(\frac{4}{5})(1 - \frac{0}{3}) = \frac{4}{5}$
63	1	0	$1^5(\frac{4}{5})(1 - \frac{0}{1}) = \frac{4}{5}$

Tableau 5.2: Données de survie de T pour les réclamations illustratives.

5.6 Exemples

En reprenant la même base de données tirée de la librairie `CASdatasets` et illustrée dans le Tableau 4.4, les résultats sont illustrés dans le Tableau 5.1. On rappelle que seulement la dernière réclamation du jeu de données (`CLM-12`) est ouverte au moment du calcul de la réserve.

Réclamation	Accident	Déclaration	Femeture	Legal	f_i	τ_i	C_i	T_i	Y_i	δ_i	M_i
`CLM-12`	1996-06-01	1996-07-01	1996-08-01	No	1996-07-15	31	45	62	45	0	397,49

Tableau 5.1: Données individuelles pour la réclamation ouverte `CLM-12`.

En utilisant l'estimateur K-M non paramétrique, on peut estimer la probabilité $P(T_i > y)$. Pour chaque valeur de y, il est également possible d'effectuer une régression logistique comme proposée dans la Section 3.4.1, mais il faudrait effectuer une régression logistique pour chaque Y_i d'une réclamation ouverte. Les principaux éléments sont présentés dans le Tableau 5.2.

Alors, les poids pour la stratégie **A1** sont

$$w_i = \begin{cases} 0, & i = \texttt{CLM-3}, \texttt{CLM-4}, \texttt{CLM-5}, \texttt{CLM-6}, \texttt{CLM-7}, \texttt{CLM-9}, \texttt{CLM-10}; \\ \left(\frac{5}{4}\right)^2, & i = \texttt{CLM-1}, \texttt{CLM-2}, \texttt{CLM-8}, \texttt{CLM-11}; \\ 0, & i = \texttt{CLM-12}. \end{cases}$$

Il en découle que les estimations des paramètres du GLM de Poisson sur-dispersé et pondérés de la stratégie **A1** sont présentées dans le Tableau 5.3.

Paramètre	Estimation
`Intercept`	6,35526
τ	−0,02995
`LegalYes`	1,04819
$\widehat{\phi}$	160,1533

Tableau 5.3: Exemple de wGLM suivant la stratégie **A1**.

La prédiction du modèle à une étape de la stratégie **A1** est

$$\widehat{M}_{12} = \exp\{6{,}35526 + 31(-0{,}02995) + 0(1{,}04819)\} = \exp\{5{,}42669\} = 227{,}40.$$

Le montant total payé prédit $\widehat{M}_{12}$ selon la stratégie **A2** est calculé par le rapport de l'espérance du montant total censuré $\widehat{M}_{12}^C$ et la probabilité de censure $\widehat{p}_{12}^C$. Encore, puisque la stratégie **A2** modélise la censure séparément l'aide d'une régression logistique, les estimés de $\widehat{M}_{12}$ diffèrent entre les stratégies **A1** et **A2**. Alors, $\widehat{p}_{12}^C = (1 + \exp\{-[2{,}07054 + 31(0{,}06916) + 0(0{,}38175)]\})^{-1} = (1 + \exp\{-0{,}92914\})^{-1} = 0{,}51836$ et $\widehat{M}_{12}^C = \exp\{4{,}36114 + 31(-0{,}01029) + 0(1{,}93185)\} = \exp\{4\,04187\} = 56{,}9326$. Il en résulte que la prédiction du montant total payé est $\widehat{M}_{12} = \dfrac{\widehat{M}_{12}^C}{\widehat{p}_{12}^C} = 109{,}83$. Les valeurs sont présentées dans le Tableau 5.4.

Paramètre	Estimation
`Intercept`	4,36114
τ	−0,01029
`LegalYes`	1,93185
$\widehat{\phi}$	652,0616

(a) Numérateur

Paramètre	Estimation
`Intercept`	−2,07054
τ	0,06916
`LegalYes`	0,38175

(b) Dénominateur

Tableau 5.4: Exemple de wGLM suivant la stratégie **A2**.

Paramètre	Estimation
`Intercept`	6,58044
τ	−0,00061
`LegalYes`	1,44524
T	−0,01527
$\widehat{\phi}$	1 128,865

Tableau 5.5: Exemple de wGLM suivant la stratégie **B**.

Si l'on réutilise la prédiction $\widehat{T}_{12} = 62{,}31$, la prédiction du modèle à deux étapes est $\widehat{M}_{12} = \exp\{6{,}58044 + 31(-0{,}00061) + 0(1{,}44524) + 62{,}31(-0{,}01527)\} = \exp\{5{,}61003\} = 273{,}15$. Les différentes valeurs sont présentées dans le Tableau 5.5.

CHAPITRE VI

RÉSULTATS NUMÉRIQUES

6.1 Introduction

Dans ce chapitre, on présente un résumé des données réelles fournies par un grand assureur canadien ainsi que le traitement de ces dernières. Avant de procéder à l'analyse des résultats obtenus par les modèles collectifs et individuels, il est nécessaire de tenir compte de l'inflation en adaptant les montants payés afin de respecter l'hypothèse i.i.d. des modèles des articles (Lopez *et al.*, 2016), (Lopez *et al.*, 2019) et (Lopez et Milhaud, 2021). Des statistiques descriptives sur les covariables utilisées pour les modèles proposés seront synthétisées à l'aide de graphiques et de tableaux. Afin d'observer la tendance de la réserve prédite par date d'évaluation, il y aura trois années d'évaluation étudiées en fonction l'erreur de prédiction pour les réclamations ouvertes et la distribution prédictive pour la réserve globale de chacun des modèles individuels wCART et wGLM. Pour les modèles wCART, de la stratégie **A2** il y a une grande augmentation du montant total payé prédit en fonction de la durée d'ouverture de la réclamation. Pour les modèles wGLM, la stratégie **A1** démontre un grand écart-type prédictif à cause des problèmes d'estimation causés par les réclamations ouvertes pour lesquelles $|\mathcal{D}_T^1(y)|$ est petit.

Pour chaque modèle présenté, il aura une étude des résultats individuels et des résultats globaux. En particulier, les réserves individuelles prédites peuvent être négatives. Afin de déterminer l'adéquation de chaque modèle, le bootstrap approprié sera employé pour simuler la distribution de la réserve globale ainsi que le calcul du RMSE basé sur les

réclamations ouvertes. Notez que puisque le modèle individuel de réclamation ouverte peut être révisé en fonction des covariables disponibles, la durée d'observation du dossier (Y_i) et l'augmentation du nombre de réclamations fermées avec lesquelles le modèle de la réclamation ouverte peuvent être améliorées.

À la prochaine section, il y aura une revue du traitement des données pour les individuels proposés qui demandent une modification relativement au processus décrit à la Section 4.4 puisque celui-ci sous-entendait une analyse basée sur une seule date d'évaluation.

6.2 Traitement des données

La base de données brute fournie pour ce mémoire par un grand assureur canadien comprend des réclamations d'accidents de véhicules avec couverture d'indemnité d'accident (AB) observée entre les années 1971 et 2017. Dans l'administration des réclamations, il y a 9 sous couvertures comprises dans la couverture AB. La couverture des indemnités d'accident indemnise un assuré blessé dans une collision ou offre du soutien financier en cas de décès. La couverture des indemnités d'accident inclut généralement des frais médicaux et de réadaptation qui ne sont pas déjà couverts par un régime d'assurance maladie.

En raison du peu de réclamations avant l'année 1992, il y a plusieurs années qui peuvent être exclues. En outre, à cause des systèmes de réclamations utilisées, il y a des covariables qui ne sont pas toujours disponibles tel que le sexe du réclamant, l'âge du réclamant, le coût historique du véhicule, l'année de fabrication du véhicule, la taille du ménage et le revenu du ménage. Ces dernières sont imputées à l'aide des forêts aléatoires de la librairie `missRanger`. Pour chaque arbre de régression pondéré et wGLM, il n'y a aucune sélection de variables et l'inclusion des variables explicatives susmentionnées est basée sur le jugement. Notamment, en fonction de l'année de calendrier, certaines variables explicatives possèdent une grande proportion de valeurs non attribuées. Malgré qu'elle ne peut pas être jugée une sélection de variables, certains modèles wGLM pour quelques réclamations ouvertes sont déficients selon le rang (*rank deficient*). En conséquence, les

variables qui ne pouvaient pas être retrouvées ont été supprimées.

Parmi les multiples années d'accidents et les multiples sous-couvertures disponibles dans la base de données fournie, on choisit deux sous-couvertures importantes : les frais médicaux (*medical*) et les dépenses (*expense*). Six années consécutives commençant à l'année 2010 ont été choisies. Pour ces dernières, les réclamations qui ont une date d'accident après la date d'évaluation de la réserve sont ignorées puisque les modèles proposés par le présent mémoire ont comme objectif de prédire la réserve RNBS individuelle.

Il y a eu 235 085 réclamants et 220 415 réclamations lors des années 1995 à 2017. Entre les années d'intérêt (de 2010 à 2015), il y a eu 69 913 réclamants et 56 038 réclamations. Il semble y avoir des réclamations avec des paiements aberrants (ou possiblement erronés) puisque le plus grand montant incrémental payé lors d'un trimestre est 781 162. Afin d'éliminer les réclamations aberrantes, celles qui ont des montants totaux payés plus élevés que le 99,9 percentile (208 935).

Afin d'assurer la convergence pour une année d'évaluation en particulier, il suffit d'éliminer les réclamations selon la première étape des algorithmes wCART et wGLM respectivement. Si $c = 25$ (c'est-à-dire, que $|\mathcal{D}_{\mathcal{T}}^{1}(y)| \geq 25$), chaque réclamation ouverte à la date d'évaluation aura au moins 25 réclamations fermées avec délais de fermeture plus élevés que la durée d'observation de la réclamation ouverte respective. Par contre, afin de maintenir le seuil c pour chaque date d'évaluation à laquelle la réserve sera calculée, il suffit d'éliminer les réclamations ouvertes ciblées à la date d'évaluation la plus lointaine dans le passé (`31-12-2013`) et d'éliminer ces réclamations pour toutes de données utilisées pour toutes évaluations futures de la réserve. Dit autrement, l'élimination des réclamations ouvertes qui ne respectent pas le seuil du nombre de réclamations fermées se fait de manière successive afin qu'une fois qu'une réclamation est éliminée à une date d'évaluation particulière, elle est éliminée pour toute date d'évaluation future.

Le Tableau 6.1 résume le nombre de réclamants et de réclamations à chaque date d'évaluation analysée dans ce chapitre ainsi que le nombre de réclamations ouvertes

Date d'évaluation		31-12-2013	31-12-2014	31-12-2015
Déclarées avant traitement	Réclamants	35 102	44 602	54 401
	Réclamations	28 250	35 773	43 575
Réclamants avec montants totaux payés aberrants		47	53	55
Ouvertes éliminées	Réclamants	95	71	68
	Réclamants (Cumul)	95	166	234
Déclarées après traitement	Réclamants	34 960	44 383	54 112
	Réclamations	28 145	35 616	43 365
	Réclamants avec dossiers fermés	29 209	38 236	47 657
	Réclamants avec dossiers ouverts	5 751	6 147	6 455
	Pourcentage de dossiers ouverts	16,5%	13,8%	11,9%

Tableau 6.1: Effectifs du nombre de réclamations déclarées relativement aux années d'évaluation 2013 à 2015.

éliminées à chaque étape de la procédure *ad hoc* décrite au chapitre précédent avant que les réclamations avec montants totaux payés aberrants (55 réclamations) soient éliminées.

Dorénavant, l'analyse individuelle sera effectuée sur la base de montants sommés par réclamant. Ceci est nécessaire, car il peut y avoir plusieurs réclamants pour une même réclamation ce qui rend difficile la modélisation avec covariables liées au réclamant.

6.2.1 Modélisation de l'inflation

Chaque année, l'inflation modifie la distribution du montant cumulatif. Un algorithme de déflation des montants est décrit par (Lopez et Milhaud, 2021) et sera résumé dans cette section. Contrairement aux poids IPCW, le taux d'inflation sera calculé seulement une fois puisqu'il est question dans ce mémoire d'avoir des réserves totales comparables d'une année d'évaluation à l'autre. Il est hors de la portée de ce mémoire d'analyser la tendance du taux d'inflation.

Soit M_i', le montant total payé pour la réclamation i avant la déflation et M_i après la

déflation. L'article précité pose

$$\log M_i' = \beta d_i + \log M_i, \tag{6.1}$$

où d_i est l'année d'accident (continue relativement à l'année 2010) de la réclamation i.

Soit $m_{i,j}'$ le montant cumulatif moyen parmi les réclamations complètes à l'année de développement j et dont l'accident est survenu à l'année $\lceil d_i \rceil$. Pour chaque année de développement j, (Lopez et Milhaud, 2021) effectue une régression linéaire pondérée par le nombre de réclamations observées $n_{i,j}$ dont l'accident est survenu à l'année $\lceil d_i \rceil$ et qui prend j années avant la fermeture.

On pose alors

$$(\widehat{\alpha}_j, \widehat{\beta}_j) = \underset{\alpha_j, \beta_j}{\operatorname{argmin}} \sum_{\mathcal{D}_{\mathcal{T}}^1} n_{i,j} (\log m_{i,j}' - \alpha_j - \beta_j d_i)^2, \tag{6.2}$$

où $\mathcal{D}_{\mathcal{T}}^1 = \{(\mathbf{X}_i, M_i') \in \mathcal{D}_{\mathcal{T}} \mid \delta_i = 1\}$. Le taux d'inflation $\widehat{\beta}$ est calculé par une moyenne pondérée

$$\widehat{\beta} = \frac{\sum_j \sqrt{n_j} \widehat{\beta}_j}{\sum_j \sqrt{n_j}},$$

où $n_j = \sum_i n_{i,j}$.

Les estimations des paramètres, le nombre de réclamations par période de fermeture et le calcul du taux d'inflation sont établis dans l'Annexe B. Pour estimer l'inflation débutant à l'année 2010, six années consécutives ont été choisies et le taux d'inflation annuel est de $-2{,}448061\%$. Alors, les montants (incrémentaux et cumulatifs) sont escomptés par le facteur $\exp\{-\widehat{\beta} d_i\}$.

6.2.2 Statistiques descriptives

La base de données fournie possède quelques caractéristiques importantes : il y a un taux élevé de réclamants avec montants accumulés nuls et la présence de montants accumulés aberrants. La sur-dispersion englobe ces deux phénomènes et cause l'écart-type du montant cumulé à être supérieur à la moyenne du montant cumulé. Notez que

les statistiques descriptives sont calculées avec les données observées des réclamations déclarées lors des années d'accident 2010 à 2015.

Par année d'accident, les boîtes à moustaches de la Figure 6.1 montrent clairement la dispersion des montants cumulés. La proportion des réclamations ouvertes a un impact direct sur la réserve et sur la précision des prédictions individuelles. Une étude brève de simulation de l'impact de la proportion des réclamations ouvertes a été effectuée par (Lopez et Milhaud, 2021) et montre que l'utilisation des poids IPCW diminue le RMSE pondéré pour les réclamations fermées.

Pour les modèles qui prennent en considération le biais de sélection induit par l'évaluation de la réserve, les différents délais (par exemple, le délai de déclaration) ont un impact direct sur la censure du délai de fermeture et la censure du montant total payé. Des quantiles pour les différents délais sont disponibles dans le Tableau 6.2.

Puisque les réclamations choisies sont celles qui ont des informations complètes relativement au montant total payé et au délai de fermeture, le montant cumulatif payé à la date d'évaluation de la réserve par réclamant relativement à la durée d'observation est illustré à la Figure 6.2. La réserve individuelle qui reste à être versée est illustrée à la Figure 6.3.

Avant de procéder à la prochaine section où les résultats finaux seront présentés, les covariables utilisées par les arbres de régression et les wGLM sont résumés dans le Tableau 6.3. La manière dont les coûts sont catégorisés est différente que la description générale de la Section 1.1. Pour la couverture AB, tous les coûts d'une réclamation sont répartis parmi les 9 sous couvertures. Les sous-couvertures *medical* et *expense* sont responsables de 33,9% et de 21,8% du coût total induit après l'année d'accident 2010.

Les variables `Medical Cov` et `Expense Cov` indiquent la présence d'un paiement pour chaque réclamant. Ces variables sont particulièrement influentes pour la prédiction de montant total payé. En général, les covariables portant les paiements individuels induisent un grand saut dans les prédictions individuelles du montant total payé. Il y a 22 305 réclamants (41,2%) qui n'ont aucun paiement positif inscrit à leurs dossiers.

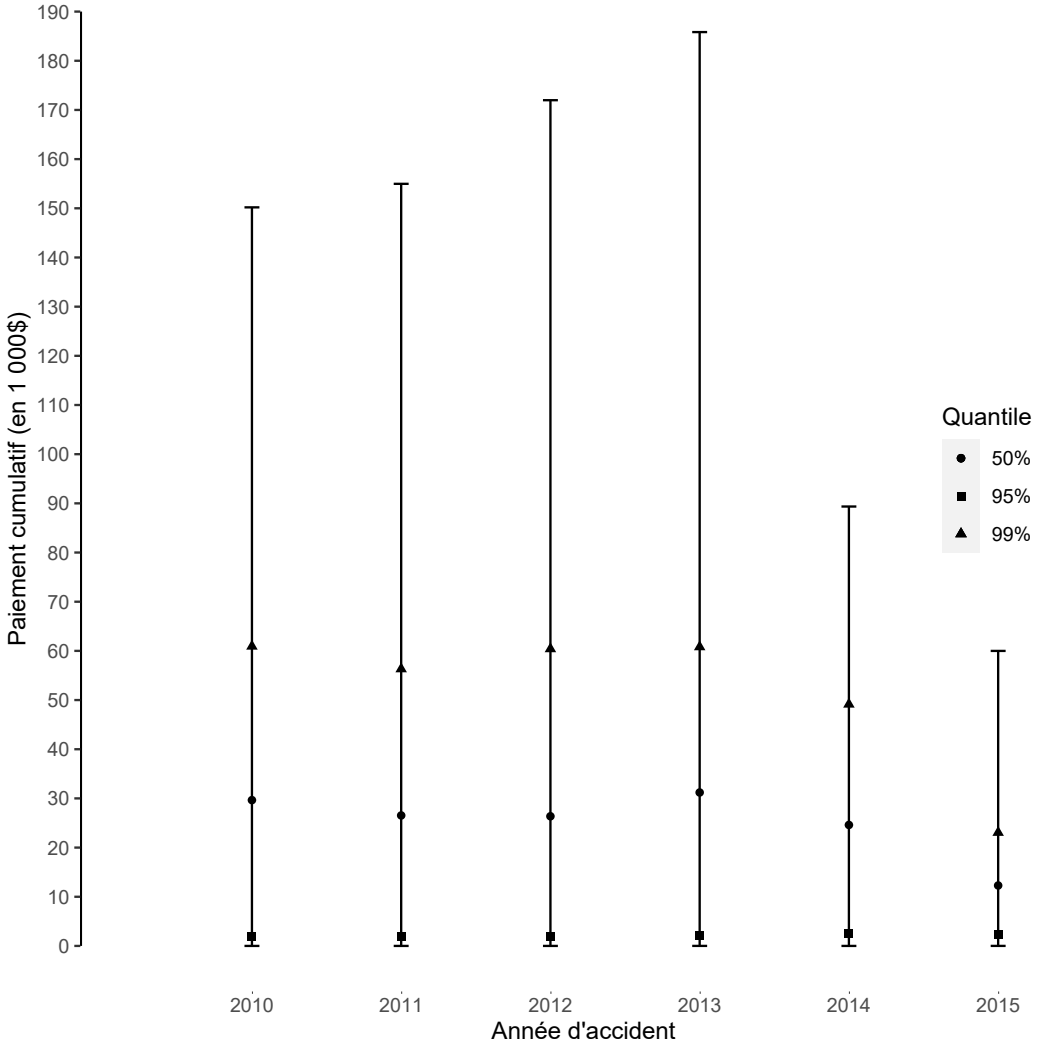


Figure 6.1: Quantiles des paiements cumulés par année d'accident.

Délai	Minimum	25^{e} quantile	Médiane	75^{e} quantile	90^{e} quantile	95^{e} quantile	99^{e} quantile	Maximum
Déclaration	0	0	1	2	6	18	106	1 340
Observation	0	50	103	254	500	706	1 166	1 778
Fermeture	0	55	110	299	592	801	1 286	2 119

(a) Toutes les réclamations

Délai	Minimum	25^{e} quantile	Médiane	75^{e} quantile	90^{e} quantile	95^{e} quantile	99^{e} quantile	Maximum
Déclaration	0	0	1	2	6	17	102	1 035
Observation	0	49	100	234	474	666	1 121	1 778
Fermeture	0	49	100	234	474	666	1 121	1 778

(b) Réclamations fermées

Délai	Minimum	25^{e} quantile	Médiane	75^{e} quantile	90^{e} quantile	95^{e} quantile	99^{e} quantile	Maximum
Déclaration	0	0	1	2	8	25	129	1 340
Observation	0	60	172	393	700	895	1 310	1 528
Fermeture	6	194	438	733	1 040	1 290	1 642	2 119

(c) Réclamations ouvertes

Tableau 6.2: Statistiques des durées pertinentes (en jours) de la base de données d'intérêt (calculées à la date d'évaluation `31-12-2015`).

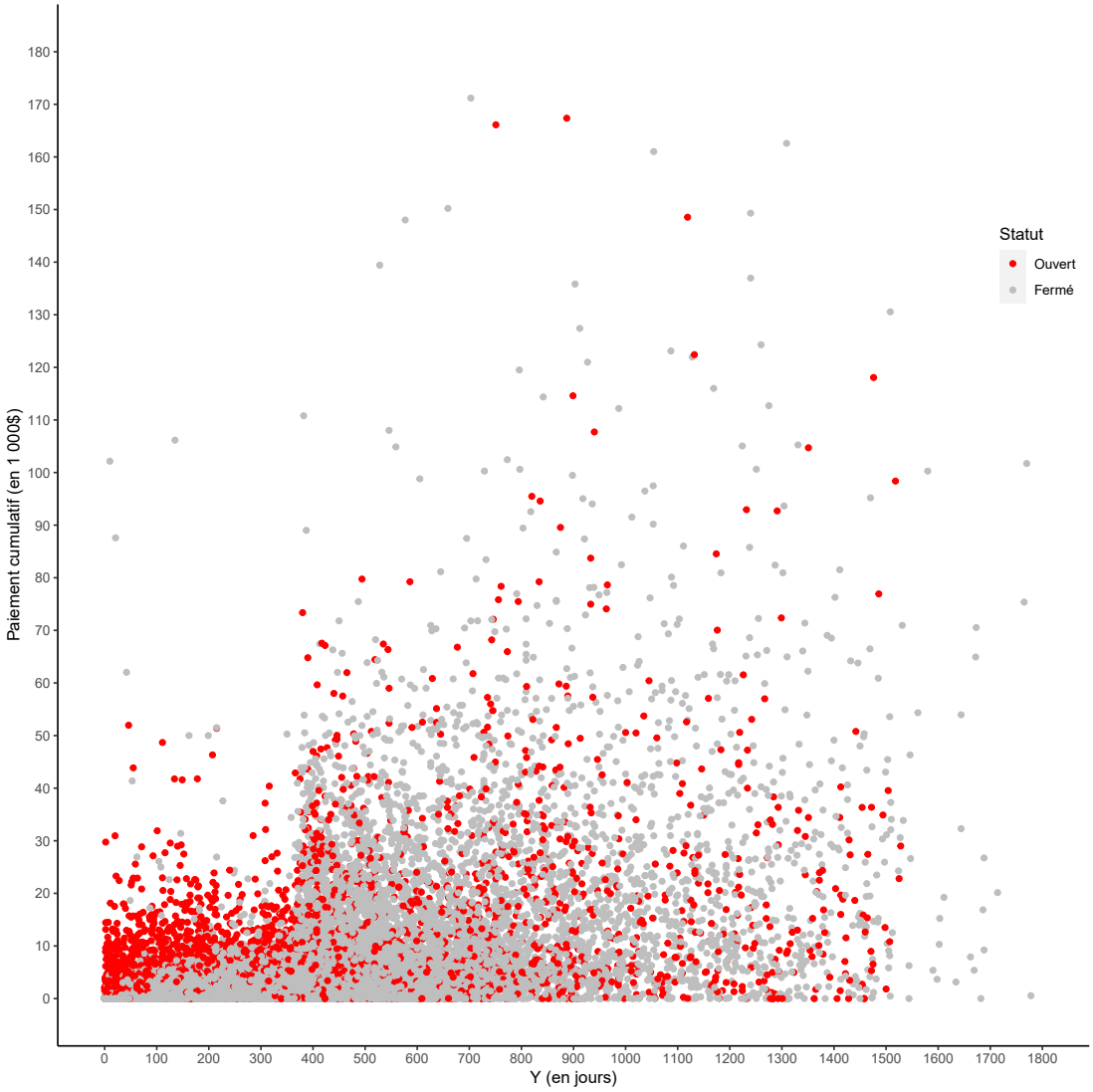


Figure 6.2: Paiement cumulatif pour chaque réclamant relativement à Y_i (calculé à la date d'évaluation `31-12-2015`).

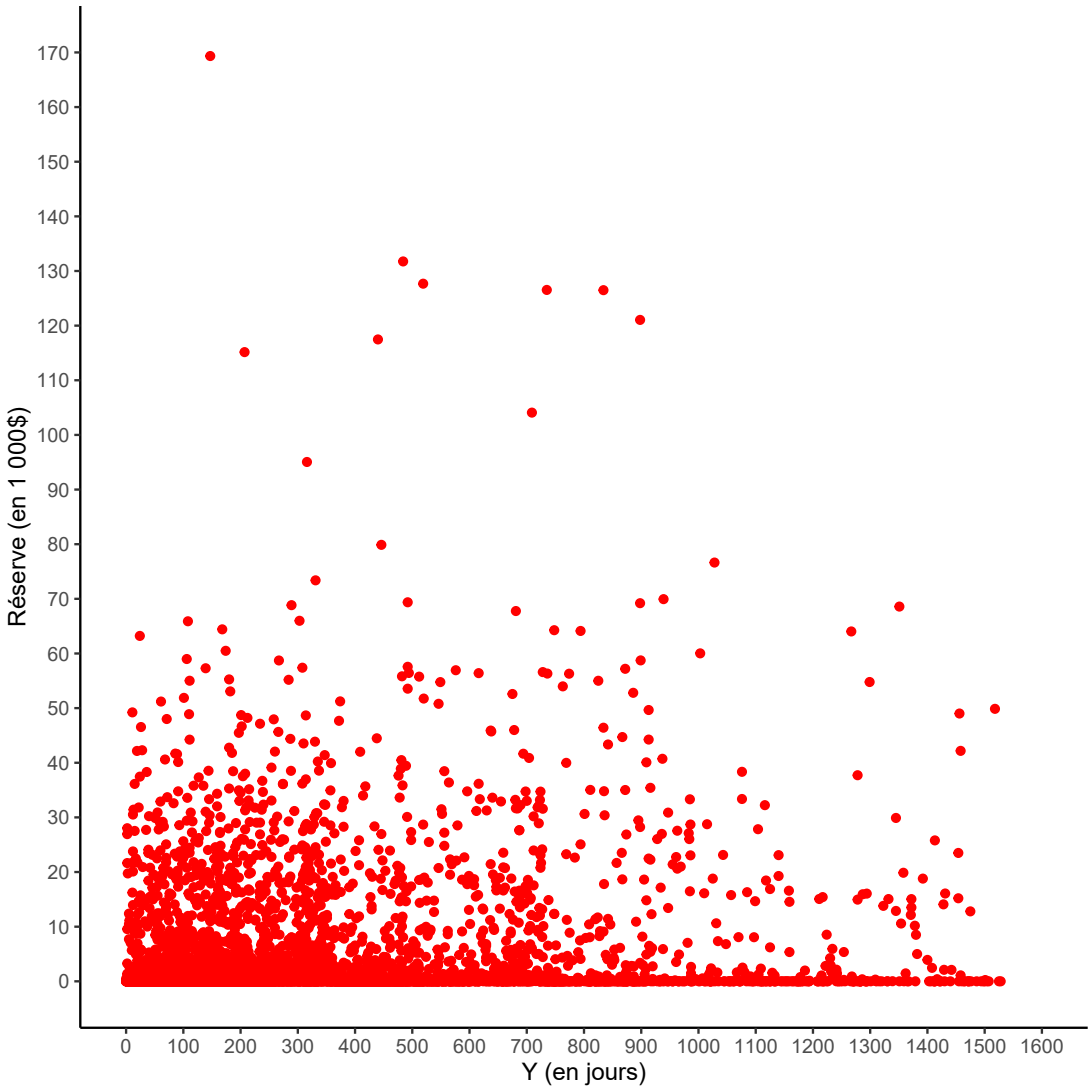


Figure 6.3: Réserve individuelle pour chaque réclamant ayant une réclamation ouverte relativement à Y_i (calculée à la date d'évaluation `31-12-2015`).

Nom	Type	Nombre de modalités	Description
`Reporting Delay Day`	Numérique	-	τ_i défini à la section 4.2
`HHincome AVG G5`	Numérique	-	Revenu du ménage du réclamant (en moyenne)
`Age at Loss`	Numérique	-	Âge du réclamant à la déclaration
`Cost Vehicle`	Numérique	-	Coût historique du véhicule
`Gender`	Catégorielle	2	Sexe du réclamant
`Legal Indic`	Catégorielle	2	Présence d'un avocat au dossier
`Risk Province Name`	Catégorielle	7	Province dans laquelle l'accident a lieu
`Loss Kind Name`	Catégorielle	7	Type d'accident
`Medical Cov`	Catégorielle	2	Retourne 1 s'il y a des frais médicaux (cumulatifs), 0 sinon.
`Expense Cov`	Catégorielle	2	Retourne 1 s'il y a des dépenses (cumulatives), 0 sinon.

Tableau 6.3: Résumé des covariables pour les modèles individuels.

		`Expense Cov`	
`Medical Cov`		0	1
	0	22 689	1 810
	1	20 828	8 785

Tableau 6.4: Nombre de réclamations ayant des paiements positifs selon la sous-couverture (calculés à la date d'évaluation `31-12-2015`).

La présence d'un paiement lié à une sous-couverture est dite une activation de la sous-couverture. Les effectifs des combinaisons d'activations des sous-couvertures des frais médicaux et des dépenses sont présentés dans le Tableau 6.4.

6.3 Modèles collectifs

Par les paiements cumulatifs du Tableau 6.5a, la réserve globale observée est de 24 604 946 (tenant compte de l'inflation). Surtout dans le cas où la variable réponse est sur-dispersée, les modèles collectifs surestiment la réserve globale. De plus, pour la base de données choisie pour ce travail, la grande majorité des réclamations ferment après deux années,

ce qui entraîne une surestimation de la réserve prédite pour les années d'accident 2014 et 2015.

Les modèles de Mack utilisés dans cette section sont résumés par l'algorithme 2.1 où les résidus sont calculés en fonction du triangle de développement incrémental. Ensuite, le processus cumulatif est simulé selon une hypothèse de distribution de Gamma ou de Poisson sur-dispersée. La fonction `BootChainLadder` a été employée avec 2 500 simulations pour arriver aux distributions prédictives de la réserve globale. La fonction `glmReserve` permet d'estimer la distribution prédictive de la réserve globale de la manière décrite par l'algorithme 2.1. Pour la famille Tweedie, la fonction `cpglm` permet d'estimer le paramètre p. Les paramètres ϕ sont estimés par la méthode des moments décrite à l'Équation (5.4) avec poids unitaires.

Comme attendu, il y a un consensus en moyenne entre les modèles de Mack et les modèles GLM. Le modèle de Mack surestime la réserve puisqu'il y a toujours des paiements élevés entre les années 2010 et 2013 tandis que les années 2014 et 2015 éprouvent une chute des paiements individuels, ce qui est illustré à la Figure 6.1.

La réserve globale prédite s'écarte en moyenne de plus en plus de la réserve observée et l'écart-type de la réserve globale prédite augmente lorsque le délai d'évaluation augmente, voir la Figure 6.4 qui illustre les distributions prédictives simulées pour trois années d'évaluation de la réserve. Ceci est une caractéristique du modèle de Mack qui peut souvent surestimer la réserve globale observée et est un désavantage de la modélisation de la réserve globale.

6.4 Modèles individuels

Comme décrit précédemment, les modèles individuels des chapitres précédents demandent différents traitements des données, différents ensembles d'entraînement et différents poids. La présente section tente de résumer l'impact des poids IPCW. Pour chaque modèle et pour chaque date d'évaluation considérée, des statistiques de la distribution prédictive

Année d'accident	Année de développement					
	1	2	3	4	5	≥ 6
2010	16 227 332	26 776 212	30 588 588	32 924 147	33 174 283	33 222 491
2011	144 62 456	20 970 440	25 069 081	26 764 394	27 400 568	27 497 867
2012	139 69 758	22 098 810	25 700 818	27 609 351	28 502 429	28 683 454
2013	159 57 880	27 005 123	32 460 846	35 515 980	35 757 228	35 757 228
2014	195 45 815	30 882 521	35 893 216	36 917 826	36 917 826	36 917 826
2015	205 83 708	32 168 573	34 666 843	34 666 843	34 666 843	34 666 843

(a) Paiements cumulatifs

Année d'accident	Année de développement					
	1	2	3	4	5	≥ 6
2010	16 227 332	26 776 212	30 588 588	32 924 147	33 174 283	33 222 491
2011	14 462 456	20 970 440	25 069 081	26 764 394	27 400 568	**27 440 386**
2012	13 969 758	22 098 810	25 700 818	27 609 351	**28 019 321**	**28 060 038**
2013	15 957 880	27 005 123	32 460 846	**34 830 582**	**35 347 779**	**35 399 146**
2014	19 545 815	30 882 521	**36 293 307**	**38 942 823**	**39 521 082**	**39 578 513**
2015	20 583 708	**32 798 337**	**38 544 784**	**41 358 665**	**41 972 797**	**42 033 791**

Note : Les prédictions sont en gras.

(b) Prédictions du modèle de Mack

Tableau 6.5: Triangle de développement cumulatif avec prédictions selon le modèle de Mack calculées à la date d'évaluation `31-12-2015`.

Type	Famille	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
Mack	ODP	3 864 466	1 373 796	662 256	6 360 430	7 657 285	10 045 370
	Gamma	3 865 263	1 345 504	845 459	6 311 670	7 449 910	9 789 290
		(4 002 400)					
GLM	ODP($\mu_{i,j}, \widehat{\phi} = 105\ 769{,}7$)	4 017 325	830 777	1 341 294	5 402 662	5 985 439	7 064 943
		(4 002 400)					
	Tweedie($\mu_{i,j}, \widehat{p} = 1{,}99, \widehat{\phi} = 0{,}0169$)	4 202 881	701 787	2 296 955	5 381 709	5 985 745	7 542 305
		(4 191 603)					
	Observé	4 714 570					

Note : Les réserves théoriques sont entre parenthèses.

(a) Évaluation à la date `31-12-2013`

Type	Famille	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
Mack	ODP	14 162 878	3 042 799	5 436 275	19 417 539	21 971 032	25 024 601
	Gamma	14 217 519	3 174 340	4 679 709	19 797 408	22 563 227	28 083 601
		(13 853 942)					
GLM	ODP($\mu_{i,j}, \widehat{\phi} = 198\ 549{,}4$)	14 042 930	2 461 450	6 735 629	18 105 153	20 203 926	23 273 899
		(13 853 942)					
	Tweedie($\mu_{i,j}, \widehat{p} = 1{,}01, \widehat{\phi} = 171\ 315{,}3$)	13 974 331	2 451 812	4 913 072	18 193 225	20 064 102	22 747 658
		(13 856 507)					
	Observé	10 541 820					

Note : Les réserves théoriques sont entre parenthèses.

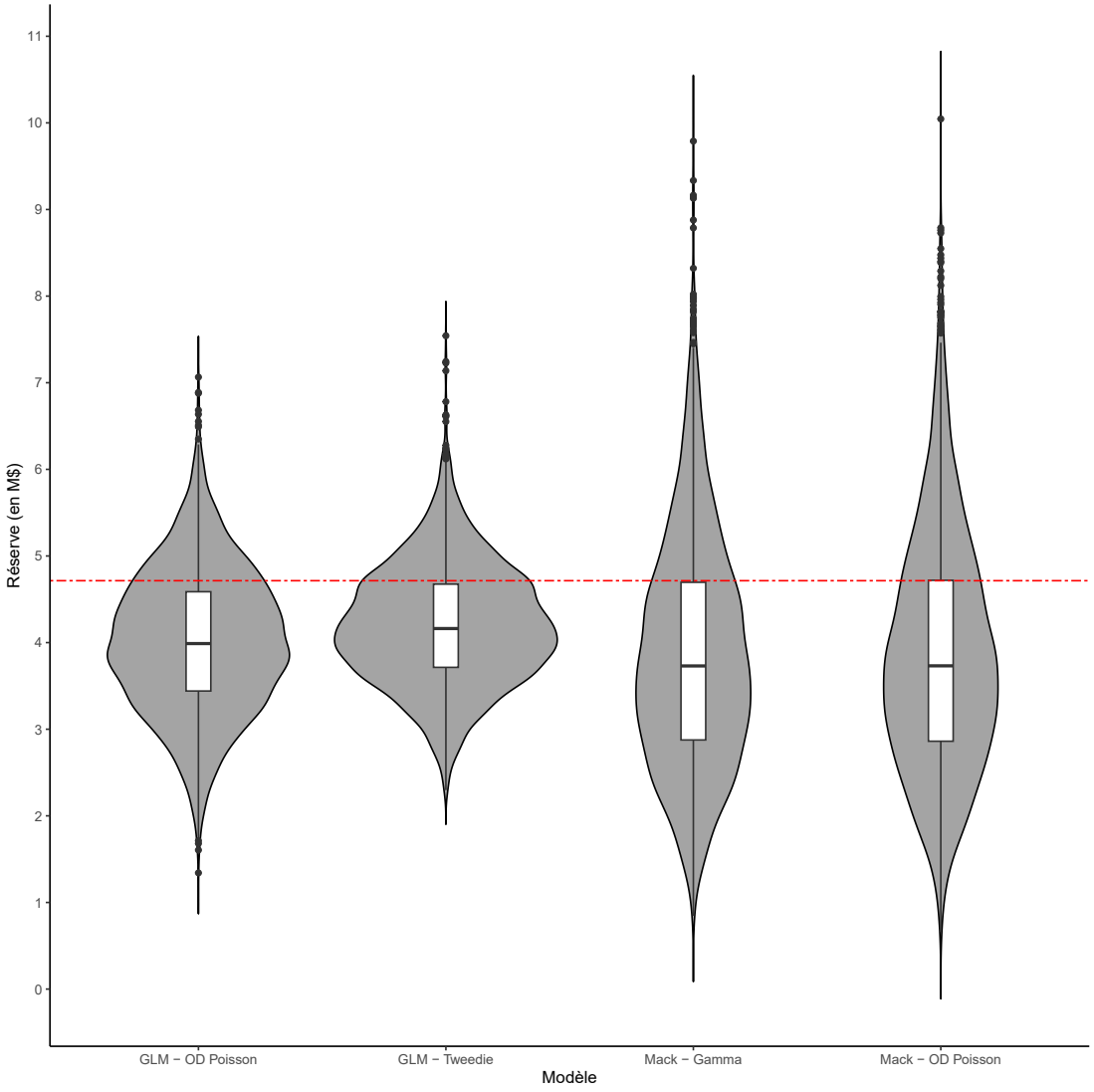
(b) Évaluation à la date `31-12-2014`

Type	Famille	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
Mack	ODP	33 585 909	3 246 377	23 737 621	38 952 962	41 152 193	46 010 192
	Gamma	33 675 664	3 341 634	23 883 784	39 066 918	41 862 803	45 306 185
		(33 574 878)					
GLM	ODP($\mu_{i,j}, \widehat{\phi} = 82\ 660{,}49$)	33 833 462	3 193 966	24 116 571	39 158 830	40 851 342	46 623 716
		(33 574 878)					
	Tweedie($\mu_{i,j}, \widehat{p} = 1{,}117, \widehat{\phi} = 18\ 849{,}62$)	33 889 731	3 399 239	23 628 123	39 399 935	41 814 758	45 765 186
		(33 590 080)					
	Observé	24 604 946					

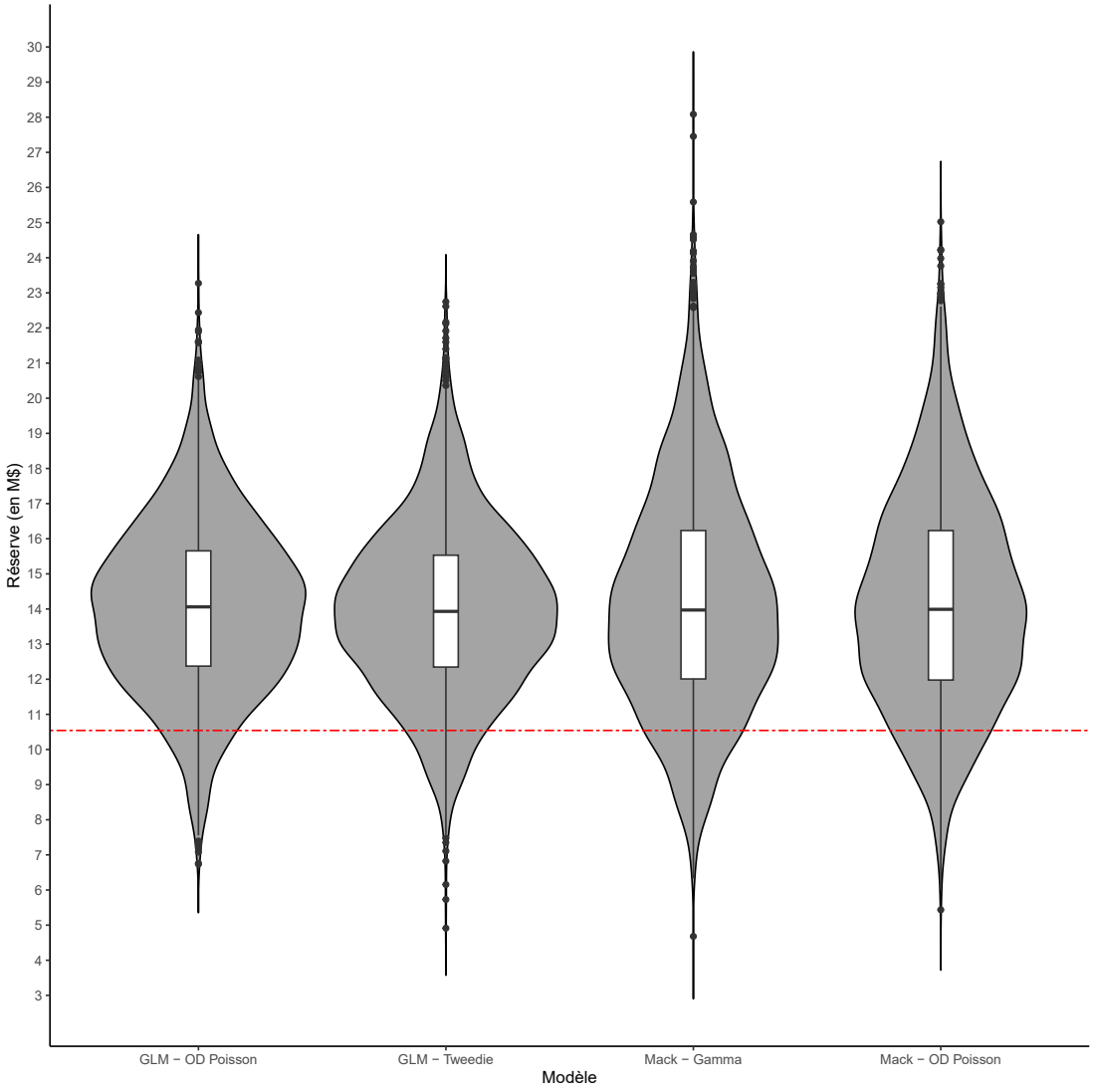
Note : Les réserves théoriques sont entre parenthèses.

(c) Évaluation à la date `31-12-2015`

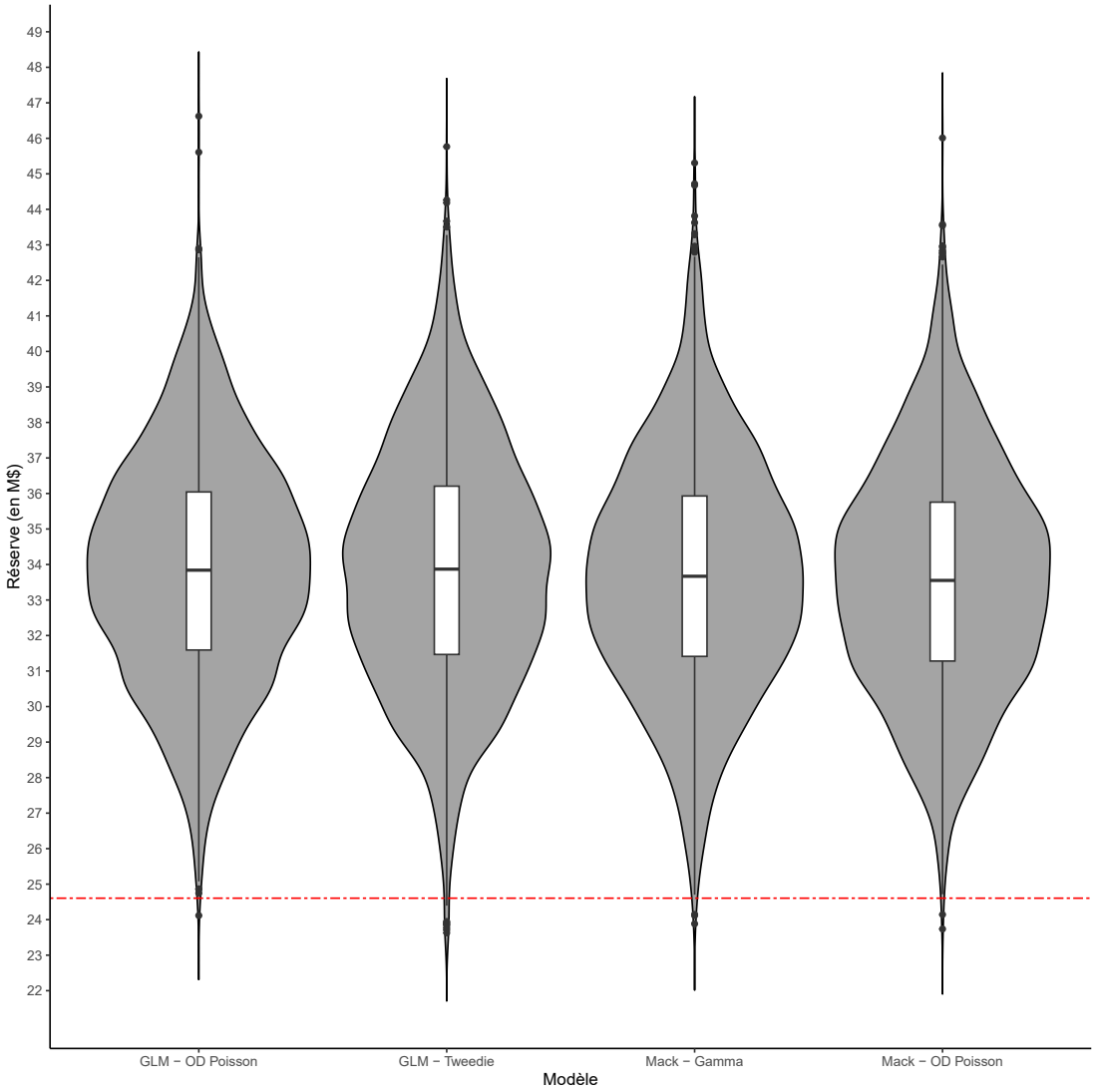
Tableau 6.6: Résultats des modèles collectifs pour la réserve globale RBNS.



(a) Évaluation à 31-12-2013



(b) Évaluation à la date `31-12-2014`



(c) Évaluation à la date `31-12-2015`

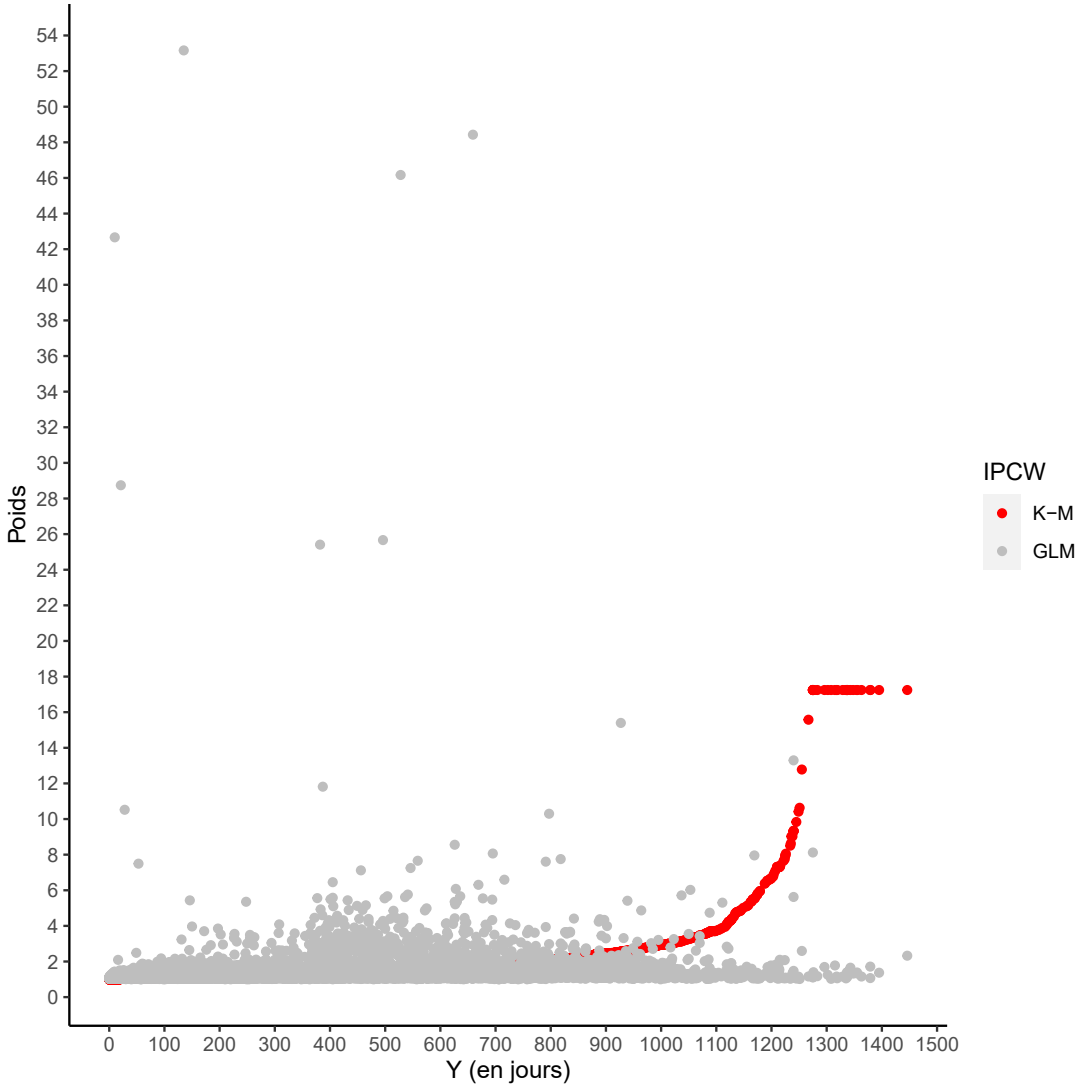
Figure 6.4: Distributions prédictives de la réserve globale RBNS pour les modèles collectifs.

de la réserve globale et de la réserve individuelle sont présentées. Les modèles wCART et wGLM tiennent compte du biais de sélection qui survient lorsque les réclamations fermées sont sélectionnées par le biais des poids IPCW. Si le montant total prédit $\widehat{M_i}$ est inférieur au montant cumulé N_i à l'évaluation, la réserve prédite définie par la différence $\widehat{M_i} - N_i$ peut s'avérer négative. Une réserve prédite individuelle négative est un défaut du modèle puisque $M_i \geq N_i$ pour chaque réclamant. Au lieu de conditionner sur une inégalité sur T_i qui entraîne un biais de sélection, il est possible d'utiliser une inégalité sur (T_i, M_i) qui amène une régression tronquée pour le montant total payé M_i. Les prédictions individuelles de la réserve RBNS les modèles wCART et wGLM seront analysés. Les stratégies **A1**, **A2** et **B** seront comparés et une brève étude longitudinale révélera certaines instabilités dans les modèles individuels et comment elles surviennent dans le temps.

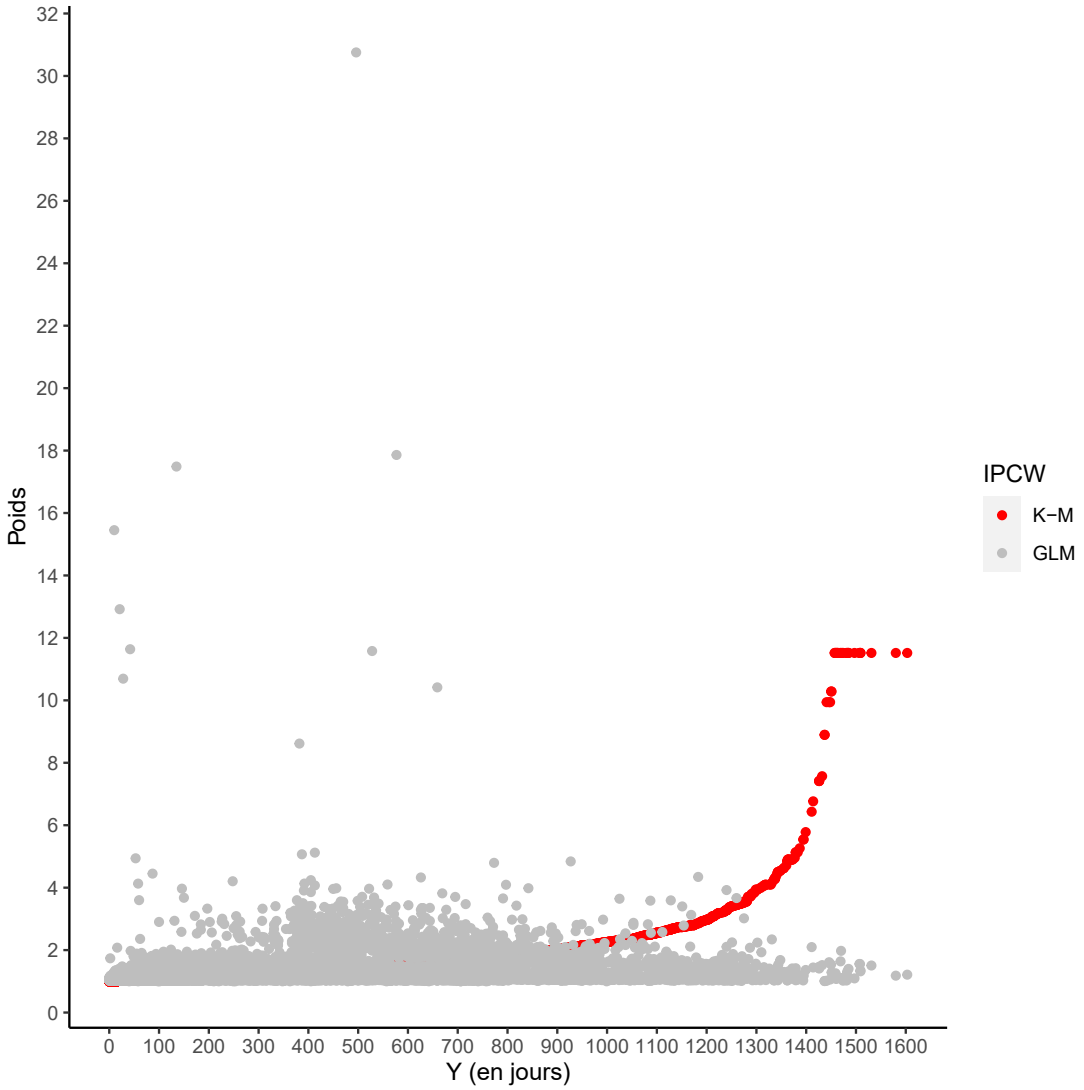
6.4.1 Poids IPCW

Pour chaque modèle à une étape et à deux étapes, les poids IPCW sont calculés sur l'ensemble d'entraînement $\mathcal{D}_{\mathcal{T}}$ pour les données créées à la manière décrite dans la Section 4.4. Il y a une ligne par réclamation qui contient les variables explicatives et le statut (ouvert ou fermé) de la réclamation. Pour chaque modèle des stratégies **A1**, **A2** et **B**, les poids selon l'estimateur Kaplan-Meier (notée par IPCW K-M) et la régression logistique (notée par IPCW GLM) sont calculés. Les paramètres du modèle de régression logistique sont disponibles dans l'Annexe C à chaque date d'évaluation considérée.

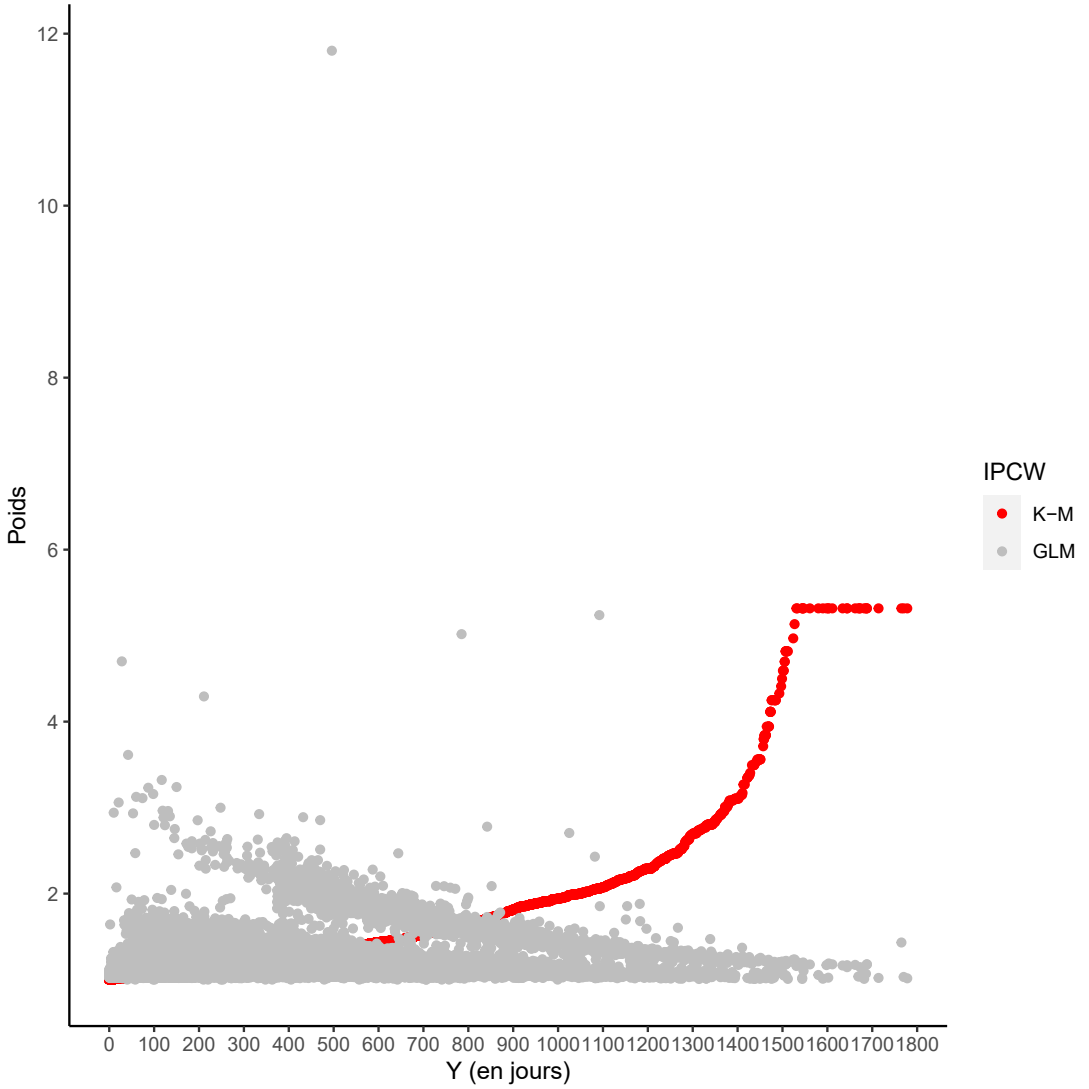
Selon la fonction de vraisemblance, les poids IPCW K-M de la stratégie **A1** pour les modèles wGLM sont assujettis à un biais de sélection double. Les poids en fonction de la durée d'observation Y_i pour chaque réclamation sont illustrés à la Figure 6.5. La fonction de survie bivariée $S_{C,T}$ est exprimée comme un produit de fonctions de survie univariées puisque $C_i \perp\!\!\!\perp T_i$ (τ de Kendall de 0,00895 à l'évaluation `31-12-2015`).



(a) Poids IPCW K-M et GLM calculés à la date d'évaluation `31-12-2013`



(b) Poids IPCW K-M et GLM calculés à la date d'évaluation `31-12-2014`



(c) Poids IPCW K-M et GLM calculés à la date d'évaluation `31-12-2015`

Figure 6.5: Poids IPCW selon les méthodes K-M et GLM pour les réclamations fermées.

6.4.2 Arbres de régression pondérés (wCART)

Les modèles de réserve individuelle d'arbres de régressions pondérés présentés au Chapitre 4 demandent des hyperparamètres pour le traitement des réclamations ouvertes et des caractéristiques des arbres. Pour le travail présent, avec le langage de la fonction `rpart`, `minbucket` $= 5$, `minsplit` $= 15$, `cp` $= 0$ et `maxdepth` est sélectionné parmi les valeurs $(5, 10, 25)$. La fonction `train` de la librairie `caret` permet de balayer plusieurs valeurs du `maxdepth` et d'optimiser l'erreur quadratique moyen du montant total payé relativement aux prédictions par validation croisée avec $k = 5$.

À titre d'illustration, les effectifs par nombre de couches des arbres pour chacune des stratégies utilisées pour estimer la réserve à chaque date d'évaluation considérée sont présentés aux Tableaux D.1, D.2 et D.3. Une profondeur de zéro est un arbre trivial qui possède une seule feuille terminale et celle-ci retourne alors la moyenne arithmétique pondérée $\frac{\sum_{i=1}^{n} w_i M_i}{\sum_{i=1}^{n} w_i}$. Les principaux résultats sont présentés au Tableau 6.8 ainsi qu'aux Figures 6.6, 6.7, 6.8 et 6.9.

La réserve RBNS globale simulée par un rééchantillonage avec remise à 2 500 reprises pour chacun des modèles wCART révèle l'instabilité évidente des modèles de la stratégie **A2**. L'instabilité individuelle semble dépendre de la durée d'observation. En particulier, pour Y_i supérieur à 800 jours, une proportion importante des réclamations prédisent une réserve croissante ou instable. Contrairement à la stratégie **A2**, la stratégie **A1** est plus conservatrice puisque les bornes de prédiction sont plus étroites et les prédictions sont également plus proches l'une des autres en général. Le choix des poids IPCW ne semble pas avoir un impact sur le contour du nuage de points. Quant à l'erreur de la réserve prédite (c'est-à-dire, la différence $\widehat{M}_i - M_i$), des déclarations analogues similaires peuvent être faites. Ceci est le cas puisqu'il y a de longs délais entre la date du dernier paiement et la fermeture. Dans le Tableau 6.7, on observe que la réserve individuelle augmente énormément si l'accident du réclamant affecte plusieurs véhicules (`Loss Kind Name` = `Multi-vehicle`), si le réclamant à un avocat aun avocat à sondossier (`Legal Indic` =

1) et s'il y a des paiements médicaux qui ont été versés par l'assureur au cours de l'année de l'évaluation (`Medical Cov` = 1).

La stratégie **B** sous-estime la réserve globale et la réserve prédite individuelle de manière que la dépendance positive entre le délai de fermeture et le montant total payé (τ de Kendall de 0,5327 à la date d'évaluation `31-12-2015`) est reflétée par les arbres de la stratégie **B**. On remarque que les arbres élagués de la stratégie **B** de l'Annexe E montrent cette tendance. Par contre, ce sont seulement pour les réclamations qui ont un délai de fermeture supérieur à 2 années où on remarque une augmentation marquée de la prédiction de la réserve. Le choix des poids IPCW semble avoir un plus grand impact sur la réserve individuelle puisqu'il y a une proportion plus élevée de réserves prédites positives lorsque les poids IPCW GLM sont choisis.

La profondeur de l'arbre élagué a un effet direct sur la précision de cette méthode. La stratégie **A1** a un écart-type inférieur à celui de la stratégie **B** puisque arbres des modèles à deux étapes sont souvent plus profonds. Notamment, la stratégie **A2** possède un écart-type large à cause de l'augmentation de la réserve prédite commensurablement avec Y_i et des prédictions aberrantes. Des tableaux qui montrent la profondeur des arbres individuels sont disponibles dans l'Annexe C et les arbres élagués de la stratégie **B** sont disponibles dans l'Annexe D pour la date d'évaluation `31-12-2015`. Les poids IPCW GLM réduisent l'écart-type de la réserve globale relativement aux poids IPCW K-M pour les stratégies **A1** et **A2**, mais non pour la stratégie **B**.

La stratégie **A1** est la seule qui tombe à l'intérieur du support de la distribution prédictive de la réserve, indépendamment du choix des poids IPCW. Évidemment, la stratégie **A2** demeure la plus instable et n'offre pas de résultats utilisables sans l'omission des prédictions extrêmes. Une analyse avec ces valeurs anormales ne sera pas effectuée dans ce mémoire puisqu'il est évident la surestimation marquée de la réserve globale persisterait. Un exemple de réclamation ouverte pour laquelle les arbres élagués des stratégies **A1** et **A2** qui montrent un grand écart dans la réserve prédite est donnée à l'Annexe F. La stratégie **A1** est alors plus applicable puisque l'effet double de la censure de la stratégie

Medical Cov	Loss Kind Name	Legal Indic	$\mathbb{1}(\widehat{M}_i - M_i > 10^5)$	Loss Year		
				2011	2012	2013
0	Hit and Run	0	0	·	1	·
		1	1	·	1	·
	Hit Pedestrian	0	0	·	1	·
		1	1	·	·	1
	Multi-vehicle	0	0	·	8	13
		1	0	·	·	7
		1	1	·	10	2
	Other	0	0	·	1	1
		1	1	·	2	1
	Single-vehicle	0	0	·	1	1
		1	0	·	·	2
		1	1	·	4	1
1	Hit and Run	1	0	·	·	1
		1	1	·	·	1
	Hit Pedestrian	0	0	·	2	5
		1	0	·	·	5
		1	1	1	5	4
	Multi-vehicle	0	0	4	28	59
		1	0	2	2	75
		1	1	8	93	78
	Other	0	0	·	1	7
		1	0	·	·	3
		1	1	1	1	4
	Single-vehicle	0	0	·	4	3
		1	0	·	1	5
		1	1	·	4	2

Tableau 6.7: Résumé des réclamations ouvertes à la date d'évaluation 31-12-2015 ayant $Y_i > 800$ pour la stratégie **A2** du wCART avec poids IPCW K-M.

A2 est nécessairement plus instable : une première fois par les poids IPCW et une deuxième fois par le rapport d'espérances. Il est possible de modifier les paramètres afin de pallier ce défaut. Notamment, on peut augmenter le nombre minimal d'observations nécessaires pour une feuille terminale. La stratégie **A1** aborde le problème de l'effet double de la censure par les poids IPCW et par la création du sous-ensemble $\mathcal{D}_{\mathcal{T}}(y)$.

6.4.3 Modèles linéaires généralisés pondérés (wGLM)

Les modèles linéaires pondérés par les IPCW manifestent plusieurs différences relativement aux modèles wCART. En premier lieu, le choix des poids IPCW a un plus grand impact sur les modèles wGLM. Deuxièmement, la tendance des écarts-types pour chaque classe de modèles est différente pour les modèles wGLM en comparaison avec les modèles wCART. Troisièmement, la dépendance positive entre T_i et M_i est plus clair pour les modèles de la stratégie **B**. La dépendance positive entre M_i et T_i est clair en observant le nuage de points entre la durée d'observation et l'erreur de la réserve. Encore une fois, l'écart des prédictions augmente avec Y_i.

Les principaux résultats sont présentés au Tableau 6.9 ainsi qu'aux Figures 6.10, 6.7, 6.8 et 6.9.

À cause de la vraisemblance de la stratégie **A1** qui tient compte d'un biais sélection double (une fois pour le statut de la réclamation, et une deuxième fois pour le délai de fermeture), il est attendu que la réserve globale est plus élevée en moyenne relativement au modèle wCART équivalent. La précision du modèle est forcément liée aux poids IPCW. La réserve globale prédite par la stratégie **A2** est significativement inférieure à celle de la stratégie **A1**. Cette diminution est également observée pour la précision de ces modèles. Il est important de retenir que le rééchantillonage de la réserve globale utilisée dans ce mémoire ne demande pas de réajustement des modèles. Alors, l'instabilité qui a été observée par les modèles wCART de la stratégie **A2** est une seule instance. Il serait possible d'évaluer l'ampleur de cette instabilité pour les modèles de la stratégie **A2** en performant un réajustement des modèles pour différentes périodes d'accident ou encore un réajustement

Stratégie	IPCW	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
A1	K-M	4 476 438	1 097 519	342 161	6 283 062	7 113 050	7 810 209
		(4 457 723)					
	GLM	4 442 449	1 095 156	385 262	6 178 179	6 943 104	8 443 042
		(4 409 068)					
A2	K-M	68 133 874	4 878 695	53 449 117	76 467 607	79 600 900	83 909 396
		(68 164 664)					
	GLM	64 904 361	4 830 304	48 522 706	73 120 330	76 310 530	85 016 606
		(64 932 912)					
B	K-M	14 618 888	1 495 997	9 250 069	17 080 878	18 026 808	20 082 975
		(14 603 891)					
	GLM	20 973 114	1 281 575	16 304 864	23 023 150	23 760 478	25 461 703
		(21 000 611)					
Observé		4 714 570					

Note : Les réserves théoriques sont entre parenthèses.

(a) Évaluation à la date `31-12-2013`

Stratégie	IPCW	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
A1	K-M	10 722 682	1 037 469	7 071 477	12 384 150	13 053 704	14 567 691
		(10 697 365)					
	GLM	10 561 138	1 028 926	7 121 932	12 278 994	12 976 982	14 496 857
		(10 549 051)					
A2	K-M	70 245 535	4 577 283	57 004 578	77 856 970	80 735 584	87 062 372
		(70 292 100)					
	GLM	70 859 769	4 645 139	55 713 582	78 813 006	82 158 800	86 043 022
		(70 888 190)					
B	K-M	4 367 650	1 070 640	804 671	6 110 014	6 790 668	7 825 575
		(4 344 172)					
	GLM	12 600 799	1 155 976	8 624 540	14 423 097	15 280 513	15 965 754
		(12 603 496)					
Observé		10 541 820					

Note : Les réserves théoriques sont entre parenthèses.

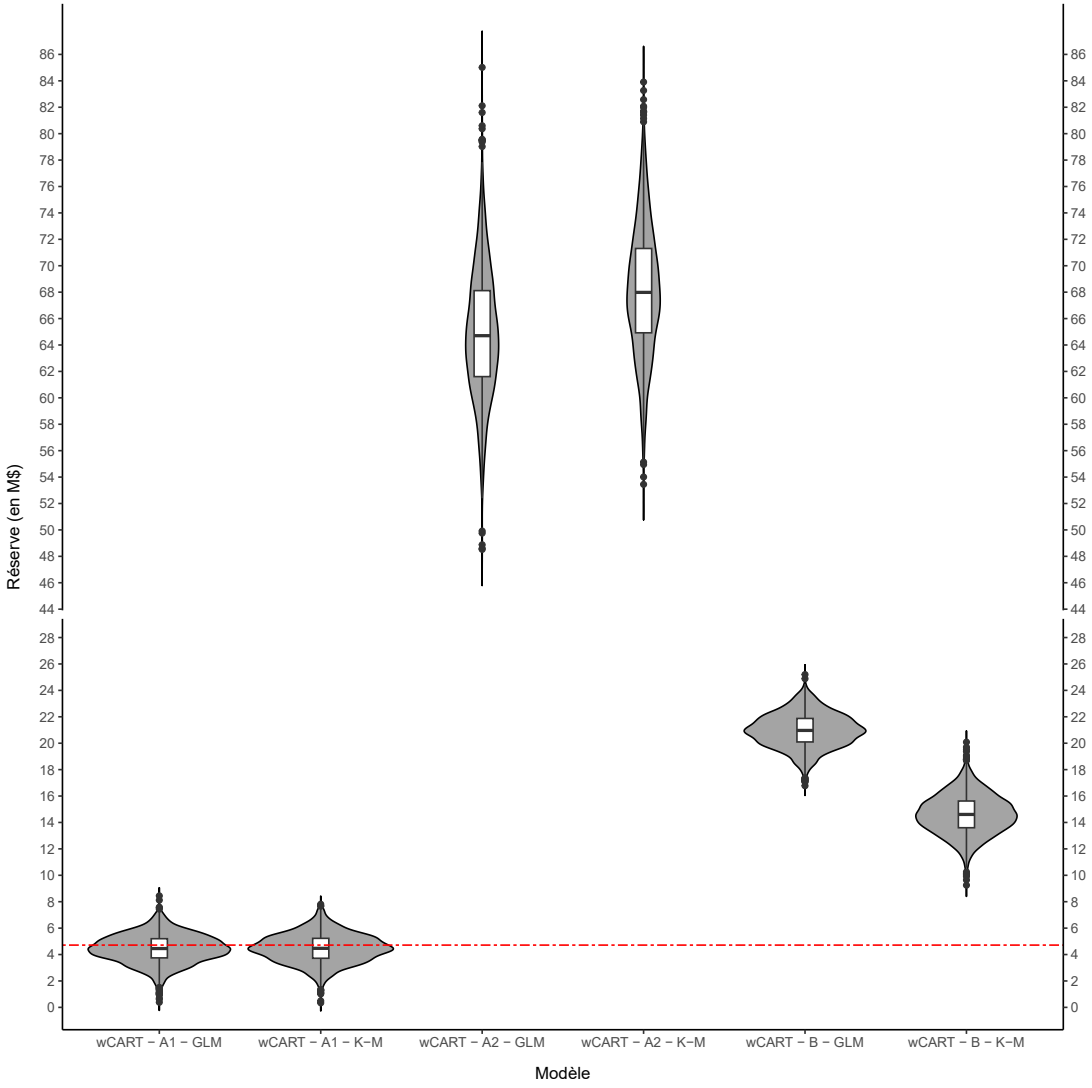
(b) Évaluation à la date `31-12-2014`

Stratégie	IPCW	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
A1	K-M	23 272 273	877 617	20 387 734	24 699 752	25 235 610	26 975 674
		(23 264 205)					
	GLM	23 438 490	865 591	20 882 415	24 866 463	25 415 052	26 040 712
		(23 456 367)					
A2	K-M	93 449 926	4 908 356	76 216 427	101 933 435	104 790 739	110 412 003
		(93 397 420)					
	GLM	94 261 038	4 245 016	80 552 540	101 327 128	104 349 914	108 423 841
		(94 228 037)					
B	K-M	19 265 624	861 635	16 258 924	20 674 512	21 260 977	21 876 524
		(19 254 104)					
	GLM	19 310 354	901 451	16 184 771	20 782 026	21 370 037	22 431 927
		(19 641 482)					
Observé		24 604 946					

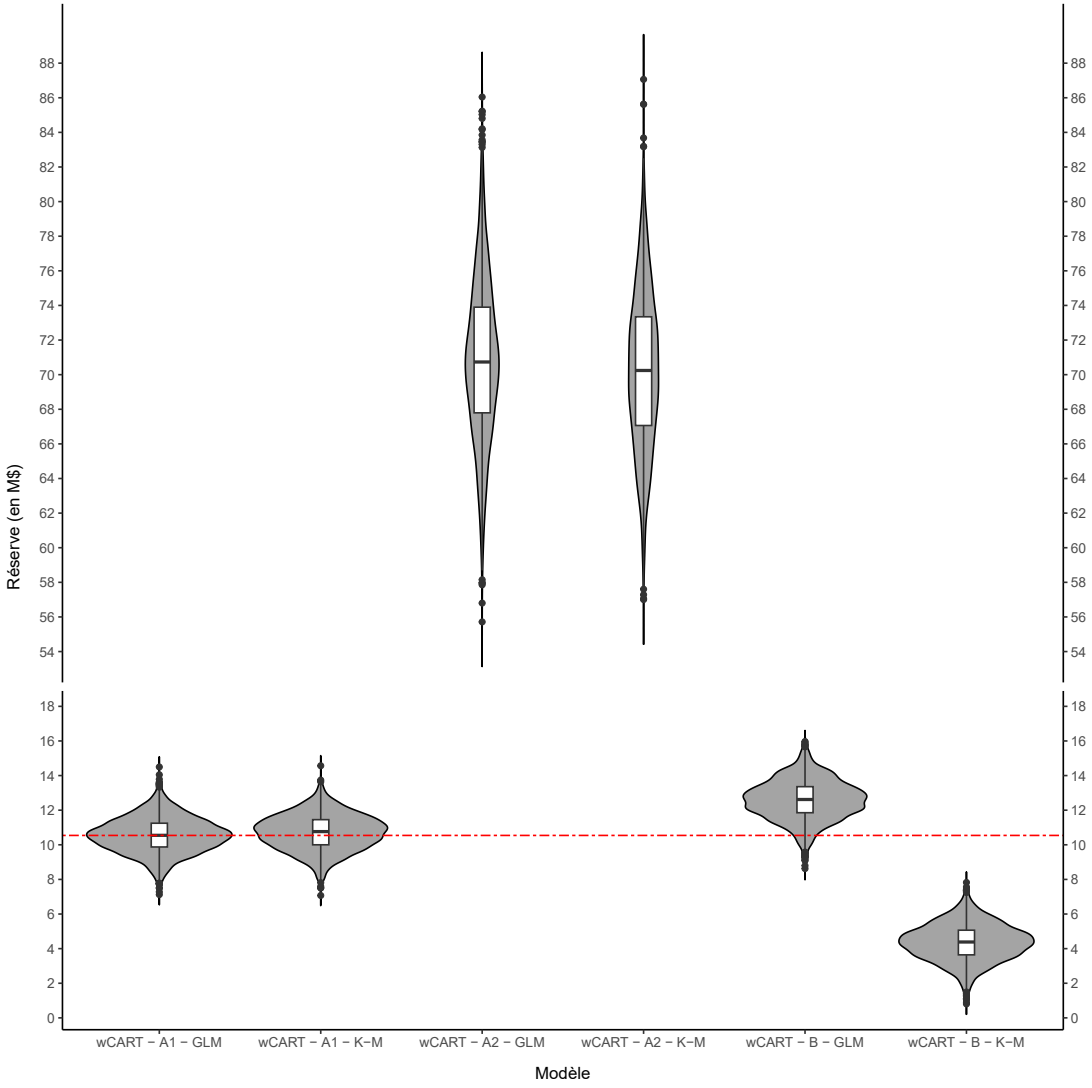
Note : Les réserves théoriques sont entre parenthèses.

(c) Évaluation à la date `31-12-2015`

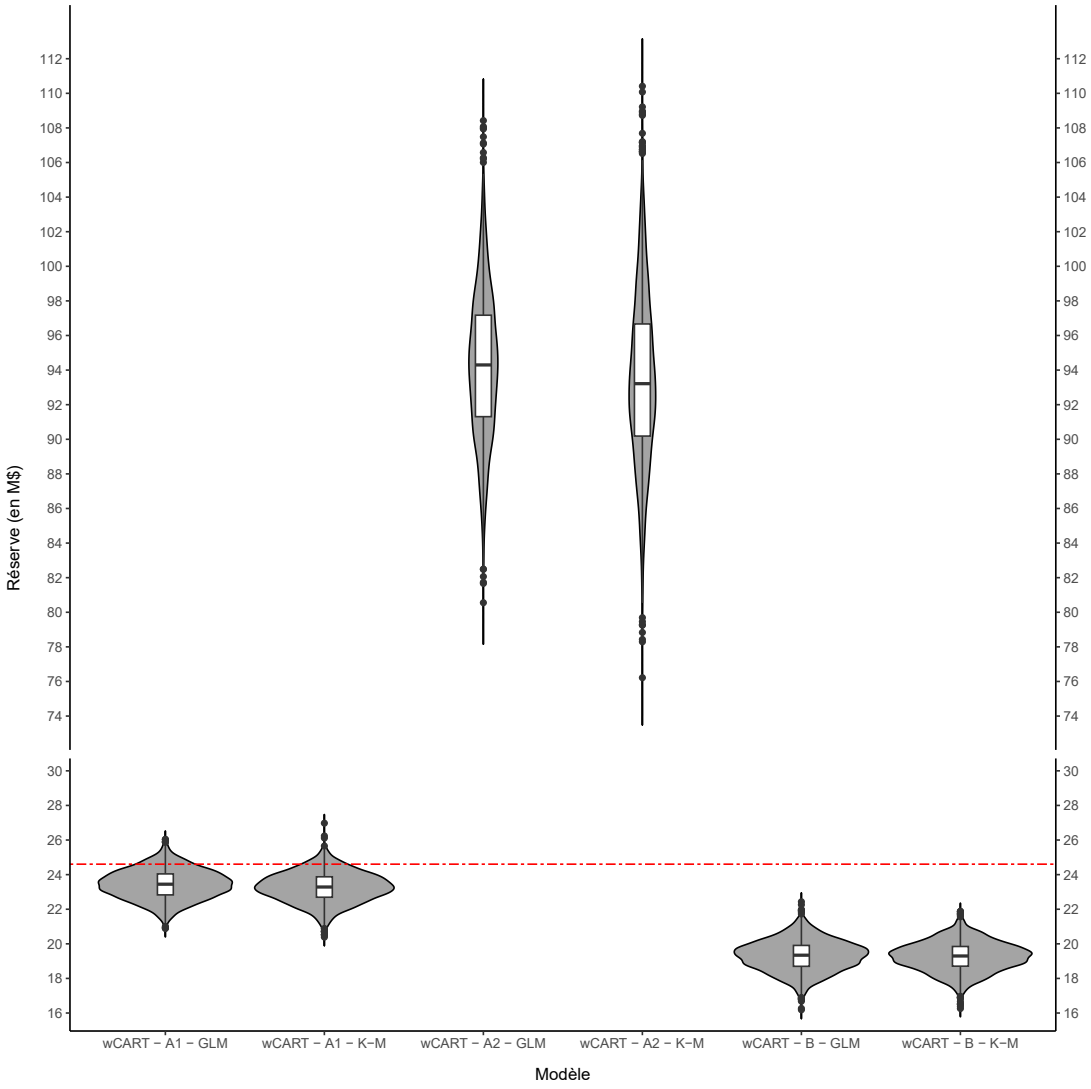
Tableau 6.8: Statistiques des simulations de la réserve RBNS pour les modèles individuels wCART.



(a) Évaluation à 31-12-2013

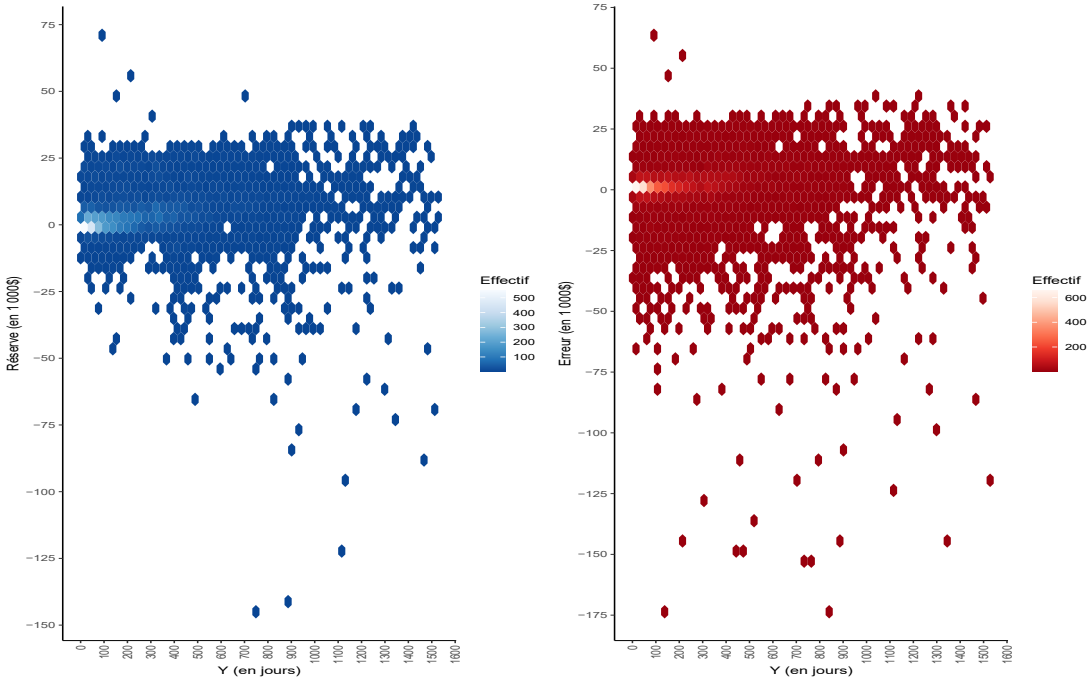


(b) Évaluation à la date `31-12-2014`

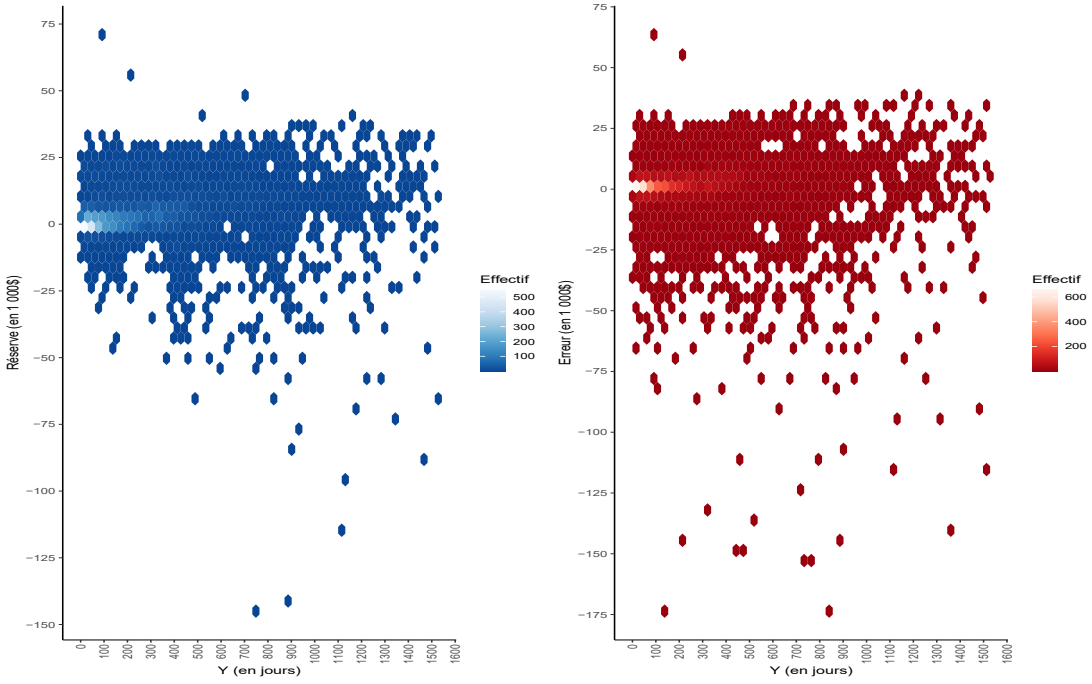


(c) Évaluation à la date `31-12-2015`

Figure 6.6: Distributions prédictives de la réserve globale RBNS pour les arbres de régression pondérés (wCART).

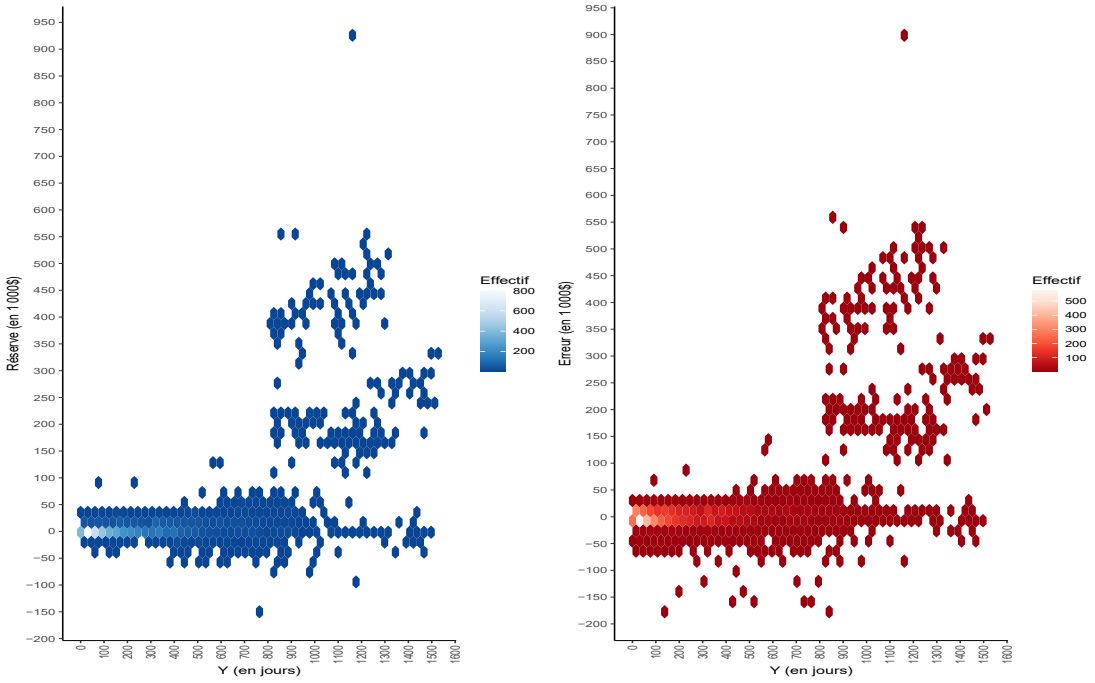


(a) Stratégie **A1** avec IPCW K-M

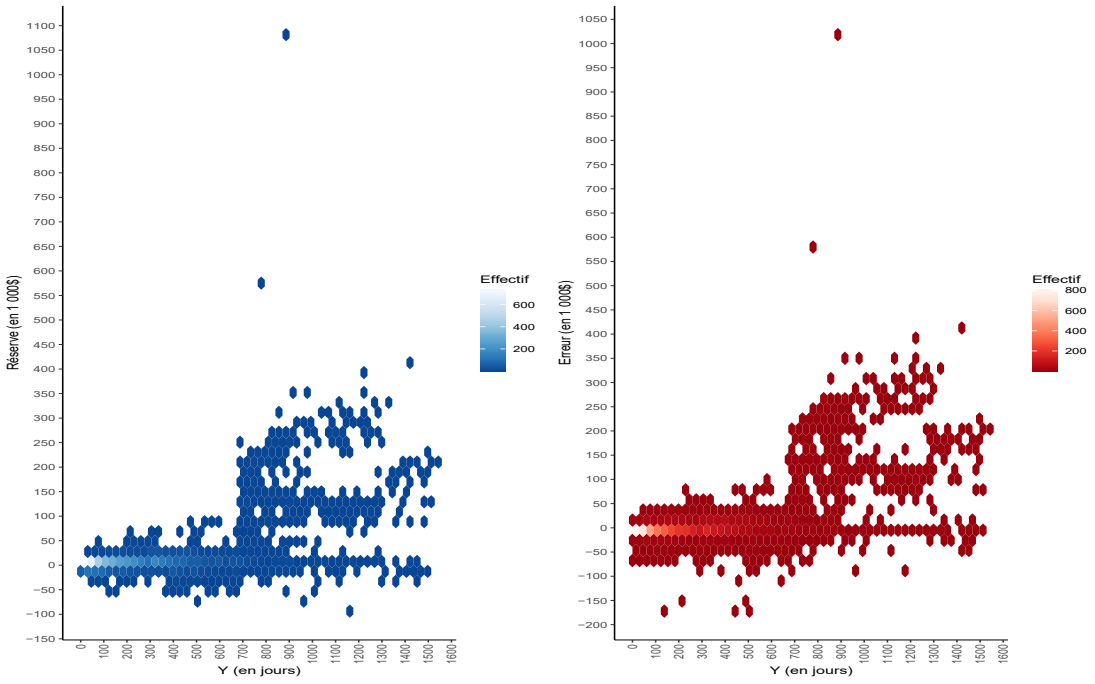


(b) Stratégie **A1** avec IPCW GLM

Figure 6.7: Réserve individuelle RBNS prédite à la date d'évaluation `31-12-2015` pour les arbres de régression pondérés (wCART) - Stratégie **A1**.

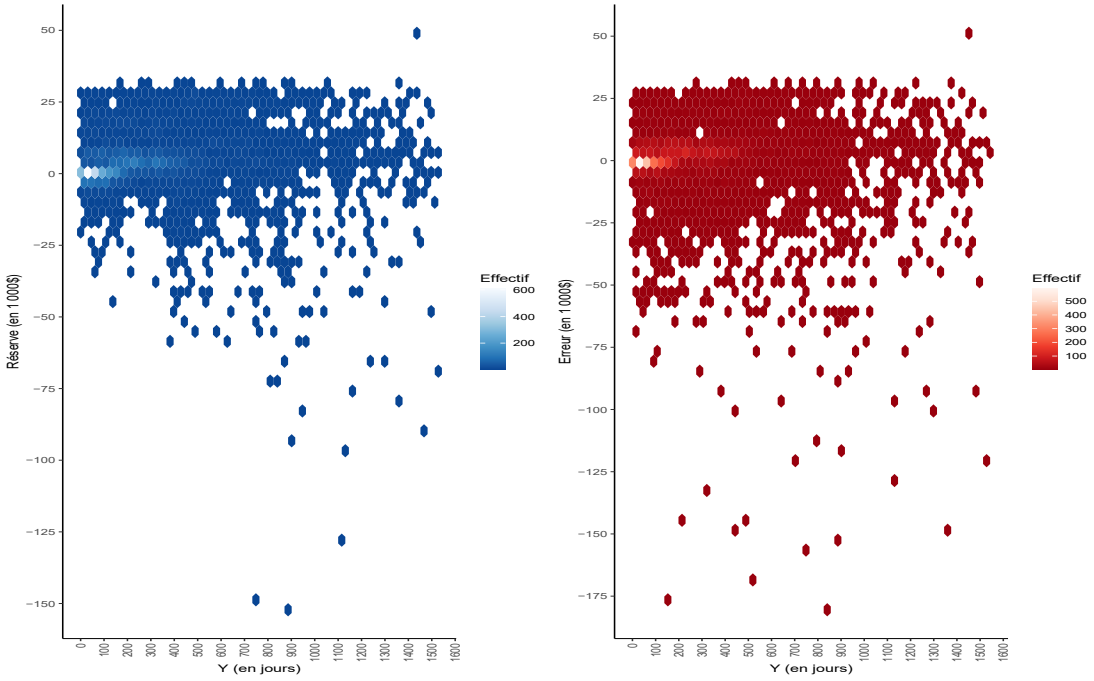


(a) Stratégie **A2** avec IPCW K-M

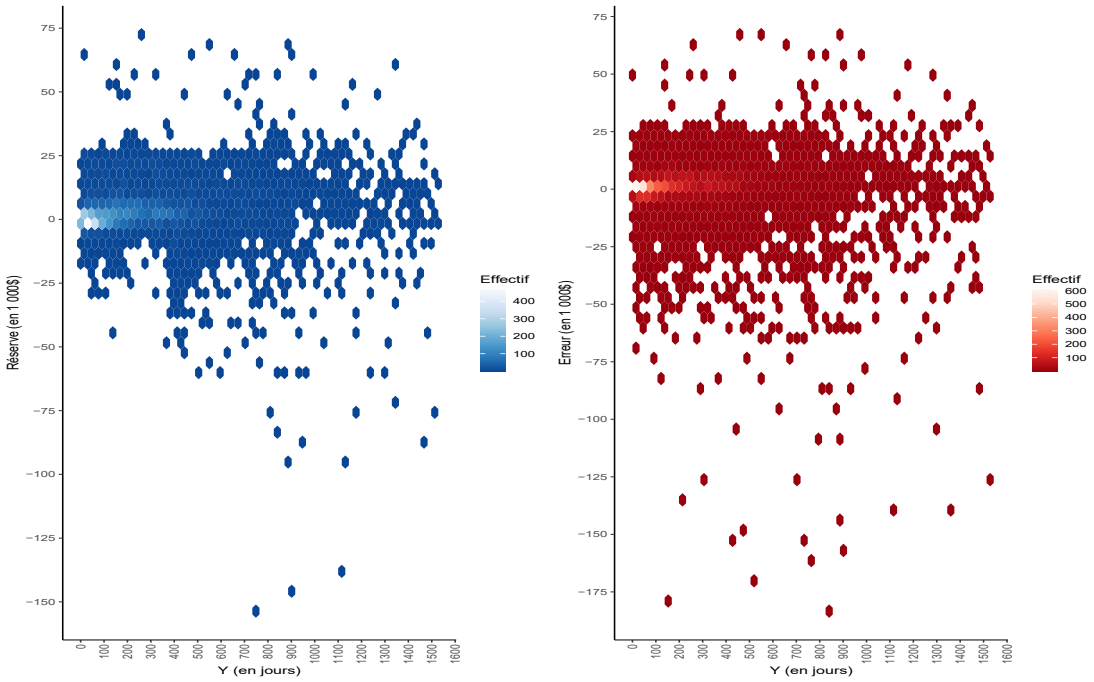


(b) Stratégie **A2** avec IPCW GLM

Figure 6.8: Réserve individuelle RBNS prédite à la date d'évaluation `31-12-2015` pour les arbres de régression pondérés (wCART) - Stratégie **A2**.



(a) Stratégie **B** avec IPCW K-M



(b) Stratégie **B** avec IPCW GLM

Figure 6.9: Réserve individuelle RBNS prédite à la date d'évaluation `31-12-2015` pour les arbres de régression pondérés (wCART) - Stratégie **B**.

pour chaque rééchantillonage de la base de données, mais ce peut s'avérer prohibitif étant donné le nombre de modèles proposés et les ressources informatiques disponibles. L'écart proportionnel diminue lorsque le nombre de réclamations fermées augmente (ou forcément si la réserve est évaluée à une date ultérieure). Par exemple, pour la stratégie **A1**, si le bootstrap de la réserve globale demandrait un réajustement des modèles, il y aurait approximativement $1{,}5 \times 10^7$ modèles à estimer pour 2 500 rééchantillonages.

Un avantage des modèles wGLM en comparaison aux modèles wCART est que la simulation est possible puisque les GLM sont assujetties à des hypothèses paramétriques. Ceci permet d'éviter le rééchantillonage des données pour les modèles des stratégies **A1** et **B**. En contrepartie, la simulation naïve n'est pas possible pour la stratégie **A2** puisque la régression logistique utilisée au dénominateur demande une réponse binaire. Ainsi, le rééchantillonage utilisé pour les modèles wCART est utilisé pour les modèles de la stratégie **A2** pour les modèles wGLM.

À cause de l'instabilité causée par l'ensemble $\mathcal{D}_T^1(y)$, il peut y avoir des prédictions élevées pour lesquelles une analyse de cas doit être effectuée. Les réclamations ayant les montants totaux prédits les plus élevés sont typiquement celles avec T_i grands. Ce qui complique l'analyse est la grande proportion des réclamations ayant des paiements totaux nuls. Par les nuages de points, il y a une concentration de l'erreur de la réserve prédite est concentrée autour d'une horizontale qui passe par zéro.

De plus, par la structure linéaire à une fonction près des GLM, il est possible de plus facilement déceler les paramètres influents sur la réserve prédite pour chaque réclamation ouverte. Les boîtes à moustaches de la Figure 6.14 montrent les distributions des estimateurs des paramètres pour les modèles wGLM individuels de la stratégie **A1**. Comme attendu, les paramètres liés aux paiements (`Medical Cov` et `Expense Cov`) sont importants par leurs moyennes élevées, mais la plus dispersée est la covariable `Loss Kind Name`, suivie de l'ordonnée à l'origine. Les GLM peuvent être généralisés afin d'expliquer certaines caractéristiques des données. Par exemple, afin de mieux capturer la proportion des réclamations avec montants payés nuls, les GLM peuvent être replacés par de modèles

gonflés à zéro ou encore pour permettre la sélection de variables et de réduire la variabilité des prédictions, on peut munir la fonction de vraisemblance de la contrainte LASSO (*least absolute shrinkage operator*).

6.4.4 L'écart quadratique moyen pour les modèles collectifs et les modèles individuels dans le temps

Typiquement en apprentissage statistique, les données sont séparées en deux ensembles nommés l'ensemble d'entraînement et l'ensemble de validation respectivement. Pour les stratégies **A1**, **A2** et **B**, cette séparation est seulement possible sur l'ensemble des réclamations fermées ce qui mène une implantation cohérente en pratique, mais étant donné le grand nombre de réclamations ouvertes à chaque date d'évaluation considérée, les données n'ont pas été isolées. Également, il serait prudent d'augmenter le seuil minimal du nombre de réclamations fermées incluses dans l'ensemble d'entraînement. Ultimement, une autre raison pour laquelle l'isolement des données n'a pas été effectué est que la réserve (globale ou individuelle) dépend évidemment des réclamations ouvertes seulement. Il serait pertinent de calculer le RMSE d'un ensemble de validation comprenant des réclamations fermées afin de connaître la performance du modèle en se disant des réclamations fermées.

Pour les fins de ce mémoire, rappelons que les réclamations qui sont ouvertes après 2017 ont été éliminées, alors toutes les réclamations ouvertes dont leurs réserves sont modélisées sont connues. En pratique, ce n'est forcément pas le cas et des méthodes d'évaluation de l'adéquation de modèles comme le *backtesting* sont utilisées. Le backtesting est une rétroaction qui demande de calculer les prédictions d'un modèle sur des données passées et de calculer des statistiques qui en découlent afin de déterminer son adéquation. Typiquement, le backtesting ne demande pas de réajuster le modèle afin d'arriver à une conclusion, mais dans le cas des modèles de provisionnement individuel présentés, il est nécessaire puisque, par exemple, une réclamation qui est fermée à la date d'évaluation `31-12-2015` sera peut-être ouverte à la date d'évaluation `31-12-2014`.

Stratégie	IPCW	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
A1	K-M	2 940 850	3 923 639	-6 922 869	9 645 851	14 569 817	29 975 601
		(2 942 793)					
A2	K-M	1 366 372	1 146 846	-2 717 290	3 242 737	3 919 576	5 809 175
		(1 339 037)					
	GLM	14 246 675	1 282 852	9 422 094	16 235 603	17 344 783	19 641 854
		(14 254 538)					
B	K-M	9 371 530	617 455	7 299 148	10 430 378	10 849 144	11 312 572
		(9 367 527)					
	GLM	28 334 587	830 353	25 326 872	29 739 298	30 307 505	30 977 679
		(28 353 446)					
Observé		4 714 570					

Note : Les réserves théoriques sont entre parenthèses.

(a) Évaluation à la date `31-12-2013`

Stratégie	IPCW	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
A1	K-M	12 382 332	2 527 424	5 608 453	16 770 108	19 243 471	23 558 013
		(12 382 532)					
A2	K-M	11 890 912	1 231 467	8 141 520	13 965 049	14 786 276	15 914 158
		(11 917 691)					
	GLM	11 165 220	1 112 207	7 858 517	12 959 252	13 750 114	15 238 389
		(11 188 628)					
B	K-M	8 589 670	595 872	6 494 312	9 584 969	10 085 066	10 663 628
		(8 592 100)					
	GLM	21 243 933	705 121	18 963 568	22 387 009	22 853 617	24 289 844
		(21 257 385)					
Observé		10 541 820					

Note : Les réserves théoriques sont entre parenthèses.

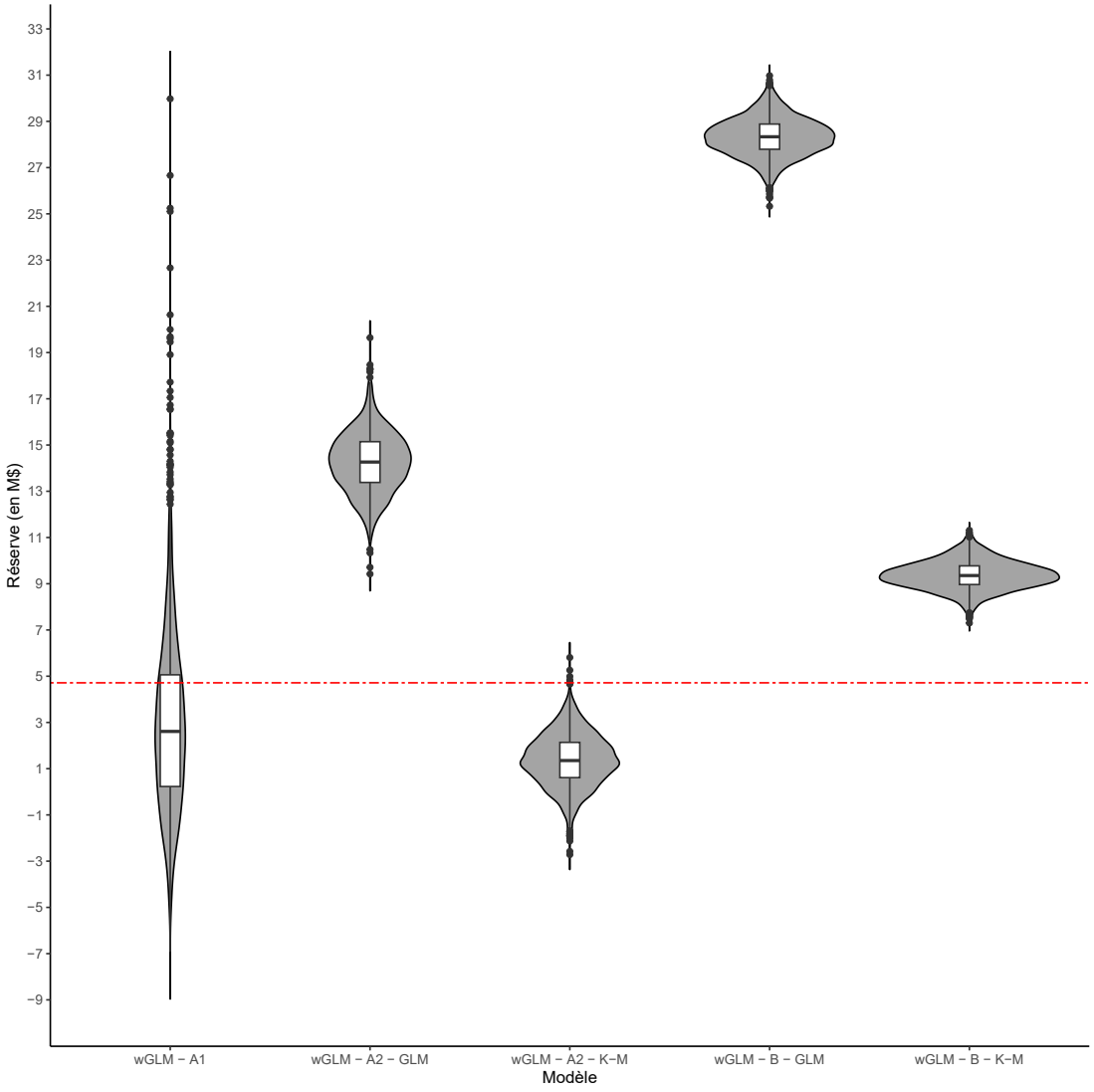
(b) Évaluation à la date `31-12-2014`

Stratégie	IPCW	Moyenne	Écart-type	Minimum	95^{e} quantile	99^{e} quantile	Maximum
A1	K-M	22 854 921	1 696 200	16 206 848	25 630 587	26 840 936	29 537 302
		(22 833 750)					
A2	K-M	22 192 882	939 487	18 745 342	23 753 461	24 345 456	26 232 823
		(22 206 725)					
	GLM	17 459 901	892 278	13 874 615	18 939 803	19 568 654	20 284 642
		(17 456 153)					
B	K-M	17 662 047	576 367	15 786 954	18 602 118	19 027 675	19 668 434
		(17 648 165)					
	GLM	21 665 081	556 572	19 826 356	22 601 944	22 966 333	23 525 574
		(21 678 158)					
Observé		24 604 946					

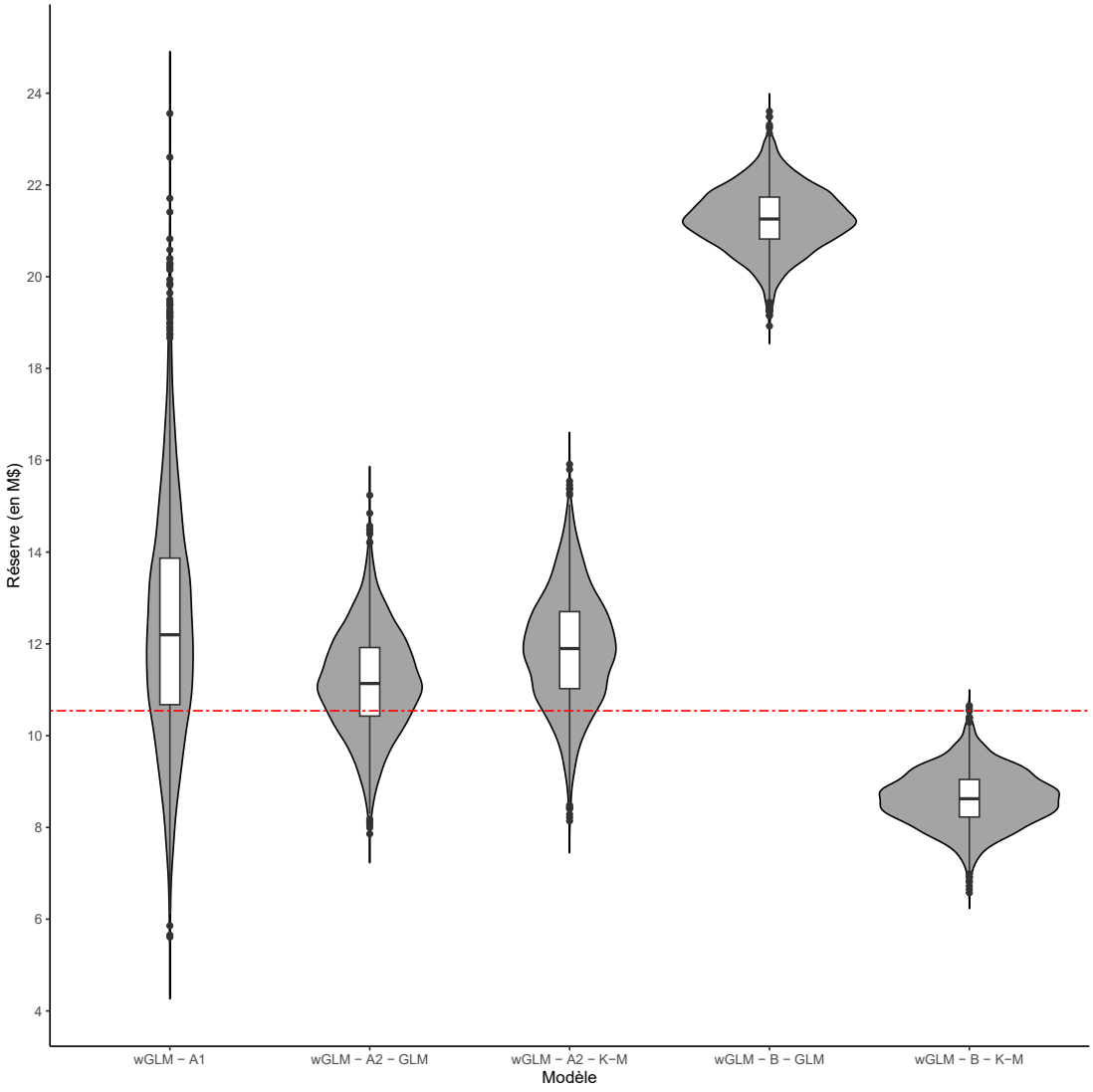
Note : Les réserves théoriques sont entre parenthèses.

(c) Évaluation à la date `31-12-2015`

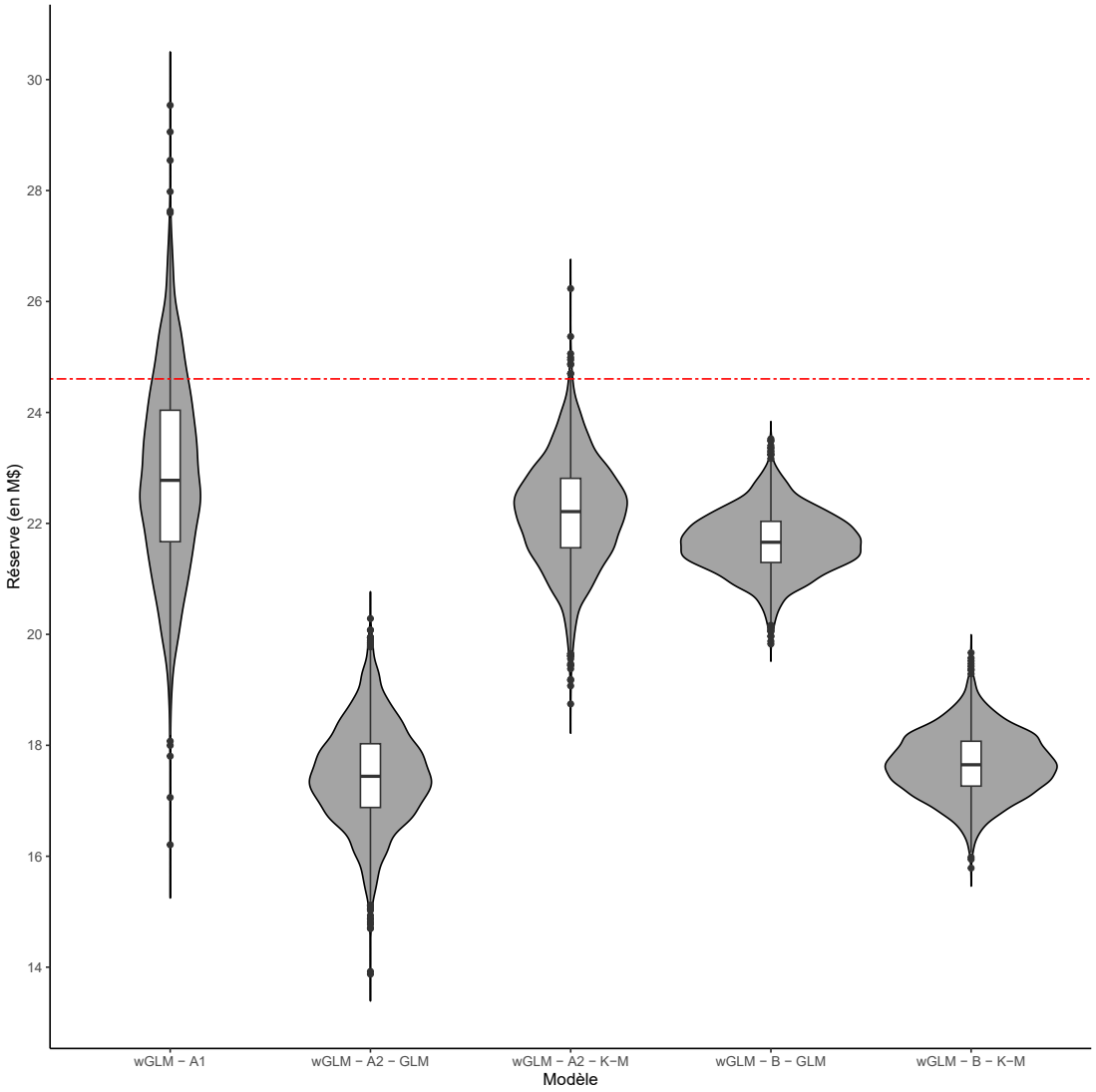
Tableau 6.9: Statistiques des simulations de la réserve RBNS pour les modèles individuels wGLM.



(a) Évaluation à 31-12-2013

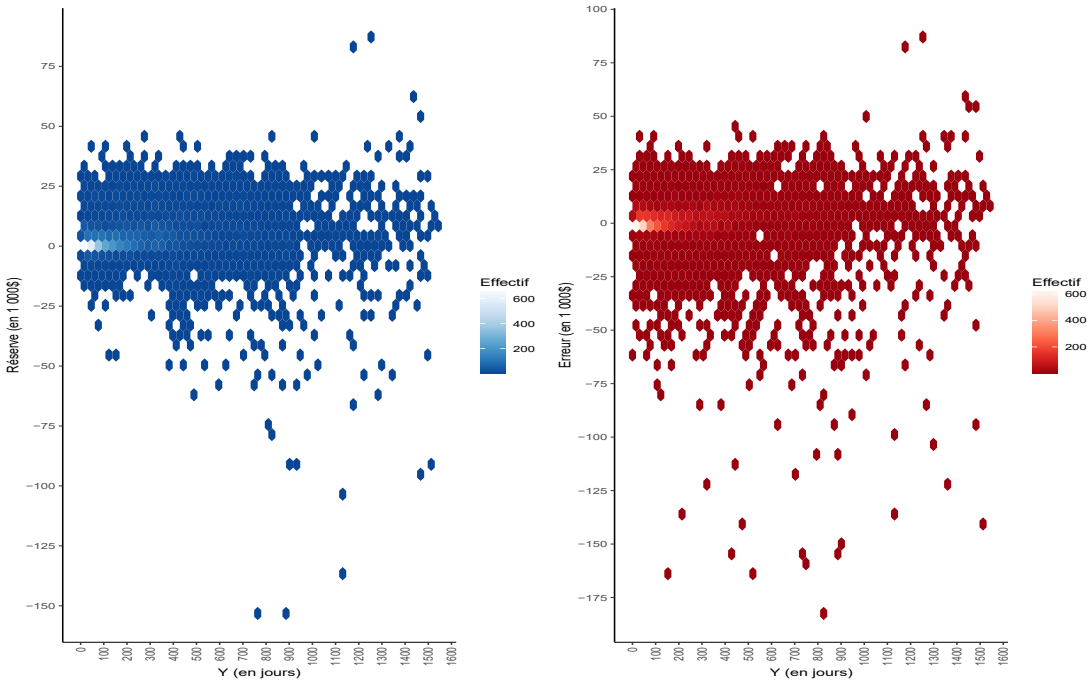


(b) Évaluation à la date `31-12-2014`



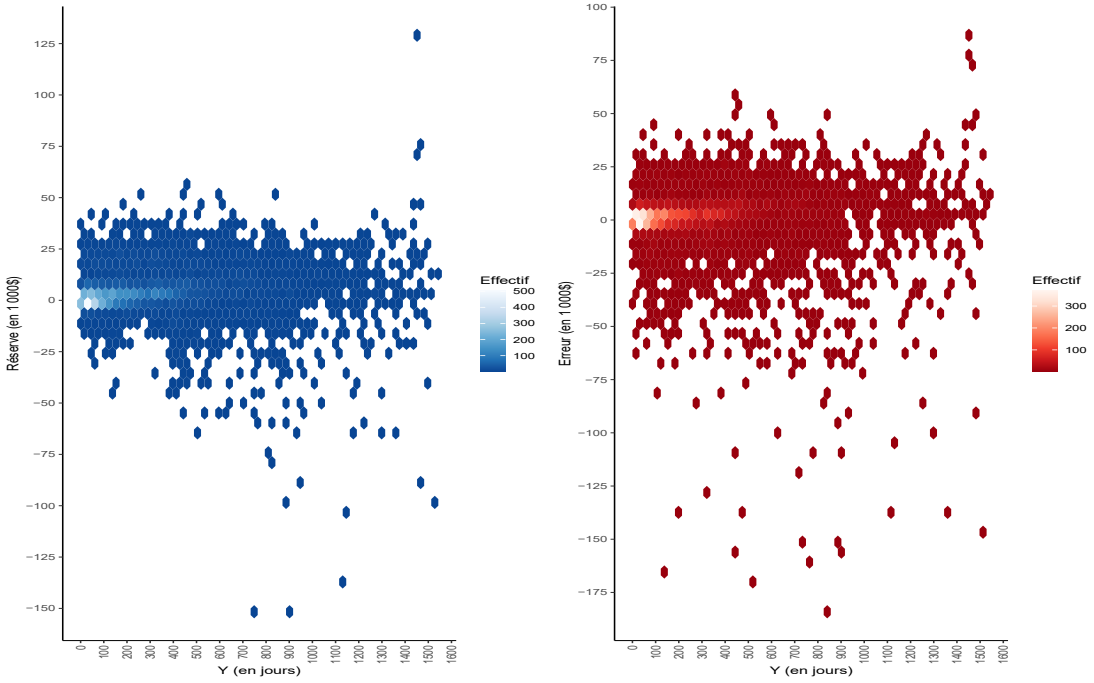
(c) Évaluation à la date `31-12-2015`

Figure 6.10: Distributions prédictives de la réserve globale RBNS pour les modéles linéaires généralisés pondérés (wGLM).

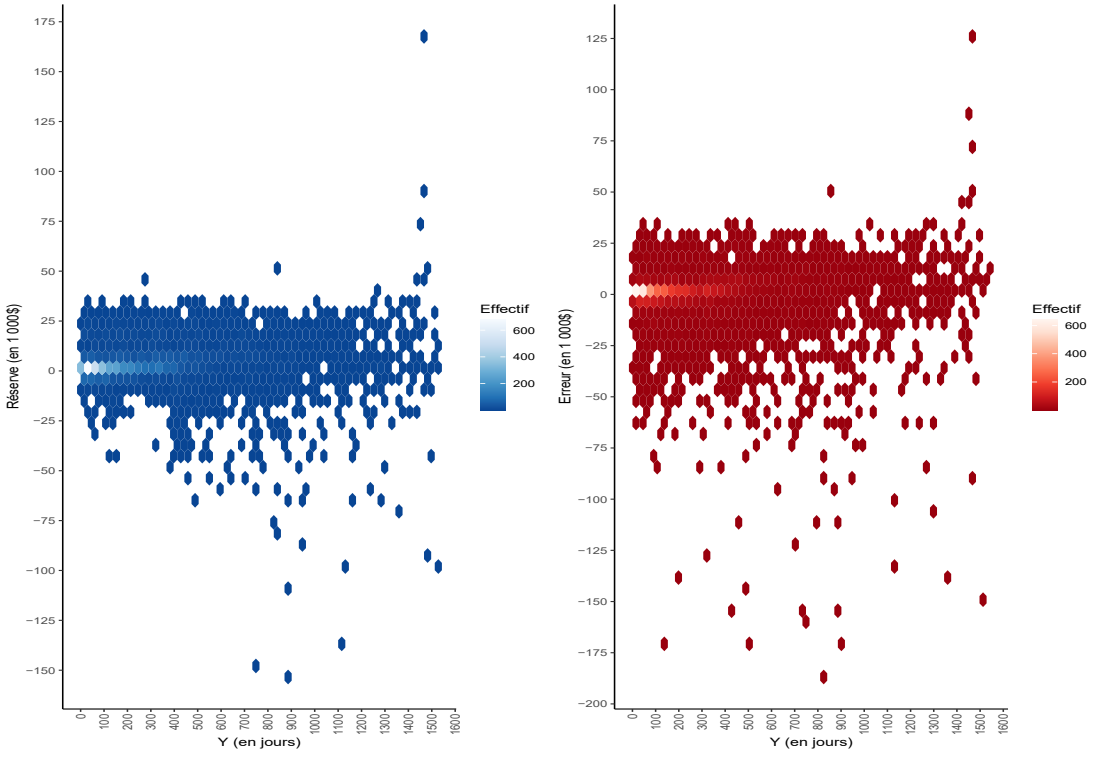


(a) Stratégie **A1** avec IPCW K-M

Figure 6.11: Réserve individuelle RBNS prédite à la date d'évaluation `31-12-2015` pour les GLM pondérés (wGLM) - Stratégie **A1**.

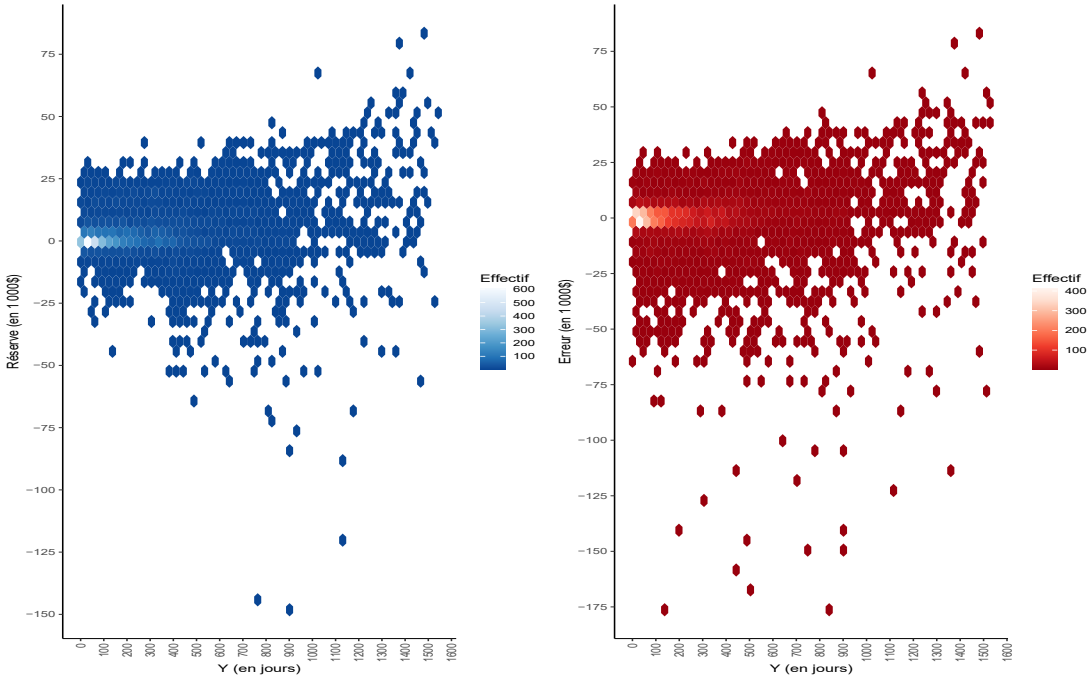


(a) Stratégie **A2** avec IPCW K-M

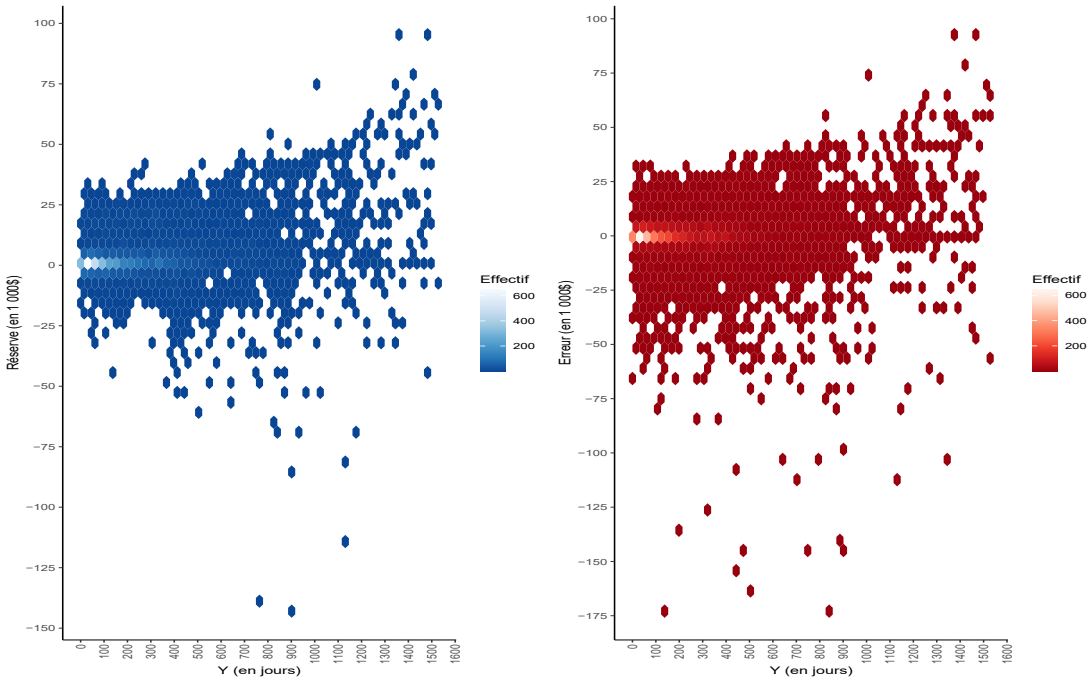


(b) Stratégie **A2** avec IPCW GLM

Figure 6.12: Réserve individuelle RBNS prédite à la date d'évaluation `31-12-2015` pour les GLM pondérés (wGLM) - Stratégie **A2**.



(a) Stratégie **B** avec IPCW K-M



(b) Stratégie **B** avec IPCW GLM

Figure 6.13: Réserve individuelle RBNS prédite à la date d'évaluation `31-12-2015` pour les GLM pondérés (wGLM) - Stratégie **B**.

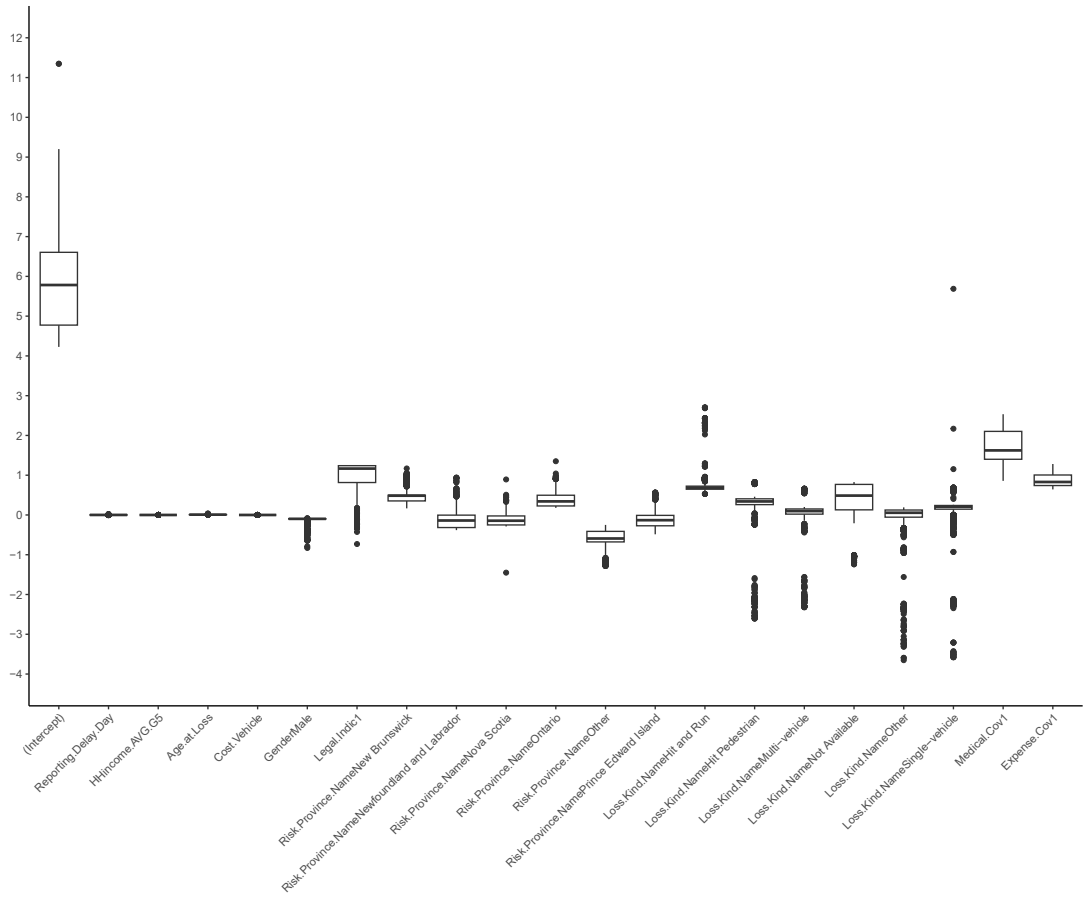


Figure 6.14: Estimés des paramètres des modèles wGLM de la stratégie **A1** calculés à la date d'évaluation `31-12-2015`.

Afin de connaître l'erreur de prédiction dans le temps, deux statistiques sont calculées pour les modèles individuels à chaque date d'évaluation de la réserve :

1. L'écart proportionnel de la réserve

$$\widehat{\epsilon} = \frac{\widehat{R} - R}{R};$$

2. La racine de l'erreur quadratique moyenne

$$\text{RMSE} = \sqrt{\frac{1}{\sum_{1i=1}^{n}(1-\delta_i)} \sum_{i=1}^{n}(1-\delta_i)(\widehat{R}_i - R_i)^2}.$$

Au Tableau 6.10, les statistiques ci-haut sont calculées pour chaque modèle (lorsque approprié) pour chaque date d'évaluation de la réserve. L'écart proportionnel est une mesure relative tandis que le RMSE est une mesure absolue. On remarque que le RMSE diminue avec le temps et que pour la majorité des modèles, l'écart proportionnel diminue également. L'exception incontestable est celle des modèles collectifs. À la date `31-12-2013`, les modèles collectifs sous-estiment la réserve globale et par après ils la sur-estime de plus en plus. Les modèles wCART ont des résultats mixtes, surtout pour les stratégies **A2** et **B**. La stratégie **A1** est la plus précise de même que la plus exacte puisqu'elle possède le plus petit RMSE et le plus petit écart relatif parmi les variantes des modèles wCART. Selon le même tableau, on observe que les modèles wGLM de la stratégie **A2** sont beaucoup plus précis que les modèles wCART équivalents et ont même des RMSE semblables au modèle wGLM de la stratégie **A1**.

En général, la stratégie **B** est l'approche la moins sensible et la stratégie **A1** offre les meilleurs résultats pour les modèles wCART et wGLM.

Modèles	Variantes	Statistiques	Date d'évaluation		
			31-12-2013	31-12-2014	31-12-2015
Collectifs	Mack	$\widehat{R}$	4 002 400	13 853 942	33 574 878
		$\widehat{\epsilon}$	-15,1%	31,4%	36,5%
	ODP	$\widehat{R}$	4 002 400	13 853 942	33 574 878
		$\widehat{\epsilon}$	-15,1%	31,4%	36,5%
	Tweedie	$\widehat{R}$	4 191 603	13 856 507	33 590 080
		$\widehat{\epsilon}$	-11,1%	31,4%	36,5%
wCART	**A1** – K-M	$\widehat{R}$	4 457 723	10 697 365	23 264 205
		$\widehat{\epsilon}$	-5,4%	1,5%	-5,4%
		RMSE	15 672,9	15 428,3	14 063,4
	A1 – GLM	$\widehat{R}$	4 409 068	10 549 051	23 456 367
		$\widehat{\epsilon}$	-6,5%	0,1%	-4,7%
		RMSE	15 758,2	15 430,1	14 030,7
	A2 – K-M	$\widehat{R}$	68 164 664	70 292 100	93 397 420
		$\widehat{\epsilon}$	1 345,8%	566,8%	279,6%
		RMSE	64 404,6	58 835,7	59 730,2
	A2 – GLM	$\widehat{R}$	64 932 912	70 888 190	94 228 037
		$\widehat{\epsilon}$	1 277,3%	572,4%	283,0%
		RMSE	62 762,0	60 287,7	50 098,1
	B – K-M	$\widehat{R}$	14 603 891	4 344 172	19 254 104
		$\widehat{\epsilon}$	209,8%	-58,8%	-21,7%
		RMSE	20 606,5	16 171,8	14 541,0
	B – GLM	$\widehat{R}$	21 000 611	12 603 496	19 641 482
		$\widehat{\epsilon}$	345,4%	19,6%	-20,2%
		RMSE	17 983,2	16 695,4	14 707,8
wGLM	**A1** – K-M	$\widehat{R}$	2 942 793	12 382 532	22 833 750
		$\widehat{\epsilon}$	-37,6%	17,5%	-7,2%
		RMSE	16 237,3	16 491,3	14 527,5
	A2 – K-M	$\widehat{R}$	1 339 037	11 917 691	22 206 725
		$\widehat{\epsilon}$	-71,6%	13,1%	-9,7%
		RMSE	16 461,7	17 427,8	14 771,7
	A2 – GLM	$\widehat{R}$	14 254 538	11 188 628	17 456 153
		$\widehat{\epsilon}$	202,4%	6,1%	-29,1%
		RMSE	17 564,4	16 309,9	14 648,0
	B – K-M	$\widehat{R}$	9 367 527	8 610 610	17 648 165
		$\widehat{\epsilon}$	98,7%	-18,3%	-28,3%
		RMSE	16 798,5	16 160,7	14 451,1
	B – GLM	$\widehat{R}$	28 353 446	21 269 773	21 678 158
		$\widehat{\epsilon}$	501,4%	101,8%	-11,9%
		RMSE	19 787,1	17 148,0	14 717,2
	Observé		4 714 570	10 541 820	24 604 946

Tableau 6.10: Prédictions et mesures d'adéquation pour les modèles collectifs et les modèles individuels pour différentes dates d'évaluation.

CONCLUSION

Dans le présent mémoire, des modèles ont été proposés pour estimer la réserve individuelle de réclamations en tenant compte de la nature censurée des données individuelles. Les modèles collectifs utilisés en industrie regroupent généralement les informations individuelles par année d'accident et par année de développement. Les modèles individuels permettent l'utilisation des informations granulaires portant sur la réclamation et/ou sur le réclamant et tentent de modéliser la réserve par une méthode d'apprentissage statistique. Il est possible que plus d'une méthode soit nécessaire dans le cas des modèles en cascade afin de saisir le processus de développement d'une réclamation.

Plusieurs stratégies ont été abordées afin de modéliser la réserve espérée conditionnellement au délai de fermeture. Les modèles d'arbres de régression de l'article (Lopez et Milhaud, 2021) ont été appliqués à une base de données `ausautoBI8999` qui comprend peu de variables explicatives utilisables et pour laquelle la construction de la base de données $\mathcal{D}_{\mathcal{T}}(y)$ pour réclamation ouverte n'est pas décrite. Avec la base de données réelles, une instabilité est produite dans les prédictions de la réserve individuelle selon le modèle de prédiction. L'article (Lopez et Milhaud, 2021) fait une étude de la stratégie **A2** tandis que la stratégie **A1** de ce projet semble plus stable tout en assurant un pouvoir de prédiction raisonnable. Ceci est validé par la simulation des distributions prédictives de ces stratégies ainsi que l'évaluation de celles-ci pour plusieurs dates d'évaluation.

Afin d'améliorer les modèles wCART, il est nécessaire de se soucier des hyperparamètres pour assurer la stabilité de chaque modèle individuel. En contrepartie, pour les modèles wGLM, les modèles pour chacune des réclamations ouvertes peuvent être modifiées au besoin. De plus, les modèles présentés pour ce projet supposent que l'historique des paiements pour chaque réclamation n'a aucun impact sur la modélisation. Au moment du calcul de la réserve, l'entrée la plus récente de la réclamation en question est prise.

L'inclusion des données passées nécessite un calcul de poids pour chaque période de développement comme propose l'article (Bang et Tsiatis, 2000) qui propose un estimateur cloisonné pour estimer des frais de soins d'une étude en cardiologie. Les poids IPCW des réclamations ouvertes sont nuls alors seulement les réclamations fermées ont des poids positifs. Les poids AIPCW (*augmented inverse probability censoring weights*) sont des poids IPCW modifiés qui permettent que les réclamations ouvertes soient attribuées un poids non nul. Plusieurs modèles peuvent améliorer le GLM en pénalisant la fonction de vraisemblance ou encore en modifiant la structure linéaire de la régression conditionnelle. Par exemple, il est possible de possible de poser un modèle GLM pondéré avec pénalisation LASSO. Ceci permettrait une sélection de variables implicite pour chaque modèle (dont chaque réclamation ouverte). Une autre amélioration proposée serait l'utilisation des GAM (*generalized additive model*) pondérés à l'aide de l'algorithme REML (*restricted maximum likelihood*) au lieu des wGLM.

L'approche de modélisation avec les poids IPCW n'est pas très avancée dans la littérature actuarielle, mais ce mémoire montre la flexibilité de cette approche pour différentes méthodes basées sur l'optimisation de la fonction de vraisemblance.

APPENDICE A

HYPOTHÈSES DE LA FONCTION DE VRAISEMBLANCE PONDÉRÉE

Les hypothèses, résultats et théorèmes suivants sont issus de (Lehmann et Casella, 2003) et permettront d'établir la convergence et la convergence en distribution à une variable aléatoire normale de l'estimateur $\mathrm{MV}_{\mathrm{IPCW}}$.

1. Le support de $\mathcal{L}^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y})$ ne dépend pas de $\boldsymbol{\theta}$.
2. Pour un sous-ensemble ouvert $\boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}$ de l'ensemble $\boldsymbol{\Theta}$ des paramètres, on suppose que pour le vecteur qui génère les données $\boldsymbol{\theta}_0 \subset \boldsymbol{\Theta}_1$.
3. Pour tout $\mathbf{y}$, les dérivées partielles

$$\frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} \mathcal{L}^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y}), \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta}_1;$$

   existent pour tous les triplets (i, j, k).
4. Les premières et secondes dérivées partielles de la fonction $\ell^{\mathbf{w}}$ respectent

$$E\left[\frac{\partial}{\partial\theta_j} \ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y})\right] = 0, \qquad j = 1, \ldots, s$$

   et

$$\mathcal{I}^{\mathbf{w}}(\boldsymbol{\theta})_{(j,k)} = -E\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y})\right], \qquad j, k = 1, \ldots, s,$$

   pour toutes paires (j, k).
5. La matrice d'information de Fisher $\mathcal{I}^{\mathbf{w}}(\boldsymbol{\theta})$ est une matrice de covariance et les $\frac{\partial}{\partial\theta_j} \ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y})$ sont linéairement indépendants.

6. Il existe des fonctions $M_{i,j,k}(\mathbf{y})$ telles que

$$\left| \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} \ell^{\mathbf{w}}(\boldsymbol{\theta} \mid \mathbf{y}) \right| \leq M_{i,j,k}(\mathbf{y}), \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$$

et $E[M_{i,j,k}(\mathbf{y})]$ existe pour chaque triplet (i, j, k).

APPENDICE B

MODÉLISATION DE L'INFLATION

Année de fermeture (j)	$\widehat{\alpha}_j$	$\widehat{\beta}_j$	n_j
1	6,612761	$3{,}975010 \times 10^{-3}$	39 413
2	9,130198	$-4{,}961594 \times 10^{-2}$	6 060
3	9,751138	$-6{,}936079 \times 10^{-3}$	1 518
4	10,208992	$-1{,}063980 \times 10^{-1}$	553
5	10,306073	$-2{,}180566 \times 10^{-1}$	185
6	10,625873	$3{,}883026 \times 10^{-16}$	56

Tableau B.1: Paramètres de la régression linéaire pondérée de $\log(M_i')$.

Il s'ensuit que le taux d'inflation ($\widehat{\beta}$) est calculé par la moyenne pondérée

$$\widehat{\beta} = \frac{\sum_j \sqrt{n_j}\widehat{\beta}_j}{\sum_j \sqrt{n_j}} = -0{,}02448061.$$

APPENDICE C

RÉGRESSION LOGISTIQUE POUR LE CALCUL DES POIDS IPCW

Paramètre	Estimation
Intercept	3,131522
Reporting Delay Day	$-2{,}464981 \times 10^{-3}$
HHincome AVG G5	$2{,}503350 \times 10^{-6}$
Age at Loss	$6{,}023058 \times 10^{-3}$
Cost Vehicle	$-1{,}171486 \times 10^{-5}$
GenderMale	$2{,}812196 \times 10^{-2}$
Legal Indic1	$-1{,}061877$
Risk Province NameNew Brunswick	$-0{,}8410766$
Risk Province NameNewfoundland and Labrador	$-1{,}447443$
Risk Province NameNova Scotia	$-1{,}040003$
Risk Province NameOntario	$-0{,}9677812$
Risk Province NameOther	$-0{,}3851104$
Risk Province NamePrince Edward Island	$-1{,}127275$
Loss Kind NameHit and Run	$0{,}3{,}064112$
Loss Kind NameHit Pedestrian	$-0{,}3159778$
Loss Kind NameMulti-vehicle	$-0{,}2453835$
Loss Kind NameNot Available	10,14479
Loss Kind NameOther	$-0{,}1668918$
Loss Kind NameSingle-vehicle	$6{,}004827 \times 10^{-3}$
Medical Cov1	$-0{,}9480709$
Expense Cov1	$-0{,}9342935$
N	$-2{,}577538 \times 10^{-5}$
Y	$1{,}642986 \times 10^{-3}$

(a) Date d'évaluation 31-12-2013

Paramètre	Estimation
Intercept	3,452358
Reporting Delay Day	$-3{,}222137 \times 10^{-3}$
HHincome AVG G5	$1{,}394740 \times 10^{-6}$
Age at Loss	$6{,}570708 \times 10^{-3}$
Cost Vehicle	$-1{,}176394 \times 10^{-5}$
GenderMale	$4{,}778123 \times 10^{-2}$
Legal Indic1	$-1{,}034952$
Risk Province NameNew Brunswick	$-0{,}9636590$
Risk Province NameNewfoundland and Labrador	$-1{,}342032$
Risk Province NameNova Scotia	$-0{,}9550900$
Risk Province NameOntario	$-0{,}8176177$
Risk Province NameOther	$-0{,}5992494$
Risk Province NamePrince Edward Island	$-1{,}154725$
Loss Kind NameHit and Run	$-2{,}496548 \times 10^{-2}$
Loss Kind NameHit Pedestrian	$-0{,}3897279$
Loss Kind NameMulti-vehicle	$-0{,}3443906$
Loss Kind NameNot Available	9,838295
Loss Kind NameOther	$-0{,}3241885$
Loss Kind NameSingle-vehicle	$-9{,}189961 \times 10^{-2}$
Medical Cov1	$-0{,}8429517$
Expense Cov1	$-1{,}289253$
N	$-1{,}668502 \times 10^{-5}$
Y	$1{,}717967 \times 10^{-3}$

(b) Date d'évaluation 31-12-2014

Paramètre	Estimation
Intercept	3,395592
Reporting Delay Day	$-2{,}889383 \times 10^{-3}$
HHincome AVG G5	$1{,}380557 \times 10^{-6}$
Age at Loss	$3{,}194297 \times 10^{-3}$
Cost Vehicle	$-1{,}018388 \times 10^{-5}$
GenderMale	$4{,}194008 \times 10^{-2}$
Legal Indic1	$-1{,}253844$
Risk Province NameNew Brunswick	$-0{,}7329279$
Risk Province NameNewfoundland and Labrador	$-1{,}084052$
Risk Province NameNova Scotia	$-0{,}6870953$
Risk Province NameOntario	$-0{,}8528809$
Risk Province NameOther	$-0{,}3002838$
Risk Province NamePrince Edward Island	$-0{,}9893575$
Loss Kind NameHit and Run	$-9{,}724480 \times 10^{-2}$
Loss Kind NameHit Pedestrian	$-0{,}1211623$
Loss Kind NameMulti-vehicle	$-0{,}1078580$
Loss Kind NameNot Available	9,878449
Loss Kind NameOther	$-0{,}1136019$
Loss Kind NameSingle-vehicle	0,1326069
Medical Cov1	$-0{,}8549465$
Expense Cov1	$-1{,}115010$
N	$1{,}775164 \times 10^{-6}$
Y	$1{,}639759 \times 10^{-3}$

(c) Date d'évaluation 31-12-2015

Tableau C.1: Paramètres du modèle logistique de l'indicatrice $\mathbb{1}(T_i \leq C_i)$.

APPENDICE D

PROFONDEURS DES ARBRES DE RÉGRESSION INDIVIDUELS PONDÉRÉS PAR LES POIDS IPCW

	Profondeur de l'arbre							
IPCW	2	3	4	5	7	8	9	10
K-M	2	134	4 188	1 377	10	3	35	2
GLM	2	138	4 172	1 381	13	1	41	3

(a) Stratégie **A1**

	Profondeur de l'arbre									
IPCW	0	2	3	4	5	6	7	8	9	10
K-M	1 013	12	119	210	740	317	399	447	415	2 079
GLM	1 012	10	114	208	751	340	434	442	416	2 024

(b) Stratégie **A2** - Dénominateur

	Profondeur de l'arbre					
IPCW	2	3	4	5	7	8
K-M	0	347	3 841	1 560	3	0
GLM	2	345	3 825	1 573	5	1

(c) Stratégie **A2** - Numérateur

	Profondeur de l'arbre										
IPCW	0	1	2	3	4	5	6	7	8	9	10
K-M	182	210	502	126	25	251	474	1 138	420	1 240	1 183

(d) Stratégie **A1** - Délai de fermeture

Tableau D.1: Profondeur des arbres (élagués) de régression pondérés du montant total payé et du délai de fermeture à la date d'évaluation `31-12-2013`.

	Profondeur de l'arbre									
IPCW	1	2	3	4	5	6	7	8	9	10
K-M	18	104	93	283	2 059	3 490	23	15	49	13
GLM	15	108	83	293	2 087	3 459	26	16	52	8

(a) Stratégie **A1**

	Profondeur de l'arbre									
IPCW	0	2	3	4	5	6	7	8	9	10
K-M	982	2	4	216	1 175	219	157	216	679	2 497
GLM	971	3	1	216	1 139	207	157	214	710	2 529

(b) Stratégie **A2** - Dénominateur

	Profondeur de l'arbre								
IPCW	1	2	3	4	5	6	7	9	10
K-M	40	64	362	286	1 066	4 189	39	87	14
GLM	29	75	357	294	1 111	4 106	45	104	26

(c) Stratégie **A2** - Numérateur

	Profondeur de l'arbre										
IPCW	0	1	2	3	4	5	6	7	8	9	10
K-M	422	5	162	80	142	708	107	746	1 496	1 794	485

(d) Stratégie **A1** - Délai de fermeture

Tableau D.2: Profondeur des arbres (élagués) de régression pondérés du montant total payé et du délai de fermeture à la date d'évaluation `31-12-2014`.

	Profondeur de l'arbre										
IPCW	0	1	2	3	4	5	6	7	8	9	10
K-M	4	31	46	145	1 402	1 514	23	45	1	3 220	24
GLM	3	26	49	131	1 398	1 551	22	42	0	3 217	16

(a) Stratégie **A1**

	Profondeur de l'arbre								
IPCW	0	3	4	5	6	7	8	9	10
K-M	1 061	15	177	1 211	99	206	424	520	2 742
GLM	1 379	12	136	1 431	76	158	326	400	2 537

(b) Stratégie **A2** - Dénominateur

	Profondeur de l'arbre										
IPCW	0	1	2	3	4	5	6	7	8	9	10
K-M	4	61	78	143	473	1 315	23	10	19	4 262	67
GLM	4	59	90	140	428	1 293	28	10	21	4 318	64

(c) Stratégie **A2** - Numérateur

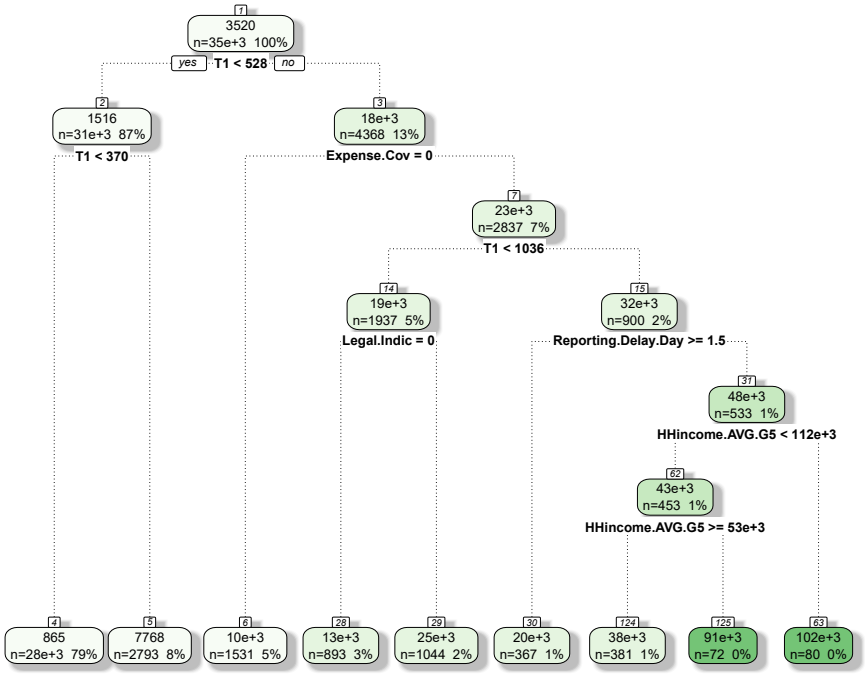
	Profondeur de l'arbre										
IPCW	0	1	2	3	4	5	6	7	8	9	10
K-M	413	49	321	151	243	353	47	492	1 576	2 325	485

(d) Stratégie **A1** - Délai de fermeture

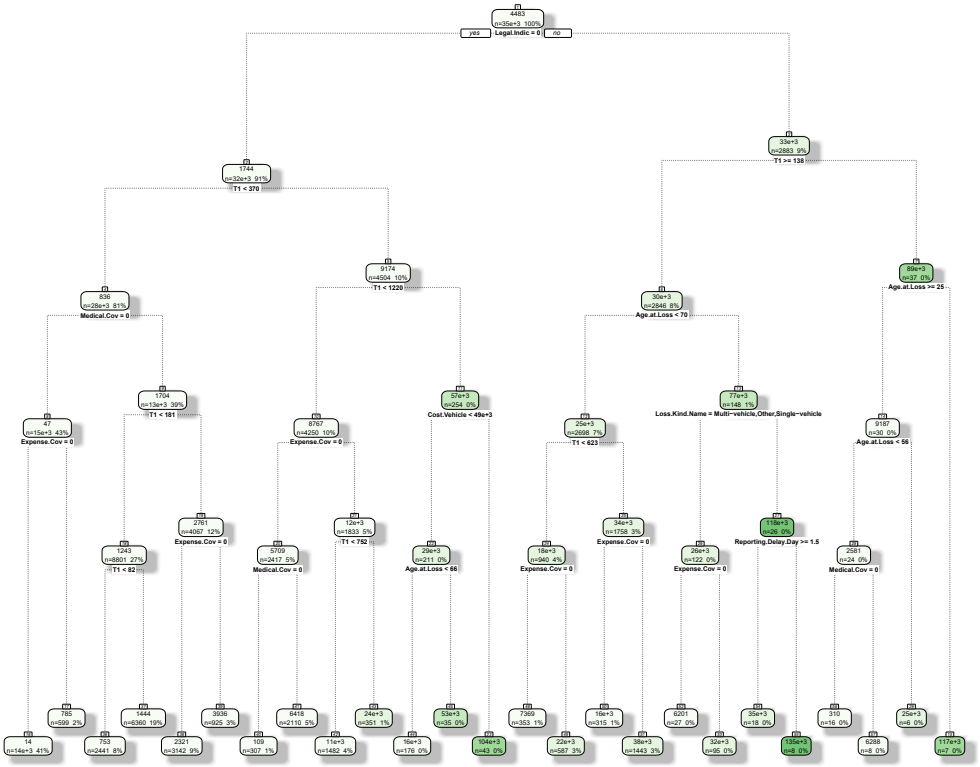
Tableau D.3: Profondeur des arbres (élagués) de régression pondérés du montant total payé et du délai de fermeture à la date d'évaluation `31-12-2015`.

APPENDICE E

ARBRES DE RÉGRESSION PONDÉRÉS DE LA STRATÉGIE **B**

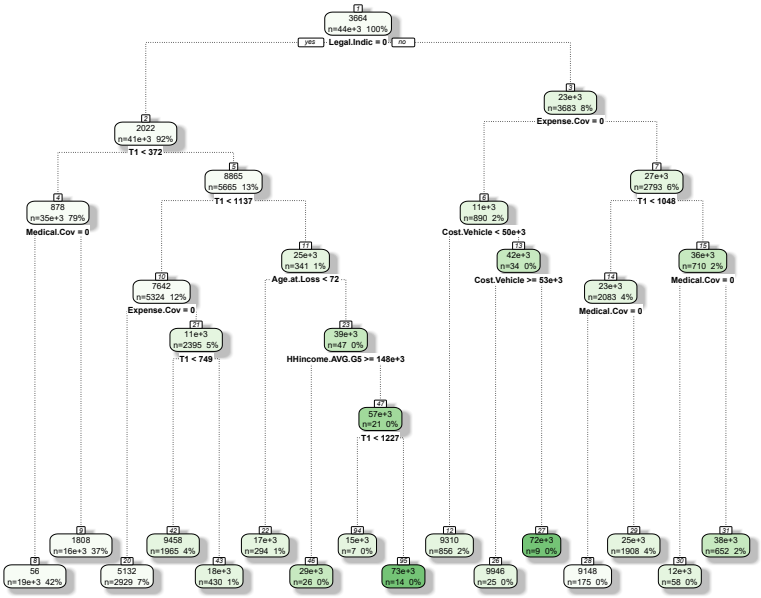


(a) Stratégie **B** avec poids IPCW K-M



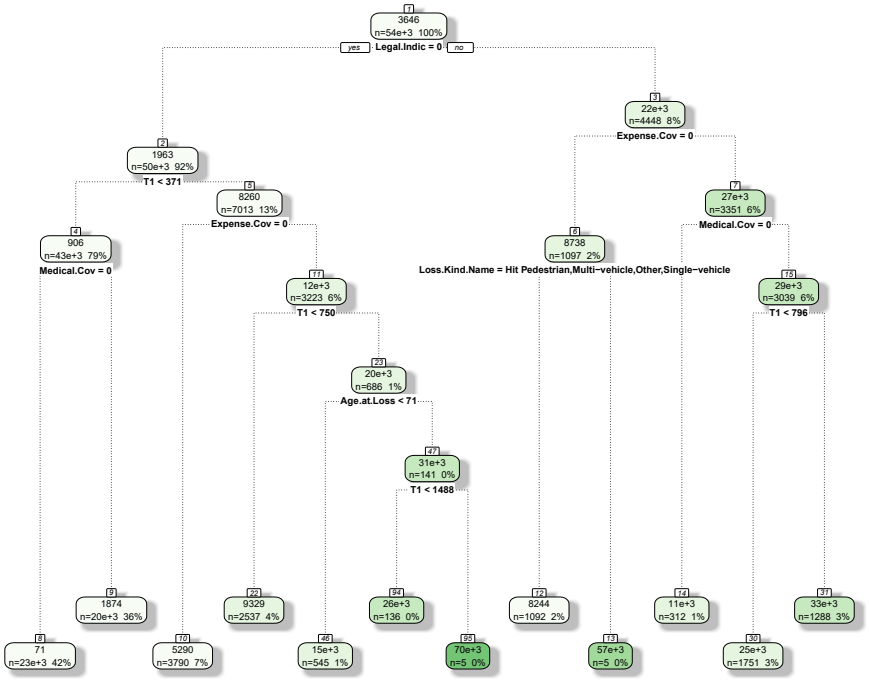
(b) Stratégie **B** avec poids IPCW GLM

Figure E.1: Date d'évaluation `31-12-2013`

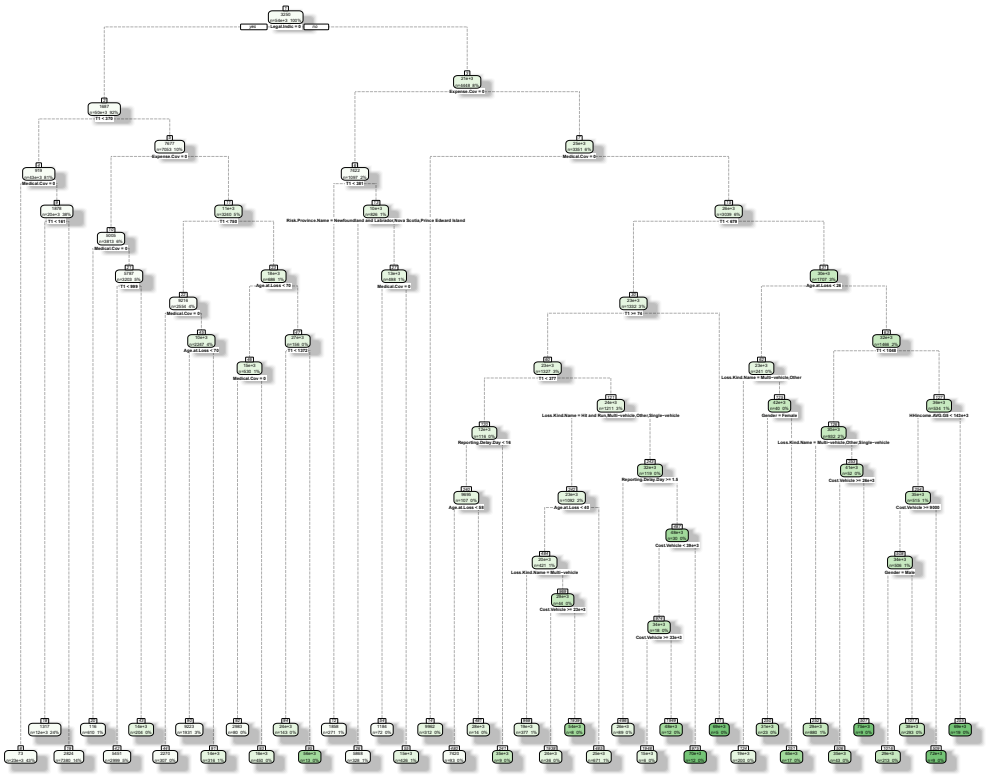


(b) Stratégie **B** avec poids IPCW GLM

Figure E.2: Date d'évaluation `31-12-2014`



(a) Stratégie **B** avec poids IPCW K-M



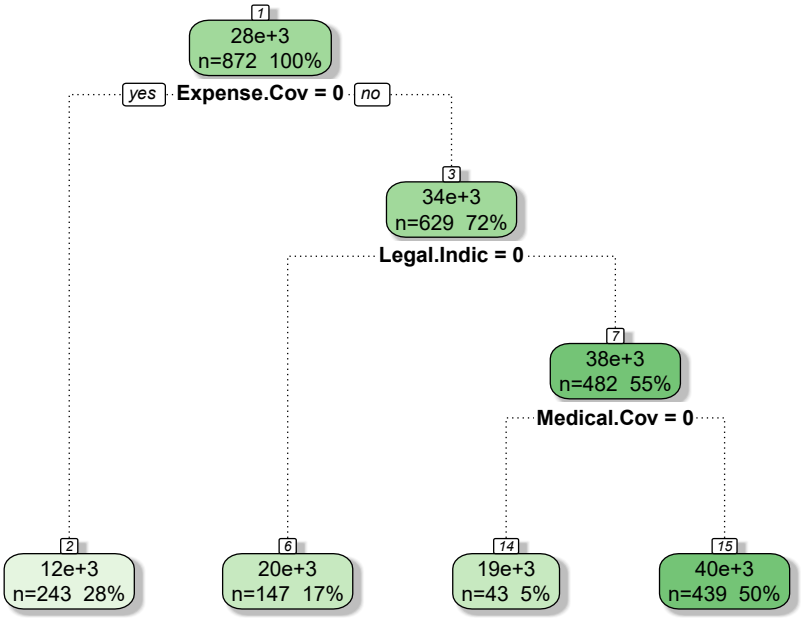
(b) Stratégie **B** avec poids IPCW GLM

Figure E.3: Date d'évaluation `31-12-2015`

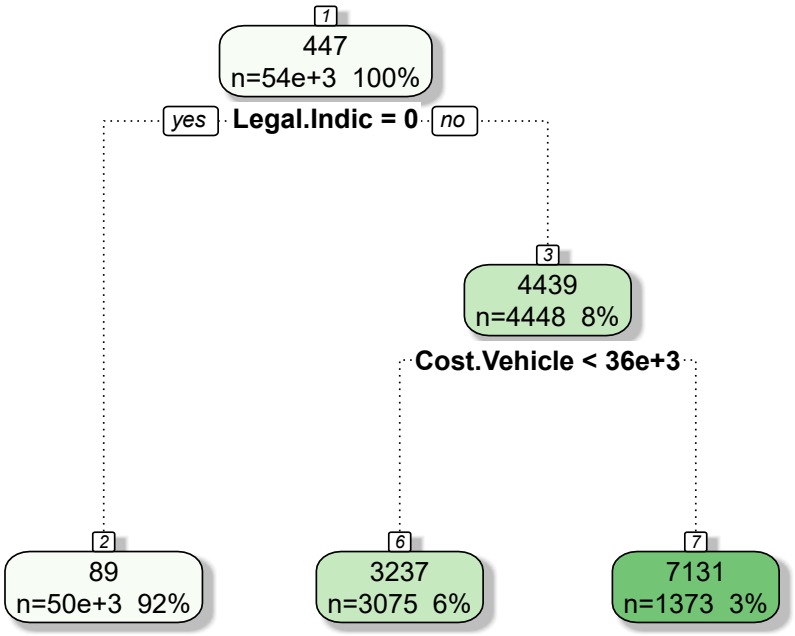
APPENDICE F

ARBRES ÉLAGUÉS DE LA STRATÉGIE **A1** ET **A2** POUR UNE RÉCLAMATION OUVERTE

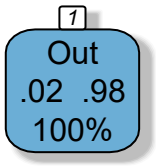
Les modèles ci-bas prédisent le montant total payé d'une réclamation ouverte à la date d'évaluation de la réserve `31-12-2015`. Les prédictions sont 39 871,33 et 200 886,40 selon les stratégies **A1** et **A2** respectivement.



(a) Stratégie **A1** avec poids IPCW K-M



(b) Numérateur de la stratégie **A2** avec poids IPCW K-M



(c) Dénominateur de la stratégie **A2** avec poids IPCW K-M

RÉFÉRENCES

Antonio, K. et Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, *2014*(7), 649–669. http://dx.doi.org/10.1080/03461238.2012.755938

Bang, H. et Tsiatis, A. (2000). Estimating medical costs with censored data. *Biometrika*, *87*(2), 329–343. http://dx.doi.org/10.1093/biomet/87.2.329. Récupéré de https://doi.org/10.1093/biomet/87.2.329

Blier-Wong, C., Cossette, H., Lamontagne, L. et Marceau, E. (2021). Machine learning in P&C insurance : A review for pricing and reserving. *Risks*, *9*(1). http://dx.doi.org/10.3390/risks9010004

Breiman, L., Friedman, J. H. et Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall.

Breslow, N. E. (2007). Generalized linear models : Checking assumptions and strengthening conclusions.

Charpentier, A. et Pigeon, M. (2016). Macro vs. micro methods in non-life claims reserving (an econometric perspective). *Risks*, *4*(2). http://dx.doi.org/10.3390/risks4020012

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York, NY, USA : Springer-Verlag.

Efron, B. et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Numéro 57 de monographs on Statistics and Applied Probability. Boca Raton, Florida, USA : Chapman & Hall/CRC.

Enders, C. (2010). *Applied Missing Data Analysis*. Methodology in the social sciences. Guilford Publications. Récupéré de https://books.google.ca/books?id=MN8ruJd2tvgC

England, P. et Verrall, R. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, *8*(3), 443–544. http://dx.doi.org/10.1017/S1357321700003809

Gabrielli, A. (2021). An individual claims reserving model for reported claims. *European Actuarial Journal*, *11*, 541–577. http://dx.doi.org/10.1007/s13385-021-00271-4

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society : Series B (Methodological)*, *49*(2), 127–145. http://dx.doi.org/https://doi.org/10.1111/j.2517-6161.1987.tb01685.x

Lehmann, E. et Casella, G. (2003). *Theory of Point Estimation*. Springer Texts in Statistics. Springer New York. Récupéré de https://books.google.ca/books?id=9St7DCbu9AUC

Little, R. J. A. et Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. USA : John Wiley & Sons, Inc.

Lopez, O. et Milhaud, X. (2021). Individual reserving and nonparametric estimation of claim amounts subject to large reporting delays. *Scandinavian Actuarial Journal*, *2021*(1), 34–53. http://dx.doi.org/10.1080/03461238.2020.1793218

Lopez, O., Milhaud, X. et Thérond, P.-E. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, *10*(2), 2685–2716. http://dx.doi.org/10.1214/16-EJS1189

Lopez, O., Milhaud, X. et Thérond, P.-E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, *49*(3), 741–762. http://dx.doi.org/10.1017/asb.2019.12

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, *23*(2), 213–225. http://dx.doi.org/10.2143/AST.23.2.2005092

Matsouaka, R. A. et Atem, F. D. (2020). Regression with a right-censored predictor using inverse probability weighting methods. *Statistics in Medicine*, *39*(27), 4001–4015. http://dx.doi.org/10.1002/sim.8704

McCullagh, P. et Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

McGuire, G., Taylor, G. et Miller, H. (2018). Self-assembling insurance claim models using regularized regression and machine learning. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.3241906

Renshaw, A. et Verrall, R. (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, *4*(4), 903–923.

Sharara, I., Hardy, M. et Saunders, D. (2010). *Regulatory Capital Standards for Property and Casaulty Insurer under U.S., Canadian and Proposed Solvency II (Standard) Formulas*. Rapport technique, Society of Actuaries.

Taylor, G. (2019). Loss reserving models : Granular and machine learning forms. *Risks*, *7*, 82. http://dx.doi.org/10.3390/risks7030082

Taylor, G. et McGuire, G. (2016). *Stochastic Loss Reserving Using Generalized Linear Models*, volume 3 de *CAS Monograph Series*.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York. Récupéré de https://books.google.ca/books?id=xqZFi2EMB40C

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G. et O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data : A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, *61*, 119–131. http://dx.doi.org/10.1016/j.jbi.2016.03.009

Wahl, F., Lindholm, M. et Verrall, R. (2019). The collective reserving model. *Insurance : Mathematics and Economics*, *87*, 34–50.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, *63*(1), 27–32. http://dx.doi.org/10.2307/2335080

Wüthrich, M. V. (2016). *Machine Learning in Individual Claims Reserving*. Swiss Finance Institute Research Paper Series 16-67, Swiss Finance Institute

Wüthrich, M. V. et Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. John Wiley & Sons.