REDECA
Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA

MODELAGEM DA OCORRÊNCIA DE SINISTROS DE VEÍCULOS PARA O ESTADO DE MINAS GERAIS VIA INFERÊNCIA BAYESIANA

MODELING THE OCCURRENCE OF VEHICLE CLAIMS FOR THE STATE OF MINAS GERAIS - BRAZIL VIA BAYESIAN INFERENCE

Luiz Otávio de Oliveira Pala [1]

Daiane Oliveira Gonçalves [2]

Bruna da Costa Silva [3]

RESUMO

Modelagens de ocorrência de sinistros são comumente realizadas com a regressão logística, mas podem falhar em situações de desbalanceamento dos dados. Como alternativa, modelos de regressão que lidem com excesso de zeros, como distribuições zero infladas ou zero ajustadas, podem ser utilizados. Objetiva-se modelar a ocorrência de sinistros do tipo colisão com perda total em veículos no estado de Minas Gerais no ano de 2019. Foi utilizada a regressão Binomial Zero Ajustada (ZABI) com inferência via abordagem Bayesiana, inserindo covariáveis relacionadas às características do segurado. Como resultado, verifica-se que os perfis de risco dos segurados caracterizados pelas variáveis sexo e idade se associaram significativamente ao sinistro. Sugerindo redução da probabilidade de ocorrência do sinistro em segurados do sexo feminino, quando comparado ao sexo masculino. Sob a abordagem Bayesiana, o modelo adotado permite a inserção do conhecimento prévio do atuário em relação ao evento analisado com uso de prioris informativas.

Palavras-chave: Inferência paramétrica; Níveis de risco; Perda total.

ABSTRACT

Claims occurrence modeling is commonly done using logistic regression, but this model can present low predictive capacity in situations of unbalanced data. Alternatively, regression models that deal with excess zeros, such as zero-inflated or zero-adjusted distributions, can be used. In this paper, we aim to model the occurrence of claims that resulted in total loss in vehicles in the state of Minas Gerais in 2019. We considered the Zero Adjusted Binomial (ZABI) regression with the Bayesian framework, inserting covariates related to the characteristics of the insured. We noted that risk profiles of policyholders characterized by the variables gender and age were significantly associated with the claim, suggesting a reduction on the probability of occurrence in female policyholders, when compared to male policyholders. Under the Bayesian approach, this model allows the insertion of the actuary's prior knowledge in relation to the analyzed event using informative priors.

Keywords: Parametric inference; Risk levels; Total loss.

1 Doutorando em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA). E-mail: luizotavio.oliveira@gmail.com https://orcid.org/0000-0002-9941-7951

2 Doutoranda em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA). E-mail: prof.daiane.oliveira@gmail.com https://orcid.org/0000-0002-0003-6965

3 Doutoranda em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA). E-mail: bruna.silva26@estudante.ufla.br https://orcid.org/0000-0002-8649-3859

DOI: 10.23925/2446-9513.2022v9id58647

1. Introdução

No Brasil, os seguros de automóveis são fornecidos pelas seguradoras, empresas autorizadas pela Superintendência de Seguros Privados (SUSEP), que recebem prêmios, assumem os riscos e garantem indenizações em caso de ocorrência de sinistros que estejam amparados pelas apólices de seguros (SUSEP, 2021).

O seguro de automóveis é uma das principais modalidades de seguro no país, sendo responsável pela arrecadação de 17,43 bilhões de reais em prêmios no primeiro semestre do ano de 2021, cujo valor foi 6,8% a mais do que o mesmo período em 2020. No entanto, ainda apenas 16% da frota de veículos no Brasil em 2019 tinham seguros (SUSEP, 2021).

Dentre os estados brasileiros, Minas Gerais ocupou, em 2007, o terceiro lugar do total de prêmios diretos no ramo de automóveis, em relação a todas as unidades federativas do país, representando 7,52%, e em 2021 esse percentual foi de 8,14, ocupando o segundo lugar (SUSEP, 2022a).

Ao analisar a série histórica da frota de veículos, entre os anos de 2007 e 2020, disponibilizada pelo IBGE (2021), o estado de Minas Gerais se manteve em segundo lugar ao contabilizar o número de veículos, perdendo somente para o estado de São Paulo. Em 2007 Minas Gerais tinha 5.271.000 veículos e passou a ter 12.053.218 em 2020, correspondendo a um aumento de 228,67% em 13 anos.

Com a competitividade no mercado, um dos pontos de grande importância dos produtos de seguro está relacionado com a etapa de precificação. Para que as seguradoras precifiquem os prêmios é necessário que elas avaliem os riscos que estão assumindo. Estas avaliações são realizadas por meio de uma série de perguntas que são feitas ao segurado, as quais podem impactar positivamente ou negativamente no valor do prêmio. Entre as perguntas realizadas tem-se, por exemplo, idade, tempo de habilitação, sexo, tipo de uso do veículo e localidade (SUSEP, 2022b).

Conforme Spedicato, Dutang e Petrini (2018), uma das formas usuais para realizar a precificação no ramo não vida, como em seguros de automóveis, é pela construção de modelos de regressão, como os modelos lineares generalizados, ou a partir de algoritmos de aprendizado de máquina. O que também pode permitir a classificação e clusterização de apólices conforme o perfil de risco do segurado a eventos como incêndio, roubo e colisão com dados parciais ou perda total.

A perda total de um veículo ocorre quando os custos de um mesmo sinistro ultrapassam o valor apurado a partir da aplicação de determinado percentual sobre o valor contratado (SUSEP, 2000). Sob o ponto de vista da Society of Actuaries (SOA, 2022), a perda total ocorre em situações de destruição total do bem ou quando este está tão danificado que não possa ser classificado como o objeto inicialmente segurado.

O evento colisão com perda total têm baixo percentual de ocorrência como evidenciado no estudo de Pala et al. (2020), sendo necessário estratégias, como o Randomly Over Sampling Examples (ROSE), para a modelagem da ocorrência do sinistro em virtude do desbalanceamento do conjunto de dados. Essa situação também foi apontada por Mota, Miquelluti e Ozaki (2020) ao construir modelos de predição de sinistros em atividades agrícolas.

Neste estudo objetiva-se modelar a ocorrência de sinistros do tipo colisão com perda total para o estado de Minas Gerais no ano de 2019. Para isso foi utilizada uma abordagem Bayesiana na regressão Binomial Zero Ajustada (ZABI), que permite modelar dados desbalanceados sem a necessidade de reamostrar o conjunto de dados, possibilitando a identificação do perfil de risco dos segurados por meio de algumas caracteríticas como idade e sexo.

2. Metodologia

Nas subseções 2.1 e 2.2 apresenta-se, respectivamente, a base de dados e a modelagem estatística utilizada.

2.1 Base de dados

Os dados utilizados por este trabalho foram disponibilizados pela Superintendência de Seguros Privados (SUSEP, 2022c), relativos a seguros de automóveis nos dois semestres de 2019, isto é, 2019B e 2020A. Foram selecionadas todas as sub-regiões do estado de Minas Gerais, sendo elas: 14 - Triângulo Mineiro, 15 - Sul de Minas, 16 - Região metropolitana de Belo Horizonte, Centro Oeste, Zona da Mata e Campos das Vertentes, 17 - Vale do Aço, Norte de Minas e Vale do Jequitinhonha, considerando segurados do sexo feminino e masculino, totalizando 487.720 observações.

A partir da seleção dos segurados, observou-se a ocorrência ou não de sinistros do tipo colisão com perda total, disponível na base de dados com o nome FREQ_SIN3, que mensura o número de sinistros para esta cobertura. Em seguida, uma variável binária associada com a ocorrência do sinistro foi criada, denominada Y, assumindo zero caso este sinistro não tenha ocorrido e um, caso contrário.

As idades dos segurados foram categorizadas em dois níveis, sendo divididas em indivíduos com até 35 anos de idade e indivíduos com idade acima de 35 anos, representada por uma variável binária, zero e um, respectivamente, sendo denominada como IDADE. Esta foi criada de modo a verificar os efeitos destas faixas etárias na ocorrência de colisões com perda total. Assim como para a variável explicativa IDADE, foi definida uma variável binária denominada SEXO, assumindo valor unitário caso o segurado seja do sexo feminino e zero para segurados do sexo masculino.

2.2 Definição do modelo

Para a modelagem estatística, foi abordado o modelo de regressão com resposta *Zero Altered Binomial*, $ZABI(n,\mu,\sigma)$, dada por:

$$p_Y(y \vee n,\mu,\sigma)=\sigma, \text{ se } y=0 \text{ e } p_Y(y \vee n,\mu,\sigma)=\frac{(1-\sigma)n!\,\mu^y(1-\mu)^{n-y}}{\left[1-(1-\mu)^n\right]y!(n-y)!}, \text{ se y} > 0,$$

para $0<\mu<1$ e $0<\sigma<1$, em que σ modela a probabilidade de $Y=0$, isto é, a não ocorrência do sinistro do tipo colisão com perda total. Maiores detalhes e propriedades sobre a distribuição $ZABI(n,\mu,\sigma)$ podem ser vistos em Rigby, Stasinopoulos, Heller e De Bastiani (2017). Foi considerado a seguinte estrutura de preditores com função de ligação logística, dada por:

$$\mu=\frac{1}{1+\exp(-\beta_0)} \quad \text{e} \quad \sigma=\frac{1}{1+\exp\left[-(S_o+S_1 SEXO+S_2 IDADE)\right]},$$

de modo que $\mu\in(0,1)$ e $\sigma\in(0,1)$. Para o conjunto de observações independentes de $Y=\{y_i, i=1,\ldots,n\}$, a função de verossimilhança do modelo, $L(\Theta \vee Y)$, é dada por

$L(\Theta \vee Y)=\prod_{i=1}^{n} p_Y(y_i \vee \cdot)$, sendo $\Theta=(\beta_0, S_o, S_1, S_2)^T$ o vetor de parâmetros desconhecidos do modelo. Além disso, foram consideradas prioris normais não informativas, ou seja, $p(\theta_i) \propto \exp\left\{\frac{(\theta_i-u_i)^2}{2\lambda_i^2}\right\}$, com hiperparâmetros definidos em $U=(0,0,0,0)^T$ e $\Lambda=(10,10,10,10)^T$. A posteriori, $\pi(\Theta \vee Y)$, foi obtida a partir do Teorema de *Bayes*, de modo que $\pi(\Theta \vee Y) \propto L(\Theta \vee Y)\, p(\Theta)$, em que $p(\Theta)$ representa a distribuição a priori conjunta. Devido a complexidade da posteriori resultante, o processo inferencial foi realizado na posteriori conjunta com o algoritmo *Metropolis Hastings*. A amostragem da posteriori foi planejada considerando 50.000 iterações, com período *burn-in* = 3.000 e *thin* = 5. E a análise de convergência foi efetuada com os diagnósticos de Geweke (1992) (G) e com o fator de dependência (FD) (RAFTERY; LEWIS, 1995), considerando convergentes os processos em que | G | < 1, 96 e o FD → 1. As análises e implementações foram realizadas no programa R (R CORE TEAM, 2021), com o auxílio dos pacotes *MHadaptive* de Chivers (2011) e *Coda* de Plummer et al. (2020).

3. Resultados e discussões

Nas subseções 3.1 e 3.2 são apresentados, respectivamente, uma análise descritiva dos dados e os resultados obtidos por meio da modelagem estatística.

3.1 Análise descritiva dos dados

O conjunto de dados abordado neste estudo se caracteriza de acordo com a Figura 1, em que 43% e 57% correspondem, respectivamente, ao percentual de segurados dos sexos feminino e masculino. Entre os segurados do sexo feminino, 25,3% tem menos que 35 anos e 74,7% tem mais de 35 anos. Já entre os segurados do sexo masculino, 23,6% tem menos que 35 anos e 76,4% tem mais que 35 anos.

DOI: 10.23925/2446-9513.2022v9id58647

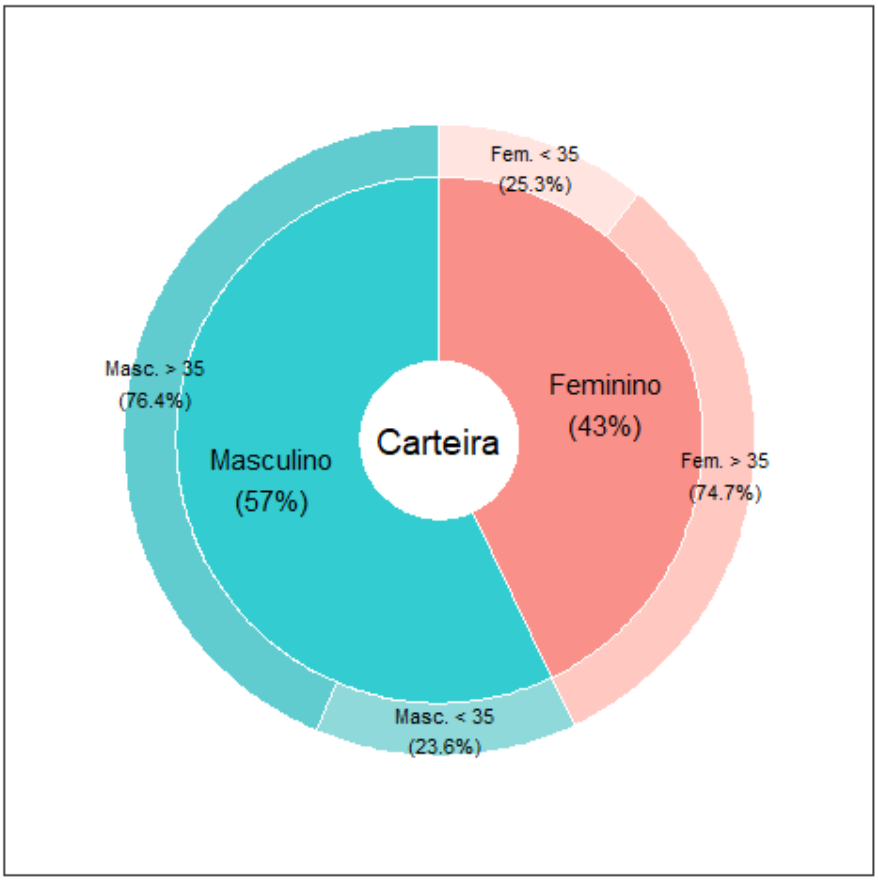


Figura 1: Composição da carteira de segurados no ano de 2019 relativos a apólices do estado de Minas Gerais, Brasil

No ano de 2019 foram reportados 9.425 sinistros do tipo colisão com perda total, representando um percentual de aproximadamente 1,932%, em relação ao total de apólices, de 487.720. Desta forma, espera-se que a cada 100 veículos segurados, no ano de 2019 em Minas Gerais, 1,932 (≈ 2) veículos colidem, resultando em perda total. Ao estratificar o número de sinistros em relação aos sexos dos segurados, ou seja, a razão entre o número de colisões pelo número de segurados, o percentual de sinistros com perda total foi de 2,09% e 1,72%, respectivamente, para os sexos masculino e feminino.

Diante deste contexto, é possível observar que o sexo do segurado é uma característica que pode influenciar no tipo de colisão, justificando a ideia de considerar tal efeito no processo de precificação de um seguro de automóvel, dado que o padrão de risco e exposição podem se diferenciar conforme as características do segurado.

Dentre as faixas etárias, observa-se que a carteira é majoritariamente composta por indivíduos com idade superior a 35 anos, o que pode ser visto na Tabela 1. Ao analisar o percentual de sinistros em relação ao grupo etário, cerca de 1,72% e 2,00% dos sinistros ocorreram, respectivamente, para os grupos com idade menor ou igual a 35 e maior que 35 anos. Ao estratificar a ocorrência de sinistros de acordo com os grupos etários e sexo, tem-se que 1,89% e 1,53% ocorreram, respectivamente, em segurados dos sexos masculino e feminino com idades menores ou iguais a 35 anos. Já para os segurados do sexo masculino e feminino com idades superiores a 35 anos, tem-se respectivamente 2,16% e 1,79% de ocorrência de sinistros.

Ao inserir a característica idade, além do sexo, é possível ter mais informações sobre a carteira de apólices, como por exemplo, a informação que os segurados com idades menores ou iguais a 35 anos têm menores taxas de ocorrência de sinistros em relação aos segurados com maiores de 35 anos. Portanto, a idade dos segurados também pode ser abordada no processo de precificação de um seguro de automóvel, com intuito de agregar informações adicionais sobre o segurado.

Sexo	Idade	Número de segurados	Número de colisões	Proporção de ocorrência (%)
Masculino	≤35	65.671	1.237	1,890
	>35	212.632	4.586	2,160
Feminino	≤35	53.050	808	1,530
	>35	156.367	2.794	1,790
Total	-	487.720	9.425	1,940

Tabela 1: Informações dos segurados por sexo e idade no ano de 2019 relativos a apólices do estado de Minas Gerais, Brasil

3.2 Análise estatística

Os resultados do modelo de regressão $ZABI(n,\mu,\sigma)$ estão apresentados na Tabela 2, com os resultados estimados das médias e desvios padrão dos parâmetros, juntamente com os intervalos de credibilidade e os testes de convergência. Note que os critérios FD e |G| sugerem convergência do processo para a distribuição estacionária e, desta forma, viabiliza a inferência estatística na distribuição a posteriori.

Verifica-se um efeito significativo das variáveis idades e sexo do segurado, ao nível de 95% de credibilidade, ao modelar a probabilidade de não ocorrência do sinistro, isto é, o parâmetro σ da distribuição $ZABI(n,\mu,\sigma)$. Em que os intervalos de credibilidade ao nível de 95% descartam o efeito de nulidade destes parâmetros.

Em relação ao sexo do segurado, os resultados indicam que o nível feminino da variável sexo contribui com o aumento da probabilidade de não ocorrência do sinistro. Para a variável idade, os resultados sugerem que segurados com idades menores ou iguais a 35 anos contribuem significativamente com o aumento da probabilidade de não ocorrência do evento

tipo colisão com perda total do veículo.

Parâmetro	Média	SD	HPD(95%)		FD	|G|
			LI	LS		
β_0	-0,096	5,016	-9,868	9,615	1,98	1,121
S_0	3,813	0,014	3,785	3,842	1,65	0,132
S_1	0,197	0,021	0,154	0,237	1,78	0,992
S_2	0,148	0,025	0,098	0,198	1,83	0,423

Tabela 2: Estimativas a posteriori dos parâmetros do modelo de regressão $ZABI(n,\mu,\sigma)$, intervalos de credibilidade e critérios para a análise de convergência das cadeias de Markov

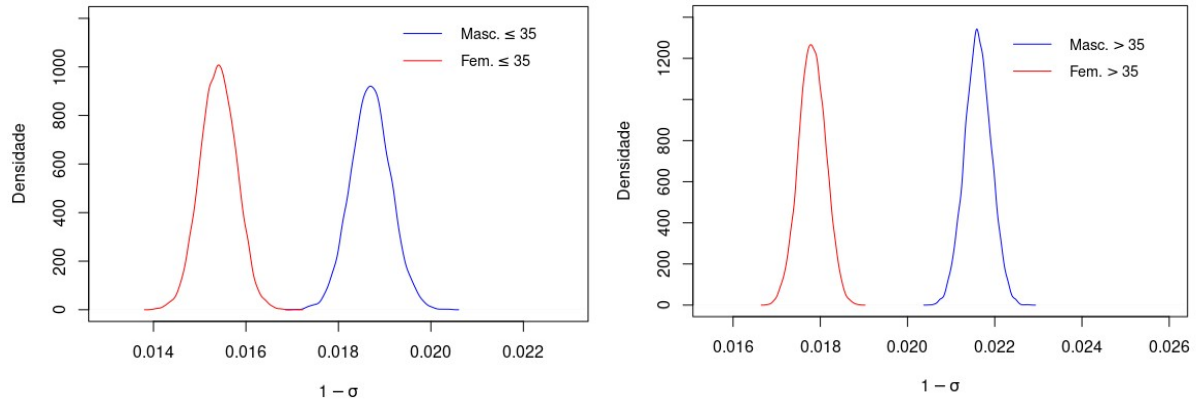
Avaliando o efeito marginal da probabilidade de ocorrência do evento, isto é, a probabilidade complementar de σ, foram obtidos os resultados apresentados nas Figuras 2(a) e 2(b). Em 2(a) tem-se a estimativa marginal de 1-σ para indivíduos com idade menor ou igual a 35 anos. Para o caso de segurados do sexo masculino neste grupo etário, a probabilidade de ocorrência estimada foi de 1,87%, variando entre 1,79% e 1,96% conforme o intervalo de credibilidade. Em termos práticos, para uma carteira com 10.000 apólices de segurados do sexo masculino, são esperados 187 sinistros do tipo colisão com perda total, podendo variar entre 179 e 196 sinistros.

Para segurados do sexo feminino com idade menor ou igual a 35 anos, apresentados na Figura 2(b), a probabilidade de ocorrência estimada foi de 1,54%, variando entre 1,47% e 1,62% conforme o intervalo de credibilidade. Para uma carteira hipotética com 10.000 apólices, espera-se 154 sinistros do tipo colisão com perda total, variando entre 147 e 162 sinistros, para o sexo feminino.

Note que ao nível de credibilidade de 95% não houve intersecção dos efeitos marginais de 1-σ entre segurados do sexo feminino e masculino com idade igual ou inferior a 35 anos, podendo ser observado na Figura 2(a). Ou seja, sugerindo um número menor de sinistros do tipo colisão com perda total no grupo feminino em relação ao grupo masculino dentro desta faixa etária.

Ao avaliar o efeito marginal de 1-σ, no grupo com idade superior a 35 anos, disposto na

Figura 2(b), tem-se para uma carteira com 10.000 apólices de segurados do sexo masculino são esperados 216 eventos, variando entre 210 e 222, ao ponto que no grupo feminino este valor é de 178 sinistros, variando entre 172 e 184, ao nível de 95% de credibilidade.



(a) (b)

Figura 2: Densidade marginal de 1-σ estimada para as faixas etárias e para os grupos masculino e feminino no ano 2019, em Minas Gerais

4. Conclusão

Durante a análise de ocorrência de sinistros e de outros fenômenos considerados raros, modelos de regressão logística são comumente utilizados, permitindo a classificação de clientes, segregação de risco e identificação de covariáveis associadas ao evento. No entanto, em situações de desbalanceamento, como o caso de colisões de veículos com perda total, situações de excesso de zeros podem resultar em problemas inferenciais.

De forma a abordar o problema de excesso de zeros, este estudo utilizou a regressão do tipo ZABI de modo a associar os perfis de risco dos segurados ao sinistro. Os perfis de risco dos segurados foram definidos pelas variáveis sexo e idade, que de acordo com os resultados obtidos se associam significativamente ao sinistro. Por exemplo, na carteira de segurados em estudo, identificou-se que pessoas com menos de 35 anos e do sexo feminino tendem a ter menores probabilidades de ocorrência do sinistro.

A análise pode ser ampliada para outras carteiras de seguros e regiões do país, bem como inseridas características individuais de cada apólice, de modo a classificar individualmente os clientes conforme a probabilidade de ocorrência do sinistro estimada pelo modelo, contribuindo com o gerenciamento do risco e em etapas de precificação.

Referências

CHIVERS, Corey. *MHadaptive: general markov chain monte carlo for Bayesian inference using adaptive Metropolis-Hastings sampling.* [S.l.], 2015. Disponível em: <https://cran.r-project.org/web/packages/MHadaptive/MHadaptive.pdf>. Data do acesso: 20 mai. 2022.

GEWEKE, *John F. Evaluating the accuracy of sampling based approaches to the calculation of posterior moments.* [S.l.], 1991. Disponível em: <https://ideas.repec.org/p/fip/fedmsr/148.html>. Data do acesso: 20 mai. 2022.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Frota de veículos*. [s.l.], 2021. Data do acesso: <https://cidades.ibge.gov.br/brasil/mg/pesquisa/22/28120?tipo=ranking&ano=2020&indicador=28120>. Data de acesso: 14 jun. 2022.

MOTA, Arthur Lula; MIQUELLUTI, Daniel Lima; OZAKI, Vitor Augusto. Predição de sinistros agrícolas: uma abordagem comparativa utilizando aprendizagem de máquina. *Economia Aplicada*, v. 24, n. 4, p. 533-554, 2020.

PALA, Luiz Otávio de Oliveira; CARVALHO, Marcela Marillac; GUIMARÃES, Paulo Henrique Sales; SÁFADI, Thelma. Vehicle claims in the south of Minas Gerais: an approach using classification models. *Semina: Exact and Technological Sciences*, v. 41, n. 1, p. 79-86, 2020.

PLUMMER Martyn; BEST Nicky; COWLES Kate; VINES Karen. *Coda: Output Analysis and Diagnostics for MCMC*. [s.l.], 2020. Disponível em: <https://cran.r-project.org/web/packages/coda/index.html>. Data do acesso: 01 jun. 2022.

R CORE TEAM. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. Data do acesso: 14 jun. 2022.

RAFTERY, Adrian E.; LEWIS, Steven M. The number of iterations, convergence diagnostics and generic metropolis algorithms. *Practical Markov Chain Monte Carlo*, v. 7, n. 98, p. 763–773, 1995. Data do acesso: 20 mai. 2022.

SPEDICATO, Giorgio Alfredo; DUTANG, Christophe; PETRINI, Leonardo. Machine learning methods to perform pricing optimization. A comparison with standard GLMs. *Variance*, v. 12, n. 1, p. 69-89, 2018.

SOA- SOCIETY OF ACTUARIES. *Actual Total Loss*. [s.l.], 2022. Disponível em: <https://actuarialtoolkit.soa.org/tool/glossary/actual-total-loss>. Data do acesso: 14 jun. 2022.

SUSEP - SUPERINTENDÊNCIA DE SEGUROS PRIVADOS. *Circular SUSEP No 145 de Novembro de 2.000. Dispõe sobre a estruturação mínima das Condições Contratuais e das Notas Técnicas Atuariais dos Contratos exclusivamente de Seguros de Automóvel [. . .].* Rio de Janeiro: SUSEP, 2000. Disponível em: <http://www2.susep.gov.br/bibliotecaweb/docOriginal.aspx?tipo=1&codigo=9058>. Data do acesso: 18 jan. 2022.

_____. *Susep simplifica seguro auto a partir de 1º de setembro*. [s.l.], 2021. Disponível em: <http://novosite.susep.gov.br/noticias/susep-simplifica-seguro-auto-a-partir-de-1o-de-setembro/> . Data do acesso: 14 jun. 2022.

_____. *Sistema de Estatísticas da SUSEP.* [s.l.], 2022a. Disponível em: <https://www2.susep.gov.br/menuestatistica/SES/premiosesinistros.aspx?id=54>. Data do acesso: 26 jun. 2022.

_____. *O que é questionário de avaliação do risco?.* [s.l.], 2022b. Disponível em: <http://www.susep.gov.br/setores-susep/cgpro/coseb/duvidas-dos-segurados-sobre-seguro-de-automoveis/o-que-e-questionario-de-avaliacao-do-risco>. Data do acesso: 26 jun. 2022.

_____. *AuToseg: sistema de estatísticas de automóveis da Susep.* [s.l.], 2022c. Disponível em: <http://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx>. Data do acesso: 10 jan. 2022.

Recebido: 07/07/2022
Aceito: 07/07/2022