Risk Loadings in Classification Ratemaking

Liang Yang[*] Zhengxiao Li[†] Shengwang Meng[‡]

Abstract

The risk premium of a policy is the sum of the pure premium and the risk loading. In the classification ratemaking process, generalized linear models are usually used to calculate pure premiums, and various premium principles are applied to derive the risk loadings. No matter which premium principle is used, some risk loading parameters should be given in advance subjectively. To overcome this subjective problem and calculate the risk premium more reasonably and objectively, we propose a top-down method to calculate these risk loading parameters. First, we implement the bootstrap method to calculate the total risk premium of the portfolio. Then, under the constraint that the portfolio's total risk premium should equal the sum of the risk premiums of each policy, the risk loading parameters are determined. During this process, besides using generalized linear models, three kinds of quantile regression models are also applied, namely, traditional quantile regression model, fully parametric quantile regression model, and quantile regression model with coefficient functions. The empirical result shows that the risk premiums calculated by the method proposed in this study are more coherent and can reasonably differentiate the heterogeneity of different risk classes.

Keywords: classification ratemaking; risk premium; quantile regression; generalized linear model

1 Introduction

Calculating the risk premium is a prime objective for insurance pricing in non-life actuarial science. The risk premium consists of two parts: the pure premium, which is used to compensate the expected value of future losses, and the risk loading, which is used to cover the excess part of future losses over the pure premium. To estimate the risk loading correctly and at the same time allow classification by tariff features, in this study, we develop a new general framework to calculate the individual risk premiums, including risk loadings, based on an arbitrary set of covariates.

A rich variety of premium principles has been proposed in the actuarial literature for predicting the risk premium of individual policies, for example, Bühlmann (1970), Mack (1997), Wang et al. (1997), Kudryavtsev (2009), and Heras et al. (2018). The standard approach for predicting the risk premium involves a separate analysis of two parts of the risk premium: the pure premium and the risk loading. The traditional approach is based on generalized linear models (GLMs) (De Jong and Heller, 2008), which provide estimates of expected losses of individual polices given a number of risk

[*] School of Insurance, Southwestern University of Finance and Economics, Chengdu, China

[†] School of Insurance, University of International Business and Economics, Beijing, China

[‡] Corresponding author, Email: mengshw@ruc.edu.cn. Center for Applied Statistics, Renmin University of China, Beijing, China

factors. Risk loading is derived in the traditional approach by applying various premium principles, for example, the expected value premium principle, standard deviation premium principle, and Wang premium principle.

Assuming that random variable Y_i denotes the aggregate claim amount for individual policy $i, i = 1, \cdots, N$, the risk premium of policy i can be expressed as a distortion function $H(Y_i)$ of the random variable Y_i. In the expected value premium principle, the risk premium equals the pure premium plus a percentage of the pure premium, that is,

$$H(Y_i) = \mathbb{E}(Y_i) + \varphi \mathbb{E}(Y_i), \tag{1.1}$$

where $\varphi > 0$ denotes the risk loading parameter and $\varphi \mathbb{E}(Y_i)$ denotes the corresponding risk loading. In the standard deviation premium principle, the risk premium equals the pure premium plus a percentage of the standard deviation, that is,

$$H(Y_i) = \mathbb{E}(Y_i) + \varphi \sqrt{\operatorname{Var}(Y_i)}. \tag{1.2}$$

An alternative approach for predicting risk premium is to consider the risk premium as a whole by applying the value at risk (VaR) premium principle and Wang premium principle; see, for example Wang (1995, 2000), Wang et al. (1997), and Kudryavtsev (2009). Based on the Wang premium principle, the risk premium is expressed as follows:

$$H(Y_i) = \int_0^\infty \Phi\left[\Phi^{-1}\left(S_{Y_i}(y)\right) + \rho\right] \mathrm{d}y, \tag{1.3}$$

where Φ and Φ^{-1} denote standard normal cumulative distribution function and its inverse function, respectively; S_{Y_i} represents the survival function of aggregate claim amount, and $\rho \in \mathbb{R}$ denotes a risk factor.

The VaR premium principle in quantile regression for ratemaking is first discussed in Kudryavtsev (2009). The risk premium is calculated as a quantile of the aggregate claim amount, as follows:

$$H(Y_i) = Q_{Y_i}(\tau) = \inf\{u \in \mathbb{R} : \ F_{Y_i}(u) \geq \tau\}, \tag{1.4}$$

where $Q_{Y_i}(\tau)$ denotes the quantile of the aggregate claim amount and τ is a given quantile level, such as 95% or 99%. Risk loading is denoted as $Q_{Y_i}(\tau) - \mathbb{E}(Y_i)$, which is expressed as the difference between the quantile and the pure premium. This premium principle explains the needs of risk loading quite well, as it estimates the maximum possible losses that an individual policy may incur with a given probability $1 - \tau$ during the forecasting period.

Following the VaR premium principle, the quantile premium principle for classification ratemaking is proposed by Heras et al. (2018), and the corresponding risk premium is calculated as follows:

$$H(Y_i) = \mathbb{E}(Y_i) + \varphi\left[Q_{Y_i}(\tau) - \mathbb{E}(Y_i)\right], \tag{1.5}$$

where $Q_{Y_i}(\tau)$ denotes the τ-th quantile of the aggregate claim amount, φ is the risk loading parameter, and $\varphi\left[Q_{Y_i}(\tau) - \mathbb{E}(Y_i)\right]$ represents the risk loading, which is the difference between the τ-th quantile of the aggregate claim amount and the pure premium. The main difference between the VaR premium principle in Eq.(1.4) and the quantile premium principle in Eq.(1.5) is that the risk loading in the quantile premium principle is adjusted by risk loading parameter φ.

Recently, Baione and Biancalana (2019) proposes a two-part quantile premium principle, that is,

$$H\left(Y_i\right)=\left(1-p_i\right)Q_{Y_i^*}\left(\tau\right), \tag{1.6}$$

where $Q_{Y_i^*}\left(\tau\right)$ denotes the τ-th quantile of aggregate claim amount given that at least one claim has been incurred and $1-p_i$ denotes the probability of incurring at least one claim.

In actuarial practice, some parameters, namely, φ, ρ, and τ in Eqs.(1.1)-(1.6), which are called risk loading parameters in this study, need to be determined in advance. To estimate the risk loading parameters, Bühlmann (1985) proposes a top-down method for insurance companies by first controlling the probability of ruin at the acceptable level in advance and then imposing this stability criterion regarding yield of invested capital. This allows insurance companies to find a total premium to be charged for the whole portfolio and then split it in a fair way among all the individual risks.

While the top-dwon method is well developed, see for example Cossette et al. (2012) and Heras et al. (2018), the use of covariate information in order to estimate the risk loading parameters through generalized linear models and quantile regression models has received much less attention. Following this line of study, Baione and Biancalana (2019) extend the work of Heras et al. (2018) by developing a down-top-down method for risk premium calculation in classification ratemaking. They first apply two-part GLMs and expected value premium principle to calculate the risk premium for each policy at the individual level and then obtains the total risk premium of the whole portfolio by simply aggregating all individual policys risk premium. Finally, the risk loading parameter is defined such that the total risk premiums for all policies are sufficient to cover the total expected losses. However, the above approach is debatable because it ignores the risk diversification effect of combining all individual policies, which might result in a over-estimated total risk premium of the whole portfolio. Moreover, the total risk premium often relies on the distribution assumption of GLMs at the individual level; for example, Baione and Biancalana (2019) apply a gamma (GA) regression to fit the non-zero aggregate claim amounts, which might be not very appropriate for practical insurance portfolios (Heller et al., 2006; Eling, 2012; Laudagé et al., 2019).

Our work is motived by the recent works of Heras et al. (2018) and Baione and Biancalana (2019). We extend this branch of the literature by developing a more general top-down framework to calculate the risk loading parameters. We first derive the total risk premium of the portfolio by implementing the bootstrap method, thereby allowing us to obtain the entire distribution of the total risk premium at the collective level, instead of exploring the distribution at individual level. Given an acceptable confidence level, this approach provides a useful tool for estimating the VaR of a portfolio.

Our method permits estimating risk loading parameters uniquely for various premium principles at the individual level. In this approach, the total risk premium is distributed to the individual policies based on the risk contribution of each policy, so that the sum of the risk premiums of all individual policies is equal to the total risk premium of the whole portfolio, which is proved to be an efficient method in ratemaking by Bühlmann (1985). The risk premiums of different tariff classes can be estimated by either GLMs or quantile regression models incorporating into the covariate information. For comparison, GLMs is used as a benchmark. Traditional quantile regression, fully parametric quantile regression, and quantile regression with coefficient functions are constructed to calculate the risk premium of each individual policy.

Thus, our approach has two advantages: (1) it controls the probability that the aggregate claim amount of the entire portfolio exceeds the total risk premium to an acceptable level; (2) it provides a general framework to determine risk loading parameters objectively for all types of models, such as two-part GLMs and two-part quantile regression models.

The remainder of the article is structured as follows. Sections 2 and 3 summarize the methods to calculate the risk premium based on two-part GLMs and two-part quantile regression models, respectively, at the individual level. Section 4 presents an analysis of the calculation of the total risk premium of a portfolio and its allocation to individual policies. Section 5 applies the proposed method to an empirical data set. Section 6 summarizes and concludes.

2 Risk Premiums Based on Two-Part GLMs

Suppose an insurance portfolio contains N policies, R_i indicates whether or not policy i has a claim submitted, Y_i represents its aggregate claim amount, w_i denotes its exposure, and $\boldsymbol{x}_i$ stands for a vector of covariates $(i = 1, \cdots, N)$.

In actuarial practice, the observed aggregate claim amounts of a portfolio usually have a probability mass at zero. In this study, we first implement the two-part GLMs to accommodate the probability mass at zero. In a two-part GLMs framework, the zero component models the probability of incurring no claim, and the continuous component models the aggregate claim amount given that at least one claim has been incurred. It is a common practice to separate claim probability and non-zero aggregate claim amount in pricing non-life insurance contracts; see, for example, Frees (2009) and Frees et al. (2013).

For claim probability, we assume that R_i follows the binomial distribution with parameter $1 - p_i$, and consider the conventional logistic regression model:

$$\text{logit}\left[\frac{1-p_i}{w_i}\right] = \boldsymbol{x}_i^R \boldsymbol{\alpha}, \tag{2.1}$$

where $\text{logit}(x) = \log\left(x/(1-x)\right)$ is the logit function, $\boldsymbol{x}_i^R = (1, x_{i1}^R, \cdots, x_{ik}^R)$ represents the $(k+1)$-dimensional vector of covariates, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_k)'$ denotes the corresponding regression coefficients to be estimated. The left-hand side of Eq.(2.1) is the log odds ratio per exposure. The logistic regression model in Eq.(2.1) is corrected for risk exposure w_i; see De Jong and Heller(2008). Correspondingly, the probability of at least one claim occurring can be obtained by $1 - p_i = w_i \frac{\exp(\boldsymbol{x}_i^R \boldsymbol{\alpha})}{1+\exp(\boldsymbol{x}_i^R \boldsymbol{\alpha})}$.

For the non-zero aggregate claim amount, we employ Gamma distribution (GA) and inverse Gaussian distribution (IG) to model its skewness and heavy tail (see Appendix A for further details). Using the log link function, we obtain the following regression model for the mean parameter of GA and IG distribution:

$$\log(\mu_i) = \boldsymbol{x}_i^\mu \boldsymbol{\beta}, \tag{2.2}$$

where $\boldsymbol{x}_i^\mu = (1, x_{i1}^\mu, \cdots, x_{ik}^\mu)$ represents a $(k+1)$-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)'$ denotes the corresponding regression coefficients to be estimated. The mean parameter is obtained by $\mu_i = \exp(\boldsymbol{x}_i^\mu \boldsymbol{\beta})$.

Under the assumption of GA and IG distribution, the pure premium of policy i is given by

$$\mathbb{E}\left[Y_i;\ p_i,\ \mu_i\right] = (1 - p_i)\,\mu_i. \tag{2.3}$$

Specifically, we derive the risk premium of policy i by applying the expected value premium principle in Eq.(1.1):

$$H\left(Y_i; p_i, \mu_i\right) = (1 - p_i)\,\mu_i + \varphi\,(1 - p_i)\,\mu_i, \tag{2.4}$$

where φ is the risk loading parameter in the expected value premium principle.

Similarly, under the standard deviation premium principle in Eq.(1.2), the risk premium of policy i is given by

$$H\left(Y_i;p_i,\mu_i,\sigma\right)=\begin{cases}\left(1-p_i\right)\mu_i+\varphi\mu_i\sqrt{\left(1-p_i\right)\left(p_i+\sigma^2\right)}, & Y_i|R_i=1\sim\text{GA}\\ \left(1-p_i\right)\mu_i+\varphi\mu_i\sqrt{\left(1-p_i\right)\left(p_i+\mu_i\sigma^2\right)}, & Y_i|R_i=1\sim\text{IG},\end{cases} \tag{2.5}$$

where σ is the scale parameter in GA and IG distribution, and φ is the risk loading parameter in the standard deviation premium principle.

The risk premium by applying the Wang transform in Eq.(1.3) is given by

$$H\left(Y_i;p_i,\mu_i,\sigma\right)=\int\limits_0^\infty\Phi\left[\Phi^{-1}\left(1-F_{Y_i}\left(y;p_i,\mu_i,\sigma\right)\right)+\rho\right]\,\mathrm{d}y, \tag{2.6}$$

where Φ is the standard normal cumulative distribution function, and Φ^{-1} is its inverse function, $F_{Y_i}\left(y\right)$ denotes the cumulative distribution function of aggregate claim amounts by applying two-part GLMs, and $\rho\in\mathbb{R}$ represents the risk aversion parameter of the Wang premium principle.

3 Risk Premiums Based on Two-part Quantile Regression Models

3.1 Risk Premiums Based on Two-part Quantile Regression Models

To assess the risk premium of individual policies, it is common practice to implement a quantile regression framework, which is introduced by Kudryavtsev (2009) and applied in actuarial practice, see Heras et al. (2018) and Baione and Biancalana (2019).

Following the two-part quantile premium principle proposed by Baione and Biancalana (2019), in this study the risk premium of policy i is simply given by

$$H\left(Y_i\right)=\left(1-p_i\right)Q_{Y_i^*}\left(\tau_i^*\left|\boldsymbol{x}_i\right.\right), \tag{3.1}$$

where $\boldsymbol{x}_i$ stands for a vector of covariates, $p_i=\Pr\left(Y_i=0|\boldsymbol{x}_i\right)$ denotes the probability that policy i incurs no claim , $Y_i^*=Y_i|Y_i>0$ represents the non-zero aggregate claim amount given that policy i incurs at least one claim, and $Q_{Y_i^*}\left(\tau_i^*\left|\boldsymbol{x}_i\right.\right)$ denotes the τ_i^*-th quantile of Y_i^*.

It is clear that

$$F_{Y_i}\left(y_i|\boldsymbol{x}_i\right)=\Pr\left(Y_i=0|\boldsymbol{x}_i\right)+\left[1-\Pr\left(Y_i=0|\boldsymbol{x}_i\right)\right]F_{Y_i^*}\left(y_i|\boldsymbol{x}_i\right), \tag{3.2}$$

which means that τ-th quantile function of Y_i is equivalent to τ_i^*-th quantile function of Y_i^*, that is

$$Q_{Y_i^*}\left(\tau_i^*\left|\boldsymbol{x}_i\right.\right)=Q_{Y_i}\left(\tau\left|\boldsymbol{x}_i\right.\right), \tag{3.3}$$

where

$$\tau_i^*=\frac{\tau-p_i}{1-p_i}, \tag{3.4}$$

for real number τ in the interval $[0,1]$, which denotes the risk loading parameter in two-part quantile premium principle and need to be given in advance. It is worth noting that though Baione and Biancalana (2019) suggests fixing a unique quantile level τ_i^* for Y_i^* associated with the i-th risk class (see Eq.(1.6)), in this study, we suggests fixing a unique quantile level $F_{Y_i}(y_i|\boldsymbol{x}_i)$ for all individual polices (see Eq.(3.1)), which follows the same assumption as the work of Heras et al. (2018). Therefore, it is quite important to directly control the risk loading by choosing the quantile level τ in the classification ratemaking process .

Using Eqs.(3.1) and (3.3), the risk premium of policy i can also be obtained as

$$H\left(Y_i\right) = \left(1-p_i\right) Q_{Y_i}\left(\tau \left| \boldsymbol{x}_i \right.\right) = \left(1-p_i\right) Q_{Y_i^*}\left(\tau_i^* \left| \boldsymbol{x}_i \right.\right). \tag{3.5}$$

In non-life ratemaking, the log link function is quite popular because it is well connected with the multiplicative framework, see Mack (1997) and Kudryavtsev (2009), among others. Similar to the GLMs in Eq.(2.2), we apply the quantile regression model by using the log link function, that is given by

$$\log Q_{Y_i^*}\left(\tau_i^* \left| \boldsymbol{x}_i^Q \right.\right) = \boldsymbol{x}_i^Q \boldsymbol{\gamma}^{\tau_i^*}, \tag{3.6}$$

where $\boldsymbol{x}_i^Q = (1, x_{i1}^Q, \cdots, x_{ik}^Q)$ represents the $(k+1)$-dimensional vector of covariates in the quantile regression and $\boldsymbol{\gamma}^{\tau_i{}^*} = (\gamma_0^{\tau_i^*}, \gamma_1^{\tau_i^*}, \cdots, \gamma_k^{\tau_i^*})'$ denotes the corresponding regression coefficients to be estimated. Note that the vectors of regression coefficients are not the same for different risk classes because of their different quantile levels.

In the following subsections, we discuss how to apply traditional quantile regression, parametric quantile regression, and quantile regression with coefficient functions to determine the risk premiums of individual policies.

3.2 Traditional Quantile Regression Model

Given the quantile level τ_i^* of policy i, we have the following traditional quantile regression:

$$\log Q_{Y_i^*}\left(\tau_i^* \left| \boldsymbol{x}_i^Q \right.\right) = \boldsymbol{x}_i^Q \boldsymbol{\gamma}^{\tau_i^*}. \tag{3.7}$$

The estimation of regression coefficients of Eq.(3.7) can be derived by solving the following minimization problem with R package `quantreg`: Quantile Regression; see Koenker and Bassett (1978) and Koenker and Hallock (2001):

$$\min_{\boldsymbol{\gamma}^{\tau_i^*} \in \mathbb{R}^{k+1}} \left[\sum_{\log\left(y_i^*\right) \geq \boldsymbol{x}_i \boldsymbol{\gamma}^{\tau_i^*}} \tau_i^* \left| \log\left(y_i^*\right) - \boldsymbol{x}_i \boldsymbol{\gamma}^{\tau_i^*} \right| + \sum_{\log\left(y_i^*\right) < \boldsymbol{x}_i \boldsymbol{\gamma}^{\tau_i^*}} \left(1-\tau_i^*\right) \left| \log\left(y_i^*\right) - \boldsymbol{x}_i \boldsymbol{\gamma}^{\tau_i^*} \right| \right]. \tag{3.8}$$

According to Eqs.(3.1) and (3.7), the risk premium of policy i is given by

$$H\left(Y_i; p_i, \boldsymbol{\gamma}\right) = \left(1-p_i\right) \exp\left(\boldsymbol{x}_i^Q \boldsymbol{\gamma}^{\tau_i^*}\right). \tag{3.9}$$

3.3 Parametric Quantile Regression Model

Parametric quantile regression models allow us to apply a wide range of skewed and heavy tailed distributions to capture flexible shapes and tail behavior in insurance claim data. These distributions include the generalized beta of the second kind distribution (Cummins et al., 1990), generalized-t distribution (McDonald and Newey, 1988), and generalized gamma (GG) distribution (Noufaily and Jones, 2013).

Compared with traditional quantile regression, parametric quantile regression allows us to consider the impact of covariates on the entire distribution, not merely on its conditional mean. Furthermore, the monotonicity of the quantile function in parametric quantile regression can be strictly guaranteed, because the inverse cumulative distribution function of a distribution is itself a quantile function, which obviates the problem of quantile crossing in the traditional quantile regression.

To develop a framework of parametric quantile regression in predicting the risk premium in non-life insurance ratemaking, we adopt the GG distribution used in Noufaily and Jones (2013). Since GG distribution is defined on a real support, we assume that the log of the aggregate claim amount of the i-th policy that has at least one claim follows the GG distribution with location parameter η_i, scale parameter ω, and shape parameter k, with its probability density function given by Stacy et al. (1962):

$$f_{\log(Y_i^*)}\left(y_i;\eta_i,\omega,k\right)=\frac{k^{k-1/2}}{\omega\Gamma\left(k\right)}\exp\left[\frac{\log(y_i)-\eta_i}{\omega}\sqrt{k}-k\exp\left(\frac{1}{\sqrt{k}}\frac{\log(y_i)-\eta_i}{\omega}\right)\right],\tag{3.10}$$

for $\log(Y_i^*)\in\mathbb{R}$ and y_i is the observed aggregate claim amount for the i-th policy that has at least one claim. We consider only the linear regression form for the location parameter of the GG distribution:

$$\eta_i=\boldsymbol{x}_i^Q\boldsymbol{\gamma},\tag{3.11}$$

where $\boldsymbol{x}_i^Q=(1,x_{i1}^Q,\cdots,x_{ik}^Q)$ represents the $(k+1)$-dimensional vector of covariates and $\boldsymbol{\gamma}=(\gamma_0,\gamma_1,\cdots,\gamma_k)'$ denotes the corresponding regression coefficients to be estimated. It should be noted that the vector of regression coefficients stays the same for different quantile levels.

The quantile function that is associated with density f in Eq.(3.10) is given by

$$Q_{Y_i^*}\left(\tau_i^*|\boldsymbol{x}_i^Q\right)=\exp\left(\eta_i\right)\left\{\frac{\Gamma\left(\tau_i^*,\ k\right)}{k}\right\}^{\omega/\sqrt{k}},\tag{3.12}$$

where $\Gamma\left(a,x\right)$ is an incomplete gamma function, that is, $\Gamma\left(a,x\right)=\int_x^\infty t^{a-1}\exp\left(-t\right)dt$.

Employing the maximum likelihood method, we obtain the estimates of parameters in the GG regression model with `optim` function in R software. The log-likelihood of the GG regression model is given by

$$\begin{aligned}\ell(\boldsymbol{\gamma},\omega,k)=\sum_{i=1}^{N}\Bigg[&\left(k-\frac{1}{2}\right)\log k-\log\omega-\log\Gamma\left(k\right)\\&+\frac{\log y_i-\boldsymbol{x}_i^Q\boldsymbol{\gamma}}{\omega}\sqrt{k}-k\exp\left(\frac{\log y_i-\boldsymbol{x}_i^Q\boldsymbol{\gamma}}{\omega\sqrt{k}}\right)\Bigg].\end{aligned}\tag{3.13}$$

According to Eqs.(3.1) and (3.12), the risk premium of policy i is given by

$$H\left(Y_i;p_i,\boldsymbol{\gamma},k,\omega\right)=\left(1-p_i\right)\exp\left(\boldsymbol{x}_i^Q\boldsymbol{\gamma}\right)\left\{\frac{\Gamma\left(\tau_i^*,k\right)}{k}\right\}^{\omega/\sqrt{k}}.\tag{3.14}$$

3.4 Quantile Regression with Coefficient Functions

One problem associated with a quantile regression model is that its coefficients depend on the quantile level; see Frumento and Bottai (2016, 2017). To solve this problem, Frumento and Bottai (2016) propose a parametric model for the coefficients in the quantile regression and adopt quantile regression coefficients modeling. Specifically, they express the regression coefficients as some parametric functions of the quantile level. Quantile regression with coefficient functions has some advantages, including parsimony, efficiency, and simple interpretation. To develop a framework of quantile regression with coefficient functions in predicting the risk premium in non-life insurance ratemaking, we adopt similar notation to that of Frumento and Bottai (2016) as follows:

$$\log\left[Q_{Y_i^*}\left(\tau_i^*\left|\boldsymbol{x}_i^Q,\boldsymbol{\theta}\right.\right)\right]=\boldsymbol{x}_i^Q\boldsymbol{\gamma}\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right), \tag{3.15}$$

where $\boldsymbol{x}_i^Q=(1,x_{i1}^Q,\cdots,x_{ik}^Q)$ represents the $(k+1)$-dimensional vector of covariates, $\boldsymbol{\gamma}\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right)=\left[\gamma_0\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right),\cdots,\gamma_k\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right)\right]'$ denotes the corresponding vector as a function of quantile level τ_i^* and finite-dimensional parameters $\boldsymbol{\theta}$, namely,

$$\boldsymbol{\gamma}\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right)=\boldsymbol{\theta}\boldsymbol{b}\left(\tau_i^*\right), \tag{3.16}$$

where $\boldsymbol{b}\left(\tau_i^*\right)=\left[b_1\left(\tau_i^*\right),\cdots,b_q\left(\tau_i^*\right)\right]'$ is a set of q known functions of quantile level τ_i^*, and $\boldsymbol{\theta}$ is a $(k+1)\times q$ matrix with entries $\theta_{m,h}(m=1,\cdots,k+1;h=1,\cdots,q)$ given by

$$\boldsymbol{\theta}=\begin{bmatrix}\theta_{11} & \theta_{12} & \cdots & \theta_{1q}\\ \vdots & \vdots & \ddots & \vdots\\ \theta_{k1} & \theta_{k2} & \cdots & \theta_{k,q}\\ \theta_{k+1,1} & \theta_{k+1,2} & \cdots & \theta_{k+1,q}\end{bmatrix}_{(k+1)\times q}.$$

Note that the quantile regression coefficient associated with the j-th covariate is given by

$$\gamma_j\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right)=\theta_{j+1,1}b_1\left(\tau_i^*\right)+\cdots+\theta_{j+1,q}b_q\left(\tau_i^*\right),\quad j=0,\cdots,k, \tag{3.17}$$

where $\boldsymbol{\theta}$ is the corresponding vector of coefficients to be estimated, and some entries of $\boldsymbol{\theta}$ may be set to 0 to allow the regression coefficient to be functions of possibly different subsets of $b\left(\tau_i^*\right)$.

Thus, the conditional quantile function is given by

$$\log\left[Q_{Y_i^*}\left(\tau_i^*\left|\boldsymbol{x}_i^Q,\boldsymbol{\theta}\right.\right)\right]=\boldsymbol{x}_i^Q\boldsymbol{\gamma}\left(\tau_i^*\left|\boldsymbol{\theta}\right.\right)=\boldsymbol{x}_i^Q\boldsymbol{\theta}\boldsymbol{b}\left(\tau_i^*\right). \tag{3.18}$$

Note that Eq.(3.18) is associated with the choice of the function $\boldsymbol{b}\left(\tau_i^*\right)$. In practice, the choice of $\boldsymbol{b}\left(\tau_i^*\right)$ must ensure that the quantile is monotonically increasing. For instance, polynomials, splines, trigonometric functions, and quantile function of standard normal distribution could be used in practice:

$$b_j\left(\tau_i^*\right)=\begin{cases}\left(\tau_i^*\right)^2\\ \left(\tau_i^*\right)^3\\ \Phi^{-1}\left(1-\tau_i^*\right)\\ \cos\left(2\pi\tau_i^*\right)\end{cases},\quad j=1,\cdots,J. \tag{3.19}$$

Estimating the τ_i^*-th quantile regression coefficients under model (3.18) requires minimizing the following loss function

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\sum_{i=1}^N\int_0^1\left(\tau_i^*-w_{\tau_i^*,i}\right)\left[\log(y_i)-\boldsymbol{x}_i^Q\boldsymbol{\theta}\boldsymbol{b}\left(\tau_i^*\right)\right]d\tau_i^*\right\}, \tag{3.20}$$

where $w_{\tau_i^*,i}=I\left[\log(y_i)\le\boldsymbol{x}_i\boldsymbol{\gamma}\boldsymbol{b}\left(\tau_i^*\right)\right]$ and $I(\cdot)$ is the indicator function. The estimation procedure can be implemented with R package `qrcm`: quantile regression coefficients modeling; see Gilchrist (2000) and Frumento and Bottai (2016, 2017).

According to Eqs.(3.1) and (3.18), the risk premium of policy i is given by

$$H\left(Y_i;p_i,\boldsymbol{\theta}\right)=\left(1-p_i\right)\exp\left[\boldsymbol{x}_i^Q\boldsymbol{\theta}\boldsymbol{b}\left(\tau_i^*\right)\right]. \tag{3.21}$$

4 Calculation of Total Risk Premium and Risk Loading Parameters

Sections 2 and 3 show that regardless of whether the two-part GLMs or the quantile regression models are used to calculate the individual risk premium, some risk loading parameters (e.g., φ, ρ, and τ) have to be given subjectively in advance. To overcome this subjective problem, we propose a top-down method to calculate the total risk premium of the portfolio and risk loading parameters in this section.

4.1 Calculating Total Risk Premium

In Solvency II regulation, the probability that the aggregate claim amount of the whole portfolio $S = \sum_{i=1}^{N} Y_i$ exceeds its total risk premium that the insurance company will charge should be controlled in a small range, such as less that 0.5%.

For the whole portfolio, suppose that the aggregate claim amount S has the cumulative distribution function F_S and its total risk premiumis denoted by C; then, the probability that the aggregate claim amount of the whole portfolio exceeds its total risk premium is given by

$$\Psi = \Pr\left[S > C\right] = 1 - F_S\left(C\right). \tag{4.1}$$

From Eq.(4.1), we obtain the total risk premium of the whole portfolio that the insurance company will charge as follows

$$C = F_S^{-1}\left(1 - \Psi\right), \tag{4.2}$$

where $F_S^{-1}\left(\varepsilon\right)$ denotes the ε-th quantile of S. In other words, if the probability that the aggregate claim amount for the whole portfolio exceeds the total risk premium C is small enough, such as $\Psi = 0.5\%$, then the total risk premium for the whole portfolio is the 99.5% quantile of its aggregate claim amount S. Hence, the key for controlling the probability Ψ and calculating the total risk premium of the whole portfolio is to derive the entire distribution of S.

In this subsection, we propose a bootstrap method to calculate the total risk premium of the whole portfolio. First, We generate a sequence of pseudo individual aggregate claim amounts and then predict the total risk premium of the whole portfolio according to the following procedure.

Step 1: Simulate a pseudo-response of the aggregate claim amount $\tilde{y}_i$ for policy i from density function $f_{Y_i}\left(y_i \left|\hat{\mu}_i, \hat{p}_i, \hat{\sigma}\right.\right)$, $i = 1, \cdots, N$. Note that density function f can be the two-part GA distribution or the two-part IG distribution of Eqs.(A.1) and (A.4) in the Appendix, respectively. Hence, a simulation of the aggregate claim amount for the whole portfolio is $\sum_{i=1}^{N} \tilde{y}_i$.

Step 2: Use the pseudo-responses to form the b^{th} bootstrap sample $\left\{\tilde{y}_1^b, \cdots, \tilde{y}_N^b\right\}$ from which to derive the bootstrap replication of $\left(\hat{\mu}_i^b, \hat{p}_i^b, \hat{\sigma}^b\right)_{i=1,\cdots,N}$ by applying the two-part GLMs framework.

Step 3: Repeating these two steps for $b = 1, \cdots, B$, we obtain a predictive distribution of aggregate claim amounts for the whole portfolio. As a result, the total risk premium for the whole portfolio is the $(1 - \Psi)$-th quantile of the aggregate claim amount for the whole portfolio, and the total pure premium for the whole portfolio is the mean of the aggregate of claim amount for the whole portfolio. The total risk loading for the whole portfolio is calculated by the difference between the total risk premium and the total pure premium.

4.2 Calculating Risk Loading Parameters

In expected value premium principle and standard deviation premium principle, the risk premium for each individual policy is related to a risk loading parameter φ. In the Wang premium principle, the risk premium is related to a risk aversion factor ρ. In the quantile premium principle, the risk premium is related to a quantile level τ. It is obvious that these relevant parameters need to be given directly or indirectly to calculate risk premiums.

In the existing literature, these parameters in premium principles are subjectively given. For instance, Heras et al. (2018) propose the quantile level $\tau = 95\%$ in quantile regression models (see Eq.(1.5)). In the VaR premium principle that Kudryavtsev (2009) proposes (see Eq.(1.4)), the 95% quantile of the aggregate claim amount of an individual policy is used as its individual risk premium and the sum of the individual risk premiums is used as the total risk premium for the whole portfolio. The shortcoming of this approach is that, while it can guarantee that the aggregate claim amount of each policy exceeds its risk premium by no more than 5%, the probability that the aggregate claim amount of the whole portfolio exceeds its total risk premium may be much less than 5%, due to a certain risk diversification effect between individual policies. In other words, the total risk premium obtained by this method may be higher than what is appropriate.

In this subsection, we first calculate the total risk premium for the whole portfolio and then distribute it to individual policies by solving the following equation:

$$\sum_{i=1}^{N} H\left(Y_i\right) = C, \tag{4.3}$$

where $H\left(Y_i\right)$ denotes the risk premium for the i-th policy. Table 1 shows the equations for calculating the risk premiums for individual policies under various premium principles. The risk loading parameters in the expected value premium principle and standard deviation premium principle can be obtained by applying two-part GLMs. The corresponding quantile level τ in the quantile premium principle and risk aversion factor ρ in the Wang premium principle may be solved by numerical algorithms. For policy i, the unique τ in Table 1 denotes the quantile level of its aggregate claim amount that contains zero claims, while the $\tau_i^* = (\tau - p_i)/(1 - p_i)$ represents the quantile level of its non-zero aggregate claim amount.

Table 1: Total risk premium and allocation in various Premium Principles

Premium Principle	**Allocation Equation**	**Relevant Parameters**
Expected value premium principle	$\sum_{i=1}^{N}\left[\mathbb{E}\left(Y_i\right) + \varphi\mathbb{E}\left(Y_i\right)\right] = C$	$\varphi = \frac{C - \sum_{i=1}^{N}\mathbb{E}(Y_i)}{\sum_{i=1}^{N}\mathbb{E}(Y_i)}$
Standard deviation premium principle	$\sum_{i=1}^{N}\left[\mathbb{E}\left(Y_i\right) + \varphi\sqrt{\operatorname{Var}\left(Y_i\right)}\right] = C$	$\varphi = \frac{C - \sum_{i=1}^{N}\mathbb{E}(Y_i)}{\sum_{i=1}^{N}\sqrt{\operatorname{Var}(Y_i)}}$
Wang premium principle	$\sum_{i=1}^{N}\int\limits_{0}^{\infty}\Phi\left[\Phi^{-1}\left(S_Y\left(y\right)\right) + \rho\right]dy_i = C$	solve ρ by the numerical algorithms
Two-part quantile premium principle	$\sum_{i=1}^{N}(1 - p_i)Q_{Y_i^*}\left(\tau_i^*\left\|Y_i\right.\right) = C$	solve τ by the numerical algorithms

5 Application to Ratemaking

The data set we use in this study contains information on full comprehensive Australian insurance policies between years 2004 and 2005, which comes from De Jong and Keller (2008); the same data set is analyzed in Heras et al. (2018) and Baione and Biancalana (2019). The insurance portfolio contains 67,856 policies, of which 4,624 have at least one claim. Each claim record consists of an aggregate claim amount (Claimcst0), claim numbers (Numclaims), occurrence of claim (Clm), exposure, and several covariates, such as age of policyholder, age of vehicle, value of vehicle, area of residence, and body type of vehicle. For simplification and comparative purposes, we consider the same covariates as Heras et al. (2018) in the following application: age of vehicle (Veh_age) and age of driver (Agecat).

The variables in the data set are listed in Table 2. For each policy, we define the aggregate claim amount as the sum of the cost of all claims submitted by the policyholder, assuming that the aggregate amount is zero if the policy has no claim. A histogram of the (positive) aggregate claim amount is given in the left panel of Figure 1. For clarity, the horizontal axis is truncated at $15,000. A total of 65 claims between $15,000 and $57,000 are omitted from this display. A bar-plot of the claim numbers for those policies that have one or more claims is given in right panel of Figure 1. In this portfolio, most of the policies, up to 93.19%, have only one claim each and only 0.002947% have four claims each.

Table 2: Description of Variables

Variables	**Type**	**Description**
Agecat	Categorical	Driver's age category: 1 (youngest), 2, 3, 4, 5, 6
Veh_age	Categorical	Age of vehicle: 1 (youngest), 2, 3, 4
Exposure	Continuous	Policy years (between 0 and 1)
Clm	Discrete	Occurrence of claim (0 = no, 1 = yes)
Numclaims	Discrete	Numbers of claims(0, 1, 2, 3,$\cdots$)
Claimcst0	Continuous	Aggregate claim amount of a policy (0 if no claim)

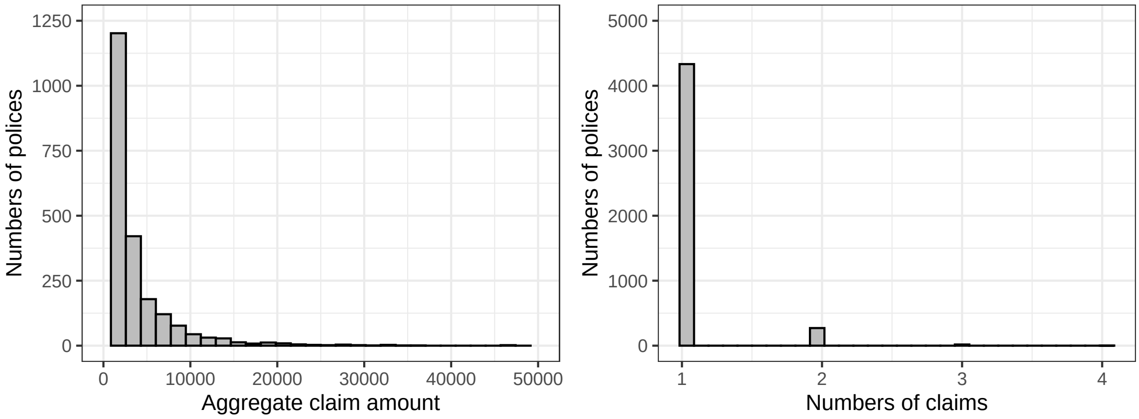

Figure 1: Predictive distribution of aggregate claim amount (left panel) and QQ-plot of aggregate claim amount (right panel) of the portfolio.

5.1 Total Risk Premium of the Portfolio

To obtain the total risk premium of the portfolio, we first establish two-part GLMs by assuming that the non-zero aggregate claim amounts follow GA distribution or IG distribution, both using two rating factors, Veh_age and Agecat.

Table 3 shows the parameter estimates and the corresponding P-values for both models. For the logistic regression part, the estimates of the two models are identical and all the parameters are highly significant, except for the first level of Veh_age and the sixth level of Agecat; this result is equivalent to that of the two-part model in Heras et al. (2018). Table 3 shows that the IG regression model is more appropriate for fitting non-zero aggregate claim amounts of individual policies, since its Akaike information criterion and Bayesian information criterion are much smaller than those of the GA regression model.

Table 3: Parameter estimates of two-part GLMs

Models	Parameters	Two-part GA regression		Two-part IG regression	
		Estimates	P-value	Estimates	P-value
Logistic regression	(Intercept)	-1.907	<0.001	-1.907	<0.001
	Veh_age: 1	-0.031	0.535	-0.031	0.535
	Veh_age: 3	-0.127	0.004	-0.127	0.004
	Veh_age: 4	-0.221	<0.001	-0.221	<0.001
	Agecat: 1	0.533	<0.001	0.533	<0.001
	Agecat: 2	0.334	<0.001	0.334	<0.001
	Agecat: 3	0.272	<0.001	0.272	<0.001
	Agecat: 4	0.230	<0.001	0.230	<0.001
	Agecat: 6	-0.003	0.966	-0.003	0.966
Non-zero aggregate claim amount regression	(Intercept)	7.420	<0.001	7.411	<0.001
	Veh_age: 1	-0.051	0.323	-0.056	0.445
	Veh_age: 3	0.027	0.546	0.033	0.608
	Veh_age: 4	0.118	0.012	0.13	0.060
	Agecat: 1	0.439	<0.001	0.453	<0.001
	Agecat: 2	0.215	<0.001	0.223	0.008
	Agecat: 3	0.104	0.072	0.106	0.179
	Agecat: 4	0.119	0.040	0.127	0.110
	Agecat: 6	0.084	0.269	0.091	0.387
Scale parameter		1.149	<0.001	0.037	<0.001
Log-likelihood		-55900.58		**-54844.71**	
AIC		111839.20		**109727.40**	
BIC		112012.50		**109900.80**	

Table 4 shows the probability of incurring no claims (p_i) for individual policies and the pure premiums for 24 risk classes by using the two-part IG regression model. The total number of policies and the total number of claims are given in columns 4 and 5 respectively. Compared with the results of Heras et al. (2018), the estimates of the probability of having no claims are the same as those of Heras et al. (2018) but the pure premiums are slightly different , because we use the IG regression model instead of the GA regression model, and the former shows better goodness of fit than the latter does.

Finally, we approximate the predictive distribution of the aggregate claim amounts by bootstrapping for 10,000 times based on the two-part IG regression model. For the current portfolio with 67,856 policies, Figure 2 shows the predictive distribution and QQ-plots for the aggregate claim amount of the whole portfolio. The mean of the predictive distribution is \$18,765,168 and the 99.5% quantile is \$20,563,196, which means that if the total risk premium is determined as C = \$20,563,196, then the probability that the aggregate claim amount of the whole portfolio is greater than the total risk premium is less than 0.5%.

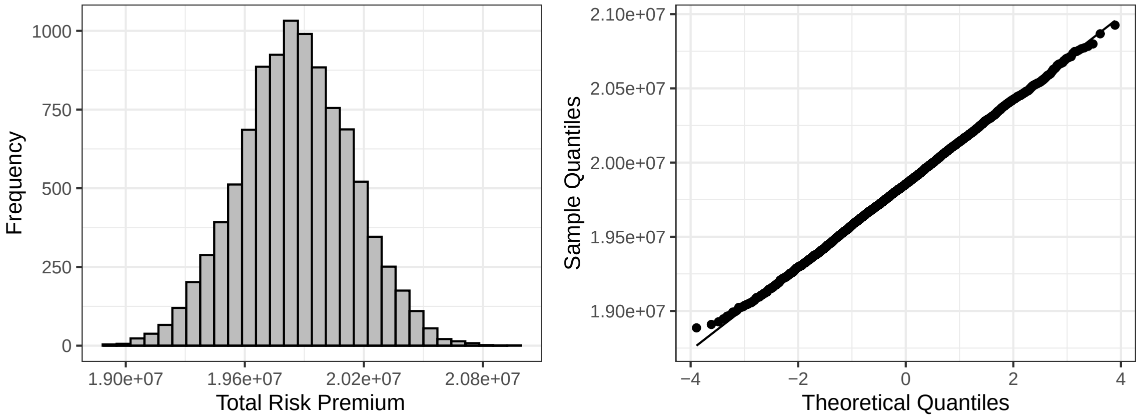

Figure 2: Predictive distribution of aggregate claim amount (left panel) and QQ-plot of aggregate claim amount (right panel) of the portfolio.

In the following subsection, we assume that the portfolio remains unchanged and that the total risk premium of the whole portfolio charged by the insurance company is \$20,563,196.

5.2 Classification Risk Premiums Based on Two-part GLMs

In two-part GLMs, we can obtain not only the pure premium, but also the standard deviation for individual policies. The risk premium for individual policies can be obtained using the expected value premium principle or the standard deviation premium principle. If the sum of the risk premiums for individual policies equals the total risk premium C of the portfolio, then the risk loading parameter in the expected value premium principle can be expressed as

$$\hat{\varphi} = \frac{C - \sum\limits_{i=1}^{n} \left[(1 - \hat{p}_i)\, \hat{\mu}_i \right]}{\sum\limits_{i=1}^{n} \left[(1 - \hat{p}_i)\, \hat{\mu}_i \right]}. \tag{5.1}$$

The risk loading parameter in the standard deviation premium principle is expressed as

$$\hat{\varphi} = \frac{C - \sum_{i=1}^{n} \left[(1 - \hat{p}_i) \hat{\mu}_i \right]}{\sum_{i=1}^{n} \left[\sqrt{(1 - \hat{p}_i) \hat{\mu}_i^2 (\hat{p}_i + \hat{\mu}_i \hat{\sigma}^2)} \right]}, \tag{5.2}$$

where $(1 - \hat{p}_i) \hat{\mu}_i$ and $\sqrt{(1 - \hat{p}_i) \hat{\mu}_i^2 (\hat{p}_i + \hat{\mu}_i \hat{\sigma}^2)}$ are the mean and standard deviation, respectively, of the aggregate claim amount of policy i.

Similarly, the risk aversion parameter ρ in the Wang premium principle can be solved from the following equation:

$$\sum_{i=1}^{N} \int_0^{\infty} \Phi \left[\Phi^{-1} \left(1 - F_{Y_i} (y_i; \hat{\mu}_i, \hat{p}_i, \hat{\sigma}) \right) + \rho \right] \mathrm{d}y_i = C, \tag{5.3}$$

where $F_{Y_i} (y_i; \hat{\mu}_i, \hat{p}_i, \hat{\sigma})$ is the cumulative distribution function of the aggregate claim amount of policy i with estimated parameters $(\hat{\mu}_i, \hat{p}_i, \hat{\sigma})$.

Table 5 presents the risk premiums of 24 risk classes predicted using the two-part IG regression model under various premium principles. We find that the risk premiums of the 24 risk classes are significantly different, and the risk loadings are very close for the expected value premium principle, standard deviation premium principle, and Wang premium principle. Although only the Wang premium principle is a coherent risk measure, the risk premiums obtained from these three premium principles make no big difference in this case.

5.3 Classification Risk Premiums Based on Two-Part Quantile Regression Models

In this section, we apply quantile regression models to calculate the risk premiums for individual policies by using the two-part quantile premium principle in Eq.(3.1). In quantile regression models, the risk loading is implicitly included in the risk premium.

The response variable in the quantile regression is the log-transformed non-zero aggregate claim amounts of individual policies (log(claimcst0)). From Eq.(3.4), we observe that to obtain the quantile of the aggregate claim amounts of individual policies that contains zeroes, we need focus only on those policies that submit at least one claim; then, the quantile level is given by

$$\tau_i^* = \frac{\tau - p_i}{1 - p_i}, \tag{5.4}$$

where ${\tau_i}^*$ is the quantile level of the non-zero aggregate claim amount for individual policy i and τ is the quantile level of the aggregate claim amount that contains zeroes. The probability of having no claim p_i is estimated using the logistic regression model in Eq.(2.1).

Before applying the quantile regression models, we need to choose an appropriate quantile level τ. In this study, given the total risk premium C, the quantile level τ can be solved from the following equation:

$$\sum_{i=1}^{N} \left[(1 - p_i) Q_{Y_i^*} \left(\tau_i^* | \boldsymbol{x}_i^Q \right) \right] = C. \tag{5.5}$$

where τ_i^* is given in Eq.(5.4).

For a given quantile level, we apply the traditional quantile regression, parametric quantile regression, and quantile regression with coefficient functions. The response variable is the log-transformed non-zero aggregate claim amounts of individual policies that submit at least one claim, and the covariates are Veh_age and Agecat, which are the same as those of the mean regression models in the previous section. In the traditional quantile regression model, the covariates are introduced into log-transformed quantile as follows:

$$\log\left[Q_{Y_i^*}\left(\tau_i^*|\boldsymbol{x}_i^Q\right)\right] = \gamma_0^{\tau_i^*} + \gamma_1^{\tau_i^*}\text{Veh\_age1}+\gamma_2^{\tau_{\mathrm{i}}^*}\text{Veh\_age3} + \cdots + \gamma_5^{\tau_i^*}\text{Agecat1} + \gamma_9^{\tau_i^*}\text{Agecat6}. \tag{5.6}$$

For parametric quantile regression, we assume that the log-transformed non-zero aggregate claim amounts follow GG distribution and the covariates are introduced into its mean parameter as follows:

$$\begin{aligned}\log\left[Q_{Y_i^*}\left(\tau_i^*|\boldsymbol{x}_i^Q\right)\right] &= \eta_i + \log\left\{\frac{\Gamma\left(\tau_i^*, k\right)}{k}\right\}^{\omega/\sqrt{k}}, \\ \eta_i &= \gamma_0 + \gamma_1\text{Veh\_age1}+\gamma_2\text{Veh\_age3}+\cdots+\gamma_5\text{Agecat1}+\gamma_9\text{Agecat6}\ . \end{aligned} \tag{5.7}$$

The quantile regression with coefficient functions is given by

$$\begin{aligned}\log\left[Q_{Y_i^*}\left(\tau_i^*|\boldsymbol{x}_i^Q\right)\right] &= \gamma_0(\tau_i^*) + \gamma_1(\tau_i^*)\text{Veh\_age1} + \gamma_2(\tau_i^*)\text{Veh\_age3} \\ &\quad + \cdots + \gamma_5(\tau_i^*)\text{Agecat1} + \cdots + \gamma_9(\tau_i^*)\text{Agecat6}, \\ \gamma_j(\tau_i^*) &= \theta_{0j} + \theta_{1j}\tau_i^* + \theta_{2j}\tau_i^{*2}, \quad j = 0, 1, \cdots, 9, \end{aligned} \tag{5.8}$$

where $\gamma_j\left(\tau_i^*\right)$ is a polynomial function for capturing the relationship between quantile levels and the coefficients of the quantile regression model.

Table 6 reports the risk premiums of 24 risk classes by using traditional quantile regression, parametric quantile regression, and quantile regression with coefficient functions. For the given total risk premium $C = 20,563,196$ of the portfolio, the appropriate quantile levels are around 96% in these three quantile regression models.

5.4 Relationship between probability Ψ and quantile level τ

The total risk premium of the portfolio should cover the actual aggregate claim amount at the $1-\Psi$ probability level or more. In this subsection, we discuss the choice of probability Ψ in Eq.(4.2) for the insurance company and check how that affects the total risk premium C of the whole portfolio and the quantile level τ. We focus on the impact of different probabilities Ψ on predicting risk premiums of different risk classes.

Figure 3 shows the range of the total risk premium C based on the parametric bootstrap method proposed in Section 4. We observe that if the probability Ψ varies between 0.5% and 25%, then the total risk premium of the portfolio is between \$20,050,581 and \$20,563,196, which shows a noticeable difference among these assumptions.

Figure 4 shows the range of quantile level τ obtained by traditional quantile regression, parametric quantile regression, and quantile regression with coefficient functions. For these three quantile regression models, as the probability $(1-\Psi)$ increases from 75% to 99.5%, the quantile level τ just increases slightly and almost remain around 96%.

Generally, as the portfolio size (number of policies) increases, the risk loading ratio, which is defined as the ratio of total risk loading to total pure premium while implementing the top-down method, should decrease due to the diversification effect. Figures 5 show the risk premiums of 24 risk classes under the different probabilities Ψ using three quantile regression models. We can see that there are

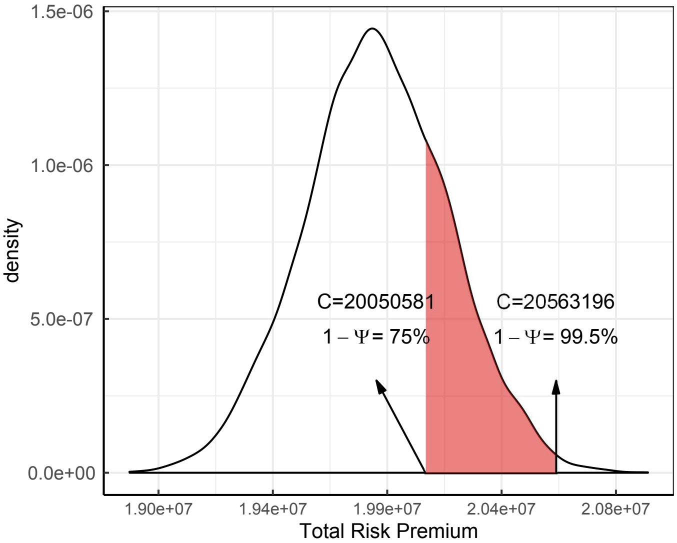

Figure 3: Total Risk Premium of the Whole Portfolio at Probabilities $1 - \Psi$ from 75% to 99.5%.

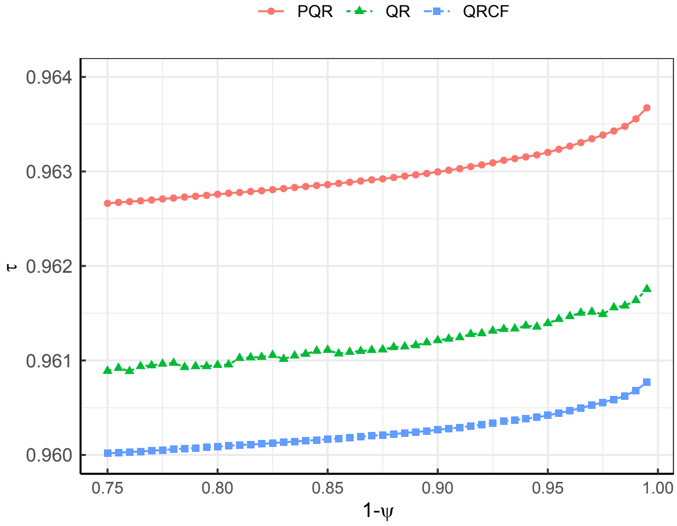

Figure 4: Relationship between probability $1 - \Psi$ and quantile level τ.

not big differences in the three cases. We conclude that, although the probability Ψ controls the risk loading of the whole portfolio at the collective level, the Ψ has small impact on the quantile level and the risk premiums for different risk classes at the individual level due to the diversification effect, which is consistent with previous conclusion in Figure 4.

It concludes that the top-down method proposed in this study guarantees that the total risk premium covers the aggregate claim amount with a probability of 75% or more, and the classification risk premiums are less affected by the probability Ψ selected in advance, which means that the method is robust.

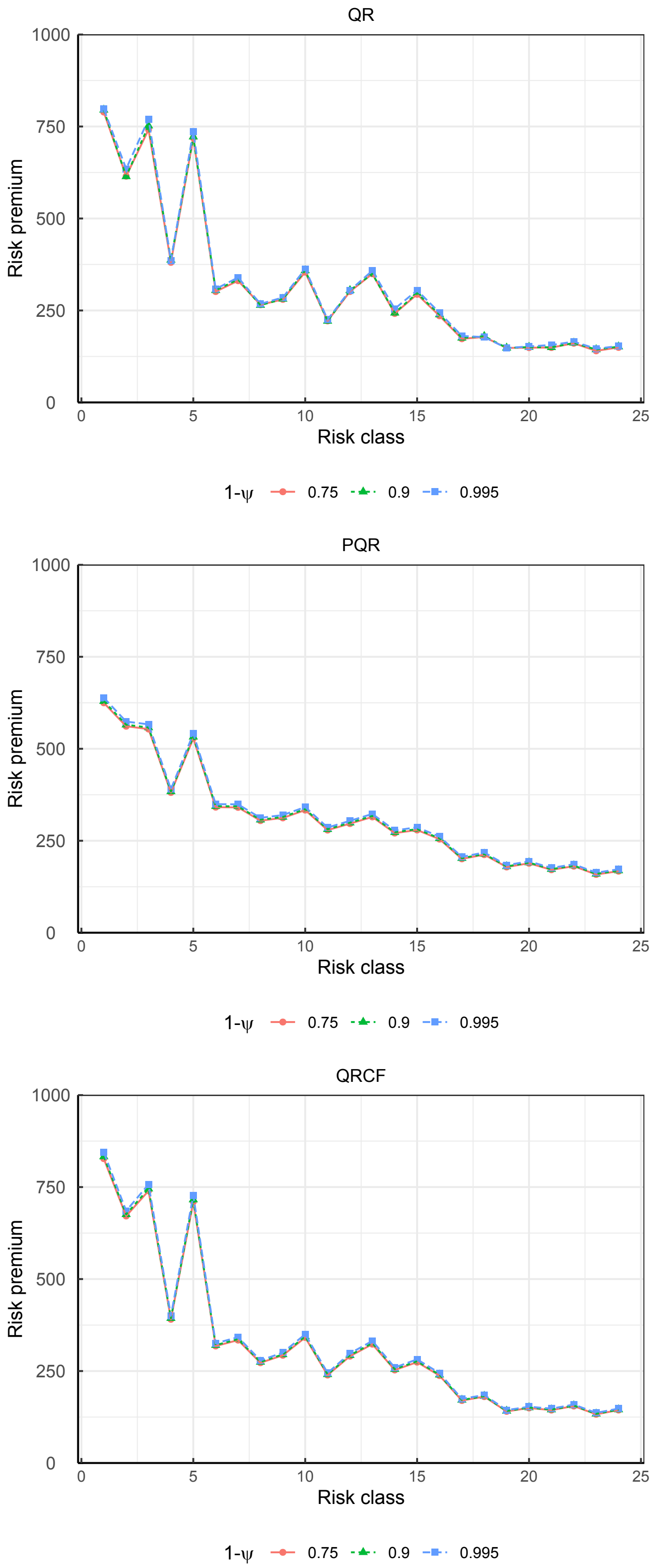

Figure 5: Corresponding risk premiums by controlling different probabilities $1 - \Psi$.

5.5 Comparative Analysis

Heras et al. (2018) propose the quantile premium principle to calculate the risk premiums of individual policies, that is

$$H\left(Y_i\right) = \mathbb{E}\left(Y_i\right) + \varphi\left[Q_{Y_i}\left(\tau_i^*\right) - \mathbb{E}\left(Y_i\right)\right], \tag{5.9}$$

where $\tau_i^* = (\tau - \hat{p}_i)/(1 - \hat{p}_i)$ is the quantile level for the i-th risk class and $\tau = 0.95$; $\hat{p}_i$ is the probability of having no claims that can be predicted by a logistic regression model; $Q_{Y_i}\left(\tau_i^*\right)$ is the τ_i^*-th quantile of the aggregate claim amount; $\mathbb{E}\left(Y_i\right)$ is the pure premium of the i-th risk class; $\varphi\left[Q_{Y_i}\left(\tau_i^*\right) - \mathbb{E}\left(Y_i\right)\right]$ represents risk loading, which is the difference between the 95% quantile of the aggregate claim amount and the pure premium; φ is the risk loading parameter.

For ease of comparison with the results of Heras et al. (2018), we assume the total risk premium of the portfolio is $C = 20,563,196$, and the risk premiums of different risk classes are recalculated using the quantile premium principle in Heras et al. (2018) with the corresponding risk loading parameter $\varphi = 0.897\%$.

Table 7 shows the risk premiums of the 24 risk classes using different models. The risk premiums using the expected value premium principle, standard deviation premium principle, and Wang premium principle are very close to those of the quantile regression model in Heras et al. (2018). In other words, for two-part GLMs, given the total risk premium of the whole portfolio C, regardless of which premium principle is used, there is little impact on the risk premiums of individual risk classes.

In order to measure the prediction accuracy, it is well known that the frequently used loss functions, eg., the root mean square error (RMSE) are not appropriate measures for capturing the difference between the predictive values and the corresponding outcomes, due to the high proportion of zeros and right heavy-tailed features in the loss distributions. In this case, the use of loss function is bounded as the observed risk premium of different risk classes is unknown. Therefore, we turn to alternative statistical measures - the ordered Lorenz curve and the associated Gini index. The Gini index is a statistical measure of distribution developed by the Italian statistician Corrado in 1912. It is often used as a gauge of economic inequality, measuring wealth distribution among a population. The index ranges from 0% to 100%, with 0% representing perfect equality and 100% representing perfect inequality. The subsequent literature is extensive. For example, Frees et al. (2011) develops theoretical properties of this Gini index and Shi and Yang (2018) applies it to measure the discrepancy between the premium and loss distributions in the non-life ratemaking. In this study, we use the original definition of Gini index developed by Corrado (1921). The ordered Lorenz curve is the plot with using the proportion of an risk exposure on the horizontal axis and a distribution function of predicted value of risk premiums on the vertical axis. The associated Gini index is defined as twice the area between the ordered Lorenz curve and the line of equality. A higher Gini index indicates greater heterogeneity of different risk classes, with high risk premium individuals receiving much larger percentages of the total risk premiums of the risk exposure.

Figure 7 displays the ordered Lorenz curves corresponding to Gini indices of the risk premium prediction reported in Table 7, which are calculated correspondingly with ranking the value of risk exposure from large to small. Relative to two-part GLMs, the Gini indices calculated by two-part quantile regression models is the largest three of all as expected, which means that the quantile regression can reveal the heterogeneity of different risk classes more efficiently, and thus, can obtain more reasonable risk premiums of individual policies. For graphical comparison that confirms the Gini indices results, we show the predivive risk premium of the 24 risk classes based on the three models proposed in Figure 6. We observed that the risk premiums calculated by quantile regression models are more significantly different between various risk classes.

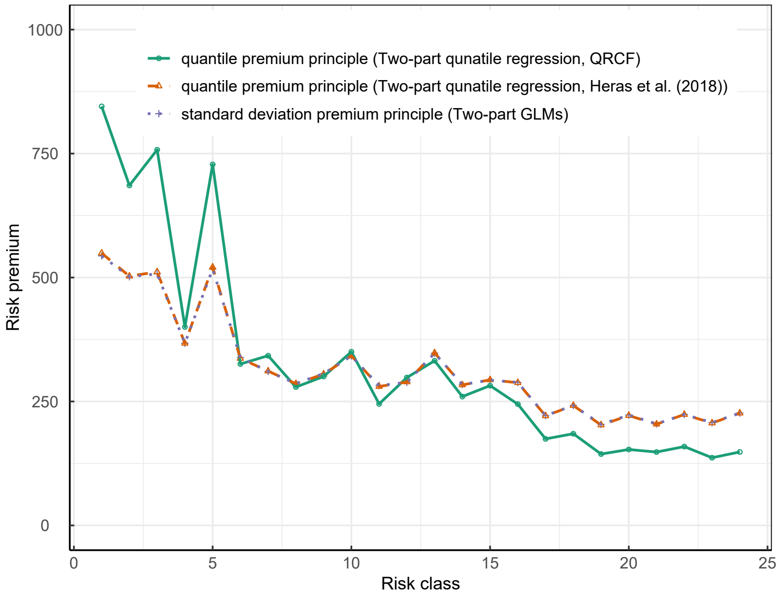

Figure 6: Classification risk premiums based on two-part quantile regression and two-part GLMs.

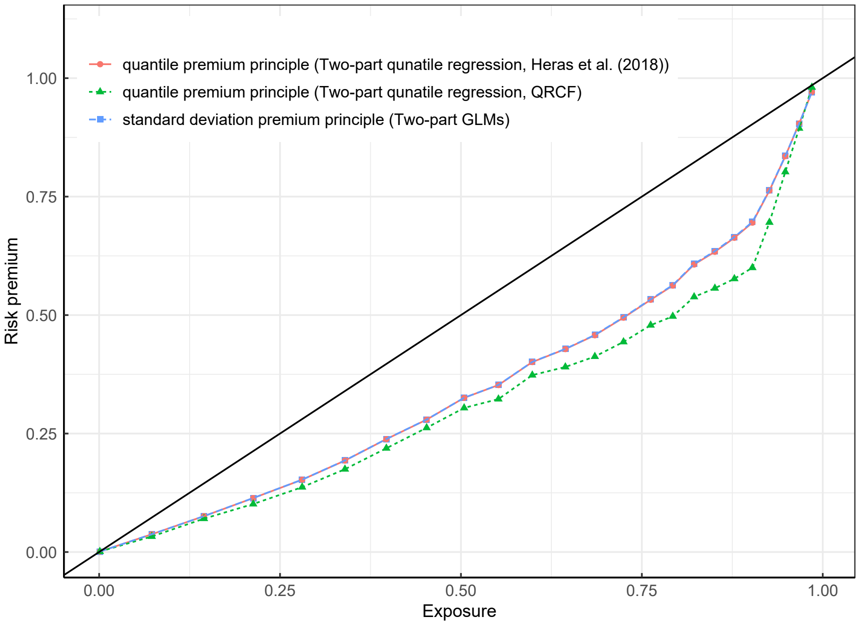

Figure 7: The ordered Lorenz curves of based on two-part quantile regression and two-part GLMs.

6 Conclusion

Risk premium calculation is an important subject in non-life actuarial applications. The popular way to calculate the risk premium is to fit the aggregate claim amounts of individual policies with regression models, such as two-part GLMs and quantile regression models. In the existing methods, the risk premium calculation depends on some parameters, such as the risk loading parameter in the expected value premium principle and standard deviation premium principle, the risk aversion factor in the Wang premium principle, and the quantile level in the quantile premium principle.

This study proposes a general top-down approach to predicting risk premiums of individual policies. First, given a small probability of the portfolio, such as 0.5%, we use a bootstrap method to obtain the total risk premium for the whole portfolio. Second, the total risk premium is allocated to individual policies in accordance with their risk measures. This approach solves the problem of subjectively selecting some relevant parameters in the existing literature. In the empirical analysis, we apply the proposed method to a data set from an Australian insurance company. Given the total risk premium, we find that the predicted risk premiums for individual policies under the expected value premium principle, standard deviation premium principle, and quantile premium principle are very close. However, the basis for calculating the risk loadings is quite different in each method. In the expected value premium principle, the risk loading is a certain proportion of the expected claim amount; in the standard deviation premium principle, the risk loading is a certain proportion of the standard deviation of the claim amount; and in the quantile premium principle, the risk loading is the difference between a quantile and the expected claim amount.

In this study, we suggest that the risk premiums of individual policies be calculated by using two-part quantile regression models. The empirical study shows that the quantile regression models can better reveal the heterogeneity of different risk classes, and thus, can yield relatively reasonable risk premiums for individual polices.

Acknowledgements

Zhengxiao Li and Shengwang Meng acknowledge the financial support from National Natural Science Fund of China (Grant No. 71901064), National Social Science Fund of China (Grant No. 16ZDA052) and MOE National Key Research Bases for Humanities and Social Sciences (Grant No. 16JJD910001)
.

Declaration of interest

We declare that there is no potential conflict of interest in the paper.

References

[1] Baione, F., and Biancalana, D. (2019). An individual risk model for premium calculation based on quantile: A comparison between generalized linear models and quantile regression. North American Actuarial Journal 23(4): 573-590.

[2] Bühlmann, H. (1970). Mathematical methods in risk theory. Berlin: Springer.

[3] Bühlmann, H. (1985). Premium calculation from top down. Astin Bull. 15: 89-101.

[4] Cossette, H., Mailhot, M., and Marceau, É. (2012). TVaR-based capital allocation for multivariate compound distributions with positive continuous claim amounts. Insurance: Mathematics and Economics 50(2): 247-256.

[5] Cummins, J. D., Dionne, G., McDonald, J. B., and Pritchett, B. M. (1990). Applications of the gb2 family of distributions in modeling insurance loss processes. Insurance: Mathematics and Economics 9: 257-272.

[6] De Jong, P., and Heller, G. Z. (2008). Generalized linear models for insurance data. Cambridge: Cambridge University Press.

[7] Dong, A. X., Chan, J. S., and Peters, G. W. (2015). Risk margin quantile function via parametric and non-parametric Bayesian approaches. Astin Bull. 45: 503-550.

[8] Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? Insurance Math. Econom. 51(2): 239-248.

[9] Frees, E. W. (2009). Regression modeling with actuarial and financial applications. Cambridge: Cambridge University Press.

[10] Frees, E. W., Meyers G., and Cummings, A. D. (2011). Summarizing Insurance Scores Using a Gini Index. Journal of the American Statistical Association 106(495): 1085-1098.

[11] Frees, E. W., Jin, X., and Lin, X. (2013). Actuarial applications of multivariate two-part regression models. Annals of Actuarial Science 7: 258-287.

[12] Frumento, P., and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. Biometrics 72: 74-84.

[13] Frumento, P., and Bottai, M. (2017). An estimating equation for censored and truncated quantile regression. Comput. Statist. Data Anal. 113: 53-63.

[14] Gilchrist, W. (2000). Statistical modelling with quantile functions. Boca Raton, FL: Chapman and Hall/CRC.

[15] Gini, C. (1921). Measurement of Inequality of Incomes. The Economic Journal, 31(121): 124-126.

[16] Heller, G., Stasinopoulos, D., and Rigby, R. (2006). The zero-adjusted inverse Gaussian distribution as a model for insurance claims. In Proceedings of the 21st International Workshop on Statistical Modelling, 226-233.

[17] Heras, A., Moreno, I., and Vilar-Zanón, J. L. (2018). An application of two-stage quantile regression to insurance ratemaking. Scand. Actuar. J. 9: 753-769.

[18] Koenker, R., and Bassett, G. (1978). Regression quantiles. Econometrica 46: 33-50.

[19] Koenker, R., and Hallock, K. F. (2001). Quantile regression. Journal of Economic Perspectives 15: 143-156.

[20] Kudryavtsev, A. A. (2009). Using quantile regression for rate-making. Insurance Math. Econom. 45: 296-304.

[21] Laudagé, C., Desmettre, S., and Wenzel, J. (2019). Severity modeling of extreme insurance claims for tariffication. Insurance Math. Econom. 88: 77-92.

[22] Mack, T. (1997). Schadenversicherungsmathematik. Karlsruhe: Versicherungswirtsch.

[23] McCullagh, P., and Nelder, J. A. (1989). Generalized linear models. 2nd Edition, Boca Raton, FL: Chapman and Hall.

[24] McDonald, J. B., and Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. Econometric Theory 4(3): 428-457.

[25] Noufaily, A., and Jones, M. (2013). Parametric quantile regression based on the generalized gamma distribution. J. Roy. Statist. Soc. Ser. C 62: 723-740.

[26] Shi, P., and L. Yang. (2017). Pair copula constructions for insurance experience rating. Journal of the American Statistical Association 113: 122-133.

[27] Stacy, E. W. (1962). A generalization of the gamma distribution. The Annals of Mathematical Statistics 33(3): 1187-1192.

[28] Wang, S. (1995). Insurance pricing and increased limits ratemaking by proportional hazards transforms. Insurance Math. Econom. 17: 43-54.

[29] Wang, S. S. (2000). A class of distortion operators for pricing financial and insurance risks. Journal of Risk and Insurance 67: 15-36.

[30] Wang, S. S., Young, V. R., and Panjer, H. H. (1997). Axiomatic characterization of insurance prices. Insurance Math. Econom. 21: 173-183.

A The framework of two-part GLMs

If g_{Y_i} is a GA density function given by

$$g_{Y_i}\left(y_i;\mu_i,\sigma\right)=\frac{1}{\Gamma\left(1/\sigma^2\ \right)}\left(\sigma^2\mu_i\right)^{-1/\sigma^2}{y_i}^{1/\sigma^2\ -1}\exp\left(-\frac{y_i}{\sigma^2\mu_i}\right), \tag{A.1}$$

where $\Gamma\left(x\right)$ is a complete GA function, that is, $\Gamma\left(x\right)=\int_0^\infty t^{x-1}\exp\left(-t\right)dt$, $\sigma>0$ and $\mu_i>0$, then for $0\leq y_i<\infty$, the mean and variance of aggregate claim amount Y_i are given by

$$\mathbb{E}(Y_i)=\left(1-p_i\right)\mu_i, \tag{A.2}$$

$$\mathrm{Var}(Y_i)=\left(1-p_i\right){\mu_i}^2\left(p_i+\sigma^2\right). \tag{A.3}$$

If g_{Y_i} is an IG density function given by

$$g_{Y_i}\left(y_i;\mu_i,\sigma\right)=\frac{1}{\sqrt{2\pi_i\sigma^2{y_i}^3}}\exp\left[-\frac{\left(y_i-\mu_i\right)^2}{2{\mu_i}^2\sigma^2y_i}\right], \tag{A.4}$$

where $\sigma>0$ and $\mu_i>0$, then for $0\leq y_i<\infty$, the mean and variance of aggregate claim amount Y_i are given by

$$\mathbb{E}(Y_i)=\left(1-p_i\right)\mu_i, \tag{A.5}$$

$$\mathrm{Var}(Y_i)=\left(1-p_i\right){\mu_i}^2\left(p_i+\mu_i\sigma^2\right). \tag{A.6}$$

Generally, IG distribution is more flexible in a skewed model with heavy-tailed data. Due to parametric nature of GLMs, we employ likelihood-based method for estimation with `optim` function in R software. Given a portfolio of N policies, the total log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma) = \prod_{i=1}^{N} \log\left[p_i I\left(y_i = 0\right) + \left(1 - p_i\right) g_{Y_i}\left(y_i; \mu_i, \sigma\right) I\left(y_i > 0\right)\right], \tag{A.7}$$

where $I\left(\cdot\right)$ is the indicator function.

Assuming the independence between the claim possibility R_i and non-zero aggregate claim amount $Y_i | R_i = 1$, the above log-likelihood function can be maximized separately. For claim probability component, the estimates of parameters are given by

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^{k+1}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \log\left[p_i I\left(y_i = 0\right)\right] \right\}. \tag{A.8}$$

For non-zero aggregate claim amount component, the estimates of parameters are given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{k+1}}{\operatorname{argmin}} \sum_{i=1}^{N} \log\left[g_{Y_i}\left(y_i; \mu_i, \sigma\right) I\left(y_i > 0\right)\right], \tag{A.9}$$

$$\hat{\sigma} = \underset{\sigma \in \mathbb{R}^{+}}{\operatorname{argmin}} \sum_{i=1}^{N} \log\left[g_{Y_i}\left(y_i; \mu_i, \sigma\right) I\left(y_i > 0\right)\right]. \tag{A.10}$$

Finally, the estimates of probability of having no claim and the mean parameter in two-part GLMs can be expressed respectively as

$$\hat{p}_i = 1 - \frac{\exp\left(\boldsymbol{x}_i^R \hat{\boldsymbol{\alpha}}\right)}{1 + \exp\left(\boldsymbol{x}_i^R \hat{\boldsymbol{\alpha}}\right)} w_i, \tag{A.11}$$

$$\hat{\mu}_i = \exp\left(\boldsymbol{x}_i^{\mu} \hat{\boldsymbol{\beta}}\right). \tag{A.12}$$

Table 4: Pure Premiums for Different Risk Classes

RiskClass	**Veh_age**	**Agecat**	**Npolicie**	**Nclaims**	**ProbNC**	**PureP**
1	2	1	1504	159	0.798	524.99
2	1	1	1283	111	0.803	484.29
3	3	1	1643	140	0.818	489.82
4	2	2	3167	288	0.828	354.88
5	4	1	1312	115	0.831	499.21
6	1	2	2160	178	0.833	327.06
7	2	3	3741	295	0.837	299.95
8	1	3	2706	212	0.841	276.37
9	2	4	3919	324	0.843	295.68
10	3	2	3956	280	0.846	329.89
11	1	4	2935	180	0.847	272.39
12	3	3	4826	386	0.853	278.54
13	4	2	3592	254	0.857	335.36
14	3	4	4760	349	0.859	274.39
15	4	3	4494	296	0.865	282.96
16	4	4	4575	332	0.870	278.61
17	2	5	2635	182	0.871	213.82
18	2	6	1621	106	0.871	233.62
19	1	5	2042	122	0.874	196.81
20	1	6	1131	73	0.875	215.02
21	3	5	3088	183	0.884	197.75
22	3	6	1791	108	0.885	216.05
23	4	5	2971	161	0.894	200.32
24	4	6	2004	103	0.894	218.85

Notes: Column 4 reports the number of policies, column 5 the number of claims, column 6 the probability of having no claims, and column 7 the pure premiums. The 24 risk classes are ordered by the probability of having no claims.

Table 5: Classification risk premiums using two-part GLMs

RiskClass	**ProbNC**	**PureP**	**EVPP** $\varphi = 3.572\%$		**SDPP** $\varphi = 0.715\%$		**WPP** $\rho = 1.592\%$	
			RiskP	**RiskL**	**RiskP**	**RiskL**	**RiskP**	**RiskL**
1	0.798	524.99	542.51	17.52	543.74	18.75	543.17	18.18
2	0.803	484.29	500.30	16.01	501.59	17.30	501.03	16.74
3	0.818	489.82	507.30	17.48	507.32	17.50	507.20	17.38
4	0.828	354.88	366.64	11.76	367.55	12.68	367.22	12.34
5	0.831	499.21	518.52	19.30	517.04	17.83	517.40	18.18
6	0.833	327.06	337.82	10.76	338.74	11.68	338.42	11.36
7	0.837	299.95	309.72	9.76	310.67	10.71	310.35	10.39
8	0.841	276.37	285.30	8.93	286.24	9.87	285.93	9.57
9	0.843	295.68	305.57	9.89	306.25	10.56	306.03	10.34
10	0.846	329.89	341.60	11.71	341.67	11.78	341.64	11.76
11	0.847	272.39	281.43	9.04	282.12	9.73	281.90	9.52
12	0.853	278.54	288.25	9.71	288.49	9.95	288.44	9.89
13	0.857	335.36	348.25	12.89	347.34	11.98	347.62	12.26
14	0.859	274.39	284.22	9.83	284.19	9.80	284.22	9.83
15	0.865	282.96	293.64	10.67	293.07	10.11	293.27	10.31
16	0.870	278.61	289.41	10.80	288.56	9.95	288.85	10.24
17	0.871	213.82	221.38	7.56	221.46	7.64	221.48	7.66
18	0.871	233.62	242.17	8.55	241.97	8.34	242.08	8.45
19	0.874	196.81	203.72	6.92	203.84	7.03	203.85	7.04
20	0.875	215.02	222.85	7.82	222.71	7.68	222.8	7.77
21	0.884	197.75	205.24	7.50	204.81	7.06	205.00	7.25
22	0.885	216.05	224.53	8.48	223.77	7.72	224.05	8.00
23	0.894	200.32	208.54	8.22	207.48	7.16	207.85	7.53
24	0.894	218.85	228.16	9.30	226.67	7.82	227.16	8.31

Notes: Column 2 reports the probabilities of having no claims. Column 3 reports the pure premium. RiskL and RiskP denotes risk loadings and risk premiums respectively. EVPP, SDPP and WPP denote expected value premium principle, standard deviation premium principle, and Wang premium principle. The risk loading factor φ in EVPP is 3.56% and in SDPP is 0.713%. The risk aversion ρ in WPP is 0.0159%.

Table 6: Classification risk premiums using two-part quantile regressions

RiskClass	ProbNC	**QR**		**PQR**		**QRCF**	
		$\tau = 96.18\%$		$\tau = 96.37\%$		$\tau = 96.08\%$	
		τ_i^*	**RiskP**	τ_i^*	**RiskP**	τ_i^*	**RiskP**
1	0.798	0.811	797.92	0.820	638.76	0.806	845.17
2	0.803	0.806	634.81	0.816	573.88	0.801	685.69
3	0.818	0.790	770.09	0.801	566.41	0.785	757.71
4	0.828	0.777	385.55	0.788	390.23	0.772	400.24
5	0.831	0.773	736.64	0.785	540.97	0.767	728.17
6	0.833	0.771	308.46	0.783	349.49	0.766	325.59
7	0.837	0.766	339.56	0.777	348.96	0.760	342.36
8	0.841	0.759	267.92	0.771	312.21	0.753	279.09
9	0.843	0.757	285.29	0.769	319.87	0.751	300.39
10	0.846	0.752	362.66	0.765	341.59	0.746	350.27
11	0.847	0.751	225.15	0.763	285.98	0.744	245.04
12	0.853	0.739	304.62	0.752	304.18	0.733	298.23
13	0.857	0.732	358.19	0.745	322.98	0.725	332.13
14	0.859	0.729	254.65	0.743	277.99	0.723	259.88
15	0.865	0.717	304.88	0.731	286.66	0.710	281.95
16	0.870	0.706	244.03	0.721	261.36	0.699	244.63
17	0.871	0.704	180.81	0.719	206.28	0.697	174.54
18	0.871	0.703	177.61	0.718	218.04	0.696	184.96
19	0.874	0.696	147.80	0.711	183.65	0.688	143.99
20	0.875	0.695	152.39	0.711	194.11	0.688	153.14
21	0.884	0.669	156.17	0.686	176.15	0.661	148.02
22	0.885	0.669	165.21	0.685	186.14	0.660	159.00
23	0.894	0.641	146.12	0.659	163.27	0.632	136.56
24	0.894	0.640	153.23	0.658	172.50	0.631	148.14

Notes: This table reports the probability of having no claims and risk premiums under different quantile regression models. RiskP denotes risk premiums. QR, PQR and QRCF denote traditional quantile regression, fully parametric quantile regression, and quantile regression with coefficient functions.

Table 7: Classification risk premiums under different models

RiskClass	**Two-part quantile regressions**			**Two-part GLMs**			**Heras et al. (2018)**
	QR	**QRCF**	**PQR**	**EVPP**	**SDPP**	**WPP**	**QPP**
1	797.92	845.17	638.76	542.51	543.74	543.17	549.11
2	634.81	685.69	573.88	500.30	501.59	501.03	502.69
3	770.09	757.71	566.41	507.30	507.32	507.20	511.46
4	385.55	400.24	390.23	366.64	367.55	367.22	367.18
5	736.64	728.17	540.97	518.52	517.04	517.40	521.00
6	308.46	325.59	349.49	337.82	338.74	338.42	336.75
7	339.56	342.36	348.96	309.72	310.67	310.35	311.23
8	267.92	279.09	312.21	285.30	286.24	285.93	285.80
9	285.29	300.39	319.87	305.57	306.25	306.03	305.12
10	362.66	350.27	341.59	341.60	341.67	341.64	342.10
11	225.15	245.04	285.98	281.43	282.12	281.90	280.20
12	304.62	298.23	304.18	288.25	288.49	288.44	289.39
13	358.19	332.13	322.98	348.25	347.34	347.62	347.94
14	254.65	259.88	277.99	284.22	284.19	284.22	283.44
15	304.88	281.95	286.66	293.64	293.07	293.27	293.61
16	244.03	244.63	261.36	289.41	288.56	288.85	287.87
17	180.81	174.54	206.28	221.38	221.46	221.48	221.01
18	177.61	184.96	218.04	242.17	241.97	242.08	241.81
19	147.80	143.99	183.65	203.72	203.84	203.85	202.55
20	152.39	153.14	194.11	222.85	222.71	222.80	221.60
21	156.17	148.02	176.15	205.24	204.81	205.00	204.18
22	165.21	159.00	186.14	224.53	223.77	224.05	223.69
23	146.12	136.56	163.27	208.54	207.48	207.85	206.50
24	153.23	148.14	172.50	228.16	226.67	227.16	226.42
Gini index	34.7%	**35.67%**	30.97%	29.68%	29.7%	29.69%	29.78%