UNIVERSIDADE FEDERAL DO CEARÁ

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE – FEAAC

DEPARTAMENTO DE ADMINISTRAÇÃO

CURSO DE CIÊNCIAS ATUARIAIS

ANTÔNIO FELIPE SILVÉRIO DA ROCHA

MODELAGEM GLM APLICADA À ATUÁRIA: UMA UTILIZAÇÃO DOS MODELOS LINEARES GENERALIZADOS NA PRECIFICAÇÃO DE SEGUROS

FORTALEZA

2015

ANTÔNIO FELIPE SILVÉRIO DA ROCHA

MODELAGEM GLM APLICADA À ATUÁRIA: UMA UTILIZAÇÃO DOS MODELOS LINEARES GENERALIZADOS NA PRECIFICAÇÃO DE SEGUROS

Monografia submetida à avaliação da banca examinadora e apresentada ao Curso de Ciências Atuariais da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Ciências Atuariais.

Orientador: Prof. Msc. Luís Gustavo Bastos Pinho.

FORTALEZA

2015

FICHA CATALOGRÁFICA

Dados Internacionais de Catalogação na Publicação

Universidade Federal do Ceará

Biblioteca da Faculdade de Economia, Administração, Atuária e Contabilidade

---

R571m Rocha, Antônio Felipe Silvério da

Modelagem GLM aplicada à atuária: uma utilização dos modelos lineares generalizados na precificação de seguros / Antônio Felipe Silvério da Rocha - 2015**.**

64 f.: il.

Monografia (graduação) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Ciências Atuariais, Fortaleza, 2015.

Orientação: Prof. Me. Luís Gustavo Bastos Pinho.

1. Modelos lineares (Estatística) 2.Ciência atuarial 3.Seguros I. Título

CDD 368.01

ANTÔNIO FELIPE SILVÉRIO DA ROCHA

MODELAGEM GLM APLICADA À ATUÁRIA: UMA UTILIZAÇÃO DOS MODELOS LINEARES GENERALIZADOS NA PRECIFICAÇÃO DE SEGUROS

Esta monografia foi submetida à avaliação por parte da banca examinadora pertencente à Universidade Federal do Ceará – UFC. O presente trabalho representa parte dos requisitos necessários à obtenção do título de Bacharel em Ciências Atuariais outorgado pela referida universidade supracitada, e encontra-se à disposição para consulta pública na Biblioteca da Faculdade de Economia, Administração, Atuária e Contabilidade – BFEAAC – por parte de qualquer pessoa que venha a se sentir interessada ou curiosa pelo tema em estudo. Ressalta-se ainda que a citação de qualquer trecho ou fragmento contido nesta monografia é livremente permitida desde que sejam observados rigorosamente os padrões éticos e as normas acadêmicas inerentes à elaboração de trabalhos científicos.

Aprovada em: ____/____/_____

BANCA EXAMINADORA

_____

Prof. Msc. Luís Gustavo Bastos Pinho: Orientador

Universidade Federal do Ceará: UFC

_____

Prof. Dr. Paulo Rogério Faustino Matos

Universidade Federal do Ceará: UFC

_____

Prof.ª Dr.ª Alane Siqueira Rocha

Universidade Federal do Ceará: UFC

DEDICATÓRIA

Aos meus queridos e amados pais, Neli e Geovanda, em retribuição ao esforço e sacrifício realizado no intuito de fazer de mim a pessoa que sou.

"A ideia revolucionária que define a fronteira entre os tempos modernos e o passado é o domínio de gestão do risco". (Peter Bernstein)

AGRADECIMENTOS

Primeiramente a Deus pelo dom da vida, e a Jesus Cristo pelos seus ensinamentos.

Aos meus amados pais, Neli e Geovanda, a minha irmã, Neliane, a todos os meus familiares, e a todos os meus amigos pela compreensão que tiveram comigo ao longo dessa jornada, tendo em consideração os inúmeros momentos nos quais precisei declinar do direito de usufruir das suas companhias a fim de concluir esta pesquisa.

Aos queridos docentes do curso de Ciências Atuariais da Universidade Federal do Ceará que tão bem souberam desempenhar seus papéis enquanto mestres no decorrer desta caminhada acadêmica, mostrando-se sempre detentores de um notório e reconhecido saber, e transmitindo com grande maestria e destreza o conhecimento de forma a avivar o interesse, e fazendo despertar, recorrentemente, a curiosidade dos alunos pela interessante e empolgante ciência atuarial. De forma direta, expresso honrosa e respeitosamente minha enorme gratidão aos professores Sérgio César de Paula Cardoso, Paulo Rogério Faustino Matos, Ana Cristina Pordeus Ramos, Alana Katielli Azevedo de Macedo, Alane Siqueira Rocha, e Iana Bezerra Jucá.

Ao nobre amigo professor Luís Gustavo Bastos Pinho por ter aceitado o pedido de orientação na realização dessa monografia, pelos oportunos e relevantes comentários, e por ter compartilhado comigo alguns importantes *insights* que vieram a contribuir significativamente ao aprimoramento e concepção desta pesquisa, principalmente no que tange às peculiaridades e especificidades relativas à análise e modelagem de dados via MLG.

Por fim, agradeço a todos que de alguma forma contribuíram para que eu chegasse até aqui.

RESUMO

Os Modelos Lineares Generalizados – MLGs – introduzidos inicialmente pelos atuários britânicos Nelder e Wedderburn (1972) correspondem a uma síntese de um numeroso conjunto de modelos de regressão linear existentes. Essa classe de modelos representa uma extensão ao modelo linear de regressão clássico tendo em vista o fato dos MLGs não restringirem a escolha da distribuição de probabilidade da variável resposta, que no caso do modelo linear clássico deve ser necessariamente a distribuição normal. Dessa forma, objetivando desenvolver um modelo de precificação para um seguro não-vida, decidiu-se aqui recorrer ao uso da modelagem GLM pelo fato das distribuições empíricas dos dados analisados em seguros se distanciarem substancialmente das hipóteses adotadas pelo modelo linear clássico, como a normalidade das observações e dos resíduos, homocedasticidade, entre outras. Nesse contexto da precificação de seguros, os MLGs foram aplicados a dados relativos à frequência e à severidade de sinistros presentes no *dataset moped insurance* obtido em Ohlsson e Johansson (2010). Por meio dos MLGs foi possível ajustar dois modelos, um para o número médio e outro para o valor médio dos sinistros, e assim estimar, a partir destes, os coeficientes e as relatividades componentes do modelo de cálculo para o prêmio de risco. Além disso, observou-se que o melhor ajuste obtido para a frequência dos sinistros foi o modelo Log-Poisson, enquanto para a severidade foi o modelo Log-Gaussiano Inverso. Finalmente foi analisado em que direção, e em que intensidade, as variáveis tarifárias afetavam o cálculo do prêmio do seguro, permitindo com isso gerar, de maneira objetiva, estimativas a *priori* acerca do prêmio a ser calculado para cada apólice com base no perfil de risco individual dos segurados.

Palavras-chave: Modelos Lineares Generalizados. MLGs. Precificação. Modelagem Atuarial. Atuária. Risco. Seguros. Frequência. Severidade. Prêmio. Sinistros. Célula tarifária. Fatores de risco. Razão chave.

ABSTRACT

The Generalized Linear Models – GLMs – introduced initially by the British actuaries Nelder and Wedderburn (1972) corresponds to a synthesis of a large set of existing linear regression models. This class of models is an extension to classical linear regression model taking into account the fact that GLMs not restrict the choice of the probability distribution of the response variable, in which case the classical linear model must necessarily be a normal distribution. Thus, in order to develop a pricing model for a non-life insurance, it was decided here resort to using the GLM modeling because of the empirical distribution of the data analyzed in safe distance themselves substantially from the assumptions adopted classical linear model, as the normality of the observations and the residuals, homocedasticity, among others. In this context the pricing of insurance, GLMs were applied to data on the frequency and severity of claims in the present dataset moped insurance Ohlsson and Johansson (2010). Through GLMs was possible to fit two models, one for the average number and other for the average value of claims, estimating from these the coefficients and the relativities components of the calculation model for the risk premium. Also, the best fit obtained for the frequency of the claims was the Log-Poisson model, while for the severity was the Log-Gaussian Inverse model. Finally it was analyzed in that direction, and in that intensity, the tariff variables affecting the calculation of the insurance premium, thereby allowing generate, objectively, estimates a priori about the premium to be calculed for each policy based on profile individual risk of the insured.

Keywords: Generalized Linear Models. GLMs. Pricing. Actuarial modeling. Actuary. Risk. Insurance. Frequency. Severity. Premium. Claims. Tariff cell. Risk factors. Key ratios.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado segurador brasileiro tem se mostrado, ao longo dos últimos anos, um setor econômico altamente competitivo e detentor de um elevadíssimo nível concorrencial entre seus *player*s. Por essa razão, cada diferença mercadológica observada entre um mesmo produto de seguro comercializado por diferentes companhias pode vir a exercer reflexos extremamente significativos em termos de competitividade e inteligência de negócio. Nesse sentido, o cálculo do preço do seguro merece total atenção quando da sua concepção, pois à medida que o processo de tarifação dos produtos incorpora um maior rigor técnico e de precisão ao cálculo, as companhias seguradoras passam a ser cada vez mais capazes de cobrar um valor justo e que represente de maneira mais fidedigna o risco médio associado ao perfil individual de cada segurado.

Várias são as técnicas estatísticas empregadas pelos atuários nos estudos onde o principal objetivo consiste em realizar a predição e a inferência sobre o comportamento de uma determinada variável aleatória de interesse. Podem ser citados como exemplos de métodos estatísticos empregáveis para tal finalidade os modelos de regressão, sobrevivência, credibilidade, multivariados, séries temporais, entre outros. A escolha da técnica a ser utilizada em cada situação dependerá sempre do objetivo que o atuário pretende alcançar, pois para cada fenômeno estatístico em específico podem coexistir técnicas menos ou mais apropriadas, ou ainda técnicas que sequer podem ser aplicadas ao fenômeno em análise.

Enveredando por essa perspectiva da modelagem estatística e lançando um olhar específico sobre a atividade de precificação dos seguros não-vida, ou seja, aqueles seguros de ramos elementares que tratam da cobertura de riscos causados a bens e ao patrimônio de particulares, é razoável supor que a tarefa de determinar uma metodologia para o cálculo do preço a ser cobrado dos segurados não seja algo trivial para as companhias seguradoras.

> O objeto da Teoria do Risco reside em estabelecer um modelo de tarifação eficiente, capaz de garantir equilíbrio em face das variações aleatórias do risco segurado e dar solvabilidade ao segurador no longo prazo. A Teoria do Risco pode ser compreendida como um sinônimo para a Matemática de Seguros Não-Vida, na qual se busca uma modelagem científica que faça frente aos sinistros que chegam ao segurador, ajustando o quanto de segurança se deve aplicar ao cálculo dos prêmios, de maneira a que o processo de ruína não ocorra (RODRIGUES, 2008).

Ohlsson e Johansson (2010) afirmam que uma apólice de seguro não-vida é um acordo entre a seguradora e o segurado, onde a seguradora indeniza o segurado pelas perdas observadas num certo período de tempo, geralmente um ano, mediante o pagamento da tarifa.

Uma apólice de seguro não-vida pode cobrir danos sofridos por um carro, casa ou outra propriedade, ou perdas decorrentes de lesões corporais com o segurado ou outra pessoa (seguro de responsabilidade civil). Para uma companhia, o seguro pode cobrir danos à propriedade, o custo pela interrupção de negócios ou de problemas de saúde com os empregados, e muito mais. Com efeito, qualquer seguro que não seja seguro de vida é classificado como seguro não-vida, também chamado de seguro geral ou, em inglês, *casualty insurance*, e em alemão, *Schadenversicherung*.

O presente trabalho busca desenvolver uma metodologia de análise e tarifação de um produto de seguro operado por uma companhia seguradora que seja capaz de proporcionar objetivamente uma diferenciação no preço de seu produto em relação ao preço dos produtos comercializados por companhias concorrentes, viabilizando assim uma redução no valor do prêmio pago pelos segurados com perfil individual de baixo risco, e uma majoração do valor do prêmio pago por parte dos segurados com perfil individual de alto risco.

Ohlsson e Johansson (2010) afirmam que atualmente o mercado tem sido desregulado em muitos países onde a legislação vem sendo modificada para permitir a livre concorrência em detrimento de um modelo de precificação uniforme. A ideia é que se uma companhia de seguros cobrar um alto prêmio para algumas apólices, estas serão perdidas para um concorrente que possui um prêmio mais justo. Supondo que uma companhia seguradora cobre um prêmio muito baixo para motoristas mais jovens e um prêmio muito alto para motoristas mais velhos, então ela tenderá a perder motoristas mais velhos para os concorrentes enquanto atrairá mais jovens motoristas. Essa *seleção adversa* resultará em perda econômica em ambas as situações: por perder rentabilidade e ganhar apólices subprecificadas. A conclusão é que, num mercado de seguros competitivo e dinâmico como o brasileiro, tornar-se-á vantajoso cobrar um prêmio justo sempre que esse prêmio corresponda, em média, às perdas esperadas transferidas pelos segurados às companhias seguradoras.

Conforme Ohlsson e Johansson (2010), a precificação de seguros não-vida é a arte de fixar o preço de uma apólice de seguro tendo em consideração várias propriedades do objeto segurado e do titular da apólice. A principal fonte na qual se baseia a decisão das companhias seguradoras está no seu próprio histórico de dados referente às apólices e às reclamações de sinistros, sendo ainda algumas vezes complementadas por dados oriundos de fontes externas. Em uma análise tarifária, o atuário utiliza esses dados com o objetivo de encontrar um modelo que descreva como o custo dos sinistros de uma apólice de seguro depende de uma série de variáveis explanatórias. Em 1990, atuários britânicos introduziram e

sugeriram os Modelos Lineares Generalizados: MLGs: como uma ferramenta de análise tarifária, tendo essa se tornado atualmente uma abordagem padrão em muitos países.

Tendo em vista a ampla utilização dos MLGs pelos atuários para a precificação dos contratos de seguro ao redor do mundo no ambiente interno das companhias seguradoras, o presente trabalho objetiva, de maneira geral, analisar e compreender a relação existente entre as variáveis tarifárias componentes do modelo de precificação a ser desenvolvido.

Em suma, essa pesquisa pretende desenvolver um modelo estatístico para o cálculo do prêmio de um seguro por meio dos Modelos Lineares Generalizados, de sorte que o prêmio calculado seja capaz de cobrir, em média, as despesas esperadas com indenizações futuras de sinistros a serem pagas aos segurados pela seguradora, considerando ainda uma margem de segurança para a contingência de possíveis flutuações aleatórias adversas do risco. Nesse sentido, conforme o exposto até aqui, pretende-se aqui investigar a seguinte questão: em que medida os Modelos Lineares Generalizados podem ser considerados uma técnica estatística eficaz para a precificação de contratos de seguro?

Adicionalmente, também figuram como objetivos específicos desse trabalho:

i. *Introduzir na modelagem estatística as variáveis caracterizadoras do perfil de risco individual de cada segurado;*

ii. *Definir as variáveis tarifárias do modelo com suas respectivas classes de risco;*

iii. *Analisar a relação causal e o poder preditivo existente entre as variáveis independentes e a variável resposta do modelo tarifário;*

iv. *Estimar e ajustar os coeficientes para os níveis de risco das variáveis preditoras com a obtenção das suas respectivas relatividades tarifárias;*

v. *Demonstrar a utilização prática do MLG ajustado no cálculo do prêmio de um segurado hipotético com base no seu perfil de risco; e*

vi. *Analisar a qualidade e a adequação do ajuste estatístico, diagnosticando e validando o modelo estimado.*

O presente trabalho é composto por essa introdução e mais cinco capítulos. No segundo capítulo são tecidas algumas considerações preliminares acerca dos aspectos básicos da tarifação dos seguros privados não-vida, além de uma breve revisão bibliográfica sobre a utilização dos MLGs no âmbito atuarial. Já o terceiro capítulo versa sobre os Modelos Lineares Generalizados propriamente ditos, onde se realiza uma abordagem teórica mais específica e detalhada sobre essa classe de métodos estatísticos, de modo a retratar e discutir

as suas propriedades matemáticas mais relevantes. O quarto capítulo descreve o procedimento metodológico empregado, além dos aspectos referentes à modalidade de seguro estudada e às características do conjunto de dados estudado. O quinto capítulo trata do exercício empírico de desenvolvimento do modelo de precificação propriamente dito por meio da utilização dos MLGs onde são realizadas análises e considerações acerca dos resultados obtidos no intuito de avaliar o nível de adequação e qualidade do ajuste dos modelos. Por fim, no sexto e último capítulo são tecidas as considerações finais, além de serem lançadas as possíveis perspectivas para pesquisas e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Esse capítulo pretende introduzir alguns conceitos básicos relativos à tarifação dos seguros privados, lançando um olhar específico sobre o ramo dos seguros não-vida, além de realizar uma breve explanação teórica a respeito das variáveis tarifárias envolvidas no processo de precificação dos mesmos. Adicionalmente, pretende-se ainda realizar uma breve revisão bibliográfica da teoria atuarial inerente à precificação dos seguros privados por meio da utilização dos Modelos Lineares Generalizados.

2.1 Revisão Bibliográfica: Modelos Lineares Generalizados em Atuária

Não são recentes as pesquisas e os estudos acadêmicos que buscam tratar de aplicações de técnicas estatísticas para o desenvolvimento de modelos de precificação dos seguros privados. A literatura atuarial vem evoluindo e desenvolvendo constantemente novas propostas metodológicas de precificação, o que tem contribuído para tornar cada vez mais eficazes os processos estatísticos de análise de risco executados pelos atuários no âmbito interno das companhias seguradoras.

De acordo com Turkman e Silva (2000) os Modelos Lineares Generalizados introduzidos inicialmente pelos atuários britânicos Nelder e Wedderburn (1972) correspondem a uma síntese de uma numerosa gama de modelos de regressão linear que contém em comum uma característica peculiar e extremamente relevante, qual seja, o fato da variável resposta seguir uma função de distribuição de probabilidade pertencente a uma família de distribuições com propriedades muito específicas: a família exponencial.

Dobson (2001) afirma que os Modelos Lineares Generalizados podem ser empregados e aplicados em fenômenos estatísticos onde se pretenda avaliar e quantificar a relação entre uma determinada variável resposta de interesse Y, e um vetor de variáveis explicativas $X^T = [x_1, x_2, \dots, x_n]$. Os MLGs diferem dos modelos lineares clássicos de regressão em função de dois aspectos:

i. A distribuição de probabilidade da variável resposta é escolhida da família exponencial (Poisson, Binomial, Gamma, Normal, Binomial Negativa e Gaussiana Inversa);

ii. Uma transformação do valor esperado da variável resposta é linearmente relacionada com as variáveis explicativas;

As funções de probabilidade da família exponencial podem ser escritas na forma:

$$f_{Y_i}(y_i;\theta_i;\phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i;\phi;w_i)\right] \qquad (1)$$

Onde θ_i e ϕ, representam, respectivamente, os parâmetros de locação e escala da variável resposta Y. Tem-se que $E(Y_i) = \mu_i = b'(\theta_i)$, e que $Var(Y_i) = \phi^{-1}V(\mu_i)$, sendo $V_i = V(\mu_i) = d\mu_i/d\theta_i$ a função de variância, e $\phi^{-1} > 0$ o parâmetro de dispersão do modelo. Um MLG é definido matematicamente por dois componentes, um estocástico expresso em 2.1, e um sistemático expresso em 2.2, qual seja:

$$g(\mu_i) = \eta_i \qquad (2)$$

Morgado (2004) afirma que com o desenvolvimento crescente de estudos focados na precificação de seguros, a técnica dos Modelos Lineares Generalizados tornou-se uma das mais conhecidas e aplicadas pelo mercado segurador.

Ainda segundo o autor, vários são os trabalhos presentes na literatura atuarial que comparam a eficiência de estimação dos Modelos Lineares Generalizados aos algoritmos de Redes Neurais e às Árvores de Decisão. Exemplificando, o mesmo cita trabalhos como o de Francis (2001), Chapados *et alli* (2001), Dugas *et alli* (2003) e Hadidi (2003).

Alguns trabalhos realizados anteriormente já testaram de maneira empírica a eficiência e robustez dos Modelos Lineares Generalizados para a modelagem de dados de sinistros e o desenvolvimento de arcabouços matemáticos em atuária, como Santos (2008), Sousa (2010), Souza e Leão (2012) e Bandeira (2013).

Santos (2008) desenvolve um processo de modelagem atuarial para o cálculo da tarifa de um seguro de responsabilidade civil automóvel partindo da análise da experiência passada dos sinistros da carteira em estudo e da sua estrutura tarifária vigente até chegar à determinação do prêmio com base nos fatores de risco dos perfis individuais dos segurados. Nesse trabalho a autora recorre à aplicação dos Modelos Lineares Generalizados, estimando a frequência média, através de um modelo Poisson, a severidade média, por meio de um modelo Gamma, e posteriormente os conjuga num modelo para o cálculo do prêmio. A mesma justifica a escolha dos MLGs pautando-se, tanto na flexibilidade existente para a escolha da variável resposta do modelo, como no crescente número de aplicações observadas para essa classe de modelos estatísticos em pesquisas envolvendo a precificação de seguros em atuária.

Já Sousa (2010) recorre à utilização dos Modelos Lineares Generalizados e dos Modelos de Dispersão no contexto de seguro agrícola. Nessa pesquisa a autora desenvolve um processo de estimação paramétrica baseado no método de máxima verossimilhança, e, em

função da pequena quantidade de dados amostrais disponíveis para análise, a mesma recorre a utilização de um método de reamostragem *Bootstrap* Não-Paramétrico. A análise estatística realizada sobre dois conjuntos de dados de sinistros ocorridos em 15 municípios do estado do Rio Grande do Sul leva a autora à conclusão de que a variável precipitação acumulada possui influência estatística sobre a variável resposta ocorrência ou não de sinistros, muito embora não consiga obter conclusões suficientes acerca de qualquer variável que exerça potencial influência estatística sobre a variável resposta montante de sinistros observados. Entretanto, recorrendo ao uso do método *Bootstrap*, a autora observa evidências de influência estatística da variável precipitação acumulada, e da variável temperatura média, sobre a frequência observada de sinistros.

Souza e Leão (2012) aplicam os Modelos Lineares Generalizados no processo de tarifação de um plano de saúde autogestão. No trabalho, os autores investigam incialmente a hipótese dos custos médicos seguirem uma distribuição de probabilidade pertencente à família exponencial, hipótese essa confirmada e validada através de testes estatísticos de aderência a respeito do ajuste para a variável resposta do modelo. Além disso os autores recorrem ao uso dos MLGs, testando variáveis de natureza qualitativa e quantitativa, no intuito de capturar os fatores de risco inerentes ao perfil de cada participante em exposição durante o período de análise. Os resultados obtidos evidenciam que a variável renda possui uma relação inversa às despesas médicas observadas, onde indivíduos com maiores níveis de renda tendem a gerar um volume de sinistros inferior à média dos indivíduos da carteira. Os autores observam ainda que beneficiários do sexo masculino incorrem em despesas médicas médias superiores às despesas médias observadas por beneficiários do sexo feminino, fato esse que imprime reflexos significativos no cálculo atuarial do prêmio a ser pago ao plano de saúde.

Nesse sentido, por meio da aplicação dos MLGs foi possível gerar tabelas de preço a serem cobrados dos beneficiários do plano de saúde, de forma a se cobrir o risco médio, ou o prêmio de risco médio esperado, decorrente da utilização de cada indivíduo exposto, tendo em consideração o seu perfil de risco individual. Esse trabalho mostrou-se um avanço para o mercado de planos de saúde na modalidade autogestão, sobretudo sob o ponto de vista atuarial, haja vista que as práticas comuns utilizadas no mercado são, por exemplo, a aplicação de um percentual fixo de contribuição aplicado sobre o salário dos participantes, a utilização de fatores moderadores do risco de utilização, como a simples utilização de tabelas de coparticipação, entre outros.

Outros importantes e abrangentes estudos focados na modelagem GLM podem ser encontrados em Cordeiro (1986), McCullagh e Nelder (1989), Dobson (2001), Paula (2004), Jong e Heller (2008), e Ohlsson e Johansson (2010).

2.2 Modelagem Atuarial e Precificação de Seguros Não-Vida

Como o próprio nome já sugere, a atividade de modelagem atuarial consiste no desenvolvimento por parte do atuário de um modelo estatístico que seja capaz de estimar de forma preditiva o comportamento de uma determinada variável aleatória de interesse. Nesse contexto, um modelo atuarial nada mais é do que um arcabouço formado por um ou mais modelos matemáticos, cuja finalidade é estimar ou prever o comportamento esperado de algum sistema real de uma forma mais simplificada através do estudo e da análise da relação existente entre variáveis, sem, no entanto, perder o seu caráter representativo da realidade. Um sistema real pode ser entendido, por exemplo, como o comportamento da variável aleatória valor dos sinistros agregados individuais observados por um segurado em um determinado período de tempo.

O desenvolvimento dos modelos de precificação atuariais decorre, basicamente, da necessidade observada pelo atuário em conhecer e inferir, *a priori*, e com o maior nível de assertividade possível, sobre o comportamento do valor esperado e da variância associada às perdas médias observadas pelos sinistros. A partir do momento em que o atuário é capaz de calcular o valor médio esperado das indenizações individuais por segurado, o mesmo também se trona capaz de determinar a tarifa de comercialização do seguro.

Corroborando com esse pensamento, Ohlsson e Johansson (2010) afirmam que através de um contrato de seguro, o risco econômico é transferido do segurado para a seguradora. Devido à lei dos grandes números, a perda de uma companhia seguradora, sendo representada pela soma de um grande número de pequenas perdas comparativamente independentes, torna-se muito mais previsível que uma perda observada de forma individual. Em termos relativos, a perda observada pela seguradora não se distanciará de maneira significativa do seu valor esperado. Geralmente, esse fato nos conduz à aplicação do princípio de que o prêmio deve ser baseado na perda média esperada que é transferida do segurado para o segurador. Deve existir também um carregamento para os custos administrativos, custos de capital, entre outros.

Seguindo esse ponto de vista, é possível chegar à conclusão de que a necessidade da seguradora recorrer ao emprego de métodos estatísticos mais avançados para o cálculo do prêmio decorre, principalmente, do fato das perdas médias esperadas variarem bastante entre as apólices. A heterogeneidade dos riscos individuais sugere a ideia da necessidade da utilização de modelos estatísticos mais robustos e que se mostrem suficientemente capazes de precificar o risco dos segurados em função de seu perfil, viabilizando inferências acerca das perdas médias esperadas por apólice com um maior nível de assertividade.

2.3 Tarifação de Seguros

O estudo de modelagem estatística desenvolvido pelos atuários objetivando calcular o preço dos seguros é normalmente denominado, segundo o jargão existente no meio atuarial, de análise de precificação ou de análise tarifária. Essa análise que se desenvolve no âmbito interno das seguradoras, via de regra, toma como base os dados cadastrais relativos às apólices e ao perfil dos segurados, os dados de sinistros observados com o desenvolvimento da carteira da seguradora, e ainda, quando possível, informações externas confiáveis coletadas em bancos de dados públicos.

2.3.1 Exposição

A exposição de uma apólice ao risco de sinistro durante o seu período de vigência representa um conceito bastante relevante dentro da atividade da tarifação de seguros. Com efeito, de uma maneira bastante objetiva, a exposição pode ser entendida como sendo a fração representativa do tempo total da análise tarifária em que uma apólice fica exposta diretamente ao risco.

"*No cálculo da exposição ao risco leva-se em consideração a relação entre o tempo em que o risco ficou exposto no período de análise e o tempo total do período de análise, mesmo que o risco tenha iniciado antes do período de análise*" (FERREIRA, 2010).

Ohlsson e Johansson (2010) afirmam que a exposição, ou duração de uma apólice, é a quantidade de tempo em que a mesma permanece vigente, sendo esse período de tempo, geralmente, mensurado em anos, podendo-se, nesse caso, utilizar o próprio prazo de vigência da apólice em anos.

2.3.2 Sinistro

Sinistro é todo e qualquer evento de natureza aleatória, futura, e incerta, que possa ser reclamado pelo segurado à seguradora com o intuito e o objetivo de reivindicar algum tipo de compensação financeira decorrente da concretização de uma perda ou de algum dano observado ao objeto segurado.

Segundo Ohlsson e Johansson (2010) um sinistro é um evento reclamado pelo titular da apólice à seguradora, através do qual o mesmo demanda um ressarcimento por meio de uma compensação econômica.

2.3.3 Frequência

A frequência de sinistros nada mais é do que o número de reclamações observadas por uma apólice em um determinado período de tempo no qual a mesma permanece vigente.

Conforme Ohlsson e Johansson (2010), a frequência de sinistros é o número de sinistros dividido pela exposição, para algum grupo de apólices em vigor durante um determinado período de tempo específico, isto é, é o número médio de sinistros por unidade de período de tempo, geralmente calculada sobre uma base anual. As frequências de sinistros são geralmente expressas em unidades relativas por mil, mensurando assim o número de sinistros observados a cada mil apólices em exposição ao risco.

$$E[N] = N = Frequência\ do\ sinistro = \frac{Número\ de\ sinistros}{Número\ de\ unidades\ expostas\ ao\ risco} \qquad (3)$$

2.3.4 Severidade

A severidade dos sinistros representa o valor médio das reclamações observadas por uma determinada apólice em um dado período de tempo específico.

Ohlsson e Johansson (2010) definem a severidade de sinistros como sendo o montante total dos sinistros dividido pelo número de reclamações observadas, ou seja, é o custo médio por sinistro. Frequentemente, uma omissão ou inclusão de sinistros zerados, ou seja, sinistros que acabam por não produzir indenização alguma por parte da seguradora, deve representar um ponto claramente relevante no processo de modelagem e ajuste da frequência e da severidade dos sinistros.

$$E[X] = X = Severidade\ do\ sinistro = \frac{Montante\ de\ sinistros}{Número\ de\ sinistros} \qquad (4)$$

2.3.5 Prêmio de Risco

O prêmio de risco, também denominado de prêmio justo ou prêmio estatístico, é o valor calculado pelo atuário com vista à obtenção do equilíbrio financeiro e atuarial entre os fluxos de receitas e despesas a serem realizados pela seguradora. Esse tipo de prêmio tem a finalidade de cobrir, exclusivamente, o risco médio relacionado às indenizações de sinistros observados e reclamados pelos segurados, ou seja, esse prêmio representa a própria esperança matemática do risco segurado.

Formalmente, Ferreira (2010) define que o prêmio de risco é um valor cobrado do segurado com a finalidade, exclusiva, de cobrir o risco médio esperado, sendo este denotado matematicamente por P_R, ou $E[S]$.

$$P_R = E[S] = E[N].E[X] \qquad (5)$$

Onde S representa a variável aleatória "valor total das indenizações ocorridas em uma carteira de seguros" em um determinado período de tempo, período esse normalmente expresso sob a base anual, enquanto $E[N]$ e $E[X]$ representam, respectivamente, o número ou a frequência média esperada de sinistros e valor médio esperado por sinistro observado.

2.3.6 Prêmio Puro

Tem-se o prêmio puro como sendo um prêmio derivado do prêmio de risco. Nesse sentido, para chegar ao prêmio puro faz-se necessário adicionar uma margem de segurança estatística ao prêmio de risco calculado atuarialmente.

"*O prêmio puro é igual ao prêmio de risco mais um carregamento de segurança estatístico θ. Esse carregamento θ funciona como uma margem de segurança que visa cobrir as flutuações estatísticas do risco, de modo que exista uma probabilidade pequena dos sinistros superarem o prêmio puro*" (FERREIRA, 2010).

$$P_P = E[S].[1 + \theta] \qquad (6)$$

2.3.7 Prêmio Comercial

O prêmio comercial, como o próprio nome sugere, é o valor financeiro a partir do qual a seguradora operacionaliza a comercialização de seus produtos no mercado junto aos clientes. A tarifa comercial é calculada incorporando-se um percentual de carregamento α ao prêmio puro atuarialmente calculado no intuito de se fazer frente às despesas administrativas, operacionais, tributárias, e ainda garantir a margem de rentabilidade desejada pela seguradora.

"*O prêmio comercial π corresponde ao prêmio puro acrescido do carregamento para as demais despesas da seguradora α, incluída uma margem para lucro*" (FERREIRA, 2010).

$$P_C = \pi = \frac{P_P}{[1-\alpha]} = \frac{E[S].[1+\theta]}{[1-\alpha]} \qquad (7)$$

3 MODELOS LINEARES GENERALIZADOS – MLGs

Esse capítulo tem por objetivo introduzir os conceitos e definições básicas acerca dos Modelos Lineares Generalizados – MLGs, expondo uma fundamentação teórica relativa aos fatores de risco, à caracterização das variáveis, às hipóteses básicas do modelo, e à família exponencial de distribuições de probabilidade, assim como aos demais aspectos igualmente relevantes e intrínsecos à classe estatística dos MLGs.

3.1 Introdução

Os Modelos Lineares Generalizados representam uma ampla classe de métodos estatísticos que vem sendo bastante difundida e empregada pelos atuários no mercado de seguros e em diversas áreas da ciência. Os MLGs foram introduzidos inicialmente pelos atuários britânicos Nelder e Wedderburn (1972) com a proposta de serem utilizados para a avaliação e mensuração do relacionamento existente entre uma variável resposta Y e as respectivas variáveis explanatórias $X_{i,s}$ de um modelo de regressão. Alguns anos mais tarde McCullagh e Nelder (1983) aplicaram os então recentes MLGs à modelagem de dados de seguro automóvel, dando continuidade à difusão desse método. Passados mais alguns anos, McCullagh e Nelder (1989) vieram a consolidar e definir formalmente, em um excelente e famoso livro, essa ampla classe de modelos que passara a englobar vários arcabouços de regressão já conhecidos até então, incluindo modelos como os do tipo probit, logit, log-lineares, e o próprio modelo linear clássico normal. Os autores propuseram um processo iterativo para a estimação dos parâmetros, introduziram o conceito de função desvio para a avaliação da qualidade do ajuste e dos resíduos, além de técnicas e medidas de diagnóstico.

Segundo a definição observada em McCullagh e Nelder (1989), um Modelo Linear Generalizado fundamenta-se em três componentes estruturais essenciais, sendo elas:

i. ***Componente estocástica:*** *variável resposta Y que se pretende modelar possui caráter estocástico, com Y seguindo uma distribuição da família exponencial.*

ii. ***Componente sistemática:*** *uma combinação linear das variáveis explanatórias.*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + +\beta_p X_{ip} \,, \forall i \in \{1,2, \dots, n\} \qquad (8)$$

iii. ***Função de ligação:*** *uma função diferenciável, invertível e monótona que aplica uma transformação no vetor de variáveis explanatórias* $X^T = [X_1, X_2, \dots, X_n]$*, e estabelece uma associação entre a componente aleatória e a sistemática. A função link pode ser*

expressa matematicamente por $g(.)$ *e objetiva realizar uma suave linearização dos dados, transformando o valor esperado da variável resposta* $\mu_i = E(Y_i)$ *com o preditor linear expresso por:*

$$g(\hat{\mu}_i) = \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip} \, , \forall i \in \{1,2, \dots, n\} \qquad (9)$$

Assim, compreendendo os Modelos Lineares Generalizados como uma extensão dos modelos lineares normais, tem-se que a ideia básica da modelagem GLM consiste em abrir o leque de distribuições de probabilidade para a variável resposta, permitindo que a mesma seja escolhida da família exponencial de distribuições, além de proporcionar uma maior flexibilidade à relação funcional entre a média da variável resposta $\hat{\mu}$ e o seu preditor linear $\hat{\eta}$.

Segundo Jong e Heller (2008), a modelagem de dados via MLG difere do uso dos modelos lineares normais de regressão em dois importantes aspectos:

i. *A distribuição da variável resposta é escolhida da família exponencial. Assim, a distribuição da variável resposta não necessita ser normal, ou aproximadamente normal, e pode ser explicitamente não-normal;*
ii. *Uma transformação da média da variável resposta relaciona-se linearmente às variáveis explanatórias.*

Jong e Heller (2008) afirmam ainda que uma consequência direta do fato da variável dependente pertencer à família exponencial é que a resposta pode ser, e geralmente é, heterocedástica. Assim, na prática, observa-se que a variância pode oscilar junto à média, a qual por sua vez pode oscilar com as variáveis explanatórias. Isso contrasta fortemente com a hipótese de homocedasticidade do modelo de regressão linear clássico normal.

Uma outra argumentação suficientemente robusta para a não adequação do modelo linear clássico de regressão à análise tarifária de seguros não-vida pode ser encontrada em Ohlsson e Johansson (2010). Os autores afirmam que o modelo clássico apresenta outras duas fragilidades, quais sejam:

i. *Assume erros aleatórios normalmente distribuídos, enquanto o número de sinistros segue uma distribuição de probabilidade discreta definida nos inteiros não-negativos, e os custos dos sinistros são não-negativos e quase sempre com assimetria à direita;*
ii. *A média é uma função linear das variáveis explanatórias, enquanto os modelos multiplicativos mostram-se geralmente mais razoáveis para a precificação.*

Os Modelos Lineares Generalizados têm se mostrado uma opção bastante viável e interessante para a análise de dados em seguros, tendo em vista que as hipóteses estabelecidas

pelo modelo linear clássico são quase sempre violadas, e, portanto, não verificáveis na prática. No caso específico dos modelos de precificação, como já discutido anteriormente, a variável resposta possui como característica uma acentuada assimetria positiva à direita, assimetria essa que de forma alguma pode ser captado e modelada por meio de uma distribuição de probabilidade simétrica como é o caso da Normal.

3.2 Risk Factors e Key Ratios

Em relação ao prêmio a ser pago por cada apólice de um portfólio de seguro, observa-se que este pode ser calculado como sendo uma função matemática existente entre um determinado vetor de variáveis explicativas e uma variável dependente de interesse, de forma que seja possível mensurar o relacionamento existente entre estas e que se possa assim estimar o custo esperado com o pagamento das indenizações de sinistros. Em particular, essas variáveis explicativas são denominadas na literatura atuarial de tarifação por *risk factors*.

Os *risk factors*, ou fatores de risco, são variáveis empregadas no desenvolvimento dos modelos de tarifação objetivando identificar e estabelecer a relação existente entre diferentes atributos, e enquadram-se geralmente em uma das três seguintes categorias a saber:

i. ***Características e propriedades do titular da apólice****: idade, sexo, renda, entre outros;*
ii. ***Características e propriedades do objeto segurado****: idade de um veículo, modelo de um veículo, marca de um veículo, entre outros;*
iii. ***Características e propriedades da região geográfica****: renda per capita, densidade populacional de segurados por região, entre outros.*

A inclusão de variáveis no modelo de tarifação, quando da concepção deste, dependerá diretamente do grau de acessibilidade e disponibilidade dos dados por parte da seguradora. Enquanto dados cadastrais como idade e sexo são bem mais fáceis de se obter, dados relativos a hábitos e comportamentos de motoristas de automóveis, ou a indivíduos em relação ao uso de cigarro, álcool, ou drogas que possam afetar a saúde, são bem mais difíceis de se obter. Além disso, deve-se observar ainda que os *risk factors* não deverão causar qualquer tipo de constrangimento ou ofensa moral aos segurados, como por exemplo, num modelo onde a seguradora decida tarifá-los em função de sua etnia ou classe socioeconômica.

Busca-se abordar nesse trabalho os aspectos da tarifação aplicáveis às situações onde as variáveis explanatórias referentes aos fatores de risco podem ser modeladas através da segmentação destas em classes ou níveis de risco.

Variáveis como frequência, severidade, prêmio, e sinistralidade, são chamadas de *key ratios* no contexto da análise tarifária. Conceitualmente, uma *key ratio* pode ser definida como sendo uma variável originada através de uma transformação aplicada a alguma outra variável já existente a fim de se obter uma nova variável resposta de interesse, variável essa que deve representar o elemento objetivo dentro da análise tarifária a ser realizada.

No Quadro 1 encontram-se exemplos de algumas das mais importantes *key ratios* estudadas no contexto da análise tarifária.

Quadro 1 – Tipos de *key ratios* mais importantes – Fonte: Ohlsson e Johansson (2010).

Exposição w	***Resposta X***	***Key Ratio*** $Y = X/w$
Duração	*Número de sinistros*	*Frequência de sinistros*
Duração	*Custo do sinistro*	*Prêmio puro*
Número de sinistros	*Custo do sinistro*	*Severidade média do sinistro*
Prêmio ganho	*Custo do sinistro*	*Sinistralidade*
Número de sinistros	*Número de grandes sinistros*	*Proporção de grandes sinistros*

A partir da análise do Quadro 1 pode-se perceber que a *key ratio* representa uma razão entre uma variável aleatória X de interesse e um componente não aleatório w, podendo assim ser interpretada como sendo uma média da variável resposta de interesse por unidade de exposição, que, via de regra, é medida em anos, podendo também eventualmente variar e ser fracionada em meses, ou até mesmo em dias em determinadas situações. Nessa pesquisa, as *key ratios* trabalhadas serão aquelas referentes à frequência e à severidade média dos sinistros, que quando combinadas permitem estimar o prêmio a ser pago pelos segurados.

Ohlsson e Johansson (2010) afirmam que todas as *key ratios* possuem uma mesma natureza, sendo estas resultantes de uma razão entre uma variável aleatória e uma medida de volume, a qual chamamos de exposição, como por exemplo a razão entre o montante total de sinistros e o número total de reclamações, o que resulta na severidade média dos sinistros. Dessa maneira, a análise tarifária passa a ser realizada em relação a *key ratio* $Y = X/w$, ao invés de ser realizada com as variáveis resposta originais. Além disso, observa-se que a exposição exerce um papel fundamental na análise tarifária, pois quanto maior for o período de exposição da massa de dados sob análise, menor será a variabilidade e a instabilidade dos *key ratios*, e maior será a credibilidade estatística observada para as estimativas e inferências realizadas em relação à variável resposta de interesse.

3.3 Hipóteses e Estrutura Básica

Nesta seção são apresentadas algumas hipóteses e formulações básicas necessárias ao desenvolvimento e aplicação dos Modelos Lineares Generalizados. Entretanto, convém observar que as hipóteses aqui apresentadas, juntamente à definição estrutural básica acerca dos MLGs, não pretendem esgotar esse tópico, representando apenas uma breve introdução ao estudo do referido método estatístico. Ohlsson e Johansson (2010) enunciam três hipóteses básicas a serem consideradas acerca dos MLGs no contexto da tarifação, sendo elas:

- ***Hipótese 1: Independência das apólices:*** *considere n diferentes apólices. Para alguma variável resposta contida no Quadro 1, e sendo X_i uma observação dessa variável para uma dada apólice qualquer i, então $X_1, \dots, X_n$ representará um conjunto de observações independentes entre si.*
- ***Hipótese 2: Independência no tempo:*** *considere n distintos intervalos de tempo. Para alguma variável resposta contida no Quadro 1, e sendo X_i uma observação dessa variável para um dado período de tempo qualquer i, então $X_1, \dots, X_n$ representará um conjunto de observações independentes entre si.*
- ***Hipótese 3: Homogeneidade:*** *considere quaisquer duas apólices pertencentes a uma mesma célula tarifária e com um mesmo período de exposição ao risco. Para alguma variável resposta contida no Quadro 1, e sendo X_i uma realização dessa variável para uma apólice qualquer i, então X_1 e X_2 possuirão a mesma distribuição de probabilidade.*

Entretanto, essas três hipóteses básicas, apesar de parecerem bastante razoáveis a priori, podem, em alguns casos, acabar por não se verificar integralmente na prática.

Ohlsson e Johansson (2010) afirmam, por exemplo, que em um seguro automóvel é possível a ocorrência de uma colisão entre dois veículos segurados pela mesma companhia, situação essa que ao ser observada viola a hipótese de independência entre as apólices. Os mesmos citam ainda que um claro exemplo de dependência entre as apólices é o caso da ocorrência de catástrofes, onde inúmeras apólices podem ser afetadas por um mesmo tornado, terremoto, ou inundação em simultâneo. Porém, segundo os autores, o risco de se negligenciar esse fato deve exercer um efeito mínimo sobre o risco global da carteira segurada.

Os autores apontam ainda supostas fragilidades existentes nas hipóteses básicas quanto à independência temporal e à homogeneidade dos riscos, apesar de afirmarem que, no geral, as hipóteses aqui expostas além de serem bastante razoáveis, proporcionam um ganho substancial de simplificação na concepção e desenvolvimento do modelo tarifário.

3.3.1 Médias e Variâncias

Em suma, os modelos de precificação desenvolvidos no contexto de uma análise tarifária têm como principal objetivo a estimação dos valores médios esperados do risco por célula tarifária, ou seja, calcular as indenizações com sinistros para um dado subconjunto de fatores de risco, sendo este o próprio perfil de risco individual dos segurados. Todavia, para a determinação de uma estimativa mais consistente do risco, além do cálculo do valor médio esperado faz-se preciso levar em consideração a variabilidade do modelo: a variância.

A seguir encontram-se descritas algumas implicações decorrentes das hipóteses 1, 2 e 3, anteriormente expostas, além das formulações para a média e a variância do modelo, estando tais relações demonstradas e detalhadas formalmente em Ohlsson e Johansson (2010).

Seja uma *key ratio* arbitrária, denotada por $Y = X/w$, para um grupo de apólices pertencentes a uma célula tarifária com exposição total w, e resposta total expressa por X.

Considerando inicialmente uma situação onde w seja o número observado de sinistros. Então, torna-se possível escrever X como sendo um somatório de w respostas individuais $Z_1, \dots, Z_w$, e assim, sendo X o custo agregado total dos sinistros, Z_k será o custo observado com o k-ésimo sinistro. Com isso, as hipóteses 1 e 2 implicam que os Z_k's são independentes, dado que os sinistros provêm de diferentes apólices em diferentes períodos de tempo. Logo, a hipótese 3 implica numa distribuição idêntica, tornando-se possível escrever que $E(Z_k) = \mu$ e $Var(Z_k) = \sigma^2$, para algum μ e σ^2. Assim, é possível definir explicitamente as expressões analíticas para a média e a variância como sendo:

$$E(X) = w\mu \qquad Var(X) = w\sigma^2 \qquad (10)$$

$$E(Y) = \mu \qquad Var(Y) = \sigma^2/w \qquad (11)$$

- ***Lema 1*** *- De acordo com as hipóteses 1, 2 e 3, sendo X alguma variável resposta contida no Quadro 1 com* $w > 0$ *e* $Y = X/w$*, então a esperança e a variância de X e de Y podem ser expressas respectivamente pelas equações (10) e (11), sendo* μ *e* σ^2*, respectivamente, a esperança e a variância para uma dada observação da variável resposta com exposição* $w = 1$.

O resultado obtido com a definição do *Lema 1* se mostra válido para as situações onde a exposição w é igual ao número de sinistros, como por exemplo, nos modelos de severidade. Entretanto, existem duas outras possíveis situações, sendo elas: quando w é igual a vigência da apólice, ou quando w é igual ao prêmio ganho, nos modelos de frequência e sinistralidade, respectivamente.

Considere-se outra situação onde w seja um número racional definido por $w = m/n$. Nessa situação específica a exposição de uma apólice pode ser fracionada em m partes de mesmo tamanho $1/n$ cada. No caso onde w é a vigência, ou exposição da apólice, esse fracionamento é realizado tomando-se m intervalos de tempo do mesmo comprimento, enquanto que no caso onde w é o prêmio ganho é possível tomar intervalos de tempo que sejam suficientemente longos para o reconhecimento da fração $1/n$ do prêmio em cada período.

Seja um conjunto de realizações da variável resposta observadas em cada intervalo de tempo m representado por $Z_1, \dots, Z_m$, as quais representam um conjunto de variáveis aleatórias independentes e identicamente distribuídas. Se adicionarmos n respostas similares Z_k então obtemos a variável Z com exposição $w = 1$. De acordo com a hipótese 3 de homogeneidade dos riscos em uma mesma célula tarifária temos que cada uma das observações Z possui a mesma esperança e variância, sendo representadas respectivamente por $E(Z) = \mu$ e $Var(Z) = \sigma^2$.

Disso, podemos obter que:

$$E(Z) = nE(Z_1) \qquad Var(Z) = nVar(Z_1) \qquad (12)$$

$$E(Z_k) = E(Z_1) = \mu/n \qquad Var(Z_k) = Var(Z_1) = \sigma^2/n \qquad (13)$$

Assim, sendo $X = \sum_{k=1}^{m} Z_k$, podemos expressar a esperança e a variância de X por:

$$E(X) = \mu m/n \qquad Var(X) = \sigma^2 m/n \qquad (14)$$

Dessa forma, temos que essas relações matemáticas validam e sustentam as hipóteses 1 e 2 de independência das apólices e das observações no tempo.

Conforme Ohlsson e Johansson (2010), isso prova o Lema 1 para os casos onde a exposição w é um número racional, caso esse que se observa recorrentemente na prática da análise tarifária. Outra consequência do Lema 1 é a necessidade da utilização consistente de ponderações para a variância dos modelos para quaisquer que sejam a *key ratios* em estudo.

3.4 Modelos Multiplicativos

Descreve-se nessa seção uma estrutura formal de modelagem multiplicativa GLM bastante utilizada na prática da análise tarifária adaptada de Ohlsson e Johansson (2010).

Seja um modelo genérico qualquer composto por M *risk factors*, cada um deles subdividido em classes, onde m_i denota o número de classes em cada *risk factor* i. Por

simplificação, supondo que esse modelo possua somente dois *risk factors*, ou seja, que $M = 2$, torna-se possível representar uma célula tarifária pela notação (i, j), onde i e j representam as classes do primeiro e segundo *risk factor*, respectivamente. Na célula (i, j) tem-se os valores observados para a exposição w_{ij} e para a resposta X_{ij}, o que implica na possibilidade de se obter a *key ratio* $Y_{ij} = X_{ij}/w_{ij}$. De acordo com o Lema 1, tem-se que $E(Y_{ij}) = \mu_{ij}$, onde μ_{ij} representa o valor médio esperado por unidade de exposição w_{ij}. Assim, o modelo multiplicativo pode ser expresso por:

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j} \qquad (15)$$

Onde $\{Y_{1i}; i = 1, \dots, m_1\}$ são parâmetros relativos às classes de risco do *risk factor* 1, enquanto que $\{Y_{2j}; j = 1, \dots, m_2\}$ são parâmetros referentes aos níveis do *risk factor* 2. Já γ_0 representa um valor base referente à célula tarifária em análise.

Visando mensurar o impacto dos parâmetros relativos aos níveis de risco sobre a variável resposta do modelo, é comum a escolha de uma célula tarifária para servir como célula de referência, ou célula base, que de preferência deve ser aquela detentora do maior volume de exposição ao risco em cada fator de risco, de forma a maximizar a credibilidade estatística e minimizar a variabilidade das estimativas do modelo.

Por simplificação, seja uma célula base denotada por $(1,1)$, e seja $\gamma_{11} = \gamma_{21} = 1$. Então γ_0 pode ser interpretado como sendo um valor base, ou seja, a *key ratio* para as apólices pertencentes à célula base, e os outros parâmetros γ_{ij} podem ser interpretados como uma medida de diferença relativa das demais células tarifárias em relação à célula base do *risk factor*, sendo essas diferenças relativas entre as células denominadas de relatividades tarifárias. Para uma melhor compreensão, supondo que $\gamma_{12} = 1{,}25$, então, tem-se assim que o valor médio esperado para a variável resposta na célula $(2,1)$ é 25% maior que o valor médio esperado para a variável resposta na célula $(1,1)$. Analogamente, a mesma interpretação pode ser estendida para as células $(1,2)$ e $(2,2)$ em relação à célula base de referência $(1,1)$.

O modelo multiplicativo apresentado acima para dois *risk factors*, ou seja, para o caso onde $M = 2$, pode facilmente ser estendido para um caso geral ao custo de uma notação algébrica um pouco mais complexa, ficando escrito da seguinte forma:

$$\mu_{i1}, \mu_{i2}, \dots, \mu_{iM} = \gamma_0 \gamma_{1i_1} \qquad (16)$$

Uma discussão mais aprofundada acerca da utilização dos modelos multiplicativos em detrimento dos modelos aditivos no contexto da atividade de tarifação de seguros pode ser obtida em Murphy, Brockman e Lee (2000).

Por fim, analisando a questão da distinção do valor do prêmio entre as apólices, é possível considerar adequada a escolha pelos modelos multiplicativos, tendo em vista o fato do nível global do prêmio calculado poder ser controlado através do ajuste do valor base do modelo representado pelo parâmetro γ_0, enquanto que os outros parâmetros γ_{ij} permitem calibrar o quanto se faz necessário cobrar a cada apólice em particular em função das medidas de relatividades tarifárias. Segundo Ohlsson e Johansson (2010), na prática, primeiramente devem ser estimadas as relatividades tarifárias do modelo γ_{ki_k}, para só então posteriormente utilizá-las no ajuste do valor base γ_0 para a obtenção final do prêmio individual a ser cobrado de cada segurado.

3.5 Família Exponencial de Distribuições

Seguindo as hipóteses já ilustradas anteriormente, seja $Y = \{y_1, y_2, \dots, y_n\}$ um conjunto de observações aleatórias independentes $Y_{i's}$ de acordo como definido na teoria geral dos MLGs, então a função distribuição de probabilidade f.d.p. de um modelo de dispersão exponencial pode ser escrito na seguinte forma:

$$f_{Y_i}(y_i; \theta_i; \phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i; \phi; w_i)\right] \qquad (17)$$

Sendo θ_i um parâmetro variável que guarda uma relação de dependência com cada caso i, enquanto que $\phi > 0$ representa um parâmetro de dispersão assumido constante para todo i. θ_i e ϕ são parâmetros de localização e escala da distribuição de probabilidade da variável resposta Y, onde $b(\theta_i)$ é a função cumulante, assumida contínua e diferenciável, sendo invertível, e admitindo derivada segunda. Para cada escolha diferente realizada para a função $b(\theta_i)$ torna-se possível obter uma família específica de distribuições de probabilidade, como por exemplo: Normal, Gamma, Binomial, Poisson, entre outras. O parâmetro variável θ_i também é chamado comumente de parâmetro natural da distribuição. Assim, tem-se que, em geral:

$$E(Y) = \mu = b'(\theta) \qquad (18)$$

$$Var(Y) = a(\phi)b''(\theta) \qquad (19)$$

No Quadro 2 encontram-se listadas algumas das distribuições de probabilidade da família exponencial mais conhecidas e empregadas na modelagem de dados via MLG. Os componentes $a(.)$, $b(.)$ e $c(.)$ representam funções conhecidas que variam em conformidade à natureza da distribuição empregada.

Quadro 2 – Funções a(.), b(.) e c(.) para a Família Exponencial – Fonte: Próprio autor.

Distribuição	$\boldsymbol{a(\phi)}$	$\boldsymbol{b(\theta)}$	$\boldsymbol{c(y;\phi)}$
Gaussiana	ϕ	$\theta^2/2$	$-1/2\,[y^2/\phi + log_e(2\pi\phi)]$
Binomial	$1/n$	$log_e(1+e^{\theta})$	$log_e\binom{n}{ny}$
Poisson	1	e^{θ}	$-log_e\, y!$
Gamma	ϕ	$-log_e(-\theta)$	$\phi^{-2}\, log_e(Y/\phi) - log\, ey - log_e\, \Gamma(\phi^{-1})$
Gaussiana Inversa	ϕ	$-\sqrt{-2\theta}$	$-1/2\,[log_e(\pi\phi y^3) + 1/(\phi y)]$

Para uma escolha particular qualquer de $b(.)$, a f.d.p. de Y pode ser suficientemente especificada pelos parâmetros θ_i e ϕ. Já a função $c(.;.)$, como não depende diretamente de θ_i, acaba gerando pouco interesse dentro da teoria dos MLGs. Outro aspecto relevante são os intervalos para os quais a resposta de Y está definida. Via de regra, as possíveis observações de resposta para a variável Y possuem suporte nos intervalos $(0, \infty)$ e $(-\infty, \infty)$, além dos inteiros não negativos para os casos discretos com variáveis aleatórias para dados de contagem.

Outras restrições para os modelos de dispersão exponenciais no contexto da análise tarifária são que o parâmetro de dispersão e de exposição devem ser respectivamente $\phi > 0$, e $w_i \geq 0$.

Jong e Heller (2008) introduzem ainda uma segunda função no contexto dos modelos de dispersão exponencial, a função *link*. Segundo os autores, a função de ligação, ou *link function*, especifica que uma transformação do valor médio da variável resposta, expressa por $g(\mu)$, relaciona-se linearmente com o vetor de variáveis explanatórias $X^T = [X_1, X_2, \dots, X_n]$.

$$g(\mu) = X'\beta \qquad (20)$$

Assim, a escolha de $g(\mu)$, determina a forma como a média da variável resposta se relaciona com as variáveis explanatórias X. No modelo linear clássico normal o relacionamento entre a média de Y e as variáveis explanatórias é dado por $\mu = X'\beta$. Para os MLGs, esse relacionamento pode ser generalizado por meio da notação $g(\mu) = X'\beta$, onde $g(.)$ representa uma função monótona e diferenciável, tal como a função logarítmica, inversa, raiz quadrada, entre outras.

3.6 Função de Ligação e Ligação Canônica

A função de ligação tem por objetivo transformar o valor médio esperado da variável resposta Y de modo com que esta estabeleça um relacionamento linear com o vetor de variáveis explanatórias $X^T = [X_1, X_2, \dots, X_n]$. Nesse sentido, para a modelagem da média de Y, representada aqui por μ, considere-se a transformação expressa em (21) por:

$$\eta_i = g(\mu_i) \qquad (21)$$

Assumindo que essa transformação da média μ siga um modelo linear normal:

$$\eta_i = X_i'\beta \qquad (22)$$

Sendo η_i o preditor linear para a média da variável resposta do modelo. Além disso, note-se que o modelo obtido em (22) é estruturalmente mais simples do que em (21). Aplicando uma transformação inversa sobre a função de ligação é possível obter que:

$$\mu_i = g^{-1}(X_i'\beta) \qquad (23)$$

No Quadro 3 encontram-se algumas das funções de ligação mais empregadas na modelagem GLM.

Quadro 3 – Funções de Ligação mais utilizadas – Fonte: Próprio autor.

Função de Ligação	$\boldsymbol{g(\mu)}$	***Ligação Canônica***
Identidade	μ	*Normal*
Log	$ln(\mu)$	*Poisson*
Power	μ^p	*Gamma* $(p = -1)$ *e Gaussiana Inversa* $(p = -2)$
Raiz quadrada	$\sqrt{\mu}$	
Logit	$ln(\mu/1-\mu)$	*Binomial*

Essas funções, ao realizarem a operação de inversão do preditor linear descrita em (23), permitem a estimação de μ_i de uma forma bem menos trabalhosa, e por conseguinte, tornam válidas as relações contidas em (24) em ambos os sentidos:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(X_i'\beta) = g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}\right) \qquad (24)$$

Um aspecto prático bastante relevante dentro da modelagem GLM ocorre quando do uso da função de ligação canônica. A ligação canônica permite que a inversão realizada em (23) resulte no próprio parâmetro natural da distribuição θ_i. Nesse caso, se diz que há ligação canônica sempre que as relações expressas em (25) forem válidas em ambos os sentidos.

$$g(\mu_i) = \theta_i \Leftrightarrow \eta_i = \theta \qquad (25)$$

3.7 Estimação Paramétrica

Uma importante propriedade dos Modelos Lineares Generalizados consiste no fato de todos eles poderem ter seus parâmetros $\hat{\beta}$ ajustados a um conjunto de dados sob interesse por meio de um estimador de máxima verossimilhança denominado *Iterative Re-Weighted Least Squares* – IRLS. Esse estimador de máxima verossimilhança, *Maximum Likelihood Estimator* – MLE, formalmente introduzido por McCullagh e Nelder (1989), baseia-se num algoritmo matemático iterativo de otimização numérica que busca efetuar reponderações dos parâmetros estimados objetivando minimizar a soma dos quadrados dos resíduos entre os valores esperados e os observados para a variável resposta do modelo.

Seja uma estimativa inicial obtida para os parâmetros $\hat{\beta}$ através dos dados em análise. A partir dessa estimativa suponha-se que seja possível estimar também o preditor linear $\hat{\eta}_i = X_i'\hat{\beta}$, e a partir disso utilizá-lo para a obtenção do ajuste de $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Suponha-se ainda que a partir desses valores seja possível estimar a variável dependente definida por:

$$Z_i = \hat{\eta}_i + (Y_i - \hat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i} = \hat{\eta}_i + (Y_i - \hat{\mu}_i)g'(\hat{\mu}_i) \qquad (26)$$

Então, torna-se possível escrever que:

$$E(Z_i) = \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} \qquad (27)$$

$$Var(Z_i) = [g'(\mu_i)]^2 a_i v(\mu_i) \qquad (28)$$

Dessa maneira, se for possível estimar os valores para Z_i, então será possível ajustar um modelo de regressão para Z em função das variáveis explanatórias X através de um método de reponderação por mínimos quadrados e da aplicação da função inversa para a $Var(Z_i)$ como fator de reponderação.

Assim, torna-se possível obter os pesos iterativos a partir da seguinte função:

$$w_i = \frac{p_i}{\left[b''(\theta_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2\right]} \qquad (29)$$

Onde $b''(\theta_i)$ representa a derivada segunda de $b(\theta_i)$, escolhida por hipótese, e assumida como $a(\phi)$ possuindo uma forma usual ϕ/p_i. Logo, o peso w_i calculado torna-se inversamente proporcional à variância da variável dependente Z_i, resultando das estimativas calculadas para os parâmetros, com sendo um fator de proporcionalidade ϕ.

Dessa forma, torna-se possível obter estimativas mais consistentes para os coeficientes $\hat{\beta}$ da regressão de Z_i em função de X utilizando os pesos w_i.

A estimativa ponderada de mínimos quadrados pode ser obtida por:

$$\hat{\beta} = (X'WX)^{-1}X'WZ \qquad (30)$$

Onde X é o vetor de variáveis explanatórias $X^T = [X_1, X_2, \dots, X_n]$, W é a matriz diagonal dos pesos com valores de entrada iniciais para w_i, e Z é o vetor contendo as observações da variável resposta com os valores de entrada para z_i.

De acordo com o algoritmo proposto por McCullagh e Nelder (1989) esse processo de reponderação iterativo deve ser repetido sucessivas vezes de forma que as estimativas para os parâmetros $\hat{\beta}$ do modelo se tornem estáveis e venham a convergir levando o erro relativo a atingir um valor menor que o estabelecido como critério de parada do algoritmo de otimização.

3.8 Função Deviance

A função *deviance*, ou função desvio, representa uma medida estatística de avaliação da qualidade em relação aos resíduos obtidos com o ajuste de um MLG. Suponha-se que $l(\hat{\mu})$ seja a função de log-verossimilhança para o conjunto de dados em análise, e que essa função possa ser por sua vez obtida a partir do próprio vetor de médias estimado $\hat{\mu}^T = [\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n]$. Suponha-se ainda a existência de um modelo capaz de proporcionar um ajuste ideal aos dados amostrais, de forma que nesse modelo o número de parâmetros k seja igual ao número de observações amostrais para a variável resposta n, com $\widehat{\mu}_i = y_i, \forall i \in \{1,2, \dots, n\}$. Esse modelo onde $k = n$ é chamado de modelo saturado e deve servir como parâmetro de referência ou *benchmark* para a avaliação da qualidade do ajuste de todos os outros submodelos possíveis detentores de um número de parâmetros diferente do modelo saturado.

De acordo com Ohlsson e Johansson (2010) a medida de teste para a função *deviance* é a *scaled deviance D**, a qual é definida como uma estatística de teste para a razão de verossimilhança, ou *Likelihood Ratio Test* – LRT, do modelo testado contra o modelo completo ou saturado. A estatística de teste LRT é definida conforme a expressão (31) como sendo duas vezes o logaritmo da razão de verossimilhança, ou seja:

$$D^* = D^*(y, \hat{\mu}) = 2[l(y) - l(\hat{\mu})] \qquad (31)$$

4 METODOLOGIA

Nesse capítulo pretende-se descrever de maneira sucinta os aspectos fundamentais relativos à metodologia empregada, apresentando-se os principais atributos caracterizadores da pesquisa, a definição do escopo de trabalho, a seleção e o meio de obtenção do conjunto de dados utilizado, bem como os ajustes necessários realizados a fim de tratar e deixar os dados aptos à manipulação através do software estatístico.

4.1 Caracterização da Pesquisa

Embora essa pesquisa possa ser classificada, predominantemente, como sendo do tipo quantitativa, deve-se considerar também o fato de que a mesma se deu, inicialmente, por meio de um levantamento bibliográfico seletivo de natureza qualitativa.

Em relação aos objetivos gerais essa pesquisa pode ser classificada como sendo do tipo descritiva e explicativa. Decide-se aqui por tal classificação tendo em vista o interesse inicial de se explorar as variáveis a serem consideradas no modelo de tarifação, se avaliar a natureza da associação entre estas, e se mensurar o nível de influência e intensidade exercido pelas variáveis explicativas sobre a variável dependente do modelo.

Corroborando com a classificação adotada, Gil (2002) afirma que também são pesquisas de natureza descritiva aquelas em que o objetivo central é identificar a existência e a natureza da associação entre variáveis, como, por exemplo, nas pesquisas eleitorais em que se busca identificar a relação existente entre a preferência político-partidária e o nível de rendimento, ou de escolaridade, dos potenciais eleitores.

Gil (2002) afirma ainda que as pesquisas de natureza explicativa têm por objetivo a identificação dos fatores que determinam ou contribuem para a ocorrência de determinado evento ou fenômeno de interesse, sendo esse o tipo de pesquisa científica que mais aprofunda o conhecimento da realidade por explicar e justificar a razão e o porquê das coisas.

4.2 Apresentação do Conjunto de Dados

O exercício empírico aqui desenvolvido emprega um conjunto de dados, *dataset*, disponível em Ohlsson e Johansson (2010), referente a um seguro do tipo *moped insurance*. Tal *dataset* contempla dados reais de sinistros reclamados no portfólio de uma companhia

seguradora sueca chamada *Wasa*, num período anterior à fusão com outra seguradora também sueca chamada *Länsförsäkringar Alliance*.

Na Suécia, o seguro ciclomotor, *moped insurance*, envolve três diferentes tipos de coberturas contra sinistros, quais sejam:

i. ***TPL (Third party liability)****: cobertura contra sinistros relativos a qualquer tipo de lesão corporal ou injúria física causada a terceiros em acidentes de trânsito;*
ii. ***Partial casco****: cobertura contra roubo e também alguns outros danos, como, por exemplo, incêndio;*
iii. ***Hull****: cobertura de danos sofridos pelo próprio veículo do titular da apólice.*

O seguro *TPL* possui caráter obrigatório na Suécia, ou seja, para que o motorista tenha permissão de dirigir é preciso que ele contrate, no mínimo, a cobertura *TPL*, enquanto que as outras duas são opcionais. Essas três coberturas são vendidas na forma de um pacote geral, todavia, na prática, as seguradoras geralmente precificam cada uma das coberturas de forma independente. Entretanto, Ohlsson e Johansson (2010) afirmam que o *dataset* utilizado refere-se apenas à cobertura do tipo *partial casco*.

O modelo de tarifação aqui desenvolvido considera como variáveis explanatórias os fatores de risco listados no Quadro 4 e contidos no conjunto de dados *moped insurance*.

Quadro 4 – Variáveis tarifárias *moped insurance* – Fonte: Ohlsson e Johansson (2010).

Fator de Risco	***Classe***	***Descrição da Classe***
Classe do veículo	*1*	*Peso superior a 60 Kg e mais de duas marchas*
	2	*Outros*
Idade do veículo	*1*	*No máximo 1 ano*
	2	*2 anos ou mais*
Zona geográfica	*1*	*Parte central e semi-central das três maiores cidades da Suécia*
	2	*Subúrbios e cidades de médio porte*
	3	*Cidades menores com exceção daquelas enquadradas em 5 e 7*
	4	*Pequenas cidades e zonas rurais, com exceção de 5 e 7*
	5	*Cidades do norte*
	6	*Zonas rurais do norte*
	7	*Gotland: maior ilha da Suécia*

O Quadro 5 apresenta, expositivamente, o *layout* dos dados contidos no conjunto de dados *moped insurance*.

Quadro 5 – *Dataset* para o seguro *moped insurance* – Fonte: Ohlsson e Johansson (2010).

Célula Tarifária			***Exposição***	***Número de Sinistros***	***Frequência de Sinistros***	***Severidade de Sinistros***	***Prêmio Puro***	***Prêmio Atual***
Classe	***Idade***	***Zona***						
1	*1*	*1*	*62.9*	*17*	*270*	*18256*	*4936*	*2049*
1	*1*	*2*	*112.9*	*7*	*62*	*13632*	*845*	*1230*
1	*1*	*3*	*133.1*	*9*	*68*	*20877*	*1411*	*762*
1	*1*	*4*	*376.6*	*7*	*19*	*13045*	*242*	*396*
1	*1*	*5*	*9.4*	*0*	*0*	*0*	*0*	*990*
1	*1*	*6*	*70.8*	*1*	*14*	*15000*	*212*	*594*
1	*1*	*7*	*4.4*	*1*	*228*	*8018*	*1829*	*396*
1	*2*	*1*	*352.1*	*52*	*148*	*8232*	*1216*	*1229*
1	*2*	*2*	*840.1*	*69*	*82*	*7418*	*609*	*738*
1	*2*	*3*	*1378.3*	*75*	*54*	*7318*	*398*	*457*
1	*2*	*4*	*5505.3*	*136*	*25*	*6922*	*171*	*238*
1	*2*	*5*	*114.1*	*2*	*18*	*11131*	*195*	*594*
1	*2*	*6*	*810.9*	*14*	*17*	*5970*	*103*	*356*
1	*2*	*7*	*62.3*	*1*	*16*	*6500*	*104*	*238*
2	*1*	*1*	*191.6*	*43*	*224*	*7754*	*1740*	*1024*
2	*1*	*2*	*237.3*	*34*	*143*	*6933*	*993*	*615*
2	*1*	*3*	*162.4*	*11*	*68*	*4402*	*298*	*381*
2	*1*	*4*	*446.5*	*8*	*18*	*8214*	*147*	*198*
2	*1*	*5*	*13.2*	*0*	*0*	*0*	*0*	*495*
2	*1*	*6*	*82.8*	*3*	*36*	*5830*	*211*	*297*
2	*1*	*7*	*14.5*	*0*	*0*	*0*	*0*	*198*
2	*2*	*1*	*844.8*	*94*	*111*	*4728*	*526*	*614*
2	*2*	*2*	*1296.0*	*99*	*76*	*4252*	*325*	*369*
2	*2*	*3*	*1214.9*	*37*	*30*	*4212*	*128*	*229*
2	*2*	*4*	*3740.7*	*56*	*15*	*3846*	*58*	*119*
2	*2*	*5*	*109.4*	*4*	*37*	*3925*	*144*	*297*
2	*2*	*6*	*404.7*	*5*	*12*	*5280*	*65*	*178*
2	*2*	*7*	*66.3*	*1*	*15*	*7795*	*118*	*119*

O conjunto de dados empregado na análise tarifária compreende 860 observações de sinistros associados a 38.508 apólices expostas ao risco em pelo menos alguma fração do intervalo de tempo compreendido entre os anos de 1994 a 1999. A base de dados original continha alguns sinistros com valor nulo, os quais foram excluídos de forma a não integrarem o conjunto de dados utilizado na análise estatística.

O *dataset* utilizado contempla variáveis referentes aos fatores de risco, exposição, número de sinistros, frequência a cada mil observações, severidade, prêmio puro teórico, e o prêmio puro atual vigente no ano de 1999 para as apólices pertencentes a cada uma das 28 células tarifárias do modelo. O conjunto de dados foi importado para o ambiente computacional do software estatístico R Core Team (2014) e salvo num objeto do tipo *data.frame(.)*, sendo disposto no formato *list form* de maneira a conter 28 células tarifárias representadas como combinação dos respectivos fatores de risco, quais sejam: classe do veículo, idade do veículo, e zona do veículo.

Em Jong e Heller (2008) os autores listam um conjunto de procedimentos que, se executados em sequência, podem ser interpretados como um algoritmo supervisionado para a execução para a modelagem GLM. Segundo os autores, dada uma *key ratio* ou uma variável resposta qualquer de interesse Y, ajustar um MLG consiste em:

i. Escolher uma distribuição de probabilidade para a variável resposta $f(y)$;

ii. Selecionar uma função de ligação $g(\mu)$;

iii. Escolher as variáveis explanatórias $X_{i's}$ em termos das quais o valor $g(\mu)$ será modelado;

iv. Coletar as observações $\{y_1, y_2, \dots, y_n\}$ para a resposta Y e os correspondentes valores $\{x_1, x_2, \dots, x_n\}$ para as variáveis explanatórias $X_{i's}$;

v. Ajustar o modelo por meio das estimativas β, e quando desconhecido, φ;

vi. Dadas as estimativas de β, gerar predições ou valores ajustados de Y para as diferentes observações de X, e examinar quão bom é o ajuste do modelo por meio do exame comparativo entre os valores ajustados e os valores observados, além de realizar outros diagnósticos referentes ao modelo.

5 EXERCÍCIO EMPÍRICO

Esse capítulo objetiva demonstrar o uso dos Modelos Lineares Generalizados na tarifação de um seguro não-vida, descrevendo os critérios estatísticos empregados no ajuste e seleção dos modelos de frequência, severidade, e do modelo encaixado para o cálculo do prêmio. Para tanto, pretende-se realizar uma análise de consistência e significância estatística dos parâmetros estimados, além de uma avaliação crítica da qualidade do ajuste e dos resíduos gerados por meio dos MLGs. Por fim, far-se-á um sucinto diagnóstico dos modelos estimados com vista à concepção final dos subsídios necessários à conclusão da pesquisa.

5.1 Modelagem da Frequência dos Sinistros

O exercício empírico de aplicação da modelagem GLM desenvolvido aqui parte de uma descrição inicial da base de dados utilizada, contemplando em seguida a modelagem da frequência e da severidade dos sinistros, até chegar à concepção do modelo de cálculo do prêmio por célula tarifária. Os resultados descritos e discutidos daqui em diante, referentes à análise tarifária, foram obtidos por meio de uma sub-rotina de programação computacional desenvolvida no software estatístico de código aberto R Core Team (2014).

Nelder e McCullagh (1989) afirmam que a distribuição Poisson, sob condições estatísticas normais, tende a surgir naturalmente como uma potencial candidata à modelagem de variáveis discretas que não possuam valores máximos em seu domínio, defendendo ainda a razoabilidade da utilização de tal distribuição para a modelagem da variável aleatória número de sinistros. Adicionalmente, tem-se que a forma assimétrica positiva à direita da Poisson faz com que esta distribuição teórica, por muitas vezes, se ajuste de maneira satisfatória à curva empírica da frequência observada para os sinistros. Logo, na classe dos MLGS, a distribuição Poisson figura como uma forte candidata ao ajuste da frequência dos sinistros em razão de se assemelhar bastante ao comportamento esperado para a variável aleatória número de sinistros.

A utilização da distribuição Poisson em estudos atuariais dessa natureza já fora proposta anteriormente por outros autores, tais como em Klugman, Panjer e Willmot (2002), Santos (2008), Jong e Heller (2008), Ohlsson e Johansson (2010), Ferreira (2010), e Pereira e Carrasco (2010). Logo, com base no exposto acima e na literatura revisitada, uma primeira distribuição candidata natural ao ajuste da frequência há de ser a distribuição Poisson.

No intuito de não tornar a escolha da distribuição de probabilidade tendenciosa, e ao mesmo tempo alargar o leque de possíveis candidatas ao ajuste do modelo de frequência, além da distribuição Poisson foram selecionadas outras duas distribuições de probabilidade para o número médio de sinistros tendo em vista a sua disseminada utilização em aplicações atuariais, sendo elas as distribuições Binomial e Binomial Negativa.

Especificando a variável aleatória para o ajuste da frequência de sinistros, pode-se assumir inicialmente por hipótese, e tomando como referência a literatura atuarial relativa à precificação de seguros, que Y segue uma distribuição Poisson, de sorte que:

$Y \sim Poisson(\lambda)$, com uma função densidade de probabilidade definida por:

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, y = \{0,1,2,\dots,n\}, e\ \lambda > 0 \qquad (32)$$

Assim, Sendo $E(Y) = \mu = \lambda = e^{-\theta}$, onde $\theta = \ln(\lambda) = \ln(\mu)$, podemos escrever:

$$f(y|\lambda) = exp\left[\ln\left(e^{-\lambda}\right) + \ln(\lambda^{y}) - \ln(y!)\right] = exp[y\ln(\lambda) - \lambda - \ln(y!)]$$

Daí, segue que:

$$f(y|\lambda) = exp\left[y\theta - e^{\theta} - \ln(y!)\right] = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right] = f(y|\theta;\phi), \quad \text{onde}$$

$a(\phi) = 1$ e $\phi = 1$.

Dessa maneira, torna-se possível definir a distribuição Poisson como sendo uma distribuição pertencente à família exponencial, onde:

Quadro 6 – Modelo Poisson para a frequência – Fonte: Próprio autor.

Modelo Funcional: Família Exponencial	***Modelo Poisson: Frequência de Sinistros***
θ	$ln(\lambda)$
$b(\theta)$	e^{θ}
ϕ	*1*
$a(\phi)$	*1*
$c(y;\phi)$	$-ln(y!)$

Dessa forma, torna-se possível escrever analiticamente a média e a variância da variável aleatória Y como sendo:

$$E(Y) = \mu = b'(\theta) = e^{\theta} = \lambda \qquad (33)$$

$$Var(Y) = b''(\theta)a(\phi) = e^{\theta} = \lambda \qquad (34)$$

Aplicando a transformação logarítmica, tem-se que: $\eta_i = \ln(\mu_i) + \varepsilon_i = \ln\left(e^{\theta_i}\right) = \theta_i$, onde, $\eta = \theta$. Logo, conclui-se ser essa a função de ligação canônica do modelo *Poisson*.

A rigor, um procedimento ideal e mais elegante consistiria na aplicação de um teste de aderência à distribuição de probabilidade teórica sob interesse, para que então se pudesse testar a hipótese estatística de adequação do seu uso no contexto da modelagem GLM. Entretanto, tendo em vista o fato dos dados utilizados já serem fornecidos de maneira agrupada por célula tarifária, e não de maneira analítica com a exibição de detalhada de cada evento individualmente observado, esse procedimento operacional de análise não se tornara viável.

Através da aplicação da função *glm(.)* disponível no R, pode-se observar que o melhor ajuste obtido para a frequência dos sinistros foi o modelo *GLM Log-Poisson*, com distribuição de probabilidade para a variável resposta Poisson e função de ligação canônica logarítmica. Ao todo foram testados seis modelos: três com ligação canônica natural, e três com ligação logarítmica, para as distribuições Poisson, Binomial, e Binomial Negativa. Os resultados obtidos para o *best fit model* através do R encontram-se descritos no quadro 7.

Quadro 7 – Modelo Log-Poisson para a frequência – Fonte: Próprio autor.

Fator de Risco	***Nível***	***GL***	***Estimativa*** $(\widehat{\boldsymbol{\beta}})$	***Erro Padrão***	***z value***	$\boldsymbol{Pr(>\lvert z\rvert)}$
Intercepto	-	*1*	*-3,829639*	*0,074997*	*-51,064*	*< 2e-16**
Classe do Veículo	*1*	*0*	*0,000000*	*0,000000*	-	-
	2	*1*	*-0,252640*	*0,073777*	*-3,424*	*0,000616**
Idade do Veículo	*1*	*1*	*0,437661*	*0,093954*	*4,658*	*3,19e-06**
	2	*0*	*0,000000*	*0,000000*	-	-
Zona do Veículo	*1*	*1*	*1,959875*	*0,101451*	*19,319*	*< 2e-16**
	2	*1*	*1,428190*	*0,099375*	*14,372*	*< 2e-16**
	3	*1*	*0,802747*	*0,111493*	*7,200*	*6,02e-13**
	4	*0*	*0,000000*	*0,000000*	-	-
	5	*1*	*0,185408*	*0,414164*	*0,448*	*0,654393***
	6	*1*	*-0,231218*	*0,219860*	*-1,052*	*0,292957***
	7	*1*	*0,000554*	*0,581627*	*0,001*	*0,999240***

Definindo o teste de hipótese para a significância ou nulidade estatística dos parâmetros β_{ij} estimados para o modelo de frequência, tem-se, de uma forma geral, que $\forall_{ij} \in \{0,1,2,\dots,p\}$, deseja-se testar:

$$H_0: \beta_{ij} = 0$$

$$H_1: \beta_{ij} \neq 0$$

A um nível de significância $\alpha = 0{,}05$, e tomando por base o *p-value* associado à estatística de teste de Wald calculada, *observa-se uma forte evidência estatística de que se deve rejeitar a hipótese nula H_0 dos coeficientes β_{ij} associados ao intercepto, e às variáveis classe do veículo e idade do veículo, serem estatisticamente nulos. Por outro lado, com base também no *p-value* calculado, **não há evidência estatística suficiente para se rejeitar H_0 em relação a todos os β_{ij} associados à variável zona do veículo. Por essa razão, considerando o resultado acima exposto, faz-se necessário avaliar de maneira mais minuciosa a relevância e a contribuição estatística dessa variável tarifária para o poder de predição e inferência global do modelo em análise.

Para testar e avaliar o efeito global da variável tarifária zona do veículo, recorreu-se à utilização do Teste de Wald, implementado na *library (aod)* do R por meio da função *wald.test(.)*. O Teste de Wald é utilizado quando há o interesse de se testar a hipótese nula de significância ou de nulidade estatística para um subconjunto particular de coeficientes do vetor estimado de parâmetros. Assim, generalizando $\forall_{ij}$, com i e $j = \{0{,}1{,}2, \dots, p\}$, o Teste de Wald deseja testar se:

$$H_0: \beta_{i1} = 0;\ ou; \beta_{i2} = 0;\ ou;\ \beta_{i3} = 0; \dots; ou;\ \beta_{ij} = 0$$

$$H_1: \beta_{i1} \neq 0; ou;\ \beta_{i2} \neq 0; ou;\ \beta_{i3} \neq 0; \dots; ou;\ \beta_{ij} \neq 0$$

O Teste de Wald indicou, para os coeficientes associados à variável tarifária zona do veículo, os resultados descritos no Quadro 8.

Quadro 8 – Teste Qui-Quadrado de Wald para a zona do veículo – Fonte: Próprio autor.

Teste de Wald	$\boldsymbol{\chi^2_{calc}}$	***GL***	$\boldsymbol{Pr(\chi^2_{calc} > \chi^2_{crit})}$
Qui-Quadrado	*448,6*	*6*	*0,00**

A estatística de teste Qui-Quadrado no valor de 448,6, para 6 graus de liberdade, possui um *p-value* associado de aproximadamente 0,00, *indicando que o efeito global da variável zona do veículo é altamente significativo a um nível de significância $\alpha = 0{,}05$. Esse resultado indica uma forte evidência estatística de que se deve rejeitar a hipótese H_0 de nulidade conjunta dos parâmetros associados à variável zona do veículo, em favor da hipótese alternativa H_1 de que pelo menos um dos coeficientes é estatisticamente não nulo.

Ainda objetivando avaliar o efeito da variável zona do veículo sobre o ajuste global do modelo, decidiu-se recorrer ao teste da razão de verossimilhança, *Likelihood Ratio Test: LRT*, executado no R por meio da função *anova(.)*, e cujos dados de saída constam no 9.

Quadro 9 – Teste da razão de verossimilhança para a frequência – Fonte: Próprio autor.

Modelo	***GL***	***Deviance***	***GL Resíduos***	***Deviance Resid.***	***F***	$\boldsymbol{Pr(>F)}$
Nulo	-	-	*27*	*520,35*	-	-
Classe do Veículo	*1*	*2,75*	*26*	*517,60*	*2,7509*	*0,0972*
Idade do Veículo	*1*	*40,26*	*25*	*477,34*	*40,2592*	*2,224e-10*
Zona do Veículo	*6*	*447,27*	*19*	*30,08*	*74,5442*	*<2,2e-16*

O *deviance* residual, calculado pela diferença entre o *deviance* do modelo testado com a inclusão da variável zona do veículo e o *deviance* do modelo saturado, apontou que a variável referente à localização geográfica contribui significativamente para o ajuste global do modelo. Observa-se, portanto, que a inclusão dessa variável contribui para a diminuição do *deviance* residual tornando-o suficientemente pequeno. Logo, o modelo com três variáveis ajusta melhor os dados que o modelo mais simples com apenas duas variáveis. Além disso, o *p-value* obtido com a inclusão dessa variável, na ordem de *<2,2e-16*, reforça a evidência estatística de que se deve rejeitar a hipótese nula H_0 do modelo mais simples ajustar melhor os dados, em favor da hipótese alternativa H_1 de que o modelo com três fatores de risco possui um melhor ajuste global.

Outro critério para a avaliação do ajustamento de um MLG proposto por Bruin (2006) e também empregado em Souza e Leão (2012) é que, segundo os autores, se um MLG proporcionar um bom ajuste aos dados, espera-se que a razão entre o *deviance* residual e o número de graus de liberdade seja próxima de 1. Para o modelo analisado, observa-se que a razão calculada foi de aproximadamente 1,58, demonstrando que o modelo *Log-Poisson* se ajusta de forma satisfatória ao conjunto de dados analisado.

5.2 Modelagem da Severidade dos Sinistros

Os primeiros trabalhos abordando a modelagem de dados de sinistros mantiveram um foco inicial sobre a variável frequência, e não sobre a severidade, apesar de alguns autores já terem introduzido, à época, a modelagem GLM para variáveis contínuas, como observa-se, por exemplo, em Nelder e Wedderburn (1972), e Nelder e McCullagh (1983).

Já os primeiros estudos que se propuseram a abordar a modelagem da severidade sugeriram, a princípio, a utilização da distribuição Normal como variável resposta, ou seja,

buscaram introduzir um arcabouço estatístico com um parâmetro de dispersão constante para todos os fatores de tarifação, como se pode verificar em Baxter, Coutts e Ross (1980).

Sabe-se que distribuições empíricas de severidade detêm um acentuado nível de assimetria em suas curvas, e, por essa razão, não se mostram propícias a serem modeladas por uma distribuição Gaussiana. Por outro lado, distribuições alternativas podem ser empregadas na modelagem deste tipo de variável aleatória contínua, como, por exemplo, as distribuições Gamma, Log-Normal, Gaussiana Inversa, Pareto, Weibull, Beta, dentre outras. Todavia, nem todas estas distribuições pertencentes à família exponencial estão implementadas e acessíveis de forma direta para serem utilizadas como argumento da função *glm(.)* no R. Além disso, algumas distribuições como a Pareto, Beta, e Weibull, por exemplo, nem sempre se mostram adequadas à modelagem de sinistros de pequenos valores por possuírem uma cauda muito densa e carregada, enquanto que, por outro lado, distribuições como a Gamma, Log-Normal, e Normal Inversa, por possuírem uma cauda mais leve e menos densa, mostram-se, na grande maioria dos casos, mais adequadas à modelagem de sinistros de pequenos valores, como é o caso do *mopped insurance*.

A exemplo da abordagem realizada para a frequência, a utilidade da distribuição Gamma para o ajuste da severidade já fora anteriormente proposta e verificada por outros autores em modelagens atuariais envolvendo dados de sinistros, tais como se observa em Klugman, Panjer e Willmot (2002), Santos (2008), Jong e Heller (2008), Ohlsson e Johansson (2010), Ferreira (2010), e Souza e Leão (2012).

Especificando a variável aleatória do modelo para o ajuste da severidade, pode-se assumir, por hipótese, e com embasamento na literatura atuarial relativa à precificação de seguros, que Y segue uma distribuição Gamma, de sorte que:

$Y \sim Gamma(\alpha;\beta)$, com uma função densidade de probabilidade definida por:

$$f(y|\alpha;\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0;\ \alpha\ e\ \beta > 0; e\ \Gamma(\alpha) = \int_0^{\infty} \beta^{\alpha} y^{\alpha-1} e^{-\beta y} dy \qquad (35)$$

Assim, sendo válida a relação a seguir, torna-se possível escrever que:

$$E(Y) = \mu = \frac{\alpha}{\beta} \Leftrightarrow \beta = \frac{\alpha}{\mu}$$

$$f(y|\alpha;\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} = exp\left[\ln\left(\beta^{\alpha} y^{\alpha-1} e^{-\beta y}\right) - \ln\Gamma(\alpha)\right]$$

$$= exp\left[\alpha \ln\left(\frac{\alpha}{\mu}\right) + (\alpha - 1)\ln y - \frac{\alpha}{\mu} y - \ln\Gamma(\alpha)\right]$$

$$= exp\left[\alpha \ln(\alpha) - \alpha \ln\mu + \alpha \ln y - \ln y - \frac{\alpha}{\mu} y - \ln\Gamma(\alpha)\right]$$

$$= exp\left[\alpha\left(-\frac{y}{\mu} - \ln\mu\right) + \alpha\ln\alpha y - \ln y - \ln\Gamma(\alpha)\right]$$

$$= exp\left[\frac{-\frac{1}{\mu}y - \ln\mu}{\frac{1}{\alpha}} + \alpha\ln\alpha y - \ln y - \ln\Gamma(\alpha)\right]$$

$$= exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right]$$

$$= f(y|\theta;\phi)$$

Pode-se escrever a distribuição Gamma como sendo uma distribuição onde:

Quadro 10 – Modelo Gamma para a severidade – Fonte: Próprio autor.

Modelo Funcional: Família Exponencial	***Modelo Gamma: Severidade de Sinistros***
θ	$-\frac{1}{\mu}$
$b(\theta)$	$ln\,\mu$
ϕ	α
$a(\phi)$	$\frac{1}{\alpha}$
$c(y;\phi)$	$\alpha\,ln\,\alpha y - ln\,y - ln\,\Gamma(\alpha)$

Dessa forma, escrevendo os dois primeiros momentos, não centrado e centrado, respectivamente, para a variável aleatória Y, tem-se que:

$$E(Y) = \mu = b'(\theta) = b'\left(-\frac{1}{\theta}\right) = b'\left[\ln\left(-\frac{1}{\theta}\right)\right] = b'[-\ln(-\theta)] = -\frac{1}{\theta} = \frac{\alpha}{\beta} \qquad (36)$$

$$Var(Y) = b''(\theta)a(\phi) = -\frac{1}{\theta^2}\frac{1}{\alpha} = \frac{\mu^2}{\alpha} = \frac{1}{\alpha}\left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2} \qquad (37)$$

Da transformação logarítmica realizada na expressão (36), observa-se que $\eta_i = \ln(\mu_i) + \varepsilon_i = \ln\left(-\frac{1}{\theta_i}\right) = -\ln(-\theta)$, e assim, que $\eta = -\ln(-\theta)$.

Nesse caso, essa é a chamada função de ligação canônica para o modelo *Gamma*.

Com base no suporte teórico desenvolvido, e na teoria atuarial sobre precificação de seguros, observa-se, a partir dos resultados obtidos com a aplicação da função *glm(.)* no R, que o melhor ajuste obtido para a severidade dos sinistros foi o *GLM Log-Gaussiano Inverso* com distribuição de probabilidade para a variável resposta Gaussiana Inversa e função de ligação não canônica logarítmica. Foram testados ao todo quatro modelos para a severidade,

sendo dois com as ligações canônicas naturais, e dois com a ligação logarítmica, para as distribuições Gamma e Gaussiana Inversa respectivamente, tomando como referência as metodologias empregadas pelos autores já citados anteriormente. Os resultados relativos ao do *best fit model* obtido através do R constam no quadro 11.

Quadro 11 – Modelo Log-Gaussiano Inverso para a severidade – Fonte: Próprio autor.

Fator de Risco	***Nível***	***GL***	***Estimativa*** $(\widehat{\boldsymbol{\beta}})$	***Erro Padrão***	***z value***	$\boldsymbol{Pr(>\lvert z\rvert)}$
Intercepto	-	*1*	*8,85075*	*0,04654*	*190,177*	*< 2e-16**
Classe do Veículo	*1*	*0*	*0,00000*	*0,00000*	-	-
	2	*1*	*-0,58491*	*0,04720*	*-12,392*	*1,29e-09**
Idade do Veículo	*1*	*1*	*0,00000*	*0,06726*	*8,089*	*4,80e-07**
	2	*0*	*0,54405*	*0,00000*	-	-
Zona do Veículo	*1*	*1*	*0,19083*	*0,06302*	*3,028*	*0,00799**
	2	*1*	*0,07635*	*0,06006*	*1,271*	*0,22181***
	3	*1*	*0,05579*	*0,06830*	*0,817*	*0,42599***
	4	*0*	*0,00000*	*0,00000*	-	-
	5	*1*	*0,12829*	*0,23617*	*0,543*	*0,59446***
	6	*1*	*0,03402*	*0,13641*	*0,249*	*0,80622***
	7	*1*	*0,37604*	*0,43484*	*0,865*	*0,39994***

Definindo o teste de hipótese para a significância ou nulidade estatística dos parâmetros β_{ij} estimados para o modelo de severidade, tem-se de forma geral que $\forall_{ij}$, com i e $j = \{0,1,2, \dots, p\}$, deseja-se testar:

$$H_0: \beta_{ij} = 0$$

$$H_1: \beta_{ij} \neq 0$$

Adotando-se um nível de significância $\alpha = 0{,}05$, e tomando por base o *p-value* associado à estatística de Wald calculada, *observa-se uma forte evidência estatística em favor da hipótese alternativa H_1, para que se rejeite a hipótese nula H_0 de que os coeficientes β_{ij} estimados, associados ao intercepto e às variáveis classe e idade do veículo, sejam estatisticamente nulos.

Por outro lado, com base no *p-value*, **não há evidência estatística suficiente para se rejeitar H_0 em favor de H_1 em todos os β_{ij} associados à variável zona do veículo.

De maneira mais criteriosa, aplicou-se o Teste de Wald para testar a hipótese nula de significância estatística do subconjunto de coeficientes estimados para a variável zona do veículo. O teste busca, de uma forma geral, testar se $\forall_{ij}$, com i e $j = \{0,1,2,\dots,p\}$:

$$H_0: \beta_{i1} = 0;\ ou; \beta_{i2} = 0;\ ou;\ \beta_{i3} = 0; \dots; ou;\ \beta_{ij} = 0$$

$$H_1: \beta_{i1} \neq 0; ou;\ \beta_{i2} \neq 0; ou;\ \beta_{i3} \neq 0; \dots; ou;\ \beta_{ij} \neq 0$$

Os resultados do Teste de Wald para a severidade estão contidos no Quadro 12.

Quadro 12 – Teste Qui-Quadrado de Wald para a zona do veículo – Fonte: Próprio autor.

Teste de Wald	$\boldsymbol{\chi^2_{calc}}$	***GL***	$\boldsymbol{Pr(\chi^2_{calc} > \chi^2_{crit})}$
Qui-Quadrado	*10,1*	*6*	*0,12*

Observa-se que a estatística de teste Qui-Quadrado obtida no valor de 10,1, para 6 graus de liberdade, possui um *p-value* associado na ordem de 0,12, indicando que o efeito global da variável zona do veículo é pouco significativo ao nível de significância $\alpha = 0{,}05$. Logo, pelo resultado obtido, conclui-se que não se pode rejeitar a hipótese H_0 de nulidade conjunta dos coeficientes associados à variável, em detrimento da hipótese alternativa H_1.

Aplicando o teste da razão de verossimilhança no R para avaliar a influência da variável zona do veículo sobre o nível de ajuste global do modelo para a severidade, foram obtidos os dados de saída contidos no Quadro 13.

Quadro 13 – Teste da razão de verossimilhança para a severidade – Fonte: Próprio autor.

Modelo	***GL***	***Deviance***	***GL Residual***	***Deviance Resid.***	***F***	$\boldsymbol{Pr(>F)}$
Nulo	-	-	*24*	*0,0158876*	-	-
Classe do Veículo	*1*	*0,0075952*	*23*	*0,0082925*	*122,1628*	*6,707e-09*
Idade do Veículo	*1*	*0,0066504*	*22*	*0,0016420*	*106,9676*	*1,716e-08*
Zona do Veículo	*6*	*0,0006569*	*16*	*0,0009851*	*1,7611*	*0,171*

Observa-se que a inclusão da variável zona do veículo contribui para a diminuição do *deviance* residual. Todavia, em sentido contrário, o *p-value* obtido com a inclusão de tal variável no modelo, na ordem de 0,171, aponta uma evidência estatística de que não se deve rejeitar a hipótese nula H_0 do modelo com dois fatores de risco ajustar melhor os dados, em detrimento da hipótese alternativa H_1 do modelo com três variáveis oferecer um melhor ajuste global. Não obstante às conclusões estatísticas supramencionadas, faz-se preciso considerar a importância da variável zona do veículo para o modelo de tarifação, tendo esta apresentado um resultado satisfatório quanto ao modelo ajustado para a frequência dos sinistros.

Pelo critério proposto em Bruin (2006) e em Souza e Leão (2012), a razão entre o *deviance* residual e o número de graus de liberdade observado para o modelo de severidade foi de aproximadamente 0,00006156, implicando no fato de que a especificação da função de regressão estimada para o modelo *Log-Gaussiano Inverso*, segundo esse critério, não modela de forma satisfatória o conjunto de dados analisado. Entretanto, tal medida representa apenas um indicador para a avaliação da qualidade do ajuste, não possuindo credibilidade estatística suficiente para invalidar ou anular a utilidade do modelo como um todo. Além disso, há de se considerar ainda o fato dos dados encontrarem-se agrupados em células tarifárias, o que penaliza fortemente o cálculo deste indicador estatístico, pois o número de graus de liberdade acaba tornando desproporcional esta medida em função do *deviance* residual. Soma-se a isso ainda o fato de algumas das células tarifárias possuírem ausência de informações, como é o caso das células 5, 19 e 21, fato esse que eleva o número de graus de liberdade do modelo, aumentando o denominador da razão calculada, sem em contrapartida contribuir de maneira positiva para a capacidade preditiva do modelo estimado.

5.3 Modelagem do Prêmio

Estimar os coeficientes, as relatividades, e consequentemente calcular o prêmio do seguro, constituem, de maneira geral e pragmática, alguns dos principais objetivos do atuário no desenvolvimento de qualquer modelo tarifação. Para tanto, na prática, após modelar os dados e estimar os elementos componentes dos modelos de frequência e severidade, dado que a estrutura estatística aqui desenvolvida possui caráter multiplicativo, calcular as relatividades para o prêmio consistirá em nada mais do que efetuar o produtório das relatividades estimadas para o modelo de frequência com as relatividades estimadas para o modelo de severidade, objetivando assim combiná-las em um arcabouço multiplicativo. As relatividades exercem fundamental importância dentro da análise tarifária por permitirem mensurar o risco das demais classes de uma determinada variável tarifária em relação à classe base de referência, conforme descrito e exemplificado anteriormente no capítulo 3.

No Quadro 14 encontra-se a descrição detalhada dos coeficientes estimados para os MLGs ajustados, já inclusas as relatividades tarifárias associadas a cada um dos níveis de risco das variáveis consideradas, relatividades essas que expressam em qual direção e em que intensidade o prêmio estatístico deve ser agravado ou suavizado em função da razão de chance calculada, *odds ratio*, conforme o risco de ocorrência de sinistros.

Quadro 14 – Coeficientes e relatividades para os MLGs estimados – Fonte: Próprio autor.

Fator de Risco	***Nível***	***Frequência***		***Severidade***		***Prêmio de Risco***	
		$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{exp}\left(\widehat{\boldsymbol{\beta}}\right)$	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{exp}\left(\widehat{\boldsymbol{\beta}}\right)$	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{exp}\left(\widehat{\boldsymbol{\beta}}\right)$
Intercepto	-	*-3,829639*	*0,021717*	*8,85075*	*6979,627596*	*5,021111*	*151,579660*
Classe do Veículo	*1*	*0,000000*	*1,000000*	*0,00000*	*1,000000*	*0,00000*	*1,000000*
	2	*-0,252640*	*0,776747*	*-0,58491*	*0,557154*	*-0,837553*	*0,432767*
Idade do Veículo	*1*	*0,437661*	*1,549079*	*0,54405*	*1,722975*	*0,981713*	*2,669026*
	2	*0,000000*	*1,000000*	*0,00000*	*1,000000*	*0,00000*	*1,000000*
Zona do Veículo	*1*	*1,959875*	*7,098439*	*0,19083*	*1,210257*	*2,150708*	*8,590940*
	2	*1,428190*	*4,171144*	*0,07635*	*1,079345*	*1,504544*	*4,502104*
	3	*0,802747*	*2,231662*	*0,05579*	*1,057376*	*0,858537*	*2,359707*
	4	*0,000000*	*1,000000*	*0,00000*	*1,000000*	*0,00000*	*1,000000*
	5	*0,185408*	*1,203708*	*0,12829*	*1,136883*	*0,313698*	*1,368477*
	6	*-0,231218*	*0,793566*	*0,03402*	*1,034606*	*-0,197196*	*0,821029*
	7	*0,000554*	*1,000554*	*0,37604*	*1,456508*	*0,376596*	*1,457315*

As relatividades foram obtidas por meio da função inversa exponencial, tendo em vista a função de ligação utilizada no ajuste dos MLGs ter sido a logarítmica.

Assim, para um dado $\beta_{ij}, \forall_{ij}$ com i e $j = \{0,1,2,\dots,p\}$, calcular a relatividade tarifária consiste em aplicar a transformação exponencial expressa em (38), qual seja:

$$\gamma_{ij} = exp\left(\hat{\beta}_{ij}\right) = e^{\widehat{\beta}_{ij}} \qquad (38)$$

Alguns importantes *insights* podem ser obtidos ao se modelar separadamente a frequência e a severidade com vista à obtenção da tarifa a ser cobrada por um seguro. Ohlsson e Johansson (2010) afirmam que a modelagem tarifária GLM padrão consiste na realização de análises iniciais distintas para a frequência e severidade, para somente então serem obtidas as relatividades do modelo para prêmio por meio da multiplicação dos resultados obtidos a *priori*. Conforme os autores, a justificativa para a análise em separado de dois MLGs apoia-se no fato de que:

i. A frequência dos sinistros é normalmente muito mais estável do que a severidade, e geralmente, grande parte do poder preditivo dos risk factors está relacionado à frequência, podendo-se dessa forma estimar o risco com mais precisão;

ii. Uma análise disjunta fornece mais detalhes e informações sobre como, e em que direção, os níveis tarifários afetam o prêmio, suavizando-o, ou agravando-o.

Uma discussão detalhada e mais aprofundada acerca da modelagem disjunta da frequência e da severidade, onde são expostos de maneira argumentativa os benefícios obtidos com a combinação de dois modelos em detrimento da opção de se aplicar diretamente um modelo para o cálculo do prêmio, pode ser encontrada em Brockman e Wright (1992) e em Murphy, Brockman e Lee (2000). Santos (2008) e Bandeira (2013) também discutem alguns aspectos relevantes no tocante a essa mesma questão.

Através dos coeficientes estatísticos estimados para os MLGs foi possível obter os coeficientes componentes do modelo para o prêmio. Por meio destes, tornou-se possível obter uma tarifa atuarialmente calculada para toda e qualquer combinação linear possível entre um vetor paramétrico de estimativas $\widehat{\beta}_{ij}$ e as respectivas variáveis *dummies* componentes do perfil de risco de cada segurado. Aplicando a formulação funcional fundamental da modelagem GLM, torna-se possível calcular o prêmio para um segurado qualquer, dado o seu perfil de risco individual, de forma que:

Se $N \sim Poisson(\lambda)$, e $X \sim Gaussiana\ Inversa(\alpha;\beta)$, da relação (5), tem-se que, sem perda de generalidade, o prêmio para uma determinada apólice i pode ser calculado por:

$$P_{R_i} = E[S_i] = E[N_i].E[X_i] \qquad (39)$$

Aplicando a transformação com a função de ligação logarítmica, tem-se que:

$$\ln\left(P_{R_i}\right) = \ln[E(S_i)] = \hat{\eta}_i = \widehat{\beta}_0 + \sum_{i=1;j=1}^{p} \widehat{\beta}_{ij} D_{ij} \qquad (40)$$

$$\ln\left(P_{R_i}\right) = \ln[E(S_i)] = \widehat{\beta}_0 + \widehat{\beta}_{11}X_{11} + \widehat{\beta}_{12}X_{12} + \widehat{\beta}_{21}X_{21} + \cdots + \widehat{\beta}_{37}X_{37} \qquad (41)$$

Onde $\widehat{\beta}_0$ representa o intercepto, os $\widehat{\beta}_{ij}$ representam os parâmetros estimados para o MLG observáveis no Quadro 14, e os D_{ij} representam as variáveis do tipo *dummy* que calibram o modelo. Aplicando uma transformação por meio da função inversa exponencial à expressão (41), o valor esperado do prêmio fica expresso por:

$$e^{\ln\left(P_{R_i}\right)} = e^{\ln[E(S_i)]} = e^{\left(\widehat{\beta}_0 + \widehat{\beta}_{11}X_{11} + \widehat{\beta}_{12}X_{12} + \widehat{\beta}_{21}X_{21} + \cdots + \widehat{\beta}_{37}X_{37}\right)} = \hat{\mu}_i \qquad (42)$$

Ademais, não há necessidade de ser aplicada qualquer transformação adicional ao prêmio de risco, pois a tarifa já é dada na base anual, sendo o período de exposição, ou duração da exposição ao risco, também expresso na base anual.

Outro fator passível de inclusão no modelo é o efeito inflacionário. Um método simples e pragmático consiste em fracionar o prêmio anual em $1/12$ avos mensal, embutir em cada uma das frações mensais uma taxa projetada de inflação, e descontar os respectivos fracionamentos a valor presente na data focal do cálculo.

Oportunamente, pode-se ainda optar pela inclusão de taxas de carregamento como θ e α para fazer frente às contingências estatísticas decorrentes das flutuações estocásticas do risco, e às despesas administrativas e de comercialização, respectivamente.

5.4 Ajuste, Sensibilidade e Interpretação dos Parâmetros

Tendo em vista o fato da função de ligação dos modelos selecionados ter sido a logarítmica, a interpretação dos parâmetros torna-se mais simples através da exponencial das estimativas obtidas. Observa-se, quanto aos coeficientes estimados para a frequência, que, por exemplo, o número médio esperado de sinistros por apólice é menor para veículos do nível tarifário 2 da variável *classe do veículo*, quando comparados com veículos do nível 1 dessa mesma variável, ao passo que o número médio esperado de sinistros é maior para veículos do nível tarifário 1 da variável *idade do veículo*, quando comparados com veículos do nível 2 da mesma variável tarifária. Já o número médio de sinistros esperado para os veículos do nível 6 da variável *zona do veículo* é menor do que o esperado para os veículos do nível de risco base dessa mesma variável, nível 4, enquanto que em sentido contrário, o número médio esperado de sinistros para os veículos dos níveis 1, 2, 3, 5, e 7, é maior quando comparado aos veículos do nível de risco base da mesma variável tarifária.

Já em relação ao MLG ajustado para a severidade, verificam-se basicamente as mesmas evidências e conclusões observadas para a frequência, no sentido do agravamento ou da suavização da severidade média esperada, considerando-se tão somente as modificações numéricas de intensidade relativas à calibragem dos parâmetros. Observa-se nesse sentido que a severidade média esperada para veículos do nível 2 da variável *classe do veículo* é menor que a esperada para os veículos do nível 1, ao passo que a severidade média para veículos do nível 1 da variável tarifária *idade do veículo* é maior que a esperada para os veículos do nível 2 da mesma variável. Além disso, tem-se ainda que a severidade média esperada para veículos dos níveis de risco 1, 2, 3, 5, 6, e 7, da variável *zona do veículo*, é maior do que a esperada para o nível de risco base dessa mesma variável tarifária, qual seja, o nível 4.

Nesse sentido, as mesmas interpretações realizadas quanto ao agravamento ou suavização do número médio de sinistros em relação ao modelo de frequência podem ser aplicadas ao modelo de cálculo do prêmio, tendo em vista que todos os sinais dos coeficientes $\hat{\beta}_{ij}$ estimados para a frequência coincidem igualmente com os dos parâmetros $\hat{\beta}_{ij}$ estimados para o modelo combinado de cálculo do prêmio de risco.

Outra interpretação bastante prática, intuitiva, e informativa para a leitura dos MLGs ajustados pode ser obtida através da análise da *odds ratio*, ou, razão de chances. Essa medida estatística visa indicar a chance ou o efeito marginal do risco observado em relação à variável dependente quando da ocorrência de variações ou alterações no comportamento das realizações de uma das variáveis independentes, *ceteris paribus*, ou seja, mantendo-se todas as demais variáveis independentes inalteradas. Em relação aos modelos aqui ajustados, a *odds ratio* encontra sua representação nas próprias relatividades tarifárias associadas a cada um dos níveis de risco das variáveis, conforme já descrito anteriormente no Quadro 14.

Observa-se, por exemplo, em relação ao modelo estimado para a frequência, que a *odds ratio* das apólices do nível de risco 2, em relação às do nível de risco base 1, da variável *classe do veículo*, é de 0,776747. Dessa forma, estima-se que o número médio de sinistros a ser observado para o nível 2 seja, aproximadamente, 0,77 vezes o número observado para o nível 1, ou, que o número médio de sinistros observado para o nível 2 seja, aproximadamente, 23,33% inferior ao número observado para o nível 1. Já a *odds ratio* das apólices do nível de risco 1, em relação às do nível de risco base 2, da variável *idade do veículo*, é de 1,549079. Assim, estima-se que o número médio de sinistros a ser observado para o nível 1 seja, aproximadamente, 1,54 vezes o número observado para o nível 2, ou, que o número médio de sinistros observado para o nível 1 seja, aproximadamente, 54,90% superior ao observado para o nível 2. Analogamente, as mesmas implicações acerca da *odds ratio* podem ser estendidas à variável *zona do veículo*.

Para o modelo estimado de severidade, a *odds ratio* das apólices do nível de risco 2, em relação às do nível de risco base 1, da variável *classe do veículo*, é de 0,557154. Dessa forma, estima-se que a severidade média dos sinistros a ser observada para o nível 2 seja, aproximadamente, 0,55 vezes a severidade observada para o nível 1, ou, que a severidade dos sinistros observada para o nível 2 seja, aproximadamente, 44,29% inferior à severidade média observada para o nível 1. Já a *odds ratio* das apólices do nível de risco 1, em relação às do nível de risco base 2, da variável *idade do veículo*, é de 1,722975. Assim, estima-se que a severidade média dos sinistros a ser observada para o nível 1 seja, aproximadamente, 1,72 vezes a severidade observada para o nível 2, ou, que a severidade dos sinistros observada para o nível 1 seja, aproximadamente, 72,29% superior à severidade média observada para o nível 2. De forma análoga, as mesmas conclusões acerca da *odds ratio* podem ser estendidas à variável *zona do veículo*.

Observa-se ainda, em relação ao modelo estimado para o prêmio, que a *odds ratio* para as apólices do nível de risco 2, em relação às do nível de risco base 1, da variável *classe do veículo*, é de 0,432767. Isso implica que a tarifa do seguro a ser pago pelos segurados do nível 2, da referida variável, será equivalente a, aproximadamente, 0,43 vezes o prêmio pago pelos segurados do nível 1, ou, que o prêmio pago pelos segurados do nível 2 sofrerá uma redução de, aproximadamente, 56,73% em relação ao prêmio pago pelos segurados do nível 1. Já a *odds ratio* do nível de risco 1, em relação à do nível de risco base 2, da variável *idade do veículo*, é de 2,669026. Isso implica que o prêmio a ser pago pelos segurados do nível 1, da referida variável, será equivalente a, aproximadamente, 2,66 vezes o prêmio pago pelos segurados do nível 2, ou ainda, que o prêmio pago pelos segurados do nível 1 será majorado em, aproximadamente, 166,90% quando comparado ao prêmio pago pelos segurados do nível 2. Por fim, as *odds ratios* para as apólices dos níveis de risco 1, 2, 3, 5, e 7, em relação às do nível de risco base 4, da variável *zona do veículo*, são de 8,590940; 4,502104; 2,359707; 1,368477; e 1,457315, respectivamente. Isso implica no fato de que os segurados pertencentes aos referidos níveis tarifários deverão pagar um prêmio de, aproximadamente, 8,59; 4,50; 2,35; 1,36; e 1,45; vezes o prêmio pago pelos segurados do nível de risco 4, ou ainda, que o prêmio pago por tais segurados será majorado em, aproximadamente, 759,09%; 350,21%; 135,97%; 136,84%; e 145,73%, respectivamente. Já a *odds ratio* do nível de risco 6, em relação à do nível de risco base 4, é de 0,821029. Isso implica que o prêmio a ser pago pelos segurados pertencentes ao nível 6 deve ser igual a 0,82 vezes o valor pago pelos segurados do nível 4, ou ainda, que o prêmio pago pelos segurados do nível 6 sofrerá uma redução aproximada de 17,90% em relação ao prêmio pago pelos segurados do nível 4.

Além das interpretações realizadas para os coeficientes e relatividades tarifárias estimadas para os MLGs, torna-se ainda possível calcular o prêmio estatístico do seguro, para um segurado qualquer, dado seu perfil caracterizador de risco individual.

Quadro 15 – Perfil de risco individual de um segurado hipotético – Fonte: Próprio autor.

Fator de Risco	***Nível***	***Parâmetro Estimado para o Modelo do Prêmio***	
		$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{exp}\left(\widehat{\boldsymbol{\beta}}\right)$
Intercepto	-	*5,021111*	*151,579660*
Classe do Veículo	*2*	*-0,837553*	*0,432767*
Idade do Veículo	*1*	*0,981713*	*2,669026*
Zona do Veículo	*7*	*0,376596*	*1,457315*

Dessa forma, o cálculo do prêmio para um segurado hipotético com perfil de risco caraterizado pelo quadro 15 ficaria, conforme o MLG estimado e as formulações 39, 40, 41, e 42, especificado da seguinte maneira, a saber:

$$P_{R_i} = E[S_i] = e^{\ln\left(P_{R_i}\right)} = e^{\ln[E(S_i)]} = e^{(\widehat{\beta}_0+\widehat{\beta}_{11}X_{11}+\widehat{\beta}_{12}X_{12}+\widehat{\beta}_{21}X_{21}+\cdots+\widehat{\beta}_{37}X_{37})} = \hat{\mu}_i$$

$$P_{R_i} = e^{(\widehat{\beta}_0+\widehat{\beta}_{12}X_{12}+\widehat{\beta}_{21}X_{21}+\widehat{\beta}_{37}X_{37})} = \hat{\mu}_i$$

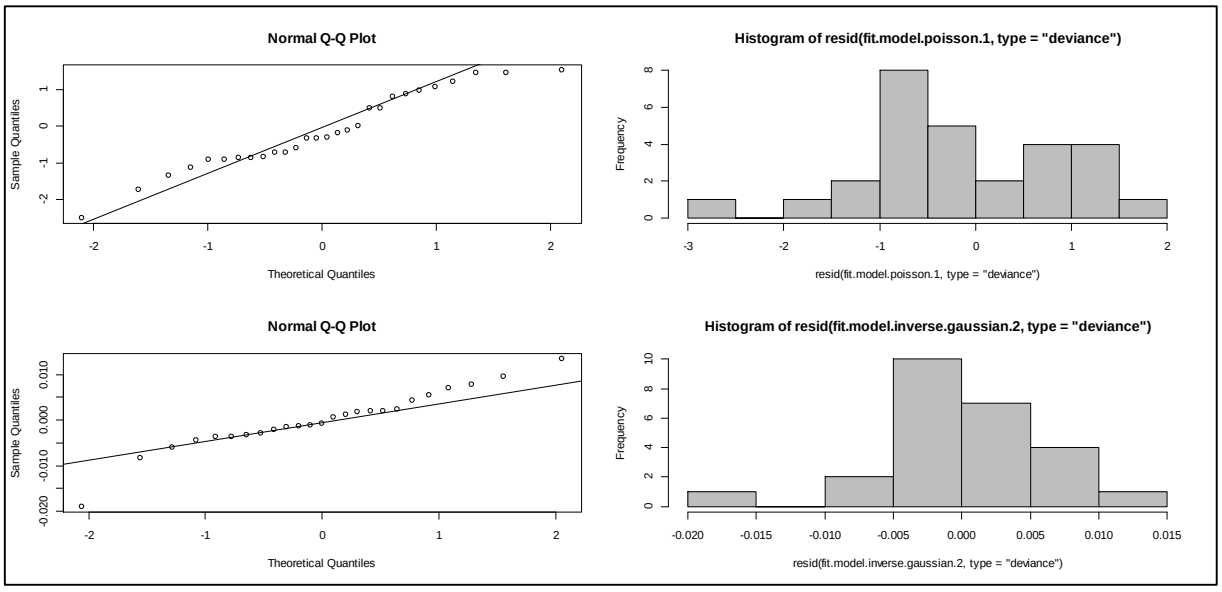
$$P_{R_i} = e^{(5{,}021111-0{,}837553+0{,}981713+0{,}376596)} = e^{(5{,}541867)} = 255{,}15$$

Logo, um segurado detentor de um veículo com peso superior a 60 Kg e com mais de duas marchas, que possua dois anos ou mais de fabricação, e que encontre sua localização geográfica na ilha de Gotland na Suécia, deverá ter seu prêmio médio estatístico calculado em KR 255,15, ou seja, 255,15 coroas suecas.

5.5 Resíduos e Diagnóstico

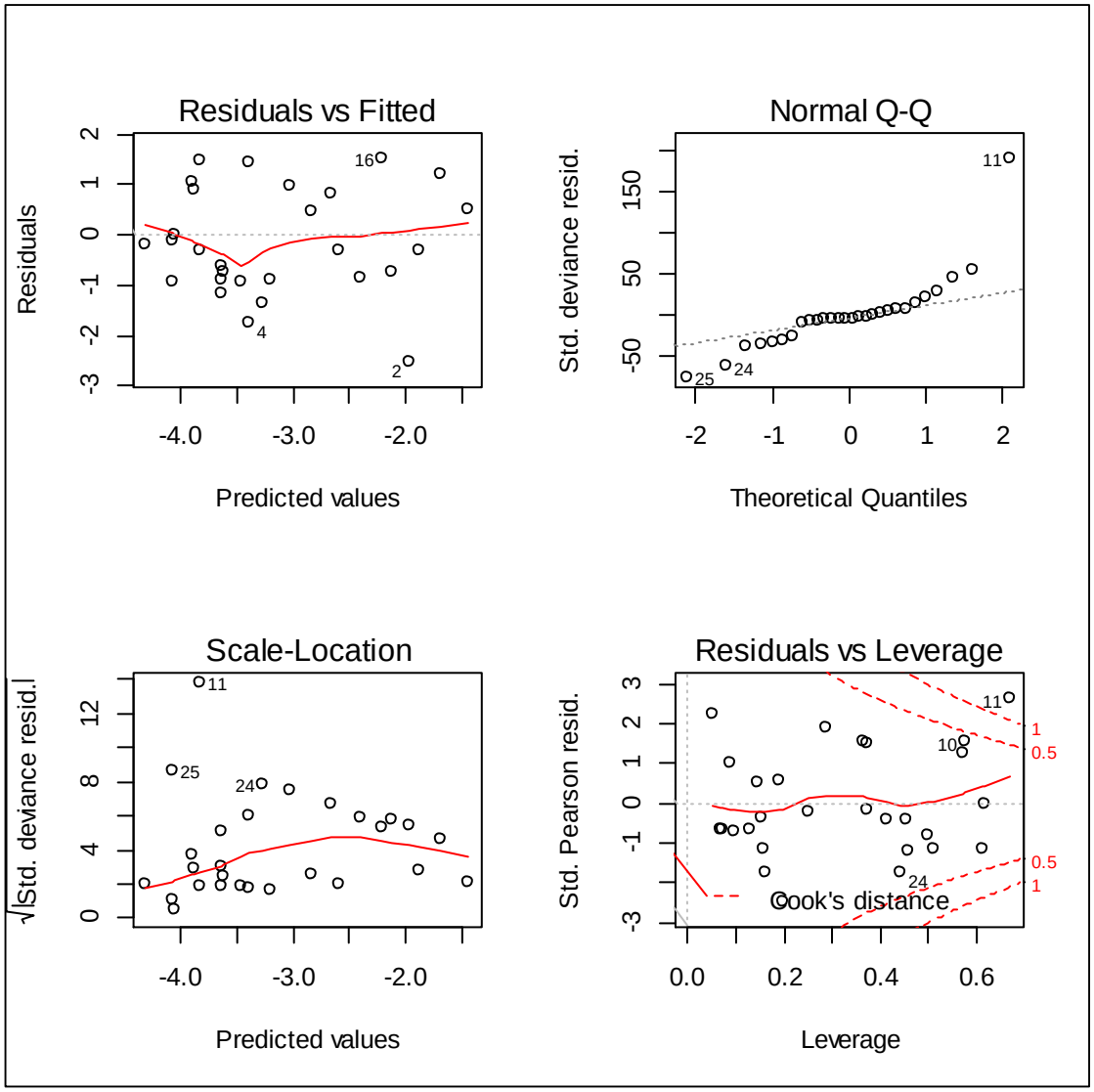
Conforme Dobson (2001), Jong e Heller (2008), e Ohlsson e Johansson (2010), o critério do *deviance* representa uma medida de seleção para a escolha do melhor ajuste GLM dentre um conjunto de modelos testados. Seguindo esse entendimento, adotou-se este como sendo o critério de decisão para a avaliação e seleção do *best fit model*, tanto para o modelo de frequência, quanto para o de severidade. Por meio do *deviance* residual é possível selecionar o melhor ajuste para os dados, de forma a minimizar os resíduos e maximizar a verossimilhança em relação ao modelo saturado.

Figura 1 – Q-Q plot e histograma dos resíduos *deviance* – Fonte: R Core Team (2014).



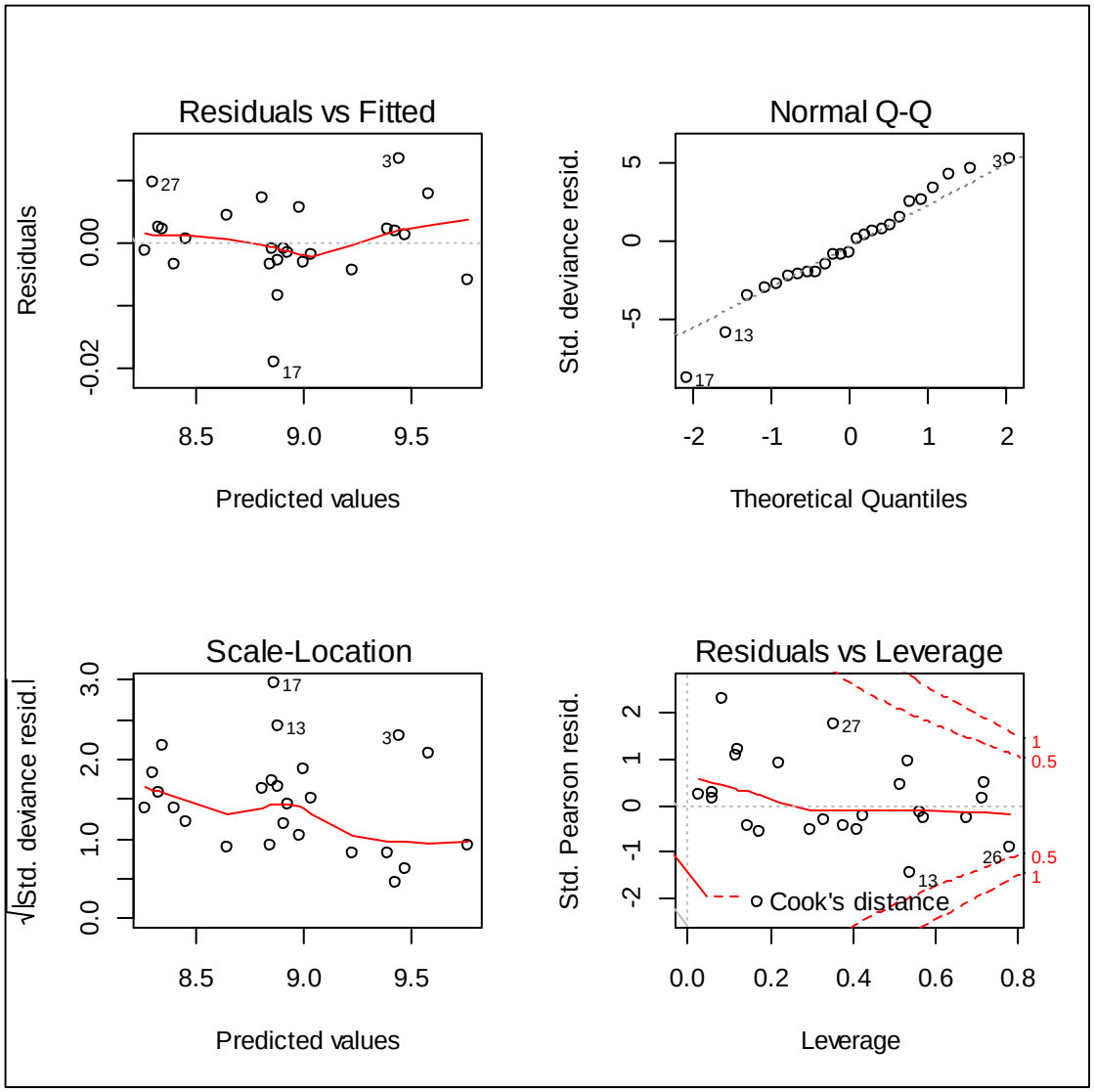
Por meio de uma análise gráfica da Figura 1, percebe-se, quanto aos Q-Q Plots obtidos, tanto para a frequência como para a severidade, que os resíduos do tipo *deviance* não se encontram alinhados de maneira satisfatória aos quantis teóricos da distribuição normal padrão, apresentando um acentuado nível de desalinhamento com muitos pontos de dispersão ao longo da reta normalizada. Já em relação aos histogramas residuais, percebe-se um elevado grau de assimetria na distribuição empírica dos resíduos do tipo *deviance*, o que contribui fortemente para o afastamento da suposição de normalidade. Tal evidência mostra-se bastante útil à rejeição da hipótese de normalidade dos dados, ao passo que sustenta razoavelmente a escolha de uma distribuição de probabilidade da família exponencial, que não seja a normal, no caso a Poisson e a Gaussiana Inversa.

Figura 2 – Diagnóstico para a frequência dos sinistros – Fonte: R Core Team (2014).



Observam-se, na Figura 2, quatro gráficos de diagnóstico para o modelo estimado de frequência dos sinistros. No gráfico superior esquerdo percebe-se a presença de três pontos destacados como sendo de alavanca, sendo estes referentes às células tarifárias #2, #4, e #16. Já o gráfico superior direito aponta três dos pontos como sendo influentes sobre as estimativas paramétricas do modelo, sendo estes relativos às células tarifárias #11, #24 e #25. O gráfico inferior esquerdo indica que os mesmos pontos apontados como influentes, #11, #24 e #25, também são pontos aberrantes. Finalmente, o gráfico inferior direito aponta para o fato de que a escolha da função de ligação logarítmica se mostra satisfatória e adequada para o modelo.

Figura 3 – Diagnóstico para a severidade dos sinistros – Fonte: R Core Team (2014).



Na Figura 3 constam outros quatro gráficos de diagnóstico para o modelo ajustado de severidade. O gráfico superior esquerdo indica a presença de três pontos de alavanca, sendo estes referentes às células tarifárias #3, #17, e #27. Ademais, o gráfico superior direito

aponta três pontos de influência sobre as estimativas dos coeficientes do modelo, sendo estes relativos às células tarifárias #3, #13 e #17. O gráfico inferior esquerdo indica que os mesmos pontos de influência apontados, #3, #13 e #17, são também pontos aberrantes. Por fim, tem-se que o gráfico inferior direito sugere a ideia de que a escolha da função de ligação logarítmica para o modelo Log-Gaussiano Inverso se mostra satisfatória e adequada.

6 CONSIDERAÇÕES FINAIS

Os Modelos Lineares Generalizados representam uma ampla classe de métodos estatísticos de regressão com diversas aplicabilidades em diversas e variadas áreas da ciência e do conhecimento, tais como, Atuária, Economia, Finanças, Biologia, e etc. Essa rica classe de modelos de regressão que fora introduzida inicialmente pelos atuários britânicos Nelder e Wedderburn, em 1972, condensa uma ampla variedade de modelos de regressão linear que tem em comum uma característica extremamente relevante e interessante, qual seja, o fato da distribuição de probabilidade da variável resposta do modelo poder ser escolhida da família exponencial.

Por essa razão, a citada classe de modelos tem despertado grande interesse prático em pesquisadores de diversas áreas, dentre elas a das Ciências Atuariais. Sob tal perspectiva, recorreu-se aqui ao uso dos MLGs visando a construção de um modelo de tarifação para um seguro do tipo *moped insurance* onde se buscou desenvolver um processo de modelagem sobre um conjunto de dados disponibilizado em Ohlsson e Johansson (2010). Para o processo de análise estatística e implementação da modelagem GLM construiu-se um projeto estatístico no software de código aberto R Core Team (2014), a fim de executar a análise do ajuste dos modelos ajustados, verificando-se os critérios relativos à qualidade global dos modelos, como por exemplo, através do critério *deviance*, conforme proposto e indicado em Dobson (2001), Jong e Heller (2008), Ohlsson e Johansson (2010), e Sousa (2010). Adicionalmente, foi possível testar hipóteses individuais e encaixadas a respeito dos parâmetros estimados para os modelos de frequência e de severidade dos sinistros, analisar a influência e a contribuição dos fatores tarifários para os modelos analisados, identificar e interpretar a influência e o impacto dos parâmetros dos modelos sobre as variáveis aleatórias dependentes, e etc.

Além disso, observou-se que, dentre as distribuições de probabilidade testadas no ajuste dos modelos, as que proporcionaram um melhor *fitting*, ou o *Goodness of fit*, aos dados foram os MLGs com distribuição de probabilidade Poisson e função de ligação logarítmica para a frequência, modelo Log-Poisson, e distribuição de probabilidade Gaussiana Inversa e função de ligação logarítmica para a severidade, modelo Log-Gaussiano Inverso. Em relação ao modelo estimado para a frequência, observou-se que o resultado empírico obtido para o melhor ajuste encontra-se alinhado aos resultados anteriormente encontrados por McCullagh e Nelder (1989), Santos (2008), e Ohlsson e Johansson (2010). Já em relação à severidade duas distribuições concorreram de forma bastante acirrada ao posto de melhor ajuste, sendo estas a

Gamma e a Gaussiana Inversa. Todavia, tomando como base o critério de decisão do *deviance residual*, além de uma série avaliações adicionais, chegou-se à conclusão de que a distribuição Gaussiana Inversa era a que melhor se ajustava os dados analisados, contrariando, portanto, alguns resultados anteriores obtidos em favor da distribuição Gamma, como, por exemplo, os verificados em McCullagh e Nelder (1989), Aitkin, Anderson, Francis, e Hinde (1989), Santos (2008), Ohlsson e Johansson (2010) e, Souza e Leão (2012).

Através da análise de sensibilidade desenvolvida para os modelos estimados das regressões, Poisson e Gaussiana Inversa, tornou-se possível proceder à interpretação das estimativas obtidas para os parâmetros de forma a compreender o efeito destes sobre o comportamento médio esperado das variáveis aleatórias referentes à frequência, à severidade, e ao prêmio de risco, analisando-se em que sentido e intensidade tais coeficientes indicavam o agravamento ou a suavização da tarifa atuarialmente calculada, e de que maneira as variáveis tarifárias consideradas e seus respectivos níveis de risco exercem influência na diferenciação do valor do prêmio a ser pago por cada segurado com base em seu perfil de risco individual.

A análise realizada revelou, de maneira geral, que as variáveis, *classe do veículo* e *idade do veículo*, impactam os valores esperados para a frequência, a severidade e o prêmio na mesma direção. Assim, por meio das *odds ratios* calculadas, chegou-se à conclusão de que o prêmio de risco médio deve ser menor para as apólices do nível de risco 2 em relação às do nível base 1, da variável *classe do veículo*, ao passo que o prêmio de risco deve ser maior para as apólices do nível de risco 1 em relação às do nível base 2, da variável *idade do veículo*. Por outro lado, foi possível concluir que o prêmio de risco médio deve ser menor para as apólices do nível de risco 6, em relação às do nível base 4, da variável *zona do veículo*, ao passo que em sentido contrário, o prêmio de risco deve ser maior para as apólices dos níveis de risco 1, 2, 3, 5, e 7, em relação às do nível base 4.

Finalmente, foi possível perceber ainda a vantagem de se combinar dois modelos de regressão, um para a frequência e outro para a severidade, em detrimento da possibilidade da aplicação direta de um único modelo para o cálculo do prêmio puro, sendo discutidas as vantagens e os benefícios de se combinar dois MLGs no contexto estratégico de modelagem e definição de uma estrutura tarifária padrão para o cálculo do prêmio de um seguro.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AITKIN, M.; ANDERSON, D.; FRANCIS, B.; HINDE, J. ***Statistical Modelling in GLIM***. 1. ed. Oxford: Oxford University Press, 1989.

BANDEIRA, M. C. O. R. M. ***Seguro De Saúde: Custos De Ambulatório - Modelização Linear Generalizada***. 2013. 77 f. Dissertação de Mestrado em Estatística, Faculdade de Ciências, Universidade de Lisboa. Lisboa, 2013.

BAXTER, L. A.; COUTTS, S. M.; ROSS, G. A. F. ***Applications of linear models in motor insurance***, Proceedings of the 21st International Congress of Actuaries, Zurich p. 11-29, 1980.

BROCKMAN, M. J.; WRIGHT, T.S. ***Statistical Motor Rating: Making Effetive Use of Your Data***. Journal of the Institute of Actuaries. 119, 457–543, 1992.

CHAPADOS, N., *et alli*. ***Estimating Car Insurance Premium: a Case Study in High-Dimensional Data Inference***. Technical Report 1199, Départament D'informatique et Recherché Opérationnelle, Université de Montreal. Montreal, 2001.

DOBSON, A. J. ***An Introduction To Generalized Linear Models***. 1. ed. London: Chapman and Hall, 1990.

DUGAS, C. *et alli*. ***Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking***. Casualty Actuarial Society Forum. Virginia, v. winter, p. 179-213, 2003

FERREIRA, P. P. ***Modelos de Precificação e Ruína para Seguros de Curto Prazo***. 2. ed. Rio de Janeiro: FUNENSEG, 2010.

FRANCIS, L. A. ***Neural Networks Demystified***. Casualty Actuarial Society Forum. Virginia, v. winter, p. 253-320, 2001.

GIL, A. C. ***Como Elaborar Projetos de Pesquisa***. 4ª Edição. Editora Atlas. São Paulo, 2002.

HADIDI, N. ***Classification Ratemaking Using Decision Trees***. Casualty Actuarial Society Forum. Virginia, v. winter, p. 253-283, 2003.

JONG, P.; HELLER, G. Z. ***Generalized Linear Models for Insurance Data***. 1. ed. New York: Cambridge, 2008.

KLUGMAN, S. A., PANJER, H. H., WILLMOT, G. E. ***Loss Models from Data to Decisions***. 2. ed. New Jersey: John Wiley & Sons, 2004.

McCULLAGH, P.; NELDER, J.A. ***Generalized Linear Models***. Mathematical Statistics of Generalized Linear Models. London: Chapman and Hall, 1989.

MORGADO, W. L. ***Método de Classificação de Risco Aplicado ao Seguro Automóvel***. 2004. 105 f. Dissertação de Mestrado em Engenharia Elétrica, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, 2004.

MURPHY, K.P.; BROCKMAN, M.J.; LEE, P.K.W. ***Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance***. Casualty Actuarial Society Forum, Virginia, v. Winter, p. 107, 2000.

NELDER, J.A.; WEDDERBURN, R.W.M. ***Generalized Linear Models***. Journal of the Royal Statistical Society. Series A (General). Vol. 135, No. 3 (1972), p. 370-384. Wiley, 1972.

OHLSSON, E.; JOHANSSON, B. ***Non-Life Insurance Pricing with Generalized Linear Models***. 1. ed. Estocolmo: Springer, 2010.

SANTOS, S. T. ***Construção de Uma Tarifa de Responsabilidade Civil Automóvel***. 2008. 91 f. Dissertação de Mestrado em Matemática e Aplicações: Actuariado, Estatística e Investigação Operacional, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa. Lisboa, 2008.

SOUZA, H. S. E.; LEÃO, L. C. S. ***Tarifação de Um Plano de Saúde Autogestão Aplicando os Modelos Lineares Generalizados***. Cadernos do IME, Universidade do Estado do Rio de Janeiro. Rio de Janeiro, v. 33, p. 01-17, 2012.

AGUIRRE, L. L.; ***Modelos de Precificação: uma Aplicação no Setor Imobiliário do DF***. Brasília, p. 01-21, 2012.

SOUSA, K. M. M. ***Modelos Lineares Generalizados e Modelos de Dispersão Aplicados à Modelagem de Sinistros Agrícolas***. 2010. 66 f. Dissertação de Mestrado em Ciências, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo. Piracicaba, 2010.

TURKMAN, M. A. A.; SILVA, G. L. ***Modelos Lineares Generalizados: da teoria à prática***. Lisboa: Universidade Técnica de Lisboa, 2000.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.URL:<www.R-project.org>.

SPRINGER. Disponível em: <www.math.su.se/GLMbook>. Acesso em: 12/09/2014.

SPRINGER. Disponível em: <www.people.su.se/~esbj/GLMbook/moppe.sas>. Acesso em: 12/09/2014.

Eidgenössische Technische Hochschule Zürich: Swiss Federal Institute of Technology Zürich. Disponível em: <www.stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html>. Acesso em: 10/10/2014.

Institute for Digital Research and Education: IDRE: UCLA. Disponível em: <www.ats.ucla.edu/stat/r/dae/nbreg.htm>. Acesso em: 12/10/2014.

Universidade Federal de Pernambuco: UFPE: Departamento de Estatística: DE. Disponível em: <www.de.ufpe.br/~cysneiros/disciplina/MES940/aulaMLGmestrado.pdf>. Acesso em: 05/10/2014.

Universidade Federal do Paraná: UFPR. Disponível em: <http://people.ufpr.br/~taconeli/CE225/sinistros.R>. Acesso em: 17/10/2014.

Institute for Digital Research and Education: IDRE: UCLA. Disponível em: <http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm>. Acesso em: 29/10/2014.