VIII SIMPÓSIO DE ATUÁRIA
10 A 12 DE ABRIL DE 2019
FORTALEZA - CE

MODELOS LINEARES GENERALIZADOS (MLG´s) E SUA APLICAÇÃO EM CIÊNCIAS ATUARIAIS

Área Temática: Aplicações em Atuária

Jaime Phasquinel Lopes Cavalcante
Universidade Federal do Ceará (UFC)
jaimephasquinell@alu.ufc.br

Luciana Moura Reinaldo
Universidade Federal do Ceará (UFC)
lucianareinaldo@ufc.br

Resumo:

A utilização de métodos estatísticos na rotina da Ciência Atuarial tem desenvolvido, historicamente, um papel central, tanto nos assuntos teóricos quanto práticos. Nesse sentido, corroborando com estudos outrora publicados, o presente estudo possui como objetivo principal demonstrar a aplicação da metodologia dos Modelos Lineares Generalizados com foco em uma problemática atuarial. A justificativa para o estudo surge do fato de que há uma vasta área de aplicações para o MLG, mas pouco exploradas, especialmente no Brasil. Diante do exposto, considerou-se a base contida em Kaas et al. (2008), que reflete a experiência anual de um portfólio de seguros de automóveis. Com isso, buscou-se relacionar a frequência de sinistros aos fatores de risco: sexo, região, tipo do carro, situação laboral. Portanto, ajustou-se um MLG com relação média-variância do tipo Poisson com função de ligação canônica. Ademais, são apresentados, como resultados, a identificação de que os fatores sexo e situação labora não são significantes para o modelo, os diagnósticos do modelo que confirmam o bom ajuste da modelagem e a análise dos desvios.

Palavras-chave: Modelos Lineares Generalizados. Ciências Atuariais.

1. INTRODUÇÃO

A gênese da ciência atuarial está centrada na gestão financeira dos sistemas destinados à redução dos impactos financeiros provenientes de eventos aleatórios que impedem o cumprimento do fluxo razoável da atividade (BOWERS et al., 1986). Contudo, tais sistemas possuem certas limitações em suas operações, pode-se citar as restrições para a redução de perdas aleatória e o fato desses sistemas não reduzirem diretamente a probabilidade de ocorrência de uma perda.

De acordo com Haberman e Renshaw (1996), a destruição de uma propriedade por incêndio ou catástrofe natural; uma doença prolongada que pode ocorrer inesperadamente e resultar em perdas financeiras em termos de redução de renda e despesas em saúde; a morte de um jovem adulto e a sobrevivência até uma idade avançada podem ser citadas como eventos aleatórios geradores de perdas financeiras. Diante do exposto, verifica-se que há um consenso na literatura ao apresentar, como uma das principais atividades do atuário, o gerenciamento da incerteza.

Dentro desse cenário, o gerenciamento do risco pode ser promovido em vários estágios distintos, por exemplo, a coleta de dados, a análise, construção de um modelo teórico, a previsibilidade do modelo e o monitoramento dos pressupostos do modelo. Com isso, é verificado, em linha com Jewell (1980), que os modelos utilizados por atuários na gestão da incerteza possuem características comuns nos principais ramos, vida e não-vida, da ciência atuarial. Uma dessas características é a existência de uma ou mais variáveis aleatórias que caracterizam as principais dimensões do risco, como duração, tamanho, número ou atraso. Com isso, é verificado a necessidade e importância da probabilidade e da estatística na identificação dos modelos atuariais.

Nesse ponto, como ferramenta de análise atuarial, é destacada a metodologia de modelos lineares generalizados (MLG´s), proposta por Nelder e Wedderburn (1972). A utilização do modelo consiste em promover a expansão das possibilidades que a distribuição da variável resposta pode assumir, fazendo, assim, com que a

mesma pertença à família de distribuições exponencial linear. Além disso, os MLG´s são capazes de fornecer uma maior versatilidade para o relacionamento funcional entre o termo médio da variável resposta e seu preditor linear.

Não obstante, Goldburd, Khare e Tevet (2016) reforçam a utilização de modelos lineares generalizados por atuários e sua crescente aderência no mercado segurador norte americano, revelando-se como um método de utilização em ascensão.

Dessa forma, o presente tem como objetivo geral apresentar a aplicação da metodologia dos Modelos Lineares Generalizados com foco em uma problemática atuarial, contribuindo para a literatura atuarial, especialmente a do Brasil. Para isso, será utilizada a base contida em Kaas et al. (2008), que reflete a experiência anual de um portfólio de seguros de automóveis.

2. INTRODUÇÃO AOS MODELOS LINEARES GENERALIZADOS

Comumente em diversos estudos estatísticos e atuariais, sejam eles da ordem experimental ou observacional, pesquisadores são confrontados com problemas cujo o objetivo principal é estudar o comportamento (relação) entre variáveis. Nesse sentido, especificamente, busca-se compreender a influência exercida por uma ou mais variáveis (explicativas) sobre uma determinada variável de interesse de nominada variável resposta. Assim, a metodologia, em geral, utilizada em tais situações é através da análise de regressão.

Em linha com Turkman e Silva (2000), o modelo linear, "desenvolvido" no século XIX por Lengendre e Gauss, foi o principal método utilizado para a modelagem estatística até meados do século XX. Contudo, ao longo de tal período, uma vasta gama de modelos não lineares e não normais foram desenvolvidos para atuar em situações onde o modelo linear normal não era adequado. Pode-se citar, como exemplos, o modelo complemento log-log (Fisher, 1922), o modelo probit (Bliss, 1935), os modelos log-lineares para dados de contagem (Birch, 1963), os modelos de regressão para análise de sobrevivência (Feigl and Zelen, 1965; Zippin and Armitage,1966; Glasser, 1967).

Tais modelos apresentados, entre outros, ditos alternativos ao modelo linear normal, compartilham as seguintes características: apresentam uma estrutura de regressão linear e a variável resposta do modelo segue uma distribuição dentro da família exponencial linear, onde são chamados de Modelos Lineares Generalizados (MLG). Além disso, Nelder e Wedderburn (1972), responsáveis pela proposta do modelo, desenvolveram um processo iterativo objetivando a estimação dos parâmetros e introduziram o conceito de desvio (amplamente utilizado para a mensuração da qualidade do ajuste do MLG).

Levando para a experiência atuarial, os modelos lineares generalizados têm apresentado uma crescente demanda em sua utilização devido à flexibilidade da variável resposta (modelagem de uma vasta gama de resultados, incluindo dados binários, dados de contagem e dados contínuos); precisão multivariada (combinação de vários efeitos de uma maneira que produza previsões consistentes e precisas) e sua capacidade de inferência (capacidade de testar formalmente as hipóteses, incluindo a significância estatística de uma variável preditora).

2.1. Família exponencial linear

Haberman e Renshaw (1996), Frees (2010) e Paula (2013) apresentam como base primordial dos MLGs (Modelos Lineares Generalizados) a suposição de que os dados são distribuídos a partir da família exponencial linear uniparamétrica. Diante disso, compreende-se que uma variável aleatória, Y, apenas é pertencente à família exponencial de dispersão se a sua função densidade de probabilidade (f.d.p) ou sua função de massa de probabilidade (f.m.p) possa ser escrito na forma abaixo:

$$f(Y;\ \theta,\phi) = exp\left\{\frac{Y(\theta) - b(\theta)}{a(\theta)} + c(\theta;\phi)\right\} \qquad (2.1)$$

Onde, a(.), b(.) e c(.) são descritos como funções reais específicas. Além disso, θ é o parâmetro de localização e Φ é tido como o parâmetro de dispersão. Com isso, é possível dizer que, de acordo com a fórmula acima, θ é a forma

canônica do parâmetro de localização e que Φ é apresentado com um parâmetro geralmente conhecido.

Ainda no que tange a família exponencial linear, definindo a função *score* e em linha com as condições de usuais de regularidade, é possível definir a média e a variância como sendo, respectivamente:

$$E(Y) = \mu = a(\Phi)E\big(S(\Phi)\big) + b'(\theta) \tag{2.2}$$

$$Var(Y) = a^2(\phi)\frac{b''(\theta)}{a(\phi)} = a(\phi)b''(\theta) \tag{2.3}$$

Logo, é possível notar que a variância decorre do produto de duas funções: função de variância, V(μ), que depende apenas do parâmetro canônico (valor médio μ) e da função a(Φ), dependendo apenas do parâmetro de dispersão. Ademais, é possível, ainda, que a função a(Φ) seja reescrita como:

$$a(\theta) = \frac{\phi}{\omega}. \tag{2.4}$$

Nesse caso, w pode ser interpretado como uma constante conhecida, fazendo com que a variância de Y seja o produto da métrica de dispersão por uma função apenas do valor médio. Assim, a Equação 2.1 pode ser apresentada como:

$$f(Y|\theta,\phi,\omega) = \exp\left\{\frac{\omega}{\phi}\big(Y(\theta) - b(\theta)\big) + c(Y,\phi,\omega)\right\}. \tag{2.5}$$

Um exemplo para o exposto acima seria considerar a função densidade de probabilidade de Poisson, apresentada de acordo com a função abaixo:

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}. \tag{2.6}$$

Logo, temos que, de acordo com a Equação 2.5, a função densidade de probabilidade pode ser escrita, sem perda de generalidade, como:

$$f(y;\lambda) = \exp\left((y\log(\lambda) - \lambda/1) - \log(y!)\right). \tag{2.7}$$

Daí, temos que θ = log(λ); b(θ) = λ = e^{θ}; a(Φ) = Φ = 1; c(y,Φ) = -log(λ!).

2.2. Especificação do Modelo Linear Generalizado (MLG)

O modelo linear generalizado pode ser compreendido como uma extensão do modelo linear, sendo definido como:

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}. \tag{2.8}$$

assim, **Z** pode ser definido como uma matriz *n* x *p*, que está relacionada a um vetor β =$(\beta_1, \dots, \beta_p)^T$ de parâmetros e ε á a componente de erros aleatoriamente distribuídos com distribuição $N_n(0, \sigma^2\boldsymbol{I})$. Diante disso, define-se que o valor esperado da variável resposta (Y) é uma função linear de suas covariáveis, ou seja, E(**Y**|Z)=μ com μ=Zβ.

O modelo pode ser estendido a partir de duas suposições. Na primeira, com relação a distribuição considerada, não é necessário que a mesma seja normal, podendo ser qualquer distribuição pertencente a família exponencial linear. Para a segunda suposição, embora seja mantida a estrutura de linearidade, a função responsável por relacionar o valor esperado e o vetor de covariáveis pode ser qualquer função diferenciável (TURKMAN e SILVA, 2000).

Dessa forma, os modelos lineares generalizados podem ser definidos de acordo com os seguintes componentes:

a) Componente aleatório: conhecendo o vetor de covariáveis ($\boldsymbol{x}_i$) as variáveis ($\boldsymbol{Y}_i$) são condicionalmente independentes com distribuição pertencente à família exponencial linear de acordo com as Equações 2.1 ou 2.5. Logo, E($\boldsymbol{Y}_i|\boldsymbol{x}_i$)= $\mu_i = b'(\theta_i)$ para i=1,...,n.
b) Componente sistemático: O valor esperado (μ_i) está relacionado com o preditor linear ($\eta_i = \boldsymbol{z}_i^T\beta$) de acordo com as seguintes relações:

$$\mu_i = h(\eta_i) = h(\boldsymbol{z}_i^T\beta) \tag{2.9}$$

e

$$\eta_i = g(\mu). \tag{2.10}$$

Daí, a partir de 2.9 e 2.10, tem-se que h é uma função monótona e diferenciavél; $g = h^{-1}$ é a função de ligação; β é um vetor de parâmetros com dimensão *p*; $\boldsymbol{z}_i$ é um vetor de parâmetros de dimensão *p*, função do vetor de covariáveis $\boldsymbol{x}_i$.

Vale ressaltar que a escolha da função de ligação depende do tipo de resposta e condução do estudo. Com isso, na Tabela 1 são apresentadas as principais funções de ligação canônicas. Em tais funções o preditor linear coincide com o parâmetro canônico.

Tabela 1 - Funções de ligação canônicas

Distribuição	Normal	Binomial	Poisson	Gama
Ligação	$\mu = \eta$	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\ln(\mu) = \eta$	μ^{-1}

Fonte: Elaborado pelos autores.

Em linha com Paula (2013), além das ligações canônicas, existem as seguintes ligações: Probito, Log-log, Box-Cox e Aranda-Ordaz.

Com relação às ligações canônicas, as vantagens de sua utilização estão no fato de que sendo o parâmetro de dispersão (Φ) conhecido, o valor de parâmetros desconhecido da estrutura linear admite uma estatística suficientemente mínima de dimensão fixa, onde as mesmas garantem uma estrutura côncava de $L(\beta)$, facilitando a obtenção de resultados assintóticos. Para mais, a estimação de β pode ser escrita como sendo um processo interativo via *Newton-Raphson*, em que a obtenção do valor estimado de verossimilhança acontece por meio da expansão da função *score* $\boldsymbol{U}_\beta$ que está definida em volta de um valor inicial definido como $\beta^{(0)}$.

Seguindo para a análise da qualidade do ajuste de um MLG, assume-se que a avaliação da mesma ocorre por meio da função desvio descrita abaixo.

$$D(\boldsymbol{y};\, \hat{\mu}) = \phi D(\boldsymbol{y};\, \hat{\mu}) = 2\{L(\boldsymbol{y};\boldsymbol{y}) - L(\hat{\mu};\boldsymbol{y})\}\,. \qquad (2.11)$$

Com isso, tem-se que essa métrica de discrepância, entre o modelo saturado e o modelo sob investigação, está baseada no modelo de razão de verossimilhanças de Wilks. Trivialmente é possível observar que o “desvio” será sempre maior ou igual a zero, podendo decrescer à medida que covariáveis (variáveis explicativas) são adicionadas ao modelo nulo. Além do mais, o mesmo é utilizado na verificação do modelo sob investigação.

2.3. DIAGNÓSTICOS DO MODELO

Com a finalidade de identificar eventuais discordâncias entre os pressupostos realizados para o modelo de regressão, uma das etapas mais relevantes em uma modelagem é o estudo do diagnóstico. Além disso, a realização efetiva do diagnóstico de um modelo permite que sejam identificados possíveis valores extremos, responsáveis por afetar de modo desproporcional ou inferencial no resultado do ajuste do mesmo.

Outro passo importante no estudo do diagnóstico é a análise de sensibilidade, podendo ser dividida em influência local e global. Ela consiste na avaliação das perturbações no modelo ajustado quando perturbações são imputadas nos dados ou nas suposições. Com isso, podem ser considerados os resíduos ordinários, resíduos de Pearson, resíduos de Pearson padronizados e as distâncias de Cook (1977) como técnicas essenciais para o diagnóstico dos modelos.

Além do exposto, vale ressaltar o gráfico de probabilidade meio-normal com envelope simulado apresentado por Atkinson (1985). Com isso, em um gráfico meio-normal, alinha-se o *i*-ésimo valor absoluto ordenado dos resíduos padronizados versus o valor esperado da métrica de ordem, de valor absoluto, do normal padrão, N (0,1). Segundo Neter et al. (1996), tal gráfico pode ser utilizado mesmo que os resíduos não apresentem uma distribuição normal.

A interpretação fundamental gerada pelo gráfico de probabilidade meio-normal com envelope simulado é que os desvios elevados, em relação ao valor mediano dos valores simulados, ou a incidência de pontos fora ou próximo dos limites da banda de confiança são indícios da não adequação do modelo.

Na seção seguinte é apresentado uma aplicação da metodologia de modelos de regressão lineares generalizados sob uma perspectiva atuarial.

3. APLICAÇÃO DO MLG EM UMA PROBLEMÁTICA ATUARIAL

O presente estudo tem por base, na construção do método cientifico aplicado, o viés fenomenológico. Buscando a identificação dos fatores geradores do fenômeno estudado, o objetivo do estudo apresenta uma natureza descritiva. Por fim, dada a utilização de recursos e métodos estatísticos, a pesquisa apresentada apresenta uma abordagem quantitativa.

Diante do exposto, utilizou-se dados secundários oriundos de Kass et al. (2008), em que os dados se referem a uma carteira de seguros de automóvel para a estimação da frequência média de sinistros por segurado de acordo com quatro variáveis tarifárias, também denominadas de fatores de risco. Os dez primeiros valores apresentados no conjunto de dados são apresentados na Tabela 2 e a classificação das variáveis de risco é apresentada na Quadro 1. Além disso, a amostra apresenta cinquenta e quatro observações, listadas ao longo de sete anos, referentes ao número de apólices com os fatores de risco apresentados (exposição) e o quantitativo de sinistros reportados à seguradora em um dado ano (qtd. Sinistros). A frequência média de sinistros, variável dependente, pode ser facilmente obtida por meio da divisão do quantitativo de sinistros reportados pelo número de apólices vezes mil.

Tabela 2 - Apresentação parcial do conjunto de dados

Exposição	Qtd.Sinistros	Sexo	T.Região	P.Veículo	S.Trabalho	Freq.Med.
70	1	1	1	1	1	14
154	8	1	1	1	2	52
210	10	1	1	1	3	48
77	8	1	1	2	1	104
105	5	1	1	2	2	48
140	11	1	1	2	3	79
175	14	1	1	3	1	80
175	12	1	1	3	2	69
161	11	1	1	3	3	68
196	10	1	2	1	1	51

Fonte: Kass et al. (2008).

Quadro 1 - Variáveis tarifárias consideradas no estudo

Fator de Risco	Descrição do Fator de Risco		
Sexo	1 – Masculino	2 – Feminino	
Tipo de Região	1 - Interior	2 - Zona de Transição	3 - Metrópole
Porte do Veículo	1 - Pequeno	2 - Médio	3 - Grande
Status de Trabalho	1 - Empregado	2- Desempregado	3 - Autônomo

Fonte: Kass et al. (2008).

Os dados, por meio do *software R*, foram reorganizados em função do valor da frequência média anual (Tabela 3), tal etapa permite uma manipulação e visualização adequada dos mesmos.

Tabela 3 - Variáveis tarifárias consideradas no estudo

P.Veículo		1			2			3		
S.Trabalho		1	2	3	1	2	3	1	2	3
Sexo	T.Região									
1	1	14	52	48	104	48	79	80	69	68
	2	51	38	78	98	82	113	116	95	118
	3	69	87	43	88	88	129	148	112	125
2	1	52	75	44	89	33	93	66	92	95
	2	86	24	71	89	53	77	75	170	127
	3	91	109	69	104	113	179	117	149	133

Fonte: Elaborado pelos autores.

Diante disso, buscou-se modelar a frequência média de sinistros por segurado, com base nas quatro variáveis tarifárias. Para tal, aplicou-se o modelo linear generalizado com distribuição de Poisson e função de ligação canônica, ou seja, Y= (qtd. Sinistros/ exposição) ~ Poisson (λ).

Assim, o modelo estatístico pode ser definido em 3.1 como:

$$\log(\mu_i) = \eta_i = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \quad (3.1)$$

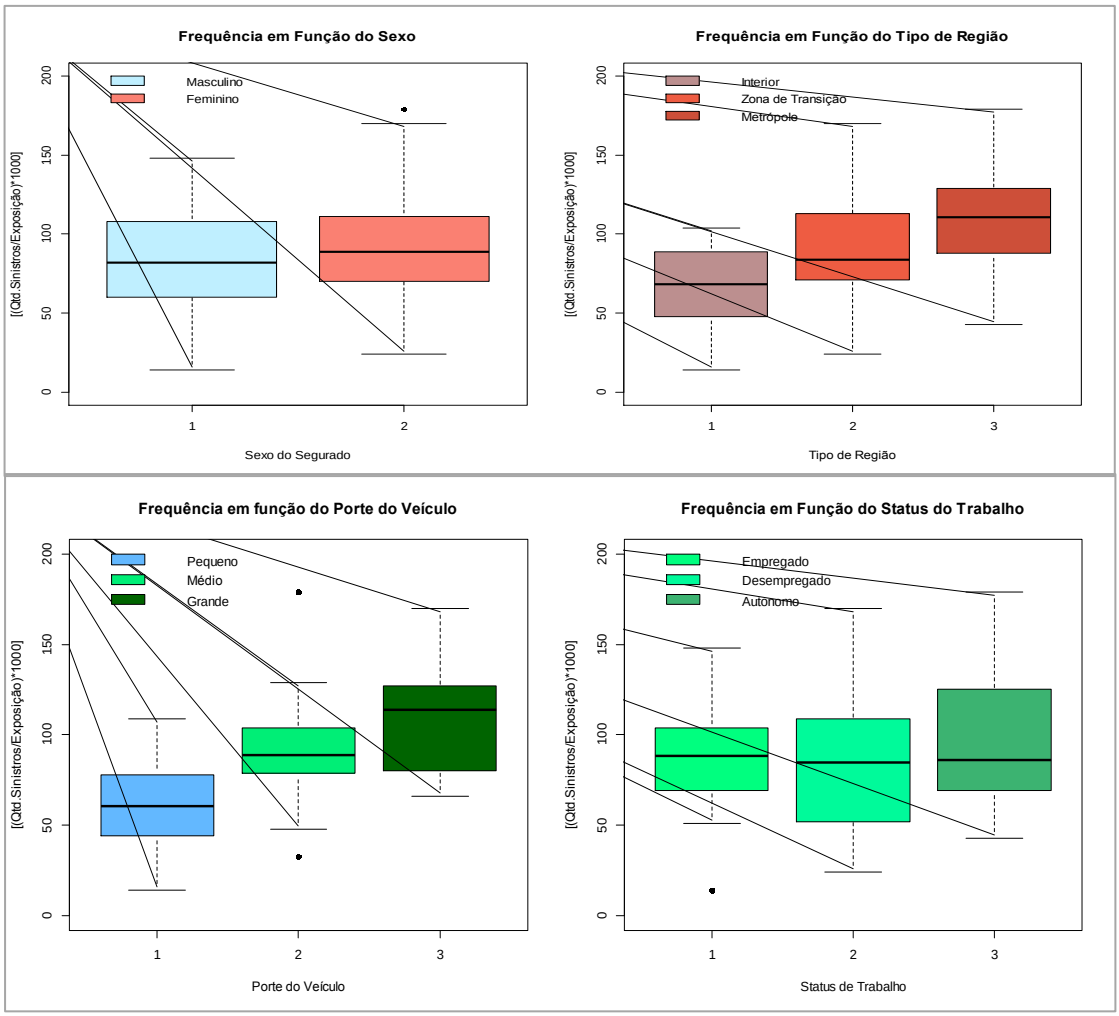
Por conseguinte, o modelo em estudo é descrito abaixo como:

$$\log(\mu_i) = \eta_i = \alpha + Sexo_i + T.Região_i + P.Veículo_i + S.Trabalho_i. \quad (3.2)$$

Daí, α representa o valor médio para o grupo de referência, o presente estudo definiu o grupo 1 como referência; as demais variáveis independentes do modelo (Sexo, Região, Veículo e Trabalho) representam as diferenças para entre suas equivalentes no grupo de controle.

Partindo para a análise das informações apresentadas, inicialmente, realizou-se um estudo exploratório dos dados onde foi possível identificar que o valor médio da frequência dos sinistros é de 87,31 ocorrências. Além disso, o desvio padrão registrado foi de 34,99, o valor mínimo e máximo foi, respectivamente, de 14 e 179 sinistros. A análise descritiva das variáveis tarifárias para a frequência de sinistros é apresentada por meio do Box Plot na Figura 1.

Figura 1 - Box Plot da frequência média em função de cada variável explicativa



Fonte: Elaborado pelos autores.

Para testar a hipótese de independência entre as variáveis do modelo, utilizou-se o teste Chi-quadrado que confirmou a independência das mesmas com um *p-valor* inferior a 0.0001. A partir disso, por meio da função *glm* contida no *software R*, o modelo de regressão linear generalizado foi ajustado para os dados e é apresentado na Tabela 5. Além disso, inicialmente, verificou-se que o desvio residual nulo foi de 104.73 em 53 graus de liberdade, o desvio residual foi de 41.93 em 46 graus de liberdade e o critério de informação de Akaike (AIC) apresentou um valor de 288.24.

Finalmente, para avaliar a significância dos parâmetros, notou-se pelo valor *pr(*> |*z*|), apresentado na Tabela 4, que os parâmetros Sexo2, T.Trabalho2 e T.Trabalho3 são não significativos para o modelo quando considerado um nível de significância de 5%, ou seja, seus valores para Pr(>|z|) são maiores que 0.05.

Tabela 4 - Estimativas dos parâmetros para o modelo

Coeficientes	Estimativa	Erro padrão	Valor *Z*	Pr(>|z|)
Intercepto	-3.099	0.122	-25.211	< 0.0001
Sexo2	0.103	0.076	1.350	0.177
Região2	0.234	0.099	2.365	0.018
Região3	0.464	0.096	4.809	< 0.0001
T.Carro2	0.394	0.101	3.881	< 0.0001
T.Carro3	0.584	0.097	6.019	< 0.0001
T.Trabalho2	-0.036	0.096	-0.373	0.709
T.Trabalho3	0.060	0.092	0.656	0.512

Fonte: Elaborado pelos autores.

Diante disso, a partir do modelo inicial, foi ajustado segundo modelo sem a inclusão dos parâmetros Sexo2, Trabalho2 e Trabalho3. Os resultados estão presentes na Tabela 5. Vale ressaltar que o novo modelo apresentou um AIC de 285.25. Tal valor evidencia que o segundo modelo é mais apropriado do que o modelo incialmente proposto, ou seja, há uma evidência de que o sexo e tipo de trabalho não são variáveis significativas para explicar o comportamento frequência média de sinistros.

Tabela 5 - Estimativas dos parâmetros para o segundo modelo

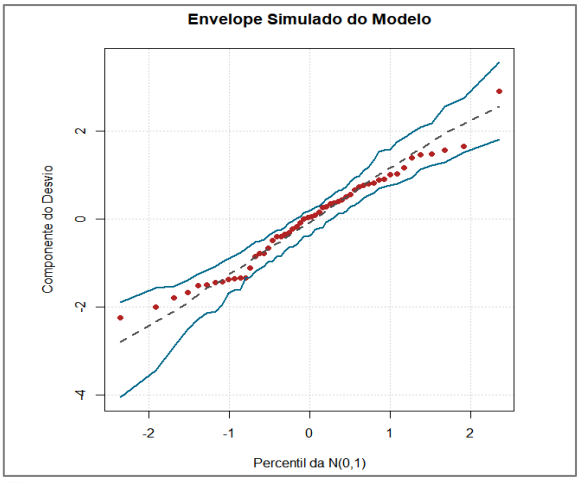
Coeficientes	Estimativa	Erro padrão	Valor *Z*	Pr(>|z|)
Intercepto	- 3.031	0.101	-29.862	< 0.0001
Região2	0.231	0.099	2.336	0.019
Região3	0.460	0.096	4.773	< 0.0001
T.Carro2	0.394	0.101	3.884	< 0.0001
T.Carro3	0.583	0.097	6.009	< 0.0001

Fonte: Elaborado pelos autores.

Para verificar o comportamento dos resíduos do segundo modelo, foi aplicado o teste Chi-quadrado (χ^2), onde a hipótese nula refere-se ao fato de que o desvio de resíduos não é significativamente grande, ou seja, o modelo é bom no que diz respeito aos resíduos; para a hipótese alternativa, tem-se que o desvio de resíduos é significativamente grande, desqualificando o modelo no que tange seus resíduos. Assim, o teste apresentou, para um desvio residual de 44.94 em 49 graus de liberdade, um valor de $\chi^2_{44.94,49} = 0.6384$ superior a 0.05. Com isso, para uma significância de 5%, não há evidencias para se rejeitar a hipótese nula.

Seguindo para o diagnóstico do modelo, foi possível identificar por meio do gráfico de probabilidade meio-normal com envelope simulado a 95% o bom ajuste dos dados à distribuição de Poisson com função de ligação canônica.

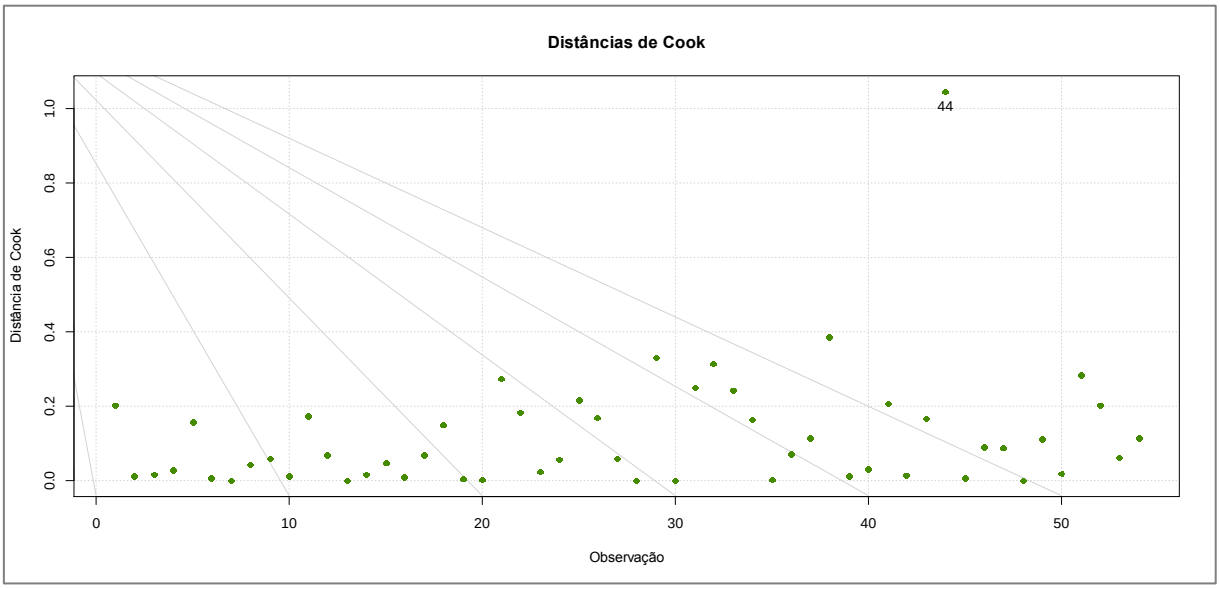
Figura 2 - Gráfico de probabilidade meio-normal com envelope simulado a 95%



Fonte: Elaborado pelos autores.

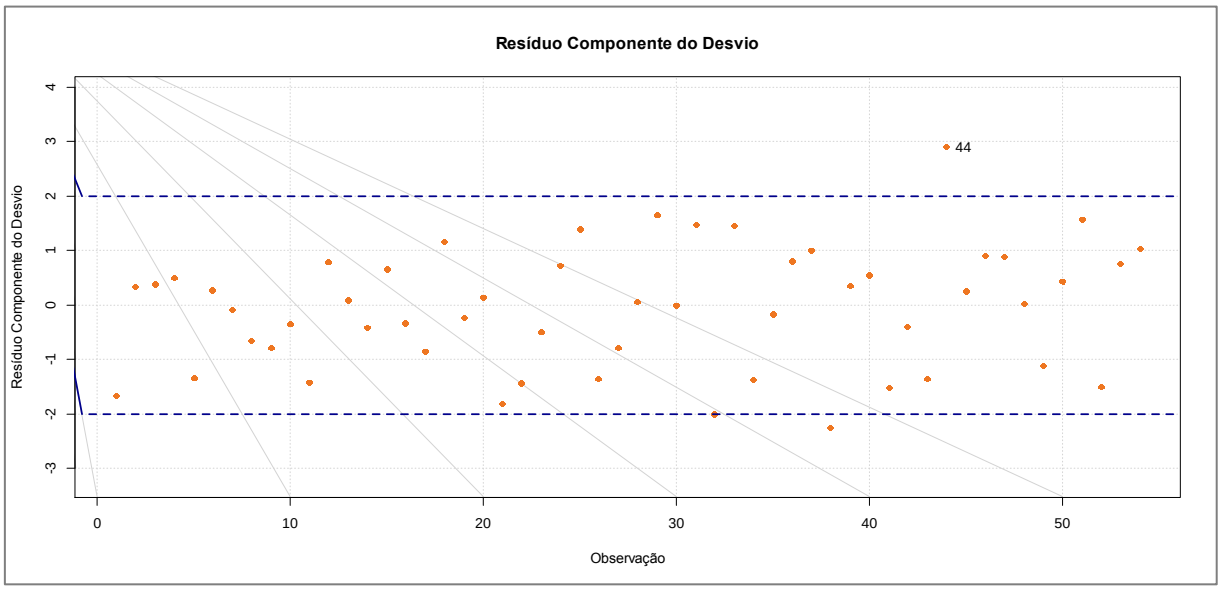
Outras técnicas de diagnósticos para o modelo foram realizadas, assim, foram analisadas as distâncias de Cook (Figura 3), o resíduo componente do desvio (Figura 4) e os resíduos versus o componente linear (Figura 5).

Figura 3 - Distâncias de Cook



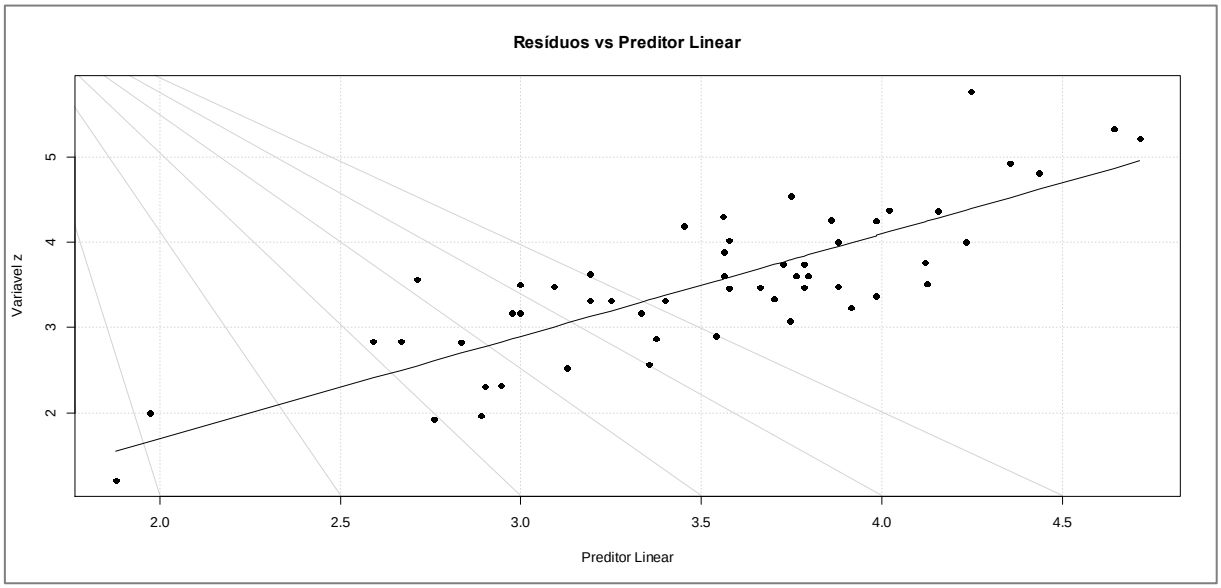
Fonte: Elaborado pelos autores.

Figura 4 - Resíduo componente do desvio



Fonte: Elaborado pelos autores.

Figura 5: Resíduos versus o componente linear



Fonte: Elaborado pelos autores.

De acordo com o diagnóstico do modelo, foi identificado a presença de um ponto influente (observação 44). Diante disso, é aplicada a utilização da medida de variação percentual descrita abaixo na Tabela 6, onde é possível avaliar a capacidade da influência exercida pelas observações distintas das demais sobre as estimativas dos parâmetros de dispersão, ou seja, verificar a influência da observação 44 sob os parâmetros. Para isso, buscou-se identificar o comportamento dos parâmetros com e sem a influência do ponto 44.

Tabela 6 - Variação percentual dos parâmetros.

Coeficientes	Estimativa	Estimativa sem obs.44	Variação %
Intercepto	- 3.031	-3.016	8.6%
Região2	0.231	0.168	27.3%
Região3	0.460	0.466	1.3%
T.Carro2	0.394	0.404	2.5%
T.Carro3	0.583	0.541	-7,2%

Fonte: Elaborado pelos autores.

Por meio da utilização do teste T, foi possível inferir com um nível de confiança de 95% que as estimativas, em média, são iguais. Isso evidencia que não há grande alteração na estimação dos parâmetros, ou seja, a influência do ponto 44 não se mostrou significante.

Finalizando com a predição do modelo, conclui-se que, dado o segundo modelo utilizado, a exposição para o grupo de referência, com os fatores de riscos iguais a um, é equivalente a $e^{(-3.03132)} = 0.0483$. Para saber o tempo médio, em anos, que um segurado do grupo de referência irá sofrer um sinistro, tem-se que $0.0483 * \mathrm{x} = 1$, assim, $\mathrm{x} = 1/0.0483 = 20.725$ anos. Similarmente, o processo pode ser estendido para os demais grupos de risco.

4. CONCLUSÕES

Com base nos resultados apresentados pelo presente estudo é possível averiguar que a metodologia de Modelos Lineares Generalizados (MLG´s) apresentou uma boa performance para a solução de problemáticas comumente vivenciadas nas rotinas atuariais. Além disso, o presente estudo, como contribuição, estendeu a metodologia presente em Kass et al. (2008) apresentando os diagnósticos para modelo. Foi possível identificar que as variáveis tarifárias referentes ao sexo e ao tipo de trabalho do segurado não foram significativas para o modelo. Com isso, dado o critério informação de Akaike, o modelo inicialmente proposto foi ajustado foi ajustado. Outro aspecto identificado no estudo foi a presença de um possível ponto influente, nesse caso, por meio da análise de variação e do teste T, foi possível inferir com um nível de confiança de 95% que o ponto não é significativamente influente.

Por meio da utilização do modelo de regressão, também foi possível a realização de predições. O modelo predisse o tempo médio, em anos, para que um segurado do grupo de referência sofresse um sinistro, revelando um tempo de 20.725 anos. Diante disso, o estudo buscou evidenciar uma metodologia capaz de operar em uma situação muito importante para os atuários, que é o sistema de classificação para definição e cálculo dos prêmios a serem recebidos dos

segurados. Com isso, foi possível demostrar que a classificação das perdas observadas de acordo com os fatores de risco apropriados é muito importante para determinar a precisão do sistema de classificação, no sentido de que os fatores de risco nos dizem exatamente qual nível de risco causa a maior perda (a ser cobrado o maior prêmio), e que causa a menor perda (a ser cobrado o menor prêmio). Vale ressaltar que além dos fatores de risco gerais (região de residência, idade do segurado, tipo de uso), algumas seguradoras tendem a classificar as perdas observadas de acordo com o chamado “sistema Bonus-Malus”.

Por fim, como sugestão de trabalhos futuros, indica-se a modelagem da frequência de sinistros por meio da aplicação dos modelos lineares generalizados para com base na distribuição binomial e funções de ligação do tipo probito, logito ou complemento log-log.

REFERÊNCIAS BIBLIOGRÁFICAS

ATKINSON, A. C. (1985). Plots, Transformations and Regressions. **Oxford Statistical Science Series**, Oxford.

BIRCH, M.W. (1963). Maximum likelihood in three-way contingency tables. **Journal of the Royal Statistical Society**, B52, 220- 233.

BLISS, C. I. (1935). The calculation of the dosage-mortality curve. **Annals of Applied Biology** 22, 134-167.

BOWERS, N. L., GERBER, H. U., JONES, D., HICKMAN, J. C. and NESBIT, C. (1986). **Actuarial Mathematics**. Chicago: Society of Actuaries.

COOK, R.D. (1977). Detection of influential observations in linear regression. **Technometrics**, 19, 15-18.

FEIGL, P. e ZELEN, M. (1965). Estimation of exponential survival probabilities with concomitant information. **Biometrics** 21, 826-838.

FISHER, R.A. (1922). On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society**, 222, 309-368.

FEIGL, P. e ZELEN, M. (1965). Estimation of exponential survival probabilities with concomitant information. **Biometrics** 21, 826-838.

FREES, E. W. (2010). **Regression Modeling with Actuarial and Financial Applications. 2 ed.** United States: Cambridge University Press, 2010. 565p.

GLASSER, M. (1967). Exponential survival with covariance. **Journal of the American Statistical Association**, 62, 561-568.

GOLDBURD, M.; KHARE, A.; TEVET, D. (2016). **Generalized linear models for insurance rating**. 1 ed. Arlington, Virginia: Casualty Actuarial Society, 2016. 96 p.

HABERMAN, S.; RENSHAW, A. E. (1996). Generalized Linear Models and Actuarial Science. **Journal of the Royal Statistical Society**, London, v. 45, n. 4, p. 407-436, jun. 1996.

JEWELL, W S. (1980) Models in insurance: paradigms, puzzles, communications and revolutions. **In Trans. 21st Int. Congr Actuaries**, vol. 5, pp. 87-141 Bern: Stampfli.

KAAS, R., GOOVAERTS, M., DHAENE, J., e DENUIT, M. (2008). **Modern actuarial risk theory: using R**. Springer Science and Business Media.

NELDER, J. A. e WEDDERBURN, R. W. M. (1972). Generalized linear models. **Journal of the Royal Statistical Society** A 135, 370-384.

PAULA, G. A. (2013). **Modelos de Regressão com apoio computacional**. São Paulo, 2013. Disponível em: < https://www.ime.usp.br/~giapaula/texto_2013.pdf >

TURKMAN, M. A. A.; SILVA, G. L. (2000). **Modelos lineares generalizados**: da teoria à prática. 1 ed. Lisboa: Universidade de Lisboa, 2000. 151p.

ZIPPIN, C. e ARMITAGE, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. **Biometrics**, 22, 665-672.