MODELAGEM DA FREQUÊNCIA DE SINISTROS DE AUTOMÓVEIS NAS REGIÕES SUL E SUDESTE DO BRASIL: UM ESTUDO DE CASO UTILIZANDO A DISTRIBUIÇÃO BELL

RESUMO

Temos que uma prática bem comum no setor de seguros, especialmente o não-vida, para o cálculo do prêmio de risco, é se ajustar um Modelo Linear Generalizado (MLG) utilizando a distribuição de Poisson para a frequência de sinistros. Esse estudo buscou comparar duas distribuições aplicáveis a dados de contagem para modelar o número de sinistros, considerando alguns fatores de risco como idade e sexo dos motoristas, além do local da ocorrência, nesse caso, representada pelos estados das regiões sul e sudeste do Brasil.

Para tanto, foram utilizados dados recentes disponíveis na página oficial da Superintendência de Seguros Privados (SUSEP), referentes a 2019, sobre a ocorrência de sinistro, indenizações pagas e características dos condutores. Para a modelagem da frequência de sinistros, no primeiro modelo se utilizou a distribuição de Poisson e no segundo modelo, a distribuição Bell. Os modelos foram estimados com função de ligação logarítmica, o que possibilita a construção de modelos multiplicativos no cálculo dos prêmios puros. Assim, estimar a frequência de sinistros torna-se um passo fundamental na construção de tabelas tarifárias e buscar alternativas que sejam mais simples e tão boas quanto as existentes, faz com que o mercado segurador se aprimore cada vez mais.

Palavras-chave: Modelos Lineares Generalizados; Frequência de Sinistros; Tarifação de Seguros; Distribuição Bell.

ABSTRACT

A very common practice in the insurance industry, especially non-life insurance, to calculate of the risk premium, is to adjust a Generalized Linear Model (GLM) using the Poisson distribution for a frequency of claims. This study sought to compare two distributions applicable to count data modelling the number of claims, considering some risk factors such as age and gender of drivers, as well as the place of the occurrence, in this case, represented by the south and southeast regions of Brazil.

For this purpose, the most recent data available on the official website of the Superintendence of Private Insurance (SUSEP), for the year of 2019, to the occurrence of accidents, paid indemnities and characteristics of drivers were used. For modelling the frequency of claims, the first model uses the Poisson distribution and the second model, the Bell distribution. The models were estimated with a logarithmic link function, which enables the construction of multiplicative models to calculate pure premiums. Thus, estimating the frequency of claims becomes a fundamental step in the construction of tariff tables and seeking alternatives that are simpler and as good as existing ones, makes the insurance industry to improve itself even more.

Keywords: Generalized Linear Models; Frequency of Claims; Insurance Pricing; Bell Distribution.

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização e Motivação

Os Modelos Lineares Generalizados (MLG's) consistem em uma classe de modelos de regressão mais abrangente que o Modelo Linear Normal Clássico, que permitem estabelecer a relação entre variáveis e permitindo a análise de dados não normais (ou não gaussianos), podendo ser utilizada em diversas áreas de estudo, como no caso deste trabalho, para a ciência atuarial.

Com os MLG's, as alterações percebidas em uma variável podem ser explicadas pelas mudanças em uma ou mais variáveis que estejam relacionadas a ela. A variável que está sendo explicada é chamada de variável "resposta", enquanto que as variáveis explicando essas alterações são chamadas de variáveis "explicativas". No âmbito atuarial, as variáveis explicativas são conhecidas como fatores de risco.

O conceito de modelagem estatística pode ser entendido como a ciência de se projetar probabilidades, ajustar e interpretar um modelo que represente uma realidade de forma simplificada. Vale lembrar que para se ter um bom modelo, é necessário que os dados disponíveis também possuam informações suficientes. Para elucidar melhor a aplicação da modelagem, podemos relacionar o número de sinistros, como os acidentes de carro, com possíveis variáveis explicativas, tais como o sexo, a idade ou o tempo de carteira do motorista. Sabendo como tais fatores de risco se relacionam com a frequência de sinistros, é possível se ter uma noção do quantitativo de acidentes esperado em determinado grupo de pessoas de acordo com suas características. Esses modelos podem ser utilizados no cálculo de seguros de carro, por exemplo.

Assim, após uma análise descritiva dos dados, e de acordo com as variáveis que decidirmos que irão compor o modelo, utilizamos funções de distribuições já conhecidas e com parâmetros definidos, cujos comportamentos se assemelham às nossas variáveis. Sabemos, a título de exemplo, que o número de sinistros se trata de uma contagem, portanto, podemos utilizar distribuições como Poisson, Binomial Negativa, ou a própria distribuição Bell, alvo do presente trabalho. É importante lembrar que o mercado segurador tem evoluído bastante ao longo dos anos, portanto, quanto mais aprimorarmos as técnicas de estimação e modelos que melhor retratem os dados, maiores as chances de se manter competitivo.

Este trabalho tem como objetivo, então, analisar a utilização da distribuição Bell como alternativa à distribuição de Poisson para se ajustar a frequência de sinistros. A Bell é uma distribuição relativamente nova e pouco estudada, mas de fácil entendimento e que, com apenas um parâmetro envolvido, torna o ajuste simples de ser executado. Em contrapartida, a Poisson também possui apenas um parâmetro, é amplamente estudada e a alternativa mais frequente no mercado segurador, mas quando consideramos os casos de sobredispersão dos dados, que é o que geralmente ocorre com dados reais, esta distribuição já não é tão interessante. Para este comparativo, ambas as distribuições serão aplicadas a uma base de dados disponibilizada pela Superintendência de Seguros Privados (SUSEP). Os ajustes encontrados para frequência dos sinistros serão analisados a partir dos resíduos e, então, poderemos concluir se de fato essa nova distribuição se adequa melhor para a situação proposta.

O conteúdo deste trabalho está organizado em 4 seções. Após a contextualização e revisão da literatura, ambas apresentadas nesta Seção 1, é mostrada a metodologia na Seção 2, onde é feita a análise da base de dados, além das metodologias utilizadas para obter os resultados mostrados e explicados na Seção 3. Por fim, na Seção 4 constam as considerações finais e sugestões para trabalhos futuros.

1.2 Revisão da Literatura

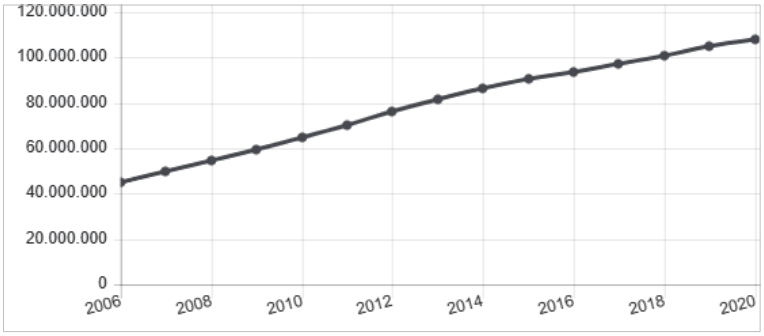
Segundo o relatório anual da Seguradora Líder, apenas durante o ano de 2020, foram pagas cerca de 310.710 indenizações com acidentes de trânsito, sendo 33.530 por morte, 210.042 por invalidez permanente e 67.138 com despesas médicas. Destas, 15,3% correspondem a veículos automotivos. Demais dados da Seguradora Líder apontam que, entre 2011 e 2020, mais de 4,7 milhões de pessoas foram indenizadas por morte, invalidez permanente e reembolso de despesas médicas. Números que contemplaram principalmente jovens na faixa dos 18 a 34 anos.

De acordo com análise de 2020 da FENACOR[1], a região Sudeste lidera o mercado segurador brasileiro, com 62% da receita total do país, vindo a região Sul em seguida, com 18%. Em relação ao mercado de seguros de automóveis[2], a região Sudeste lidera com 41%, enquanto a região sul possui uma participação de 19%.

1 A análise é feita com os dados estatísticos disponibilizados pela SUSEP
2 Não considera o Seguro DPVAT

Analisando o gráfico disponibilizado pelo Instituto Brasileiro de Geografia e Estatística - IBGE (2021), vemos que a frota de veículos automotivos no Brasil está em tendência de crescimento, portanto, podemos pensar que o mercado segurador ainda tem muito para se desenvolver.

Figura 1 - Quantitativo da frota de veículos automotivos no Brasil



Fonte: Gráfico elaborado pelo IBGE 2021, com dados disponibilizados pelo Ministério da Infraestrutura, Departamento Nacional de Trânsito - DENATRAN – 2020

Conforme exposto por David e Jemna (2015), em objetivo fundamental das seguradoras é calcular um preço de seguro adequado ou prêmio correspondente a um segurado, a fim de cobrir um determinado risco. Um método bem conhecido para calcular o prêmio é multiplicar a expectativa condicional da frequência dos sinistros pelo custo esperado dos sinistros. Assim, a modelagem da frequência de sinistros representa uma etapa essencial da precificação de seguros não vida. A análise de regressão de contagem permite a identificação dos fatores de risco e a previsão da frequência esperada de sinistros dadas as características de risco.

De acordo com Jong e Heller (2008), a modelagem linear generalizada é usada para avaliar e quantificar a relação entre uma variável de resposta e variáveis explicativas e foi proposta por Nelder e Wedderburn em 1972. Esta modelagem difere da modelagem de regressão linear comum em dois aspectos importantes:

(i) A distribuição da variável resposta é escolhida a partir de uma família exponencial. Assim, sua distribuição não precisa ser normal ou próxima da normal, e pode ser explicitamente não-normal.

(ii) Uma transformação da média da resposta está linearmente relacionada às variáveis explicativas. Uma consequência de se permitir que a resposta seja um membro da família exponencial é que a resposta pode ser, e geralmente é, heterocedástica. Assim, a variância variará com a média que pode, por sua vez, variar com as variáveis. Isso contrasta com a suposição de homocedasticidade da regressão normal.

Como exemplo de aplicação, os autores ajustam o número de filhos considerando a idade da mãe utilizando a distribuição de Poisson, alternando as funções de ligação entre logarítmica e identidade para os ajustes. Outro exemplo é relativo ao número de mortes devido ao diabetes em New South Wales, Austrália em 2002. Nesse último caso, o tamanho da população n entra no modelo como um termo de compensação, pois a exposição influencia diretamente no número de mortes. O ajuste é feito então considerando essa correção na variável.

Ainda na definição dos autores, os modelos lineares generalizados são muito importantes na análise de dados de seguros. Com os dados de seguros, as premissas do modelo normal frequentemente não são aplicáveis. Por exemplo, a severidade de sinistros, frequência de sinistros e a ocorrência de um sinistro em uma única apólice são resultados que não são normais. Além disso, a relação entre resultados e fatores de risco costuma ser multiplicativa, em vez de aditiva.

Para Ohlsson e Johansson (2010), o atuário utiliza dados históricos para encontrar um modelo que descreve como o custo do sinistro de uma apólice de seguro depende de uma série de variáveis explicativas. Na década de 1990, os atuários britânicos introduziram os MLG's como uma ferramenta para a análise de tarifas e isso se tornou a abordagem padrão em muitos países. Em seus estudos, os autores apresentam como exemplo de utilização dos MLG's para seguros de ciclomotores considerando como variáveis explicativas a classe, idade e local, considerando o número de sinistros por milha.

Ainda em relação a aplicação de MLG's no mercado segurador, Frees (2010) apresenta a aplicação do ajuste para acidentes de automóveis ocorridos na Califórnia durante o ano de 1963. As variáveis explicativas consideradas foram a densidade do tráfego, a idade do motorista, o número de condenações e número de envolvimento em outros acidentes.

Em outro exemplo, o autor cita o caso de seguros para negligências médicas, onde utiliza a distribuição de Poisson para ajustar as reinvindicações dos médicos que cometem erros e acabam sendo processados pelas partes prejudicadas por esses erros. No ajuste, ele

mostra que a área de atuação dos médicos, região, extensão da prática e características pessoais dos médicos (experiência e gênero) são determinantes importantes do número de processos por negligência médica. Como as seguradoras desejam precificar com precisão esse tipo de cobertura, tal ajuste torna-se fundamental.

2. METODOLOGIA

Para a utilização dos MLG's, a variável resposta deve estar distribuída a partir da família exponencial linear. Assim, dada uma variável resposta *y* qualquer, dizemos que ela pertence à família exponencial se a sua função densidade de probabilidade (f.d.p) possa ser escrita na forma abaixo:

$$f(y)=\exp\left\{\frac{y\theta-b(\theta)}{\varphi}+c(y,\varphi)\right\}(1)$$

$$g(\mu)=\eta=x'\beta(2)$$

A equação 1 para f(*y*) especifica que a distribuição da resposta está na família exponencial. A equação 2 especifica que uma transformação da média em uma função, g(μ), está linearmente relacionada às variáveis explicativas contidas no preditor linear η, ou seja, no vetor de valores de variáveis explicativas x. Já β é um vetor de parâmetros desconhecidos a serem estimados. O parâmetro φ é o parâmetro de dispersão (precisão, geralmente escolhemos distribuições que possuem esse parâmetro conhecido) da família exponencial.

A escolha de b(θ) determina qual a distribuição da variável resposta. Já a escolha de g(μ), chamada de função de ligação, determina como a média μ está relacionada às variáveis explicativas x. No modelo linear normal, a relação entre a média de y e as variáveis explicativas é dada simplesmente por μ = x'β. No caso dos MLG's, este conceito é estendido para g(μ) = x'β, onde g é uma função monotônica e diferenciável (como o caso das funções log ou raiz quadrada).

A configuração dada na equação 2 afirma que, dado x, μ é determinado por meio de g(μ). Dado μ, θ é determinado por b(θ) = μ. Finalmente, dado θ, y é determinado como um resultado da família exponencial especificada em b(θ).

Ainda no que diz respeito à família exponencial, é possível definir a média e a variância como sendo, respectivamente:

$$E(y)=b'(\theta)(3)$$

$$Var(y)=\varphi b''(\theta)(4)$$

Importante lembrar que as observações em y são consideradas independentes entre si.

Os MLG's têm se mostrado cada vez mais aplicáveis à ciência atuarial, pois a modelagem consegue abranger tanto dados de contagem quanto contínuos, permite a combinação de efeitos de forma consistente e precisa nas previsões, com a possibilidade avaliar a significância estatística da(s) variável(is) preditora(s).

Pela flexibilidade do modelo, percebemos que podem haver várias funções de ligação adequadas para uma distribuição específica. Para facilitar a escolha, segundo Frees (2010), um caso intuitivo ocorre quando o preditor linear é igual ao parâmetro de interesse ($\eta = \theta$). Isso porque $\eta = g(\mu)$ e $\mu = b(\theta)$. Então, é fácil ver que se $g^{-1} = b$, então $\eta = g(b(\theta)) = \theta$. A escolha de g(.) que seja a inversa de $b(\theta)$ é chamada de função de ligação canônica.

As funções de ligação mais usadas para $g(\mu)$ são fornecidas a seguir. Com exceção da função de ligação logit, as funções são da forma $g(\mu) = \mu p$, com o caso logarítmico sendo o limite de $(yp - 1) / p$ como $p \rightarrow 0$.

Figura 2 - Distribuições e respectivas Ligações Canônicas

Distribution	Mean function $b'(\theta)$	Canonical link $g(\mu)$
Normal	θ	μ
Bernoulli	$e^{\theta}/(1+e^{\theta})$	$\text{logit}(\mu)$
Poisson	e^{θ}	$\ln \mu$
Gamma	$-1/\theta$	$-1/\mu$
Inverse Gaussian	$(-2\theta)^{-1/2}$	$-1/(2\mu^2)$

Fonte: Regression Modelling with Actuarial and Financial Applications, Frees, 2010, p.388

A modelagem de contagens, como o número de sinistros ou mortes em um grupo de risco, exige que seja feita uma correção para o número n de expostos ao risco (Jong e Heller, 2008). Se μ é a média da contagem y, então a taxa de ocorrência μ / n de interesse ficará da forma:

$$g\left(\frac{\mu}{n}\right) = x'\beta \qquad (5)$$

Quando utilizamos a função logarítmica, temos:

$$\ln\left(\frac{\mu}{n}\right)=x'\beta=\ln n+x'\beta \qquad (6)$$

A variável n é chamada de exposição e ln n é chamada de "offset". Um offset é efetivamente outra variável x na regressão, com um coeficiente β igual a um. Com o deslocamento, y tem um valor esperado diretamente proporcional à exposição:

$$\mu=n e^{x'\beta} \qquad (7)$$

Os offsets são usados então para corrigir o tamanho do grupo ou diferentes períodos de observação. A função offset é muito útil, não somente na aplicação para o cálculo de prêmios para a área de seguros, mas em diversas outras áreas que se aplica, como no caso do número de pessoas vacinadas por estado. Sabemos que o número de doses aplicadas dependerá da população suscetível a este procedimento, então devemos olhar para a taxa de aplicação, e não somente o número em si.

A distribuição de Poisson, bem como sua média e variância são dadas por:

$$f(y)=\frac{e^{-\mu}\mu^{y}}{y!}, y=0,1,2\ldots n \qquad (8)$$

$$E(y)=Var(y)=\mu \qquad (9)$$

No modelo de regressão de Poisson, temos Poisson na forma da família exponencial:

$$f(y)=\exp\left(\left(\frac{ylog(\mu)-\mu}{1}\right)-\log(y!)\right) \quad (10)$$

Rearranjando os termos conforme apresentado na equação (1), temos então:

$\theta = \log(\mu)$; $b(\theta) = \mu = e^{\theta}$; $\varphi = 1$; $c(y, \varphi) = -\log(y!)$.

A equidispersão, uma das considerações da distribuição Poisson, nem sempre é observada em dados de contagem. Diante disso, muitos pesquisadores propuseram distribuições baseadas na de Poisson, dando origem a várias derivações dessa com sobredispersão (o valor da variância é maior do que o da média) ou subdispersão (o valor da variância é menor do que o da média) (PEREIRA, 2016).

Em geral, quando a variância é diferente da média, temos algumas alternativas, como a distribuição Binomial Negativa (BN) ou, a distribuição de Poisson Generalizada. Entretanto,

essas distribuições possuem um parâmetro adicional, tornando-as mais flexíveis, mas, ao mesmo tempo, mais susceptíveis a erros na estimação desses parâmetros. No caso da BN, o parâmetro introduz a sobredispersão e, no caso da Poisson Generalizada, o parâmetro apura tanto subdispersão quanto sobredispersão. Como alternativa à distribuição de Poisson e com apenas um parâmetro, utilizaremos a distribuição Bell.

Como apresentado por Castellares et al. (2018), a distribuição Bell uni-paramétrica pode ser aplicada a variáveis resposta que se tratam de contagem, sendo possível a aplicação de um modelo de regressão relacionado a um preditor linear por meio de uma função de ligação, na mesma configuração que um MLG. Este modelo não depende de funções complicadas e se apresenta como uma boa alternativa aos modelos que utilizam a distribuição de Poisson.

Uma variável aleatória discreta Y tem uma distribuição de Bell com parâmetro θ se sua função densidade de probabilidade for dada por:

$$f(y)=\frac{\theta^{y} e^{e^{\theta}+1}}{y!} B_{Y}, y=0,1,2 \ldots n, \theta>0 \qquad (11)$$

O termo B_y corresponde aos números de Bell, dados por:

$$B_n=\frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}, \quad \text{iniciando com } B_0 = B_1 = 1 \qquad (12)$$

A média e a variância de Y ~ Bell (θ) são dadas por:

$$E(y)=\theta e^{\theta} \qquad (13)$$

$$V(y)=\theta(1+\theta)e^{\theta} \qquad (14)$$

Em modelos de regressão, normalmente é mais útil modelar a média da variável de resposta. Assim, para obter uma estrutura de regressão para a média da distribuição de Bell, temos uma parametrização diferente da f.d.p de Bell apresentada em (11). Seja $\mu = \theta e^{\theta}$, temos $\theta = W_0(\mu)$, onde $W_0(\cdot)$ é a função de Lambert[3]. Segue-se então que $E(y) = \mu$ e $V(y) = \mu (1 + W_0(\mu))$.

3 Mais informações sobre a utilização da função de Lambert aplicada à distribuição Bell pode ser vista no artigo dos autores. Informações sobre o pacote LambertW está disponível em: https://cran.r-project.org/web/packages/LambertW/index.html

Assim, reescrevendo a distribuição Bell na forma da família exponencial para aplicação em MLG:

$$f(y)=\exp\left(1-e^{W_0(\mu)}\right)\frac{W_0(\mu)^y B_y}{y!}, y=0,1,2\ldots \text{ e } \mu>0 \qquad (15)$$

Temos que $W_0(\mu) > 0$ para $\mu > 0$ e, portanto, $V(y) > E(y)$. Isso implica que a distribuição de Bell pode ser adequada para modelar dados de contagem com sobredispersão, assim como a distribuição da BN de dois parâmetros. Uma vantagem da distribuição de Bell em relação à distribuição BN é que nenhum parâmetro adicional (dispersão) é necessário para acomodar a sobredispersão (CASTELLARES, 2018).

Determinadas as variáveis, as funções de distribuição e implementado o offset, o ajuste do modelo é determinado pelo vetor de parâmetros β, sendo o método de máxima verossimilhança (MVS) a base teórica para estimação desses parâmetros. Ocorre um processo iterativo na solução das equações de máxima verossimilhança, onde a variável dependente não é mais y, mas sim z, uma forma linearizada da função de ligação agora aplicada em y, e os pesos W são funções dos valores ajustados de μ. Para Paula (2016), o processo iterativo de Newton-Raphson[4] para a obtenção da estimativa de máxima verossimilhança de β é de tido expandindo uma função escore U em torno de um valor inicial $\beta_{(0)}$.

Nos modelos de regressão, os resíduos são definidos como a diferença entre os valores observados e os estimados. Aplicando-se ao contexto dos MLG's, os resíduos são utilizados para verificar a adequação do modelo ajustado em relação à escolha da função de variância, da função de ligação e dos termos do preditor linear. Além disso, os resíduos são também úteis para identificar a presença de pontos atípicos, que podem ser influentes ou não no modelo final (CORDEIRO e NETO, 2004).

Conforme exposto por Frees, 2010, em modelos de regressão com Poisson, antecipamos as variáveis dependentes heterocedásticas por causa da relação $Var(y_i)=\mu_i$

Esta característica significa que os resíduos comuns ($y_i - \mu_i$) são de menor utilidade, de modo que é mais comum examinar os resíduos de Pearson, definidos como:

$$r_i=\frac{y_i-\hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}} \qquad (16)$$

4 Mais informações sobre o processo iterativo para o processo de estimação por máxima verossimilhança podem ser consultadas no material do autor.

Assim, os resíduos de Pearson são aproximadamente homocedásticos, ou seja, possuem uma distribuição de frequência regular. Gráficos de resíduos de Pearson podem ser usados para identificar observações incomuns ou para detectar se variáveis adicionais de interesse podem ser usadas para melhorar a especificação do modelo. Neste trabalho, utilizaremos os resíduos de Pearson, que tem uma interpretação mais direta, para avaliar os valores ajustados em cada modelo.

A qualidade do ajuste de um MLG é comumente avaliada através da função desvio (*deviance*), que pode ser entendida como a distância entre o logaritmo da função de verossimilhança do modelo saturado (n=p parâmetros) e do modelo sob investigação (com p<n parâmetros) avaliado na estimativa de máxima verossimilhança β. Um valor pequeno para a função desvio indica que, para um número menor de parâmetros, obtemos um ajuste tão bom quanto o ajuste com o modelo saturado (PAULA, 2013). Para os modelos utilizados neste trabalho, a função da *deviance* é dada por:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n} \{ y_i \log\left(y_i / \hat{\mu}_i \right) - \left(y_i - \hat{\mu}_i \right) \} \qquad (17)$$

para a distribuição de Poisson e;

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n} \{ e^{W_0(\mu_i)} - e^{W_0(y_i)} + y_i \log\left(\frac{W_0(y_i)}{W_0(\mu_i)} \right) \}, \qquad y>0 \qquad (18)$$

para a distribuição da Bell.

Podemos usar a *deviance* para avaliar a qualidade do ajuste e como base de comparação de diferentes MLG's propostos. Quanto maior o valor da *deviance*, uma maior discrepância existe entre os ajustes dos modelos proposto e saturado, ou seja, maior evidência de falta de ajuste do modelo proposto.

Caso o modelo proposto seja adequado, o ajuste tem distribuição assintótica Qui-Quadrado χ^2 com (n – p) graus de liberdade. Nesse caso, pode-se testar a hipótese nula de que o modelo se ajusta aos dados comparando o valor de $D(y;\hat{\mu})$ com o quantil (1−α) da distribuição$\chi^2_{(n-p)}$. O modelo é rejeitado, ao nível de significância α, se $D(y;\hat{\mu}) > \chi^2_{(n-p-\alpha)}$.

2.1. Informações e tratamento da base de dados

Esta seção visa apresentar os métodos utilizados neste trabalho com o propósito de responder à questão proposta na introdução. Como ponto de partida, foi necessária a captação dos dados divulgados no sistema AUTOSEG, que é uma plataforma criada pela SUSEP para disponibilizar consultas on-line referentes a dados estatísticos brasileiros no que diz respeito a seguro de automóveis.

Para avaliar o impacto da utilização de duas distribuições de probabilidades distintas na modelagem da frequência de sinistros, foram considerados os veículos automotivos do modelo Uno 1.0[5]. Considera-se como sinistros os registros de roubo, incêndio, colisões e outras causas registrados pela SUSEP no período de janeiro a dezembro de 2019. Os dados foram selecionados seguindo os seguintes critérios (as variáveis explicativas):

- Estados da região sul e sudeste (ES, MG, PR, RJ, RS, SC e SP);
- Apenas registros com sexo do motorista informado (feminino ou masculino);
- Faixa etária informada a partir de 18 anos.

Além das categorias selecionadas acima, a base de dados possuía as seguintes informações: Importância Segurada Média (em R$), Quantidade de Expostos, Prêmio Médio (em R$), Frequência e Valor das Indenizações para os sinistros (sendo estas duas últimas as variáveis de interesse aquelas escolhidas quando se pretende modelar para o cálculo do prêmio puro).

No total, havia 70 observações com 50.281 ocorrências de sinistros, totalizando um valor de quase cento e um milhões de reais pagos em indenizações. Dessa forma, a base de dados se mostrou suficiente, com as informações necessárias para construir as tabelas tarifárias. Para uma melhor visualização dos dados, algumas das categorias foram agregadas, por exemplo, as ocorrências nos estados que estavam separados em regiões metropolitanas e interior, foram somadas.

Não havia valores zerados para nenhum dos dados obtidos, nem valores negativos, que poderiam prejudicar o ajuste. A tabela a seguir apresenta uma síntese das variáveis explicativas coletadas no sistema. Importante ressaltar que todas as variáveis escolhidas são categóricas (não numéricas).

5 A base de dados abrange todos os anos de fabricação deste modelo.

Tabela 1 - Variáveis explicativas consideradas na modelagem

Idade do Condutor
Entre 18 e 25 anos
Entre 26 e 35 anos
Entre 36 e 45 anos
Entre 46 e 55 anos
Maior que 55 anos
Sexo do Condutor
Feminino
Masculino
Região
ES - Espírito Santo
MG - Minas Gerais
PR – Paraná
RJ - Rio de Janeiro
RS - Rio Grande do Sul
SC - Santa Catarina
SP - São Paulo

Fonte: Elaborado pela autora

2.2. Análise Descritiva

Antes de prosseguir para os cálculos do trabalho, é fundamental fazer uma análise descritiva dos dados utilizados, para entender melhor o contexto deles e como as variáveis escolhidas se apresentavam no banco de dados. Verificamos que o banco de dados é composto por variáveis qualitativas, como a região em que se encontra o veículo, o sexo e a faixa etária do condutor, e também variáveis quantitativas, como a importância média segurada, prêmio médio, valor das indenizações (severidade) e número de sinistros, alvo da modelagem neste estudo.

Os dados foram ajustados e agregados em uma tabela formato csv para ser lida no software R, o sistema utilizado para realizar a modelagem. Utilizando o comando *summary*,

podemos obter uma análise rápida das variáveis, com sua classificação e principais estatísticas, como valores mínimos, máximos e média.

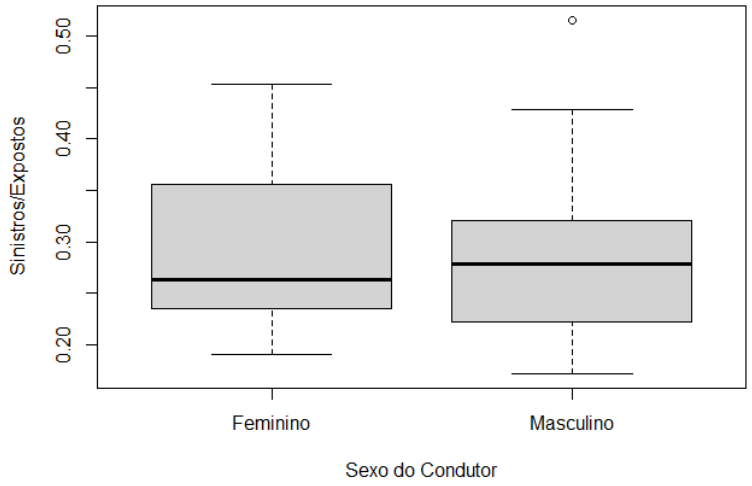
Figura 3 - Função *Summary* aplicada ao banco de dados

```
    Região              Sexo              Idade            Sinistros        Severidade         Expostos
 Length:70          Length:70          Length:70          Min.   :  13.0   Min.   :  29488   Min.   :   33.0
 Class :character   Class :character   Class :character   1st Qu.: 117.5   1st Qu.: 340353   1st Qu.:  453.2
 Mode  :character   Mode  :character   Mode  :character   Median : 411.0   Median : 864650   Median : 1415.0
                                                          Mean   : 718.3   Mean   :1442210   Mean   : 2368.0
                                                          3rd Qu.: 716.5   3rd Qu.:1702902   3rd Qu.: 2936.0
                                                          Max.   :4564.0   Max.   :7468357   Max.   :15047.0
```

Fonte: Elaborado pela autora

A seguir, temos os gráficos gerados também no software R, do comportamento de cada variável explicativa em relação à variável escolhida como resposta. No caso, como queremos modelar a taxa de sinistros, ou seja, a correção devido à exposição, os gráficos apresentados foram plotados considerando o número de sinistros dividido pelos expostos.

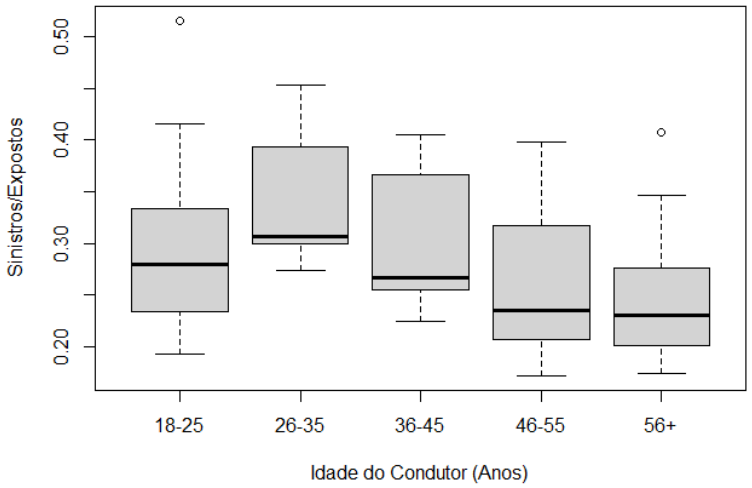
Gráfico 1 - Boxplot da variável Sexo do Condutor



Fonte: Elaborado pela autora

Vemos que a taxa de sinistros, apesar de apresentar uma mediana inferior para as mulheres, possui maior variabilidade para este grupo. Já para o grupo de homens, percebemos que há um *outlier* dentre os dados.

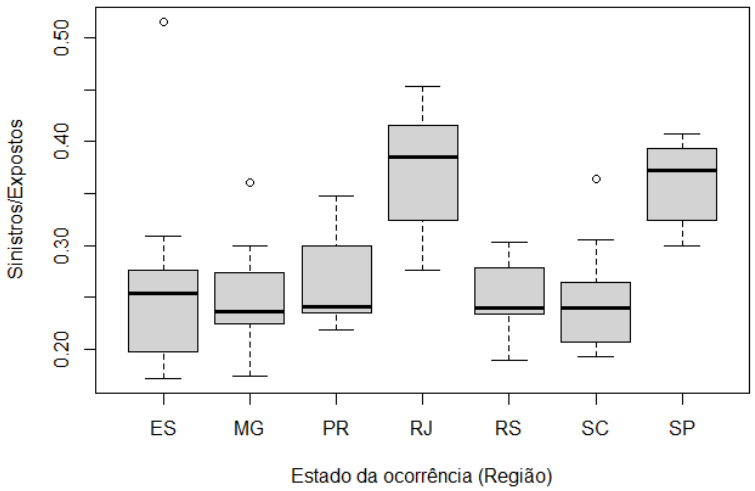
Gráfico 2 - Boxplot da variável Idade do Condutor



Fonte: Elaborado pela autora

Em relação à idade, vemos um comportamento esperado, onde as idades mais jovens registram as maiores taxas de sinistros. Com as idades mais avançadas, esperamos que essa taxa diminua.

Gráfico 3 - Boxplot da variável Região

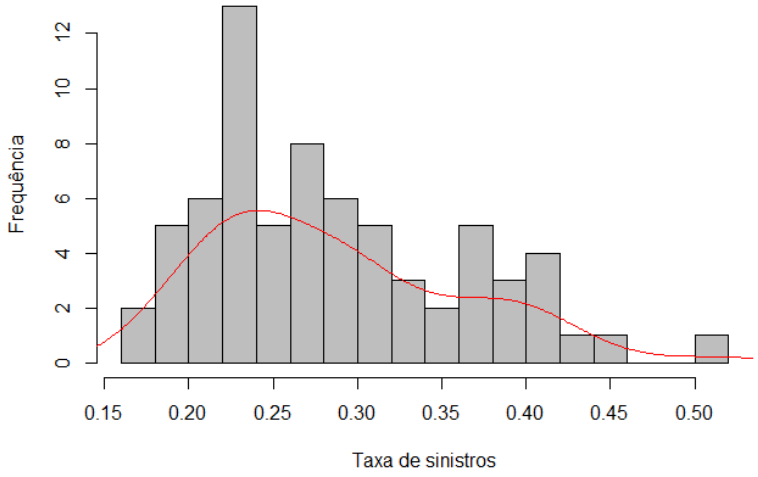


Fonte: Elaborado pela autora

Quando analisamos o boxplot para as regiões, vemos que muitas delas possuem medianas bem próximas. Os estados que se destacam com altas taxas de sinistros são Rio de Janeiro e São Paulo, o que também era esperado, uma vez que são estados que possuem uma grande frota de veículos e uma grande população. Alguns *outliers* são observados no Espírito Santo, Minas Gerais e Santa Catarina, o que pode interferir no ajuste.

Por último, fazemos o histograma da taxa de sinistros, para saber como nossos dados estão distribuídos.

Gráfico 4 - Histograma da taxa de sinistros



Fonte: Elaborado pela autora

2.3. Ajustes

No caso do ajuste dos sinistros por Poisson, foi utilizado o pacote glm, que já está estruturado com as principais distribuições e funções de ligação, além de fornecer os resultados, parâmetros estimados e resíduos de forma rápida e direta. Dessa forma, os dados são lidos aplicando-se a família exponencial Poisson com a função de ligação canônica, que é a própria função log (*default* no programa para esta família).

No caso do número de sinistros, a modelagem necessita ser corrigida para a exposição, conforme explicado anteriormente. Então, aplicaremos o número de expostos no preditor linear, conforme equação (5), para estimar os parâmetros para a Bell. O pacote glm possui a opção de adicionar offset com um comando, entretanto, o código disponibilizado para a regressão Bell[6] ainda é recente e não possuía essa implementação, que foi feita então no

6 O código pode ser acessado em: https://cran.r-project.org/web/packages/bellreg/bellreg.pdf

presente trabalho, também a partir da utilização da função de ligação log (equação 6). O cálculo de estimação por MVS utilizou as funções fornecidas no próprio código da distribuição Bell disponibilizado pelos autores. O código para o ajuste utilizando a distribuição Bell aplicada aos modelos lineares generalizados, implementado o offset, encontra-se no apêndice A deste trabalho e pode ser consultado para melhor entendimento.

Para analisar a significância dos modelos estimados com Poisson e com a Bell, testamos então se a *deviance* possuía um valor menor do que o valor tabelado para uma distribuição Qui-Quadrado com n-p graus de liberdade. Isso ocorre quando o modelo com p parâmetros (o modelo corrente) é considerado correto a um determinado nível de significância. O pacote glm já possui a função *deviance* para avaliarmos o modelo Poisson. Para o ajuste com a Bell, foi implementada a equação (18) no código.

Uma vez estimados os parâmetros β por cada um dos modelos considerados, foi essencial avaliar a qualidade dos ajustes e fazer uma comparação entre eles. Para tal, utilizou-se da análise de resíduos de Pearson. Novamente, os resíduos para a Poisson foram obtidos diretamente pelo pacote glm, enquanto para a Bell, foram calculados conforme código disponibilizado no Apêndice A.

3. RESULTADOS

3.1 O modelo Poisson

O ajuste com o modelo Poisson retornou os seguintes valores na saída do pacote glm:

```
Call:
glm(formula = Sinistros ~ Sexo + Idade + Região + offset(log(Expostos)),
    family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.7128  -1.4630  -0.0819   1.7973   9.5797

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.450373   0.042857 -33.842  < 2e-16 ***
SexoM           -0.116146   0.009040 -12.848  < 2e-16 ***
Idade26-35 anos  0.182213   0.027566   6.610 3.84e-11 ***
Idade36-45 anos  0.068291   0.027286   2.503   0.0123 *
Idade46-55 anos -0.007207   0.027538  -0.262   0.7935
Idade56 anos +  -0.053064   0.026846  -1.977   0.0481 *
RegiãoMG         0.001528   0.036026   0.042   0.9662
RegiãoPR         0.116010   0.038048   3.049   0.0023 **
RegiãoRJ         0.449531   0.036657  12.263  < 2e-16 ***
RegiãoRS         0.059808   0.037742   1.585   0.1130
RegiãoSC         0.097394   0.039433   2.470   0.0135 *
RegiãoSP         0.459954   0.034745  13.238  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3129.87  on 69  degrees of freedom
Residual deviance:  443.59  on 58  degrees of freedom
AIC: 1006.1

Number of Fisher Scoring iterations: 4
```

Se considerarmos um nível de confiança de 5%, as variáveis de Idade de 46 a 55 anos de idade e as regiões de Minas Gerais e Santa Catarina não se mostram significativas quando comparadas ao *baseline*[7].

Como vemos na saída, a *deviance* residual é de 443,59 para esse modelo, com 58 graus de liberdade (n - p, que corresponde n=70 e p=12).

3.2 O modelo Bell

O ajuste com a distribuição Bell nos retorna a seguinte saída:

7 O baseline deste modelo é considerado o grupo de 18 a 25 anos de idade, da região do Espírito Santo e do sexo feminino

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.539771	0.092198	-16.700762	0.000000
SexoM	-0.099894	0.022692	-4.402159	0.000011
Idade26-35 anos	0.244352	0.062100	3.934831	0.000083
Idade36-45 anos	0.136039	0.061399	2.215651	0.026715
Idade46-55 anos	0.036622	0.062091	0.589808	0.555320
Idade56 anos +	-0.006689	0.060342	-0.110855	0.911731
RegiãoMG	0.049356	0.077639	0.635709	0.524966
RegiãoPR	0.489047	0.079230	6.172501	0.000000
RegiãoRJ	0.492929	0.074292	6.634993	0.000000
RegiãoRS	0.128689	0.085186	1.510676	0.130871
RegiãoSC	0.095064	0.081619	1.164719	0.244133
RegiãoSP	0.171820	0.082086	2.093181	0.036333

Novamente, se considerarmos um nível de confiança de 5%, agora temos que as variáveis de Idade de 46 a 55 anos, 56 anos e mais e as regiões de Minas Gerais, Rio Grande do Sul e Santa Catarina não se mostram significativas quando comparadas ao *baseline*, que neste caso, também possui as mesmas variáveis que o modelo Poisson.

Quando calculamos a *deviance* residual para a Bell, encontramos o valor de 78.22.

3.3 Comparação dos modelos

Quando comparamos as duas saídas dos modelos, vemos que os dois ajustes se assemelham em relação aos valores dos parâmetros estimados, entretanto, vemos que algumas variáveis são significativas no modelo Poisson, enquanto que no modelo Bell elas são rejeitadas. Se considerarmos que o modelo Poisson não está apto a captar a sobredispersão, este modelo acaba subestimando o erro padrão, ocasionando numa área maior para a estatística de teste. Com isso, pode ser que o p-valor calculado para os parâmetros se torne menor, sendo plausível aceitar uma variável quando na verdade ela deveria ser rejeitada.

Se calcularmos o valor crítico com 95% de confiança e 58 graus de liberdade, para testar a aceitação do modelo a partir da *deviance*, temos um valor de 76.78. Aumentando o nível de confiança para 99%, o valor crítico chega a 85.95. Isso significa que podemos considerar o modelo Bell adequado, uma vez que sua *deviance* é de 78.22, enquanto o modelo Poisson não se mostra adequado, com uma *deviance* de 443.59.

Para avaliar se existe a sobredispersão no ajuste da Poisson, podemos utilizar a função *dispersiontest*, que está no pacote AER. Sabemos que o modelo Poisson restringe a variância dos dados a ser igual à média, condicional às variáveis explicativas. A falha desta restrição tem consequências semelhantes às da heterocedasticidade no modelo de regressão linear: desde que a função de regressão seja especificada corretamente, as estimativas dos parâmetros

dos mínimos quadrados ordinários são consistentes, mas as variâncias para as estimativas dos parâmetros são estimadas de forma inconsistente e os testes de hipótese serão inválidos (CAMERON e TRIVEDI,1990)[8]. Assim, ao aplicar esse teste, temos a seguinte saída:

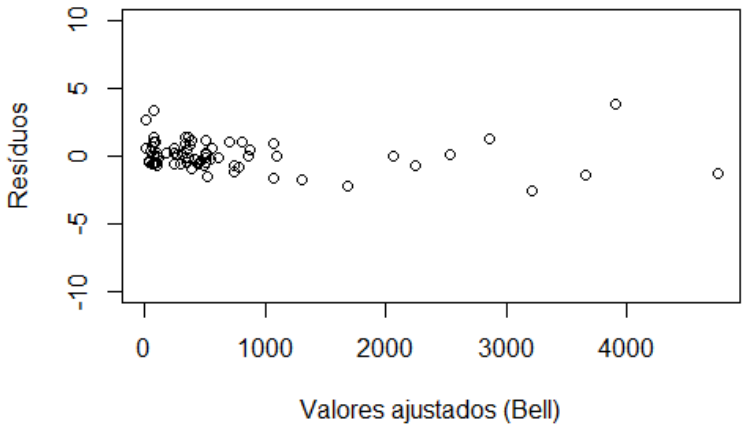
```
        Overdispersion test

data:  teste
z = 3.4019, p-value = 0.0003346
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  6.541495
```

Pelo resultado, temos que 6.541495 > 1, portanto, há fortes indícios de sobredispersão nos dados.

Assim, quando fazemos a análise dos modelos considerando os resíduos de Pearson versus valores ajustados nos gráficos 5 e 6, apresentados a seguir, podemos ver claramente que o ajuste para a Bell se mostrou melhor, os resíduos estão distribuídos de forma mais regular e com menor distância em torno de 0. Se observarmos, no modelo Poisson alguns resíduos atingem o valor de 10. Atribuímos esse melhor ajuste do modelo Bell justamente pela sobredispersão apurada no teste anterior (*overdispersion test*), que dificulta encontrar bons resultados na utilização da Poisson.

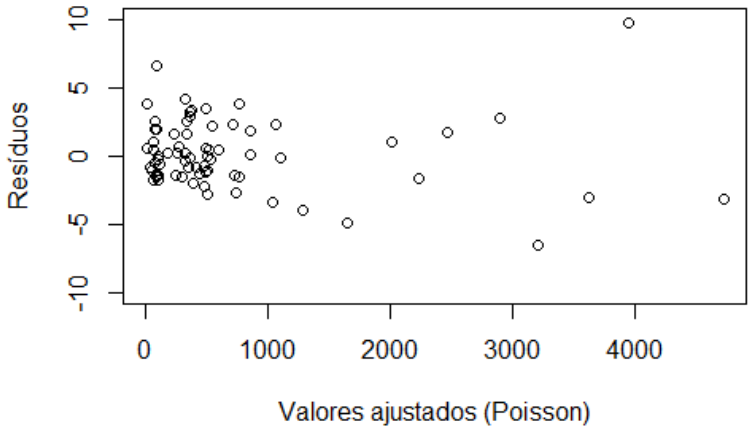
Gráfico 5 – Resíduos de Pearson x Valores Ajustados para o modelo Bell



Fonte: Elaborado pela autora

8 O artigo com mais detalhes sobre o teste de dispersão pode ser acessado em: https://doi.org/10.1016/0304-4076(90)90014-K

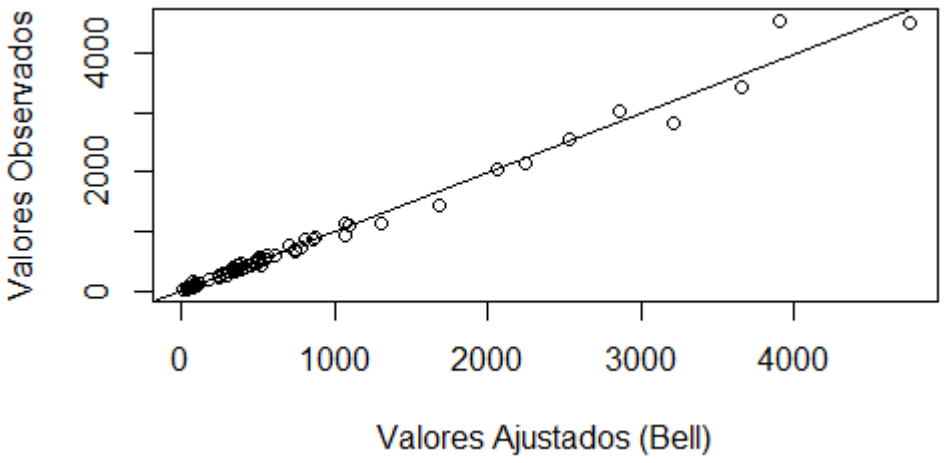
Gráfico 6 – Resíduos de Pearson x Valores Ajustados para o modelo Poisson



Fonte: Elaborado pela autora

Por fim, após verificar que o modelo está adequado na comparação da *deviance* e que os resíduos se apresentam de forma satisfatória no modelo Bell, analisamos agora o gráfico 7 considerando os valores observados com os valores ajustados. Temos então que o modelo se ajusta bem aos dados utilizados.

Gráfico 7 - Valores Observados x Valores Ajustados para o modelo Bell



Fonte: Elaborado pela autora

4. CONSIDERAÇÕES FINAIS

Nesse trabalho, procurou-se propor a utilização de uma nova distribuição de probabilidades para o ajuste de frequência de sinistros de automóveis, considerando idade e sexo do condutor e as regiões sul e sudeste do Brasil. De acordo com os resultados obtidos nos ajustes e análise de resíduos, observamos que a distribuição Bell se adequa muito bem aos dados e é tão simples de se estimar quanto a um modelo que utiliza Poisson, possuindo também apenas um parâmetro e ainda capaz de lidar com a sobredispersão.

Assim, as estimativas mais precisas obtidas com o ajuste Bell podem impactar diretamente nos preços para apólices, uma vez que o mercado segurador está se tornando cada vez mais competitivo, as empresas precisam continuamente buscar melhores resultados. Este trabalho mostra-se então uma importante referência para a área seguradora nacional, contribuindo com um novo modelo para ajuste dos sinistros, sendo também disponibilizado o código utilizando a distribuição Bell considerando a opção de offset, que ainda não havia sido implementado no pacote existente bellreg.

Conclui-se que com os dados retirados da plataforma AUTOSEG da SUSEP foi possível chegar a um ajuste significativo, tanto na adequação do modelo quanto na análise de resíduos, com uma boa capacidade de predição dos valores. Assim, este ajuste representa um primeiro passo para a precificação de seguros no mercado segurador utilizando uma nova distribuição para contagens.

Como sugestão de trabalhos futuros, propõe-se a construção de tabelas tarifárias fazendo a análise da severidade em conjunto com o ajuste do número de sinistros utilizando a distribuição Bell. Considerando a tendência de crescimento na frota de veículos no Brasil de forma geral, entendemos que a construção de tabelas tarifárias mais precisas venha a ser útil no futuro, estimulando estudos mais aprofundados que busquem a aplicação de novas distribuições e aprimoramento dos métodos de modelagem existentes.

5. BIBLIOGRAFIA

CAMERON AC, TRIVEDI PK. **Regression-based Tests for Overdispersion in the Poisson Model**. Journal of Econometrics, 46, 347–364, 1990.

CASTELLARES, F. et al. **Applied Mathematical Modelling 56**, p. 172–185, Elsevier Inc., 2018.

CORDEIRO, G. M.; NETO, E. A. L. **Modelos Paramétricos.** Anais do XVI Simpósio Nacional de Probabilidade e Estatística (SINAPE), Caxambu, MG, 2004.

DAVID, Mihaela and JEMNA, Dănuţ-Vasile. **Modeling the frequency of auto insurance claims by means of poisson and negative binomial models.** Annals of the Alexandru Ioan Cuza University-Economics, v. 62, n. 2, p. 151-168, 2015.

Federação Nacional dos Corretores de Seguros Privados e de Resseguros, de Capitalização, de Previdência Privada, das Empresas Corretoras de Seguros e de Resseguros (FENACOR) – **Análise Estatística Fenacor** – Rio de Janeiro, 2021.

FREES, E. W. **Regression Modeling with Actuarial and Financial Applications**. 2 ed. United States, Cambridge University Press, 2010.

JONG, P.; HELLER, G. Z. **Generalized Linear Models for Insurance Data.** 1 ed. Cambridge University Press, 2008.

Instituto Brasileiro de Geografia e Estatística (IBGE). Consulta de dados disponível em: https://cidades.ibge.gov.br/brasil/pesquisa/22/28120; Acesso em 11/08/2021.

Instituto de Pesquisa Aplicada (IPEA). **Custos dos acidentes de trânsito no Brasil: Estimativa simplificada com base na atualização das pesquisas do IPEA sobre custos de acidentes nos aglomerados urbanos e rodovias** – Brasília, 2020.

OHLSSON, E.; JOHANSSON, B. **Non-Life Insurance Pricing with Generalized Linear Models**. Springer-Verlag Berlin Heidelberg, 2010.

PAULA, Gilberto A. **Modelos de regressão com apoio computacional**. Instituto de Matemática e Estatística - Universidade de São Paulo, 2013.

PEREIRA, Hemilhana Tolentina. **Estudo da distribuição de Poisson generalizada. 2016. x, 86 f., il. Dissertação (Mestrado em Estatística)** — Universidade de Brasília, Brasília, 2016.

SEGURADORA LÍDER. **Relatório Anual 2020** – Disponível em: https://www.seguradoralider.com.br/Centro-de-Dados-e-Estatisticas/Relatorio-Anual; Acesso em 08/08/2021.

Superintendência de Seguros Privados (SUSEP). Base de dados disponível em: http://www2.susep.gov.br/menuestatistica/Autoseg/menu1.aspx ; Acesso em 23/06/2021.

APÊNDICE A

```
# Para utilização da distribuição Bell estruturada com a função W de Lambert
require(LambertW)

#Variável resposta que corresponde ao número de sinistros no banco de dados
y <- Sinistros

# Consideração para a exposição ao risco, vetor com a quantidade de expostos
E <- Expostos

# Função de log-verossimilhança
loglik <- function(beta){
  mu <- E*exp(X%*%beta)
  sum(y*log(W(mu)) - exp(W(mu)))
}

# Função Score
fscoreBell <- function(beta){
    mu <- E*exp(X%*%beta)
    vt <- 1/(1 + W(mu))
    mT <- diag(as.vector(vt))
  score <- t(X)%*%mT%*%(y - mu)
  score
```

```
}

# Função de Fisher para a Bell
fFisherBell <- function(beta){
  mu <- E*exp(X%*%beta)
  vw <- mu/(1 + W(mu))
  mW <- diag(as.vector(vw))
  mF <- t(X)%*%mW%*%X
  mF
}

# Processo de otimização
chute <- glm(Sinistros~Sexo+Idade+Região+offset(log(Expostos)), family = poisson(link =
"log"))$coef
est.betas <- optim(chute, loglik, gr = fscoreBell, control=list(fnscale=-1), hessian = TRUE)
est.betas$par

# Ajuste com a distribuição Bell
se <- sqrt(diag(solve(fFisherBell(est.betas$par))))
z.value <- est.betas$par/se
p.value <- 2*(1 - pnorm(abs(z.value)))
rval <- cbind( round(est.betas$par, 6), round(se, 6), round(z.value, 6), round(p.value, 6) )
colnames(rval) <- c("Estimate", "Std. Error", "z value", "Pr(>|z|)")
rval
```

```
# Visualização do ajuste Bell

MuHat <- E*exp(X%*%est.betas$par)

plot(MuHat, Sinistros, ylab = "Valores Observados", xlab = "Valores Ajustados (Bell)")

abline(0,1)


# Cálculo dos Resíduos de Pearson

vw <- MuHat*(1 + W(MuHat))

rd1 <- (y-MuHat)/sqrt(vw)

plot(MuHat, rd1, ylab = "Resíduos", xlab = "Valores ajustados (Bell)", ylim=c(-10,10))


# Deviance para Bell

d1 <- 2*sum((y*log(W(y)/W(MuHat))+exp(W(MuHat))-exp(W(y))))

rdd1 <- sign(y-MuHat)*sqrt(2*(y*log(W(y)/W(MuHat))+exp(W(MuHat))-exp(W(y))))

plot(MuHat, rdd1)
```