STOCHASTIC MODELING OF VEHICLE INSURANCE CLAIMS: AN APPLICATION IN THE SOUTH OF MINAS GERAIS/BRAZIL

Luiz Otávio de Oliveira PALA[1]
Marcela de Marillac CARVALHO[1]
Thelma SÁFADI[1]

▪ *ABSTRACT: Risk and exposure factors are important features to be considered, providing financial and actuarial information for the insurer. Pricing methods are supported by the mutualism theory, ensuring a level of indemnity and expected cost, making possible to constitute monetary reserves. The aim of our paper is to model and analyze the distribution of vehicle insurance claims in the south of Minas Gerais/Brazil. The data represents policies with a claim occurrence in the year of 2018. Under the Bayesian approach, we consider the Gamma and Log-normal distributions that allow asymmetric data modeling and they can be used in loss models. The Jeffreys's prior class was applied considering the data of the first semester of 2018. The information level was updated to construct an informative prior to analyze the data of the second semester. To compare models, we estimated the Bayes Factor and the logarithm of the marginal likelihood, that showed the Log-normal more likely. After selecting a model, we estimate metrics as the Conditional Tail Expectation (CTE) and the percentiles of the adjusted distribution to evaluate extreme costs. The results showed the applicability of Bayesian inference to fit insurance data, allowing to insert prior knowledge as the portfolio experience and to use a wide class of probability distributions.*

▪ *KEYWORDS: Bayesian inference; financial reserves; risk levels.*

1 Introduction

The development of statistical models in actuarial science has contributed to the pricing process, evaluating of financial reserves and risk management

[1] Universidade Federal de Lavras - UFLA, Departamento de Estatistica, CEP: 37200-900, Lavras, MG, Brasil. E-mail: *luizotavio.oliveira@gmail.com, marcela-carvalho_@hotmail.com, safadi@ufla.br*

(HABERMAN; RENSHAW, 1996; MAKOV; SMITH; LIU, 1996; Ravi, 2009; LALLY; HARTMAN, 2018; SRIRAM; SHI, 2020; HUANG; MENG, 2020). According to Klugman, Panjer and Willmot (2012) an important factor in the actuarial context is related to construct models that can predict future costs, allowing to estimate financial reserves and generating information for risk policies and resource allocation.

In the non-life market, specifically in car insurance, regression models can be applied to predict premiums and to segment portfolios according to risk levels such as fire, theft and collision. Pala *et al.* (2020), for example, have explored insurance data to predict the occurrence of a claim using algorithms of supervised learning as random forest and logistic regression, and Teixeira and Scalon (2016) have analyzed the spatial dependence of car insurance premiums in *Minas Gerais*, Brazil.

In the case of collisions, there is a segmentation in partial collisions and total loss. Peres, Maldonado and Cândido (2019) have pointed the importance of the insurance market for the Brazilian economy, analyzing conditions as demand and competitive; as well as Teixeira and Scalon (2016), have indicated that the state of *Minas Gerais* has contributed to the Brazilian insurance market.

An important topic is the modeling of loss distributions, frequently heavy-tailed distributions, with the purpose of inferring about the risk level (AHMAD; MAHMOUDI; HAMEDANI, 2019). In the literature, we found studies using the Pareto, Gamma, Log-normal, Weibull and Burr distributions (GUILLEN; PRIETO, 2011; AHMAD; MAHMOUDI; HAMEDANI, 2019). Besides to the analysis of risk measures, that are related to the distributions tails, it is possible to build confidence intervals to the monetary reserves (MEYERS, 2007).

As seen in Klugman, Panjer and Willmot (2012), in several actuarial studies the main interest is associated with the positive tail and the presence of extreme values. The occurrence of extreme costs has a considerable impact on the total loss. For this, stochastic models are used to estimate and construct metrics as the Value-at-Risk (VaR), Conditional-Value-at-Risk (CVaR), Conditional Tail Expectation (CTE), and the quantiles of the fitted distribution.

The CTE is a risk measure that quantify the expected loss of a random variable Y, given that Y exceeds a specific limit, it is, $CTE_Y(t) = E[Y|Y > F^{-1}(t)]$ for $t \in (0,1)$, associating the risk of (1 - t)100% of loss above of $F^{-1}(t)$ (BRAZAUSKAS *et al.*, 2008). In the insurance context, this metric allows to estimate the expected value of extreme indemnities setting values for t, as 0.90, 0.95, and 0.99. In addition to this measure, the Value-at-Risk is extensively used in the insurance market and in financial industry, being defined as the t quantile of the distribution of Y (CHUN; SHAPIRO; URYASEV, 2012).

Stochastic models have been applied in actuarial modeling for some time (GAO, 2018), permitting solutions for the measurement of uncertain events (SCOLLNIK, 2001). The Bayesian framework has enabled the estimation of non-life reserves (GAO, 2018; HUANG; MENG, 2020), with the insertion of a prior knowledge, such as the portfolio experience, risks and patterns. As an example of applications, we found the study of Makov (2001), using Bayesian inference; Alba

(2002), modeling the severity and frequency, Hunt and Blake (2020) estimating the mortality rates and Huang and Meng (2020) using the Bayesian framework to predict the loss.

In the Bayesian approach, the inferential process is carried out based on a learning process, relating the data to the previously available information (KODUVELY, 2015), resulting in a posterior distribution. In analyzes where there is a low level of knowledge about the phenomenon, non-informative priors can be used as the Jeffreys's class, being invariant to transformations 1 to 1.

The posterior distribution is obtained by applying the Bayes Theorem, and the inference is performed with the marginal posterior distribution of each parameter, that is obtained using integration. However, in most cases, these integrals are complex, and algorithms as Markov Chain Monte Carlo (MCMC) are run to get moments of the marginal distributions (NASCIMENTO *et al.*, 2011). Details of these sampling algorithms are available in Casella and George (1992), Gamerman and Lopes (2006).

In this study, we present aspects of the Bayesian inference for modeling vehicle insurance claims related to collision with total loss, which occurred in the south of *Minas Gerais*/Brazil. The aim is to fit and analyze the distributions that are used by actuarial studies: Gamma and Log-normal. These distributions are flexible and allow to fit insurance data, as seen in Marlin (1984), Zuanetti, Diniz and Leite (2006), Gilenko and Mironova (2017) and they can be used in loss models, see for instance the studies of Punzo, Bagnato and Maruotti (2017) using compound unimodal distributions such as Gamma, Log-normal and Inverse Gaussian; Punzo, Mazza and Maruotti (2018) fitting insurance and economic data with contaminated Gamma distributions and Goffard and Laub (2021) comparing insurance loss models.

2 Risk modeling and Inference procedure

We use the data provided by Superintendência de Seguros Privados (2020), corresponding to indemnities in vehicle claims in the south of *Minas Gerais*/Brazil for total loss. We consider the observations of 2018, totaling 806 observations in the first semester and 881 in the second semester. Policies with no loss, that is, without indemnities, were disregarded from the analysis and for policies with more than one occurrence, the total amount of the indemnity was considered. Our interest is to model the total cost and not the severity, which is the average cost per claim.

2.1 Models

In this section, we are going to describe the models and the prior distributions that were considered in our study. The presented models have the total amount of the indemnity as a response variable, namely y, and we do not consider explanatory variables. We also present the algebraic aspects for obtaining these distributions in the attached section.

We performed the analysis of the first semester of 2018 adopting the Gamma and Log-normal distributions for the response variable, presented in subsections 2.1.1 and 2.1.2, considering an approach with a non-informative prior.

2.1.1 Gamma model

The Gamma(α, β) is a flexible distribution, used to fit asymmetric and survival data. The distribution is related to the exponential, normal and chi-square distributions (CASELLA; BERGER, 2014), defined in the parametric space $\mathbf{\Phi} = \{\alpha, \beta | \alpha, \beta > 0\}$, being $y > 0$. The probability density function is given by 1.

$$p(y|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{(\alpha-1)} e^{(-\beta y)}. \quad (1)$$

Considering the set of independent observations of $y = \{y_i, i = 1, \dots, n\}$, the likelihood function can be written as 2:

$$L(\alpha, \beta|y) = \frac{\beta^{\alpha n}}{[\Gamma(\alpha)]^n} \left(\prod_{i=1}^{n} y_i^{\alpha-1} e^{-\beta \sum_{i=1}^{n} y_i} \right). \quad (2)$$

The data analysis of the first semester of 2018 was performed considering a non-informative distribution for $\boldsymbol{\theta} = (\alpha, \beta)^{\top}$ with the Jeffreys's method, that is a kind of non-informative prior being invariant under one-to-one transformations, defined from the Fisher information matrix, where p($\boldsymbol{\theta}$) $\propto det[I(\boldsymbol{\theta})]^{\frac{1}{2}}$. Hence, considering the logarithm of $L(\alpha, \beta|y)$, we have the result presented by 3.

$$I(\boldsymbol{\theta}) \propto \begin{bmatrix} \Psi^{'}(\alpha) & \frac{-1}{\beta} \\ \frac{-1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix} \therefore p(\boldsymbol{\theta}) \propto \frac{\sqrt{\alpha \Psi'(\alpha) - 1}}{\beta}, \quad (3)$$

being $\Psi^{'}(\alpha) = \sum_{n=0}^{\infty} \frac{1}{(\alpha+n)^2}$ the trigamma function. In Figure 1(a) there is a representation of $p(\boldsymbol{\theta})$ for pairs of α and β.

Therefore, the posterior distribution, $p(\boldsymbol{\theta}|y)$, using the Bayes rule is given by 4:

$$p(\boldsymbol{\theta}|y) \propto \beta^{(n\alpha-1)} e^{-\beta \sum_{i=1}^{n} y_i} \prod_{i=1}^{n} yi^{\alpha-1} \sqrt{\Psi'(\alpha)\alpha - 1} \frac{1}{[\Gamma(\alpha)]^n}. \quad (4)$$

2.1.2 Log-normal model

The Log-normal(μ, σ) distribution is commonly used in applications of positive asymmetric data as income and lifetime, see for instance Casella and Berger (2014), the probability density function is given by 5.

$$p(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma y} e^{\frac{-(\log(y)-\mu)^2}{2\sigma^2}}. \quad (5)$$

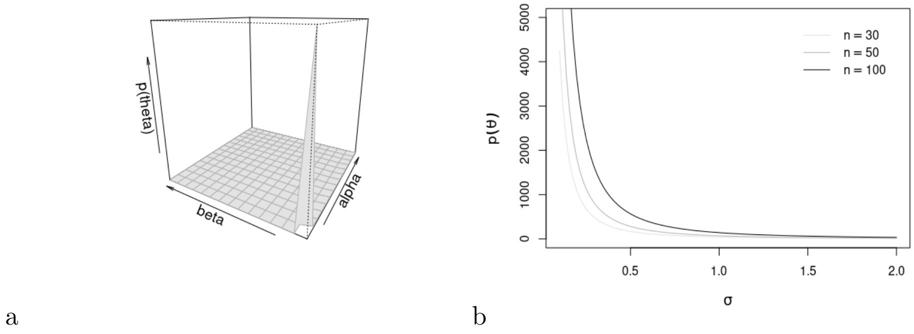

Figure 1: Behavior of the prior distributions obtained with the Jeffreys's method, considering the Gamma (a) and Log-normal (b) likelihoods.

It is defined in the parametric space $\boldsymbol{\Phi} = \{\mu, \sigma | -\infty < \mu < +\infty, \sigma > 0\}$. So, considering the set of observations $y = \{y_i, i = 1, ..., n\}$, the likelihood function can be written as showed in the equation 6.

$$L(\mu, \sigma | y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{\prod_{i=1}^n y_i}\right) e^{\frac{-\sum_{i=1}^n (\log(y_i) - \mu)^2}{2\sigma^2}}. \tag{6}$$

Specifying the p($\boldsymbol{\theta}$) distribution to analyze the data of the first semester, being $\boldsymbol{\theta} = (\mu, \sigma)^\top$, with the Jeffreys's method and considering the expected value of the expressions: $E\left(\sum_{i=1}^n (\log(y_i) - \mu)^2\right) = n\sigma^2$ and $E\left(\sum_{i=1}^n \log(y_i)\right) = n\mu$, we have that $p(\boldsymbol{\theta}) \propto \frac{n\sqrt{2}}{\sigma^2}$. In figure 1(b) we also have illustrated $p(\boldsymbol{\theta})$ considering n = $\{30, 50, 100\}$. Then, using the Bayes rule, the posterior distribution is given by 7.

$$p(\boldsymbol{\theta}|y) \propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{\prod_{i=1}^n y_i}\right) \left(\frac{n\sqrt{2}}{\sigma^2}\right) \left(e^{\frac{-\sum_{i=1}^n (\log(y_i) - \mu)^2}{2\sigma^2}}\right). \tag{7}$$

To analyze the data of the second semester and update the information level of $\boldsymbol{\theta}$, we selected the better model (Gamma or Log-normal) according to the results obtained in the analysis of the first semester and used it for the next. The Bayes factor (BF) was applied to selected a model, through the ratio of the marginal likelihoods, given by 8:

$$BF_{ij} = \frac{p(y|M_i)}{p(y|M_j)}, \tag{8}$$

for the models i and j. The marginal likelihood was estimated using the average and mode of the distributions. We also use the AICM selection criteria (RAFTERY *et al.*, 2007), which is estimate by AICM $= 2(\bar{l} - s_l^2)$, being $\bar{l}$ e $s^2 l$ the average and the variance of the marginal likelihood, respectively.

After selecting a model, we verified the behavior of the marginal densities of $\boldsymbol{\theta}$ and used a normal kernel with parameters settled according to the results obtained previously, resulting in an informative way and enabling to consider the past information as source of information.

2.2 Inference and convergence

To estimate the unknown parameters of each considered model, we have used the Metropolis-Hastings (MH) algorithm with a transition kernel, q($\boldsymbol{\theta}, \beta$), bivariate normal centered on $\boldsymbol{\theta}$ and with a small standard deviation. For details and properties about Markov Chain Monte Carlo (MCMC) methods, see Korn; Korn and Kroisandt (2010). The inference was based on the joint posterior distribution and we consider 10,000 samples with a burn-in period of 5,000 samples and thin = 5, in a way to decrease the autocorrelation.

The convergence analysis was performed using the Dependence Factor (ϕ) proposed by Raftery and Lewis (1992) and the diagnostic of Heidelberger and Welch (1983), i.e. (HW). The Dependence Factor is understood as the proportional increase in the number of iterations attributable to autocorrelation, where values of $\phi > 5$ can indicate convergence problems and values of $\phi \gg 1$ suggest a higher dependence (RAFTERY; LEWIS, 1992). And the HW criteria is based on the spectral density, constructing a confidence interval for the mean of $\boldsymbol{\theta}$ and testing if the sequence can be considered stationary.

2.3 Risk measures

To analyze and evaluate the risk of extreme events, we estimate the 0.95 and 0.99 quantiles of the fitted distribution, allowing to evaluate the tails and construct exposure metrics for this portfolio. We also estimate the Conditional Tail Expectation considering the 0.95 and 0.99 levels, making possible to evaluate the average cost of extreme indemnities. For this, we use the CTE definition showed in Brazauskas *et al.* (2008) available in 9:

$$CTE_Y(t) = E\left[Y \mid Y > F_Y^{-1}(t)\right], \quad t \in (0,1), \tag{9}$$

being F_Y^{-1} the quantile function, it is $F_Y^{-1}(t) = inf\{y : F_Y(y) \geq t\}$. According to Necir, Rassoul and Zitikis (2010), since the cumulative distribution function is continuous, we have the following result showed in 10:

$$CTE_Y(t) = \frac{1}{1-t}\int_t^1 F_y^{-1}(s)ds. \tag{10}$$

Necir, Rassoul and Zitikis (2010) added that, in some cases, can be desirable to construct confidence intervals for CTE. Additionally, we analyze the semivariance under (ν_1) and above (ν_2) the mean and the ratio of them, as performed by Pla-santamaria and Bravo (2013) in portfolio optimization. ν_1 and ν_2 can be estimated as shown in 11:

$$\nu_1(Y) = \sum_{y_i < \bar{y}} \frac{(y_i - \bar{y})^2}{n}, \text{ and } \nu_2(Y) = \sum_{y_i > \bar{y}} \frac{(y_i - \bar{y})^2}{n}. \tag{11}$$

According to Francis and Kim (2013), the semivariance is a measure of dispersion that can be comparable to the variance. The quantities ν_1 and ν_2 measure the dispersion under and above the mean, respectively. The ratio of them, it is $\frac{\nu_1}{\nu_2}$, allows to quantify the risk in the tails and to compare portfolios according to the risk. In this study, we use the R software, R Core Team (2021), with the package Coda of Plummer *et al.* (2006).

3 Data analysis

In the first half of 2018 were reported 806 claims of total loss in the south of *Minas Gerais*, Brazil. This fact generates an average indemnity of R$ 46,855.15. As expected, the skewness and kurtosis metrics suggested an asymmetric positive and the existence of a heavy tail, shown in Table 1. We found that there are several amounts of indemnities around R$24,053.23, also displayed in Table 1.

Table 1: Descriptive statistics of the collision indemnities reported in the first half of 2018, south of *Minas Gerais*, Brazil

Mean	**Mode**	**Asymmetry**	**Kurtosis**	**SD**
46,885.140	24,053.230	3.245	15.321	45,182.340

The indemnities are arranged in the interval between R$54 and R$386,900.00, and approximately 55.30% of the claims came from male policyholders, 31.30% from female and 13.40% from polices with no information about the gender. Looking at the mean by group, the group with no information about gender had the largest indemnity average, reaching R$69,333.19. For female and male groups, the estimated averages were R$ 37,870.42 and R$46,784.96.

In the Figures 5(a) and 5(b) we present the chains of the marginal distributions for the Gamma model. Similarly, in 5(c) and 5(d) are displayed the chains for parameters μ and σ in the Log-normal model, respectively. Graphically, we can verify a stationary behavior of the simulated chains, shown in Figure 5. In Table 2 are shown the statistics of the models, as the posterior means, credibility intervals and the convergence criteria. In relation to the convergence, all simulated chains have presented $\phi < 5$ and the HW criteria also indicates a stationary distribution, considering a α level of 5%.

According to the Table 2, the shape parameter, α, of the Gamma model was estimated in 1.693 and the β parameter was estimated in 3.6e-5. Taking the expected value of the distribution, we have a mean of R$46,912.13 in the range of R$ 45,924.09 and R$ 47,419.84, according to the credibility interval. Additionally, the Log-normal distribution was estimated with parameters $\mu = 10.432$ and $\sigma =$

Table 2: Summary of the Gamma and Log-normal models considering Jeffreys's priors, applied to the data of the first semester of 2018

Model	**Parameter**	**Mean**	**SD**	**HPD**(95%)		ϕ	HW
				LL	UL		
Gamma	α	1.693	0.077	1.541	1.841	3.180	0.148
	β	3.6e-5	2e-6	3.2e-5	4.0e-5	1.870	0.258
Log-normal	μ	10.432	0.031	10.374	10.493	1.340	0.996
	σ	0.862	0.021	0.818	0.903	1.220	0.470

0.862, resulting in an average cost of R$49,244.02, varying in the interval between R$44,588.88 and R$54,253.61, according to the credibility interval.

We also have estimated the mode of the fitted distributions. For the case of the most frequent monetary amount, we obtained the following result: R$23,173.97 and R$21,030.96, for the Gamma and Log-normal models. Comparing with the estimated mode, the Gamma model performed better when compared to the Log-normal.

In Figure 2 (a) are shown the estimated densities of the distributions compared with the indemnities. We noted that both models were close to the behavior of the indemnities. However, the Log-normal seems to be closer to the mode of the values.

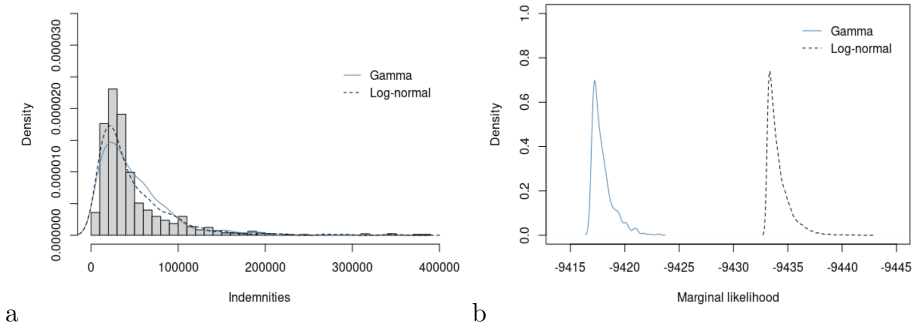

Figure 2: Behavior of the fitted distributions compared to the real values of indemnities (a) and marginal likelihood of the Gamma and Log-normal models (b).

To compare them we use the Bayes Factor with the ratio of the logarithm of the marginal likelihood function, graphically shown in the Figure 2 (b). Considering the average of the marginal likelihood distributions, the BF = 0.998, indicating an evidence to the Log-normal. The same decision is reported when we estimate the BF using the mode of the distributions (BF = 0.998). In addition, the BF in both cases → 1, point in which there is no evidence of superiority in the models considered.

Besides, the AICM criteria were estimated in -18,837.88 and -18,870.56 for Gamma and Log-normal models.

Therefore, selecting the Log-normal model to fit the insurance indemnities, it is possible to create metrics to analyze the portfolio risk, as the quantiles of the fitted distribution. The quantile (95%) was estimated in R$140,388.00 demonstrating that in the analyzed period there was a probability of 5% to an indemnity be greater than R$140,388.00. For the quantile 99%, there was a probability of 1% to an indemnity be greater than R$253,192.30. These amounts interest the insurer, to establish risk and pricing policies for the portfolio.

We also calculated the CTE considering the Log-normal distribution, estimated at R$ 214,271.30 for $\alpha_1 = 0.95$ and R$ 354,093.40 for $\alpha_2 = 0.99$. It is, given that the indemnity value exceeds the quantile α_1, the average indemnity is expected to be R$ 214,271.30. For a extreme value, α_2, the average indemnity is expected to be R$354,093.40.

In addition to these measures, the ratio of the semivariance above the mean to semivariance below the mean was estimated in 4.36, indicating that the dispersion of the indemnities above the mean are 4.36 higher than the dispersion below the mean. Regarding to the estimated median, it shows that 50% of the total cost with indemnities are lower than R$ 33,948.79.

Considering the result obtained in the data analysis of the first semester, we proceed with the analysis of the second semester. In the second half, were reported 881 claims of total loss, with an average cost of R$ 47,697.72. A brief summary of the data is available in Table 3, where 55.05% of the claims came from male policyholders, 31.78% from female and 13.17% from polices with no information about gender.

Table 3: Descriptive statistics of the collision indemnities reported in the second half of 2018, south of *Minas Gerais*, Brazil

Mean	**Mode**	**Asymmetry**	**Kurtosis**	**SD**
47,697.72	24,776.27	2.509	8.334	41,571.42

With the information level constructed previously for the Log-normal model, available in Table 2, we updated $\boldsymbol{\theta} = (\mu, \sigma)^\top$ considering normal priors centered on the estimated values as hyperparameters. It is: $\mu \propto exp\left[\frac{-1}{2}\left(\frac{\mu-10.432}{0.031}\right)^2\right]$ and $\sigma \propto exp\left[\frac{-1}{2}\left(\frac{\sigma-0.862}{0.021}\right)^2\right]$. Hence, the posterior distribution considering a Log-normal likelihood is given by 12:

$$p(\boldsymbol{\theta}|y) \propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{\prod_{i=1}^n y_i}\right) \left(e^{\frac{-\sum_{i=1}^n(\log(y_i)-\mu)^2}{2\sigma^2} - \frac{1}{2}\left(\frac{(\mu-10.432)^2}{0.031} + \frac{(\sigma-0.862)^2}{0.021}\right)}\right). \tag{12}$$

In the Table 4 we presented the estimated values considering the data of the second semester, including the highest posterior density intervals and the convergence criteria. According to the values of ϕ and HW, a stationary distribution is reached considering an α level of 5%. The trace of the simulated chains of each parameter is shown in Figure 3. Comparing these results with the Log-normal model fitted to the first semester (presented in Table 2), we noted an increase of 0.029 in the average parameter (μ) and a reduction in the standard deviation parameter (σ). Reductions of 48.75% and 8.10% were observed in the range of the credibility intervals for μ and σ, respectively.

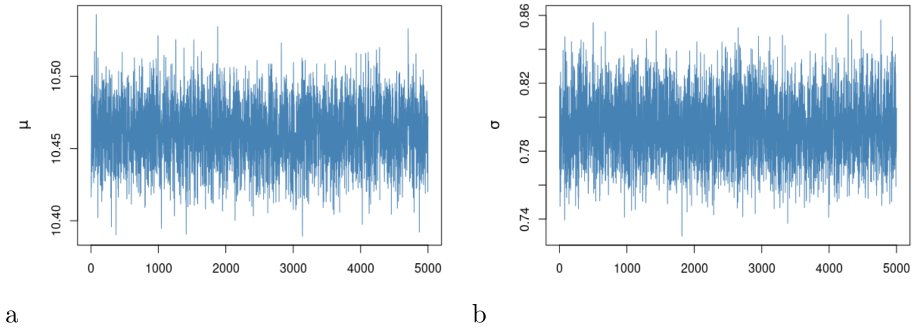

Figure 3: Trace of the chains for the parameters μ (a), σ (b) of the model presented in the Table 4.

Table 4: Summary of the Log-normal model with an informative prior, applied to the data of the second semester of 2018

Model	**Parameter**	**Mean**	**SD**	**HPD**(95%)		ϕ	HW
				LL	UL		
Log-normal	μ	10.461	0.020	10.421	10.501	1.570	0.116
	σ	0.793	0.019	0.757	0.831	1.280	0.510

In Figure 4 (a) are shown the fitted distribution and the real values of indemnities. Additionally, in 4 (b) we constructed the Q-Q plot, that shows a deviation in tails. Taking the expected value of the fitted distribution we have an average cost of R$ 47,858.48, in an interval between [R$ 44,748.44; R$51,440.28]. Comparing with the real average of R$ 47,697.72, we obtained an error of 0.337%.

Regarding to the risk metrics, the quantile (95%) was estimated in R$ 125,585.1, a reduction of 7.946% compared with the first semester of 2018. The quantile (99%) also was reduced, estimated in R$ 218,742.5, a variation of -15.182%. Consequently, the CTE has been reduced to R$182,781.00 considering $\alpha_1 = 0.95$ and R$ 286,208.90 for $\alpha_2 = 0.99$, indicating a reduction in the occurrence of extreme values.

The ratio of the semivariance above the mean to semivariance below the mean was 3.76. It indicates that the dispersion above the mean is approximately 3.76 higher than the payments below the mean. Comparing this result with the estimated ratio in the first semester, we noted a reduction in the ratio. However, the estimated median was R$ 34,840.60, indicating an increase in the median value.

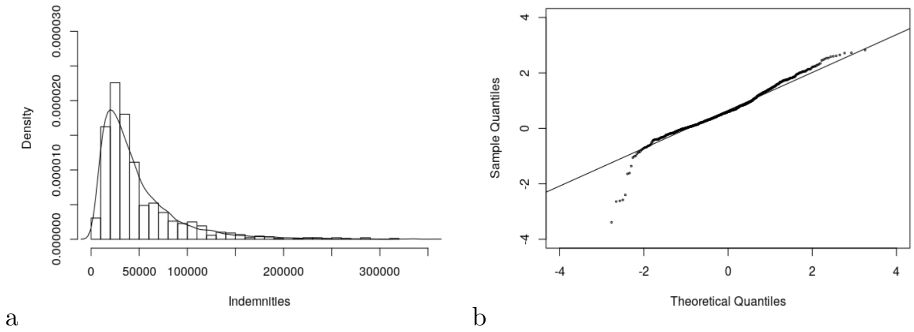

Figure 4: (a): Behavior of the Log-normal distribution compared to the real data of the second semester, and (b): Quantile-quantile plot of the residuals of the model fitted in (a).

Conclusions

In this study, we have pointed aspects of stochastic modeling in insurance claims, through the Bayesian approach, which provides the insertion of a prior knowledge, such as the experience of portfolios that can contribute to the analysis and enable the updating of the information.

Two distributions used for the modeling of positive asymmetric phenomena were considered and applied to collision insurance data for total loss. However, these models can be used in other regions of the country or periods, assuming the posterior behavior presented in our study as a probable source of information.

In addition, new distributions can be applied and proposed in order to analyze the adherence of the tails of the distribution, such as the Burr family, Generalized Gamma and Weibull, allowing the comparison and analysis of the tails, associating risk measures and expected levels of extreme events.

Acknowledgments

This study was financed in part by the *Coordenção de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

PALA, L. O. O.; CARVALHO, M. M.; SAFADI, T. STOCHASTIC MODELING OF VEHICLE INSURANCE CLAIMS: AN APPLICATION IN THE SOUTH OF MINAS GERAIS/BRAZIL. *Braz. J. Biom.,* Lavras, v.40, n.3, p.256-272, 2022.

- RESUMO: Fatores de risco e exposiçao são pontos importantes a serem considerados, oferecendo informações financeiras e atuariais para seguradoras. Métodos de precificação são fundamentados na teoria do mutualismo, possibilitando níveis de indenização, custos esperados e a constituição de reservas montetárias. O objetivo deste trabalho é modelar e analisar a distribuição de indenizações de seguro de veículos no sul de Minas Gerais, Brasil. Os dados representam apólices com ocorrência de sinistros no ano de 2018. Sob o enfoque Bayesiano, foram considerados os modelos Gama e Lognormal, que possibilitam a modelagem de dados assimétricos e são comumente utilizados em modelos de perda. Para o estabelecimento de distribuições a priori, recorreu-se a classe de prioris não informativas de Jeffreys considerando os dados do primeiro semestre de 2018. O nível de informação foi atualizado, construindo informações a priori para analisar os dados do segundo semestre. A comparação dos modelos foi realizada a partir do Fator de Bayes e da razão entre o logaritmo das verossimilhanças marginais, que indicaram o modelo Lognormal mais plausível. Posteriormente, foram calculadas métricas como a Conditional Tail Expectation (CTE) e os percentis da distribuição ajustada, que permitem avaliar níveis de risco, custos extremos e medidas de gerenciamento de reservas monetárias. Ademais, os resultados mostraram a aplicabilidade da inferência Bayesiana na modelagem de dados de seguro, permitindo a inserção de informações a priori, como o histórico de carteiras, e uso de diversas famílias de distribuições.

- PALAVRAS-CHAVE: Inferência Bayesiana, reservas financeiras, níveis de risco.

References

AHMAD, Z. A.; MAHMOUDI, E. M.; HAMEDANI, G. G. A family of loss distributions with an application to the vehicle insurance loss data. *Pakistan Journal of Statistics and Operation Research*, v.15, n.3, p.731–744, 2019.

ALBA, H. Bayesian estimation of outstanding claim reserves. *North American Actuarial Journal*, v.6, n.4, p.1–20, 2002.

BRAZAUSKAS, V.; JONES, B.; PURI; M., ZITIKIS, R. Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, Elsevier, v.138, n.11, p.3590–3604, 2008.

CASELLA, G.; BERGER, R. *Inferência Estatística.* [S.l.]: Cengage Learning, 2014.

CASELLA, G.; GEORGE, E. I. Explaining the gibbs sampler. *The American Statistician*, v.46, n.3, p.167–174, 1992.

CHUN, S.; SHAPIRO, A.; URYASEV, S. Conditional value-at-risk and average value-at-risk: Estimation and asymptotics. *Operations Research*, v.60, n.4, p.739-756, 2012.

FRANCIS, J.; KIM, D. *Modern Portfolio Theory: Foundations, Analysis, and New Developments*. John Wiley & Sons, 2013.

GAMERMAN, D.; LOPES, H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. [S.l.]: Chapman & Hall, 2006.

GAO, G. *Bayesian Claims Reserving Methods in Non-life Insurance with Stan*. Berlin: Springer, 2018.

GILENKO, E.; MIRONOVA, E. Modern claim frequency and claim severity models: An application to the Russian motor own damage insurance market. *Cogent Economics & Finance*, v.5, n.1311097, p.1–12, 2017.

GOFFARD, P.; LAUB, P.Approximate Bayesian Computations to fit and compare insurance loss models. *Insurance: Mathematics and Economics*, v. 100, p.350-371, 2021.

GUILLEN, M.; PRIETO, F. Modelling losses and locating the tail with the pareto positive stable distribution.*Insurance Mathematics & Economics*, v.49, p.454–461, 11 2011.

HABERMAN, S.; RENSHAW, A. Generalized linear models and actuarial science. *Journal of the Royal Statistical Society*, v.45, p.407–436, 1996.

HEIDELBERGER, P.; WELCH, P.Simulation run length control in the presence of an initial transient. *Operations Research*, v. 31, n.6, p.1109–1144, 1983.

HUANG, Y.; MENG, S. A bayesian nonparametric model and its application in insurance loss prediction.*Insurance: Mathematics and Economics*, v. 93, p.84 – 94, 2020.

HUNT, A.; BLAKE, D. A Bayesian approach to modeling and projecting cohort effects. *North American Actuarial Journal*, p.1–20, 2020.

KLUGMAN, S.; PANJER, H.; WILLMOT, G. *Loss Models*. [S.l.]: John Wiley & Sons, 2012.

KODUVELY, H. M. *Learning Bayesian Models with R*. [S.l.]: Birmingham: Packt Publishing, 2015.

KORN, R.; KORN, E.; KROISANDT, G. *Monte Carlo methods and models in finance and insurance*. [S.l.]: CRC press, 2010.

LALLY, N.; HARTMAN, B. Estimating loss reserves using hierarchical Bayesian gaussian process regression with input warping. *Insurance: Mathematics and Economics*, v.82, p.124 – 140, 2018.

MAKOV, U. Principal applications of Bayesian methods in actuarial science. *North American Actuarial Journal*, v.5, n.4, p.53–57, 2001.

MAKOV, U. E.; SMITH, A.; LIU, Y.-H. Bayesian methods in actuarial science. *Journal of the Royal Statistical Society*, v.45, n.4, p. 503–515, 1996.

MARLIN, P.Fitting the log-normal distribution to loss data subject to multiple deductibles. *The Journal of Risk and Insurance*, v.51, n.4, p.687–701, 1984.

MEYERS, G. G. Estimating predictive distributions for loss reserve models. *Variance*, v.1, n.2, p.248–272, 2007.

NASCIMENTO, M.; FONSECA E SILVA, F.; SÁFADI, T.; NASCIMENTO, A.; FERREIRA, R.; CRUZ, C. Abordagem bayesiana para avaliação da adaptabilidade e estabilidade de genótipos de alfafa. *Pesquisa Agropecuária Brasileira*, v.46, p.26 – 32, 2011.

NECIR, A.; RASSOUL, A.; ZITIKIS, R. Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics*, v.2010, 2010.

PALA, L.; CARVALHO, M.; GUIMARÃES, P.; SÁFADI, T. Vehicle claims in the south of Minas Gerais: An approach using classification models. *Semina: Ciências Exatas e Tecnológicas*, v.41, n.1, p.79–86, 2020.

PERES, V.; MALDONADO, W.; CÂNDIDO, O. Seguros de automóvel no Brasil: Concentração e demanda de mercado. *Revista Contabilidade e Finanças*, v.30, n.81, p.396–408, 2019.

PLA-SANTAMARIA, D.; BRAVO, M. Portfolio optimization based on downside risk: a mean-semivariance efficient frontier from Dow Jones blue chips. *Annals of Operations Research*, v.205, n.1, p.189-201, 2013.

PLUMMER, M.; BEST N. COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for MCMC. *R News*, v.6, n.1, p.7–11, 2006.

PUNZO, A.; BAGNATO, L.; MARUOTTI, A. Compound unimodal distributions for insurance losses. *Insurance: Mathematics and Economics*, v.81, p.95-107, 2017.

PUNZO, A.; MAZZA, A.; MARUOTTI, A. Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions. *Journal of Applied Statistics*, v.45, n.14, p. 2563-2584, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2021. Available at: ⟨https://www.R-project.org/⟩.

RAFTERY, A.; NEWTON, M.; SATAGOPAN, J.; KRIVITSKY, P.Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian statistics*, v.8, p.1–45, 2007.

RAFTERY, A. E.; LEWIS, S. M. One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statist. Science*, The Institute of Mathematical Statistics, v.7, n.4, p.493–497, 1992.

RAVI, S. Statistical and Probabilistic Methods in Actuarial Science. *Journal of the Royal Statistical Society Series A*, v.172, n.2, p.530–530, 2009.

SCOLLNIK, D. Actuarial modeling with mcmc and bugs. *North American Actuarial Journal*, v.5, n.2, p.96–124, 2001.

SRIRAM, K.; SHI, P. Stochastic loss reserving: A new perspective from a dirichlet model. *Journal of Risk and Insurance*, p.1–36, 2020.

Superintendência de Seguros Privados. *Autoseg: Sistema de Estatísticas de Automóveis*. [S.l.], 2020. Available at: http://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx.

TEIXEIRA, F. J.; SCALON, J. D. A dependência espacial do valor do prêmio de automóvel. *Revista Brasileira Risco e Seguro*, v.11, n. 20, p.29–54, 2016.

ZUANETTI, D.; DINIZ, C.; LEITE, J. A lognormal model for insurance claims data. *REVSTAT*, v.2, 2006.

4 Appendices

4.1 $p(\boldsymbol{\theta})$ distribution in the Gamma model with non-informative prior

Considering the derivatives of the likelihood function, we have:

$$\begin{aligned}
\frac{\partial \log L(\theta|y)}{\partial \alpha} =& n\log(\beta) - n\frac{\partial}{\partial \alpha}\log(\Gamma(\alpha)) + \sum_{i=1}^{n}\log y_i,\\
\frac{\partial^2 \log L(\theta|y)}{\partial \alpha^2} =& -n\frac{\partial^2}{\partial \alpha^2}\log\left(\Gamma(\alpha)\right),\\
\frac{\partial \log L(\theta|y)}{\partial \beta} =& \frac{n\alpha}{\beta} - \sum_{i=1}^{n} y_i,\\
\frac{\partial^2 \log L(\theta|y)}{\partial \beta^2} =& \frac{-n\alpha}{\beta^2},\\
\frac{\partial^2 \log L(\theta|y)}{\partial \alpha \partial \beta} =& \frac{n}{\beta}.
\end{aligned}$$

Arranging the results in the Fisher information matrix to obtain the distribution:

$$I(\theta) \propto \begin{bmatrix} \Psi'(\alpha) & \frac{-1}{\beta} \\ \frac{-1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix} \therefore p(\theta) \propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\beta},$$

4.2 $p(\boldsymbol{\theta})$ distribution in the Log-normal model with non-informative prior

Considering the derivatives of the likelihood function and the expected value of $E\left(\sum_{i=1}^{n}(\log(y_i)-\mu)^2\right)=n\sigma^2$ and $E\left(\sum_{i=1}^{n}\log(y_i)\right)=n\mu$:

$$\begin{aligned}
\frac{\partial \log L(\theta|y)}{\partial \mu} &= \frac{\sum_{i=1}^{n}(\log(y_i)-\mu)}{\sigma^2},\\
\frac{\partial^2 \log L(\theta|y)}{\partial \mu^2} &= \frac{-n}{\sigma^2},\\
\frac{\partial \log L(\theta|y)}{\partial \sigma} &= \frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(\log(y_i)-\mu)^2}{\sigma^3},\\
\frac{\partial^2 \log L(\theta|y)}{\partial \sigma^2} &= \frac{n}{\sigma^2} - \frac{3\sum_{i=1}^{n}(\log(y_i)-\mu)^2}{\sigma^4},\\
\frac{\partial^2 \log L(\theta|y)}{\partial \sigma \partial \mu} &= \frac{-2\sum_{i=1}^{n}(\log(y_i)-\mu)}{\sigma^3}.
\end{aligned}$$

Similarly,

$$I(\theta) \propto \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix} \therefore p(\theta) \propto \frac{n\sqrt{2}}{\sigma^2}.$$

4.3 Plots for Markov Chain Monte Carlo simulations

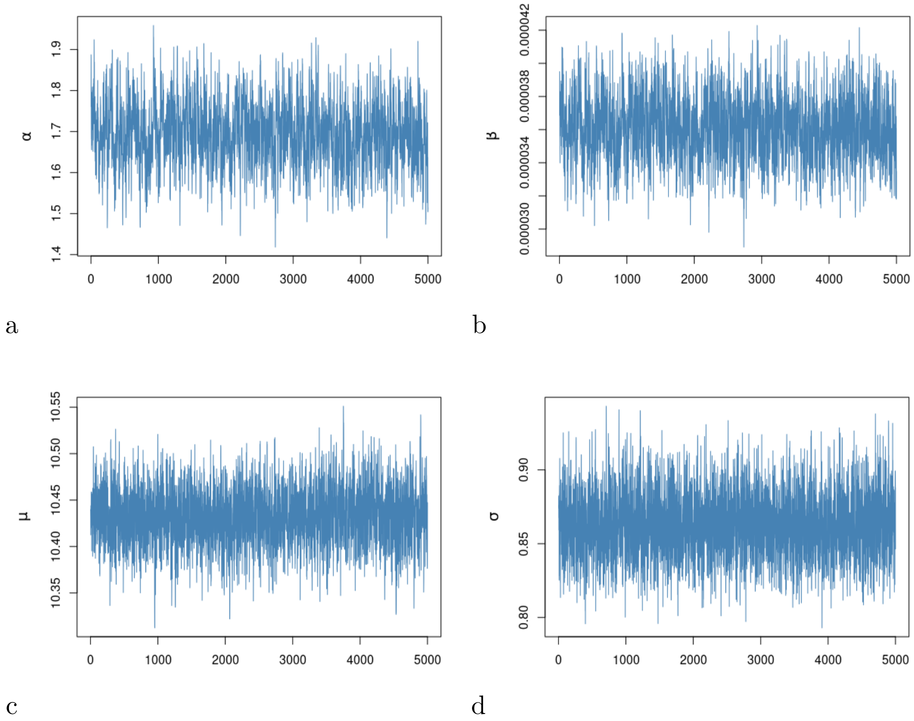

Figure 5: Trace of the chains for the parameters α (a), β (b), μ (c) and σ (d) of the models presented in Table 2.

Received in 10.03.2021.

Approved after revised in 05.01.2022.