UNIVERSITÉ DU QUÉBEC À MONTRÉAL

EXPÉRIENCE D'ASSURANCE ET DURÉE DE COUVERTURE DANS LA TARIFICATION SOUS UNE LOI DE TWEEDIE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN MATHÉMATIQUES

PAR

RAÏSSA COULIBALY

JUIN 2026

REMERCIEMENTS

J'adresse mes sincères remerciements à Jean-Philippe Boucher, mon directeur de doctorat à l'UQAM, pour toutes les connaissances qu'il m'a transmises ainsi que pour le soutien qu'il m'a apporté tout au long de cette thèse. Je remercie également son collègue Mathieu Pigeon, professeur à l'UQAM, et, à travers lui, toute l'équipe de la Chaire Co-operators en analyse des risques actuariels de l'UQAM ainsi que la compagnie Co-operators pour le soutien qui a permis la réalisation de cette thèse.

Je remercie aussi chaleureusement tous les anciens et nouveaux étudiants de la chaire pour les bons moments passés ensemble et pour leur contribution à cette thèse. Un clin d'œil également à tous les autres étudiants de l'UQAM et d'ailleurs qui ont contribué, directement ou indirectement, à ce parcours doctoral.

J'adresse également mes sincères remerciements à Philippe Lambert, mon directeur de mémoire de master à l'UCLouvain (Belgique), sans qui ce projet de doctorat n'aurait probablement pas vu le jour. Toujours en Belgique, je souhaite remercier Julien Truffin, professeur à l'ULB, qui a contribué à rehausser la qualité de cette thèse par son implication dans mon troisième article scientifique. Enfin, j'exprime ma gratitude envers Michel Denuit, professeur à l'UCLouvain, pour ses précieux commentaires sur mon deuxième article. J'adresse également un clin d'œil à toutes les personnes et organisations en Belgique ayant contribué, directement ou indirectement, à mon parcours académique.

À travers Hélène Guérin, professeure à l'UQAM, je remercie tout le département de mathématiques de l'UQAM pour les connaissances apprises. J'en profite aussi pour remercier toutes les personnes sympathiques que je croisais dans les couloirs menant à mon bureau au pavillon Président-Kennedy, ainsi que celles qui me servaient à manger et à boire au restaurant du même pavillon.

Je remercie également toutes les personnes et organisations que j'ai côtoyées au Canada, ainsi que dans les pays où j'ai vécu, notamment le Mali, le Sénégal et le Cameroun. J'en profite enfin pour remercier mon professeur de mathématiques du Lycée Bâ Aminata Diallo de Bamako, en classe de terminale sciences exactes (promotion 2003), que nous surnommions Z-barre (Danséry). Paix à son âme ; il m'aimait beaucoup.

Enfin, je remercie toute ma famille, en particulier ma grand-mère Fatoumata Cissé, dite Tata (vivante au Mali), qui m'a élevée, et ma mère Hawa Diabaté (vivante en France), qui a été une confidente précieuse tout au long de cette thèse. Je rends également un grand hommage à mon grand-père Moussa Balla Coulibaly, chef des chasseurs du Mali (1992–2002), paix à son âme, pour tout l'amour que lui et sa famille m'ont donné.

Je dédie cette thèse à mes deux tantes décédées, Awa Coulibaly et Habibatou Coulibaly, paix à leurs âmes, pour tout l'amour qu'elles m'ont donné.

Si tu es issu(e) d'une minorité, sache que cette thèse t'est aussi dédiée. Ta différence n'est pas un défaut, et il existe de bonnes personnes prêtes à t'aider à réaliser tes projets.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

GAM Generalized Additive Model.

GLM Generalized Linear Model.

BMS BONUS-MALUS SCALE.

ABC Area Between the Curves.

CPG Compound Poisson-gamma.

DGLM Double Generalized Linear Model.

AIC Akaike Information Criterion.

BIC Bayesian Information Criterion.

EDF Exponential Dispersion Family.

IRLS Iteratively Reweighted Least Squares.

QMLE Quasi-Maximum Likelihood Estimator.

GWM Gamma-Weight Model.

CWM Constant-Weight Model.

EWM Exposure Weighted Model.

NOTATION

Distribution

Tweedie's CP paramétrisation composée Poisson–Gamma de la distribution de Tweedie.

Mathématique

KL Kullback–Leibler divergence.

RÉSUMÉ

Le contrat d'assurance automobile est généralement conclu pour une durée d'un an, de sorte que le calcul de la prime s'effectue également sur une base annuelle, en fonction de la sinistralité observée. Dans cette thèse, la charge totale de réclamations d'un contrat est considérée comme une mesure de sinistralité. Nous supposons que cette variable suit une loi de Tweedie et dépend, dans un premier temps, de l'expérience d'assurance associée à la police — définie par l'intermédiaire de l'ensemble des contrats détenus par un même assuré. Cette expérience est quantifiée à l'aide de l'historique de réclamations. Nos analyses empiriques mettent en évidence une dépendance positive significative de cet historique sur la charge totale.

La thèse s'intéresse également à la prise en compte de la durée de couverture dans le calcul de la prime. Même si la durée de couverture est généralement d'une année, certains contrats présentent une durée d'exposition au risque inférieure à une année, notamment lorsqu'ils sont résiliés en cours de terme. Cette durée d'exposition est mesurée par l'exposition au risque, qui correspond au nombre de jours d'exposition d'un contrat rapporté à une année. Dans un premier temps, nous supposons que cette exposition au risque est proportionnelle à la prime et, sous l'hypothèse d'une loi de Tweedie, nous comparons deux approches de modélisation : l'approche *offset* et l'approche *ratio*. Bien que l'approche *offset* conduise à des estimateurs présentant de bonnes propriétés asymptotiques, l'approche ratio apparaît particulièrement pertinente du point de vue opérationnel, notamment en raison de sa proximité avec l'équilibre financier, selon lequel la somme des charges totales observées doit, en moyenne, correspondre aux primes collectées au niveau du portefeuille.

Dans un second temps, nous introduisons une famille de modèles fondés sur la loi de Tweedie, qui suppose une forme plus générale pour décrire le lien entre la prime et l'exposition au risque. Cette approche permet de mieux représenter le risque associé aux contrats que les deux approches traditionnelles *offset* et *ratio*. Elle permet également d'introduire une structure de pénalité adaptée aux résiliations en cours de contrat.

Les résultats théoriques et empiriques présentés dans les trois volets de cette thèse sont illustrés à l'aide d'un même jeu de données réelles provenant d'une grande compagnie d'assurance canadienne.

INTRODUCTION

Dans cette thèse, nous nous intéressons principalement à la modélisation de la charge totale de réclamations sous une loi de Tweedie, en considérant simultanément l'expérience d'assurance de la police et la durée d'exposition des contrats qui la composent. Nous démontrons que l'historique de réclamations d'une police, généralement utilisé pour quantifier son expérience d'assurance, a un impact sur la charge totale des contrats qui la composent. Nous montrons également que la durée d'exposition d'un contrat constitue une variable pertinente pour l'évaluation de son risque.

Notre intérêt pour la modélisation de la charge totale se justifie par le fait que la plupart des études en tarification d'assurance automobile tenant compte de l'expérience d'assurance d'une police ou de la durée d'exposition des contrats se sont principalement concentrées sur la modélisation du nombre total de réclamations, plutôt que sur d'autres mesures de risque telles que la sévérité d'une réclamation ou la charge totale. Ces deux dernières mesures, représentant des coûts, sont généralement caractérisées par une forte asymétrie à droite et une grande variabilité, ce qui explique en partie le nombre relativement limité d'études actuarielles qui leur sont consacrées.

Cette variabilité est encore plus marquée pour la charge totale. En effet, la fréquence des réclamations en assurance automobile étant généralement faible, une proportion importante des contrats ne génère aucune réclamation, ce qui conduit à un grand nombre de charges totales nulles dans les bases de données de tarification. La charge totale est ainsi une variable semi-continue, caractérisée par une masse de probabilité en zéro et une composante continue sur les valeurs positives, ce qui limite le nombre de distributions statistiques susceptibles de la modéliser adéquatement.

Même si la loi de Tweedie est largement utilisée en tarification pour modéliser la charge totale de réclamations d'un contrat, l'apport principal de cette thèse réside dans la prise en compte simultanée de l'expérience d'assurance et de la durée de couverture dans cette modélisation. Afin de mettre en évidence cette contribution, nous avons structuré cette thèse en trois chapitres, chacun proposant des approches de modélisation de la charge totale sous une loi de Tweedie.

Avant de présenter un résumé de chacun des chapitres, nous rappelons le principe du calcul de la prime d'assurance. La modélisation de la charge totale de réclamations a pour objectif le calcul de cette prime, qui

constitue l'objectif central de la tarification. Celle-ci est généralement calculée sur une période de couverture d'un an. Pour ce faire, les assureurs définissent une variable aléatoire $Y \geq 0$ représentant la sinistralité observée d'un contrat sur cette période de couverture. Ils considèrent ensuite un vecteur de caractéristiques, noté $\mathbf{x} \in \mathbf{R}^q$, contenant les variables explicatives du contrat, observables ou non. En notant $\mu(\mathbf{x}) = \mathrm{E}[Y \mid \mathbf{x}]$, la charge totale moyenne conditionnellement aux caractéristiques du contrat, la prime est définie par

$$\mu = \mu(\mathbf{x}) = \mathrm{E}[Y \mid \mathbf{x}]\,.$$

L'idée sous-jacente est de construire un portefeuille hétérogène de contrats dans lequel chaque prime reflète le niveau de risque du contrat auquel elle est associée. Autrement dit, deux contrats présentant des caractéristiques différentes ne devraient généralement pas se voir attribuer la même prime.

Lorsque la charge totale conditionnelle $Y \mid \mathbf{x}$ est supposée suivre une loi de Tweedie, sa variance est reliée à sa moyenne par

$$\mathrm{Var}[Y \mid \mathbf{x}] = \frac{\phi}{w}\mu^p,$$

où

- μ est le paramètre de moyenne ;
- $\phi > 0$ est le paramètre de dispersion ;
- $w > 0$ est le paramètre de poids ;
- $p \in \mathbf{R}$ est le paramètre de variance.

Cette relation traduit l'idée intuitive qu'un contrat auquel est associée une prime plus élevée présente également une variabilité plus importante de sa charge totale de réclamations. C'est notamment cette propriété qui explique la popularité de la loi de Tweedie en tarification d'assurance automobile.

En outre, lorsque le paramètre de variance de la loi de Tweedie vérifie $1 < p < 2$, il est démontré que cette loi est équivalente en distribution à une loi de Poisson composée Gamma (Compound Poisson-gamma (CPG)).

Cette représentation constitue l'une des principales justifications de l'utilisation de la loi de Tweedie pour modéliser la charge totale de réclamations, puisqu'elle reflète naturellement son caractère semi-continu.

Il faut également noter que cette équivalence en distribution justifie deux approches de modélisation largement utilisées dans la littérature :

- Approche Tweedie : la charge totale de réclamations est modélisée directement par une loi de Tweedie, seule ou conjointement avec une autre variable d'intérêt, généralement le nombre total de réclamations.
- Approche CPG : le nombre de réclamations est modélisé par une loi de Poisson tandis que le coût individuel des réclamations est modélisé par une loi Gamma. La charge totale est ensuite obtenue par composition de ces deux modèles.

Ces deux approches sont largement utilisées en assurance automobile et ont récemment fait l'objet d'une étude comparative. Toutefois, cette étude ne tient pas compte de l'expérience d'assurance de la police. Nous nous sommes donc demandé si les conclusions obtenues demeuraient valables lorsque l'historique de réclamations est intégré à la modélisation de la charge totale. Cette question constitue le point de départ du premier chapitre de cette thèse, dont le résumé est présenté ci-dessous.

Le chapitre 1 compare les approches Tweedie et CPG en intégrant l'historique de réclamations de la police dans la modélisation de la charge totale. La méthodologie proposée est appliquée à un échantillon de données provenant d'une grande compagnie d'assurance canadienne. Nous évaluons les modèles à la fois en termes de qualité d'ajustement et de capacité prédictive. Les résultats montrent que les deux approches constituent des alternatives pertinentes pour la modélisation de la charge totale, mais que la prise en compte de l'expérience d'assurance permet d'intégrer certaines considérations opérationnelles essentielles à la tarification.

Même si la modélisation statistique d'un risque en tarification a pour objectif principal le calcul de la prime μ, il convient de noter que, lorsque ce risque est représenté par la charge totale suivant une loi de Tweedie, tous les paramètres de cette distribution font généralement l'objet d'une inférence statistique, à l'exception du paramètre de poids w, qui est considéré comme une information connue a priori. En pratique, les assureurs relient généralement ce paramètre de poids à l'exposition au risque du contrat, notée t, avec $0 < t \leq 1$.

Cette grandeur correspond à la proportion de l'année pendant laquelle le contrat est exposé au risque. Dans le chapitre 1, nous adoptons cette même stratégie en assimilant le paramètre de poids à l'exposition au risque. Toutefois, ce choix n'y est pas justifié d'un point de vue théorique. Cette question constitue précisément le point de départ du deuxième chapitre de cette thèse, dont le résumé est présenté ci-dessous.

Le chapitre 2 compare les deux approches traditionnelles utilisées pour prendre en compte l'exposition au risque lorsque la charge totale suit une loi de Tweedie : l'approche ratio et l'approche *offset*. Ces deux approches supposent que la prime μ est proportionnelle à l'exposition au risque t, mais elles diffèrent par le choix du paramètre de poids w. Cette différence fait que ces deux approches constituent deux stratégies d'estimation de la prime μ. Dans ce chapitre, nous montrons que cette différence ne se manifeste que lorsqu'il existe dans le portefeuille des contrats dont l'exposition au risque vérifie $t < 1$. Nous développons une analyse théorique et des études de simulation montrant que l'approche *offset* est asymptotiquement plus efficace que l'approche ratio. En revanche, du point de vue de l'équilibre financier du portefeuille, critère recherché en tarification, l'approche ratio présente de meilleures performances. Ces résultats sont finalement illustrés à l'aide d'une analyse empirique réalisée sur un portefeuille constitué d'un échantillon des données utilisées dans le premier chapitre et comportant un nombre important de contrats tels que $t < 1$.

Même si la comparaison présentée dans le deuxième chapitre a permis d'apporter une justification théorique au choix du paramètre de poids w adopté dans le premier chapitre, une question demeure : l'hypothèse de proportionnalité entre la prime μ et l'exposition au risque t est-elle réaliste ?

Cette interrogation est motivée par une analyse empirique montrant que cette hypothèse ne semble pas être vérifiée pour tous les contrats dont l'exposition au risque satisfait $t < 1$. En effet, notre base de données présente la particularité d'observer, sur une même période annuelle de couverture, plusieurs contrats de véhicules rattachés à une même police d'assurance, assimilée dans cette thèse à un assuré. Ces contrats ne possèdent pas nécessairement la même durée d'exposition au risque. Nos analyses montrent notamment que certains contrats vérifiant $t < 1$ correspondent à des ajouts de véhicules en cours de période de couverture, tandis que d'autres résultent d'une annulation (ou résiliation) de contrat avant son échéance. C'est principalement pour cette seconde catégorie de contrats que l'hypothèse de proportionnalité entre la prime et l'exposition au risque ne semble plus être vérifiée. Cette observation constitue le point de départ du troisième chapitre de cette thèse, dont le résumé est présenté ci-après.

Le chapitre 3 propose une famille de modèles sous une loi de Tweedie dans laquelle la prime μ n'est plus supposée proportionnelle à l'exposition au risque t. Nous introduisons une forme plus générale pour décrire le lien entre la prime et l'exposition au risque, permettant de mieux représenter le risque associé aux contrats résiliés en cours de période de couverture. À partir du même jeu de données que celui utilisé dans les deux chapitres précédents, nous mettons en évidence les avantages de la famille de modèles proposée par rapport aux approches traditionnelles *offset* et ratio. Les analyses montrent que cette famille de modèles décrit plus adéquatement la relation entre la prime et l'exposition au risque lorsque des contrats sont résiliés en cours de période de couverture. Nous montrons également que cette approche peut procurer un avantage stratégique à l'assureur en permettant de compenser indirectement certaines pertes importantes par une pénalité associée aux résiliations en cours de contrat, tout en préservant la cohérence actuarielle du modèle et les propriétés de convergence des estimateurs.

Au-delà des contributions propres à chacun des chapitres, cette thèse propose un cadre unifié de modélisation de la charge totale de réclamations sous une loi de Tweedie, intégrant simultanément l'expérience d'assurance et la durée de couverture des contrats. Compte tenu de l'originalité de ces contributions et de leur complémentarité, chacun des chapitres a été rédigé indépendamment sous la forme d'un article scientifique et soumis à une revue spécialisée.

- **Chapitre 1 :** Boucher, J.-P. et Coulibaly, R. (2024). *Bonus-malus scale premiums for Tweedie's compound Poisson models*. *Annals of Actuarial Science*, 1–25. `http://dx.doi.org/10.1017/S1748499524000113`.
- **Chapitre 2 :** Boucher, J.-P. et Coulibaly, R. (2026). *Comparison of offset and ratio weighted regressions in Tweedie models with application to mid-term cancellations*. *European Actuarial Journal*. `http://dx.doi.org/10.1007/s13385-026-00452-z`.
- **Chapitre 3 :** Boucher, J.-P., Coulibaly, R. et Trufin, J. (2026). *Varying Risk Exposure in Auto Insurance : A Weighted Tweedie Framework for Experience Rating and Cancellation Penalties*, soumis pour publication.

Les trois chapitres sont reproduits en anglais dans ce manuscrit, conformément aux versions publiées ou soumises aux revues scientifiques.

CHAPITRE 1

BONUS-MALUS SCALE PREMIUMS FOR TWEEDIE'S COMPOUND POISSON MODELS

1.1 Introduction

Experience Rating and predictive ratemaking refer to ratemaking models that use claims information from past insurance contracts to predict the future total amount of claims (also known as "loss costs"). From a ratemaking point of view, the idea of experience rating is to compute a premium for insured i, for a contract of period T, that will consider all the insured's past insurance contracts $1, \ldots, T-1$.

Historically, research on this type of predictive ratemaking has focused on modeling the annual claims number of a contract. For an insured i, one can use panel data theory to model the joint distribution of the annual claims number for each T contract, denoted by $N_{i,t}$ for contract t, $t = 1, \ldots, T$. This is expressed as the product of predictive distributions :

$$\Pr(N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}) = \prod_{t=1}^{T} \Pr(N_{i,t} | \boldsymbol{n}_{i,(1:t-1)}),$$

where $\boldsymbol{n}_{i,(1:t-1)} = \{n_{i,1}, n_{i,2}, \ldots, n_{i,t-1}\}$ is the vector of annual past claims numbers observed at the beginning of each contract t.

Considering each conditional random variable $N_{i,t}|\boldsymbol{n}_{i,(1:t-1)}$, we can calculate the frequency component premium of the contract t ($t = 1, \ldots, T$), denoted $\pi_{i,t}^{(F)}$, using the following equation : $\pi_{i,t}^{(F)} = E[N_{i,t}|\boldsymbol{n}_{i,(1:t-1)}]$. Therefore, the premium of contract t of policy i can be interpreted as a function of $\boldsymbol{n}_{(1:t-1)}$, which materializes the dependency between the T contracts of insured i.

Usually, the classic actuarial approach to introducing dependency between the T contracts of an insured i is to introduce a random term common to all of the insured's T contracts. See Turcotte et Boucher (2023) or Pechon *et al.* (2019) for a review of this approach. Several other approaches in the literature have also been tried, such that of Bermúdez *et al.* (2018) for models based on time series or Shi et Valdez (2014) for models using a copula by the 'jittering' method.

More recently, instead of using the random effects approach, several papers have highlighted the advantages of using two families of models that take advantage of the fact that the average frequency of claims in property and casualty insurance is often between 0 and 1. Called Kappa-N model and BONUS-MALUS SCALE (BMS) model, these families propose to use a claims history function directly in the mean parameter of a counting distribution to model the decrease in premiums for insureds who do not file claims and the increase in premiums for insureds who do. See Simon *et al.* (2023),Villacorta Iglesias *et al.* (2021) or Adillon *et al.* (2020) for a review of the BMS models. Research on modelling the total annual claims number across several different databases and insurance products has shown that the BMS models can generate an excellent fit of training data and excellent prediction on test data, and often outperform the results of Bayesian random-effects models (Boucher, 2022; Boucher et Inoussa, 2014).

In this paper, we generalize this BMS approach by working with data with a slightly more complex structure, close to what is used in practice, at least in North America. In Canada, some families contain multiple insured vehicles. From an experience-rating standpoint, the premium of one specific vehicle insured by a family could be calculated using the number of past claims for all others in this family, not just the number of past claims for that particular vehicle. In fact, in a family, the different drivers all use one or the other of the vehicles, so it makes sense to use the experience of all vehicles for rating. If we take as an example a family with two insured vehicles, let $N_{i,j,t}$ denote the annual claims number of vehicle j ($j = 1, 2$) during period t. We then obtain

$$\begin{aligned} E[N_{i,1,t} + N_{i,2,t}|\boldsymbol{n}_{i,1,(1:t-1)}, \boldsymbol{n}_{i,2,(1:t-1)}] &= E[N_{i,1,t}|\boldsymbol{n}_{i,1,(1:t-1)}, \boldsymbol{n}_{i,2,(1:t-1)}] + E[N_{i,2,t}|\boldsymbol{n}_{i,1,(1:t-1)}, \boldsymbol{n}_{i,2,(1:t-1)}] \\ &\neq E[N_{i,1,t}|\boldsymbol{n}_{i,1,(1:t-1)}] + E[N_{i,2,t}|\boldsymbol{n}_{i,2(1:t-1)}], \end{aligned}$$

where $\boldsymbol{n}_{i,j,(1:t-1)} = \{n_{i,j,1}, n_{i,j,2}, \ldots, n_{i,j,t-1}\}$ is the vector of annual past claims numbers observed at the beginning of the contract t of vehicle j for this family i.

Instead of counting the number of claims per vehicle in a family, we can count the number of claims per coverage/warranty (see Boucher et Inoussa, 2014, for an illustration). We could also analyze the number of claims per insurance product, counting the number of home insurance claims to model the number of auto insurance claims. One can also consult Verschuren (2021) for this type of application of BMS model.

In this paper, we also propose to generalize the type of random variables modeled. Instead of using only the annual claims number for each contract t of an insured i, we propose to develop a structure to model the cost of each claim k, $Z_{i,j,k,t}$ and the annual claims amount $Y_{i,j,t}$. First, we deal with the joint distribution of the annual claims number and the costs of each of these claims for the T contracts of insured i. Second, we deal with the joint distribution of the annual claims number and the annual claims amount for these same T contracts of insured i. The joint modelling of the annual claims number and the costs of each of these claims is called frequency-severity modelling. See Jeong et Valdez (2020); Oh *et al.* (2020); Lee et Shi (2019), and Shi et Yang (2018) for a literature review about this model. Even if, in the loss cost model, the target variable is the annual claims amount of a contract, researchers recommend that this target variable and the annual claims number be modeled jointly (Delong *et al.*, 2021; Frees, 2014).

1.1.1 Terminologies and Definitions

Similar to what has been done for the annual claims number modelling, one can also model the conditional distribution of the annual claims amount $Y_{i,j,t}$ according to the $t-1$ past annual claims amounts. However, in keeping with the idea of the Kappa-N and BMS models that are built conditionally on the number of past claims, our approach will be based on the analysis of the distribution of the annual claims amount, conditional on the number of past claims. In such a case, for an insured i, the premium of contract t of vehicle j is calculated as :

$$\pi_{i,j,t}^{(Y)} = E[Y_{i,j,t}|\boldsymbol{n}_{i,1,(1:t-1)}, \ldots, \boldsymbol{n}_{i,J,(1:t-1)}],$$

where J is the total number of insured vehicles in the past. As shown in Section 1.2 and Section 1.3, the severity could also be modeled using this approach.

More generally, to cover all of these possibilities, Boucher (2022) defined two types of variables to be used in experience rating :

1. The variable to model, named the **target variable**;
2. The information used to define what we consider the past claim experience, named the **scope variable**.

Using these definitions means that for this paper, three target variables will be modeled : the annual claims number (the claims frequency), the claims costs (the claims severity) and the annual claims amount (also called the loss cost). In contrast, all three target variables will be modeled based only on one type of scope variable : the number of past claims.

1.1.2 Summary

In section 1.2 of the paper, we present some contextual elements and hypotheses to better introduce the models we propose. The models' notation is revised so that predictive pricing approaches can be used for the two new target variables, severity and loss cost. In the recent paper by Delong *et al.* (2021), the Compound Poisson-gamma (CPG) and Tweedie distributions were studied for their practical advantages in loss cost modelling. Our paper has the same objective as that of Delong *et al.* (2021), but we also consider the past insurance experience of each insured vehicle in calculating the premium of its future contracts. Details on how to use these two distributions in our proposed models is given in section 1.3. In section 1.4, we apply our proposed models to an auto insurance database of a major Canadian insurer. A variable selection step based on the *Elastic-net* regularization is also introduced to measure the impact of adding a pricing component per experience. Section 1.5 concludes the paper.

1.2 Data Structure and Hypotheses

1.2.1 Definitions and Form of Available Data

We assume a hierarchical data structure in which the claims experience of the policies, vehicles and insurance contracts associated with each vehicle is observed. To ensure a common vocabulary, we have retained some of the terms used in the introduction but have clarified their definitions further :

- A **policy** is usually associated with a single insured. In insurers' databases, a policy is usually identified by a unique number.
- An insured vehicle, or simply a **vehicle**, is associated with a policy. For an in-force policy, the minimum number of insured vehicles is one, but many insured (particularly in North America) might have several vehicles.
- A policy is often made up of several insurance **contracts**. An insurance contract is also often referred to as an insurance term, and it is usually one year long. Insurance contracts are sometimes shorter, for

example three or six months. Some insurance policies contain only one term, but a significant portion of policies contain multiple contracts.

For a given policy $i, i = 1, \ldots, m$, we assume that the claims of a $j, j = 1 \ldots, J_i$ vehicle are observable through $T_{i,j}$ contracts. We denote by t, the index associated with the T contracts (i and j are removed to simplify reading). Our variables of interest are therefore the claims experience of the T contracts of each vehicle in a policy.

To better capture the form of the available data, Table 1.1 provides an illustration of a sample of three insurance policies. It shows that policy #1 contained only one insured vehicle for the 2018 and 2021 policies, but two vehicles were insured in 2019 and 2020. Thus, the claims experience of the first vehicle in policy #1 was observed during four annual contracts, while the claims experience of the second vehicle was observed during only two annual contracts. Policy #2 in this sample contains only one insured vehicle, and that vehicle was insured for only one annual contract. Finally, two vehicles insured on a single annual contract were observed in the third and final policy in this table.

It should be noted that the contract number is obtained according to its associated policy and its effective date. That is, in a given policy, all contracts with the same effective year have the same contract number. The first contract for vehicle #2 of policy #1 illustrates this situation. Such a notation is important for the rest of the paper.

As can be seen in the sample in the same table, the characteristics of the insured and the insured vehicle are also available. Finally, the frequency of claims and the cost of each claim are also available information. The loss cost, representing the sum of the costs of each claim, is shown in the last column of the table. An insured who has not made any claims during their contract execution period has a loss cost of zero.

Policy Number	Vehicle Number	Contract Number	Contract Date	Risk Characteristics			Number of Claims	Costs of Claim			Loss Costs
				Age	Sex	...		#1	#2	#3	
1	1	1	2018-01-15	42	M	...	0	.	.	.	0
1	1	2	2019-01-15	43	M	...	2	6,592	11,520	.	18,112
1	1	3	2020-01-15	44	M	...	1	24,151	.	.	24,151
1	1	4	2021-01-15	45	M	...	0	.	.	.	0
1	2	2	2019-01-15	40	F	...	2	1,490	24,505	.	25,995
1	2	3	2020-01-15	41	F	...	0	.	.	.	0
2	1	1	2018-02-05	24	M	...	0	.	.	.	0
3	1	1	2018-02-08	34	F	...	0	.	.	.	0
3	2	1	2018-02-08	30	M	...	1	8,150	.	.	8,150

Table 1.1 – Illustration of frequency and severity data

1.2.2 Target Variables

For vehicle j of an insured i, the random variable $N_{i,j,t}$ represents the annual claims number of its contract t. If the observed annual claims number $n_{i,j,t} = n > 0$, the random vector $Z_{i,j,t} = (Z_{i,j,t,1}, ..., Z_{i,j,t,n})^{'}$ will represent the vector of each the insured's n claims costs. This is not defined if the associated observed annual claims number $n_{i,j,t} = n = 0$.

Thus, we calculate the loss cost, denoted by the random variable $Y_{i,j,t}$ as follows :

$$Y_{i,j,t} = \begin{cases} \sum_{k=1}^{N_{i,j,t}} Z_{i,j,t,k} & \text{if } N_{i,j,t} > 0 \\ 0 & \text{if } N_{i,j,t} = 0 \end{cases} = \left(\sum_{k=1}^{N_{i,j,t}} Z_{i,j,t,k} \right) I\left(N_{i,j,t} > 0\right).$$

We define our three target variables by the following three random variables : $N_{i,j,t}$, $Z_{i,j,t}$ and $Y_{i,j,t}$.

1.2.2.1 Premiums

For the m insured in the portfolio, assuming a minimization of the square distance for the calculation of the premium of contract t for each vehicle j in the policy i, the parameter of interest corresponds to $\mathrm{E}[Y_{i,j,t}]$, $\forall i = 1, ..., m$, $\forall j = 1, \ldots, J_i$, $\forall t = 1, \ldots, T_{i,j}$. This parameter can be calculated in two ways : (1) By the multiplying the frequency component premium and the severity component premium according

to some assumptions ; (2) By considering the conditional distribution of the loss cost denoted by $f_{Y_{i,j,t}}(.)$. Formally, these two ways are expressed as :

$$\mathrm{E}[Y_{i,j,t}] = \begin{cases} \mathrm{E}[N_{i,j,t}]\,\mathrm{E}\big[Z_{i,j,t,k}\big] & (1) \\ \\ \int y f_{Y_{i,j,t}}(y)\,dy & (2). \end{cases}$$

The assumptions to obtain a premium according to (1) are generally defined as follows :

1. The independence between the claims frequency and the claims severity for the same contract t :

$$N_{i,j,t} \perp\!\!\!\perp Z_{i,j,t,k},\ \forall i = 1,\dots,m,\ \forall j = 1,\dots,J_i,\ \forall t = 1,\dots,T_{i,j},\ \forall k = 1,\dots,n_{i,j,t}.$$

2. The independence of the costs of each claim across distinct policies :

$$Z_{i,j,t,k_1} \perp\!\!\!\perp Z_{i,j,t,j,k_2},\ \forall i = 1,\dots,m,\ \forall j = 1,\dots,J_i,\ \forall t = 1,...,T_{i,j}, k_1 \neq k_2.$$

3. For the same contract t, the costs of each claim are identically distributed.

In order to include some form of segmentation in the rating (Frees, 2014), it should be noted that the premium is generally calculated considering specific observable characteristics of each contract, such as those illustrated in Table 1.1. We denote these characteristics by the following vector $\boldsymbol{X}_{i,j,t} = (x_{i,j,t,0}, ..., x_{i,j,t,q})^{'} \in \mathbb{X} \subset \{1\} \times \mathbb{R}^q, q > 0$. We are finally interested in these quantities : $\pi^{(Y)}_{i,j,t} = \mathrm{E}[Y_{i,j,t}|\boldsymbol{X}_{i,j,t}]$, $\pi^{(N)}_{i,j,t} = \mathrm{E}[N_{i,j,t}|\boldsymbol{X}_{i,j,t}]$, $\pi^{(Z)}_{i,j,t} = \mathrm{E}[Z_{i,j,t}|\boldsymbol{X}_{i,j,t}]$.

1.3 Experience Rating with CPG and Tweedie Models

For the experience rating, the Kappa-N and BMS models are generally proposed (Boucher, 2022), which model the conditional distribution of a target variable according to the scope variables. In this section, the CPG and Tweedie are used as an underlying distribution in each model. Before presenting Kappa-N and BMS models in our context, we start with an example to better explain how the scope variables are calculated in practice.

1.3.1 Scope Variables

It is known in actuarial science that insureds who make claims will have a higher frequency of claims in their future contracts. This can be explained in several ways : some insureds behave more riskily than others, some insureds live in areas that are more prone to disasters, and some insured property is more likely to be damaged. Individual characteristics used as segmentation variables may partly explain this situation. However, many of these variables cannot be measured and modeled directly in rating. Thus, past claims experience can be used to approximate the effect of these non-measurable characteristics on premiums. This is why, in addition to conditioning on characteristics $\boldsymbol{X}_{i,j,t}$, we price an insured according to their claims history, defined as a scope variable in the introduction.

To illustrate the situation adequately, we use Table 1.1 as an example, which we generalize to Table 1.2. For each vehicle in Table 1.1, we can calculate the number of claims from past contracts, i.e $\boldsymbol{n}_{i,j,(1:t-1)} = \{n_{i,j,1}, n_{i,j,2}, \dots, n_{i,j,t-1}\}$. This is shown in columns 5, 6 and 7 of Table 1.2. However, our scope variable of frequency will not only be composed of the number of past claims for the same vehicle, but also the number of past claims of the entire policy. Thus, in the last three columns of Table 1.2, the sum of the past claims for all vehicles of the same policy is shown for the previous contracts, namely :

$$\boldsymbol{n}_{i,\bullet,(1:t-1)} = \left\{ \sum_{j=1}^{J_i} n_{i,j,1}, \sum_{j=1}^{J_i} n_{i,j,2}, \dots, \sum_{j=1}^{J_i} n_{i,j,t-1} \right\} = \{n_{i,\bullet,1}, n_{i,\bullet,2}, \dots, n_{i,\bullet,t-1}\}.$$

Policy	Vehicle	Contract		Scope Variables					
i	j	t	$n_{i,j,t}$	$n_{i,j,t-1}$	$n_{i,j,t-2}$	$n_{i,j,t-3}$	$n_{i,\bullet,t-1}$	$n_{i,\bullet,t-2}$	$n_{i,\bullet,t-3}$
1	1	1	0	.	.	.	.	.	.
1	1	2	2	0	.	0	0	.	.
1	1	3	1	2	0	2	4	0	.
1	1	4	0	1	2	0	1	4	0
1	2	2	2	.	.	0	0	.	.
1	2	3	0	2	.	.	4	0	.
2	1	1	0	.	.	.	.	.	.
3	1	1	0	.	.	.	.	.	.
3	2	1	1	0	.	.	1	0	.

Table 1.2 – Illustration of scope variables

1.3.2 Kappa-N Models

For a loss cost model, we are looking for the joint conditional distribution of $(N_{i,j,t}, Y_{i,j,t})$ according to $\boldsymbol{n}_{i,\bullet,(1:t-1)}$ and $\boldsymbol{X}_{i,j,t}$. For a frequency-severity model, we are looking for the joint conditional distribution of $(N_{i,j,t}, Z_{i,j,t})$ according to $\boldsymbol{n}_{i,\bullet,(1:t-1)}$ and $\boldsymbol{X}_{j,t}$.

Using a logarithmic link between the co-variables and the mean parameter such as in Generalized Linear Models (GLMs) (Delong *et al.*, 2021; Jong et Heller, 2008), the expectations for our three variables of interest are expressed as given by Equations (1.3.1), (1.3.2) and (1.3.3) where $h^{(Y)}(.)$, $h^{(N)}(.)$ et $h^{(Z)}(.)$ represent the functions of historical claims.

$$\pi_{i,j,t}^{(N)} = \exp\left(\boldsymbol{X}_{i,j,t}'\beta^{(N)} + h^{(N)}(n_{i,\bullet,1},\ldots,n_{i,\bullet,t-1})\right),\ \beta^{(N)} = (\beta_0^{(N)},\ldots,\beta_q^{(N)})' \in \mathbb{R}^{q+1}; \tag{1.3.1}$$

$$\pi_{i,j,t}^{(Z)} = \exp\left(\boldsymbol{X}_{i,j,t}'\beta^{(Z)} + h^{(Z)}(n_{i,\bullet,1},\ldots,n_{i,\bullet,t-1})\right),\ \beta^{(Z)} = (\beta_0^{(Z)},\ldots,\beta_q^{(Z)})' \in \mathbb{R}^{q+1}; \tag{1.3.2}$$

$$\pi_{i,j,t}^{(Y)} = \exp\left(\boldsymbol{X}_{i,j,t}'\beta^{(Y)} + h^{(Y)}(n_{i,\bullet,1},\ldots,n_{i,\bullet,t-1})\right),\ \beta^{(Y)} = (\beta_0^{(Y)},...,\beta_q^{(Y)})' \in \mathbb{R}^{q+1}. \tag{1.3.3}$$

It should be noted that several possibilities exist to define these historical claims functions. Boucher (2022) listed some of them, and the problems that they could create. Taking advantage of the fact that the average automobile insurance claim frequency is between 0 and 1, and that insureds expect a discount when they do not claim and a surcharge if they report a claim, Boucher proposed instead to define a new indicator variable $\kappa_{i,j,t} = I(n_{i,j,t} = 0)$ that identifies claims-free contracts. We thus have two new variables summarizing the claims experience :

$$n_{i,\bullet,\bullet} = \sum_{\tau=1}^{t-1} n_{i,\bullet,\tau} \text{ , and } \kappa_{i,\bullet,\bullet} = \sum_{\tau=1}^{t-1} \kappa_{i,\bullet,\tau} = \sum_{\tau=1}^{t-1} I(n_{i,\bullet,\tau} = 0), \text{ and so}$$
$$h^{(.)}(n_{i,\bullet,1},\ldots,n_{i,\bullet,t-1}) = -\gamma_0^{(.)}\kappa_{i,\bullet,\bullet} + \gamma_1^{(.)} n_{i,\bullet,\bullet},$$

where $\gamma_0^{(.)}, \gamma_1^{(.)} \in \mathbb{R}$. The negative sign in front of the positive parameter $\gamma_0^{(.)}$ is used to highlight that an additional year without a claim will decrease the premium. This simple way of summarizing the claims history in the mean parameter of a random variable is called Kappa-N modelling. The idea is to consider $\kappa_{i,\bullet,\bullet}$ and $n_{i,\bullet,\bullet}$ as co-variables in premium modelling.

1.3.2.1 Claims Score

For each policy i and each vehicle's contract t, we define a positive quantity $\ell_{i,t}^{(.)}$ called the claims score based on the function $h^{(.)}(.)$ and an initial score ℓ_1. This initial score is interpreted as the maximum number of years for which a contract can remain without a claim from its effective date to its end date. Boucher (2022) sets $\ell_1 = 100$ for a simple aesthetic reason :

$$\begin{aligned}
\ell_{i,t}^{(.)} &= \ell_1 + \frac{1}{\gamma_0^{(.)}} h^{(.)}(n_{i,\bullet,1}, \ldots, n_{i,\bullet,t-1}) \\
&= \ell_1 + \frac{1}{\gamma_0^{(.)}} \left(-\gamma_0^{(.)} \kappa_{i,\bullet,\bullet} + \gamma_1^{(.)} n_{i,\bullet,\bullet} \right) \\
&= \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(.)} n_{i,\bullet,\bullet}, \text{ with } \Psi^{(.)} = \frac{\gamma_1^{(.)}}{\gamma_0^{(.)}},
\end{aligned}$$

where $\Psi^{(.)}$ is the jump parameter and γ_0 is the relativity parameter of penalties related to past claims.

For a Kappa-N model using the claims score, the expectations of our three variables of interest can be expressed as :

$$\pi_{i,j,t}^{(N)} = \exp\left(\boldsymbol{X}_{i,j,t}' \beta^{(N)} + \gamma_0^{(N)} \ell_{i,t}^{(N)} \right); \tag{1.3.4}$$

$$\pi_{i,j,t}^{(Z)} = \exp\left(\boldsymbol{X}_{i,j,t}' \beta^{(Z)} + \gamma_0^{(Z)} \ell_{i,t}^{(Z)} \right); \tag{1.3.5}$$

$$\pi_{i,j,t}^{(Y)} = \exp\left(\boldsymbol{X}_{i,j,t}' \beta^{(Y)} + \gamma_0^{(Y)} \ell_{i,i,t}^{(Y)} \right). \tag{1.3.6}$$

From these equations, one can quickly assess the impact of the claims score on the insurance premium. This has several desirable qualities in terms of the contract's rating structure :

- For an insured i without insurance experience, $n_{i,\bullet,\bullet} = 0$ and $\kappa_{i,\bullet,\bullet} = 0$, which means a claim score of $\ell_t = 100 = \ell_1$;
- Each annual contract without a claim will decrease the claim score ℓ by 1;
- Each claim increases the claim score ℓ by Ψ ;

— The impact of a single claim on the premium is then roughly equal to Ψ years without claims ;

— The penalty for a claim is an increase of $(\exp(\Psi\gamma_0) - 1)$% of the premium ;

— Each year without a claim decreases the premium by $(1 - \exp(-\gamma_0))$%.

1.3.2.2 Kappa-N For Independent Poisson Annual Claims Numbers

Considering the risk exposure of contract t, denoted by $d_{i,j,t}$, we assume that its annual claims number $N_{i,j,t}|\ell_{i,t}^{(N)}$ is Poisson distributed, such that :

$$\pi_{i,j,t}^{(N)} = d_{i,j,t} \exp\left(\boldsymbol{X}_{i,j,t}'\beta^{(N)} + \gamma_0^{(N)}\ell_{i,t}^{(N)}\right), \tag{1.3.7}$$

$$\ell_{i,t}^{(N)} = \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(N)} n_{i,\bullet,\bullet}, \text{ with } \Psi^{(N)} = \frac{\gamma_1^{(N)}}{\gamma_0^{(N)}}. \tag{1.3.8}$$

For inference purposes , we assume the independence between the contracts' annual claims number of distinct policies :

$$N_{i_1,j,t} \perp\!\!\!\perp N_{i_2,j,t}, \ \forall i_1, i_2 = 1, ..., n | i_1 \neq i_2.$$

Further, given the use of the claims score $\ell^{(N)}$, a form of dependency will exist between contracts for the same vehicle, and contracts for vehicles of the same policy. More formally, we have :

$$N_{i,j_1,t_1} \not\perp\!\!\!\perp N_{i,j_2,t_2}, , \ \forall j_1, j_2, t_1, t_2.$$

The likelihood contribution for the claims frequency of a single policy i is expressed as follows (i is removed for easy reading) :

$$\prod_{j=1}^{J}\prod_{t=1}^{T} \exp\left(-\pi_{j,t}^{(N)} + n_{j,t}\log\left(\pi_{j,t}^{(N)}\right) - \log\left(n_{j,t}!\right)\right). \tag{1.3.9}$$

Finally, the idea is to estimate the parameters $\beta^{(N)}$, $\gamma_0^{(N)}$ and $\gamma_1^{(N)}$ by maximizing the likelihood function built by multiplying the contribution (1.3.9) for m policies. This optimization can also be done using the *glm* function in R.

1.3.2.3 Kappa-N for Independent Gamma Claims Costs

For a contract t, we assume that the common distribution of the costs of each of its $n_{i,j,t} = n$ claims, $Z_{i,j,t,k}$, is gamma such that :

$$\pi_{i,j,t}^{(Z)} = \exp\left(\boldsymbol{X}'_{i,j,t}\beta^{(Z)} + \gamma_0^{(Z)}\ell_{i,t}^{(Z)}\right), \tag{1.3.10}$$

$$\ell_{i,t}^{(Z)} = \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(Z)} n_{i,\bullet,\bullet}, \text{ with } \Psi^{(Z)} = \frac{\gamma_1^{(Z)}}{\gamma_0^{(Z)}}. \tag{1.3.11}$$

It is important to note that the cost of a claim, $Z_{i,j,t,k}, k = 1, \ldots, n$ depends on the score $\ell^{(Z)}$ which is set at the beginning of period t. Thus, the first, second, and third claims of contract t are all dependent on the same score $\ell^{(Z)}$. This score, as expressed in equation (1.3.11), will only be updated at the end of contract t for the rating of contract $t + 1$.

Similar to the frequency part, we assume that the claims severities of contracts of distinct policies are independent. Further, we take into account the dependency between the severity of contracts for the same vehicle and between those of vehicles of the same policy :

$$Z_{i_1,j,t,k} \perp\!\!\!\perp Z_{i_2,j,t,k}, \ \forall i_1, i_2 = 1, ..., n | i_1 \neq i_2;$$
$$Z_{i,j_1,t_1,k} \not\perp\!\!\!\perp Z_{i,j_2,t_2,k}, , \ \forall j_1, j_2, t_1, t_2.$$

We therefore evaluate the contribution of the likelihood for the severity of a policy i as follows (i is removed for easy reading) :

$$\prod_{j=1}^{J}\prod_{t=1}^{T}\prod_{k=1}^{n} \exp\left(-\frac{\gamma}{\pi_{j,t}^{(Z)}} z_{j,t,k} - \gamma \log\left(\pi_{j,t}^{(Z)}\right) - \log\left(\frac{z_{j,t,k}^{\gamma-1}\gamma^{\gamma}}{\Gamma(\gamma)}\right)\right), \tag{1.3.12}$$

where $z_{j,t,k}$, is the observed cost of claim k of contract t. Thus, the inference consists in maximizing the likelihood function built from the likelihood contributions (1.3.12) of the m policies. Similar to the Poisson model, the *glm* function in R can also be used to estimate the parameters $\beta^{(Z)}$, $\gamma_0^{(Z)}$ and $\gamma_1^{(Z)}$.

1.3.2.4 Remarks

In the CPG model, for each contract, it is assumed that the individual's annual claims number is Poisson distributed and the costs of each of the insudred's claims is gamma distributed. This model also assumes independence between the frequency and the severity of claims of each contract. Therefore, to calculate the annual premium of a contract, one can model its frequency and severity components separately (Delong *et al.*, 2021).

For the severity modelling, we consider the costs of each claim, unlike Delong *et al.* (2021), who considered the average claims amount of each contract as a target variable. Note that these two approaches lead to the same inference results (Delong *et al.*, 2021).

1.3.2.5 Kappa-N for Independent Tweedie Annual Claims Amount

Instead of using the CPG approach to model the loss cost of a contract, another alternative is to use the distribution of its annual claims amount directly and calculate the expectation of this distribution to obtain the annual premium. This is the idea of the Tweedie model.

Consistent with Delong *et al.* (2021), we consider for each contract t, the couple of random variables $(N_{i,j,t}, Y_{i,j,t})$ representing the annual claims number and the annual claims cost : $Y_{i,j,t}$ is Tweedie distributed and $N_{i,j,t}$ is Poisson distributed.

For inference purposes, we are interested in the conditional distribution of $(N_{i,j,t}, Y_{i,j,t})$ according to the $\ell_{i,t}^{(Y)}$ and $\boldsymbol{X}_{i,j,t}$. We also assume the following equations for the mean $\mu_{i,j,t}$ and the dispersion $\phi_{i,j,t}$ parameters of the Tweedie distribution :

$$\pi_{i,j,t}^{(Y)} = d_{i,j,t} \exp\left(\boldsymbol{X}_{i,j,t}'\beta^{(Y)} + \gamma_0^{(Y)}\ell_{i,t}^{(Y)}\right) = \mu_{i,j,t}, \tag{1.3.13}$$

$$\phi_{i,j,t} = \exp\left(\boldsymbol{X}_{\boldsymbol{i,j,t}}'\beta^{(D)} + \gamma_0^{(D)}\ell_{i,t}^{(Y)}\right), \tag{1.3.14}$$

$$\ell_{i,t}^{(Y)} = \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(Y)} n_{i,\bullet,\bullet}, \text{ with } \Psi^{(Y)} = \frac{\gamma_1^{(Y)}}{\gamma_0^{(Y)}}, \tag{1.3.15}$$

where $\beta^{(D)}$ is a real vector of the same dimension as $\beta^{(Y)}$ and $\gamma_0^{(D)}$ is a real parameter.

According to Delong *et al.* (2021), we can also obtain the probability density function of $(N_{i,j,t}, Y_{i,j,t})$ as follows (i is removed for easy reading) :

$$f_{j,t}(y,n) = \begin{cases} \prod_{t=1}^{T} \exp\left\{\frac{w_{j,t}}{\phi_{j,t}}\left(y\frac{\mu_{j,t}^{1-p}}{1-p} - \frac{\mu_{j,t}^{2-p}}{2-p}\right) + \log\left(\frac{\left(\left(\frac{w_{j,t}}{\phi_{j,t}}\right)^{\gamma+1} y^{\gamma}\right)^{n}}{n!\Gamma(n\gamma)y(p-1)^{\gamma n}(2-p)^{n}}\right)\right\} & n > 0 \\ \prod_{t=1}^{T_{i,j}} \exp\left\{-\frac{w_{j,t}\mu_{j,t}^{2-p}}{(2-p)\phi_{j,t}}\right\} & n = 0, \end{cases}$$

where $w_{j,t}$ and p represent respectively the weight and the Tweedie variance parameter. p is a positive real and verifies : $1 < p < 2$. The parameter γ is obtained as : $\gamma = \frac{2-p}{p-1}$.

By assuming the independence between the loss costs of contracts of distinct policies, we can calculate the contribution of policy i to the likelihood as (i is removed for easy reading) :

$$\prod_{j=1}^{J}\prod_{t=1}^{T} f_{j,t}\left(y_{j,t}, n_{j,t}\right). \tag{1.3.16}$$

To estimate all parameters, one can use the maximum likelihood strategy considering the (1.3.16). The idea is to define for each policy i, each vehicle j and each contract t, a response variable for the dispersion parameter $\phi_{i,j,t}$ as follows :

$$D_{i,j,t} = \frac{2}{\nu_{i,j,t}} \left(-w_{i,j,t} \left(Y_{i,j,t} \frac{\mu_{i,j,t}^{1-p}}{1-p} - \frac{\mu_{i,j,t}^{2-p}}{2-p} \right) - \phi_{i,j,t} \frac{N_{i,j,t}}{p-1} \right) + \phi_{i,j,t},$$

where $\nu_{i,j,t} = \frac{2w_{i,j,t}}{\phi_{i,j,t}} \frac{\mu_{i,j,t}^{2-p}}{(p-1)(2-p)}$.

The main motivation of this response variable is that we obtain : $\mathrm{E}[D_{i,j,t}] = \phi_{i,j,t}$. Therefore, the optimization of the likelihood function can be seen as two connected GLM : 1) GLM for the mean parameter ; and 2) GLM for the dispersion parameter. This approach is called Double-GLM (Double Generalized Linear Model (DGLM)) (Delong *et al.*, 2021).

1.3.3 Remarks

We close this section on Kappa-N models with a few remarks about the underlying distributions used.

1.3.3.1 Tweedie Case

For practical reasons, in our analysis, we carefully distinguish between the risk exposure $d_{i,j,t}$ of a contract and the weight $w_{i,j,t}$ of the Tweedie distribution. As previously noted, $d_{i,j,t}$ is a correction factor for the annual premium, whereas weight is a parameter that influences the Tweedie distribution modelling. In our analysis, we find that some values of the weight lead to an overestimation of the amount of total losses. For this reason, we suggest to consider the weight $w_{i,j,t} = d_{i,j,t}^{p-1}$. However, there will always be a difference between the total amount of losses and the total amount of expected losses. Even so, the difference between these two quantities is not very large. To correct this, one can adopt the off-balance correction as mentioned in the paper by (Denuit *et al.*, 2021).

1.3.3.2 Tweedie vs CPG

By considering the log-likelihood criterion, Delong *et al.* (2021) showed that the Tweedie model is preferable to the CPG model. However, to use this log-likelihood criterion, the data representation should be the same in each model. Delong *et al.* (2021) proposed a theorem (*Theorem 3.8* in their paper) to compare the log-likelihood obtained in each model. We consider the same theorem in our analysis (i is removed for easy

reading) :

$$\tilde{\boldsymbol{X}}'_{\boldsymbol{j,t}}\beta^{*Y} = \boldsymbol{X}'_{\boldsymbol{j,t}}\beta^{N} + \boldsymbol{X}'_{\boldsymbol{j,t}}\beta^{Z} + \gamma_0^{(N)}\ell_t^{(N)} + \gamma_0^{(Z)}\ell_t^{(Z)}, \tag{1.3.17}$$

$$\tilde{\boldsymbol{X}}'_{\boldsymbol{j,t}}\beta^{*D} = -\log(2-p) - (p-1)\boldsymbol{X}'_{\boldsymbol{j,t}}\beta^{N} + (2-p)\boldsymbol{X}'_{\boldsymbol{j,t}}\beta^{Z} - (p-1)\gamma_0^{(N)}\ell_t^{(N)} + (2-p)\gamma_0^{(Z)}\ell_t^{(Z)}, \tag{1.3.18}$$

where $\tilde{X}_{j,t} = \left(x_{j,t,0}, ..., x_{j,t,q}, \ell_t^{(N)}, \ell_t^{(Z)}\right)'$, β^{*Y} and β^{*D} are respectively the mean and the dispersion parameters of CPG in Tweedie parametrization.

Even if $\tilde{X}_{j,t} \neq \left(x_{j,t,0}, ..., x_{j,t,q}, \ell_t^{(Y)}\right)'$, we are able to compare the two models using (1.3.17) and (1.3.18).

1.3.4 Bonus-Malus Scale Models

One of the practical problems with the Kappa-N model is the excessive increase and decrease in premiums due to annual penalties and discounts. In order to limit the variation of premiums over time and to allow some form of forgiveness of old claims in the rating, another approach constrains the score ℓ to be between two limits for all the past contracts. Boucher (2022) called this approach the BMS Models. See Lemaire (2012) and Denuit *et al.* (2007a) for a historical review of BMS models. Instead of interpreting $\ell_{i,t}$ as a claims score, it can simply mean the BMS level. For an insured i, we define this BMS level as :

$$\ell_{i,t} = \ell_{i,t-1} - \sum_{j=1}^{J_i} \kappa_{i,j,t-1} + \Psi \times \sum_{j=1}^{J_i} n_{i,j,t-1}, \text{ with } \ell_{min} \leq \ell_{i,t} \leq \ell_{max}, \forall t = 1, \ldots, T. \tag{1.3.19}$$

Recursively, for any policy i and any contract t, the BMS level, $\ell_{i,t}$, is obtained as follows :

$$\ell_{i,t} = \ell_{i,1} - \sum_{\tau=1}^{t-1}\sum_{j=1}^{J_i} \kappa_{i,j,\tau-1} + \psi \sum_{\tau=1}^{t-1}\sum_{j=1}^{J_i} n_{i,j,\tau-1} = \ell_{i,1} - \kappa_{i,\bullet,\bullet} + \psi n_{i,\bullet,\bullet}.$$

It should be noted that these recursive equations are true regardless of the ℓ_{min} and ℓ_{max} limits, and confer the Markov property on the Kappa-N and BMS models. The starting BMS level ℓ_1 is set at 100 as in the Kappa-N model. Thus, for our three variables of interest, the corresponding premiums are calculated using

Table 1.3. We note that the difficulty of the inference is the definition of the BMS level, which is a co-variable and an endogenous variable to the model simultaneously. This BMS level depends on the parameters Ψ, ℓ_{min} and ℓ_{max} which are all unknown in the model. However, for the inference, one can use the profile maximization of the likelihood function on the three parameters : Ψ, ℓ_{min} and ℓ_{max}. The idea is to use all possible values of these three parameters to estimate the other parameters of the model. For the jump parameter, Ψ, the use of natural numbers is required.

Model	Premium	BMS level
Frequency	$\pi_{i,j,t}^{(N)} = d_{i,j,t} \exp\left(\boldsymbol{X}_{i,j,t}' \beta^{(N)} + \gamma_0^{(N)} \ell_{i,t}^{(N)}\right)$	$\ell_{i,t}^{(N)} = \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(N)} n_{i,\bullet,\bullet}$ $\ell_{min}^{(N)} \leq \ell_{it}^{(N)} \leq \ell_{max}^{(N)}$
Severity	$\pi_{i,j,t}^{(Z)} = \exp\left(\boldsymbol{X}_{i,j,t}' \beta^{(N)} + \gamma_0^{(Z)} \ell_{i,t}^{(Z)}\right)$	$\ell_{i,t}^{(Z)} = \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(Z)} n_{i,\bullet,\bullet}$ $\ell_{min}^{(Z)} \leq \ell_{it}^{(Z)} \leq \ell_{max}^{(Z)}$
Loss Cost	$\pi_{i,j,t}^{(Y)} = d_{i,j,t} \exp\left(\boldsymbol{X}_{i,j,t}' \beta^{(Y)} + \gamma_0^{(Y)} \ell_{i,t}^{(Y)}\right)$	$\ell_{i,t}^{(Y)} = \ell_1 - \kappa_{i,\bullet,\bullet} + \Psi^{(Y)} n_{i,\bullet,\bullet}$ $\ell_{min}^{(Y)} \leq \ell_{it}^{(Y)} \leq \ell_{max}^{(Y)}$

Table 1.3 – Premiums in BMS models

1.4 Numerical Application

1.4.1 Description of data

We consider a non-random sample of an automobile insurance database of a major insurer in Canada over a total period of 13 consecutive years. The data concern the Canadian province of Ontario and contain more than 2 million observations. Each observation corresponds to an annual contract for one vehicle. The form of the database is similar to Table 1.1 introduced at the beginning of our paper. For each observation in the database, we have a policy number, a vehicle number as well as the effective and end date of the vehicle contract. Several characteristics of the insured or the insured vehicle are also available. Finally, for each

contract for each vehicle, the number of claims and the cost of each claim are available.

The database is also divided into coverage type, which provides information on third-party liability, collision and comprehensive claims. To illustrate the approach described in this paper, we focused on a single cover. Thus, in connection with the defined terms in the introduction, we illustrate our pricing model by experience with :

- A target variable based on collision coverage, representing the property damage protection of auto insurance for accidents for which the driver is at fault ;
- A scope variable also based on collision coverage.

As mentioned earlier, however, the proposed pricing approach is very flexible, and any combination of target and scope variables would be possible. For example, the analysis of the sum of liability and collision coverage could be analyzed by conditioning on the experience of past claims of comprehensive coverage.

For confidentiality reasons, the full description of the data will not be detailed. That being said, we can provide some summary information for the studied sample :

- The observed annual claims frequency is approximately 2% ;
- The average severity of a claim is around $7,500 ;
- The average annual loss cost is about $160 for all available years ;
- The average number of vehicles insured per contract is around 1.70 ;
- On average, a vehicle is insured for 2.75 contracts.

We also split the data into a training set and a test set. To maintain the dependency between the contracts and the vehicles of the same policy, the training and test set were formed by policy number selection. For example, if a policy is in the training set, it means that all vehicles and contracts in that policy are in the same training set. Thus, 75% of the policies were assigned to the training set and 25% to the test set. These correspond respectively to 75% and 25% of all observations. We made these splits by ensuring that we had the same claims frequency and the same average claims severity in each of the two sets.

1.4.1.1 Available Covariates

We have several characteristics for each vehicle and for each contract. In order to illustrate the impact of segmentation in rating, we select eight of these characteristics as covariates. For confidentiality reasons, but also because this is not the focus of the paper, these covariates are simply labelled as $X1 - X8$ and their meaning is not explained. A summary of the proportions of each modality of these variables is given by Figure 1.1. To be consistent with the rating approach usually used in practice, which is also often used in the actuarial scientific literature, we have chosen classic risk characteristics related to the sex and age of the insured, the use of the vehicle or the type of vehicle driven, for example. We did not seek to artificially inflate residual heterogeneity from risk characteristics that are not used in pricing. Thus, we consider that the pricing model developed with the chosen covariates is representative of standard pricing models.

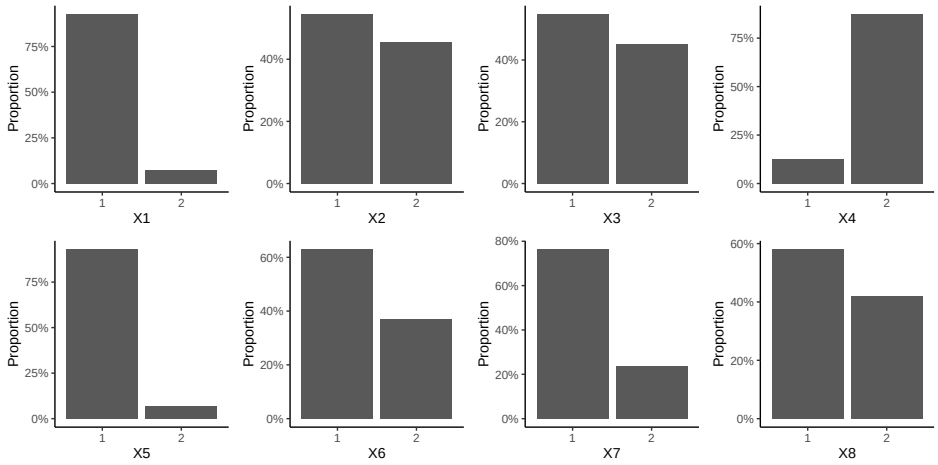


Figure 1.1 – Distribution of covariates

In addition to the selected covariates, indicator variables for each of the calendar years of the contracts were included in the modelling.

1.4.1.2 Impact of Past Insurance Experience

Although we have 13 years of experience, we use a portion of the database to create a claims history for all insureds. It should be noted that many insureds in the database during the first year have a claims history with the insurer. However, this claims history is not available owing to the structure of the data. Therefore, the first six years of the database are used exclusively to obtain the claims history of each insured, and only

the following seven years are used for modelling purposes. For consistency purposes, a fixed window of six years in the past is always used to calculate past claims statistics, $n_{i,\bullet,\bullet}$ and $\kappa_{i,\bullet,\bullet}$, for each of the insureds and each contract. In other words, this means the BMS level of a contract t depends only on the claims experience of contracts $t-1,\ldots,t-6$ and not on the clains experience of the contracts $t-7$, $t-8\ldots$. The impact of this choice of window on the models used is minimal, but it does mean that the Markovian property for a single contract, defined in Equation (1.3.19), no longer holds. See Boucher (2022) for a study of the window of experience to be used in predictive ratemaking.

A classic quote from Lemaire (2012) is that if only one segmentation variable were to be used for the rating, it should be based on claims experience. For our preliminary analysis of the impact of claims history on premiums, we create six groups of contracts according to their past experience. The first three groups of contracts are based on the number of past contracts, categorized into the intervals [0,1], [2,3] and [4,5] according to the number of previous contracts, and contain only those insureds who could be qualified as inexperienced. We can also call them the new insureds or new policies. The last three groups of policies include only insureds who have been observed for six years or more. The difference between the three groups is based on claims history : the insured in group E and F have filed claims at least once in the past while group D insureds have not filed claims in the last six years. Table 1.4 summarizes the groups of contracts and indicates the frequency, severity and loss cost ratios. These ratios are obtained by dividing the average claims frequency, average claims cost and average loss cost by the corresponding average of each group.

For each of the seven years studied, Figure 1.2 shows the frequency and severity ratios for each group. For a given calendar year or for all years combined, the **Frequency Ratio** is defined as the ratio of the observed frequency for a group to the observed frequency of the portfolio. This value indicates how much better or worse a group of policyholders is than the portfolio average. The **Severity Ratio** and the **Loss Cost Ratio** are defined in the same way.

	Type	Past Experience	$n_{i,\bullet,\bullet}$	Frequency Ratio	Severity Ratio	Lost Cost Ratio
New insureds	A	$[0, 1]$	-	1.303	1.110	1.446
	B	$[2, 3]$	-	1.025	1.027	1.052
	C	$[4, 5]$	-	0.915	0.950	0.870
Other insureds	D	≥ 6	0	0.743	0.863	0.641
	E	≥ 6	1	1.061	0.936	0.993
	F	≥ 6	≥ 2	1.572	0.989	1.555

Table 1.4 – Group of contracts by past experience

Although the impact of covariates may need to be considered in order to better understand the statistics shown in Table 1.4 and Figure 1.2, it is still relevant to comment directly on each group.

1.4.1.2.1 Type A

We observed that new policyholders in Group A have a much worse claims experience than other groups, in terms of both frequency and severity. With a claims frequency 30.3% higher than average, and an 11.0% higher severity, the total burden of Group A policyholders is approximately 45% higher than the portfolio average.

1.4.1.2.2 Type B

Group B insureds, with only 1 or 2 years of experience more than the to Group A insureds, seem to differ from the latter. Indeed, the curves illustrated in Figure 1.2, representing their loss experience in frequency and severity compared to the portfolio average, are close to one. The insureds of Group B have a higher claims frequency and claims severity than the insureds of Group C, who have one or two years more experience than Group B.

1.4.1.2.3 Type C

Group C policyholders have four or five years of past experience. They may or may not have had claims during those years. However, when we look at Figure 1.2, their average claims frequency and severity are better than the averages of the portfolio. We can see, based on the number of years of experience in the

company, that a minimum of about four years is necessary to have insureds with claims experience similar to the average of the portfolio.

1.4.1.2.4 Type D

Insureds in this group have insurance experience but have never filed a claim in the last six years. What can be quickly noticed from the figure and the table is that experienced insureds who have not claimed in the last six years (Type D) have a lower claims frequency than other insureds. Surprisingly, this same group of insured also has a better severity than the others.

1.4.1.2.5 Type E

Group E policyholders are those who have insurance experience but have filed a claim once in the last six years. These insureds have a claim frequency comparable to the new insureds with two and three years of insurance experience. In contrast, their average claims costs are generally lower than new insureds and the average claims cost of the portfolio.

1.4.1.2.6 Type F

Finally, Group F insureds also have insurance experience but have made at least two claims in the last six years. These insureds are the ones who produce the most interesting claims statistics. In fact, they have a 57% higher frequency of claims than the portfolio average. They claim more than the new policies of Group A. However, their average claims cost is lower than the average claims cost of the portfolio compared to the insureds in Group A.

Through these analyses, we show how important it is to correctly identify the contracts of new insureds (especially those in Group A) from those of Group D because the insureds of these two groups have the same value for $n_\bullet$. The use of a covariate, counting the number of past contracts without claims $\kappa_\bullet$ is then justified.

Finally, to better understand how the past insurance experience impacts each target variable, it is necessary to model their distribution. One can also use other covariates to have flexible rating models.

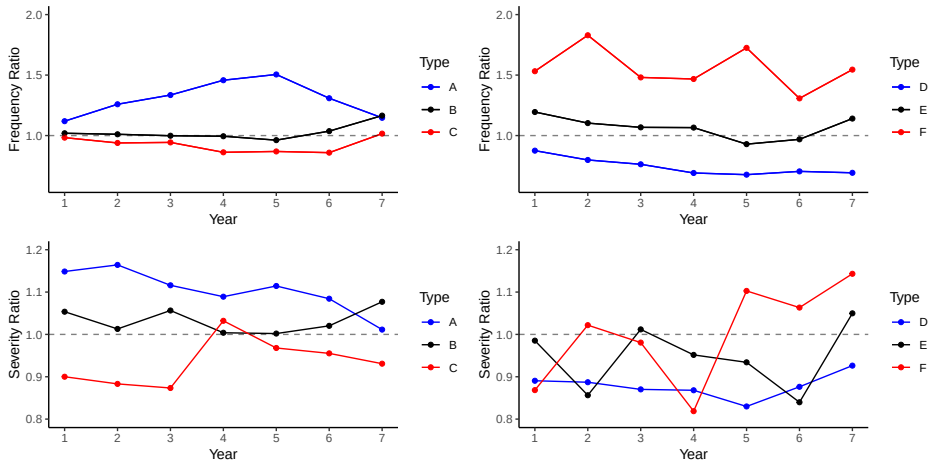


Figure 1.2 – Average claim frequency and severity by group

1.4.2 Covariate selection

The data were used to adjust three types of models for frequency, severity, and loss cost :

1. A model that will be called **standard**, which has no component related to past claims experience ;
2. The **Kappa-N** model (Section 1.3.2) using covariates $n_\bullet$ and $\kappa_\bullet$;
3. The **Bonus-Malus Scale** (Section 1.3.4) limiting the claims Score to ℓ_{min} and ℓ_{max}.

For each of the models, we considered the same vector of characteristics except for the Kappa-N and BMS models which use also use the covariates $\kappa_{i,\bullet,\bullet}$ and $n_{i,\bullet,\bullet}$. However, not all risk characteristics consistently have the same impact in our three variables of interest. For example, the frequency of claims is greatly impacted by the characteristics of the insured, such as age and gender, while the severity of collision coverage will usually be more influenced by the characteristics of the vehicle, mainly the value of the vehicle. Therefore, a statistical procedure for selecting covariates seems necessary.

We adopted the *Elastic-net* regularization to select the covariates. This method is seen as a combination of Lasso and Ridge regressions. See Hastie *et al.* (2009) and Hastie *et al.* (2015) for more details about this approach. One of the advantages if this approach is that it solves the redundancy of variables and the multicollinearity of risk factors. The idea of the procedure is to impose constraints on the coefficients of the model. Excluding the intercept from the procedure, the constraint to be added to the log-likelihood score to

be maximized is expressed as :

$$\left(\alpha \sum_{j=1}^{q+1} \mid \boldsymbol{\beta}_j^{(.)} \mid + \frac{(1-\alpha)}{2} \sum_{j=1}^{q+1} \boldsymbol{\beta}_j^{(.)2} \right) \leq \lambda, \;\; \lambda > 0, \;\; 0 \leq \alpha \leq 1.$$

This penalty constraint depends on the values chosen for the parameters α and λ. If $\alpha = 0$, the *Elastic-net* is equivalent to a Lasso regression. In contrast, if $\alpha = 1$, it is equivalent to a Ridge regression. For each studied model, the optimal values of λ and α were obtained by a cross-validation using deviance as a selection criterion.

1.4.3 Fitting Statistics and Prediction Scores

Table 1.5 shows the fit results of the models based on the training set, and the prediction quality based on the test set. The number of parameters used in the model (after applying the procedure by elastic net), the log-likelihood, as well as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for each of the models are indicated. To evaluate the prediction quality based on the test set we avoided using least squares because they are not always adequate for frequency, severity, or loss cost statistics. A logarithmic score SL representing the negative loglikelihood value on the test set with the parameters estimated on the training set is used instead. More formally, if we denote by $\hat{\boldsymbol{P}}$ the estimated parameters for each of the models, the logarithmic prediction score is calculated as : $SL = -\log(f(\hat{\boldsymbol{P}}|Test\ set))$, where $f()$ is the probability density of function of each target variable. As with least squares, the aim is to obtain the smallest value of SL on the test set in order to control the over-fitting resulting from the estimation of the parameters on the training set. Optimal use of this score implies the estimation of all parameters by maximizing the likelihood function and not only the parameters associated with the mean of each target variable.

Model		Number of parameters	Log-likelihood	Train AIC	BIC	Test Score
Standard	*Poisson*	*15*	*-97,786*	*195,602*	*195,782*	*32,939*
	Gamma	*15*	*-187,229*	*374,489*	*374,606*	*63,396*
	CPG	30	-284,252	568,564	568,924	95,982
	Tweedie	31	-281,849	563,760	564,131	95,221
Kappa-N	*Poisson*	*17*	*-97,426*	*194,887*	*195,090*	*32,811*
	Gamma	*16*	*-187,189*	*374,410*	*374,536*	*63,384*
	CPG	33	-283,885	567,836	568,232	95,851
	Tweedie	33	-281,490	563,046	563,441	95,092
BMS	*Poisson*	*19*	*-97,421*	*194,876*	*195,079*	*32,812*
	Gamma	*18*	*-187,189*	*374,413*	*374,555*	*63,382*
	CPG	37	-283,881	567,836	568,280	95,852
	Tweedie	34	-281,487	563,042	563,449	95,097
	Tweedie's CP	37	-282,151	564,376	564,819	95,260

Table 1.5 – Model comparisons

1.4.4 Comparison between Models

1.4.4.1 CPG vs Tweedie

The first angle of analysis is to compare the fit and prediction quality of the CPG model with the Tweedie model. Some analyses of the insurance data showed that the gamma distribution was not rejected by a hypothesis test (the QQ-plots are available in Appendix A.2), but the Poisson distribution for the number of claims is not ideal. Given the convergence of the Poisson model estimators, however, the assumption of a Poisson distribution for the annual claims number will be retained, which will allow us to continue the analysis with Tweedie.

As mentioned by Delong *et al.* (2021), to use likelihood based criteria to compare the two approaches (CPG and Tweedie), the data samples must be the same in each model. So, using our remark from 1.3.3, we get the adjustment statistics and the prediction scores of the two models as given by Table 1.5.

Delong *et al.* (2021) also argued that the fit quality of the Tweedie is always better than that of the CPG if the same covariates are used in both models and under other assumptions that we have considered here as well. However, we do not have the same covariates in the two models due to the Elastic-net procedure and

the addition of the claims score or BMS level in the mean parameter modelling in each model. Considering the adjustment statistics and the prediction scores in Table 1.5, the Tweedie models are better than the CPG models.

For only the BMS case, we consider a Tweedie model (**Tweedie's CP**) with the same covariates selected as in the CPG model in line with Delong *et al*. (2021). This Tweedie model is better than CPG, as Delong *et al*. (2021) assert, but our proposed Tweedie model is still better than Tweedie's CP model.

1.4.4.2 Standard vs Kappa-N and BMS

Another angle of analysis, more specific to what we have presented in the paper, is related to the modelling of past experience. First, the analysis should be divided according to the chosen distribution, and then a summary analysis should be done of all the models combined.

1.4.4.2.1 Poisson (Frequency) :

The introduction of a claim score or BMS level into the mean parameter of a Poisson distribution is not new. Nevertheless, it is still interesting to see that the Kappa-N and BMS models significantly improve the AIC and BIC statistics, compared to the standard models. The prediction score is also improved if one switches from the standard model to the Kappa-N or BMS model. The differences in the fit statistics and prediction score of the Kappa-N model and the BMS model are minimal. Knowing that the Kappa-N model is difficult to use in practice, it is interesting to see that the cost of having a predictive pricing model that has a potential for use is negligible.

1.4.4.2.2 Gamma (Severity) :

Modelling severity based on the number of past claims is not a very common approach in actuarial science. As we saw earlier in the severity analysis based on five groups of insureds, severity approaches that included a loss experience component were expected to perform well. This is what we are seeing : the Kappa-N and BMS models produced better values of AIC, BIC, and a better logarithmic prediction score than did the standard model. Considering that the standard model does not use claims history in the premium calculation, this result is very interesting since it seems to indicate that there is value for insurers in including a component of discount and surcharge on severity based on past claims. As with frequency, the observed differences

between the values of AIC, BIC and logarithmic SL score are very small between the BMS approach and the Kappa-N approach, showing once again that the requirement to have a practical approach is not very restrictive.

1.4.4.2.3 CPG (Loss cost) :

The CPG model is the combination of the frequency and severity approach discussed above. Both the frequency and severity approaches favor the Kappa-N and BMS models. Thus, it is obvious that both of these approaches will be better than the standard approaches. We also notice that the results are much better with the BMS model except for the BIC criterion, which causes a greater number of parameters to be penalized, at 37 compared to 33 for Kappa-N.

1.4.4.2.4 Tweedie (Loss cost) :

As with severity modelling, the use of the number of past claims to model the loss cost is not a very common approach in actuarial science and it is therefore very interesting to check whether this generalization of the approach is relevant. Analysis of the AIC, BIC and SL tends to show that the addition of elements related to past insurance experience helps to better segment the risk for collision coverage. Indeed, the values obtained for the Kappa-N and BMS approaches strongly favor these models, compared to the standard approach. Just like before, the BMS model performed better than the Kappa-N except for the BIC criteria and the prediction score, where a difference was observed for the Kappa-N model.

1.4.5 Analysis of Premiums

1.4.5.1 Estimated Parameters

Table A.1 in Appendix A.1 shows the estimated values of the β used with the selected covariates for the Standard model and the BMS model. For frequency, severity, and loss cost, there is a marked difference between some estimators. The impact of adding components linked to a BMS level on parameters related to segmentation variables had already been observed and analyzed by Boucher et Inoussa (2014) for frequency analysis. We will not elaborate further, especially since the same explanations of the reasons for these differences apply to the analysis of severity and loss cost.

Above all, it is necessary to analyze in a little more detail the values of the estimators of the parameters

related to the claim experience. Table 1.6 shows the value of the parameters related to past claims experience for the Kappa-N model and the value of the structural parameters for the BMS approach. For frequency, severity, and loss cost, the table shows that the impact of estimating the parameters γ_0 and Ψ is small when minimum and maximum BMS limits are added, i.e. ell_{min} and ℓ_{max}.

	Kappa - N			BMS		
Parameters	Poisson	Gamma	Tweedie	Poisson	Gamma	Tweedie
γ_0	0.081	0.025	0.107	0.094	0.026	0.112
ψ	2.899	2.073	2.728	3	2	3
$[\ell_{min}, \ell_{max}]$	-	-	-	[95,106]	[94,100]	[95,104]

Table 1.6 – Estimation of the other parameters of the Kappa-N and BMS models

The results in Table 1.6 indicate that the jump parameter for a past claim is the same for the frequency and loss cost ($\Psi^N = \Psi^Y = 3$), but this parameter is different for severity ($\Psi^Z = 2$). The relativity parameter γ_0 of the frequency is also closer to that of the loss cost than to that of the frequency. Finally, the frequency model proposes BMS level limits that are slightly larger than the severity or loss cost. To better understand how the experience of past claims impacts policyholders' premiums according to the studied models, we can refer to Figure 1.3 which illustrates the relativity curve of the frequency (Poisson), the severity (gamma) and loss cost (Tweedie) as a function of BMS level. The impact of the parameters Ψ, γ_0, ℓ_{min} and ℓ_{max} can all be observed simultaneously in the same figure :

— It shows that the range of possible penalties for severity (in blue) is much smaller than those for frequency (red) and loss cost (black).

— The maximum penalty for frequency is higher than the maximum penalty for loss cost, which is much higher than the penalty for severity. We reach the same conclusion by comparing the maximum discount obtained in each model.

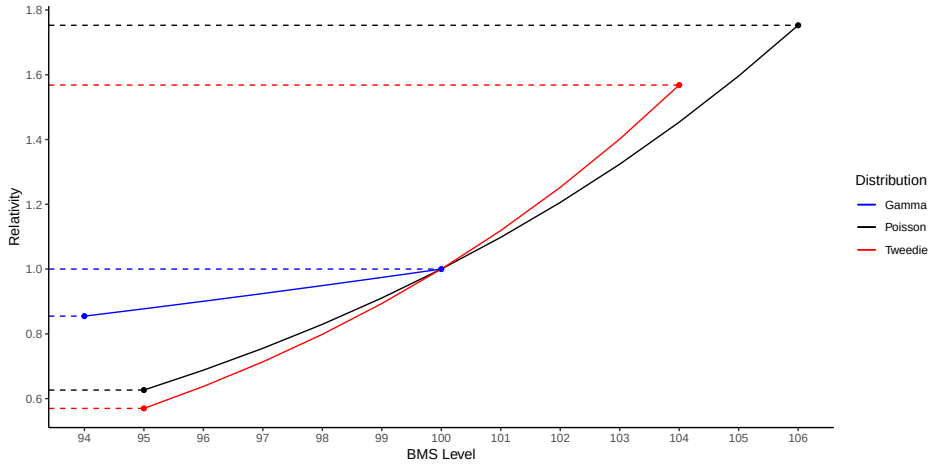


Figure 1.3 – BMS relativites

In the CPG model, an insured's total premium is the frequency multiplied by the severity. For the premium calculation of this model, the BMS levels of frequency and severity must be calculated. Thus, it is not possible to illustrate the BMS relativities of the CPG model in two dimensions, as is done in Figure 1.3. To compare BMS surcharges, discounts, and minimum and maximum relativities, BMS models of frequency and severity should be combined. The result of this comparison is shown in Table 1.7. The direct comparison between the CPG model and Tweedie model shows that the surchages of the two approaches for a claim are similar : a premium increase of 40.1% compared to an increase of 39.5%. The discounts for a claims-free year are also similar : -11.3% versus -10.6%. The most striking difference between the two models is revealed above all at the level of relativity of each model. In the CPG approach, an insured with a lot of past claims might have a surchage of more than 175.3% while this surchage is limited to 156.8% for the Tweedie model. In contrast, the maximum discount is comparable in both models.

	Poisson	Gamma	CPG	Tweedie
Surcharge per Claim	0.324	0.054	0.395	0.401
Claims Free Discount	-0.089	-0.026	-0.113	-0.106
Min. & Max. Relativities	[0.626, 1.753]	[0.855, 1.000]	[0.535, 1.753]	[0.570, 1.568]

Table 1.7 – Impacts of past claims for all BMS Models

Insured	Years (t)											
i	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	1	2	0	1	2	0	1	2	0	1
3	4	1	3	0	1	0	0	0	0	0	2	0
4	0	2	0	0	0	0	0	1	0	3	1	4

Table 1.8 – Insureds with claims experience

1.4.5.2 Numerical Example

To better illustrate the similarities and differences between the CPG model and the Tweedie model, we will use the estimated parameters of the models and thus assume four insureds with the claim history shown in Table 1.8. Each insured was observed for twelve consecutive years. The first insured has not filed a claim during the 12-year periods, insured #2 is a bad driver who claims frequently, insured #3 filed many claims in the first three years, but the number diminished during the other years, and the last insured has a deteriorating driving experience. It will be assumed that all insured persons start at the BMS level of 100 at year 0, for the bonus-malus scale of frequency, severity and loss cost. As was done with the insurance data used earlier, a 6-year window is assumed for the calculation of levels and the first 6 years are used to create a claims history. We will analyze the resulting premiums for each insured in years 7 to 12.

At the beginning of each year t, Figure 1.4 shows the evolution of BMS levels of frequency, severity and loss cost for each of the four fictitious insureds. The grey area of each graph corresponds to the first six years used for the calculation of the initial BMS level.

Figure 1.5 shows the resulting BMS relativity, where the BMS relativity of the CPG model is the combined effect of frequency and severity. For all years after the sixth contract ($t \geq 7$), we can see that the BMS relativities for each of the two models are similar for the four insureds in the example. Thus, even though the BMS levels sometimes appear to be different, the combination of severity and frequency means that the relativity obtained is close to that of Tweedie. Of course, there are some differences between the two curves, but the general trend is always the same.

Obviously, the four insureds in the example are fictitious situations, and the supposed history may not have taken place in the database used. The main purpose of the example is to show that despite the imposition of

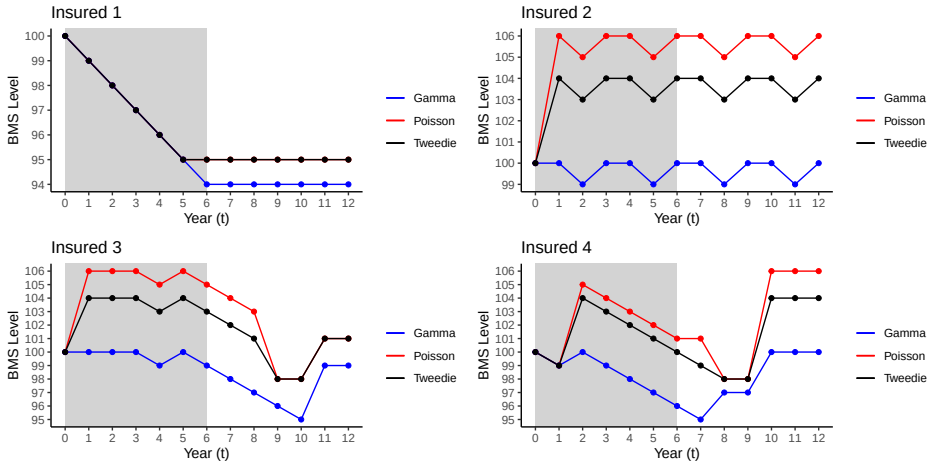


Figure 1.4 – BMS levels for all four fictitious insureds

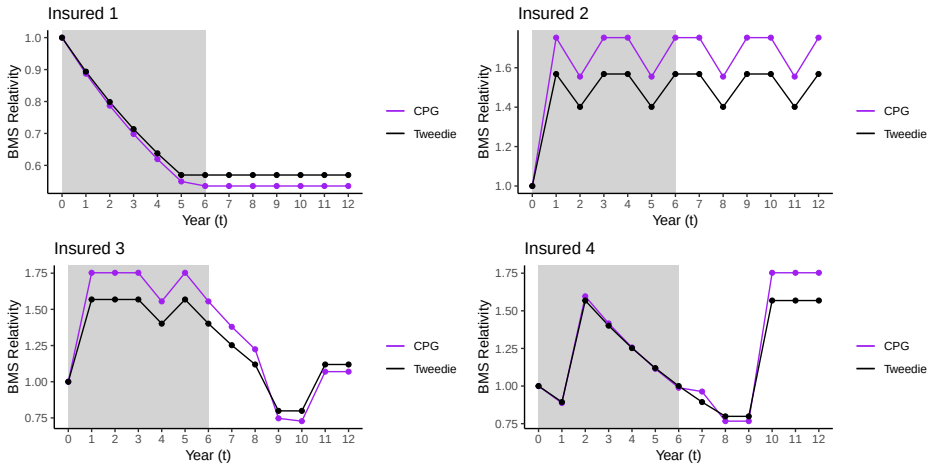


Figure 1.5 – BMS relativities for all four fictitious insureds

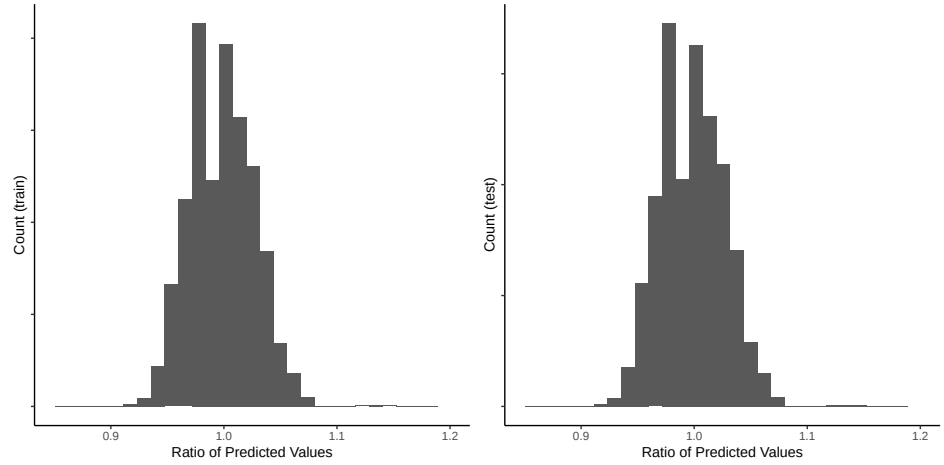


Figure 1.6 – Premium ratio (left : training set, right : test set)

a rigid predictive rating structure, the models are still comparable.

1.4.5.3 CPG vs Tweedie

Figure 1.6 shows the ratio between the CPG premium and Tweedie for the training set and test set. The distribution is similar for both parts of the dataset. We can also see that the premium ratio is around 1 but that spreads of minus 95% or more than 110% exist in the portfolio. The choice of a type of model for predictive pricing thus has a significant potential impact.

1.4.6 Predicted and Observed Loss Cost

1.4.6.1 Types of Insureds

A relevant way to compare BMS models with the two underlying distributions (CPG and Tweedie) is to check the fit between what is observed and what has been predicted for each of the models according to the type of contracts. By taking the five types of policyholders defined in Section 1.4.1.2, we can potentially see the type of contract for which the two Tweedie models could be improved. Table 1.9 below shows the ratio of the predicted loss cost to the annual average for the five types of policyholders. Figure 1.7, in contrast, illustrates the observed pure charge ratio and those predicted for both models : the two graphs at the top are for CPG and the ones at the bottom are for Tweedie. For the test portion of the database, the graphs on the left refer to type A, B and C insureds (and therefore what could be called new insureds), and those on the right indicate the result for type D and E insureds, i.e. insured with at least six years of insurance experience.

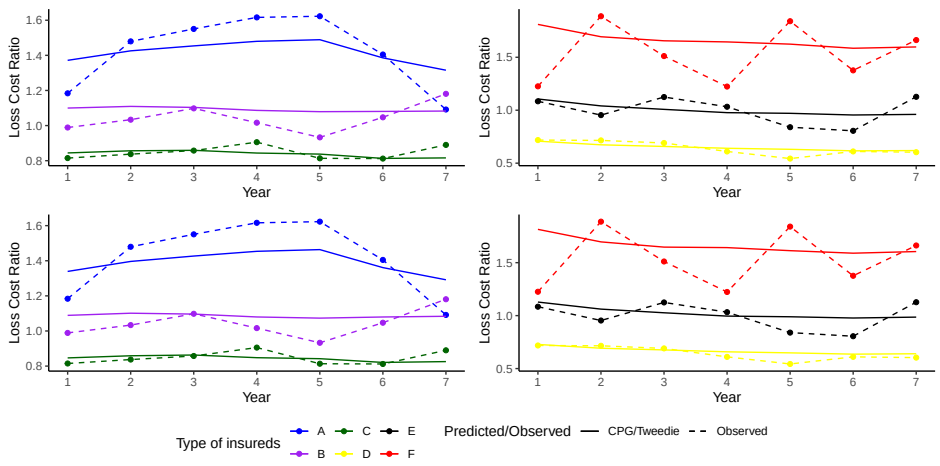


Figure 1.7 – Loss Cost for the test set (top : CPG, bottom : Tweedie, left : Type A-B-C, right : Type D-E-F)

This analysis by type of insured makes it possible to see the type of insured that seems to be best or worst predicted by the different models. The two approaches, CPG and Tweedie, predict the loss costs of policyholders whose average total charge is close to or smaller than that of the portfolio. This prediction is almost perfect for Group D. In contrast, for policyholders who have an average loss cost higher than that of the portfolio, the costs are generally underestimated by both approaches.

	Training Set			Test Set		
Type	Observed	CPG	Tweedie	Observed	CPG	Tweedie
A	1.482	1.407	1.396	1.434	1.408	1.399
B	1.068	1.108	1.105	1.047	1.102	1.099
C	0.896	0.854	0.856	0.861	0.851	0.853
D	0.644	0.650	0.670	0.641	0.649	0.668
E	0.933	1.012	1.034	1.013	0.999	1.020
F	1.393	1.658	1.657	1.613	1.669	1.668

Table 1.9 – Loss Cost Ratio for all types of insureds

1.4.6.2 BMS Levels

It may also be interesting to check the fit between the observed and the predicted values according to the BMS level of the contract. Figure 1.8 illustrates this fit for frequency, severity, and loss cost. The blue curves represent the predicted mean relativity while the red curve represents the observed relativities. The solid lines are for the training set while the dotted lines represent the results of the test set. For the frequency of

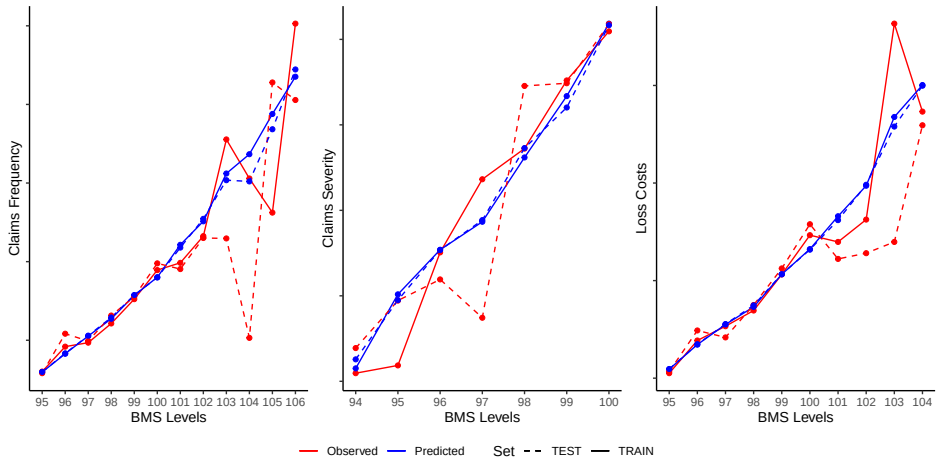


Figure 1.8 – Predicted vs Observed for the claims frequency (left) and the claims severity (right)

claims, we see that the difference between predicted and observed relativities is minimal for contracts with BMS levels below 103. For levels 103 and above, where there are far fewer insureds, we can see that the general trend of the model is in line with the average of what has been observed. For the severity model, the relativity curve clearly shows a decrease when the BMS level decreases, which is what the pricing model also assumes. The difference between the observed and the predicted values is more variable for severity than for frequency. Finally, the gap between the observed relativities and the relativities obtained by the Tweedie model for the total charge is close to what was observed for the frequency model : the difference is minimal for lower BMS levels and more variable for higher levels.

1.5 Conclusion

We generalized the paper of Delong *et al.* (2021) by including a predictive ratemaking component in the premium. Our approach can also be seen as an extension of the work of Boucher (2022) by considering the severity and the loss cost as target variables in addition to the frequency. In other words, our objective was to compare the BMS models to the standard models when the CPG and Tweedie are used as underlying distributions. Our main conclusions and remarks are as follows.

First, in the BMS model with CPG as an underlying distribution, we found that BMS level impacts the frequency and severity components of the CPG differently. Although both are positively related to the BMS level, the impact is stronger for the frequency component than the severity component. This results in the surcharge and discount being significant in the frequency component than severity. Finally, by comparing the relativities,

the frequency component penalizes insureds who have made a lot of claims in the past more than the severity component does.

Second, in the BMS model with Tweedie as an underlying distribution, we found positive dependency between the BMS level and the corresponding premium. We also noted that the surcharge by claim and the claims-free discount are comparable in both the Tweedie and CPG cases. However, the CPG model penalized insureds who have made a lot of claims in the past more than the Tweedie model did.

Finally, we obtained the same conclusions as Boucher (2022) about the BMS model. All BMS models considered in our analysis have better quality data adjustment and data prediction than the standard approaches do. These statistics are better when the Tweedie is used as an underlying distribution compared with the CPG. In addition, the Tweedie model (**Tweedie**) with a unique BMS level is better than the Tweedie model (**Tweedie's CP**), which uses the BMS levels obtained in the CPG model.

The results obtained in this paper were indeed found thanks to a vast automobile insurance database (more than 2 million contracts). It might be interesting to check whether the BMS model proposed in this paper is robut and fits nicely into small and medium-sized insurance portfolios.

We conclude the paper with some remarks about the underlying distributions used. First, there are other excellent distributions for the frequency and the severity modelling. For example, the Negative-Binomial is preferable to the Poisson. Yet to compare the CPG and Tweedie models, the frequency and severity components of the CPG must be modelled by the Poisson and gamma distributions. Finally, due to the positive probability density function of loss cost when this loss cost is zero, a mixture distribution (discrete and continuous) may be an adequate choice for the loss cost modellling. This is why we make an appropriate choice of the variance parameter of the Tweedie distribution to allow loss cost modelling.

CHAPITRE 2

COMPARISON OF OFFSET AND RATIO WEIGHTED REGRESSIONS IN TWEEDIE MODELS WITH APPLICATION TO MID-TERM CANCELLATIONS

2.1 Introduction

In insurance, particularly within the property and casualty sector like automobile insurance, risk exposure has traditionally been associated with the duration of coverage (see Denuit *et al.* (2007b)). Risk exposure refers to the extent to which an insured entity is subject to potential claims or losses over the policy period. For example, under a linear interpretation of exposure, a one-year policy would be considered twice as risky as a six-month policy, assuming identical characteristics for the insured. This concept is central to many actuarial research projects and practical applications. Given its importance, it becomes essential to develop a comprehensive framework for incorporating various dimensions of risk exposure into the modelling of random variables used for ratemaking. Such models ensure that the premiums charged accurately reflect the underlying risks. More formally, we are interested in modelling the random variable Y_i, representing either the total number of claims or the total claims amount (also referred to as the loss cost) for contracts $i = 1, \ldots, n$, where n denotes the total number of contracts in the portfolio. Assuming that the expected value corresponds to the premium (as discussed in Denuit *et al.* (2007b) and Frees *et al.* (2014)), the premium can be written as :

$$\mathrm{E}[Y_i|\mathbf{X_i}] = g^{-1}\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), \tag{2.1.1}$$

where $\boldsymbol{\beta}$ is a parameter vector to be estimated, and $\mathbf{X_i} = (x_{i,0}, \ldots, x_{i,q})^\top$ represents the vector of risk characteristics associated with contract i (with q being a positive integer). These risk characteristics—such as age, sex, and the value of the insured goods—are used as covariates in the regression model, allowing the premium to vary based on these factors. The function $g(\cdot)$ serves as a link function that connects the score $\mathbf{X}_i^\top \boldsymbol{\beta}$ to the expected value, similar to the approach in Generalized Linear Models (GLMs) (see Nelder et Wedderburn (1972)).

An analysis of historical motor insurance data reveals significant variability in exposure periods among insured individuals. This variation often occurs when coverage is cancelled before the contract ends due to

reasons such as the sale or replacement of a vehicle, theft, or a total loss resulting from an accident. For modelling purposes, insurers typically treat the exposure period as a fraction of the total number of days in a year. This fraction, referred to as **risk exposure** and denoted by t, is a positive value between 0 and 1, reflecting the proportion of the year that the contract was active. Formally, we consider a portfolio of n insured contracts, indexed by i, some of which have a risk exposure, denoted by t_i, that is less than 1. To account for this variation in exposure when calculating premiums and estimating model parameters, it is necessary to generalize Equation (2.1.1). This paper examines two approaches within the maximum likelihood estimation framework to address this generalization :

1. The first approach assumes that the exposure t_i is proportional to the mean of the conditional distribution $Y_i|\mathbf{X_i}$, where $f_{Y_i}(\cdot)$ denotes the assumed density or probability mass function of $Y_i|\mathbf{X_i}$, and y_i the observed loss cost for contract i. Under this assumption, the objective function for estimating the parameter vector $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}|y_1,\ldots,y_n) = \sum_{i=1}^{n} \log\left(f_{Y_i}(y_i)\right),$$

   where the mean of Y_i is linearly proportional to t_i,

$$\mathrm{E}[Y_i|\mathbf{X_i}] = g^{-1}\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) t_i.$$

   With the logarithmic link function $g(\cdot)$, this yields

$$\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) t_i = \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} + \log(t_i)\right),$$

   so that $\log(t_i)$ acts as an offset variable. This method is therefore referred to as the **offset approach**.
2. The second approach, referred to as the **ratio approach**, annualizes the random variable Y_i to facilitate comparisons of policyholders' claims on an annual basis. A new random variable $Z_i = \frac{Y_i}{t_i}$ is defined for each contract $i = 1,\ldots,n$, producing normalized claim amounts $z_1,\ldots,z_n$.

In this framework, the mean of $Z_i|\mathbf{X_i}$ no longer depends on t_i, while the assumed density or probability mass function incorporates the exposure through weighting. The objective function for estimating $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}|z_1,\ldots,z_n) = \sum_{i=1}^{n} t_i \log\left(f_{Z_i}(z_i)\right),$$

where $f_{Z_i}(\cdot)$ is denotes the assumed density or probability mass function of $Z_i|\mathbf{X_i}$, and

$$\mathrm{E}[Z_i|\mathbf{X_i}] = g^{-1}\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right).$$

With a logarithmic link function, the expected loss cost for an insured exposed to risk over t_i years is

$$\mathrm{E}[Y_i|\mathbf{X_i}] = \mathrm{E}[t_i Z_i|\mathbf{X_i}] = \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) t_i = \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} + \log(t_i)\right),$$

which has the same form as in the offset approach.

It is important to emphasize that in the offset approach, the random variable Z_i can be used to estimate the parameter vector $\boldsymbol{\beta}$, but this method should not be confused with the ratio approach. Similarly, in the ratio approach, the random variable Y_i can also be used, but this method should not ne mistaken for the offset approach. We refer to the offset approach when risk exposure is only incorporated into premium modelling as an offset variable. Conversely, we refer to the ratio approach when risk exposure is only included as a weight in the premium modelling process. These two approaches represent distinct methods for integrating risk exposure, ensuring that premiums are appropriately aligned with the insured's actual level of risk.

2.1.1 Examples for claim counts

To clarify the issue, we can illustrate the situation with a preliminary comparison of the two approaches, assuming that Y_i represents the number of claims for insured i.

Example 1 *Under the ratio approach, we assume that $Z_i, i = 1, \ldots, n$ follows a Poisson distribution with mean $\zeta_i = \exp(\mathbf{X}_i^\top \boldsymbol{\beta})$. In this framework, it can be shown that the ratio approach corresponds to a Poisson weighted regression and yields the same inference results of the parameter vector $\boldsymbol{\beta}$ as in the offset approach, where $Y_i, i = 1, \ldots, n$ is Poisson distributed with mean $\mu_i = t_i \exp(\mathbf{X}_i^\top \boldsymbol{\beta}) = \exp(\mathbf{X}_i^\top \boldsymbol{\beta} + \log(t_i))$.*

Preuve. The objective function function of the offset and ratio approaches are given as follows :

— ***Offset Approach :***

$$\begin{aligned}
\ell(\boldsymbol{\beta}|y_1, \ldots, y_n) &= \sum_{i=1}^{n} \log\left(\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}\right) \\
&= \sum_{i=1}^{n} \left(-t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) + y_i \left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) + y_i \log(t_i) - \log(y_i!)\right) \\
&\propto \sum_{i=1}^{n} \left(-\exp(\mathbf{X}_i^\top \boldsymbol{\beta}) + z_i \left(\mathbf{X}_i^\top \boldsymbol{\beta}\right)\right) \times t_i.
\end{aligned}$$

— ***Ratio Approach :***

$$\begin{aligned}
\ell(\boldsymbol{\beta}|z_1, \ldots, z_n) &= \sum_{i=1}^{n} t_i \log\left(\frac{\exp(-\zeta_i)\zeta_i^{z_i}}{z_i!}\right) \\
&= \sum_{i=1}^{n} \left(-t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) + t_i z_i \left(\mathbf{X}_i^\top \boldsymbol{\beta}\right) - \log((t_i z_i)!)\right) \\
&\propto \sum_{i=1}^{n} \left(-\exp(\mathbf{X}_i^\top \boldsymbol{\beta}) + z_i \left(\mathbf{X}_i^\top \boldsymbol{\beta}\right)\right) \times t_i.
\end{aligned}$$

For the Poisson distribution, the offset and the ratio approaches are equivalent, meaning there is no need to choose between them. □

Example 2 *In the offset approach, we assume that $Y_i, i = 1, \ldots, n$ follows a zero-inflated Poisson distribution characterized by the parameters ϕ (for the zero-inflation) and $\mu_i = \exp(\mathbf{X}_i^\top \boldsymbol{\beta} + \log(t_i))$ (from the Poisson distribution). In the ratio approach, we assume that $Z_i, i = 1, \ldots, n$ follows a zero-inflated Poisson distribution characterized by the parameters ϕ (for the zero-inflation) and $\zeta_i = \exp(\mathbf{X}_i^\top \boldsymbol{\beta})$ (from the Poisson distribution). It can be shown that these two approaches yield different inference results for the estimation of the parameter vector $\boldsymbol{\beta}$.*

Preuve. For a portfolio of n contracts, we can develop the objective function of the offset approach to see that both approaches are not equivalent :

$$\begin{aligned}\ell(\boldsymbol{\beta}|y_1,\ldots,y_n) &= \sum_{i=1}^{n}\log\left(I(y_i=0)\phi+(1-\phi)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}\right)\\ &\neq \sum_{i=1}^{n}\log\left(I(z_i=0)\phi+(1-\phi)\frac{\exp(-\zeta_i)\zeta_i^{z_i}}{z_i!}\right)\times t_i = \ell(\boldsymbol{\beta}|z_1,\ldots,z_n),\end{aligned}$$

In contrast to the Poisson distribution, a decision must be made between the offset and ratio methods when we want to suppose that the number of claims follows this form of zero-inflated distribution. □

Depending on the distribution used for Y_i, it is not straightforward to determine whether both approaches are equivalent. Even when the underlying distribution is assumed to be the same, the estimated parameters obtained from each approach may differ, potentially leading to different premiums for individual policyholders. Ideally, if the data-generating process for insurance claims were fully understood, one could determine which modelling approach is most appropriate. However, in practice, this is rarely the case, making it difficult to justify one formulation over the other on theoretical grounds alone. Consequently, the choice between approaches often relies on a combination of empirical performance and desirable mathematical properties.

In this context, it is worth noting that Denuit et Trufin (2019) already provide some comparisons between these two approaches. In particular, they show that the offset approach can be interpreted as a weighted regression when the response variable follows a Tweedie distribution. Building on their work, we extend the comparison further by emphasizing the role of stochastic orders and the principle of financial balance, which provide additional insights into the structural differences between the two modeling approaches.

2.1.2 Structure of the paper

This paper emphasizes the differences between the offset and ratio regression approaches in modeling contract loss costs using the Tweedie distribution, rather than focusing on claim frequency. Section 2.2 introduces the notation and model assumptions, detailing key properties of the Tweedie distribution and the parameter inference methods used. It also provides the technical framework required to establish the equivalence results and the asymptotic comparisons developed in the remainder of the paper. Section 2.3

then explains the offset and ratio approaches, as well as the link between them. We show that both approaches can be viewed as weighted regressions, where only the form of the weight differentiates the approaches. Section 2.4 presents an analysis of the consistency of the parameter vector estimators for both approaches, demonstrates the asymptotic efficiency of the offset-approach parameter vector estimator, and examines the gap between observed aggregate loss costs and the sum of estimated premiums based on simulated data. In Section 2.5, we provide an empirical illustration of both approaches using a sample of automobile insurance data from a Canadian insurer. Finally, Section 2.6 concludes the paper.

2.2 GLM with Tweedie-distributed Response Variable

The loss cost of a contract is the total amount of claims, calculated as the aggregate sum of individual claim costs. Formally, let N denote the total number of claims for a contract, and let Z_k represent the cost of the k-th claim, where $k, k = 1, \ldots, N$ if $N > 0$. Thus, assuming Y denotes the loss cost of a contract, as Y is obtained as follows :

$$Y = \begin{cases} \sum_{k=1}^{N} Z_k & \text{if } N > 0 \\ \\ 0 & \text{if } N = 0. \end{cases}$$

Even though individual claim costs $Z_1, \ldots, Z_N$ are continuous random variables, we can demonstrate that the loss cost Y is not a continuous random variable. Specifically, the probability of a null loss cost is non-zero, indicating that the loss cost has a mass at zero and is therefore a semi-continuous random variable. As a result, insurers often use the Tweedie distribution to model a contract premium when the loss cost is the response variable. The rationale for using the Tweedie distribution is its ability to represent a mixture distribution (both continuous and discrete), as noted by Tweedie *et al.* (1984). Additionally, Delong *et al.* (2021) provides a comprehensive summary of this mixture distribution, with relevant applications in actuarial science.

Accordingly, we model the premium using the loss cost as the response variable, assuming that the latter follows a Tweedie distribution. This distribution is viewed as a compound Poisson–Gamma mixture and belongs to the linear exponential family, which justifies the use of the GLM framework (see Tweedie, 1984; Jørgensen, 1987). Before introducing the GLM approach with the Tweedie distribution, we briefly review its

properties by referring to Tweedie *et al.* (1984), Jørgensen (1987), and Delong *et al.* (2021). It is also important to note that several results presented here can also be found in Wüthrich et Merz (2023) and Denuit et Trufin (2019) ; however, this review is essential for a proper understanding of the offset and ratio regression approaches.

2.2.1 Exponential Family Form of Tweedie Distribution

Définition 2.1 *The density function of the Tweedie distribution as a mixture distribution is given by :*

$$f(y|\mu, w, \phi) = \exp\left(\frac{w}{\phi}\left(\frac{\mu^{1-p}}{1-p}y - \frac{\mu^{2-p}}{2-p}\right) + a(y, w, \phi)\right), y \in \mathbf{R}, \tag{2.2.1}$$

where :

- $\mu > 0$ *denotes the mean parameter ;*
- $\phi > 0$ *denotes the dispersion parameter ;*
- $w > 0$ *denotes the weight parameter ;*
- $p \in (1, 2)$ *denotes the variance parameter ;*
- $a(\cdot)$ *denotes a function that does not depend on* μ.

It is important to note that assuming the variance parameter lies within the interval $]1, 2[$ ensures that the Tweedie distribution density is defined as described above. As noted by Tweedie *et al.* (1984) and Jørgensen (1987), the Tweedie distribution can be generalized for any positive variance parameter $p > 0$. However, they also mention that when the variance parameter lies within the interval $]0, 1[$, the Tweedie distribution does not belong to the exponential family. The density function $f(\cdot)$ mentioned above is expressed in the form of the exponential family density function. In the case of the Tweedie distribution, this family is sometimes referred to as Exponential Dispersion Family (EDF) due to the inclusion of the dispersion parameter (see Jørgensen (1997)).

It should also be noted that $f(\cdot)$ lacks a specific form because the function $a(\cdot)$ is unknown. To determine $a(\cdot)$, one can refer to Delong *et al.* (2021), which provides a comprehensive overview of the Tweedie distribution. In this paper, we do not focus on determining $a(\cdot)$ since it does not affect the regression approaches discussed

here. Instead, we concentrate on the canonical parameter and the canonical link, which are relevant to insurance pricing. These two quantities for the Tweedie distribution are defined as follows :

— θ denotes the canonical parameter and is identified by : $\theta = \frac{\mu^{1-p}}{1-p}$;
— The canonical link function is defined by : $s > 0 \mapsto \frac{s^{1-p}}{1-p}$.

It is important to note that the canonical link connects the canonical parameter to the mean parameter of distributions within the exponential family. In car insurance pricing, the canonical link function is often employed within the framework of GLM to ensure the balance property, as described by Nelder et Wedderburn (1972). The balance property ensures that the sum of the estimated premiums equals the sum of the observed loss costs. The rationale behind this equality is that pricing models must replicate the observed risk at each portfolio level as accurately as possible. Consequently, the primary objective for insurers is to develop models that closely reproduce the observed risk across all portfolio levels.

However, when the response variable follows a Tweedie distribution, interpreting the premium through the canonical link becomes challenging. Therefore, as discussed in the following section, choosing the log link to model the premium instead of the canonical link offers a more straightforward interpretation of the premium. It is important to note, however, that using the log link in premium modelling introduces a gap between the sum of the observed loss costs and the sum of the estimated premiums. This paper will analyze this gap in detail. For the remainder of this paper, we denote the parameters of the Tweedie distribution by (μ, w, ϕ, p) and we define $\mathcal{T}$ as the set of Tweedie distributions (μ, w, ϕ, p) where the variance parameter p lies within the interval $(1, 2)$.

2.2.2 Notations

We consider n independent insurance contracts. For each contract i (where $i = 1, \ldots, n$), the following variables are defined :

— Y_i : A random variable representing the loss cost for contract i.
— y_i : The observed loss cost for contract i.
— t_i The risk exposure for contract i.

Additionally, we consider q risk factors, denoted by $\boldsymbol{x_1}, \ldots, \boldsymbol{x_q}$. For each contract i, the following are also defined :

— $x_{i,j}$: The observed characteristic for contract i corresponding to risk factor $\boldsymbol{x_j}$, where $j = 1, \ldots, q$.

— $\mathbf{X}_i$: The vector of risk characteristics associated with contract i, defined as $\mathbf{X}_i = (x_{i,0}, x_{i,1}, \ldots, x_{i,q})^\top$, where $x_{i,0} = 1, \forall i = 1, \ldots, n$.

Finally, the matrix $\boldsymbol{X}$, representing the risk factors, is given by :

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,q} \\ 1 & x_{2,1} & \ldots & x_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,q} \end{pmatrix}$$

This matrix $\boldsymbol{X}$ captures all the observed risk characteristics across the contracts.

2.2.3 Assumptions and Properties

The main assumptions of GLM are as follows :

1. The distribution of $Y_i|\mathbf{X}_i$ belongs to the exponential linear family for all contract i where $i = 1, \ldots, n$.
2. $Y_{i_1}|\mathbf{X}_{i_1} \perp\!\!\!\perp Y_{i_2}|\mathbf{X}_{i_2}, \forall i_1, i_2 = 1, \ldots, n$ such that $i_1 \neq i_2$.
3. $\boldsymbol{X}$ has full rank $q + 1$.

The first assumption in the case of the Tweedie distribution is equivalent to stating that $Y_i|\mathbf{X}_i$ follows a Tweedie distribution with parameters (μ_i, w_i, ϕ, p). The second assumption implies the independence of the loss costs between contracts, given the risk characteristics. It is important to note that these first two assumptions are fundamental for constructing the likelihood function. The third assumption is essential for calculating the variance of the estimator of μ_i.

The concept of GLM revolves around statistical inference on μ_i, which is considered the premium for contract i in insurance pricing. This consideration is justified by the fact that the mean parameter μ_i, regardless of the distribution of $Y_i|\mathbf{X}_i$, corresponds to the constant closest to $Y_i|\mathbf{X}_i$ in terms of Euclidean distance (for further details, see Denuit *et al.* (2007b)).

Thus, for modelling the premium of contract i, we assume that the logarithm of this premium (μ_i) linearly depends on the characteristics vector $\mathbf{X}_i$ as follows :

$$\mu_i = \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right) = \exp\left(\beta_0\right) \prod_{j=1}^{q} \exp\left(\beta_j x_{i,j}\right) = b_0 \times \prod_{j=1}^{q} r_j, \tag{2.2.2}$$

where :

- $\boldsymbol{\beta}$: The parameter vector, defined as $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^\top$ where $\beta_j \in \mathbf{R}$ for $j = 0, 1, \ldots, q$;
- b_0 : The basic premium, defined as $b_0 = \exp\left(\beta_0\right)$.
- r_j : The relativity factors associated with each characteristic, defined as $r_j = \exp\left(\beta_j x_{i,j}\right)$.

The advantage of using the logarithm function as the link function is that the premium μ_i for contract i can be expressed as the product of the basic premium b_0 and relativity factors $r_j, j = 1, \ldots, q$. While the basic premium b_0 is consistent across all contracts, the relativity factors depend on the specific risk characteristics of each contract. Consequently, statistical inference on μ_i is equivalent to estimating the parameter vector $\boldsymbol{\beta}$. To conclude, it is pertinent to introduce an interesting property of the Tweedie distribution, as discussed in Denuit et Trufin (2019), which will be valuable for further developments.

Property 2.2.3.1 (Tweedie Invariance Property) *If a random variable Z is Tweedie-distributed with parameters (μ, w, ϕ, p) and belongs to the set $\mathcal{T}$, then tZ, for any positive t, is also Tweedie-distributed with parameters $(t\mu, \frac{w}{t^{2-p}}, \phi, p)$ and belongs to the same set $\mathcal{T}$. Conversely, if tZ, for any positive t, is Tweedie distributed with parameters (μ, w, ϕ, p) and belongs to the set $\mathcal{T}$, then Z is also Tweedie-distributed with parameters$(\frac{\mu}{t}, wt^{2-p}, \phi, p)$ and belongs to the same set $\mathcal{T}$. Formally, we have :*

$$Z \in \mathcal{T} \Longleftrightarrow tZ \in \mathcal{T}, \forall t > 0.$$

2.2.4 Log-likelihood Function

The GLM employs the maximum likelihood approach to estimate the parameter vector $\boldsymbol{\beta}$ (Nelder et Wedderburn, 1972). The maximum likelihood approach maximizes the likelihood function, which depends on the

parameter vector $\boldsymbol{\beta}$ given the observations $y_1, \ldots, y_n$. For a detailed discussion on the theory of maximum likelihood, see Casella et Berger (2024). In the case of the Tweedie distribution, by utilizing the conditional independence between contracts and Definition 2.1, the likelihood function is expressed as follows :

$$L(\beta_0, \ldots, \beta_q) = \prod_{i=1}^{n} \exp\left(\frac{w_i}{\phi}\left(\frac{\mu^{1-p}}{1-p}_i y_i - \frac{\mu^{2-p}}{2-p}_i\right) + a(y_i, w_i, \phi)\right).$$

In practice, the maximum likelihood approach often uses the logarithm of the likelihood function as the objective function instead of the likelihood function itself. This is because a function and all strictly increasing transformations of that function reach their maximum at the same point (See Boyd et Vandenberghe (2004) for the theory on function optimization). Therefore, in the Tweedie case, the logarithm of the likelihood function (commonly referred to as the log-likelihood function) is given by (2.2.3).

$$\ell(\beta_0, \ldots, \beta_q) = \sum_{i=1}^{n}\left(\frac{w_i}{\phi}\left(\frac{\mu_i^{1-p}}{1-p}y_i - \frac{\mu_i^{2-p}}{2-p}\right) + a\left(y_i, w_i, \phi, p\right)\right) \tag{2.2.3}$$

$$\propto \frac{1}{\phi}\sum_{i=1}^{n} w_i\left(\frac{\mu_i^{1-p}}{1-p}y_i - \frac{\mu_i^{2-p}}{2-p}\right). \tag{2.2.4}$$

Additionally, considering equation (2.2.4), we note that maximizing the log-likelihood in the Tweedie case is equivalent to performing a weighted regression, where each contract i is weighted by w_i. The log-likelihood maximization can also be interpreted as finding the zeros of the gradient function associated with the model. The gradient function is the vector of the first partial derivatives of the log-likelihood function ($\frac{\partial \ell(\beta_0,\ldots,\beta_q)}{\partial \beta_j}, j = 0, \ldots, q$). For the Tweedie case, the gradient function is derived as follows :

$$\begin{aligned}
\nabla(\boldsymbol{\beta}) &= \left(\frac{\partial \ell(\beta_0, \ldots, \beta_q)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_0}, \ldots, \frac{\partial \ell(\beta_0, \ldots, \beta_q)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right)^T \\
&= \frac{1}{\phi}\left(\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\mu_i^{p-1}}, \ldots, \sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\mu_i^{p-1}} x_{i,q}\right)^\top \\
&= \frac{1}{\phi}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{R};
\end{aligned}$$

The vector $\boldsymbol{R}$ and the matrix $\boldsymbol{D}$ are defined as follows :

$$\boldsymbol{R} = \left(\frac{y_i}{\mu_i} - 1\right)_{i=1,\dots,n} \quad \boldsymbol{D} = \operatorname{diag}\left(w_i \mu_i^{2-p}\right)_{i=1,\dots,n}. \tag{2.2.5}$$

2.2.4.1 Iterative Algorithm

Generally, GLM theory adopts the Iteratively Reweighted Least Squares (IRLS) algorithm for maximizing the log-likelihood, which is a variant of the Newton-Raphson algorithm (For a detailed discussion on the Newton-Raphson algorithm, refer to Boyd et Vandenberghe (2004)). It is important to note that the IRLS algorithm is based on the Fisher Information Matrix, the theory of which can be found in Casella et Berger (2024). In this section, we review the Fisher Information Matrix in the context of the Tweedie distribution and justify its invertibility. We also recall the definition of positive definite matrices and their properties, as discussed by Horn et Johnson (2012).

Let $\boldsymbol{\beta}_{(k)}$ denote the estimate of $\boldsymbol{\beta}$ at iteration k, The IRLS algorithm is defined as follows :

$$\boldsymbol{\beta}_{(k+1)} = \boldsymbol{\beta}_{(k)} + \left(I_n\left(\boldsymbol{\beta}_{(k)}\right)\right)^{-1} \nabla\left(\boldsymbol{\beta}_{(k)}\right); \tag{2.2.6}$$

$I_n(\cdot)$ is the Fisher information matrix and is defined as follows :

$$\begin{aligned} I_n(\boldsymbol{\beta}) &= -\mathrm{E}[\mathsf{Hess}\,(\boldsymbol{\beta})] \\ &= \frac{1}{\phi}\left(\sum_{i=1}^{n} w_i \mu_i^{2-p} x_{i,j_1} x_{i,j_2}\right)_{j_1,j_2=0,\dots,q} \\ &= \frac{1}{\phi}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}, \end{aligned}$$

where $\mathsf{Hess}\,(\boldsymbol{\beta})$ is the Hessian matrix. Referring to equation (2.2.6), we note that the IRLS algorithm relies on the invertibility of the Fisher Information Matrix. The inverse of the Fisher Information Matrix is also used to calculate the asymptotic covariance matrix of the parameter vector estimator. As demonstrated in this paper, comparing the covariance matrices of the parameter vector estimators is crucial for comparing

premium estimators in the offset and ratio approaches. Therefore, analyzing the invertibility of the Fisher Information Matrix is of prime importance.

We now review the properties of positive definite matrices and introduce the following notations :

— $\mathcal{M}_{n,m}$ denotes the set of matrices with n rows and m columns.
— $\mathcal{P}_m$ denotes the set of positive definite matrices.

Définition 2.2 *A matrix $M \in \mathcal{M}_{m,m}$ is positive definite ($M \succ 0$), if the following condition holds :*

$$x^\top M x > 0; \forall x \in \mathbb{R}^m.$$

Property 2.2.4.1 *With $\mathcal{M}_{n,m}$ and $\mathcal{P}_m$ defined above, the following relationships can be stated :*

$$A \in \mathcal{P}_m \iff A^{-1} \in \mathcal{P}_m; \tag{2.2.7}$$

$$\forall A, B \in \mathcal{P}_m : A - B \succ 0 \iff B^{-1} - A^{-1} \succ 0; \tag{2.2.8}$$

$$\forall A \in \mathcal{P}_m, \forall X \in \mathcal{M}_{m,q} : X^\top A X \succ 0 \iff X \text{ has rank } q. \tag{2.2.9}$$

Property (2.2.7) serves as a necessary and sufficient condition for proving the invertibility of a matrix. Thus, to prove the existence of the inverse of the Fisher Information Matrix, we must show that this matrix is positive definite. Using the GLM theory assumption regarding the rank of the matrix $\boldsymbol{X}$ and the necessary and sufficient condition given in (2.2.9), we conclude that the Fisher Information Matrix is indeed positive definite. The necessary and sufficient condition given in (2.2.8) is also helpful when comparing premium estimators in the offset and ratio approaches.

2.2.4.2 Remarks

The variance and weight parameters exclusively influence the estimates of the parameter vector in the IRLS algorithm. However, as stated in equation (2.2.10), the dispersion parameter does not affect these

estimates, and therefore, it is not a focus of this paper. Finally, if the weight of each contract is known and the variance parameter is also known, the GLM in the Tweedie case involves using the IRLS algorithm to obtain an estimate of the parameters vector.

$$\left(I_n(\boldsymbol{\beta})\right)^{-1} \nabla\left(\boldsymbol{\beta}\right) = \left(X^\top DX\right)^{-1} X^\top DR. \tag{2.2.10}$$

It is also noteworthy that the IRLS algorithm is implemented in the R software through the *glm* function.

To end these remarks, it should be noted that the GLM theory based on the maximization of the likelihood function assumes that the model is well specified. In the Tweedie case, a well-specified model means that the true conditional distribution of the loss cost also follows a Tweedie distribution with the same specification as that assumed by the model. When the model is misspecified, however, the objective function used to estimate $\boldsymbol{\beta}$ remains unchanged. In this case, the resulting estimator is referred to as the quasi-maximum likelihood estimator in the sense of White (1982).

2.3 The Offset and the Ratio Approaches

To present the offset and ratio regression approaches within the context of a Tweedie distribution and to examine their similarities and differences, we continue to consider n independent contracts, along with the same notations and assumptions as in the previous section. The primary distinction from the previous section is that, in these two regression approaches, the premium of contract i is adjusted to account for its risk exposure t_i. We first present the estimation procedure for the parameter vector under the assumption that the corresponding model is well specified.

2.3.1 Offset Approach

As mentioned in Section 2.1.1, the statistical inference on the parameter vector $\boldsymbol{\beta}$ in offset approach is performed with the loss costs $Y_1, \ldots, Y_n$. To account for the logarithm of a contract's risk exposure as the offset variable, we assume that $Y_i|\mathbf{X}_i$ follows the Tweedie distribution with parameters $(\mu_i, 1, \phi, p)$ where μ_i is defined as $\mu_i = t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right)$. We also denote an estimate of μ_i in the offset regression by $\widehat{\mu}_i^O$, obtained as follows :

$$\widehat{\mu}_i^O = \exp\left(\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}^O + \log(t_i)\right), \tag{2.3.1}$$

where $\widehat{\boldsymbol{\beta}}^O$, an estimate of $\boldsymbol{\beta}$ in offset approach, is obtained using the IRLS algorithm with the following quantities :

— **Log-likelihood Function :**

$$\ell^O(\beta_0, \ldots, \beta_q) = \sum_{i=1}^{n} \left(\frac{1}{\phi} \left(\frac{\mu_i^{1-p}}{1-p} y_i - \frac{\mu_i^{2-p}}{2-p} \right) + a^O(y_i, \phi, p) \right). \tag{2.3.2}$$

— **Gradient Function :**

$$\nabla^O(\boldsymbol{\beta}) = \frac{1}{\phi} \left(\sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i^{p-1}}, \ldots, \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i^{p-1}} x_{i,q} \right)^\top \tag{2.3.3}$$

$$= \frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{D}^O \boldsymbol{R}^O; \tag{2.3.4}$$

The vector $\boldsymbol{R}^O$ and the matrix $\boldsymbol{D}^O$ are defined as follows :

$$\boldsymbol{D}^O = \operatorname{diag}\left(t_i^{2-p} \exp\left((2-p) \mathbf{X}_i^\top \boldsymbol{\beta} \right) \right)_{i=1,\ldots,n}, \boldsymbol{R}^O = \left(\frac{y_i}{\mu_i} - 1 \right)_{i=1,\ldots,n}.$$

2.3.2 Ratio Approach

We saw in Section 2.1.1 that the ratio-based method can be interpreted as a form of weighted regression. In this framework, the response variable is defined as the ratio of the contract's loss cost to its risk exposure, with each contract being assigned a weight proportional to its respective risk exposure. We denote the ratio of contract i's loss cost Y_i and its risk exposure t_i by Z_i, defined as $Z_i = \frac{Y_i}{t_i}$, for $i = 1, \ldots, n$. To account for the risk exposure t_i as the weight of contract i, we also assume $Z_i | \mathbf{X}_i$ follows a Tweedie distribution with parameters (ζ_i, t_i, ϕ, p), where the mean parameter ζ_i is assumed to be :

$$\zeta_i = \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), \tag{2.3.5}$$

and $\widehat{\zeta_i}$ is an estimate of ζ_i, equal to $\widehat{\zeta_i} = \exp\left(\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}^R\right)$. The vector $\widehat{\boldsymbol{\beta}}^R$, an estimate of $\boldsymbol{\beta}$ in ratio approach, is obtained using the IRLS algorithm with the following quantities :

— **Log-likelihood Function :** This function is constructed with the observations $z_1, \ldots, z_n$ where $\zeta_i = \frac{y_i}{t_i}, i = 1, \ldots, n$, as follows :

$$\ell^R(\beta_0, \ldots, \beta_q) = \sum_{i=1}^{n} \left\{ \frac{t_i}{\phi} \left(\frac{\zeta_i^{1-p}}{1-p} z_i - \frac{\zeta_i^{2-p}}{2-p} \right) + a^R(z_i, \phi, p) \right\}. \tag{2.3.6}$$

— **Gradient Function :**

$$\nabla^R(\boldsymbol{\beta}) = \frac{1}{\phi} \left(\sum_{i=1}^{n} t_i \frac{z_i - \zeta_i}{\zeta_i^{p-1}}, \ldots, \sum_{i=1}^{n} t_i \frac{z_i - \zeta_i}{\zeta_i^{p-1}} x_{i,q} \right)^\top \tag{2.3.7}$$

$$= \frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{D}^R \boldsymbol{R}^R; \tag{2.3.8}$$

The vector $\boldsymbol{R}^R$ and the matrix $\boldsymbol{D}^R$ are defined as follows :

$$\boldsymbol{D}^R = \operatorname{diag}\left(t_i \exp\left\{ (2-p) \mathbf{X}_i^\top \boldsymbol{\beta} \right\} \right)_{i=1,\ldots,n}, \boldsymbol{R}^R = \left(\frac{z_i}{\zeta_i} - 1 \right) = \left(\frac{y_i}{\mu_i} - 1 \right)_{i=1,\ldots,n},$$

where $\mu_i = t_i \zeta_i = t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} \right)$.

2.3.3 Link between Offset and Ratio Approaches

Building on the Tweedie Invariance Property defined in Proposition 2.2.3.1, the equivalence of the two approaches can be demonstrated as follows :

1. **The offset approach viewed in the form of the ratio approach**

   In the offset approach, $Y_i | \mathbf{X}_i^\top$ follows a Tweedie distribution with parameters $\left(t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} \right), 1, \phi, p \right)$. We can examine the ratio of the loss cost to the risk exposure t_i, denoted as $Z_i = \frac{Y_i}{t_i}$, and apply the Tweedie Invariance Property. It can then be demonstrated that $Z_i | \mathbf{X}_i^\top$ follows a Tweedie distribution with parameters $\left(\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} \right), t_i^{2-p}, \phi, p \right)$. More formally, we have the following equivalence :

$$Y | \mathbf{X}_i^\top \sim \text{Tweedie}\left(t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} \right), 1, \phi, p \right) \equiv Z | \mathbf{X}_i^\top \sim \text{Tweedie}\left(\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta} \right), t_i^{2-p}, \phi, p \right).$$

Using equation (2.2.5) for the calculation of the matrix $\boldsymbol{D}$, we can demonstrate that the two Tweedie distributions introduced earlier share the same expression :

$$\boldsymbol{D} = \operatorname{diag}\left(t_i^{2-p} \exp\left((2-p)\mathbf{X}_i^\top \boldsymbol{\beta}\right)\right)_{i=1,\ldots,n} \equiv \boldsymbol{D}^O$$

This result provides an additional verification that these two regression approaches are equivalent, leading to identical inference outcomes for the parameter vector $\boldsymbol{\beta}$.

2. **The ratio approach viewed in the form of the offset approach**

Similarly, in the ratio approach, where $Z_i|\mathbf{X}_i^\top$ is assumed to follow a Tweedie distribution with parameters $(\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), t_i, \phi, p)$, we can apply the Tweedie Invariance Property to demonstrate that $Y_i|\mathbf{X}_i^\top$ follows a Tweedie distribution with parameters $(t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), t_i^{p-1}, \phi, p)$. This equivalence is expressed as :

$$Z|\mathbf{X}_i^\top \sim \text{Tweedie}(\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), t_i, \phi, p) \equiv Y|\mathbf{X}_i^\top \sim \text{Tweedie}(t_i \exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), t_i^{p-1}, \phi, p).$$

For each form of the Tweedie distributions, we also have the following equivalence for the matrix $\boldsymbol{D}$:

$$\boldsymbol{D} = \operatorname{diag}\left(t_i \exp\left((2-p)\mathbf{X}_i^\top \boldsymbol{\beta}\right)\right)_{i=1,\ldots,n} \equiv \boldsymbol{D}^R.$$

The previous results facilitate a direct comparison of the two approaches. By focusing on the random variable Z, the choice is between the following two distributions :

$$\text{Tweedie}(\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), t_i^{2-p}, \phi, p) \quad \textit{versus} \quad \text{Tweedie}(\exp\left(\mathbf{X}_i^\top \boldsymbol{\beta}\right), t_i, \phi, p)$$

The choice between the offset and ratio formulations corresponds to different specifications of the Tweedie dispersion structure. While both formulations lead to identical estimating equations for the parameter vector

$\boldsymbol{\beta}$, they are not probabilistically equivalent models. The observed equivalence concerns the estimation of $\boldsymbol{\beta}$, not the full distributional assumptions.

2.3.4 Weights Analysis

Since the distinction between the two approaches primarily hinges on the choice of the weight parameter w_i, it is crucial to analyze its impact on modelling loss costs with the Tweedie distribution in greater detail. An initial evident result is the dominance of the weight in the ratio approach compared to that in the offset approach, given by :

$$w_i^O = t_i^{2-p} \geq t_i = w_i^R, \quad \forall p \in]1, 2[,\ t \in]0, 1].$$

In Figure 2.1, we analyze in more detail the differences between the weights used in the offset and ratio approaches. The diagonal dashed line represents the weights in the ratio approach $(w_i^R = t_i)$, while the other coloured lines illustrate the weights t_i^{2-p} employed in the offset approach for various values of the variance parameter p.

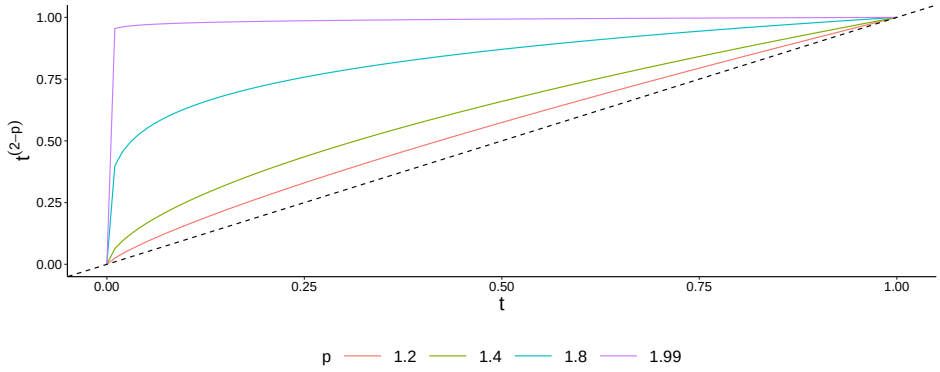


Figure 2.1 – Comparison of the Weight Parameters in the Ratio and Offset Approaches for Different Risk Exposures (t)

For a given value of p, we observe that the difference in weights between the offset and ratio approaches diminishes as the risk exposure increases. This indicates that when the contract exposure period is short, the disparity between the weights in both approaches is significant compared to a contract with an exposure

period close to one year. In the case where the exposure period equals one year $(t = 1)$, the weights in each approach equal 1, which means that these two regression approaches are equivalent. This equivalence between the methods is also evident in the equality between $\boldsymbol{D}^R$ and $\boldsymbol{D}^O$ when $t = 1$:

$$\boldsymbol{D}^R = \boldsymbol{D}^O = \operatorname{diag}\left(\exp\left((2-p)\mathbf{X}_i^\top \boldsymbol{\beta}\right)\right)_{i=1,\ldots,n}.$$

We can also interpret Figure 2.1 by varying the variance parameter p. As p decreases and approaches 1, the difference between the weights in the offset and ratio approaches becomes negligible. Since the Tweedie distribution converges to the Poisson distribution as $p \to 1$, it is noteworthy that this observation aligns with the equivalence of the methods, corresponding to the result obtained regarding claim numbers in Example 1 presented at the beginning of the paper. Conversely, as p approaches 2, the disparity in weights between the two approaches becomes more pronounced, indicating that the choice of approach has a more significant impact. Notably, when $p \to 2$, the Tweedie distribution converges to a gamma distribution, which has historically been employed to model the severity of claims in actuarial science (see Denuit *et al*. (2007b)). In the context of severity modelling, it is well known that the risk exposure t_i should not be utilized, implying that the weight should remain constant regardless of the value of the risk exposure t_i. The ratio approach with $p = 1.999$ tends to demonstrate this characteristic. Thus, from the perspective of the weighting structure, these results tend to favour the ratio approach over the range $p \in]1, 2[$.

2.4 Comparison of the Approaches

According to the analysis presented in Section 2.3.3, for a fixed value of the variance parameter p, the difference between the weights w_i^O and w_i^R associated with the offset and ratio approaches, respectively, depends on the risk exposure t_i. This difference is substantial for contracts with short exposure periods, while it becomes negligible for contracts with exposure periods close to one year. However, a limitation of this analysis is that it does not, by itself, allow us to determine which approach is more appropriate for insurance pricing.

To address this issue, we proceed in two steps. First, we compare the parameter vector estimators derived under the two approaches, relying on the quasi-maximum likelihood framework of White (1982). Second, using a numerical illustration, we highlight the practical merits of the ratio approach relative to the offset

approach in the context of automobile insurance pricing. To this end, we consider as the response variable the loss cost Y per unit of risk exposure t, namely $Z = \frac{Y}{t}$, and assume that the true data-generating process of Z is characterized by a density function g.

For all $\boldsymbol{x} \in \mathbb{R}^{q+1}$ and $t \in (0, 1]$, representing the covariate vector (risk characteristics) and the risk exposure, respectively, the true expected value of Z is given by

$$\zeta^{\text{True}} = \int_0^{+\infty} z\, g\big(z \mid \boldsymbol{\beta}^{\text{True}}, t\big)\, dz = \exp\Big\{\boldsymbol{x}^\top \boldsymbol{\beta}^{\text{True}}\Big\} < +\infty,$$

where $\boldsymbol{\beta}^{\text{True}} \in \mathbb{R}^{q+1}$ denotes the true parameter vector.To clarify our strategy, we also recall the definition of estimator consistency, following Casella et Berger (2024). An estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^{\text{True}}$ is said to be consistent (or convergent in probability) if

$$\widehat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^{\text{True}} \quad \text{as } n \to +\infty,$$

that is, if the following condition holds :

$$\lim_{n\to\infty} \Pr\Big(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\text{True}}\| > \epsilon\Big) = 0, \quad \text{for all } \epsilon > 0.$$

2.4.1 Parameter Vector Estimators Comparison

Recall that the estimation of the true parameter vector $\boldsymbol{\beta}^{\text{True}}$ under the two approaches is carried out by assuming a Tweedie distribution with different weights in each case : the offset approach uses $w = t^{2-p}$, whereas the ratio approach uses $w = t$. We denote the corresponding density by f, which is defined as

$$f(z \mid \boldsymbol{\beta}, t) = \exp\left(\frac{w}{\phi}\left(\frac{\exp\{(1-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{1-p}\, z - \frac{\exp\{(2-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{2-p}\right) + a(z, w, \phi)\right), \qquad z \geq 0, \qquad \boldsymbol{\beta} \in \mathbb{R}^{q+1}.$$

We then define the Kullback–Leibler (KL) divergence between the true density $g(\cdot)$ and the density $f(\cdot)$ as

$$
\begin{aligned}
\mathsf{KL}(\boldsymbol{\beta}) &= \mathbb{E}\left[\log \frac{g(z \mid \boldsymbol{\beta}^{\mathsf{True}}, t)}{f(z \mid \boldsymbol{\beta}, t)}\right] \\
&= \int_0^{+\infty} \log\big(g(z \mid \boldsymbol{\beta}^{\mathsf{True}}, t)\big)\, g(z \mid \boldsymbol{\beta}^{\mathsf{True}}, t)\, dz \\
&\quad - \int_0^{+\infty} \log\big(f(z \mid \boldsymbol{\beta}, t)\big)\, g(z \mid \boldsymbol{\beta}^{\mathsf{True}}, t)\, dz.
\end{aligned}
$$

Proposition 2.3 *According to White (1982), if the Kullback–Leibler divergence between the true density g and the density f assumed under either the offset or the ratio approach exists, and if the true conditional mean ζ^{True} satisfies the mean specification assumed by both approaches, then the corresponding estimator of the true parameter vector is consistent under each approach. More formally, if*

$$
\mathsf{KL}(\boldsymbol{\beta}) < +\infty \quad \textit{for all } \boldsymbol{\beta}, \quad \textit{and} \quad \zeta^{\mathsf{True}} = \exp\Big\{\boldsymbol{x}^\top \boldsymbol{\beta}^{\mathsf{True}}\Big\},
$$

then

$$
\widehat{\boldsymbol{\beta}}^{O} \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^{\mathsf{True}}, \qquad \widehat{\boldsymbol{\beta}}^{R} \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^{\mathsf{True}} \quad \textit{as } n \to +\infty,
$$

where $\widehat{\boldsymbol{\beta}}^{O}$ and $\widehat{\boldsymbol{\beta}}^{R}$ denote the estimators of $\boldsymbol{\beta}^{\mathsf{True}}$ obtained from the offset and ratio approaches, respectively.

Preuve.

The function $\mathsf{KL}(\cdot)$ is always nonnegative and finite whenever the probability measures induced by $g(\cdot \mid \boldsymbol{\beta}^{\mathsf{True}}, t)$ and $f(\cdot \mid \boldsymbol{\beta}, t)$ are mutually absolutely continuous (see Gray (2011)). Since both $g(\cdot \mid \boldsymbol{\beta}^{\mathsf{True}}, t)$ and $f(\cdot \mid \boldsymbol{\beta}, t)$ are densities, this condition is equivalent to requiring that their associated probability measures be mutually absolutely continuous, which follows from the Radon–Nikodym theorem.

Suppose that the minimum of $\mathsf{KL}(\boldsymbol{\beta})$ exists and denote it by $\boldsymbol{\beta}^*$. According to White (1982), the estimators obtained under the working density f converge in probability to this minimizer $\boldsymbol{\beta}^*$. Therefore, it suffices to

show that this minimizer coincides with the true parameter vector, that is,

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^{\textit{True}}.$$

We start from the following equivalence :

$$\arg\min_{\boldsymbol{\beta}} \textit{KL}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \int_0^{+\infty} \log\big(f(z \mid \boldsymbol{\beta}, t)\big)\, g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz.$$

For notational convenience, we write

$$\int_0^{+\infty} \log\big(f(z \mid \boldsymbol{\beta}, t)\big)\, g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz \;=\; \int_0^{+\infty} \log(f)\, g\, dz.$$

Using the expression of the Tweedie density, we obtain

$$\begin{aligned}\int_0^{+\infty} \log(f)\, g\, dz &= \int_0^{+\infty} \left[\frac{w}{\phi}\left(\frac{\exp\{(1-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{1-p}\, z - \frac{\exp\{(2-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{2-p}\right) + a(z,w,\phi)\right] g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz \\ &\propto \int_0^{+\infty} \left(\frac{\exp\{(1-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{1-p}\, z - \frac{\exp\{(2-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{2-p}\right) g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz,\end{aligned}$$

where the proportionality follows from the fact that $a(z, w, \phi)$ does not depend on $\boldsymbol{\beta}$.

By linearity of the integral, this expression becomes

$$\frac{\exp\{(1-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{1-p} \int_0^{+\infty} z\, g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz - \frac{\exp\{(2-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{2-p} \int_0^{+\infty} g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz.$$

Since

$$\int_0^{+\infty} z\, g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz = \zeta^{\textit{True}} = \exp\left\{\boldsymbol{x}^\top\boldsymbol{\beta}^{\textit{True}}\right\}, \qquad \int_0^{+\infty} g(z \mid \boldsymbol{\beta}^{\textit{True}}, t)\, dz = 1,$$

the objective function can be written as

$$K(\boldsymbol{\beta}) = \frac{\exp\{(1-p)\boldsymbol{x}^\top\boldsymbol{\beta} + \boldsymbol{x}^\top\boldsymbol{\beta}^{\textit{True}}\}}{1-p} - \frac{\exp\{(2-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}}{2-p}.$$

The gradient of $K(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$\nabla_{\boldsymbol{\beta}} K(\boldsymbol{\beta}) = \Big(\exp\{(1-p)\boldsymbol{x}^\top\boldsymbol{\beta} + \boldsymbol{x}^\top\boldsymbol{\beta}^{\textit{True}}\} - \exp\{(2-p)\boldsymbol{x}^\top\boldsymbol{\beta}\}\Big)\boldsymbol{x}.$$

At the minimizer $\boldsymbol{\beta}^*$, the first-order condition yields

$$\nabla_{\boldsymbol{\beta}} K(\boldsymbol{\beta}^*) = \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{\beta}^* = \boldsymbol{\beta}^{\textit{True}},$$

which completes the proof. □

2.4.2 Asymptotic Variance Approximation

Another important result established by White (1982) is that, under both the offset and ratio approaches, the estimators of the true parameter vector $\boldsymbol{\beta}^{\text{True}}$ are asymptotically Gaussian with a common mean equal to $\boldsymbol{\beta}^{\text{True}}$. Once consistency is established, differences between the two approaches therefore arise solely through their asymptotic dispersion rather than through their limiting value.

From a practical standpoint, differences in asymptotic dispersion translate into differences in the precision and stability of the estimated covariate effects. A smaller asymptotic covariance matrix implies tighter confidence intervals for the regression coefficients, which facilitates the comparison, selection, and validation of pricing variables, particularly when some effects are modest but economically relevant. At the same time, reduced dispersion limits the sensitivity of the estimated parameters to random portfolio fluctuations.

In this context, the comparison of asymptotic covariance matrices provides a natural and meaningful criterion for assessing the relative efficiency of the two estimation strategies. Using the asymptotic expansion derived in Appendix B.1, we may therefore state the following proposition.

Proposition 2.4 *Consider the leading Gaussian terms of the Edgeworth expansion of the asymptotic covariance matrices of $\widehat{\boldsymbol{\beta}}^O$ and $\widehat{\boldsymbol{\beta}}^R$ associated with the offset and ratio approaches. Assume that these leading terms are given by*

$$\Sigma^O = \phi(\boldsymbol{X}^\top \boldsymbol{D}^O \boldsymbol{X})^{-1}, \tag{2.4.1}$$

$$\Sigma^R = \phi(\boldsymbol{X}^\top \boldsymbol{D}^R \boldsymbol{X})^{-1}. \tag{2.4.2}$$

Under the assumptions of Proposition 2.3 and for risk exposure satisfying $t \leq 1$, these leading Gaussian terms satisfy

$$\Sigma^R - \Sigma^O \succ 0.$$

Preuve. First, consider n observations of the loss cost per unit of exposure $Z_1, \ldots, Z_n$, with $Z_i \sim \text{Tweedie}(\zeta_i, w_i, \phi, p)$, $w_i \in \{t_i, t_i^{2-p}\}$, and $\zeta_i = \exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}$. According to White (1982), the asymptotic covariance matrix of the estimator is given by $\Sigma = A^{-1}BA^{-1}$, where $A = -I_n(\boldsymbol{\beta})$ and $B = \mathbb{E}[\nabla(\boldsymbol{\beta})\nabla(\boldsymbol{\beta})^\top]$, with

$$\nabla(\boldsymbol{\beta}) = \frac{1}{\phi}\Big(\sum_{i=1}^{n} w_i \frac{Z_i - \zeta_i}{\zeta_i^{p-1}} x_{i,0}, \ldots, \sum_{i=1}^{n} w_i \frac{Z_i - \zeta_i}{\zeta_i^{p-1}} x_{i,q}\Big)^\top.$$

As shown in Appendix B.1, $B = \phi^{-1}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X} = -A$*, yielding*

$$\Sigma = A^{-1}BA^{-1} = -A^{-1} = \phi(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1}.$$

We now compare the two covariance matrices. As noted in Section 2.2.4.1, $\Sigma^O, \Sigma^R \in \mathcal{P}_{q+1}$*. Define* $M = \Sigma^R - \Sigma^O$*. To prove that* M *is positive definite, it suffices, by the results of Equation (2.2.8), to show that* $\boldsymbol{D}^O - \boldsymbol{D}^R$ *is positive definite. From Figure 2.1, we have*

$$\forall v \in \mathbb{R}^n, \quad v^\top(\boldsymbol{D}^O - \boldsymbol{D}^R)v = \sum_{i=1}^{n}(t_i^{2-p} - t_i)\zeta_i^{2-p}v_i^2 > 0.$$

Moreover, since $\boldsymbol{X}$ *has full rank* $q+1$*, equation (2.2.9) implies*

$$\boldsymbol{X}^\top(\boldsymbol{D}^O - \boldsymbol{D}^R)\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{D}^O\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{D}^R\boldsymbol{X} \succ 0,$$

It then follows from Equation (2.2.8) *that* $M = \Sigma^R - \Sigma^O$ *is positive definite.*

□

Taken together, the previous propositions show that each approach can be implemented independently and yields a consistent (i.e., convergent) estimator of the true parameter vector $\boldsymbol{\beta}^{\text{True}}$. However, based solely on the result stated in Proposition 2.4, one might be led to conclude that the estimator obtained under the offset approach is asymptotically more efficient than the one obtained under the ratio approach.

It is important to emphasize that Proposition 2.4 is formulated exclusively at the level of the leading Gaussian term of the Edgeworth expansion and therefore concerns only first-order asymptotic quantities. More specifically, the matrices Σ^O and Σ^R represent the leading $O(1/n)$ terms in the covariance expansions

of $\widehat{\boldsymbol{\beta}}^O$ and $\widehat{\boldsymbol{\beta}}^R$. Consequently, the difference $\Sigma^R - \Sigma^O$ identified in Proposition 2.4 is itself a first-order quantity of order $O(1/n)$. The apparent superiority of the offset approach therefore relies on a comparison between leading terms only and is meaningful insofar as the full asymptotic covariance matrices of $\widehat{\boldsymbol{\beta}}^O$ and $\widehat{\boldsymbol{\beta}}^R$ are sufficiently well approximated by their leading Gaussian components Σ^O and Σ^R, as defined in equations (2.4.1) and (2.4.2). Since higher-order contributions in the Edgeworth expansion are neglected, the resulting approximation error is of smaller order asymptotically, but may nevertheless be of the same order of magnitude as the first-order difference $\Sigma^R - \Sigma^O$ in finite samples. As a result, it is likely inappropriate to draw definitive conclusions regarding the relative performance of the two approaches based solely on Proposition 2.4. A natural extension would therefore consist in deriving and comparing higher-order terms in the Edgeworth expansion. While such an analysis would be of clear theoretical interest, it lies beyond the scope of the present paper.

2.4.2.1 Simulations Study

Instead, we focus on assessing the accuracy of the asymptotic approximation and on comparing the two methods through a simulation study. The simulation study is designed solely to assess the empirical relevance of Proposition 2.4, which provides a first-order asymptotic comparison between the offset and ratio approaches based on the leading Gaussian term of the Edgeworth expansion of the estimators' covariance matrices. To this end, independent samples are generated under a data-generating process that is fully consistent with the modeling assumptions underlying the theoretical result. For each scenario, defined by a combination of sample size n and Tweedie power parameter p, repeated Monte Carlo replications are performed, and both the offset and ratio estimators are computed using the same Tweedie specification as that used for data generation.

For each approach, the resulting collection of parameter estimates is used to construct empirical covariance matrices, which provide finite-sample approximations of the true sampling covariances of the estimators. The empirical counterpart of the theoretical comparison is then obtained by examining the difference between these covariance matrices and by evaluating its spectral properties, in particular through the sign of its smallest eigenvalue. Across all scenarios considered, the empirical results are consistent with the ordering predicted by Proposition 2.4. In particular, the empirical difference $\widehat{\Sigma}^R - \widehat{\Sigma}^O$ is found to be positive definite in all tested configurations.

From the standpoint of parameter estimation accuracy, the offset approach may be preferable to the ratio

approach in finite samples, as it consistently yields smaller empirical variances for the estimator of β. This empirical finding is in line with the asymptotic ordering established in Proposition 2.4.

2.4.3 Sum of Individual Observed Gaps Comparison

In the context of automobile insurance pricing, one of the primary objectives is not only parameter estimation but also the accurate determination of premiums. From this perspective, we instead focus on proximity to financial balance as a key criterion for evaluating the different approaches. As discussed in Denuit *et al*. (2024), financial balance broadly refers to the situation in which the sum of observed total losses coincides with the sum of premiums computed under a given pricing method. Since this balance is not exactly achieved under either approach due to the use of a logarithmic link function in premium modeling, financial balance naturally emerges as an important criterion to consider when selecting between the two approaches.

We consider n independent contracts with observed loss costs $y_i = t_i z_i$ for $i = 1, \dots, n$, where z_i denotes the loss cost per unit of risk exposure t_i. We define the *individual observed gap* as the difference between the observed loss cost and the estimated premium. For contract i, these gaps are denoted by Δ_i^O and Δ_i^R under the offset and ratio approaches, respectively :

$$\Delta_i^O = t_i\big(z_i - \widehat{\zeta}_i^O\big), \qquad \Delta_i^R = t_i\big(z_i - \widehat{\zeta}_i^R\big),$$

where $\widehat{\zeta}_i^O$ and $\widehat{\zeta}_i^R$ are the estimated premiums obtained from the offset and ratio approaches. These quantities can be interpreted as realizations of the corresponding premium estimators. Our quantities of interest are therefore the aggregated gaps

$$\sum_{i=1}^{n} \Delta_i^O \quad \text{and} \quad \sum_{i=1}^{n} \Delta_i^R. \tag{2.4.3}$$

Although these sums are generally not equal to zero—reflecting the use of a logarithmic link function rather than the canonical link of the Tweedie model in premium modeling—we show that, under the ratio approach, the sum of individual observed gaps tends to be closer to zero than under the offset approach.

To analyze this criterion, we rely on the same simulation framework as in the previous section and examine, for each replication with $p = 1.5$, the ratio between the total predicted premiums and the total observed losses under both approaches. This comparison is illustrated in Figure 2.2. Across all considered sample sizes, the distribution associated with the ratio approach is markedly more concentrated around the value 1—corresponding to a situation close to financial balance—than the distribution obtained under the offset approach. In contrast, the offset approach exhibits substantially greater dispersion, reflecting higher variability around the global balance condition, even as the sample size increases.

A complementary perspective consists in identifying, for each simulation, which approach yields a total predicted premium closer to the observed total losses. According to this criterion, and for all scenarios under consideration, the ratio approach is closer to financial balance in approximately 87% of the replications. This result further supports the view that, from the standpoint of aggregate financial balance, the ratio approach displays a more stable and robust behavior than the offset approach in finite samples.

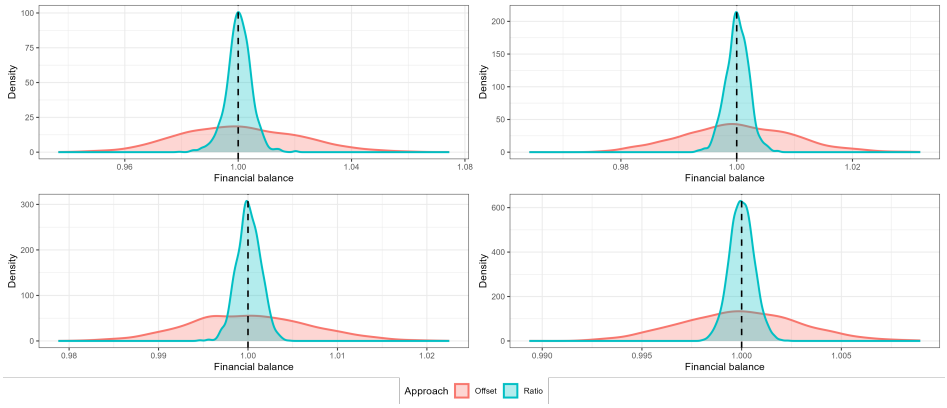


Figure 2.2 – Financial balance density for $n = 1{,}000$, $5{,}000$, $20{,}000$, and $50{,}000$ under the offset and ratio approaches.

Nevertheless, rather than relying solely on simulation results, we argue that additional insight into this financial balance criterion can be gained by examining the parameter estimation equations themselves. To this end, we first consider a homogeneous setting in which the sum of individual observed gaps is exactly zero under the ratio approach. We then extend the analysis to a heterogeneous setting in order to compare the behavior of this aggregate gap under the ratio and offset approaches.

2.4.3.1 Homogeneous Portfolio

First, we will assume the analysis of a homogeneous portfolio for which no explanatory variable is used in the mean parameter of the two approaches under study. More formally, we assume that the loss costs $Z_1, \ldots, Z_n$ are independent and have a common mean parameter ζ. We also assume that $\zeta_i = \zeta$ for all contracts $i = 1, \ldots, n$. This allows us to express the log-likelihood function for ζ as follows :

$$\ell(\zeta) = \frac{1}{\phi}\left(\frac{\zeta^{1-p}}{1-p}\sum_{i=1}^{n} w_i z_i - \frac{\zeta^{2-p}}{2-p}\sum_{i=1}^{n} w_i\right) + \sum_{i=1}^{n} a\left(z_i, w_i, \phi, p\right). \tag{2.4.4}$$

The first derivative of this log-likelihood function is given by :

$$\ell^{'}(\zeta) = \frac{1}{\phi\zeta^{p}}\left(\sum_{i=1}^{n} w_i z_i - \zeta\sum_{i=1}^{n} w_i\right). \tag{2.4.5}$$

Thus, the Quasi-Maximum Likelihood Estimator (QMLE) of ζ, denoted by $\widehat{\zeta}$, is obtained as follows : $\widehat{\zeta} = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}$.

1. **Offset Approach :** Under the offset approach, the QMLE of the homogeneous risk parameter ζ is given by
$$\widehat{\zeta}^{O} = \frac{\sum_{i=1}^{n} t_i^{2-p} z_i}{\sum_{i=1}^{n} t_i^{2-p}}. \tag{2.4.6}$$
The corresponding aggregate estimated premium is therefore
$$\sum_{i=1}^{n} t_i \widehat{\zeta}^{O} = \left(\sum_{i=1}^{n} t_i\right)\frac{\sum_{i=1}^{n} t_i^{2-p} z_i}{\sum_{i=1}^{n} t_i^{2-p}}. \tag{2.4.7}$$
Exact financial balance, i.e.
$$\sum_{i=1}^{n} t_i \widehat{\zeta}^{O} = \sum_{i=1}^{n} t_i z_i,$$
holds if and only if the t_i^{2-p}-weighted average of the observed loss costs z_i coincides with their t_i-weighted average. This condition is automatically satisfied in degenerate cases such as $p = 1$, or when all exposures t_i are equal, in which case the two weighting schemes coincide. However, under

the general setting considered in this paper, where $1 < p < 2$ and exposures satisfy $t_i \leq 1$ with heterogeneous durations, there is no structural mechanism in the offset model that enforces this identity. As a result, aggregate financial balance is not guaranteed under the offset approach and may fail in general, although it can still occur for particular realizations of the portfolio. Consequently, the aggregate deviation

$$\sum_{i=1}^{n} \Delta_i^O = \sum_{i=1}^{n} t_i \widehat{\zeta}^O - \sum_{i=1}^{n} t_i z_i$$

is not constrained to vanish.

2. **Ratio Approach :** In the ratio approach, where Z_i also serves as the response variable, the weight for contract i is defined as $w_i = t_i$. The QMLE for ζ, denoted by $\widehat{\zeta}^R$, is calculated as :

$$\widehat{\zeta}^R = \frac{\sum_{i=1}^{n} t_i z_i}{\sum_{i=1}^{n} t_i}. \tag{2.4.8}$$

In this case, the aggregate observed loss costs matches the sum of the expected premiums, as shown by :

$$\begin{aligned}
\sum_{i=1}^{n} t_i \widehat{\zeta}^R &= \sum_{i=1}^{n} t_i \left(\frac{\sum_{i=1}^{n} t_i z_i}{\sum_{i=1}^{n} t_i} \right) \\
&= \frac{\sum_{i=1}^{n} t_i z_i}{\sum_{i=1}^{n} t_i} \sum_{i=1}^{n} t_i \\
&= \sum_{i=1}^{n} t_i z_i.
\end{aligned}$$

So, we conclude that $\sum_{i=1}^{n} \Delta_i^R = 0$. This result suggests that the ratio approach inherently ensures that the sum of estimated premiums aligns with the aggregate observed loss costs, making it advantageous for achieving balance at the portfolio level.

2.4.3.2 Heterogeneous Portfolio

The homogeneous framework can be extended to account for heterogeneity by incorporating the covariate vector $\mathbf{X}_i$ in the premium model for each contract i. This allows for contract-specific premium estimation and

captures variation across contracts. In contrast to the homogeneous portfolio, where a closed-form expression for the QMLE of $\boldsymbol{\beta}$ was available, the presence of covariates implies that the QMLE of the parameter vector $\boldsymbol{\beta}$ no longer admits a closed-form solution. This is due to its dependence on $\mathbf{X}_i$. Consequently, the QMLE of $\boldsymbol{\beta}$ is obtained using the IRLS algorithm (see Section 2.2.4.1) for both the offset and ratio approaches, as discussed in the previous sections.

To better understand the gap between the aggregate observed loss costs and the sum of estimated premiums in each approach, we initialize the IRLS algorithm with values of $\boldsymbol{\beta}$ based on the homogeneous portfolio assumption. This initialization promotes consistency in comparing the offset and ratio approaches and clarifies how incorporating $\mathbf{X}_i$ affects the alignment of estimated premiums with aggregate loss costs.

1. **Offset Approach :** We initialize the IRLS algorithm with :

$$\beta_0 = \log\left(\widehat{\zeta}^O\right), \text{ and } \beta_j = 0, \text{ for } j = 1, \ldots, q.$$

By denoting the initial value by $\boldsymbol{\beta}^O_{(0)}$, the estimate of the parameter vector at iteration K, $\boldsymbol{\beta}^O_{(K)}$, is obtained as follows :

$$\boldsymbol{\beta}^O_{(K)} = K\boldsymbol{\beta}^O_{(0)} + \underbrace{\sum_{k=1}^{K-1}\left(I_n^O\left(\boldsymbol{\beta}^O_{(k)}\right)\right)^{-1}\nabla^O\left(\boldsymbol{\beta}^O_{(k)}\right)}_{\delta^O_{K-1}} = K\boldsymbol{\beta}^O_{(0)} + \delta^O_{K-1},$$

where $I_n^O(\cdot)$ is the Fisher information matrix for the offset approach. Therefore, we calculate the estimate premium at iteration K, denoted by $\widehat{\zeta}^O_{i(K)}$, as follows :

$$\widehat{\zeta}^O_{i(K)} = \left(\widehat{\zeta}^O\right)^K \exp\left\{\mathbf{X}_i^T\delta^O_{K-1}\right\},$$

where $\widehat{\zeta}^O$, defined by equation (2.4.6), represents the annual premium for all insureds in the offset approach for the homogeneous portfolio. We can conclude from this last equation that the sum of the estimated premiums at iteration K, denoted as $\sum_{i=1}^n t_i\widehat{\zeta}^O_{i(K)}$, does not necessarily equal the sum of

the observed loss costs, $\sum_{i=1}^{n} t_i z_i$. This discrepancy is typical in the offset approach, given its weighted nature and the way the IRLS updates progress over iterations.

2. **Ratio Approach :** We initialize the IRLS algorithm with :

$$\beta_0 = \log\left(\widehat{\zeta}^R\right), \text{ and } \beta_j = 0, \text{ for } j = 1, \ldots, q.$$

By denoting the initial value by $\boldsymbol{\beta}^R_{(0)}$, the estimate of the parameter vector at iteration K, $\boldsymbol{\beta}^R_{(K)}$, is obtained as follows :

$$\boldsymbol{\beta}^R_{(K)} = K\boldsymbol{\beta}^R_{(0)} + \underbrace{\sum_{k=1}^{K-1}\left(I_n^R\left(\boldsymbol{\beta}^R_{(k)}\right)\right)^{-1}\nabla^R\left(\boldsymbol{\beta}^R_{(k)}\right)}_{\delta^R_{K-1}} = K\boldsymbol{\beta}^R_{(0)} + \delta^R_{K-1},$$

where $I_n^R(\cdot)$ is the Fisher information matrix for the ratio approach. Therefore, we calculate the estimate premium at iteration K as follows :

$$\widehat{\zeta}^R_{i(K)} = \left(\widehat{\zeta}^R\right)^K \exp\left\{\mathbf{X}_i^T\delta^R_{K-1}\right\},$$

where $\widehat{\zeta}^R$, defined by equation (2.4.8), represents the annual premium for all insureds in the ratio approach for the homogeneous portfolio. This allows us to calculate the sum of the estimated premiums at iteration K, $t_i\widehat{\zeta}^R_{i(K)}$, leading to the following equation :

$$\begin{aligned}\sum_{i=1}^{n} t_i\widehat{\zeta}^R_{i(K)} &= \left(\sum_{i=1}^{n} y_i\right) \times \left(\frac{\left(\widehat{\zeta}^R\right)^{K-1}}{\sum_{i=1}^{n} t_i}\sum_{i=1}^{n} t_i \exp\left\{\mathbf{X}_i^T\delta^R_{K-1}\right\}\right) \\ &= \sum_{i=1}^{n} y_i \times \epsilon_{(K-1)},\end{aligned}$$

highlighting the connection between the total premiums and the aggregate claims $\sum_{i=1}^{n} y_i$. That means that unless $\epsilon_{(K-1)}$ equals 1, there is no financial equilibrium for this approach. While the sum

of estimated premiums may not exactly match the observed aggregate loss costs, this formulation allows for an analysis of $\epsilon_{(K-1)}$, the gap between these two quantities.

2.4.3.3 Numerical Illustration

To better understand the implications of the parameter estimation equations underlying the two approaches—particularly in the heterogeneous case, where no explicit analytical solution is available—we propose a simple numerical illustration. The objective of this illustration is not to assess the statistical performance of the estimators, but rather to highlight the structural differences between the estimation equations and clarify how these differences translate into distinct behaviors with respect to the financial balance criterion.

The setup is as follows :

1. **Portfolio setup.** In the homogeneous case, we consider a portfolio of $n = 100$ independent contracts without covariates. In the heterogeneous case, two risk factors are incorporated into the mean specification of each contract. These covariates are generated independently from binomial distributions :
   - $x_1 \sim$ Binomial$(100, 0.75)$,
   - $x_2 \sim$ Binomial$(100, 0.15)$.

   This specification induces heterogeneity in the portfolio while remaining sufficiently simple to allow a transparent interpretation of the estimation equations.
2. **Risk exposure.** Exposure levels t_i are drawn from a uniform distribution on the interval

$$\left[\tfrac{30}{365}, \tfrac{335}{365}\right],$$

   and then sorted in increasing order to assign contract indices $i = 1, \ldots, n$. This ordering reflects a common actuarial situation in which higher indexed contracts correspond to larger exposure levels.
3. **Loss cost.** Rather than assuming a stochastic loss-generating mechanism, we consider a deterministic specification given by

$$y_i = i, \qquad i = 1, \ldots, n.$$

   This choice is deliberate and aims to isolate the effect of the estimation equations themselves. By eliminating random noise, the resulting comparison focuses exclusively on the structural differences between the offset and ratio approaches and on their implications for aggregate financial balance.

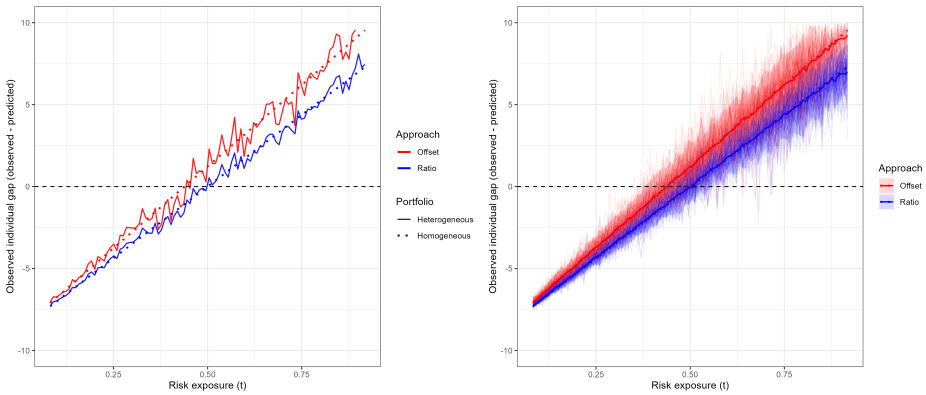


Figure 2.3 – Individual observed gaps : single heterogeneous realization (left) and across multiple heterogeneous realizations (right).

We then apply the corresponding parameter estimation equations to this portfolio. For a single realization, the individual observed gaps are displayed in the left panel of Figure 2.3. This panel allows a direct comparison across portfolio types (homogeneous versus heterogeneous) and estimation approaches (offset versus ratio). Several qualitative insights can be drawn from this figure :

— **Ratio approach (solid and dashed blue curves).**

Recall that, for a homogeneous portfolio, the ratio approach satisfies the exact aggregate balance condition $\sum_{i=1}^{n} \Delta_i^R = 0$, as established in Section 2.4.3.1. Although this aggregate quantity is not explicitly reported in the figure, the dashed blue curve corresponds to the individual observed gaps obtained under the homogeneous assumption and therefore sums to zero by construction.

This homogeneous benchmark provides a natural reference for interpreting the solid blue curve, which represents the individual observed gaps obtained under the heterogeneous specification. The two curves are remarkably close over the entire range of exposure levels, indicating that the introduction of covariates does not materially alter the aggregate balance property of the ratio approach. In particular, this visual proximity suggests that the sum $\sum_{i=1}^{n} \Delta_i^R$ in the heterogeneous case remains close to zero.

At the same time, the small local deviations and step-like variations observed along the solid blue curve reflect the fact that exact financial balance cannot generally be achieved in the heterogeneous

setting, except in accidental cases where the sum of these local variations happens to cancel out. These discrepancies arise from the contract-specific nature of the estimated premiums and from the iterative nature of the estimation procedure. Nevertheless, the overall behavior remains strongly anchored to the homogeneous equilibrium, highlighting the robustness of the ratio approach with respect to portfolio heterogeneity.

— **Offset approach (solid and dashed red curves).**

As established in Section 2.4.3.1, the offset approach does not satisfy an exact aggregate balance condition, that is, $\sum_{i=1}^{n} \Delta_i^O \neq 0$. This feature is directly visible in Figure 2.3, where both red curves display a systematic displacement relative to the ratio approach, which serves as a natural benchmark since its aggregate gap is zero by construction.

In the homogeneous case (dashed red curve), this displacement is already apparent, as the offset curve lies uniformly above its ratio counterpart, indicating a departure from financial equilibrium. In the heterogeneous case (solid red curve), the offset curve remains centered around its homogeneous counterpart, suggesting that the imbalance persists even in the presence of additional local variability arising from contract-specific covariates.

To further assess the impact of parameter estimation on financial balance, we consider a Monte Carlo experiment based on 1,000 heterogeneous portfolios generated from varying realizations of x_1 and x_2. The corresponding results are reported in the right panel of Figure 2.3. In this panel, all 1,000 individual gap curves are displayed simultaneously as semi-transparent lines, providing a direct visualization of the outcomes associated with heterogeneous portfolios. To summarize this variability, a pointwise 95% envelope is added in the form of a shaded ribbon, while the bold solid curve represents the median gap across replications. We can see that the individual gap curves obtained across replications are strongly concentrated around their respective homogeneous benchmarks for both approaches. This is reflected by the close proximity of the median heterogeneous curve to the corresponding homogeneous curve shown in the left panel, as well as by the relatively narrow width of the 95% ribbons over the entire range of risk exposure. These patterns indicate that heterogeneity primarily introduces dispersion around the homogeneous structure, rather than inducing systematic shifts in the gap curves themselves.

An important implication of this concentration phenomenon is that the aggregate imbalance observed under

the homogeneous specification provides a meaningful first-order diagnostic for the heterogeneous case. Since heterogeneity mainly adds random fluctuations around the homogeneous benchmark, a substantial lack of financial balance at the homogeneous level is unlikely to be systematically corrected once heterogeneity is introduced. Conversely, when the homogeneous imbalance is small, near-balance under heterogeneous modeling becomes more plausible, albeit still not structurally enforced. In this respect, the two approaches differ fundamentally. Exact aggregate balance holds by construction for the homogeneous ratio approach, whereas no such property holds for the homogeneous offset approach. As a result, the concentration of heterogeneous gap curves around their homogeneous counterparts implies that, in heterogeneous portfolios, the ratio approach is typically closer to financial equilibrium than the offset approach.

Finally, we note that the offset approach may occasionally exhibit a better financial balance than the ratio approach for specific realizations, as observed in the simulation results reported in Figure 2.2, where it is selected in approximately 13% of the replications. As illustrated in Figure 2.3, such outcomes can partly be attributed to more pronounced random fluctuations, whose aggregate effect may incidentally compensate for the residual imbalance. However, these situations are also more likely to occur in scenarios where the homogeneous offset specification itself happens to be close to financial balance. In such cases, the additional dispersion introduced by heterogeneity may suffice to produce a near-balanced aggregate outcome under the offset approach. Since this compensation mechanism is not structural but contingent on the particular data realization and on the initial homogeneous imbalance, the ratio approach generally provides a more reliable and robust financial balance across heterogeneous portfolios.

2.5 Empirical Illustration

The dataset used in this chapter corresponds to a subsample of the dataset analyzed in Chapter 1, Section 1.4. However, to ensure the independence of contracts, we restrict the original dataset to include only one vehicle contract per policy. In this revised dataset, each policy corresponds to a single contract, resulting in a dataset of approximately 161,390 contracts.

2.5.1 Description of Data

2.5.1.1 Description of Contracts

A standard car insurance contract typically carries a one-year risk exposure ; however, cancellations can occur within this period. To offset the administrative costs associated with such cancellations, the insured is subject to a financial penalty for breaching the contract. Several scenarios can lead to mid-term cancellations :

1. The insured sells the vehicle and no longer requires insurance coverage.
2. The insured opts to switch to another insurer, potentially attracted by a lower cost despite the cancellation penalty.
3. The insured experiences a significant loss (e.g., theft or accident) and subsequently changes vehicles, rendering the original insurance coverage unnecessary.

In cases #1 and #3, it could be argued that the situation constitutes a vehicle change rather than an outright cancellation of insurance. In practice, insurers often do not impose penalties for such changes, provided the insured continues coverage with the same insurer for the new vehicle. However, a vehicle change frequently enables the insured to explore options with other insurers, which can quickly lead to scenario #2.

2.5.1.2 Summary of Data

In Table 2.1, we present a summary of the data available for the numerical application. It shows that the majority of the observations in the database come from insured individuals covered for a full year. However, a significant portion of the contracts observed in the database corresponds to mid-term cancellations. It can be observed that the average exposure of contracts without cancellations is 1, while contracts with mid-term cancellations have an average exposure of approximately 0.5. The table also indicates the average risk exposure for each group. The last column presents what we refer to as the **Loss Cost Reference**, which is derived by dividing the average loss cost of each group by the portfolio's average [1]. A notable difference emerges between the two groups, indicating that insureds with mid-term cancellations tend to have higher insurance damages than those in the other group.

1. The use of the *Loss Cost Reference* helps prevent excessive disclosure of confidential information from the insurer who provided us with the data.

Group	Proportion of contracts	Average Risk Exposure	Loss Cost Reference
Full exposure	64%	1.00	0.63
Mid-term cancellation	36%	0.51	2.45

Table 2.1 – Descriptive statistics by group of contracts

Figure 2.4 presents a detailed view of the risk exposure distribution for insured individuals with mid-term cancellations, while the risk exposure for other insured individuals remains constant at 1. The distribution shows that risk exposure is fairly uniform throughout the year ; however, there is still considerable variation.

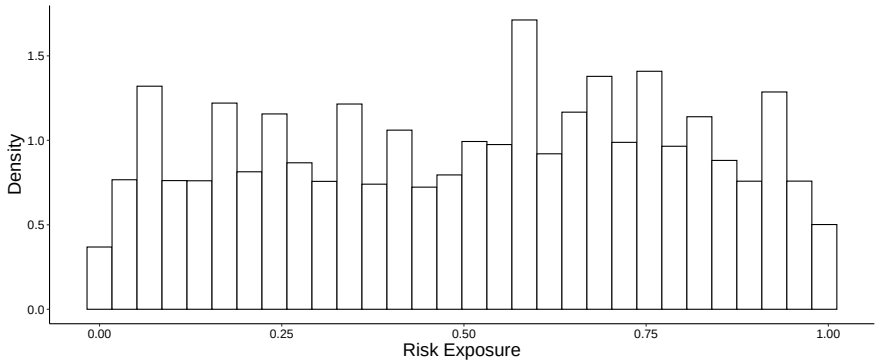


Figure 2.4 – Distribution of risk exposures

To further examine the relationship between risk exposure and loss costs, we grouped contracts by the total number of coverage months, from 1 to 11. For each coverage duration, we calculated the *loss cost reference* and plotted these values against the average risk exposure, as shown in Figure 2.5. Insured individuals with full exposure (coverage for 12 months) do not appear in the groups for months 1 to 11 ; thus, only their average value is displayed in Figure 2.5 as a dashed line.

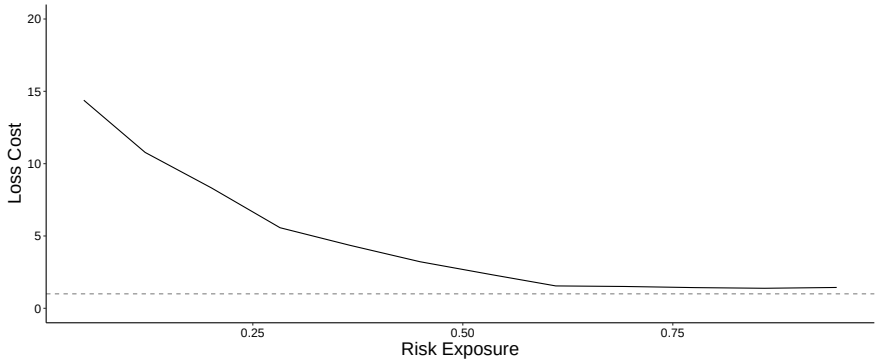


Figure 2.5 – Loss cost references by risk exposures

For insured individuals with mid-term cancellations, the loss cost reference curves exhibit a clear downward trend as risk exposure increases, which clearly aligns with the Decreasing Scenario in Section 2.4.3.3. This seems to imply that policyholders who cancel at the beginning of their contract have more claims, or claims with greater severity, than others. This finding highlights the significant role of risk exposure in explaining contract costs. Although traditional pricing models face challenges in predicting an insured's risk exposure or the likelihood of cancellation before the end of their contract, incorporating this information into pricing models is crucial for accurately estimating parameters and setting premiums.

2.5.2 Offset and Ratio Approaches

To compare the offset and ratio regression approaches, we estimate the parameters using a Tweedie distribution for each segment of the portfolio (insureds with full exposure and those with mid-term cancellation). The same eight covariates as those described in Chapter 1, Section 1.4.1.1, are used in the estimation. Furthermore, for each approach (ratio and offset), we calculate the premium for each contract using both methods. Accordingly, we use the same estimated variance parameter, $p = 1.42$, as obtained in the numerical analysis of Chapter 1 for modelling the Tweedie distribution (this estimate was not explicitly reported in the results of Chapter 1, as it was not the primary focus of that analysis). As discussed in Section 2.2, the dispersion parameter does not affect premium modelling ; therefore, we do not emphasize its role in this analysis.

2.5.2.1 Estimated Parameters

The estimated parameters $\boldsymbol{\beta}$ for covariates $X_1 - X_8$ can be different for both approaches. To compare the estimated values, we compute a coefficient ratio for each covariate by dividing the estimated coefficient from the offset approach by the corresponding coefficient from the ratio approach. If the estimated values were identical for both approaches, this ratio would be 1 for all covariates. These results are displayed in Figure 2.6. Without surprise, the estimated coefficients are identical in both the offset and ratio approaches, which was expected because the offset and ratio approaches yield equivalent results when the risk exposure is equal to 1.

On the other hand, for contracts with mid-term cancellations, the estimated coefficients differ noticeably between the offset and ratio approaches. This divergence is anticipated because, as discussed in Section 2.3, the log-likelihood and gradient functions differ for the two approaches when risk exposure is less than 1. The significant differences, particularly in coefficients associated with covariates X_1, X_2, and X_4, highlight how the choice of approach impacts premium estimation. These differences imply that premiums for certain policyholder profiles will vary substantially based on the chosen approach.

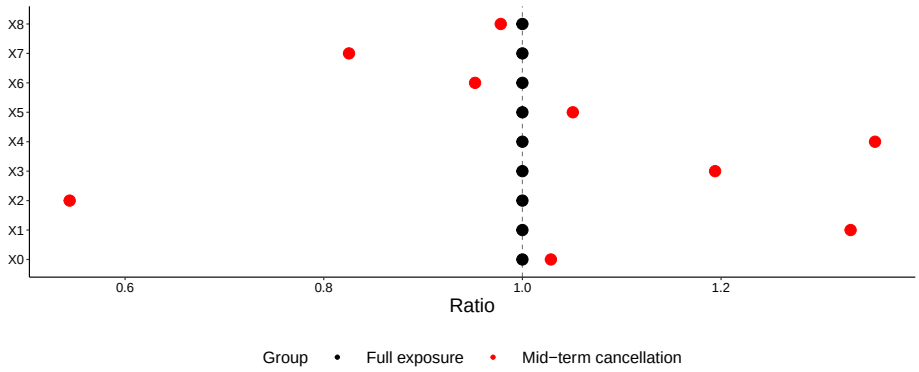


Figure 2.6 – Ratio of estimated parameter vectors

2.5.2.2 Estimated Premiums

We also compared the distribution of premiums for each approach. To do so, we calculated the premium ratio for each method by dividing the estimated premium for each contract under the offset approach by the premium for the same contract under the ratio approach. The distribution of the ratio of estimated premiums, with box-plots, is presented in Figure 2.7. For the group of insured individuals covered for a one-year period, we observe identical estimated premiums for all contracts in this group (Figure 2.7). This outcome is expected,

as all contracts in this group have a risk exposure of 1, and, as previously discussed, the offset and ratio approaches yield equivalent results in this situation. In contrast, for contracts with mid-term cancellations, we observe premium ratios greater than 1, indicating that the offset approach produces higher premium estimates compared to the ratio approach. This finding is consistent with the Decreasing Scenario discussed in Section 2.4.3.3, where the average claim amount decreases as the average risk exposure increases.

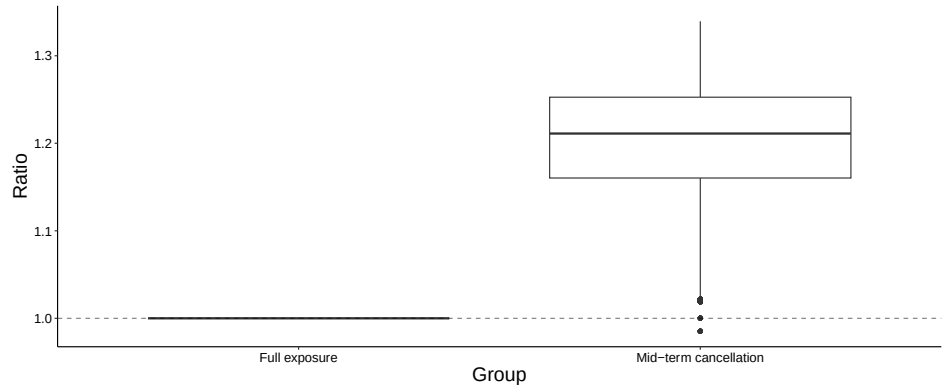


Figure 2.7 – Box-plot of the ratio of estimated premiums

2.5.2.3 Financial Equilibrum Analaysis

Financial equilibrium requires that the aggregate observed loss costs closely align with the sum of estimated premiums at each level of the risk factors in the regression model. To assess this, we calculated the aggregate observed loss costs and the sum of estimated premiums for each level of our eight risk factors, across all contract groups and approaches. These values were then organized in ascending order based on the aggregate observed loss costs, and we computed ratios by dividing the sum of estimated premiums (from both the offset and ratio approaches) by the aggregate observed loss costs. These ratios are plotted against aggregate observed loss costs in Figure 2.8, with risk classes labeled as integers from 1 to 17.

In the full-exposure group, the sum of individual observed gaps is identical for both the offset and ratio approaches, which is explained by the fact that the estimated premiums are identical under the two methods. As noted previously, it is nevertheless interesting to observe that, in the heterogeneous data, the sums of predicted and observed losses by risk class may differ even when the risk exposure equals one. A similar phenomenon is observed under homogeneous data when the offset approach is used. For the mid-term cancellation group, all ratios are reported in the right panel of Figure 2.8. The ratio approach yields a closer alignment between the sum of estimated premiums and the aggregate observed loss costs than the offset approach. In particular, the ratio of the sum of estimated premiums to the aggregate observed loss cost

under homogeneous data using the offset approach is equal to 1.22, whereas the ratio approach yields exact aggregate balance under homogeneity. As discussed in Section 2.4.3.3, ratios of similar magnitude are observed for both approaches in the heterogeneous case.

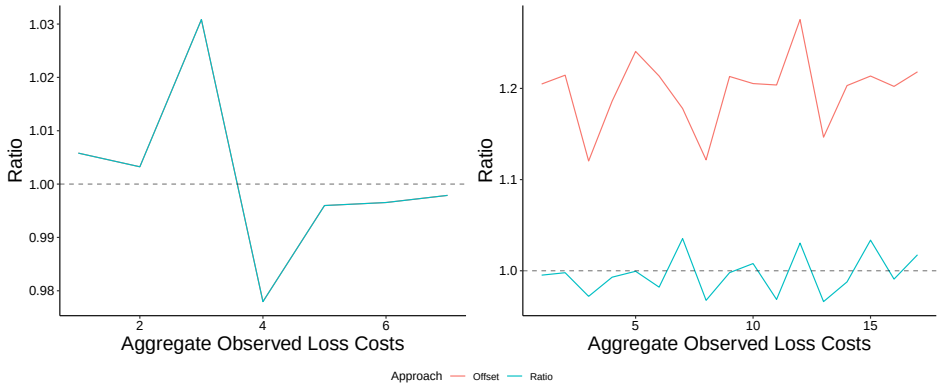


Figure 2.8 – Sum of individual observed gaps comparison (left : Full exposure, right : Mid-term cancellation)

2.5.3 Discussion

Even though neither of the two approaches studied is perfect in every respect, the advantages of using the ratio approach for a Tweedie regression are clear for the insurance portfolio we analyzed. However, the analysis of the modeling choices considered in this paper also brings to light several important issues, which may lead to different conclusions or modeling strategies depending on the portfolio structure and the underlying risk characteristics.

First, the relevance of the offset or ratio approach strongly depends on the composition of the insurance portfolio. As illustrated in Figure 2.1, the impact of the chosen approach remains relatively limited for portfolios characterized by a low frequency of mid-term cancellations, such as home insurance or certain commercial insurance lines, where most contracts exhibit full-year exposures. In such contexts, the offset and ratio approaches yield similar results. In contrast, for portfolios with a substantial proportion of partial-year exposures—typical of certain insurance markets, such as North American automobile insurance—the choice between the two approaches becomes critical and can materially affect the modeling outcomes.

These differences are not only technical but also have direct implications for premium determination and policyholder fairness. As shown by the estimated parameters reported in Figure 2.6, the offset and ratio approaches may lead to substantially different premiums for the same insured risk. In a competitive

insurance market, such discrepancies may influence which policyholder profiles are favored or disadvantaged by each approach. Understanding how these pricing differences affect fairness perceptions and market competitiveness is therefore an important consideration when selecting a ratemaking methodology.

The empirical results further emphasize the central role of mid-term cancellations in shaping loss costs. Table 2.1 and Figure 2.4 show that cancellations have a significant impact on observed loss costs, underlining the need to address this phenomenon more explicitly in pricing models. Rather than incorporating exposure mechanically within a single Tweedie framework, as done in this paper, a more detailed analysis of cancellation behavior could allow insurers to develop more refined modeling strategies. In particular, distinguishing between different causes of cancellations may help capture heterogeneous risk dynamics and improve premium adequacy.

Finally, the analysis raises broader questions regarding the assumption of strict linear proportionality between exposure and expected loss cost, an assumption that underlies both the offset and ratio approaches. While this assumption is widely adopted in actuarial practice, the descriptive results presented in Figure 2.5 suggest that loss costs do not necessarily scale linearly with exposure. Shorter exposure periods, in particular, appear to be associated with systematically different loss dynamics, consistent with several scenarios discussed in Section 2.5.1.1. Although both approaches considered here impose linearity by construction, these empirical patterns indicate that relaxing this assumption could provide additional modeling flexibility and lead to a more accurate representation of exposure-related heterogeneity.

2.6 Conclusion

This paper has examined two widely used strategies for incorporating risk exposure into premium estimation under Tweedie regression models : the offset approach and the ratio approach. Although both methods rely on the same proportionality assumption between exposure and expected loss, we have shown that they induce fundamentally different weighting structures and therefore lead to distinct statistical and financial properties. From a purely inferential perspective, both approaches yield consistent estimators of the true parameter vector under mild regularity conditions. A first-order asymptotic analysis based on the leading Gaussian term of the Edgeworth expansion suggests that the offset approach may exhibit greater efficiency, in the sense of a smaller asymptotic covariance matrix. Simulation results seem to confirm this comparison, indicating that in finite samples the offset approach also displays smaller empirical variance. Beyond parameter estimation, this paper emphasizes that premium modeling in insurance is inherently a financial exercise,

for which aggregate balance considerations play a central role. From this standpoint, a key contribution of this work is to show that the ratio approach exhibits a structurally stronger proximity to financial equilibrium. In homogeneous portfolios, it satisfies an exact aggregate balance condition by construction, while in heterogeneous settings it remains systematically closer to balance than the offset approach. Importantly, this property is not driven by asymptotics or incidental compensation effects, but follows directly from the structure of the estimating equations. In the final section of the paper, we complement the theoretical and simulation-based analyses with an empirical study based on a real Canadian automobile insurance portfolio. This application allows for a direct comparison of the offset and ratio approaches both in terms of parameter estimates—where the two approaches yield different values—and in terms of aggregate financial balance. The empirical results indicate that the ratio approach achieves a closer alignment between the sum of estimated premiums and the aggregate observed loss costs, whereas the offset approach exhibits a systematic imbalance.

The empirical results also suggest that a more refined examination of the relationship between risk exposure and observed loss costs may be warranted. In particular, the assumption of a strictly linear proportionality between exposure and expected loss appears to be restrictive in the presence of mid-term cancellations and heterogeneous exposure patterns, even though such proportionality is implicitly imposed by both modeling approaches considered in this paper. One important source of this heterogeneity arises when vehicles are replaced during the policy term, either following the sale and purchase of another vehicle or, as discussed in Section 2.5, after a significant loss that renders the original coverage unnecessary. In this context, more detailed tracking of vehicle replacements within insurers' databases could help improve the representation of exposure-related risk. More generally, these findings indicate that, while risk exposure t remains a vehicle-level quantity, its relationship with risk may depend on the broader exposure configuration of the policy to which the vehicle belongs. In particular, when multiple vehicles are insured sequentially or concurrently under the same policy, considering interactions among vehicle-level exposures may improve the representation of the underlying risk structure.

CHAPITRE 3

VARYING RISK EXPOSURE IN AUTO INSURANCE : A WEIGHTED TWEEDIE FRAMEWORK FOR EXPERIENCE RATING AND CANCELLATION PENALTIES

3.1 Introduction

In property and casualty insurance—automobile insurance in particular—contracts are typically written for a one-year term. At renewal, a policyholder may either remain with the same insurer or move to a competitor. Although most policy modifications occur at renewal, mid-term cancellations—situations in which a policy is terminated before its scheduled expiry date—are also common in North America. Insurers often levy a financial penalty in such cases, primarily to recover administrative costs. Several situations may lead to mid-term cancellations :

1. The insured sells the vehicle, and coverage is no longer required ;
2. The insured experiences a major loss (e.g., theft or total loss) and replaces the vehicle, thereby requiring a new policy ;
3. The insured switches to another insurer, often motivated by lower premiums despite the cancellation charge.

Cases (1) and (2) are sometimes treated operationally as vehicle substitutions rather than true cancellations, and insurers frequently waive cancellation penalties when coverage continues for the replacement vehicle. From an actuarial standpoint, however, substitutions introduce important data and modelling challenges. Standard pricing databases are not designed to track vehicle swaps : the cancellation of one contract and the creation of another often occur at different dates (sometimes weeks apart), and the designation of the principal driver may change across vehicles within the same household. These timing inconsistencies complicate the identification of substitutions and hinder the construction of coherent longitudinal exposure histories.

Because commonly used pricing datasets do not explicitly record substitutions, actuaries typically treat each vehicle contract independently and rely solely on the contract's recorded exposure duration. Recent work has begun to exploit dependencies among vehicles held by the same policyholder (Boucher et Coulibaly, 2024; Turcotte et Boucher, 2023; Pechon *et al.*, 2019), but these approaches do not fully resolve the practical

difficulties created by asynchronous cancellations and additions. Until richer, transaction-level data become widely available, pricing practice will continue to rely on modelling each contract's observed exposure.

It is therefore essential to distinguish contracts cancelled mid-term from those affected by vehicle additions or replacements. Although both contract types have an exposure duration shorter than one year, they arise from different operational mechanisms. As detailed in Section 3.2, mid-term cancellations correspond to contracts whose termination date differs from the initially scheduled end date, whereas vehicle additions and replacements retain the original policy end date. Regardless of the underlying cause, accurate modelling of such contracts requires careful consideration of risk exposure.

However, most pricing models that account for exposure rely on the assumption that the premium μ—defined as the expected value of the response variable—is proportional to the exposure t, typically specified as (see, e.g., Denuit *et al.* (2007b); Frees *et al.* (2014))

$$\mu = t \, \exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\}, \tag{3.1.1}$$

where $\boldsymbol{\beta}$ denotes the vector of parameters to be estimated and $\mathbf{x}$ the covariate vector. In actuarial science, loss costs are commonly modeled using distributions from the Tweedie family (see Tweedie *et al.* (1984)), which is parameterized by two components through which exposure may affect premiums : the mean μ and the weight parameter w. In this context, Chapter 2 examines two classical specifications—the offset formulation ($w = 1$) and the ratio formulation ($w = t^{p-1}$)—both grounded in the proportional mean structure given in (3.1.1). The analysis establishes that the offset approach is asymptotically more efficient than the ratio approach. However, when performance is assessed in terms of portfolio-level financial equilibrium, the ratio approach proves to be more effective. Although these findings are insightful, the underlying proportionality assumption between premiums and exposure appears overly restrictive. Unlike Chapter 2, which retains the proportional mean structure and modifies only the weight specification, we relax the proportionality assumption itself and allow the exposure to enter the mean through a flexible function. In particular, the empirical evidence reported in Section 3.2 of this chapter challenges this assumption. The objective of the present study is therefore twofold :

1. To generalize the specification of the mean premium μ to allow more flexible relationships between

exposure t and expected loss, i.e. $\mu = \gamma(t)\ \exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\}$,

2. To identify an appropriate weight function w consistent with this generalized structure.

This extension is motivated by both empirical evidence and practical considerations in premium calculation and risk classification. It is also worth noting that mid-term cancellations resulting from major losses (such as total losses requiring vehicle replacement) may create additional strategic considerations for insurers. In such cases, insurers may benefit from structuring cancellation penalties to partially reflect the occurrence of severe claims. This mechanism effectively allows the insurer to indirectly identify major losses through the cancellation surcharge, capturing part of the expected cost associated with the claim. From a competitive perspective, this ensures that policyholders maintaining full-year coverage implicitly benefit from a reduced annual premium, whereas those cancelling mid-term bear a larger share of their corresponding expected cost. Consequently, it is important to consider what adjustments may be introduced to accommodate this flexible framework.

The structure of the paper is as follows. Section 3.2 presents a detailed analysis of an automobile insurance dataset from a Canadian insurer and highlights the strong relationship between mid-term cancellations and poor driving behavior, thereby motivating the need for a more flexible pricing framework. Section 3.3 introduces several practical strategies designed to better penalize mid-term cancellations. Section 3.4 formally develops the Tweedie distribution theory underlying our ratemaking models. Section 3.5 illustrates the proposed methods using a numerical case study based on the same dataset, assessing their impact on insurer profitability and risk selection. Section 3.6 concludes.

3.2 Data Description and Objectives

We use the same automobile insurance dataset described in Chapter 1, Section 1.4, along with the same training and test sets. The analysis focuses on collision coverage, which insures property damage to the policyholder's vehicle in at-fault accidents. To inform the specification of the mean function μ introduced in the previous section, we briefly describe the dataset and present exploratory analyses highlighting the empirical relationship between recorded exposure and aggregate cost. The primary aim of this section is to document the data and the main variables used in the study, and to provide preliminary evidence on the functional form of μ that will guide subsequent modeling choices (e.g., offset vs. ratio vs. flexible exposure functions, and the selection of an appropriate weighting scheme).

3.2.1 Description of Contracts

To comprehensively account for all situations involving mid-term cancellations, we classify all contracts in the dataset into four distinct groups according to their exposure period :

- $\mathcal{XX}$: Contracts with full risk exposure for the entire duration of the policy.
- $\mathcal{OX}$: Contracts initiated during a policy period and exposed to risk until the period ends.
- $\mathcal{XO}$: Contracts exposed to risk at the start of a policy period but canceled before the policy expires.
- $\mathcal{OO}$: Contracts initiated during a policy period and canceled before the policy expires.

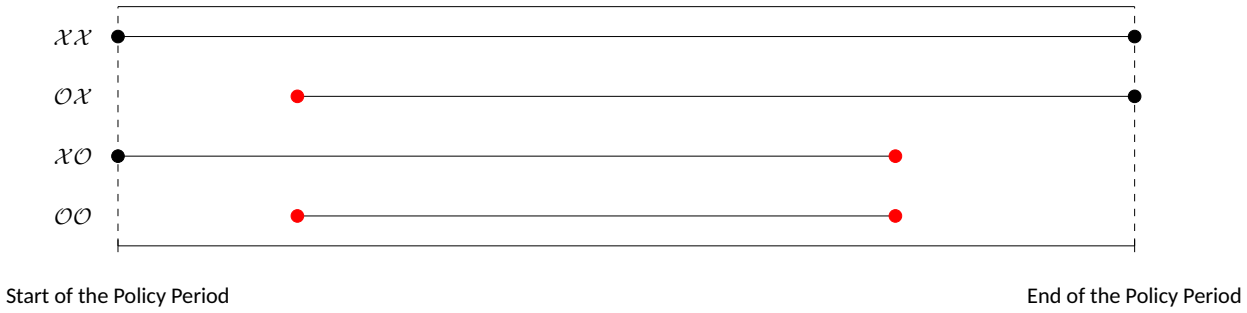


Figure 3.1 – *All truncation possibilities for a one-year policy period*

These groups can be better understood by referring to Figure 3.1. This classification provides a structured framework to capture the effects of mid-term cancellations and vehicle substitutions, thereby facilitating a more accurate analysis of insurance contract dynamics. However, since our objective is to model the penalty applied to policyholders who cancel before the end of their contract—rather than to address cases where additional vehicles (or coverages) are added during the term—we restrict our analysis to policyholders of types $\mathcal{XX}$ and $\mathcal{XO}$ only. Table 3.1 reports the proportions of policyholders in the $\mathcal{XX}$ and $\mathcal{XO}$ classes for both the training and testing datasets.

Datasets	Contract type	
	$\mathcal{XX}$	$\mathcal{XO}$
Training set	323 231 (34.7%)	608 261 (65.3%)
Test set	138 852 (34.8%)	260 360 (65.2%)

Table 3.1 – Descriptive statistics by contract type

3.2.1.1 Distribution of exposure

The histogram of Figure 3.2 presents the distribution of risk exposure specifically for the $\mathcal{XO}$ group in the training set, i.e., contracts that were canceled before the end of the policy period. All full-year exposures are excluded since they correspond to the $\mathcal{XX}$ group. This visualization highlights the variation in partial-year exposures and provides insight into the timing and prevalence of mid-term cancellations.

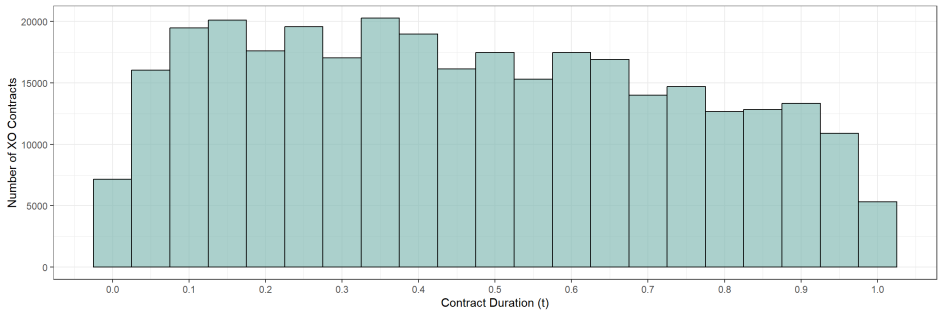


Figure 3.2 – Distribution of the risk exposure for the $\mathcal{XO}$ group

3.2.2 Empirical Impact of Mid-Term Cancellations on Contract Risk

To better understand the impact of mid-term cancellations on contract risk, and to challenge the standard pricing assumption of linearity between contract duration (risk exposure) and the premium, we first recall the definition of the loss cost, defined as the total amount paid in claims over the coverage period.

Consider n insurance contracts. Let n_i denote the total number of claims reported for contract i, for $i = 1, \ldots, n$, and let $z_{i,k}$ denote the cost (i.e., claim severity) of the k-th claim, for $k = 1, \ldots, n_i$ if $n_i > 0$. The observed loss cost for contract i, denoted by y_i, is therefore given by

$$y_i = \begin{cases} \sum_{k=1}^{n_i} z_{i,k}, & \text{if } n_i > 0, \\ 0, & \text{if } n_i = 0. \end{cases}$$

Accordingly, a first way to quantify the risk associated with a contract using the total loss cost is to compute the average loss cost, defined as $\frac{1}{n}\sum_{i=1}^{n} y_i$. However, since some contracts may terminate before the end of the coverage period, it is more appropriate to consider an exposure-adjusted indicator, referred to as the average annualized loss cost, defined as $\frac{1}{\sum_{i=1}^{n} t_i}\sum_{i=1}^{n} y_i$, where t_i denotes the risk exposure of contract i.

Using the definition of the loss cost, the average annualized loss cost can be expressed as

$$\frac{1}{\sum_{i=1}^{n} t_i} \sum_{i=1}^{n} y_i = \left(\frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} t_i} \right) \left(\frac{\sum_{i=1}^{n} \sum_{k=1}^{n_i} z_{i,k}}{\sum_{i=1}^{n} n_i} \right),$$

where, by convention, $\sum_{k=1}^{n_i} z_{i,k} = 0$ when $n_i = 0$, corresponding to the product of the claim frequency and the average claim severity.

Our empirical strategy therefore consists in analyzing each component of this decomposition in order to better characterize contracts that are canceled mid-term. It is also important to note that this analysis is conducted using the training sample. We begin by examining the claim frequency, defined as the ratio of the total number of claims to the total exposure, namely $\frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} t_i}$. More specifically, we study how this indicator varies by policy year and by contract type. Figure 3.3 presents the corresponding empirical distribution based on the training sample from the seven-year analysis period (i.e., the last seven years of the dataset). Although a more detailed model will eventually need to incorporate policyholder characteristics, we can already observe, consistently across years, that the claim frequency of policyholders who cancel their contracts mid-term is much higher than that of policyholders who maintain coverage until the end of the policy period.

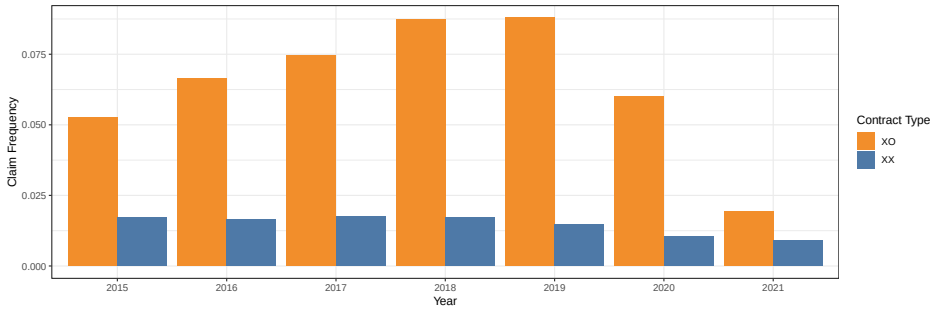


Figure 3.3 – Claim frequency by policy year and contract type

A similar analysis can be conducted for the average claim severity component. However, since claim severity is a continuous variable, it is particularly informative to examine its distribution across the two contract groups. This distributional perspective provides additional insights beyond those obtained from the claim frequency analysis.

Accordingly, for each group, we compute the average cost per claim, defined as $\frac{1}{n_i} \sum_{k=1}^{n_i} z_{i,k}$ when $n_i > 0$, and estimate the corresponding kernel density, as illustrated in Figure 3.4. Interestingly, policyholders who

cancel mid-term not only exhibit a higher claim frequency (relative to their risk exposure) but also tend to have higher claim severity, which further suggests that more severe claims may trigger policy cancellations.

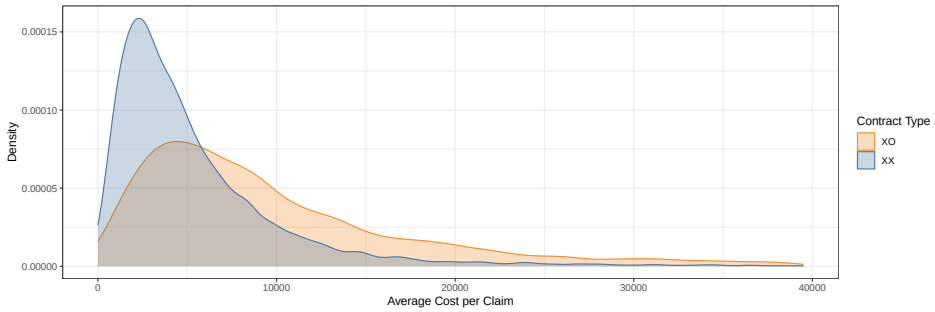


Figure 3.4 – Average cost per claim by contract type

Finally, Figure 3.5 displays the average loss cost and the average annualized loss cost as functions of risk exposure (with all policyholders in the $\mathcal{XX}$ group grouped at an exposure level of 1). For these plots, exposure is discretized into intervals of width 0.05, and policyholders are aggregated within each interval. The size of each point reflects the number of policyholders in the corresponding exposure group. Consistent with the patterns observed in the analyses of claim frequency and cost per claim, contract duration has a clear impact on claim outcomes.

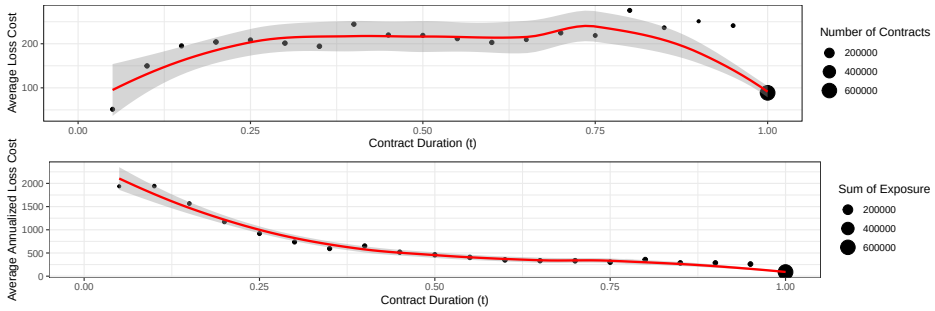


Figure 3.5 – Average loss cost by risk exposure

3.2.3 Available Covariates

The dataset includes multiple attributes for each contract and vehicle. To investigate how segmentation affects pricing, we focus on five primary covariates, labeled X_1 through X_5 for confidentiality reasons. These variables are selected from the eight covariates described in Chapter 1, Section 1.4.1.1.

It should be emphasized, however, that the selected covariates are not intended to be fully representative of those commonly used in practice, as their detailed impact is not the primary focus of our analysis. In particular, our objective is not to assess finely the contribution of each covariate to risk differentiation, but rather to use them as illustrative controls within the modeling framework.

Rather than focusing solely on understanding risk exposure and mid-term cancellations in the portfolio as a function of standard covariates, we place greater emphasis on the individual policyholder's experience with the insurer and their claims history. While it is clear that conventional covariates, such as age and gender, vehicle type, and usage patterns, affect the modeling of risk exposure, we argue that the policyholder's behavioral and historical profile is even more informative.

3.2.4 Claims Experience

Policyholders with higher inherent risk tend to report claims more frequently, which increases not only their expected loss cost but also the likelihood of experiencing a severe claim that may lead to a mid-term contract termination. To better distinguish among policyholders with different risk profiles, we follow the approach proposed by Boucher et Coulibaly (2024) and focus on a key covariate available in the dataset : the BONUS-MALUS SCALE (BMS) level, denoted by ℓ. By summarizing past claims experience, this level provides a structured and interpretable measure of individual risk, which is essential for isolating the behavior of exposure groups such as $\mathcal{XX}$ and $\mathcal{XO}$.

In practice, insurers routinely adjust premiums based on past claims, rewarding claim-free histories and penalizing frequent claimants. The BMS formalizes this mechanism by assigning each policyholder a level ℓ according to their historical claims. This level is updated at each contract renewal, increasing after a claim and decreasing following a claim-free period. Consequently, the BMS condenses a policyholder's entire claims record into a single, actuarially meaningful indicator of risk.

In the context of analyzing mid-term cancellations, the BMS level is particularly informative : it not only reflects underlying claim risk but may also influence the probability of early contract termination. Because the BMS parameters for this portfolio were already calibrated and optimized in Boucher et Coulibaly (2024), we directly use the resulting BMS levels as observed covariates. The system applied in the dataset is characterized by an initial level of 100 for new policyholders, a jump parameter of 3 for each reported claim, and admissible levels ranging from 95 to 104. Thus, policyholders with few or no claims tend to accumulate levels close to

$\ell = 95$, whereas those with multiple claims can reach levels up to $\ell = 104$; policyholders without significant claims generally remain around $\ell = 100$. This relationship is illustrated in Figure 3.6, which displays the average annualized loss cost across BMS levels separately for contracts of type $\mathcal{XX}$ and $\mathcal{XO}$.

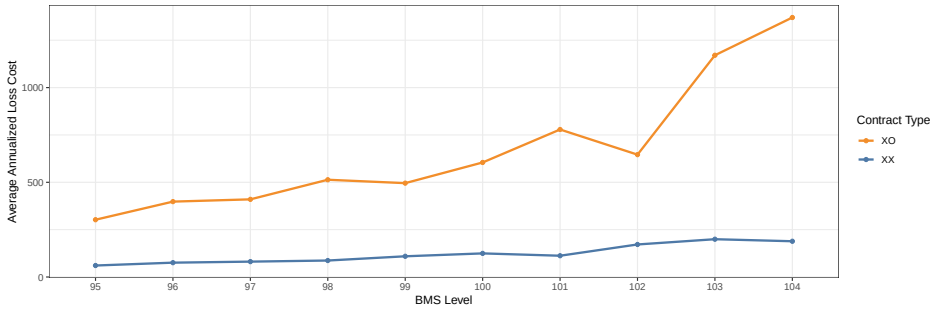


Figure 3.6 – Average annualized loss cost by BMS level and contract type

Consistent with the findings of Boucher et Coulibaly (2024), the figure reveals a clear positive association between the BMS level and the average annualized loss cost : lower BMS scores are associated with lower annualized loss costs, whereas higher BMS scores correspond to higher annualized loss costs. In addition, a pronounced separation between the two contract types is observed. Across all BMS levels, contracts of type $\mathcal{XX}$ exhibit consistently lower risk levels than contracts of type $\mathcal{XO}$, which are characterized by higher annualized loss costs and a greater propensity for mid-term cancellation.

To further characterize the relationship between the BMS score and contract type, Figure 3.7 displays the proportion of $\mathcal{XO}$ contracts across BMS levels, represented by the black line, while the light-blue bars indicate the corresponding number of contracts. A large share of policyholders is concentrated at $\ell = 95$, the most favorable level, reflecting a substantial segment of low-risk individuals. This group exhibits the lowest proportion of $\mathcal{XO}$ contracts. At the opposite end of the spectrum, policyholders with levels $\ell > 100$ constitute a relatively small portion of the portfolio and are generally considered the riskiest. Interestingly, their proportion of mid-term cancellations is not particularly high. Conversely, policyholders at intermediate levels, particularly $\ell = 100$ (often corresponding to newer clients) and $\ell = 99$ (those with limited claims history) display the highest proportion of $\mathcal{XO}$ contracts. Although the distribution of $\mathcal{XO}$ contracts across BMS levels appears relatively homogeneous, it remains informative regarding the likelihood of future contract cancellation. In particular, policyholders who are aware that a recent claim may substantially increase their future premium — due to the penalty induced by a deterioration in their BMS level — may be more inclined to

terminate their contract following such an event. Moreover, the BMS score may implicitly capture additional behavioral factors influencing the decision to cancel coverage. It is also important to note that, while the BMS level is treated as an exogenous covariate in our empirical analysis, it is inherently endogenous to the underlying pricing and claims process.

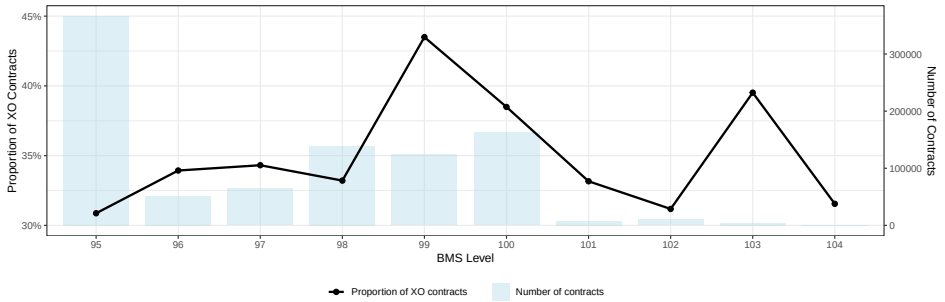


Figure 3.7 – Proportion of $\mathcal{XO}$ contracts by BMS level and number of contracts

3.3 Characterizing the Mean Response to Risk Exposure

3.3.1 Ratemaking Strategies

When incorporating risk exposure into premium calculations, two main approaches may be considered :

1. **Traditional Approach** : For contracts that terminate before the end of the policy year, the conventional actuarial assumption is that the expected claim amount is proportional to exposure. Denoting by t the fraction of the year during which the contract was active, the standard specification is :

$$\mu = t \times \exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\} \equiv \pi^{\text{Trad}}(t),$$

   where $\exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\}$ represents the annualized premium corresponding to a full year of coverage for a policy with covariate vector $\mathbf{x}$. This formulation implies that, in the absence of administrative fees, an insured who cancels halfway through the year would receive a pro rata refund equal to one half of the annual premium—that is, the premium scales linearly with recorded exposure t.

2. **Flexible Approach** : Empirical observations from the data suggest that the relationship between risk exposure t and the expected loss cost may be more complex than assumed under the traditional approach. Motivated by this finding, we consider the following more flexible specification :

$$\mu = \gamma(t) \exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\} \equiv \pi^{\text{Flex}}(t),$$

where $\gamma(t)$ is a smooth exposure function capturing deviations from linearity. Estimating $\gamma(t)$ directly can be challenging. As demonstrated in the Numerical Application of Section 3.5, we propose to use the framework of Generalized Additive Models (GAMs), following the approach introduced by Wood (2017). GAMs extend generalized linear models by incorporating non-linear relationships through smooth functions, making them well suited for capturing complex exposure–risk patterns commonly observed in insurance data.

Adopting a flexible exposure function is particularly appealing for insurers, as it enables a more refined treatment of mid-term cancellations—especially those arising after major loss events. By allowing differentiated refunds or surcharges that more accurately reflect actual exposure, such a specification improves the alignment between premium and risk, reduces cross-subsidization among policyholders, and may ultimately confer a competitive advantage.

3.3.2 Penalty Structures

While the flexible approach extends the traditional method and is likely to provide a better fit to the data—both on the training set and on new data, as suggested by our numerical application—this does not necessarily imply that it represents the optimal choice in practice. The actual risk exposure of a vehicle under a given contract is unknown at the time the policy is issued, and regulatory authorities often require insurers to disclose the full premium upfront, even if the contract is later canceled mid-term. Such requirements aim to prevent non-transparent pricing practices, for example those that could disproportionately penalize policyholders who cancel after experiencing an accident. However, these regulations do not completely rule out the possibility of implementing more sophisticated penalty structures. If all potential penalties are clearly stated in the contract prior to subscription, the insurer could, at least in principle, apply a premium scheme that better reflects expected risk exposure, even when cancellation occurs early. Within this framework, contractual transparency requirements may still allow insurers to design more adequate and actuarially grounded penalty structures, which is the perspective adopted in the analysis below.

For simplicity and to streamline notation, let us first assume that the policyholder pays the full annual premium $\pi^{\text{Flex}}(1)$, denoted by π^{Flex}, at the start of the contract. Under standard industry conventions, if the policyholder cancels the contract at time t, they expect a refund equal to $(1-t)\%$ of the annual premium π^{Flex}. However, in a flexible pricing framework where risk exposure t is incorporated into the premium through a function $\gamma(t)$, a cancellation occurring at time t would typically imply a refund of $\pi^{\text{Flex}} - \gamma(t)\pi^{\text{Flex}}$.

This model-based refund does not necessarily coincide with the conventional refund.

To evaluate whether it is at least partially feasible to implement a premium structure that explicitly accounts for mid-term cancellations, we compare these two refund amounts : the refund expected by the policyholder and the refund generated by the flexible model. We define the difference between these quantities as the *cancellation penalty*, denoted by $\rho(t)$:

$$\begin{aligned}\rho(t) &= \gamma(t)\pi^{\text{Flex}} - t\pi^{\text{Flex}} \\ &= \pi^{\text{Flex}}(\gamma(t) - t)\end{aligned}$$

Intuitively, the premium $\pi^{\text{Flex}}(t)$ and the penalty $\rho(t)$ associated with a cancellation at time t should satisfy some fundamental principles. To formalize these requirements, we impose the following two constraints :

- **(C1)** : $\pi^{\text{Flex}}(t) \geq \pi^{\text{Flex}}(t')$ for all $t \geq t'$, meaning that the refund prescribed by the flexible approach decreases as the contract progresses. Equivalently, the function $\gamma(t)$ must be increasing in the risk exposure t.
- **(C2)** : $\rho(t) \geq 0$, ensuring that the cancellation penalty is non-negative. At a minimum—and for strategic consistency—the premium charged for coverage lasting until time t should match the conventional benchmark $t\pi^{\text{Flex}}$ naturally expected by the policyholder. Furthermore, the pricing model must be protected against offering mid-term premiums below this benchmark, as doing so would undermine the intended structure and artificially inflate the annual premium.

One may further argue that the penalty never exceeds the remaining portion of the premium. This condition guarantees that a policyholder never pays more by canceling than by maintaining the contract until its maturity. However, given that the function $\gamma(t)$ must be increasing in the risk exposure t in (C1), this requirement is automatically satisfied.

As we will show in the empirical application, analyzing the premium $\pi^{\text{Flex}}(t)$ and the penalty $\rho(t)$ remains essential, and in some situations it may even be desirable to transform the penalty structure so as to ensure logically coherent and actuarially meaningful values. This point is crucial : imposing penalty structures

different from the conventional one may alter policyholder behavior. By analogy with Lemaire's *bonus-hunger effect* (Lemaire, 2012), which highlights how changes in claim-related penalties affect subsequent claim frequency, one must recognize that modifying penalties for mid-term cancellations will influence both the realized claims experience as a function of t and the cancellation probability at duration t.

The central idea is that insurers could, in principle, adopt alternative mid-term cancellation penalty structures. More generally, suppose that the current penalty function $\rho(t)$ is replaced by a modified structure $\rho'(t)$. Under such a change, the premium charged to a policyholder for coverage lasting a duration t could take the form

$$t\,\pi^{\text{Flex}} + \rho'(t),$$

which naturally raises the question of how to quantify the modified penalty $\rho'(t)$. This specification will be examined in detail in Section 3.5.

3.4 Understanding the Tweedie Distribution in Actuarial Models

To operationalize the ratemaking strategies and penalty structures discussed in Section 3.3, it is necessary to specify an appropriate statistical distribution for the response variable in the proposed models. As motivated in Section 3.1, we adopt the Tweedie distribution for this purpose. The objective of this section is therefore to provide theoretical justification for the performance of the flexible approach under the assumption that the loss cost follows a Tweedie distribution. Before presenting these results, we first discuss the main reasons for the widespread use of the Tweedie distribution in insurance pricing, in order to build intuition for the subsequent theoretical developments.

3.4.1 Justification in Auto Insurance Pricing

3.4.1.1 Variance–Mean Relationship

The Tweedie distribution is commonly characterized by four parameters : the mean parameter, denoted by μ, which represents the expected value of the distribution ; the dispersion parameter, denoted by ϕ; the weight parameter, denoted by w ; and the variance parameter, denoted by p. While the dispersion and weight parameters must be strictly positive, the mean parameter is, in principle, real-valued. However, in insurance pricing applications, the mean parameter is required to be positive, and additional restrictions are typically imposed on the variance parameter p, as discussed in the next subsection.

A key feature of the Tweedie distribution in insurance pricing is the functional relationship linking its variance to its mean parameter. More formally, if a random variable $Y \geq 0$ follows a Tweedie distribution, its variance is given by

$$\mathrm{Var}[Y] = \frac{\phi}{w}\mu^p.$$

When the response variable in a pricing model follows a Tweedie distribution, this variance specification implies a positive relationship between the mean parameter μ (interpreted as the premium) and the variability of the response. This feature allows the model to reflect the empirically observed increase in variability associated with higher expected response levels. Moreover, the flexibility provided by the variance parameter p enables the model to accommodate different degrees of risk escalation as the mean increases.

Finally, the presence of both the dispersion parameter ϕ and the weight parameter w in the variance structure plays an important role, as it allows the model to capture additional sources of heterogeneity across observations.

3.4.1.2 Loss Cost Variable

We now consider the stochastic counterpart of the observed loss cost introduced in Section 3.2.2. Let N denote the number of claims reported for a given contract, and let Z_k denote the cost of the k-th claim, for $k = 1, \ldots, N$ when $N > 0$. The loss cost, denoted by Y, is defined as

$$Y = \begin{cases} \sum_{k=1}^{N} Z_k, & \text{if } N > 0, \\ 0, & \text{if } N = 0. \end{cases}$$

Although each claim amount Z_k is a continuous random variable, the aggregate quantity Y is not itself continuous. In particular, the probability that no claim occurs ($N = 0$) is strictly positive, implying that the distribution of Y has a point mass at zero. The loss cost is therefore a semi-continuous random variable.

This feature motivates the widespread use of the Tweedie distribution in insurance pricing models. In particular, when the loss cost Y follows a Tweedie distribution with variance parameter satisfying $1 < p < 2$, it admits a compound Poisson–Gamma representation, as discussed in (Tweedie *et al.*, 1984; Jørgensen, 1997; Delong *et al.*, 2021). More precisely, the claim count N follows a Poisson distribution, while the individual claim amounts Z_k are assumed to be independent and gamma-distributed.

Under this representation, the Tweedie distribution can be interpreted as a mixed discrete–continuous distribution, making it a natural candidate for modeling loss cost Y. In addition, an exponential dispersion formulation can be derived, which facilitates statistical inference and premium modeling. The corresponding probability density function is given by

$$f(y) = \exp\left(\frac{w}{\phi}\left(\frac{\mu^{1-p}}{1-p}y - \frac{\mu^{2-p}}{2-p}\right) + a(y, w, \phi)\right), \qquad y \in \mathbb{R}, \tag{3.4.1}$$

where $a(\cdot)$ is a normalizing function that does not depend on μ.

Throughout the remainder of the paper, we assume that the variance parameter satisfies $1 < p < 2$. For notational convenience, we write

$$Y \sim \mathrm{Tweedie}(\mu, w, \phi, p).$$

3.4.1.3 Weight interpretation in auto insurance pricing

To better understand the role of the weight parameter in the Tweedie distribution, we recall its equivalence with the compound Poisson–Gamma representation, as summarized by Delong *et al.* (2021). Suppose that a random variable Y follows a Tweedie distribution, that is, $Y \sim \mathrm{Tweedie}(\mu, w, \phi, p)$. Then Y can be represented as a compound Poisson–Gamma random variable with the following components :

— The mean of the Poisson component is

$$w\frac{\mu^{2-p}}{(2-p)\phi}.$$

— The shape parameter and the mean of the Gamma component are given respectively by

$$\frac{2-p}{p-1}, \qquad \frac{(2-p)\phi}{w\mu^{1-p}}.$$

In this representation, the weight parameter w enters both components of the compound Poisson–Gamma distribution. It appears proportionally in the mean of the Poisson component, whereas it is inversely proportional to the mean of the Gamma component. This dual role is intuitive for the Poisson component, since w may naturally be interpreted as a measure of exposure or duration, consistent with the construction of a Poisson process.

In the special case where the weight parameter is set equal to the risk exposure, that is, $w = t$, the resulting Tweedie model has limited practical appeal. In premium modeling, exposure t is typically incorporated within a Tweedie framework through either the offset approach or the ratio approach, as introduced earlier. Under both approaches, the premium μ is assumed to be proportional to the risk exposure t, in line with traditional pricing practice.

The distinction between the two approaches lies in the specification of the weight parameter w : the offset approach sets $w = 1$, whereas the ratio approach specifies $w = t^{p-1}$. The ratio formulation is arguably more intuitive, as exposure t enters exclusively through the Poisson component of the compound Poisson–Gamma representation as an offset variable. This is consistent with actuarial pricing principles.

3.4.2 A Flexible Specification for Incorporating Risk Exposure

As illustrated in the top panel of Figure 3.5, the empirical relationship between exposure and observed loss cost does not follow the simple linear proportionality between premium and risk exposure assumed in traditional approaches. In particular, the average loss cost does not increase at a constant rate with respect to exposure, suggesting that the traditional specification may underestimate expected losses for contracts with low exposure. This departure from proportionality indicates that exposure does not act merely as a mechanical scaling factor, but instead interacts with the claim-generating process in a more intricate manner.

To account for this empirical behavior, we relax the proportionality constraint and allow exposure to affect the premium through a flexible adjustment function, as discussed in Section 3.3. Accordingly, the premium can be written in the additive form

$$\mu = \exp\Big\{ \mathbf{x}^\top \boldsymbol{\beta} + \log\big(\gamma(t)\big) \Big\},$$

which makes explicit the smooth contribution of exposure through the term $\log(\gamma(t))$.

To estimate $\boldsymbol{\beta}$, we assume that the response variable follows a Tweedie distribution with weight parameter w, while the dispersion parameter ϕ and the variance power parameter p are treated as fixed constants. Under this specification, the estimation problem naturally falls within the framework of GAM, as introduced in Section 3.3, with $\log(\gamma(t))$ approximated using spline basis functions ; see Wood (2017) for further details.

3.4.2.1 Consistency Under the Traditional and Flexible Approaches

The advantages of the flexible approach over the traditional one stem from the statistical properties of the corresponding estimators of the parameter vector $\boldsymbol{\beta}$. We consider the same n independent loss costs as in Section 3.2.2 and assume that they are generated from a common true distribution with density $g(\cdot \mid \boldsymbol{\beta}^{\text{True}}, t)$, where $\boldsymbol{\beta}^{\text{True}}$ denotes the true parameter vector. Although no specific distributional form is imposed on g, we assume that the true expected premium is linked to the covariate vector $\mathbf{x}$ and the risk exposure t through

$$\mu^{\text{True}} = \int_0^{+\infty} y\, g(y \mid \boldsymbol{\beta}, t)\, dy = \delta(t)\, \exp\Big\{\mathbf{x}^\top \boldsymbol{\beta}^{\text{True}}\Big\},$$

where $\delta(t)$ is a positive function of t. Our objective is to show that when traditional approaches are used to estimate $\boldsymbol{\beta}^{\text{True}}$, the resulting estimator is consistent *if and only if* $\delta(t) = t$. In other words, consistency is obtained only when the true mean is proportional to the exposure, as assumed in traditional approaches.

We recall the definition of consistency from Casella et Berger (2024) : an estimator $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ if

$$\lim_{n\to\infty} \Pr\Big(|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}| > \epsilon\Big) = 0 \qquad \text{for all } \epsilon > 0.$$

Because the true data-generating process g differs in general from the model assumed under traditional approaches, the model may be misspecified. We therefore rely on the theory of quasi–maximum likelihood estimation (or *pseudo–maximum likelihood*), developed by White (1982). Let $f(\cdot \mid \boldsymbol{\beta}, t)$ denote the Tweedie density used in traditional approach, whose mean parameter is $\mu = t\, \exp\big(\mathbf{x}^\top \boldsymbol{\beta}\big)$.

The Kullback–Leibler divergence between the true density g and the assumed Tweedie density f is defined as

$$KL(\boldsymbol{\beta}) = \int_0^{+\infty} \log\left(\frac{g(y \mid \boldsymbol{\beta}^{\text{True}}, t)}{f(y \mid \boldsymbol{\beta}, t)}\right) g(y \mid \boldsymbol{\beta}^{\text{True}}, t)\, dy.$$

According to White (1982), the probability limit of $\widehat{\boldsymbol{\beta}}$ is the minimizer of $KL(\boldsymbol{\beta})$, which is equivalent to maximizing the expected log-likelihood :

$$\int_0^{+\infty} \log f(y \mid \boldsymbol{\beta}, t)\, g(y \mid \boldsymbol{\beta}^{\text{True}}, t)\, dy = \int_0^{+\infty} \left(\frac{w}{\phi} \left(\frac{\mu^{1-p}}{1-p} y - \frac{\mu^{2-p}}{2-p} \right) + a(y, t, \phi) \right) g(y \mid \boldsymbol{\beta}^{\text{True}}, t)\, dy,$$

where w denotes any weight parameter used in the traditional approaches. In this setting, maximizing the expected log-likelihood is equivalent to maximizing the criterion

$$K(\boldsymbol{\beta}) = \frac{\mu^{1-p}}{1-p}\, \mu^{\text{True}} - \frac{\mu^{2-p}}{2-p}. \tag{3.4.2}$$

The gradient of $K(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$\nabla_{\boldsymbol{\beta}} K(\boldsymbol{\beta}) = \Big(\delta(t)\, t^{1-p} \exp\big\{ (1-p) \mathbf{x}^\top \boldsymbol{\beta} + \mathbf{x}^\top \boldsymbol{\beta}^{\text{True}} \big\} - t^{2-p} \exp\big\{ (2-p) \mathbf{x}^\top \boldsymbol{\beta} \big\} \Big) \mathbf{x}. \tag{3.4.3}$$

It follows that the true parameter vector $\boldsymbol{\beta}^{\text{True}}$ satisfies the first-order condition *if and only if* $\delta(t) = t$. Consequently, estimators obtained from traditional approaches converge to the true parameter vector only when the true mean is proportional to the risk exposure, exactly as assumed in these methods. It is worth noting that this convergence does not depend on the weight, dispersion, or variance parameters specified under these traditional approaches.

Similarly, suppose that the function f corresponds to the Tweedie density used in the flexible approach, whose mean parameter is specified as

$$\mu = \gamma(t)\, \exp\Big(\mathbf{x}^\top \boldsymbol{\beta} \Big).$$

In this case, we obtain another maximization criterion of the form given in Equation 3.4.2, based on this alternative mean specification. The associated gradient of this criterion is

$$\nabla_{\boldsymbol{\beta}} K(\boldsymbol{\beta}) = \Big(\delta(t)\, \gamma(t)^{\,1-p} \exp\big\{ (1-p)\, \mathbf{x}^\top \boldsymbol{\beta} + \mathbf{x}^\top \boldsymbol{\beta}^{\text{True}} \big\} - \gamma(t)^{\,2-p} \exp\big\{ (2-p)\, \mathbf{x}^\top \boldsymbol{\beta} \big\} \Big) \mathbf{x}.$$

It follows that the true parameter vector $\boldsymbol{\beta}^{\text{True}}$ satisfies the first-order condition if and only if $\delta(t) = \gamma(t)$. This result implies that, when the true mean loss cost is correctly specified—as assumed under the

flexible approach—the corresponding estimator of the parameter vector is consistent. Determining which specification of the true mean is most appropriate can be guided by examining the empirical relationship between the average loss cost and the risk exposure, as we did in the data description section.

It is also important to note that, although the possibility that $\delta(t) = t$ cannot be excluded, flexible methods reduce the risk of functional misspecification. In particular, the GAM specification adopted in the flexible approach enables the estimation procedure to capture, or closely approximate, the true functional relationship $\delta(t)$ through spline-based approximations, regardless of whether $\delta(t)$ coincides with t.

3.4.3 Choice of the Weight Parameter

As shown above, the weight parameter w of the underlying Tweedie distribution does not affect the consistency of the parameter vector estimator under the flexible specification. However, in order to obtain an estimate of this parameter vector, a specific value of w must be supplied. A natural choice in an insurance pricing context is to let w depend on the risk exposure t, that is, to set $w = \omega(t)$. In this regard, extending the approach proposed in Chapter 2 for a flexible function $\gamma(t)$, we consider two specifications for the weight function $\omega(\cdot)$ within the proposed framework :

1. **Constant-Weight Model (CWM)** This specification sets $\omega(t) = 1$. It can be viewed as a generalization of the *offset approach*, where all observations receive the same weight regardless of the exposure duration.
2. **Gamma-Weight Model (GWM)** This specification sets $\omega(t) = \gamma(t)^{\,p-1}$, which generalizes the *ratio approach*. This choice is motivated by the fact that the weight parameter appears exclusively in the Poisson component of the compound Poisson–Gamma representation of the Tweedie distribution, making $\omega(t)$ interpretable as a measure of exposure duration in the underlying Poisson process.

We further assume that $\omega(1) = 1$. This assumption sets the one-year contract as the reference level, ensuring that policies with shorter or longer exposure durations are evaluated relative to the standard annual contract. Accordingly, we redefine the weight for a general exposure level t as

$$\omega(t) = \left(\frac{\gamma(t)}{\gamma(1)}\right)^{p-1}.$$

It is also important to note that the estimation of the parameter vector in the GWM is inherently more challenging than in the CWM. In the CWM, the weights are constant across contracts, and the only non-linear component to estimate is the spline-based function $\gamma(t)$. In contrast, fitting a Tweedie GAM under the weighting scheme $\omega(t) = (\gamma(t)/\gamma(1))^{p-1}$ is non-standard because the same unknown function $\gamma(t)$—represented through a spline basis—appears both in the mean specification and in the weights. This creates an implicit dependence between the smoothing component and the estimation weights, requiring a tailored estimation strategy to ensure numerical stability and convergence.

To address this difficulty, we adopt the following iterative estimation scheme :

- **Initialization.** Fit an initial Tweedie GAM with mean $\mu = s_1(t)\exp\{\mathbf{x}^\top\boldsymbol{\beta}\}$, and weights $\omega(t) = t$, where $s_1(t)$ is an initial spline approximation of $\gamma(t)$.
- **Iteration.** Given the spline estimate $s_k(t)$ at iteration k, fit a model with $\mu = s_{k+1}(t)\exp\{\mathbf{x}^\top\beta\}$, and weights $\omega(t) = (s_k(t)/s_k(1))^{p-1}$.
- **Convergence.** Repeat this procedure until the sequence $s_k(t)$ stabilizes, yielding the final estimate of $\gamma(t)$ and the corresponding regression parameter vector $\boldsymbol{\beta}$.

This iterative algorithm ensures that the spline approximation of $\gamma(t)$ is coherently incorporated into both the mean specification and the weight structure, thereby enabling a stable and internally consistent estimation of the Gamma-Weight Model.

We also argue that a third weighting structure deserves consideration. To motivate this proposal, we recall the ratio approach introduced in Chapter 2. As shown therein, this framework can be equivalently formulated by considering the aggregate loss cost Y as the response variable and assuming that it follows a Tweedie distribution whose mean parameter is proportional to the risk exposure t, and whose weight parameter is given by t^{p-1}.

The proposed third weighting structure then generalizes this approach by retaining the same weighting scheme $\omega(t) = t^{p-1}$, while adopting a flexible specification for the mean of Y. In this interpretation, we obtain a closed-form weighting function $\omega(t)$ that is non-constant and enjoys practical appeal, as it is motivated by the same exposure-based weighting principle underlying the ratio approach. In line with the two models introduced above, we therefore define this third specification as follows :

3. **Exposure Weighted Model (EWM)**. A generalization of the ratio-based approach obtained by specifying $\omega(t) = t^{p-1}$ while allowing for a flexible exposure effect in the mean structure.

3.4.4 Model Selection

Model comparison in actuarial science is commonly performed using scoring rules derived from the predictive distribution of risk, as discussed in Gneiting et Raftery (2007). Among these, the logarithmic score is particularly appealing. However, the log-likelihood of the Tweedie distribution does not admit a closed-form expression, which restricts its direct application. For this reason, we rely on a deviance-based model selection criterion to guide the comparison among CWM, GWM, EWM, and the traditional ratio approache.

It should be emphasized, however, that the deviance-based selection criterion is purely statistical in nature, as recalled below. For this reason, we also consider complementary evaluation measures. In particular, we examine the Area Between the Curves (ABC) criterion, derived from concentration and Lorenz curves, as summarized in Denuit *et al.* (2019), and a criterion based on Murphy diagrams, which relies on Bregman dominance, as presented in Denuit *et al.* (2025).

3.4.4.1 Assessing Model Fit and Comparison Using the Deviance

We consider the same n independent loss costs $y_1, \ldots, y_n$ as previously introduced. We denote by μ_i the corresponding premium for observation i, and by t_i its risk exposure. According to Nelder et Wedderburn (1972), the deviance of a regression model is defined as follows :

$$\mathcal{D}(\widehat{\boldsymbol{\beta}}) = 2\phi \left(\ell^* - \ell(\widehat{\boldsymbol{\beta}}) \right),$$

where ℓ^* denotes the log-likelihood of the saturated model, and $\ell(\widehat{\boldsymbol{\beta}})$ is the log-likelihood evaluated at the estimated model. For the Tweedie model, the deviance takes the explicit form

$$\mathcal{D}(\widehat{\boldsymbol{\beta}}) = 2 \sum_{i=1}^{n} \omega(t_i) \left[y_i \left(\frac{y_i^{1-p} - \widehat{\mu}_i^{1-p}}{1-p} \right) - \frac{y_i^{2-p} - \widehat{\mu}_i^{2-p}}{2-p} \right],$$

where $\omega(t_i)$ is replaced by the weight used in the corresponding model, and $\widehat{\mu}_i$ denotes the premium estimate obtained under that same model. The deviance provides a quantitative measure of lack of fit : smaller values indicate a model that better captures the structure of the data. Consequently, when comparing

two competing models, the preferred model is the one with the lower deviance. In the numerical application, we explicitly perform this comparison by splitting the dataset into training and test samples, and by computing the deviance on both samples. The purpose of this split is to assess whether the competing models exhibit overfitting, which would typically manifest itself through a substantial increase in the deviance when moving from the training sample to the test sample. Ideally, one would expect only a small gap between the deviances computed on the two samples.

3.4.4.2 Model Comparison Using Concentration and Lorenz Curves

To compare competing premium estimators, we use the concentration curve, the Lorenz curve, and the associated ABC; see Denuit *et al.* (2019). Let μ_i denote the true premium for contract i, let $\widehat{\mu}_i$ be a candidate premium estimator, and let y_i be the observed loss cost. We write μ, $\widehat{\mu}$ and Y for the corresponding generic random variables, and assume that $\mathrm{E}[\mu] < +\infty$, $\mathrm{E}[\widehat{\mu}] < +\infty$ and $\mathrm{E}[Y] < +\infty$.

Let F denote the distribution function of $\widehat{\mu}$, that is,

$$F(x) = \Pr(\widehat{\mu} \leq x)\,, \qquad x \in \mathbf{R},$$

and let

$$F^{-1}(\theta) = \inf\{x \in \mathbf{R} : F(x) \geq \theta\}, \qquad 0 \leq \theta \leq 1,$$

be its generalized inverse. The concentration curve is defined by

$$CC[\mu, \widehat{\mu}; \theta] = \frac{\mathrm{E}\big[\mu\, \mathbf{1}_{\{\widehat{\mu} \leq F^{-1}(\theta)\}}\big]}{\mathrm{E}[\mu]}, \qquad 0 \leq \theta \leq 1. \tag{3.4.4}$$

As shown in Denuit *et al.* (2019), it can equivalently be written as

$$CC[Y, \widehat{\mu}; \theta] = \frac{\mathrm{E}\big[Y\, \mathbf{1}_{\{\widehat{\mu} \leq F^{-1}(\theta)\}}\big]}{\mathrm{E}[Y]}, \qquad 0 \leq \theta \leq 1, \tag{3.4.5}$$

which is the formulation used in practice since μ is not observable.

The Lorenz curve associated with $\widehat{\mu}$ is

$$LC[\widehat{\mu}; \theta] = \frac{\mathrm{E}\big[\widehat{\mu}\, \mathbf{1}_{\{\widehat{\mu} \leq F^{-1}(\theta)\}}\big]}{\mathrm{E}[\widehat{\mu}]}, \qquad 0 \leq \theta \leq 1. \tag{3.4.6}$$

The ABC criterion is then defined as

$$ABC(\widehat{\mu}) = \int_0^1 \Big(CC[Y, \widehat{\mu}; \theta] - LC[\widehat{\mu}; \theta] \Big)\, d\theta.$$

Following Denuit et Trufin (2024) and Denuit *et al.* (2024), the comparison of competing premium estimators through the ABC criterion is meaningful once they are put on a common globally balanced scale. For this reason, a candidate estimator $\widehat{\mu}$ may first be rescaled as

$$\widehat{\mu}_c = \frac{\mathrm{E}[Y]}{\mathrm{E}[\widehat{\mu}]}\, \widehat{\mu},$$

so that $\mathrm{E}[\widehat{\mu}_c] = \mathrm{E}[Y]$. Since this transformation preserves the ranking induced by $\widehat{\mu}$, it does not modify the concentration curve. The competing estimators can then be compared through their ABC values, smaller values indicating a closer proximity to local balance.

3.4.4.3 Murphy Diagrams

We adopt the same notation as in the previous subsection. In particular, μ denotes the true premium, Y the loss cost, and $\widehat{\mu}^{(1)}$ and $\widehat{\mu}^{(2)}$ two competing premium estimators. Following Denuit *et al.* (2025), we say that $\widehat{\mu}^{(2)}$ Bregman-dominates $\widehat{\mu}^{(1)}$ if

$$\mathrm{E}\Big[L(Y, \widehat{\mu}^{(2)}) \Big] \;\leq\; \mathrm{E}\Big[L(Y, \widehat{\mu}^{(1)}) \Big]$$

for every Bregman loss function L. This dominance relation provides a comparison criterion that does not depend on the choice of a particular loss function within the Bregman class.

Such a comparison can be examined graphically by means of Murphy diagrams. To this end, consider the elementary loss functions

$$L_m(Y, \widehat{\mu}^{(k)}) = \frac{1}{2}(m - Y)\, \mathbf{1}_{\{\widehat{\mu}^{(k)} > m\}}, \qquad k = 1, 2,$$

indexed by $m \in \mathbf{R}_+$. For each estimator $\widehat{\mu}^{(k)}$, the Murphy diagram is the graph of

$$m \longmapsto \mathrm{E}\Big[L_m(Y, \widehat{\mu}^{(k)}) \Big], \qquad m \in \mathbf{R}_+.$$

If

$$\mathrm{E}\Big[L_m(Y, \widehat{\mu}^{(2)}) \Big] \;\leq\; \mathrm{E}\Big[L_m(Y, \widehat{\mu}^{(1)}) \Big] \qquad \text{for all } m \in \mathbf{R}_+,$$

then $\widehat{\mu}^{(2)}$ dominates $\widehat{\mu}^{(1)}$ in the sense of Murphy diagrams, and therefore Bregman-dominates it ; see Denuit *et al.* (2025). Murphy diagrams thus provide a convenient graphical tool for comparing premium estimators over the whole class of Bregman losses. In contrast with the ABC criterion, which evaluates proximity to local balance once global balance has been enforced, Murphy diagrams assess dominance directly through expected loss.

3.5 Numerical Application

To illustrate the proposed modeling framework, we apply the different Tweedie specifications to the dataset introduced in Section 3.2. Recall that this dataset consists of a non-random sample extracted from a Canadian automobile insurance portfolio. It includes contracts of type $\mathcal{XX}$ (fully exposed to risk for the entire policy period) and $\mathcal{XO}$ (initially exposed but canceled before expiration). In addition to the covariates described in Section 3.2, the BMS level is included as an additional rating factor, as discussed in Section 3.2.4.

3.5.1 Adjusting the Models Under Different Weighting Schemes

We examine the three proposed flexible specifications—CWM, GWM, and EWM (introduced in Section 3.4.3)—and compare them with the traditional ratio approach, which has been recommended as superior to the offset approach for incorporating risk exposure t in conventional actuarial models (Boucher et Coulibaly, 2026).

For the GWM , Figure C.1 in Appendix C.1 illustrates the evolution of the weights at each iteration k of our algorithm, providing an informal visualization of its convergence. Convergence is extremely rapid : after only two or three iterations, the sequence of weights already appears to have reached its final shape.

All three models are calibrated using the training dataset, while the test dataset is reserved for evaluating their predictive performance. We do not estimate the dispersion parameter ϕ and instead fix the variance parameter at $p = 1.42$, using the estimate obtained in Chapter 1. To ensure a fair comparison across approaches, we use a cubic regression spline with the same set of knots. Although the choice of spline type does not affect our conclusions, enforcing the same knots and basis functions across models facilitates meaningful comparisons.

3.5.1.1 Weights and Spline-Based Estimation Analysis

The left panel of Figure 3.8 displays the weight functions $\omega(t)$ obtained for the three models under study. In the CWM, all observations contribute equally to the estimation of the Tweedie model parameters. In other words, each policy has the same influence on the estimation process, even if it has been in force for only a short period—sometimes just a few weeks or even days. In contrast, the two other approaches, namely the GWM and the EWM, assign different weights to observations with exposures lower than one. The EWM behaves intuitively, assigning increasing weights as exposure increases. The GWM , however, exhibits a more peculiar pattern : it assigns relatively higher weights to policies that were canceled mid-term.

Although the exposure t does not appear explicitly in the weight function $\omega(t)$ under the CWM, it is still incorporated into the model through the exposure function $\gamma(t)$, which is estimated using spline basis functions and shown in the right panel of Figure 3.8. This figure reveals that, although the estimated exposure curves have similar shapes across the three models, small but noticeable differences remain. For instance, the curves associated with the CWM and EWM tend to lie slightly below the one obtained under the GWM .

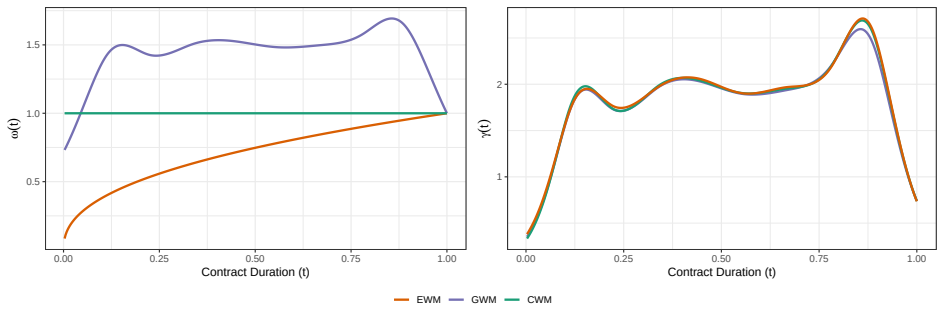


Figure 3.8 – Weight function $\omega(t)$ (left) and spline-based estimation of $\gamma(t)$ (right) for the flexible approaches

3.5.1.2 Mean Parameter Adequacy Test

We verify our assumption regarding the mean structure in the proposed flexible approaches by comparing the average loss cost and the average estimated premiums in both the training and test datasets, as illustrated in Figure 3.9. Before commenting on this figure, we examine the estimated coefficients reported in Table 3.2. This table presents the estimated β parameters included in the mean structure, corresponding to the five segmentation covariates as well as the effect of the BMS level. It is interesting to observe that the estimated

β coefficients exhibit more noticeable variation across models than the spline basis coefficients. This suggests that the weighting scheme primarily affects the relative importance of the covariates rather than the overall shape of the exposure adjustment.

Parameters	Traditional approach	Flexible approach		
		CWM	GWM	EWM
β_1	-0.158	-0.216	-0.235	-0.185
β_2	0.232	0.162	0.153	0.170
β_3	0.728	0.878	0.839	0.926
β_4	0.252	0.360	0.340	0.380
β_5	-0.239	-0.145	-0.158	-0.131
β_{BMS}	0.130	0.108	0.108	0.107

Table 3.2 – Estimated parameter vectors for all models

The mean specification assumed under the flexible approach appears to provide an adequate fit to the loss-cost data, as illustrated in Figure 3.9. This adequacy is observed for both the training and test datasets. We also note that the ratio approach—represented by the yellow curve—which assumes proportionality between the premium and the risk exposure, does not fit the data as well as the three flexible models, even though its increasing trend with respect to exposure is consistent with theoretical expectations.

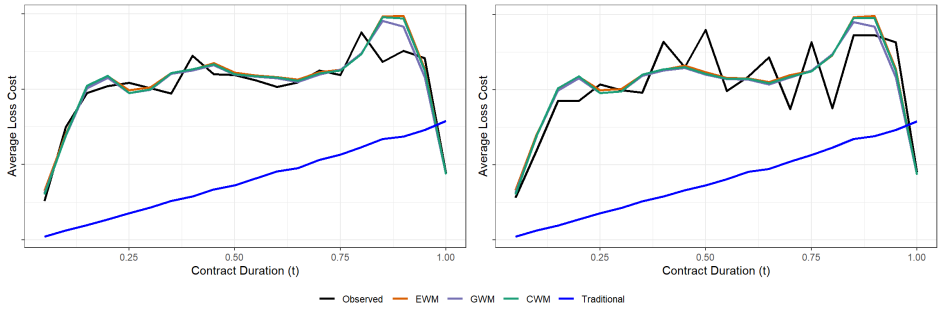


Figure 3.9 – Observed average loss costs and estimated average premiums (left : training set, right : test set)

3.5.1.3 Analysis of Alternative Weight Selection Criteria

As mentioned earlier, we rely on two additional criteria for selecting the weight specification : the deviance-based criterion (evaluated on both the training and test datasets) and the criterion derived from concentration and Lorenz curves (evaluated on training set).

For the deviance criterion, since the four models under comparison do not share the same weight function $\omega(t)$—although they use the same response variable—we introduce normalized observation weights. Let $t_1, \ldots, t_n$ denote the risk exposures associated with the n observations in the dataset on which the deviance is evaluated. The normalized weight associated with observation i is defined as

$$w_i = \frac{\omega(t_i)}{\sum_{j=1}^{n} \omega(t_j)}, \qquad i = 1, \ldots, n.$$

The advantage of this normalization is that it places all models on a comparable scale, as the scale effect of the original weight function is removed. Based on the normalized deviances in Table 3.3, we observe that the flexible approaches outperform the traditional ratio approach, yielding lower deviance scores for both the training and test datasets. Among the flexible specifications, EWM performs best, achieving the lowest deviance values on both datasets. We can also observe that the deviance of the test dataset is comparable to that of the training dataset, which suggests that the proposed flexible framework does not suffer from overfitting.

Datasets	Traditional approach	Flexible approach		
		CWM	GWM	EWM
Training set	114.64	102.57	102.60	102.56
Test set	114.77	102.77	102.80	102.76

Table 3.3 – Comparison of deviance scores across the four models for the training and the test set

Figure 3.10 is in line with the conclusions drawn from the deviance comparison. To construct this figure, we use the empirical counterparts of the concentration and Lorenz curves, computed on the training set ; see Denuit *et al.* (2019). For $0 \leq \theta \leq 1$, these are defined by

$$\widehat{CC}(\theta) = \frac{\sum_{i:\, \widehat{\mu}_i < \widehat{F}^{-1}(\theta)} y_i}{\sum_{i=1}^{n} y_i}, \qquad \widehat{LC}(\theta) = \frac{\sum_{i:\, \widehat{\mu}_i < \widehat{F}^{-1}(\theta)} \widehat{\mu}_i}{\sum_{i=1}^{n} \widehat{\mu}_i},$$

where y_i and $\widehat{\mu}_i$ denote the observed values of Y and of the candidate premium estimator $\widehat{\mu}$, respectively, and where $\widehat{F}^{-1}(\theta)$ is the empirical quantile of order θ computed from the empirical distribution of $\widehat{\mu}$.

We then rescale $\widehat{\mu}$ for all competing models, as described in Section 3.4.4.2. This rescaling ensures that global balance is achieved for each model. We subsequently compute the actual area between the curves using

the following expression :

$$\text{Area} = \int_0^1 \left| \widehat{CC}(\theta) - \widehat{LC}(\theta) \right| \, d\theta.$$

In the present setting, the Area measure is preferred to the ABC criterion because the empirical concentration and Lorenz curves seem to intersect. In such circumstances, the ABC criterion is less informative, since it depends on the signed difference between the two curves and may therefore be subject to compensations between positive and negative discrepancies. Consequently, a small, or even null, ABC value need not reflect a genuine proximity between the curves, but may simply result from offsetting deviations. The Area measure is therefore more suitable here, as it provides a more faithful quantification of the discrepancy between the two curves.

The curves obtained from the three flexible approaches are nearly indistinguishable graphically (see Figure 3.10). The Area measure shown in Figure 3.10 yields a substantially larger value for the traditional ratio approach (Area ≈ 0.13) than for the flexible specifications. Among the latter, the Area values are very close to one another. Although the differences are minor, GWM produces the smallest Area, suggesting a slight advantage in terms of local balance.

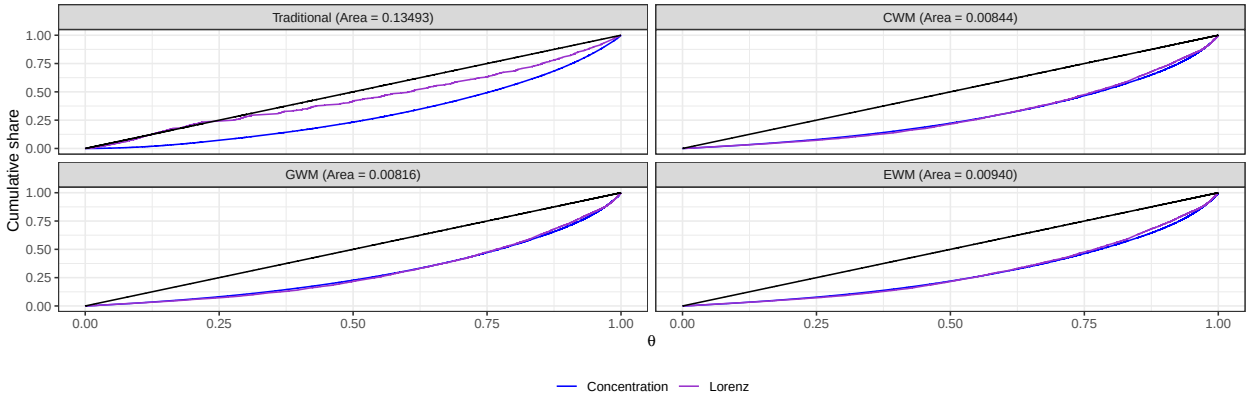


Figure 3.10 – Concentration and Lorenz curves

The strong performance of the flexible approaches is further clarified by Figure 3.11, which displays the mean exposure together with the proportion of full-year contracts along the same ordering induced by the estimated premiums. A marked contrast appears between the traditional and flexible specifications. For the flexible approach (illustrated here for the EWM), contracts with the lowest estimated premiums tend to

have an average exposure close to 1, and thus correspond predominantly to policies running to maturity. By contrast, the traditional specification imposes a predetermined relationship between the premium and the exposure t, and therefore cannot fully adjust the premium level for contracts that terminate before one year. As a result, short-exposure contracts tend to be underpriced relative to the flexible models, which helps explain the weaker overall performance of the traditional approach.

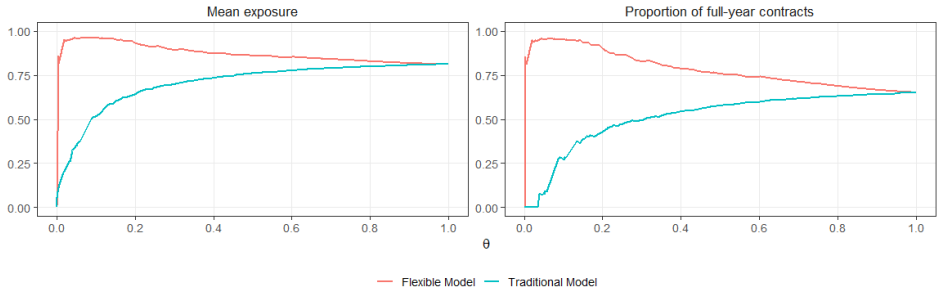


Figure 3.11 – Mean exposure and proportion of full-year contracts

3.5.2 Penalties and Ratemaking Strategies

The right-hand panel of Figure 3.8 displays the function $\gamma(t)$ as a function of exposure t. With respect to the ratemaking constraints introduced in Section 3.3.2, it is clear that the resulting pattern is not practically applicable, even though the three models with a flexible specification of the mean outperform the traditional models on both the training and test datasets. In particular, under this specification, a policyholder may face a higher mid-term cancellation penalty when canceling after only 0.2 years of coverage than if the policy were not canceled at all, which occurs when $\gamma(t)$ exceeds 1. To obtain a more realistic and operational structure, it is therefore preferable to impose the ratemaking constraints directly on the spline basis functions used in the three flexible Tweedie models.

For illustrative purposes, however, we propose to transform the function $\gamma(t)$ into a new function $\gamma_{\text{con}}(t)$ that satisfies the ratemaking constraints C1 and C2 introduced in Section 3.3.2, and to re-estimate all model parameters using this transformed function. This approach allows all parameters—in particular the regression coefficients β—to be estimated simultaneously, thereby enabling a partial adjustment to the imposed constraints. It should be noted that imposing structural constraints on $\gamma(t)$ modifies the estimation framework considered in Section 4. In particular, once monotonicity and ratemaking requirements are enforced, the estimator no longer arises from the unrestricted quasi–maximum likelihood setting of White (1982), and

consistency is therefore no longer guaranteed ; this reflects a trade-off between statistical optimality and practical implementability. In practice, however, the imposed constraints are mild and primarily affect the shape of the penalty function rather than the exposure adjustment in the interior of the domain. As a result, the core exposure–risk relationship captured by $\gamma(t)$ remains largely preserved.

Based on the results obtained under the EWM, Figure 3.12 displays the optimal penalty as a function of the risk exposure t. According to our definition, the penalty function $\rho(\cdot)$ corresponds to the area between the diagonal dashed line and the green curve in Figure 3.12. The green curve represents the transformed function $\gamma_{\text{con}}(t)$ obtained by imposing both ratemaking constraints (C1) and (C2). In our numerical application, imposing constraint (C2) has only a negligible effect on the results, as illustrated in the figure.

It should be emphasized, however, that imposing $\gamma_{\text{con}}(t)$ to be increasing in the exposure t and to lie everywhere above the identity function ($t \mapsto t$) does not guarantee that the penalty function $\rho(\cdot)$ is itself monotone in t. As illustrated in Figure 3.12, it may occur that a contract canceled after nine months incurs a smaller penalty than one canceled earlier in the year. Nevertheless, once the additional structural constraints of our pricing framework are imposed, we ensure that the total premium paid by a policyholder canceling after nine months cannot fall below that of policyholders who cancel earlier in the year.

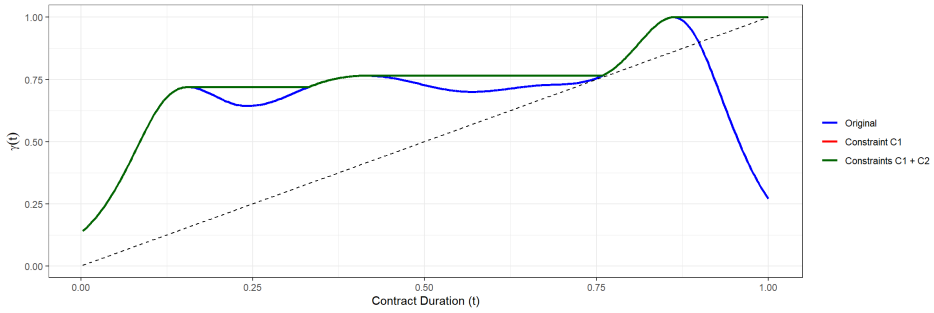


Figure 3.12 – Optimal penalty function as a function of exposure t

We compare the performance of the model using $\gamma_{\text{con}}(t)$, obtained by imposing both ratemaking constraints (C1) and (C2), with the Exposure-weight approach, represented by the blue (original) line in Figure 3.13. This figure displays the Murphy diagrams of the two models based on the following empirical version of the loss function :

$$L_m = \frac{1}{2n} \sum_{i=1}^{n} (m - y_i) \, \mathbf{1}_{\{\widehat{\mu}_i > m\}}.$$

We observe that the EWM outperforms the constrained model for all values of m, which is consistent with our theoretical findings. Indeed, we previously established that the parameter vector estimator is consistent under the Exposure-weight specification, ensuring that it captures the same structural relationship as the true premium.

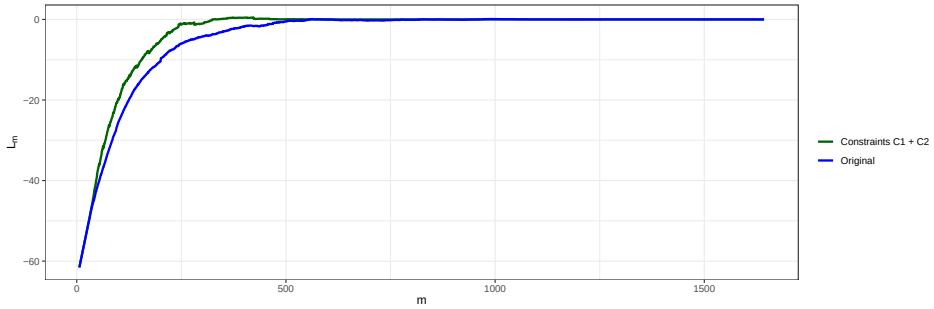


Figure 3.13 – Murphy diagram of the optimal penalty

3.5.2.1 Smoothing the Penalties

We observe that imposing constraint (C1) still produces extreme outcomes. After only a few weeks of insurance coverage, a policyholder would already be required to pay more than half (almost 75%) of the full annual premium if they cancel their policy. Although this may appear unfair at first glance, it is important to emphasize that, based solely on the empirical results and without imposing constraint (C1), the situation would be even worse : the policyholder would have to pay more than the full annual premium in order to cancel mid-term.

It is worth noting that, as with many other products involving annual subscriptions or, more generally, continuous usage over time, it may not be unreasonable for the insurance industry to reconsider its current commercial practices and reflect on the possibility of requiring a minimum premium, even when coverage is canceled after only a few days. Although such an approach could eventually be proposed, our purpose here is not to recommend a new pricing policy. Instead, our objective is to introduce a correction to the penalty function obtained earlier, in order to produce a more realistic premium structure. To that end, we propose

the following functional form :

$$\gamma_{\text{adj}}(t, a) = a\,\gamma_{\text{con}}(t) + (1 - a)\,t, \qquad 0 \leq a \leq 1.$$

Figure 3.14 illustrates several values of a and their corresponding effects on the penalty function. As can be seen, setting $a = 100\%$ recovers the optimal penalty function previously discussed, whereas $a = 0\%$ corresponds to a null penalty. This figure also highlights that the penalty increases as a increases. Hence, insurers could use the parameter a to adopt a more or less stringent penalty for mid-term cancellations.

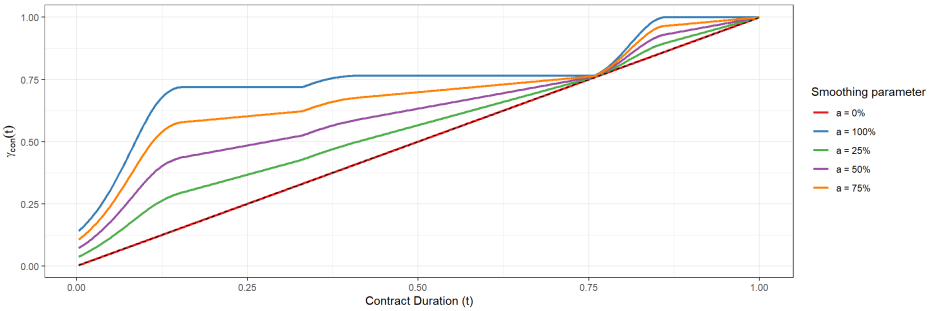
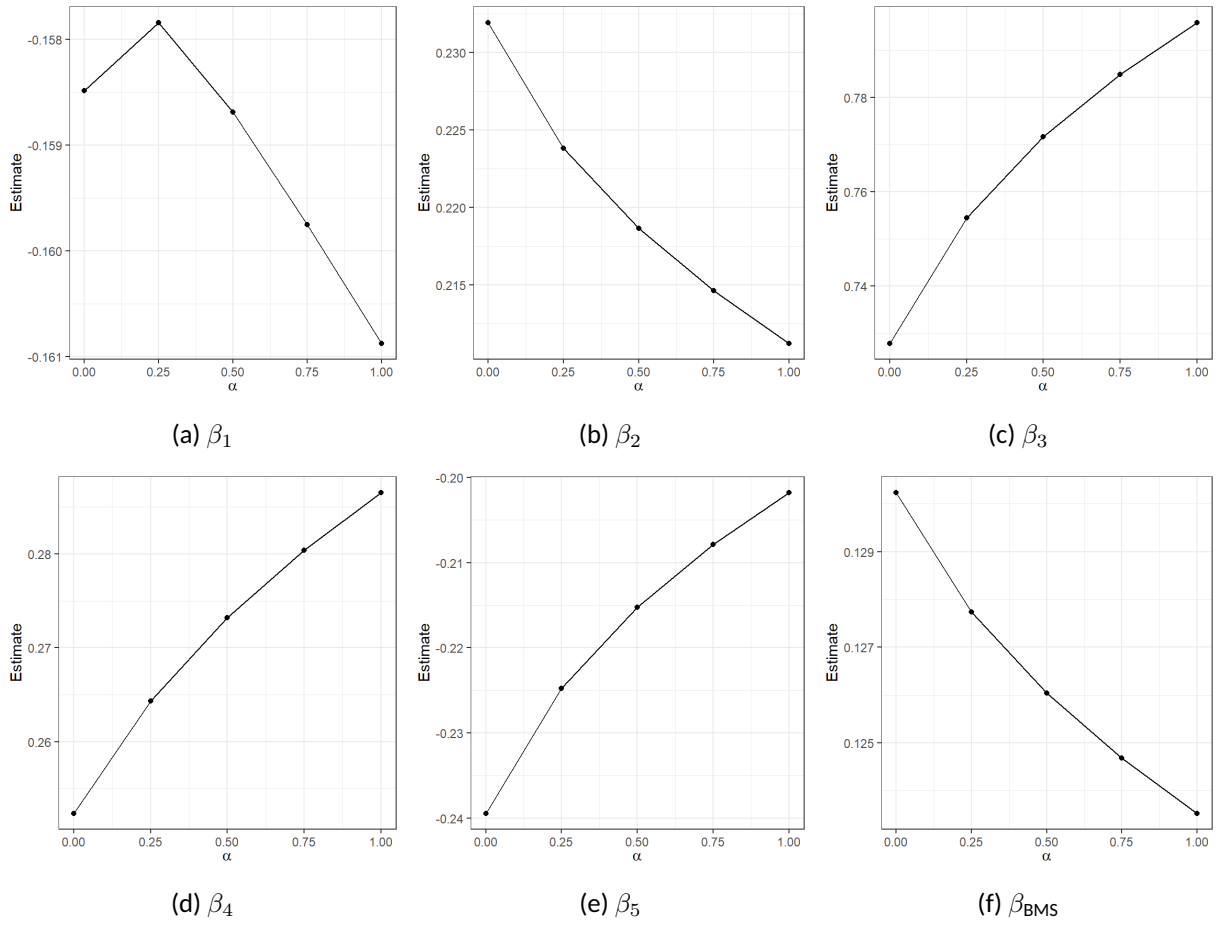


Figure 3.14 – Effect of the adjustment parameter a on the penalty functions

3.5.2.2 Tweedie Model with an Imposed Penalty Structure

It then becomes possible to apply a Tweedie model under constraints (C1) and (C2), together with the various penalty shapes proposed through $\gamma_{\text{adj}}(t, a)$. This is achieved by directly incorporating the function $\log(\gamma_{\text{adj}}(t, a))$ as an offset term in the mean structure of the Tweedie model. Formally, the mean parameter is specified as $\mu = \exp\big(\mathbf{x}^\top \boldsymbol{\beta} \;+\; \log(\gamma_{\text{adj}}(t, a))\big)$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_5, \beta_{\text{BMS}})^\top$ denotes the parameter vector.

For our numerical illustration, we restrict attention to the Exposure-Weighted specification, although the other approaches could also be applied. The panels in Figure 3.15 display the evolution of the components of the estimator $\widehat{\boldsymbol{\beta}}$ as a function of the adjustment parameter a, with the intercept excluded from the figure for readability. This representation facilitates the visualization of how the mid-term cancellation penalty affects the relativities associated with the covariates. In particular, by examining the estimated values of β_{BMS} across different values of a, one can observe that stronger penalties for early cancellations tend to reduce the influence of the claims experience. It can also be noted that the evolution of the estimators is not always monotonic, as illustrated, for instance, by the behavior of the estimate of β_1.



(a) β_1 (b) β_2 (c) β_3

(d) β_4 (e) β_5 (f) β_{BMS}

Figure 3.15 – Evolution of the estimated coefficients as a function of the adjustment parameter a.

Finally, Table 3.4 presents the predictive performance on the test dataset, based on the deviance and the area between the concentration and Lorenz curves for each value of a. As can be observed, these scores tend to improve as a increases. Therefore, $a = 100\%$ corresponds to the best predictive performance. It is also worth noting that $a = 100\%$ is precisely the value associated with the optimal penalty function.

	Smoothing penalty factor a				
Score	0%	25%	50%	75%	100%
Area	0.1237	0.0934	0.0678	0.0470	0.0306
Deviance	114.77	111.76	110.09	108.96	108.12

Table 3.4 – Predictive scores for different values of the smoothing penalty factor a

3.5.3 Loss Allocation through Mid-Term Cancellation Penalties

One important reason why policyholders may cancel their insurance coverage mid-term is the occurrence of a major claim—for example, when the insured vehicle becomes unusable and must be replaced. Within a pricing structure that penalizes mid-term cancellations, the insurer may design penalties that partially reflect the occurrence of such severe claims. In doing so, the insurer indirectly identifies large losses through the cancellation surcharge, thereby recovering part of the expected cost associated with the claim.

Let π^{Flex} denote the annual premium corresponding to $\gamma(1)$. Under the proposed model, the premium charged to a policyholder with exposure t is expressed as $\gamma(t)\,\pi^{\text{Flex}}$, where $\gamma(t)$ is the adjustment factor associated with mid-term cancellation behavior. This quantity naturally decomposes into two components :

1. the *pro rata* premium corresponding to the effective period of coverage, $t\pi^{\text{Flex}}$;
2. the additional amount associated with the cancellation *penalty*, $\rho(t)$.

As introduced in Section 3.3.2, this decomposition can be written as

$$\gamma(t)\,\pi^{\text{Flex}} \;=\; t\,\pi^{\text{Flex}} \;+\; \rho(t).$$

Empirically, approximately 11% of the total premium collected in the portfolio is attributable to this penalty component (10.89% in the training set and 10.80% in the test set). Moreover, as illustrated by the curves in Figure 3.9, the proposed ratemaking structure naturally imposes a higher surcharge on policyholders who cancel their contracts mid-term than the traditional approach. More precisely, conditional on the XO group (i.e., contracts canceled mid-term), the proposed model recovers approximately 15% more premium from

these policyholders compared with the traditional specification (15.28% in the training set and 15.01% in the test set).

Finally, using the testing dataset, Figure 3.16 displays the cumulative accumulation of premiums as a function of risk exposure t. The blue curve represents the proposed premium $\gamma(t)\,\pi$; the green curve corresponds to the traditional model; and the red curve shows the cumulative penalty $\rho(t)$, which is included in the blue curve. Comparison with the total claim costs (black curve) indicates that discrepancies in model fit remain—very likely caused by the pricing constraints C1 and C2 introduced in Section 3.3.2. This suggests that it may be beneficial for insurers to develop refined approaches capable of capturing a larger share of premium from policyholders who cancel their contracts mid-term.

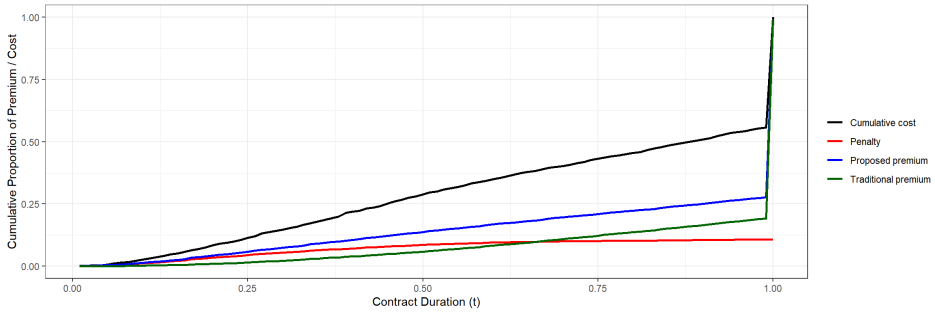


Figure 3.16 – Cumulative proportion of premiums and costs by exposure level

3.5.3.1 Competitive Market

A final perspective that deserves closer examination concerns the potential competitive advantage that such a model could offer to an insurer implementing it. The annual premium for full coverage is typically the main factor considered by policyholders when choosing an insurer, whereas the structure of mid-term cancellation penalties is unlikely to play a significant role in their decision. Consequently, increasing the mid-term cancellation penalties effectively leads to a reduction in the annual premium for full-year coverage.

To further illustrate this phenomenon, Figure 3.17 displays the ratio between the full-year premiums predicted by the proposed full-penalty approach ($a = 1$) and those produced by the traditional model, evaluated across all policyholder profiles generated from the covariates used in this study. The size of the green dots reflects the number of contracts observed in the test dataset for each profile, while the red dots (of which there are very few) correspond to unobserved profiles for which the model nevertheless produces a premium

estimate. The figure clearly shows that the proposed approach yields substantially lower annual premiums, with ratios consistently below one. In some cases, the reduction is quite substantial : for certain profiles that typically generated annual premiums around $300, the new pricing structure offers discounts of up to 25%.

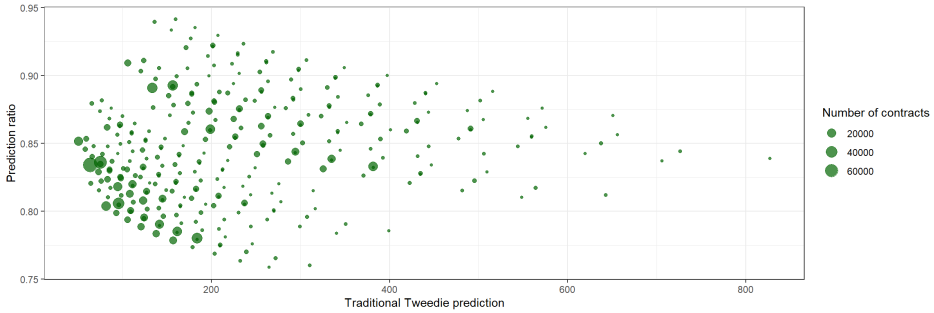


Figure 3.17 – Annual premium ratio between the flexible and the traditional models

3.5.4 Tweedie with Group-Specific Spline

Although the new ratemaking approaches that explicitly account for mid-term cancellations have produced promising results, there remains room for further investigation. In Section 3.2.4, we highlighted one policyholder characteristic that appears particularly relevant for understanding both loss costs and mid-term cancellations : the BMS level. As an initial step, Figure 3.18 presents a comparison between the predicted amounts from the proposed pricing model and the observed costs across BMS levels. Although the model fit appears reasonably good, it is well known that policyholders with higher BMS levels—corresponding to riskier profiles—tend to file more claims and are therefore more likely to cancel their policies mid-term following a loss event. This observation suggests that introducing group-specific spline functions based on the BMS level could represent a natural extension of the proposed framework. Such an approach would allow the penalty function to vary with policyholder risk characteristics, making it possible to allocate a slightly larger share of premium to higher-risk individuals who are statistically more prone to mid-term cancellations, thereby improving both fairness and profitability, as previously illustrated in Section 3.5.3.

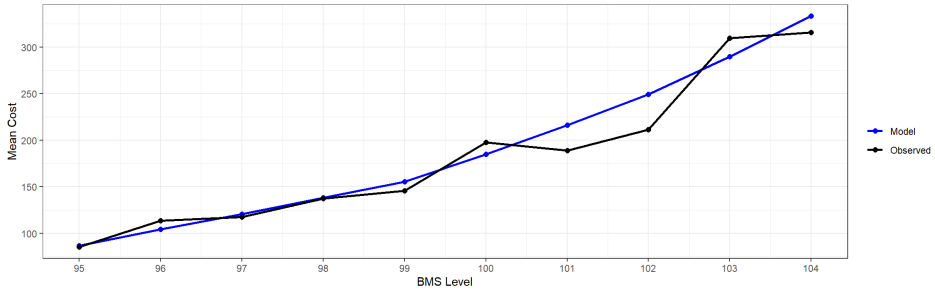


Figure 3.18 – Comparison between observed average costs and predicted values across BMS levels

Our objective is therefore to develop a more general function $\gamma(t)$ that explicitly incorporates policyholder characteristics—using the BMS level as a first step in our application. By allowing the mid-term cancellation penalty to vary with individual attributes, the model can better capture heterogeneity in policyholder behavior, enhance predictive accuracy, and provide a more refined understanding of the relationship between exposure and observed outcomes.

3.5.4.1 Grouping by BMS Level

To generalize the smoothing model of the risk exposure according to the policyholder's risk profile defined by BMS level, we aim to partition policyholders into groups based on their BMS level. More formally, let $k \in \{1, 2\}$ denote the BMS group associated with contract, and t its exposure. We then consider a Tweedie-type model with a logarithmic link, in which the penalty function depends nonlinearly on exposure and varies across BMS groups. Specifically, the mean parameter is specified as :

$$\mu = \exp\Big\{\mathbf{x}^\top \boldsymbol{\beta} + f_k(t)\Big\}, \quad k \in \{1, 2\}, \tag{3.5.1}$$

where $\mathbf{x}$ includes the six explanatory covariates used previously, $\boldsymbol{\beta}$ is the corresponding parameter vector, and $f_1(\cdot)$ and $f_2(\cdot)$ are two distinct smooth functions (penalized splines) of exposure t, estimated separately for groups 1 and 2.

In other words, instead of assuming a single function $\gamma(t)$ common to the entire portfolio, we allow the shape of the relationship between exposure and expected cost to vary across risk profiles as defined by the BMS level. This specification enables the model to capture differentiated penalty structures for low- and

high-risk policyholders, while maintaining the Tweedie generalized linear modeling framework.

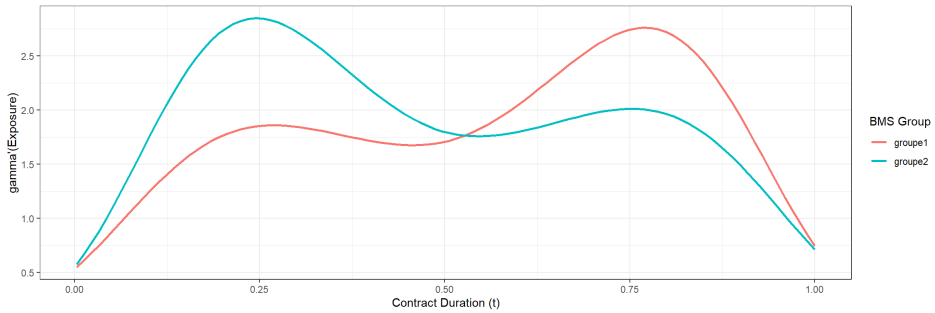


Figure 3.19 – Estimated smooth functions $\gamma_{\text{con}}(d)$ by BMS group

To identify the most appropriate grouping, we proceeded iteratively by dividing policyholders into two subsets according to their BMS level. Given the ordinal nature of the BMS scale—where lower levels (e.g., 95) correspond to safer drivers and higher levels (e.g., 104) to riskier ones—we searched for the optimal cut point that maximizes the difference between the two estimated smooth functions. By testing all eight possible divisions, the best segmentation, illustrated in Figure 3.19, was obtained using the following BMS groups :

— **Group 1** : levels $95 \leq$ BMS ≤ 99,

— **Group 2** : levels $100 \leq$ BMS ≤ 104.

The difference between the estimated smooth functions for the two groups, together with the corresponding standard errors, is shown in Figure 3.20. It is interesting to observe that the smooth function for Group 1 lies below that of Group 2 around mid-year, while it becomes higher thereafter. Except for the very beginning of the policy period and around mid-year, the smooth functions are statistically different, confirming that the effect of exposure varies significantly across BMS levels. Further attempts to subdivide either group did not yield additional statistically significant differences.

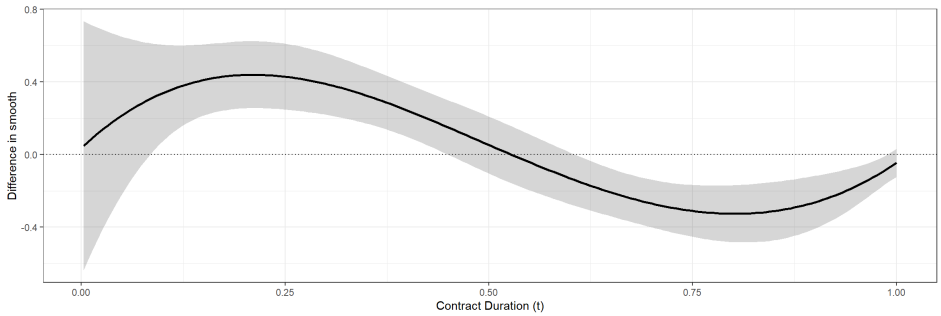


Figure 3.20 – Difference between smooth functions for BMS groups

It is particularly interesting to note how the division between the two groups of policyholders emerged. Group 1 corresponds to policyholders who have maintained at least one claim-free year of insurance, while Group 2 includes both new policyholders and those who have filed a claim in recent years. In other words, this implies that insurers would benefit from imposing stronger mid-term cancellation penalties (as illustrated by the blue spline in Figure 3.19) for newly insured policyholders. Given that new policyholders typically represent a higher administrative cost segment, this form of penalty is also operationally relevant. For future work, it would be worthwhile to generalize the proposed framework by allowing the mid-term cancellation penalty to depend on additional covariates, thereby capturing more complex interactions between individual characteristics and cancellation behavior.

The next step in the practical application of an approach with penalty curves differentiated by policyholder profile—in this case, by BMS level—should involve applying the ratemaking constraints C1 and C2, as introduced in Section 3.3.2 and implemented in our initial pricing model in Section 3.5.2, illustrated in Figure 3.12. As shown in Figure 3.19, enforcing a consistent penalty structure for policyholders in Group 2 would result in a very steep curve : the full annual premium would be charged after only one quarter of coverage. Although this may appear extreme, it is not entirely out of line with many one-year subscription-based products. In several industries—such as fitness memberships, telecommunications services, or annual software licenses—front-loaded fees or non-refundable upfront costs imply that a large fraction of the annual price is effectively charged within the first months of the contract.

For policyholders in Group 1, the penalty structure would be more similar to the one developed for the initial model. A smoothing of the two penalty curves using a factor a, as proposed in Section 3.5.2.1, could also be

considered to achieve a more balanced adjustment between the groups.

3.6 Conclusion

This paper has proposed a new family of Tweedie-based ratemaking models that explicitly account for mid-term policy cancellations, addressing an aspect of insurance pricing that is often overlooked in traditional frameworks. By introducing a flexible penalty function $\gamma(t)$ and alternative weighting schemes, the proposed approach provides a more realistic representation of the earned premium when coverage is interrupted before policy maturity. Through a combination of theoretical developments and empirical analysis, we have shown that the specification of both the mean and weight functions has a significant impact on parameter estimation and model stability.

Empirical results based on real automobile insurance data confirm that policyholders who cancel their coverage mid-term exhibit higher claim frequencies and severities, reinforcing the need to incorporate cancellation behavior into ratemaking models. The proposed constrained Tweedie models, together with the adjustment penalty parameter a, demonstrate that it is possible to reconcile statistical performance with practical interpretability, offering a flexible mechanism to penalize early cancellations while preserving fairness toward policyholders who maintain full-year coverage.

Beyond the models developed in this paper, several extensions could be explored to further enhance the flexibility and interpretability of the proposed framework. In particular, the penalty function $\gamma(t)$—currently estimated as a single smooth function of exposure—could itself be modeled as varying across policyholder characteristics. A first attempt in this direction was made by allowing distinct spline functions for different BMS levels. This specification implicitly assumes that mid-term cancellation penalties differ according to the policyholder's risk profile, which is consistent with actuarial intuition : safer drivers should face less stringent penalties than higher-risk individuals.

Preliminary results suggest that this extension is both statistically and conceptually promising. By conditioning the shape of $\gamma(t)$ on covariates such as the BMS level or driving experience, it becomes possible to capture heterogeneous behavioral responses to mid-term cancellations and to refine the alignment between premium structure, risk exposure, and observed claim patterns. Future research will aim to formalize this multi-dimensional modeling of the penalty spline and to assess its implications for risk classification, portfolio segmentation, and insurer competitiveness.

CONCLUSION

Dans cette thèse, nous nous sommes donné pour objectif de modéliser la charge totale d'un contrat d'assurance automobile sous la loi de Tweedie, en vue de proposer une prime d'assurance. Notre approche s'est principalement articulée autour des quatre défis identifiés dans l'introduction. Nous présentons ci-après quelques précisions sur les réponses apportées par cette thèse à ces défis.

Plus formellement, dans le chapitre 1, nous avons supposé une loi de Tweedie dont tous les paramètres, à l'exception du paramètre de poids w, ont fait l'objet d'une inférence statistique. Comme indiqué dans l'introduction, ce paramètre de poids peut être interprété comme une information a priori caractérisant la structure de la distribution. Nous avons choisi d'associer cette information a priori à l'exposition au risque. Afin d'estimer les autres paramètres de la distribution, nous avons adopté une modélisation conjointe de la charge totale de réclamations sous l'hypothèse d'une loi de Tweedie et du nombre total de réclamations sous l'hypothèse d'une loi de Poisson. L'avantage de cette approche est qu'elle permet d'obtenir une forme explicite de la densité de la loi de Tweedie dans le cas $1 < p < 2$. En effet, lorsque la modélisation est effectuée uniquement à partir d'une variable réponse supposée suivre une loi de Tweedie, la densité ne dispose pas d'expression fermée simple.

Ce premier chapitre met également en évidence l'intérêt de modéliser le paramètre de dispersion ϕ de la loi de Tweedie. En effet, bien que l'objectif principal en tarification soit l'estimation du paramètre de moyenne — correspondant à la prime — la dispersion constitue une information pertinente, notamment pour évaluer la stabilité des estimations et effectuer certains tests statistiques. Concernant le paramètre de variance p, il convient de noter que son estimation n'est pas aisée, en particulier dans un contexte où l'on utilise un modèle BONUS-MALUS SCALE (BMS). Toutefois, l'avantage de modéliser conjointement ces paramètres réside dans la possibilité d'utiliser des critères de sélection de modèle fondés sur la log-vraisemblance. C'est d'ailleurs ce qui nous a permis de comparer l'approche Tweedie et l'approche Compound Poisson-gamma (CPG) à l'aide de ces critères. Il convient également de noter que les chapitres 2 et 3 répondent eux aussi au premier défi mentionné dans l'introduction de cette thèse ; toutefois, dans ces chapitres, l'accent est davantage mis sur l'estimation de la prime que sur celle des autres paramètres de la loi de Tweedie. Par conséquent, moins d'attention y est accordée à l'estimation de p et de ϕ, bien que ces paramètres constituent également des composantes importantes de la loi de Tweedie.

Pour la prise en compte de l'expérience d'assurance à travers l'historique de réclamations, il convient de noter que nous suggérons l'approche Tweedie, bien que celle-ci présente une performance légèrement inférieure à celle de l'approche CPG. Son avantage réside dans le fait qu'elle est plus intuitive et plus naturelle pour modéliser la charge totale et calculer la prime d'un contrat. De plus, en raison de son lien structurel avec le nombre total de réclamations, l'impact de l'expérience d'assurance conserve la même signification lorsque le nombre total de réclamations est considéré comme variable tarifaire, approche qui a fait l'objet de nombreuses études dans la littérature. Cela ne signifie pas pour autant que l'approche CPG ne soit pas intéressante ; toutefois, si l'expérience d'assurance était quantifiée par une variable davantage liée au coût des réclamations, il est possible que des résultats encore meilleurs soient obtenus. Il convient également de préciser que le chapitre 3 contribue indirectement à ce défi, bien que celui-ci n'en constitue pas l'axe principal. Dans ce chapitre, la composante de dépendance — également connue sous le nom de niveau BMS — est calculée à partir de celle obtenue sous la loi de Tweedie dans le chapitre 1. Cette hypothèse nous a permis d'éviter de reproduire l'ensemble du processus d'estimation de la composante de dépendance, tout en conservant la cohérence du cadre méthodologique.

Les chapitres 2 et 3 répondent principalement à la problématique de la prise en compte de la durée de couverture d'un contrat dans le calcul de sa prime. Il convient de rappeler que le chapitre 2 repose sur une hypothèse de proportionnalité entre l'exposition au risque t et la prime. Ainsi, ce chapitre consiste principalement en une comparaison théorique et numérique des propriétés des estimations obtenues à partir des deux approches considérées. Bien que, dans le chapitre 3, nous ayons adopté une démarche similaire, l'avantage de ce chapitre est qu'il se positionne comme une généralisation du chapitre 2, en proposant une famille de modèles qui introduit une structure plus flexible entre la prime et l'exposition au risque qu'une simple hypothèse de proportionnalité. L'idée sous-jacente était de mieux prendre en compte les risques associés aux contrats qui se résilient en cours de terme. Ainsi, nous avons proposé une structure de pénalité liée aux résiliations, constituant un outil opérationnel permettant de limiter les transferts implicites de risque entre contrats.

Il convient également de noter que, dans l'ensemble des chapitres, une attention particulière est accordée à la notion de prime appropriée. Dans le chapitre 1, bien que les critères de sélection de modèle utilisés ne soient pas explicitement conçus dans cette optique, l'adoption du modèle BMS permet néanmoins de prendre en compte cette notion, notamment en limitant la sélection adverse par l'octroi de rabais aux assurés présentant un faible niveau de risque. Dans le chapitre 2, l'un des critères de sélection de modèle retenus est

la proximité à l'équilibre financier. Cet équilibre, généralement atteint dans les modèles linéaires généralisés (Generalized Linear Model (GLM)) lorsque le lien canonique est utilisé, correspond à la situation où les primes collectées financent le coût total des réclamations d'un portefeuille. Enfin, dans le chapitre 3, nous allons plus loin en introduisant la notion d'équilibre local, selon laquelle, pour chaque sous-ensemble de contrats, la somme des primes doit couvrir le coût total des réclamations du sous-portefeuille correspondant. Bien que cet équilibre ne soit atteint que sous certaines conditions — qui ne sont pas nécessairement vérifiées par les modèles proposés — la comparaison de leur proximité à cet équilibre nous a permis d'identifier un modèle adéquat répondant à cette considération pratique de la tarification.

À la lumière de l'ensemble de ces résultats, nous pouvons affirmer que l'objectif poursuivi dans cette thèse a été atteint. Nous avons proposé un cadre cohérent de modélisation de la charge totale sous la loi de Tweedie, intégrant l'expérience d'assurance, la durée de couverture et les considérations d'équilibre financier. Toutefois, plusieurs questions demeurent ouvertes, notamment quant à la prise en compte de la durée de couverture dans le calcul de la prime. Bien que l'hypothèse de proportionnalité entre la prime et l'exposition au risque ne semble pas adéquate au regard des données analysées, cela ne remet pas en cause l'intérêt des approches proposées sous la loi de Tweedie. Une piste de recherche naturelle consisterait à établir théoriquement la dominance de l'approche ratio par rapport à l'approche offset lorsque des critères d'équilibre financier ou d'équilibre local sont explicitement considérés. Une autre perspective serait de modéliser l'exposition au risque comme une variable aléatoire et d'analyser l'impact de cette hypothèse sur les modèles proposés, notamment dans la famille généralisée introduite au chapitre 3. Une telle extension permettrait d'intégrer, dès le calcul de la prime, l'incertitude liée à la durée effective de couverture en tenant compte du risque de résiliation à travers une distribution appropriée de l'exposition au risque. Toutefois, cette approche devrait être mise en œuvre avec prudence, car elle pourrait engendrer des effets indésirables, notamment un transfert implicite de risque conduisant à une majoration de prime pour certains assurés à faible risque.

Ces travaux contribuent ainsi à enrichir le cadre méthodologique de la tarification automobile sous la loi de Tweedie, tout en ouvrant la voie à de nouveaux développements théoriques et appliqués.

ANNEXE A

A.1 Estimated Coefficients of the Mean Parameter

	Standard		BMS		Standard		BMS	
Parameter	Poisson	Gamma	Poisson	Gamma	CPG	Tweedie	CPG	Tweedie
β_0	-4.409	8.770	-13.541	6.227	4.361	4.330	-7.314	-6.626
β_1	0.135	0.067	0.121	0.062	0.202	0.207	0.183	0.189
β_2	0.031	-0.078	0.044	-0.068	-0.047	-0.047	-0.024	-0.032
β_3	-0.154	0.000	-0.144	-	-0.154	-0.157	-0.144	-0.145
β_4	0.192	-	0.220	0.005	0.192	0.187	0.224	0.225
β_5	0.474	0.109	0.405	0.086	0.583	0.587	0.491	0.505
β_6	0.548	0.186	0.528	0.184	0.735	0.735	0.711	0.716
β_7	0.231	0.098	0.188	0.085	0.329	0.342	0.273	0.283
β_8	-0.128	-0.150	-0.054	-0.113	0.199	0.238	0.160	0.176
β_9	0.181	0.018	0.160	-	0.199	0.238	0.160	0.176
β_{10}	0.343	0.104	0.316	0.084	0.447	0.486	0.400	0.417
β_{11}	0.378	0.160	0.347	0.136	0.538	0.576	0.482	0.499
β_{12}	0.223	0.190	0.191	0.165	0.413	0.450	0.356	0.371
β_{13}	-0.095	0.190	-0.151	0.160	0.095	0.130	0.009	0.189
β_{14}	0.106	0.150	0.042	0.126	0.256	0.298	0.168	-
β_{15}	-	-	0.094	0.026	-	-	0.094	0.112
β_{16}					-	-	0.026	-

Table A.1 – Estimated parameters

A.2 Residual Analysis

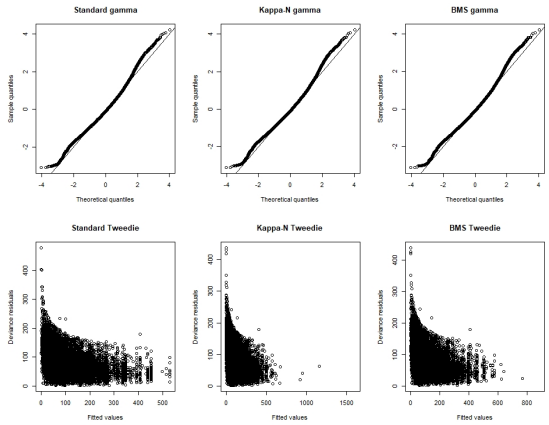


Figure A.1 – Cox-Snell residuals for severity and Anscombe residuals for loss cost

ANNEXE B

B.1 Calculation of Asymptotic Covariance Matrices

This appendix summarizes results from White (1982) that are directly relevant to Proposition 2.4. We provide additional details on the calculation of the asymptotic covariance matrices of the parameter estimators in the Offset and Ratio approaches. We begin by recalling the expression of the response variance in both approaches :

$$\mathrm{Var}[Z_i \mid \mathbf{X}_i, w_i] = \frac{\phi}{w_i}\,\zeta_i^p, \qquad i = 1, \ldots, n,$$

where $\zeta_i = \exp\{\mathbf{X}_i^\top \boldsymbol{\beta}\}$, and $w_i = t_i^{2-p}$ under the Offset approach while $w_i = t_i$ under the Ratio approach. For convenience, we also introduce

$$\boldsymbol{D}(W) = \mathrm{diag}\Big(w_i \zeta_i^{2-p}\Big)_{i=1,\ldots,n},$$

with $\boldsymbol{W} = (w_1, \ldots, w_n)^\top$. Following White (1982), the asymptotic covariance matrix is

$$\Sigma = A^{-1} B A^{-1},$$

where $A = -\frac{1}{\phi}\boldsymbol{X}^\top \boldsymbol{D}(W)\boldsymbol{X}$ and $B = \mathbb{E}\big[\nabla(\boldsymbol{\beta})\nabla(\boldsymbol{\beta})^\top\big]$, with score vector

$$\nabla(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} w_i \frac{Z_i - \zeta_i}{\zeta_i^{p-1}}\,\mathbf{X}_i.$$

Then B is given by

$$\begin{aligned}
B &= \frac{1}{\phi^2}\mathrm{E}\left[\left(\sum_{i=1}^{n} w_i \frac{Z_i - \zeta_i}{\zeta_i^{p-1}} x_{i,j_1}\right)\left(\sum_{i=1}^{n} w_i \frac{Z_i - \zeta_i}{\zeta_i^{p-1}} x_{i,j_2}\right)\right]_{j_1, j_2 = 0, \ldots, q} \\
&= \frac{1}{\phi^2}\mathrm{E}\left[\sum_{i=1}^{n} w_i^2 \frac{(Z_i - \zeta_i)^2}{\zeta_i^{2p-2}} x_{i,j_1} x_{i,j_2}\right]_{j_1, j_2 = 0, \ldots, q} \\
&= \frac{1}{\phi^2}\sum_{i=1}^{n} w_i^2 \frac{\mathrm{Var}[Z_i \mid \mathbf{X}_i, w_i]}{\zeta_i^{2p-2}} x_{i,j_1} x_{i,j_2} \\
&= \frac{1}{\phi^2}\sum_{i=1}^{n} w_i^2 \frac{\frac{\phi}{w_i}\zeta_i^p}{\zeta_i^{2p-2}} x_{i,j_1} x_{i,j_2} \\
&= \frac{1}{\phi}\sum_{i=1}^{n} w_i \zeta_i^{2-p} x_{i,j_1} x_{i,j_2} \\
&= \frac{1}{\phi}\boldsymbol{X}^\top \boldsymbol{D}(W)\boldsymbol{X} = -A.
\end{aligned}$$

Hence, $B = -A$, and the asymptotic covariance matrix reduces to $\Sigma = -A^{-1}$.

ANNEXE C

C.1 Weight iterations

Figure C.1 illustrates the evolution of the weight function ω_i across the successive iterations of the estimation algorithm described in Section 3.4. The figure provides a graphical representation of how the spline-based approximation of $\gamma(d_i)$ stabilizes as the iterative procedure updates both the mean and weight components. Each curve corresponds to a given iteration k, starting from the initial Exposure Weighted Model weighting scheme. A small horizontal offset was applied to each curve to improve visual distinction. The rapid convergence observed after only a few iterations confirms the numerical stability and internal consistency of the Gamma Weighted estimation process.

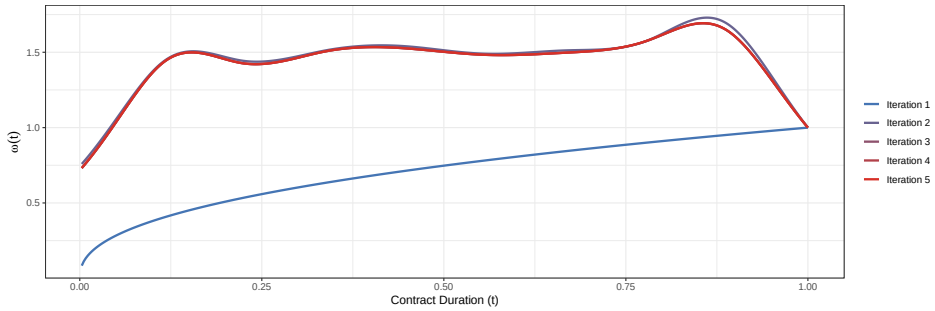


Figure C.1 – Evolution of the weight function ω_i over the iterative estimation process

BIBLIOGRAPHIE

Adillon, R., Jorba, L. et Mármol, M. (2020). Modal interval probability : Application to bonus-malus systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *28*(05), 837–851.

Bermúdez, L., Guillén, M. et Karlis, D. (2018). Allowing for time and cross dependence assumptions between claim counts in ratemaking models. *Insurance : Mathematics and Economics*, *83*, 161–169.

Boucher, J.-P. (2022). Bonus-malus scale models : creating artificial past claims history. *Annals of Actuarial Science*, p. 1–27. http://dx.doi.org/10.1017/S1748499522000100

Boucher, J.-P. et Coulibaly, R. (2024). Bonus-malus scale premiums for Tweedie's compound poisson models. *Annals of Actuarial Science*, 1–25. http://dx.doi.org/10.1017/S1748499524000113

Boucher, J.-P. et Coulibaly, R. (2026). Comparison of offset and ratio weighted regressions in tweedie models with application to mid-term cancellations. *European Actuarial Journal*. http://dx.doi.org/10.1007/s13385-026-00452-z

Boucher, J.-P. et Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin : The Journal of the IAA*, *44*(3), 587–612.

Boyd, S. P. et Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Casella, G. et Berger, R. (2024). *Statistical inference*. CRC Press.

Delong, Ł., Lindholm, M. et Wüthrich, M. V. (2021). Making Tweedie's compound Poisson model more accessible. *European Actuarial Journal*, *11*(1), 185–226.

Denuit, M., Charpentier, A. et Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance : Mathematics and Economics*, *101*, 485–497.

Denuit, M., Huyghe, J., Trufin, J. et Verdebout, T. (2024). Testing for auto-calibration with lorenz and concentration curves. *Insurance : Mathematics and Economics*, *117*, 130–139.

Denuit, M., Maréchal, X., Pitrebois, S. et Walhin, J.-F. (2007a). *Actuarial modeling of claim counts : Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Denuit, M., Maréchal, X., Pitrebois, S. et Walhin, J.-F. (2007b). *Actuarial modelling of claim counts : Risk classification, credibility and bonus-malus systems*. Wiley, West Sussex.

Denuit, M., Sznajder, D. et Trufin, J. (2019). Model selection based on lorenz and concentration curves, gini indices and convex order. *Insurance : Mathematics and Economics*, *89*, 128–139.

Denuit, M. et Trufin, J. (2019). *Effective statistical learning methods for actuaries I*. Springer, Cham.

Denuit, M. et Trufin, J. (2024). Convex and lorenz orders under balance correction in nonlife insurance pricing : Review and new developments. *Insurance : Mathematics and Economics*.

Denuit, M., Trufin, J. et Verdebout, T. (2025). Comparison of predictors' performance in insurance pricing : testing for bregman dominance based on murphy diagrams. *European Actuarial Journal*, *15*(2), 493–504.

Frees, E. W. (2014). Frequency and severity models. In E. W. Frees, R. A. Derrig, et G. Meyers (dir.), *Predictive Modeling Applications in Actuarial Science*, volume 1 138–164. Cambridge : Cambridge University Press

Frees, E. W., Derrig, R. A. et Meyers, G. (dir.) (2014). *Predictive Modeling Applications in Actuarial Science*. Cambridge University Press.

Gneiting, T. et Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). The elements of statistical learning. *Cited on*, *33*.

Hastie, T., Tibshirani, R. et Wainwright, M. (2015). *Statistical Learning with Sparsity : The Lasso and Generalizations*. CRC Press.

Horn, R. A. et Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Jeong, H. et Valdez, E. A. (2020). Predictive compound risk models with dependence. *Insurance : Mathematics and Economics*, *94*, 182–195.

Jong, P. D. et Heller, G. Z. (2008). *Generalized Linear Models for Insurance Datas*. Cambridge University Press. `http://dx.doi.org/https://doi.org/10.1017/CBO9780511755408`

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, *49*(2), 127–145.

Jørgensen, B. (1997). *The theory of dispersion models*. CRC Press.

Lee, G. Y. et Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance : Mathematics and Economics*, *87*, 115–129.

Lemaire, J. (2012). *Bonus-malus systems in automobile insurance*, volume 19. Springer science.

Nelder, J. A. et Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A : Statistics in Society*, *135*(3), 370–384.

Oh, R., Shi, P. et Ahn, J. Y. (2020). Bonus-malus premiums under the dependent frequency-severity modeling. *Scandinavian Actuarial Journal*, *2020*(3), 172–195.

Pechon, F., Denuit, M. et Trufin, J. (2019). Multivariate modelling of multiple guarantees in motor insurance of a household. *European Actuarial Journal*, *9*, 575–602.

Shi, P. et Valdez, E. A. (2014). Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal*, *2014*(2), 159–179.

Shi, P. et Yang, L. (2018). Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association*, *113*(521), 122–133.

Simon, P.-A., Trufin, J. et Denuit, M. (2023). Bivariate poisson credibility model and bonus-malus scale for claim and near-claim events. *North American actuarial journal*.

Turcotte, R. et Boucher, J. P. (2023). Gamlss for longitudinal multivariate claim count models. *North American Actuarial Journal*, 1–24.

Tweedie, M. C. *et al.* (1984). An index which distinguishes between some important exponential families. Dans *Statistics : Applications and new directions : Proc. Indian statistical institute golden Jubilee International conference*, volume 579, 579–604.

Verschuren, R. M. (2021). Predictive claim scores for dynamic multi-product risk classification in insurance. *ASTIN Bulletin : The Journal of the IAA*, *51*(1), 1–25.

Villacorta Iglesias, P. J., González-Vila Puchades, L. et de Andrés-Sánchez, J. (2021). Fuzzy markovian bonus-malus systems in non-life insurance. *Mathematics*, *9*(6), 647. `http://dx.doi.org/10.3390/math9060647`

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica : Journal of the econometric society*, 1–25.

Wood, S. N. (2017). *Generalized additive models : an introduction with R*. Chapman and Hall/CRC.

Wüthrich, M. V. et Merz, M. (2023). *Statistical foundations of actuarial learning and its applications*. Springer, Cham.