

Karolina Decnop Soares

**Avaliação de modelos estatísticos para a
precificação de seguros de automóveis na
cidade do Rio de Janeiro**

Niterói - RJ, Brasil

07 de fevereiro de 2022

Karolina Decnop Soares

**Avaliação de modelos estatísticos
para a precificação de seguros de
automóveis na cidade do Rio de
Janeiro**

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientadora: Profa. Dra. Estelina Serrano de M. Capistrano

Niterói - RJ, Brasil

07 de fevereiro de 2022

Karolina Decnop Soares

**Avaliação de modelos estatísticos para a
precificação de seguros de automóveis na
cidade do Rio de Janeiro**

Monografia de Projeto Final de Graduação sob o título “*Avaliação de modelos estatísticos para a precificação de seguros de automóveis na cidade do Rio de Janeiro*”, defendida por Karolina Decnop Soares e aprovada em 07 de fevereiro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Estelina Serrano de M. Capistrano
Departamento de Estatística – UFF

Prof. Dr. Luis Guillermo Coca Velarde
Departamento de Estatística – UFF

Profa. Dra. Ana Beatriz Monteiro Fonseca
Departamento de Estatística – UFF

Niterói, 07 de fevereiro de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

S676a	<p>Soares, Karolina Decnop Avaliação de modelos estatísticos para a precificação de seguros de automóveis na cidade do Rio de Janeiro / Karolina Decnop Soares ; Estelina Serrano de M. Capistrano, orientadora. Niterói, 2022. 85 f. : il.</p> <p>Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.</p> <p>1. Ciências Atuariais. 2. Modelos Lineares Generalizados. 3. Prêmio. 4. Inferência Bayesiana. 5. Produção intelectual. I. Capistrano, Estelina Serrano de M., orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.</p>
	CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O mercado de seguro de automóveis no Brasil é extremamente competitivo e obriga as seguradoras a realizarem uma tarifação correta e bem ajustada de acordo com o perfil do segurado. Neste trabalho, foram analisados modelos estatísticos para a precificação de seguros de automóveis na cidade do Rio de Janeiro, usando informações de uma carteira de apólices de seguros de uma determinada seguradora brasileira, no ano de 2015. Além disso, buscou-se determinar quais são os fatores relevantes para determinar o preço pago pelo segurado, o qual denominou-se por prêmio, a fim de que a seguradora assuma o risco do pagamento de indenizações, caso ocorra algum sinistro (como acidentes, perda de bens, morte etc.). De um modo geral, o prêmio deve ser suficiente para cobrir os sinistros esperados e as demais despesas da seguradora, incluindo uma margem de lucro. Devido à natureza assimétrica dos dados de severidade, foram adotados os modelos lineares generalizados para modelar a precificação de seguros, levando em consideração as características individuais do segurado, bem como as informações disponíveis sobre sua região de residência e seu respectivo automóvel. Foi realizado um paralelo entre as abordagens frequentista e Bayesiana e, em especial, foram considerados os modelos log-normal e gama para ajustar o prêmio do seguro de automóveis. Considerando o AIC como medida de comparação entre os modelos frequentistas e, o DIC entre os modelos Bayesianos, foi observado que o modelo log-normal se ajustou melhor ao conjunto de dados analisados em ambas as abordagens. Pelo fato de terem sido adotadas distribuições *a priori* vagas, os resultados das estimativas dos coeficientes foram similares para os métodos frequentista e Bayesiano. Além disso, todas as covariáveis consideradas (idade, sexo, estado civil, categoria do veículo e área de planejamento) mostraram-se relevantes para a modelagem de prêmio em ambas as abordagens.

Palavras-chave: Ciências Atuariais. Modelos lineares generalizados. Prêmio. Inferência Bayesiana.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Definições	p. 10
1 Introdução	p. 11
1.1 Objetivos	p. 13
1.2 Organização	p. 14
2 Materiais e Métodos	p. 15
2.1 Base de dados	p. 15
2.2 Modelos de Regressão Múltipla	p. 16
2.3 Família Exponencial	p. 19
2.3.1 Distribuição Log-Normal	p. 19
2.3.2 Distribuição Gama	p. 20
2.4 Modelos Lineares Generalizados	p. 21
2.5 Procedimento de Inferência	p. 22
2.5.1 Abordagem Frequentista	p. 22
2.5.2 Abordagem Bayesiana	p. 26
2.6 Comparação dos Modelos	p. 30
2.6.1 AIC (Akaike Information Criterion)	p. 30

2.6.2 DIC (Deviance Information Criterion)	p. 30
3 Resultados	p. 32
3.1 Dados Simulados	p. 32
3.1.1 Resultados obtidos para o conjunto de dados 1	p. 33
3.1.2 Resultados obtidos para o conjunto de dados 2	p. 36
3.2 Análise exploratória dos dados reais	p. 38
3.3 Modelagem do valor do prêmio	p. 42
3.3.1 Ajuste Frequentista	p. 43
3.3.2 Ajuste Bayesiano	p. 46
4 Conclusão	p. 49
Apêndice A – Tabela da relação de bairros e áreas de planejamento da cidade do Rio de Janeiro	p. 51
Apêndice B – Estimador de β por mínimos quadrados	p. 53
Apêndice C – Propriedades da Família Exponencial	p. 54
Apêndice D – Propriedades da Função Escore	p. 57
Apêndice E – Resultados dos Dados Ajustados para uma Distribuição Normal	p. 59
Apêndice F – Código R utilizados no Estudo de Simulação	p. 61
Referências	p. 84

Listas de Figuras

1	Mapa das Áreas de Planejamento da Cidade do Rio de Janeiro	p. 16
2	Gráfico da distribuição de densidade das funções a <i>posteriori</i> dos betas, com dados gerados e ajustados por uma distribuição Log-Normal.	p. 34
3	Gráfico da distribuição de densidade das funções a posteriori dos betas, com dados gerados por uma distribuição Log-Normal e ajustados por uma Gama.	p. 34
4	Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados e ajustados por uma distribuição Log-Normal.	p. 35
5	Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados por uma distribuição Log-Normal e ajustados por uma Gama. .	p. 35
6	Gráfico da distribuição de densidade das funções a <i>posteriori</i> dos betas, com dados gerados e ajustados por uma distribuição Gama.	p. 36
7	Gráfico da distribuição de densidade das funções a posteriori dos betas, com dados gerados por uma distribuição Gama e ajustados por uma Log-Normal.	p. 37
8	Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados e ajustados por uma distribuição Gama.	p. 37
9	Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados por uma distribuição Gama e ajustados por uma Log-Normal. .	p. 37
10	Mapa da cidade do Rio de Janeiro segundo o valor médio do prêmio por área de planejamento de acordo com a base de dados em estudo.	p. 38
11	Histograma para a variável idade. A linha tracejada corresponde à média.	p. 39
12	<i>Boxplot</i> do valor do prêmio na carteira de apólices (geral) e segregados de acordo com o sexo e o estado civil dos segurados.	p. 41

13	<i>Boxplot</i> do valor do prêmio de acordo com a categoria do veículo	p. 41
14	<i>Boxplot</i> do valor do prêmio de acordo com a área de planejamento	p. 42
15	Gráfico de resíduos e Q-Q Plot para o modelo de regressão múltipla . . .	p. 43

Listas de Tabelas

1	Exemplos de funções de ligação	p. 22
2	Estimativas frequentistas e bayesianas de dados simulados considerando uma variável resposta com distribuição Log-Normal.	p. 33
3	Estimativas Frequentistas e Bayesianas de Dados Simulados considerando uma Variável Resposta com Distribuição Gama.	p. 36
4	Medidas descritivas das características presentes em uma carteia de apólices de seguros de automóveis.	p. 40
5	Estimativas frequentistas (média e intervalo de confiança) dos parâmetros com ajuste pelas funções Log-Normal e Gama.	p. 44
6	Exponenciais das médias dos coeficientes de regressão obtidas pela ajuste Log-Normal através da abordagem frequentista.	p. 45
7	Estimativas Bayesianas (média e intervalo de credibilidade) dos parâmetros com ajuste pelas funções Log-Normal e Gama.	p. 47
8	Exponenciais das médias dos coeficientes de regressão obtidas pela ajuste Log-Normal através da abordagem Bayesiana.	p. 48
10	Relação de bairros e APs da cidade do Rio de Janeiro	p. 52
11	Estimadores dos parâmetros para o modelo de regressão múltipla . . .	p. 60

Listas de Definições

Apólice: documento emitido por uma seguradora, que formaliza a aceitação do risco objeto do contrato de seguro.

Precificação: processo de definição do valor monetário a ser cobrado ao cliente por um produto ou serviço.

Prêmio (comercial): valor monetário pago pelo segurado para a contratação do seguro.

Segurado: pessoa física ou jurídica pela qual a seguradora assume a responsabilidade sobre determinados riscos.

Seguradora: instituição que assume riscos previstos em um contrato, mediante o recebimento de um prêmio, obrigando-se a indenizar prejuízos ou reparar danos ocorridos em bens e pessoas.

Severidade: magnitude da perda monetária devido à ocorrência de sinistro.

Sinistro: qualquer evento em que o bem segurado sofre um acidente ou prejuízo material.

1 Introdução

De acordo com o Fórum Brasileiro de Segurança Pública, entre 2013 e 2019, o Brasil registrou 3.495.366 roubos ou furtos de veículos, o que nos leva a uma média de aproximadamente 499 mil roubos ou furtos por ano. Portanto, pode-se perceber que é essencial contratar o seguro de automóveis. Segundo Souza (2018), o seguro é garantido através de um documento emitido pela empresa seguradora, conhecido como apólice, onde a companhia seguradora garante uma indenização ao segurado caso ocorra o sinistro, isto é, algum tipo de dano, como perda de bens, acidentes, morte etc. Apesar de um contrato de seguro ser de extrema importância, de acordo com o diretor executivo da Federação Nacional de Seguros Gerais (FenSeg), Julio Cesar Rosa, aproximadamente 70% dos carros que circulam no Brasil não possuem seguro (CNSEG, 2017). Um dos fatores que influenciam neste fato é o preço do seguro, chamado de prêmio, que tem grande variabilidade em diversas regiões do país.

Considerando este problema, é de suma importância compreender os fatores que influenciam a precificação do seguro, ou seja, o processo que determina o custo a se cobrar pelo seguro, para que assim, o valor do prêmio proposto pela seguradora seja competitivo, de forma a cobrir o valor de possíveis indenizações e prováveis oscilações. Além da competitividade do mercado, há vários outros indicadores que influenciam na precificação dos seguros, tais como as características do segurado, do automóvel e do local de residência; portanto, a consideração de certas características individuais também é fundamental na modelagem do prêmio.

Na literatura atuarial, existem diversos conceitos de prêmio, que diferem em seus cálculos. De acordo com Ferreira (2002), o prêmio de risco representa o valor esperado total das indenizações ocorridas em uma carteira de seguros em determinado tempo, obtido através do produto da probabilidade de ocorrência de sinistros e do valor médio dos mesmos. Já o prêmio puro equivale ao prêmio de risco acrescido de um carregamento

de segurança, que cobre as flutuações estatísticas do risco. Por fim, o prêmio comercial corresponde ao prêmio puro acrescido do carregamento para as demais despesas da seguradora e uma margem de lucro. A base de dados utilizada neste trabalho apresenta os valores de prêmio comercial.

Conforme Duncan (2007), os modelos lineares generalizados (MLGs) são os métodos usuais na precificação da linha de seguros de automóveis na União Europeia e, cada vez mais, ganham popularidade no Brasil. Esses modelos são capazes de evidenciar quais variáveis são relevantes no comportamento da variável resposta, além de analisar a relação entre esta variável e as demais variáveis explicativas. Sendo assim, o MLG é ideal para modelar o prêmio do seguro discutido neste trabalho. Outros autores já utilizaram um MLG para avaliar o comportamento de variáveis relacionadas ao seguro, como Bandeira (2013), Santos (2008) e Farias e Jesus (2020).

Bandeira (2013) avaliou o custo da cobertura do ambulatório no Seguro Saúde em Portugal. Esta cobertura, juntamente com a internação e a estomatologia são as principais coberturas do Seguro Saúde. No entanto, de acordo com a autora, apesar de a internação ser a cobertura de maior mutualidade entre os clientes, por ter uma probabilidade relativamente pequena (3%), a cobertura que torna o seguro saúde mais apelativo é o seguro ambulatório, pois é utilizada com mais frequência. Por este motivo, foi realizado um estudo focado no custo da cobertura de ambulatório (que consiste em consultas, urgências, análises, exames e fisioterapia) através dos modelos de regressão, onde foi utilizado o modelo de regressão logística para melhor compreensão da ocorrência de sinistros e os modelos gama e normal para modelar a severidade dos mesmos. A autora observou um comportamento assimétrico positivo na variável prêmio, como esperado nesse tipo de dado, e, portanto, comparou os resultados dos modelos log-normal e gama. Nestes modelos, as variáveis idade, sexo, tipo de seguro e número de coberturas do produto foram relevantes para explicar o prêmio.

Santos (2008) optou por modelar separadamente as variáveis frequência e severidade, para obter um entendimento mais detalhado do processo. Além disso, a autora relatou a importância de se realizar análise de dados que abranjam, no mínimo, um espaço de tempo de três anos, devido a possíveis sinistros não encerrados. E assim, para modelar o número de sinistros, ela utilizou o modelo Poisson enquanto que, para o custo dos mesmos, foi utilizado o modelo Gama. Por fim, Santos realizou uma junção dos modelos construídos separadamente para a severidade e a frequência de sinistros, resultando no modelo do

prêmio.

Já em Farias e Jesus (2020), foi realizado um estudo com objetivo de compreender a tarifação de seguros de automóveis no mercado do Nordeste. Para isso, foi realizada uma combinação de modelos lineares generalizados, entre o Poisson (modelo para a frequência de sinistros) e o log-normal (para a severidade), para que assim chegassem em um modelo final, tendo como variável resposta o prêmio. Os dados utilizados pelos autores são referentes às estatísticas fornecidas pela SUSEP, no primeiro semestre de 2018. No trabalho, os autores não observaram homogeneidade da variância nos resíduos da distribuição log-normal, o que indica que esta distribuição não seria a mais indicada para a construção do modelo de severidade.

Após a discussão da relevância da modelagem do prêmio do seguro e de como os MLGs são métodos importantes para este fim, foi apresentada a motivação desse trabalho, que consiste em modelar o prêmio de seguros de automóveis na cidade do Rio de Janeiro, identificando as variáveis que influenciam na precificação do mesmo. Há anos que a criminalidade é um problema cotidiano para a população carioca. Dentre os crimes que impedem a tranquilidade nas ruas da capital, está o furto de veículos.

De acordo com o ISP (2021), as áreas integradas de segurança pública (AISP) da cidade do Rio de Janeiro que mais registraram furto de veículos entre os anos de 2018 e 2020 foram: em primeiro lugar, as áreas onde estão localizadas as delegacias de Realengo e Bangu, com 2.252 furtos de veículos neste período; o segundo maior registro foi feito nas áreas onde se localizam as delegacias da Barra da Tijuca e Guaratiba, com 2.051 furtos de veículos e, por fim, as áreas onde se localiza a delegacia de Campo Grande, com 1.873 furtos de veículos e onde se localizam as delegacias de Vicente de Carvalho, Ricardo de Albuquerque e Pavuna, com 2.163 registros deste tipo de furto (ISP, 2021).

1.1 Objetivos

O objetivo principal deste trabalho é modelar o valor do prêmio de seguros de automóveis na cidade do Rio de Janeiro utilizando os modelos lineares generalizados e avaliar os efeitos das características individuais dos segurados no valor do prêmio. Mais especificamente, descrever o prêmio a partir das informações disponíveis na base de dados sobre o segurado, o automóvel e a região de residência através dos modelos log-normal e gama para. Além disso, tem-se como objetivo comparar os resultados da modelagem sob

abordagem frequentista e Bayesiana.

1.2 Organização

Este trabalho está organizado da seguinte forma: o Capítulo 2 introduz a base de dados de seguros utilizada e apresenta uma revisão sobre os modelos de regressão múltipla, a família exponencial e os modelos lineares generalizados. Em seguida são explicitados os métodos utilizados para as estimações frequentista e Bayesiana dos parâmetros, mais especificamente o método iterativo de Newton-Raphson e o método de MCMC. Em seguida, no Capítulo 3, foi realizado um estudo com dados simulados e uma análise exploratória dos dados reais. Em seguida, são apresentados os modelos adotados e os resultados das análises realizadas. Por fim, no Capítulo 4, segue a conclusão com as considerações finais. O Apêndice A apresenta uma tabela com a relação dos bairros e áreas de planejamento da cidade do Rio de Janeiro. O Apêndice B, contém os cálculos necessários para encontrar o estimador de β para o modelo de regressão múltipla através do método dos mínimos quadrados. Já o apêndice C possui cálculos que demonstram as propriedades da família exponencial, enquanto o Apêndice D demonstra as propriedades da função escore. O Apêndice E contém uma tabela com os resultados dos dados ajustados para uma distribuição normal sob a ótica frequentista. Por fim, o Apêndice F contém os códigos em R utilizados nesse trabalho.

2 Materiais e Métodos

2.1 Base de dados

A base de dados a ser estudada contém informações sobre uma determinada seguradora de veículos brasileira que, por motivos de confidencialidade, não será identificada. Os dados referem-se ao ano de 2015 e foram fornecidos pela própria seguradora. A área de estudo adotada foi a cidade do Rio de Janeiro, estando os bairros dessa cidade agrupados em 5 áreas de planejamento.

O banco de dados contém informações sobre as características individuais do segurado, bem como informações sobre a região de residência e o tipo de automóvel segurado. Além do valor do prêmio comercial, que será considerada a variável resposta, a base de dados disponibiliza as seguintes variáveis: sexo (1 - masculino e 0 - feminino); estado civil (1 - solteiro e 0 - casado); idade do segurado; categoria de veículo assegurado (16 categorias ao total) e área de planejamento referente ao bairro de residência do segurado. O mapa das áreas de planejamento da cidade do Rio de Janeiro, pode ser observado na Figura 1, já a relação de bairros por área de planejamento pode ser observada no Apêndice A.

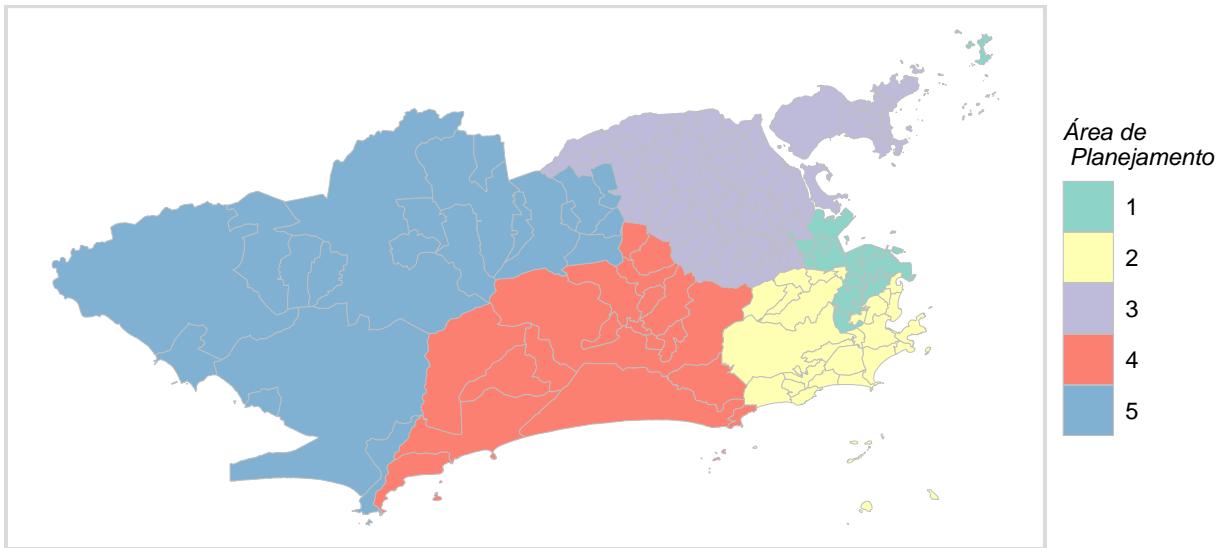


Figura 1: Mapa das Áreas de Planejamento da Cidade do Rio de Janeiro

A seguir, serão apresentadas as metodologias que serão utilizadas neste trabalho, a fim de identificar quais são as características relevantes, dentre as disponíveis, para ajustar o prêmio de seguros de automóveis. Inicialmente, serão introduzidos os modelos de regressão múltipla, bem como a notação a ser adotada ao longo deste trabalho, destacando ainda as suposições necessárias para utilização destes modelos.

2.2 Modelos de Regressão Múltipla

De acordo com Yan e Su (2009), o objetivo dos modelos de regressão múltipla é compreender o efeito de duas ou mais variáveis independentes sobre a variável de interesse de forma simultânea. Em outras palavras, a regressão pode ser utilizada para examinar quanto um determinado conjunto de variáveis independentes podem explicar suficientemente o resultado. Para tal fim, inicialmente são selecionadas as variáveis explicativas que sejam de fato relevantes para o modelo e, em seguida, estimados seus coeficientes.

O modelo de regressão múltipla é capaz de adquirir capacidade preditiva, ou seja, realizar previsões de valores futuros para a variável de interesse, um fator determinante para a eficiência do modelo. Tal característica pode ser conferida através da comparação entre o valor de cada observação da variável de interesse, presente na amostra, e o valor

predito obtido através do modelo.

Seja Y_i a variável resposta para o i -ésimo indivíduo ($i = 1, \dots, n$). O vetor de variáveis respostas que representa o componente aleatório do modelo, é denominado por \mathbf{Y} onde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

Considerando X_{ij} a j -ésima variável explicativa ($j = 1, \dots, k$) para o i -ésimo indivíduo, tem-se a matriz de variáveis explicativas \mathbf{X} , onde a primeira coluna é constituída de um vetor de 1's. Ou seja,

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}.$$

Tais variáveis explicativas podem ser tanto quantitativas, quanto qualitativas. Caso sejam qualitativas e assumam vários níveis, são construídas variáveis *dummies*, ou seja, variáveis binárias que assumem valor 1 caso pertençam a um determinado nível desta variável e, valor 0 caso contrário.

A combinação linear entre as variáveis explicativas e seus respectivos coeficientes, representa o componente sistemático do modelo, denominado por $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, sendo $\boldsymbol{\beta}$ o vetor de coeficientes que define os pesos, isto é,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

onde β_0 é o intercepto e β_j representa o coeficiente associado à j -ésima variável explicativa ($j = 1, \dots, k$).

Por fim, considerando ϵ_i o erro aleatório da observação do i -ésimo indivíduo ($i = 1, \dots, n$), tem-se que o vetor de erros aleatórios $\boldsymbol{\epsilon}$ do modelo pode ser escrito como

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Portanto, de forma geral, o modelo de regressão múltipla pode ser escrito na forma matricial como $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

De acordo com Dobson e Barnett (2018), para que o modelo seja aplicado, os dados devem satisfazer as seguintes suposições:

- **linearidade:** o vetor de variáveis resposta \mathbf{Y} é uma função linear da matriz de variáveis explicativas \mathbf{X} ;
- **multicolinearidade:** as variáveis explicativas não apresentam correlação alta entre si;
- **normalidade, homocedasticidade e independência:** o vetor de erros aleatórios $\boldsymbol{\epsilon}$ segue distribuição normal tal que $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$ e $Cov(\epsilon_i, \epsilon_{i'} = 0)$, sendo $i \neq i'$, onde $i, i' = 1, \dots, n$;

De acordo com Rao e Toutenburg (1999), o vetor $\boldsymbol{\beta}$ pode ser estimado através do método dos mínimos quadrados ao definir o vetor $\hat{\boldsymbol{\beta}}$ que minimiza a soma dos quadrados dos resíduos. Tal soma, apresentada na Equação 2.1 representa as distâncias quadráticas entre os valores observados de Y e seus valores ajustados pelo modelo, denotados por \hat{Y} .

$$SQE = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (2.1)$$

Igualando a derivada da Equação 2.1 a zero, obtém-se que $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. Tal demonstração pode ser encontrada no Apêndice B.

Muitas das vezes, o conjunto de dados a ser analisado não satisfaz esses pré-requisitos, o que faz com que os modelos de regressão múltipla não sejam adequados. Portanto, é necessário adotar modelos que independem de algumas condições acima, como os modelos lineares generalizados, que serão apresentados na Seção 2.4. Porém, antes de explorar tais modelos, será apresentado o conceito de família exponencial e suas propriedades para facilitar a compreensão dos mesmos.

2.3 Família Exponencial

Considere uma variável aleatória Y com distribuição de probabilidade indexada pelo parâmetro de interesse θ . Diz-se que esta distribuição pertence à família exponencial caso possa ser escrita da seguinte forma:

$$f_y(y; \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}, \quad (2.2)$$

onde a e d são funções de y independentes do parâmetro θ e b e c são funções apenas de θ . Em especial, quando $a(y) = y$, é dito que a distribuição expressa na Equação 2.2 possui forma canônica.

Segundo Dobson e Barnett (2018), caso haja outros parâmetros, além do parâmetro de interesse θ , eles são considerados como parâmetros de incômodo que fazem parte das funções a , b , c e d , e são tratados como se fossem conhecidos.

Dentre as propriedades da família exponencial, pode-se mostrar que:

$$E[a(y)] = \frac{-c'(\theta)}{b'(\theta)} \quad (2.3)$$

e

$$Var(a(y)) = \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{[b'(\theta)]^3}, \quad (2.4)$$

onde b' , c' , b'' e c'' representam a primeira e a segunda derivada das funções b e c , respectivamente, com relação ao parâmetro θ .

As demonstrações da média e da variância descritas nas Equações 2.3 e 2.4, respectivamente, encontram-se disponíveis no Apêndice C. Para exemplificar tais propriedades, serão consideradas as duas distribuições pertencentes à família exponencial que são de interesse nesse trabalho.

2.3.1 Distribuição Log-Normal

Uma variável aleatória Y segue distribuição log-normal quando $\ln(Y)$ segue distribuição normal. Considerando que $Y \sim LogNormal(\mu, \sigma^2)$, a função de densidade pode ser escrita como:

$$f_y(y|\mu, \sigma) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\ln(y) - \mu)^2}{2\sigma^2}\right\}, \quad y > 0. \quad (2.5)$$

Admitindo que σ é conhecido, note que a função de densidade definida na Equação 2.5 pode ser reescrita como:

$$\begin{aligned} f_y(y|\mu, \sigma) &= \exp\{-\ln(y^2 2\pi\sigma^2)^{1/2}\} \exp\left\{\frac{-(\ln(y)^2 - 2\mu\ln(y) + \mu^2)}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{\mu}{\sigma^2}\ln(y) - \frac{1}{2\sigma^2}\ln(y)^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(y^2 2\pi\sigma^2)\right\}, \end{aligned}$$

onde $a(y) = \ln(y)$, $b(\mu) = \frac{\mu}{\sigma^2}$, $c(\mu) = \frac{\mu^2}{2\sigma^2}$ e $d(y) = \frac{1}{2\sigma^2}\ln(y)^2 - \frac{1}{2}\ln(y^2 2\pi\sigma^2)$.

Seguindo as propriedades apresentadas nas Equações 2.3 e 2.4, consegue-se determinar a esperança e variância desta distribuição. Derivando em relação a μ , tem-se que $b' = \frac{1}{\sigma^2}$ e $c' = \frac{-\mu}{\sigma^2}$. Portanto,

$$E(a(Y)) = E(\ln(Y)) = \frac{\mu/\sigma^2}{1/\sigma^2} = \mu.$$

Fazendo a segunda derivada em relação a μ , segue que $b'' = 0$ e $c'' = \frac{-1}{\sigma^2}$. Logo,

$$Var(a(Y)) = Var(\ln(Y)) = \frac{(1/\sigma^2)^2}{(1/\sigma^2)^3} = \frac{1}{1/\sigma^2} = \sigma^2.$$

2.3.2 Distribuição Gama

Suponha que $Y \sim Gama(\alpha, \alpha/\mu)$, com função de densidade de probabilidade parametrizada da seguinte maneira:

$$f_y(y|\mu, \alpha) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha y^{\alpha-1} e^{-\alpha(y/\mu)}, \quad y > 0. \quad (2.6)$$

Considerando μ o parâmetro de interesse e α como parâmetro de incômodo, a função de densidade definida na Equação 2.6 pode ser reescrita como

$$\begin{aligned} f_y(y|\mu, \alpha) &= \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha \exp\{(\alpha-1)\ln(y)\} \exp\left\{-\left(\frac{\alpha}{\mu}\right)y\right\} \\ &= \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha \exp\left\{(\alpha-1)\ln(y) - \left(\frac{\alpha}{\mu}\right)y\right\} \\ &= \exp\left\{\ln\left(\frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha\right) + (\alpha-1)\ln(y) - \left(\frac{\alpha}{\mu}\right)y\right\} \end{aligned}$$

onde $a(y) = y$, $b(\mu) = \frac{-\alpha}{\mu}$, $c(\mu) = \ln\left(\frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right)^\alpha\right)$ e $d(y) = (\alpha - 1)\ln(y)$.

Calculando a derivada da funções b e c em relação a μ , tem-se que $b' = \frac{\alpha}{\mu^2}$ e $c' = \frac{-\alpha}{\mu}$. Logo, pela Equação 2.3,

$$E(a(Y)) = E(Y) = \frac{\alpha/\mu}{\alpha/\mu^2} = \mu.$$

Além disso, pela segunda derivada, segue que $b'' = \frac{-2\alpha}{\mu^3}$ e $c'' = \frac{\alpha}{\mu^2}$. Logo, pela Equação 2.4, tem-se que:

$$Var(a(Y)) = Var(Y) = \frac{-2\alpha}{\mu^3} \frac{-\alpha}{\mu} - \frac{\alpha}{\mu^2} \frac{\alpha}{\mu^2} = \frac{\mu^2}{\alpha}.$$

Observe que, nessa parametrização, a variância depende do termo quadrático da média do processo.

Uma vez definido o conceito de uma distribuição pertencente à família exponencial, é possível explorar os modelos lineares generalizados, que não exigem suposições como linearidade e normalidade para que sejam aplicáveis, o que possibilita a sua utilidade em situações práticas.

2.4 Modelos Lineares Generalizados

Originalmente propostos em Nelder e Wedderburn (1972), os modelos lineares generalizados (MLGs) surgiram com o objetivo de possibilitar a construção de um modelo, a partir de uma base de dados, independentemente se a variável de interesse a ser modelada segue uma distribuição normal. Este tipo de modelo pode ser aplicado desde que as variáveis resposta que tenham mesma distribuição que pertençam à família exponencial.

Considerando o vetor de variáveis aleatórias independentes $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ de mesma distribuição da família exponencial, tem-se que a função de densidade conjunta de \mathbf{Y} pode ser escrita como

$$f_y(y_1, \dots, y_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(y_i; \boldsymbol{\theta}).$$

onde $\boldsymbol{\theta}$ é o vetor paramétrico que contém o vetor de parâmetros de interesse.

No MLG, a condição de linearidade entre a variável resposta e as variáveis explicativas não é essencial. Portanto, é realizada uma transformação da média μ_i ($i = 1, \dots, n$), através de uma função de ligação, denotada por g , a fim de obter uma relação linear entre a média transformada da variável e suas variáveis explicativas X_i . De acordo com Dobson e Barnett (2018), a relação entre a função de ligação e as variáveis explicativas é dada por

$$g(\mu_i) = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta} \quad (2.7)$$

e

$$\mu_i = E[Y_i]. \quad (2.8)$$

Algumas funções de ligações usuais na literatura e suas respectivas distribuições podem ser encontradas na Tabela 1.

Distribuição de Y	Funções de Ligação
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \ln \lambda$
Bernoulli	Logística: $\eta = \ln \left(\frac{p}{1-p} \right)$
Gama	Inversa: $\eta = \frac{1}{\mu}$ Logarítmica: $\eta = \ln \mu$

Tabela 1: Exemplos de funções de ligação

O vetor de paramétrico $\boldsymbol{\beta}$ é desconhecido e, portanto, precisa ser estimado. Ao longo da literatura estatística, foram desenvolvidos diversos métodos de estimação sob o ponto de vista frequentista e Bayesiano. A seguir, serão apresentados os métodos adotados neste trabalho para obtenção das estimativas de $\boldsymbol{\beta}$.

2.5 Procedimento de Inferência

2.5.1 Abordagem Frequentista

Um método plausível para estimar o vetor $\boldsymbol{\beta}$ seria o de máxima verossimilhança, mas existe a possibilidade de que o mesmo resulte em diversos valores ou até mesmo em nenhum. Ademais, caso exista um único resultado para o método, é provável que o mesmo não pertença a um intervalo admissível para $\boldsymbol{\beta}$. Por este motivo, será utilizado um

método iterativo, de acordo com Dobson e Barnett (2018), a fim de encontrar um valor aproximado para o estimador de máxima verossimilhança.

A função de verossimilhança de Y_i ($i = 1, \dots, n$) na forma canônica é dada por:

$$\begin{aligned} L(\theta_i; Y_i) &= f_{y_i}(y_i; \theta_i) \\ &= \exp\{y_i b(\theta_i) + c(\theta_i) + d(y_i)\}. \end{aligned} \quad (2.9)$$

Denotando a função de log-verossimilhança de Y_i por $l_i(\theta_i; Y_i) = \ln(L(\theta_i; Y_i))$ e considerando todas as observações, tem-se que a função de log-verossimilhança conjunta é dada por:

$$\begin{aligned} l(\theta; Y) &= \sum_{i=1}^n l_i(\theta_i; Y_i) = \sum_{i=1}^n \ln(L(\theta_i; Y_i)) \\ &= \sum_{i=1}^n y_i b(\theta_i) + \sum_{i=1}^n c(\theta_i) + \sum_{i=1}^n d(y_i) \end{aligned} \quad (2.10)$$

Para determinar os estimadores de máxima verossimilhança dos parâmetros, precisa-se encontrar os valores que maximizam a Equação (2.10) através da função escore. Definindo U_j como o vetor de escores do j -ésimo coeficiente, tem-se que:

$$U_j(\beta) = \frac{\partial l(\theta; Y)}{\partial \beta_j} = 0. \quad (2.11)$$

Logo, tem-se que a relação entre a função de log-verossimilhança e β_j pode ser obtida por:

$$\begin{aligned} U_j(\beta) &= \frac{\partial l(\theta; Y)}{\partial \beta_j} \\ &= \sum_{i=1}^n \left(\frac{\partial l_i}{\partial \beta_j} \right) = \sum_{i=1}^n \left(\frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \right), \end{aligned} \quad (2.12)$$

pois θ_i está relacionado a μ_i de acordo com as Equações 2.3 e 2.8 e μ_i está relacionado a β_j através do preditor linear, como apresentado na Equação 2.7. A partir da Equação 2.10, é possível encontrar as derivadas contidas na Equação 2.12. Segue que:

- $\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i);$

e, de acordo com as propriedades da família exponencial apresentadas nas Equações 2.3 e 2.4:

- $\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = \left(\frac{-c''(\theta_i)b'(\theta_i) + b''(\theta_i)c'(\theta_i)}{[b'(\theta_i)]^2} \right)^{-1} = \frac{1}{b'(\theta_i)Var(Y_i)};$

- $\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$

Assim, aplicada a regra da cadeia, a Equação 2.12 transforma-se em

$$U_j(\beta) = \sum_{i=1}^n \left[\frac{(y_i - \mu_i)}{Var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]. \quad (2.13)$$

Na forma matricial, tem-se que

$$U_j(\beta) = X^T \cdot D \cdot V^{-1} (y - \mu),$$

onde X^T é a matriz transposta de X ; V^{-1} é a matriz inversa da variância e D é a diagonal da matriz de $\frac{\partial \mu}{\partial \eta}$.

Algumas propriedades da função escore estão listadas a seguir. Suas demonstrações podem ser encontradas no Apêndice D.

- $E[U(\theta; y)] = 0;$
- $Var(U(\theta; y)) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta);$
- $Var(U) = E[U^2] = -E[U']$

Dada uma função escore $U_j(\beta)$, a sua matriz de variâncias e covariâncias, também conhecida como matriz de informação de Fisher e é composta pelos elementos

$$\tau_{jk} = E[U_j U_k] = E \left[\frac{\partial l}{\partial \beta_j} \frac{\partial l}{\partial \beta_k} \right] = -E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]. \quad (2.14)$$

para todo $j, k = 1, \dots, p$. Ou seja,

$$\tau_{jk} = -E \left[\sum_{i=1}^n \frac{\partial^2 li}{\partial \beta_j \partial \beta_k} \right] = \sum_{i=1}^n E \left[\frac{\partial li}{\partial \beta_j} \frac{\partial li}{\partial \beta_k} \right] = E \left[\left(\sum_{i=1}^n \frac{\partial li}{\partial \beta_j} \right) \left(\sum_{i=1}^n \frac{\partial li}{\partial \beta_k} \right) \right] \quad (2.15)$$

Substituindo a Equação (2.13) na expressão anterior, tem-se que:

$$\begin{aligned} \tau_{jk} &= E \left[\sum_{i=1}^n \left(\frac{(y_i - \mu_i)}{Var(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{(y_i - \mu_i)}{Var(Y_i)} x_{ik} \frac{\partial \mu_i}{\partial \eta_i} \right) \right] \\ &= \sum_{i=1}^n \left\{ E \left[\left(\frac{(y_i - \mu_i)^2 x_{ij} x_{ik}}{Var(Y_i)^2} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right] \right\} \end{aligned}$$

Como $E[(y_i - \mu_i)^2] = Var(Y_i)$, tem-se que:

$$\tau_{jk} = \sum_{i=1}^n \frac{x_{ij}x_{ik}}{Var(Y_i)^2} \left(\frac{\partial \mu_i}{\partial \eta_j} \right)^2. \quad (2.16)$$

Então, a matriz de informação de Fisher pode ser escrita como $\tau = X^T W X$, onde W é uma matriz diagonal $n \times n$ com elementos

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \quad (2.17)$$

Por fim, será utilizado o algoritmo de Newton-Raphson, mais amplamente explicado em Dobson e Barnett (2018), com intenção de aproximar o estimador de máxima verosimilhança de β , que pode ser obtidos através da Equação 2.11. Para isso, inicialmente, é escolhido um valor que pertence ao intervalo da função e, através da derivada, calcula-se o valor da reta tangente deste ponto, até que seja encontrado o ponto pertencente à função que seja o mais próximo àquele que maximiza a mesma. No caso dos MLGs, tem-se que:

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} - \frac{\mathbf{U}^{(m-1)}}{\mathbf{U}'^{(m-1)}}. \quad (2.18)$$

Pode-se ainda aproximar U' para o seu valor esperado $E[U']$. Já que $\tau = E[-U']$, sabe-se que:

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + \frac{\mathbf{U}^{(m-1)}}{\tau^{(m-1)}}, \quad (2.19)$$

onde

- $\hat{\beta}^{(m)}$ é o vetor de estimativas do parâmetro β na m -ésima iteração;
- $\mathbf{U}^{(m-1)}$ é o vetor dos elementos dados em (2.13), avaliados em $\hat{\beta}^{(m-1)}$;
- $\tau^{(m-1)}$ é a matriz de informação de Fisher com elementos dados em 2.16.

A Equação 2.19 pode ser escrita de forma matricial como

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + \mathbf{U}^{(m-1)} [\tau^{(m-1)}]^{-1}, \quad (2.20)$$

onde $[\tau^{(m-1)}]^{-1}$ é a inversa da matriz de informação de Fisher.

Reescrevendo a Equação 2.20, tem-se que:

$$\hat{\beta}^{(m)} [\tau^{(m-1)}] = \hat{\beta}^{(m-1)} [\tau^{(m-1)}] + \mathbf{U}^{(m-1)}, \quad (2.21)$$

onde, de acordo com as Equações 2.16 e 2.13

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m-1)}[\boldsymbol{\tau}^{(m-1)}] + \mathbf{U}^{(m-1)} &= \sum_{k=1}^p \sum_{i=1}^n \frac{x_{ij}x_{ik}}{Var(Y_i)^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \hat{\beta}_k^{(m-1)} + \sum_{i=1}^n \left[\frac{(y_i - \mu_i)}{Var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \\ &= \mathbf{X}^T \mathbf{W} \mathbf{z},\end{aligned}\quad (2.22)$$

onde \mathbf{z} tem elementos

$$z_i = \sum_{j=1}^k x_{ij} \hat{\boldsymbol{\beta}}_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right). \quad (2.23)$$

Logo, as equações iterativas dadas em 2.20 podem ser reescritas como

$$\hat{\boldsymbol{\beta}}^{(m)} = \frac{\mathbf{X}^T \mathbf{W} \mathbf{z}}{\mathbf{X}^T \mathbf{W} \mathbf{X}}. \quad (2.24)$$

O processo iterativo apresentado na Equação 2.24 é repetido até que o valor de $\hat{\boldsymbol{\beta}}$ convirja, ou seja, até que a diferença entre as consecutivas aproximações de $\hat{\boldsymbol{\beta}}^{(m-1)}$ e $\hat{\boldsymbol{\beta}}^{(m)}$ seja suficientemente pequena.

2.5.2 Abordagem Bayesiana

De acordo com Casella (2001), na abordagem Bayesiana, o parâmetro de interesse θ é considerado uma quantidade aleatória cuja variabilidade pode ser descrita por uma distribuição de probabilidade, chamada de distribuição *a priori* e denotada por $\pi(\theta)$. Esta é uma distribuição subjetiva, baseada na crença do experimentador, e é estabelecida antes que os dados sejam observados. Considerando um vetor de parâmetros de interesse $\boldsymbol{\theta}$, a distribuição *a priori* conjunta é denotada por $\pi(\boldsymbol{\theta})$.

A inferência Bayesiana combina o conhecimento prévio acerca do vetor paramétrico $\boldsymbol{\theta}$ com a informação proveniente dos dados observados, resumida através da função de verossimilhança $L(\boldsymbol{\theta}; y)$. Os parâmetros são, portanto, estimados com base na distribuição *a posteriori* conjunta, denotada por $\pi(\boldsymbol{\theta}|y)$. Considerando o Teorema de Bayes, tem-se que:

$$\pi(\boldsymbol{\theta}|y) = \frac{\pi(\boldsymbol{\theta})L(\boldsymbol{\theta}; y)}{p(y)}. \quad (2.25)$$

Note que, como $p(y)$ não depende de $\boldsymbol{\theta}$, a Equação 2.25 pode ser reescrita como

$$\pi(\boldsymbol{\theta}|y) \propto \pi(\boldsymbol{\theta})L(\boldsymbol{\theta}; y).$$

No entanto, realizar inferência diretamente da distribuição a *posteriori* conjunta $\pi(\boldsymbol{\theta}|y)$ pode ser uma tarefa árdua devido à dimensão do vetor paramétrico $\boldsymbol{\theta}$. Para contornar essa dificuldade, em geral, considera-se as distribuições condicionais completas a *posteriori*. Assuma $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)'$ onde cada um dos componentes $\boldsymbol{\theta}_j$ ($j = 1, \dots, d$) pode ser um escalar, um vetor ou uma matriz. As distribuições condicionais completas a *posteriori* são definidas por $\pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j})$, onde $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \dots, \boldsymbol{\theta}_d)$.

Mesmo considerando as distribuições condicionais completas, o procedimento de inferência ainda pode ser complexo. Como estas distribuições podem não ter forma conhecida, pode ser difícil ou até impossível encontrar as estimativas para os parâmetros, como a média ou intervalo de credibilidade, a partir das distribuições condicionais completas. Sendo assim, é comum utilizar o Método de Monte Carlo via Cadeia de Markov (MCMC), que é um método computacional iterativo para facilitar a estimação dos parâmetros de interesse. Os principais algoritmos deste método são o Amostrador de Gibbs e o de Metropolis-Hastings, explicitados a seguir.

- **Amostrador de Gibbs:** De acordo com Geman e Geman (1984), o algoritmo do amostrador de Gibbs gera amostras da distribuição conjunta de interesse diretamente das distribuições condicionais completas. Para isso, portanto, é necessário que as condicionais completas sejam conhecidas. O algoritmo pode ser descrito da seguinte forma:

1. Inicializar o contador de iteração da cadeia $j = 1$ e definir os valores iniciais $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})'$;
2. Obter um novo valor $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \dots, \theta_d^{(j)})'$ por $\boldsymbol{\theta}^{(j-1)}$ pela sucessiva geração de distribuições condicionais completas:

$$\begin{aligned}\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \dots, \theta_d^{(j-1)}), \\ \theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_d^{(j-1)}), \\ &\vdots \\ \theta_d^{(j)} &\sim \pi(\theta_d|\theta_1^{(j)}, \dots, \theta_{d-1}^{(j)});\end{aligned}$$

3. Atualizar o contador $j = j + 1$ e voltar para o passo 2 até que a convergência seja atingida.

A convergência da cadeia pode ser analisada visualmente através do gráfico de traços do método MCMC. Além disso, pode-se adotar mais de uma cadeia iniciando de pontos distintos e verificar se ambas convergem para o mesmo valor.

Este processo da Cadeia de Markov, ou seja, a repetição desse algoritmo até que a convergência seja atingida, gera valores que permitem formar uma amostra da distribuição conjunta de $\boldsymbol{\theta}$. Desta forma, portanto, é possível obter as estatísticas do parâmetro a partir desta amostra.

- **Metropolis-Hastings:** O algoritmo de Metropolis-Hastings foi desenvolvido inicialmente por Metropolis (1953) e aprimorado por Hastings (1970) para ser utilizado caso a geração não iterativa da distribuição de interesse fosse muito complexa ou cara. E, diferentemente do Amostrador de Gibbs, este método pode ser utilizado quando uma ou mais distribuições condicionais completas a *posteriori* forem desconhecidas.

Para inicializar o algoritmo, precisa-se escolher uma distribuição proposta, denominada por $q(\cdot)$, da qual seja mais simples obter amostras. Em seguida, por meio de um processo iterativo, decide-se se o vetor de valores amostrados $\boldsymbol{\phi}$ através distribuição proposta deve ser incluído ou não na cadeia. Em geral, opta-se por distribuições que dependam dos valores atuais da cadeia. Dessa forma, a distribuição proposta transita até convergir para a distribuição de interesse.

De acordo com Gamerman e Lopes (2006), para saber se o novo vetor de valores será incluído na cadeia, Hastings (1970) propôs definir a probabilidade de aceitação deste vetor de forma que, quando combinada com a distribuição proposta, aceitasse o novo valor proposto e, por consequência movimentasse a cadeia, ou rejeitasse tal valor de forma que a cadeia não avançasse. Sendo assim, a expressão mais comum para a probabilidade de aceitação pode ser escrita como:

$$a(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min \left\{ 1, \frac{p(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})}{p(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})} \right\}, \quad (2.26)$$

onde $p(\cdot)$ denota a distribuição de interesse, que no caso deste trabalho é a distribuição a *posteriori*, $\boldsymbol{\phi}$ é o novo vetor de valores sorteados pelo algoritmo, ou seja,

o vetor com os valores propostos. O algoritmo pode ser executado pelo seguinte passo-a-passo:

1. Inicializar o contador como $j = 1$ e escolher um valor inicial arbitrário para $\boldsymbol{\theta}^{(0)}$;
2. Gerar um novo valor ϕ por $q(\boldsymbol{\theta}^{(j-1)}, \cdot)$;
3. Avaliar a probabilidade de aceitação da transição $a(\boldsymbol{\theta}^{(j-1)}, \phi)$, dada pela Equação 2.26. Para isso, é sorteado um valor u de uma distribuição uniforme independente. Se $u \leq a$, a transição é aceita e $\boldsymbol{\theta}^{(j)} = \phi$, caso contrário, a cadeia não se move e $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$;
4. Alterar o contador para $j+1$ e retornar para o passo 2 até que a cadeia converja.

A convergência pode ser analisada através do gráfico de traços do método MCMC.

Este processo da Cadeia de Markov, ou seja, a repetição desse algorítimo até que a convergência seja atingida, gera valores que permitem formar uma amostra da distribuição a *posteriori* de $\boldsymbol{\theta}$. Desta forma, portanto, é possível obter as estatísticas do parâmetro a partir desta amostra.

Entretanto, existem dois fatores que devem ser ponderados em relação a estes algoritmos: a amostra da distribuição a *posteriori*, gerada pelo MCMC depende dos valores iniciais escolhidos para os parâmetros, o que eventualmente pode gerar estimativas ruins no início do algoritmo. Além disso, os valores de θ são correlacionados, pois são gerados pelo processo da cadeia de Markov, onde as estimativas na iteração atual $\boldsymbol{\theta}^{(j)}$ dependem das estimativas da iteração anterior $\boldsymbol{\theta}^{(j-1)}$.

Sendo assim, torna-se essencial a utilização do período de aquecimento (*burn-in*), onde a parte inicial da cadeia é descartada, de forma que os valores iniciais definidos tenham influência reduzida sobre as estimativas. Caso haja autocorrelação significativa entre as amostras subsequentes dos componentes do vetor paramétrico $\boldsymbol{\theta}$, isso pode ser solucionado utilizando um espaçamento na cadeia, onde considera-se apenas 1 a cada k amostras e descartando as restantes.

Uma vez que o procedimento de inferência foi estabelecido e os métodos de estimação dos parâmetros foram apresentados, é essencial avaliar se o modelo proposto consegue se adequar ao conjunto de dados analisado. Em geral, são considerados alguns modelos que são julgados apropriados e posteriormente é realizada a comparação entre eles. Na

próxima seção, serão apresentadas as medidas utilizadas para a comparação de modelos adotadas nesse trabalho.

2.6 Comparação dos Modelos

Após o ajuste de diversos modelos, é de extrema importância que seja verificado qual modelo se adequa melhor aos dados. Para isso, existem alguns métodos bastante utilizados na literatura estatística. Neste trabalho, foi considerado o AIC, apresentado na Subseção 2.6.1, como medida de comparação entre os modelos frequentistas e o DIC, apresentado na Subseção 2.6.2, como medida de comparação para os modelos Bayesianos.

2.6.1 AIC (Akaike Information Criterion)

O AIC, proposto originalmente por Akaike (1973), é uma medida baseada na função de verossimilhança, definida por:

$$AIC = -2\ln(L(\hat{\theta}; \mathbf{y})) + 2d,$$

onde $\ln(L(\hat{\theta}; \mathbf{y}))$ é a função de log-verossimilhança avaliada nos parâmetros estimados ($\hat{\theta}$) e d é o número de parâmetros.

Para a comparação de modelos, pode-se dizer que um valor mais baixo, significa um melhor ajuste. Para mais informações sobre o critério de informação de Akaike, consultar Bozdogan (1987).

2.6.2 DIC (Deviance Information Criterion)

De acordo com Spiegelhalter, Best, Carlin e Linde (2002), o critério de informação da deviance é capaz de identificar modelos que melhor expliquem os dados observados e pode ser escrita como:

$$DIC = \bar{D} + 2p_D,$$

onde \bar{D} é o desvio esperado a *posteriori* e p_D é dado por $\overline{D(\theta)} - D(\bar{\theta})$ (média da deviance Bayesiana menos a deviance da média da *posteriori* dos parâmetros).

Assim como o AIC, quanto menor o valor desta medida, melhor o modelo se adequa aos dados. Mais informações sobre o DIC podem ser encontradas em Spiegelhalter, Best,

Carlin e Linde (2002).

3 Resultados

Inicialmente, para verificar a capacidade dos modelos propostos em determinar os verdadeiros valores dos parâmetros, foi elaborado um estudo com dados simulados através do *software R Core Team* (2021). Os pacotes utilizados para tal estudo foram o *stats*, *R2jags* e o *mcmcplots*.

3.1 Dados Simulados

Para realizar o estudo com dados simulados, foram gerados dois conjuntos de dados, ambos de tamanho amostral igual a 1.000, considerando duas variáveis explicativas X_1 e X_2 , com distribuições escolhidas de forma arbitrária, tais que $X_1 \sim N(5, 2)$ e $X_2 \sim Bin(0.3)$. Para o primeiro conjunto de dados, a variável independente Y foi gerada através da distribuição Log-Normal tal que $\log(Y) \sim N(\mu, \sigma^2)$, onde $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ e $\sigma^2 = 1$. Já para o segundo conjunto de dados, foi considerado que $Y \sim Gama\left(\alpha, \frac{\alpha}{\mu}\right)$, onde $\mu = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ e $\alpha = 2$. Em ambos os casos, foi considerado o vetor de coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T = (0.5, 0.2, -0.1)^T$, definido também de forma arbitrária, apenas atentando para que a média da distribuição Gama resultasse em um valor positivo, em razão de suas propriedades.

Os parâmetros dos modelos acima foram estimados tanto pelo método frequentista quanto pelo método Bayesiano. Para o ajuste de ambos conjuntos de dados, foram adotados os modelos Log-normal e Gama.

Para o procedimento de inferência Bayesiana, foram estabelecidas as seguintes distribuições a *priori*: $\pi(\beta_j) \sim N(0, 100)$ para os coeficientes β_j ($j = 0, 1, 2$); $\pi(\tau) \sim Gama(0.01, 0.01)$ para o parâmetro de precisão $\tau = \frac{1}{\sigma^2}$ do modelo Log-Normal; e $\pi(\alpha) \sim U(0, 100)$ para o parâmetro de forma α da distribuição Gama, conforme a parametrização definida na Equação 2.6.

Além disso, nos métodos de MCMC, foram consideradas 2 cadeias com valores iniciais distintos. Para cada cadeia, foram realizadas 40.000 iterações, com período de aquecimento (*burn-in*) igual a 10.000, ou seja, as 10.000 primeiras iterações foram descartadas para que os valores iniciais não afetem as estimativas. Após o *burn-in*, foi adotado um espaçamento de 30 iterações para evitar valores altamente correlacionados, ou seja, foram consideradas apenas 1 a cada 30 iterações. Ao final, obteve-se uma amostra de tamanho 1.000 das estimativas dos parâmetros em cada cadeia.

Os resultados obtidos para ajuste dos dados simulados gerados a partir da distribuição Log-Normal e da Gama serão apresentados nas Subseções 3.1.1 e 3.1.2, respectivamente.

3.1.1 Resultados obtidos para o conjunto de dados 1

Inicialmente, considerou-se um conjunto de dados cuja variável independente Y segue uma distribuição Log-Normal. Para a modelagem desses dados, foram adotados os modelos Log-Normal e Gama com função de ligação logarítmica.

A Tabela 2 apresenta as estimativas para o vetor de coeficientes β obtidas através da inferência frequentista (média e intervalo de 95% de confiança) e da inferência Bayesiana (média a *posteriori* e intervalo de 95% de credibilidade). Pode-se perceber que as estimativas Bayesianas e frequentistas são muito próximas. Isto ocorre pelo fato de que, no método Bayesiano, as funções *a priori* escolhidas para os parâmetros do modelo, foram *priori*s não-informativas. De acordo com Berger (1985), *priori*s não-informativas são *priori*s que não possuem informação sobre θ , ou, mais informalmente, que não favorecem um valor de θ sobre outros. Segundo Paulino, Turkman, Murteira e Silva (2018), tal função, conhecida também como *priori* vaga, permite que informações da *posteriori* sejam deduzidas, mesmo quando o conhecimento sobre os dados é escasso.

Valor dos Coeficientes	Ajuste	Estimação Frequentista	Estimação Bayesiana
$\beta_0 = 0.5$	Log-Normal	0.660 (0.49;0.83)	0.657 (0.48;0.83)
		0.172 (0.14;0.20)	0.173 (0.14;0.20)
$\beta_1 = 0.2$	Gama	-0.170 (-0.30;-0.04)	-0.171 (-0.30;-0.04)
		1.204 (1.00;1.41)	1.205 (1.05;1.36)
$\beta_2 = -0.1$		0.160 (0.12;0.20)	0.160 (0.13;0.19)
		-0.197 (-0.36;-0.03)	-0.193 (-0.31;-0.06)

Tabela 2: Estimativas frequentistas e bayesianas de dados simulados considerando uma variável resposta com distribuição Log-Normal.

Sobre a estimativa pelo método Bayesiano, foram gerados gráficos de densidade das amostras referentes às distribuições a *posteriori* dos parâmetros, tanto para o ajuste com o modelo Log-Normal, quanto para o ajuste com o modelo Gama, que podem ser observados respectivamente nas Figuras 2 e 3. As linhas azuis e vermelhas do gráfico, representam um intervalo de credibilidade de 95% de cada cadeia e a linha preta tracejada, o valor real de cada parâmetro. Em ambas figuras, pode-se perceber que a distribuição a *posteriori* dos parâmetros β_0 , β_1 e β_2 parecem seguir distribuição normal. Percebe-se, portanto, na Figura 2 que o valor real se encontra dentro do intervalo de credibilidade para todos os parâmetros, diferente da Figura 3, onde os intervalos de credibilidade de β_0 e β_1 não contém o valor real dos mesmos.

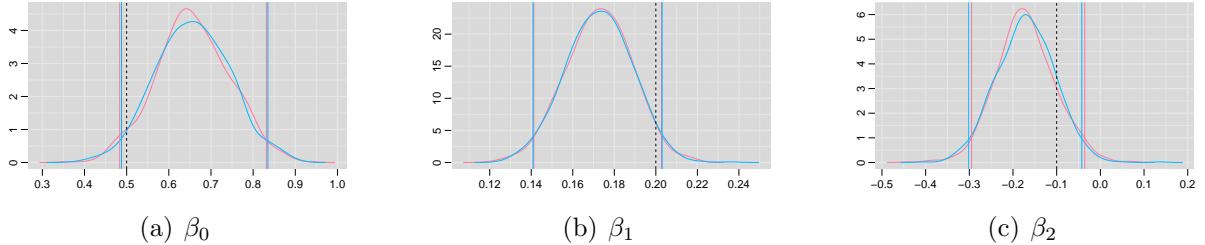


Figura 2: Gráfico da distribuição de densidade das funções a *posteriori* dos betas, com dados gerados e ajustados por uma distribuição Log-Normal.

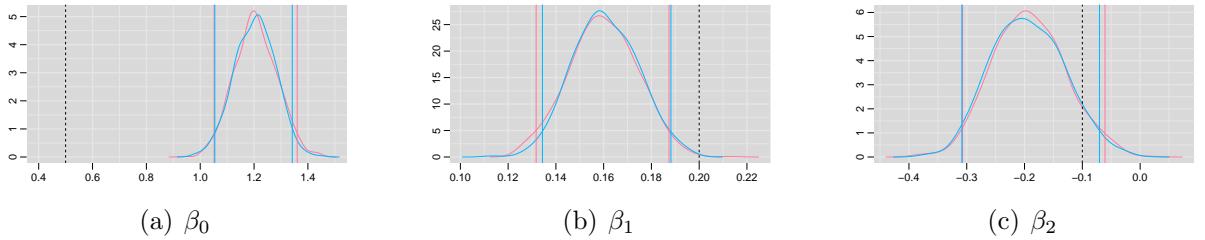


Figura 3: Gráfico da distribuição de densidade das funções a *posteriori* dos betas, com dados gerados por uma distribuição Log-Normal e ajustados por uma Gama.

Além disso, foram gerados gráficos de iteração das cadeias de Markov a fim de verificar se os mesmos cumpriram a condição de convergência determinado pelo método de MCMC, explicitado na Subseção 2.5.2. Os gráficos para o ajuste Log-Normal pode ser observado na Figura 4, já os gráficos para o ajuste Gama com ligação logarítmica, podem ser observados na Figura 5. Em ambas as Figuras, a linha preta tracejada representa o valor real do parâmetro, enquanto que as linhas azuis e vermelhas do gráfico, representam os intervalos de credibilidade de 95% de cada cadeia. Percebe-se que, na 4, para os três parâmetros

$(\beta_0, \beta_1$ e $\beta_2)$, os traços estão concentrados em torno de um mesmo valor do parâmetro ao longo das iterações, confirmando a convergência da cadeia. Mas, ao observar a Figura 5, apesar da convergência da cadeia para os três parâmetros, fica claro que os traços não convergem para o valor real de β_0 .

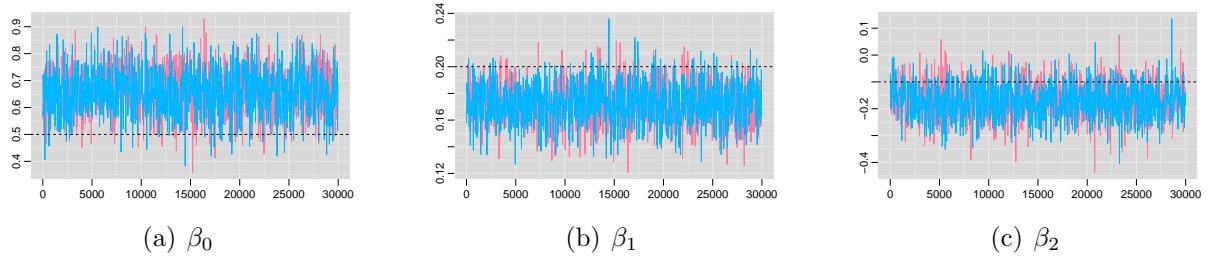


Figura 4: Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados e ajustados por uma distribuição Log-Normal.

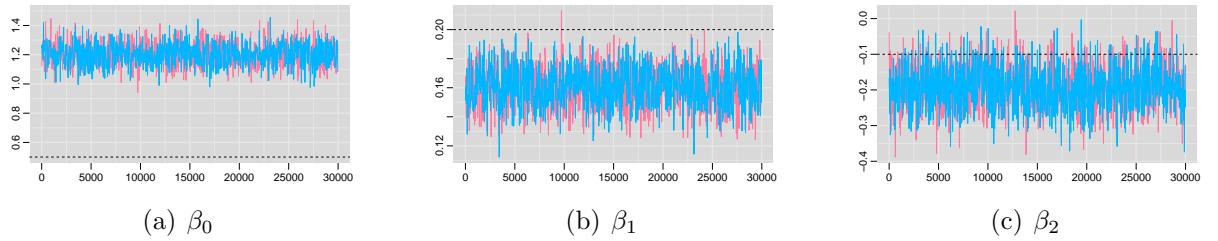


Figura 5: Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados por uma distribuição Log-Normal e ajustados por uma Gama.

Além dos resultados exibidos na Tabela 2, também foram obtidos o AIC e o DIC tanto pelo ajuste pela função Log-Normal quanto pela função Gama com função de ligação logarítmica. Para o ajuste pela Log-Normal, foi obtido um AIC de 2810, já pela Gama, o resultado foi de 5900. Sobre o DIC, foram observados os valores de 5756 e 5890 para os ajustes Log-Normal e Gama, respectivamente. Ou, seja, para dados gerados por uma distribuição Log-Normal, ambas as medidas apresentaram valores menores para ajuste Log-Normal. Além disso, os intervalos de credibilidade do modelo com ajuste Log-Normal são os únicos que contêm os verdadeiros valores dos parâmetros. Desta forma, concluímos que este modelo se adequa melhor ao conjunto de dados quando comparado ao modelo com ajuste Gama.

3.1.2 Resultados obtidos para o conjunto de dados 2

Em seguida, considerou-se o segundo conjunto de dados, cuja variável independente Y foi gerada através de uma distribuição Gama utilizando a função de ligação logarítmica. Este conjunto de dados foi ajustado considerando o modelo Gama com função de ligação logarítmica e também o modelo Log-Normal. As estimativas frequentistas (média e intervalo de 95% confiança) e Bayesianas (média e intervalo de 95% credibilidade) dos ajustes também podem ser observados na Tabela 3. Nota-se que os resultados para os ajustes frequentista e Bayesiano são semelhantes. Como foram utilizadas *prioris* vagas, isso já era esperado, dados os motivos explicitados na Subseção 3.1.1.

Valor dos Coeficientes	Ajuste	Estimação Frequentista	Estimação Bayesiana
$\beta_0 = 0.5$	Gama	0.471 (0.35;0.59)	0.469 (0.35;0.59)
		0.207 (0.18;0.23)	0.207 (0.19;0.23)
		-0.150 (-0.24;-0.05)	-0.147 (-0.24;-0.05)
$\beta_1 = 0.2$	Log-Normal	0.193 (0.06;0.33)	0.200 (0.06;0.34)
		0.207 (0.18;0.23)	0.205 (0.18;0.23)
		-0.095 (-0.20;-0.01)	-0.090 (-0.21;0.00)

Tabela 3: Estimativas Frequentistas e Bayesianas de Dados Simulados considerando uma Variável Resposta com Distribuição Gama.

Os gráficos de densidade das distribuições a *posteriori* dos parâmetros para os modelo com ajuste Gama e ajuste Log-Normal, podem ser observados nas Figuras 6 e 7, respectivamente. Nota-se que todos parecem apresentar distribuição Normal, como esperado. Além disso, observando a Figura 6, notou-se que o intervalo de credibilidade dos três parâmetros de interesse (β_0 , β_1 e β_2) contém o valor real dos mesmos, enquanto ao observar a Figura 7, é possível perceber que isso não ocorre para β_0 .

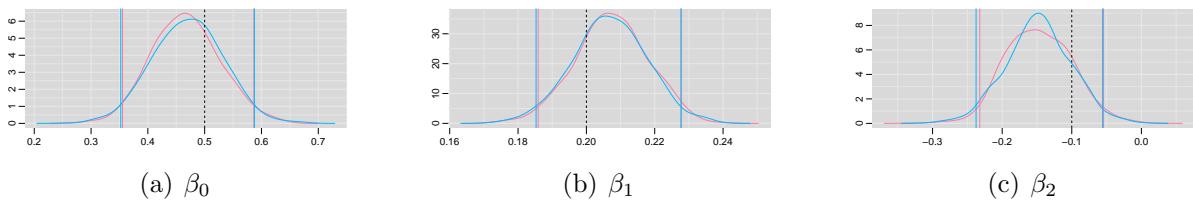


Figura 6: Gráfico da distribuição de densidade das funções a *posteriori* dos betas, com dados gerados e ajustados por uma distribuição Gama.

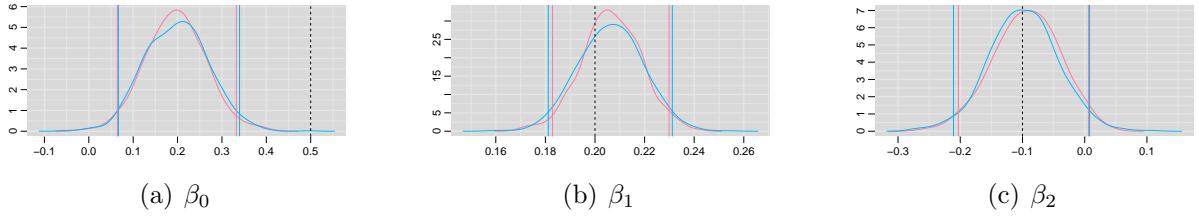


Figura 7: Gráfico da distribuição de densidade das funções a posteriori dos betas, com dados gerados por uma distribuição Gama e ajustados por uma Log-Normal.

Já os gráficos da iteração das cadeias de Markov podem ser observados nas Figuras 8 e 9 para os ajustes Gama e Log-Normal, respectivamente. Pode-se observar que em ambas as figuras, os traços estão concentrados em torno de um mesmo valor do parâmetro ao longo das iterações, cumprindo a condição de convergência determinada pelo método MCMC.

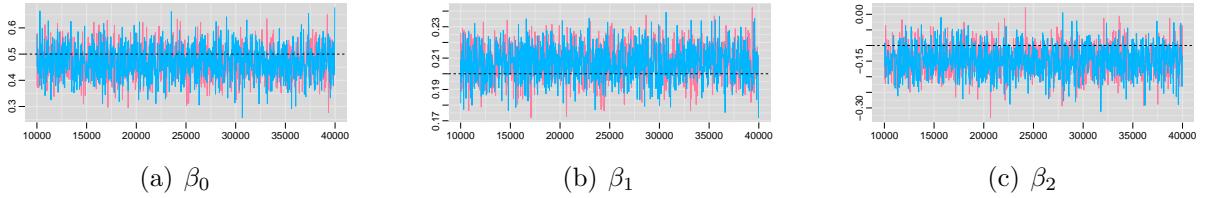


Figura 8: Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados e ajustados por uma distribuição Gama.

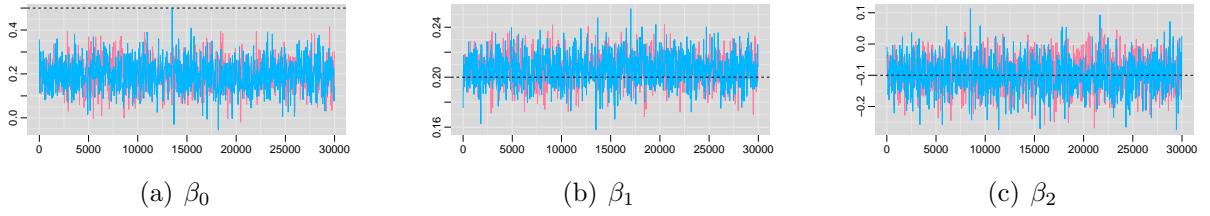


Figura 9: Gráfico de traços da iteração das cadeias para os parâmetros de dados gerados por uma distribuição Gama e ajustados por uma Log-Normal.

Além disso, foram calculadas as medidas AIC e DIC para cada ajuste. O AIC encontrado para os ajustes Gama e Log-Normal foram, respectivamente, 4687 e 2373. Já o DIC para o Gama foi de 4684 e para o Log-Normal, 4785. Neste caso, apesar de o modelo com ajuste Log-Normal possuir um AIC bem menor, o modelo com ajuste Gama possui DIC menor, além de conter o verdadeiro valor do parâmetro em todos os intervalos

de confiança e de credibilidade, o que nos leva a crer que o modelo com ajuste Gama se adequa melhor aos dados, como esperado.

Na próxima Seção, será apresentada uma análise exploratória do conjunto de dados reais a fim de esclarecer e resumir as principais características das variáveis utilizadas na construção dos modelos propostos para a variável de interesse valor do prêmio.

3.2 Análise exploratória dos dados reais

Ao explorar a base de dados fornecida pela seguradora, foi constatado que 2.176 apólices não informavam a categoria do veículo assegurado, o equivalente a 4,3% da base original. Com a intenção de facilitar as análises, estas observações foram descartadas. Sendo assim, neste trabalho, foram consideradas as informações de um conjunto de dados composto por 46.326 apólices de seguros de automóveis.

Conforme citado na Seção 2.1, a área de estudo é a cidade do Rio de Janeiro, cujos bairros estão agrupados em 5 áreas de planejamento. A Figura 10 apresenta um mapa da cidade do Rio de Janeiro segundo o valor médio do prêmio em cada área de planejamento, calculado a partir da base de dados analisada. As regiões com montantes mais altos para o valor médio do prêmio são, respectivamente, as áreas de planejamento 3, 4, 1, 2 e 5.

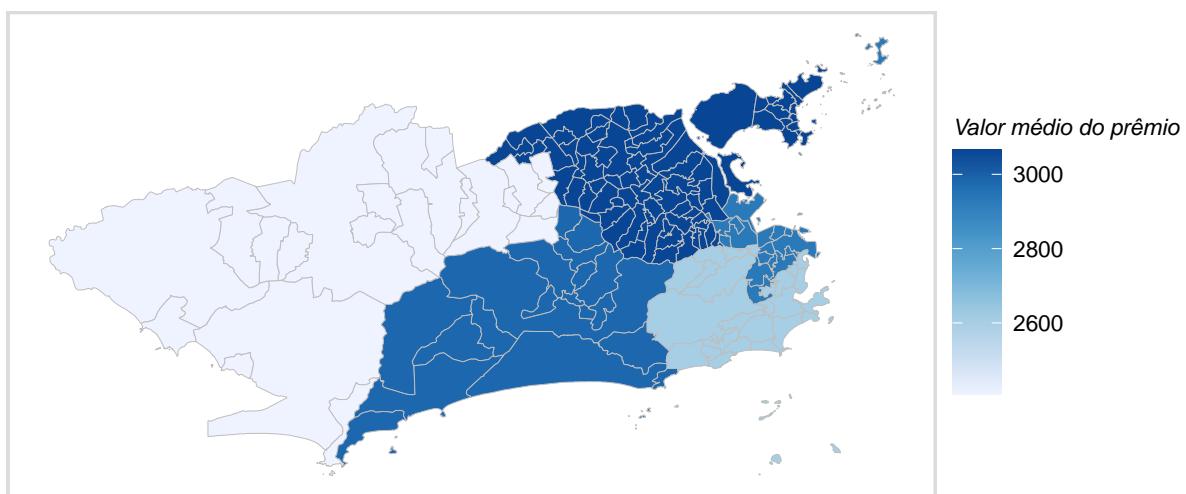


Figura 10: Mapa da cidade do Rio de Janeiro segundo o valor médio do prêmio por área de planejamento de acordo com a base de dados em estudo.

O banco de dados contém informações sobre o sexo, o estado civil e a idade (em anos)

do segurado, bem como informações sobre a região de residência conforme as áreas de planejamento, a categoria do automóvel assegurado e o valor do prêmio (em reais). A Tabela 4 apresenta as medidas descritivas dessas características presentes nas apólices. Para as variáveis qualitativas, são apresentadas as frequências absoluta e relativa; para as variáveis quantitativas, a média e desvio padrão.

Vale ressaltar que, os indivíduos com idade superior a 75 anos foram agrupados pela própria seguradora. A fim de realizar o cálculo para a média e desvio-padrão, considerou-se a idade de 76 anos em todas apólices cujos segurados tinham idade acima de 75 anos. A Figura 11 apresenta o histograma para a idade dos segurados, onde a linha tracejada em vermelho representa a média dessa variável. Pelo histograma, observa-se um pico entre as idades 31 e 40 anos, o que significa que este intervalo de idade é o mais comum na base de dados em estudo.

A Figura 12 apresenta um diagrama de caixas (*boxplot*) com os valores dos prêmios em reais, assim como a comparação deste valor dentro dos grupos das variáveis sexo e estado civil. A mediana do prêmio está próxima de R\$1.300,00, ou seja, 50% das observações possuem valores de prêmio menores que R\$1.300. Além disso, é possível observar a presença de muitos *outliers*, o que eleva o preço médio do prêmio. Dentro dos grupos da variável sexo, é possível observar que os valores de prêmio são um pouco mais altos para o sexo masculino, assim como sua variabilidade. Por outro lado, em relação ao estado civil, não parece haver diferenças entre os valores dos prêmios para os grupos de segurados solteiros e casados.

Já a Figura 13 mostra os *boxplots* dos valores dos prêmios de acordo com as categorias de veículos e, através dela, fica evidente que os automóveis do tipo “Coupe/Roadster”

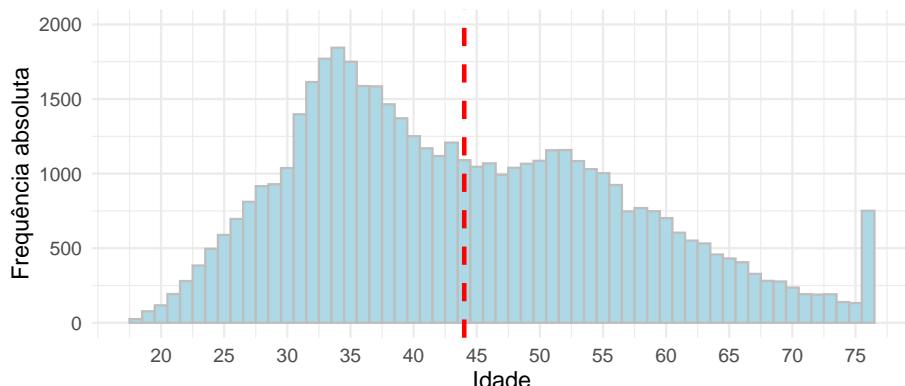


Figura 11: Histograma para a variável idade. A linha tracejada corresponde à média.

Tabela 4: Medidas descritivas das características presentes em uma carteia de apólices de seguros de automóveis.

Características	N=46.326
Sexo, no. masculino (%)	29.275 (63,19%)
Estado civil, no. solteiro (%)	13.020 (28,11%)
Idade, média (dp)	44 (12.85)
Categoria do veículo, no. (%)	
Coupe/Roadster	83 (0,18%)
Crossover	1.182 (2,55%)
Hatchback Compacto	15.259 (32,94%)
Hatchback Medio	3.219 (6,95%)
Jipe	11 (0,02%)
Minivan Compacto	3.300 (7,12%)
Minivan Médio ou Grande	650 (1,40%)
Multivan/Furgão	307 (0,66%)
Picape Cabine Dupla	1.531 (3,30%)
Picape	625 (1,35%)
Sedan Compacto	8.369 (18,06%)
Sedan Médio ou Grande	5.361 (11,57%)
Station-Wagon Compacto	420 (0,91%)
Station-Wagon Médio ou Grande	318 (0,69%)
SUV Compacto	1.792 (3,87%)
SUV Médio ou Grande	3.899 (8,42%)
Área de Planejamento, no. (%)	
Área de planejamento 1	4.269 (9,22%)
Área de planejamento 2	9.214 (19,89%)
Área de planejamento 3	13.421 (28,97%)
Área de planejamento 4	15.002 (32,38%)
Área de planejamento 5	4.420 (9,54%)
Prêmio, média (dp)	2.319,88 (1.812,57)

possui valores de prêmio muito mais elevados que os restantes. Destaca-se também, os valores de prêmio para veículos do tipo “ Picape Cabine Dupla”, cuja mediana é muito superior às demais categorias de automóveis, com exceção de “Coupe/Roadster”.

Por fim, a Figura 14 expõe os *boxplots* do valor do prêmio de acordo com as áreas de planejamento. Percebe-se que as áreas 2 e 5 apresentam ter valores de prêmio menores do que as restantes. Além disso, a Área de Planejamento 5 parece ter variabilidade pouco menor do prêmio.

Através das Figuras 12, 13 e 14, pode-se perceber a presença de forte assimetria à direita na distribuição dos valores do prêmio, tanto de forma geral, considerando todas as

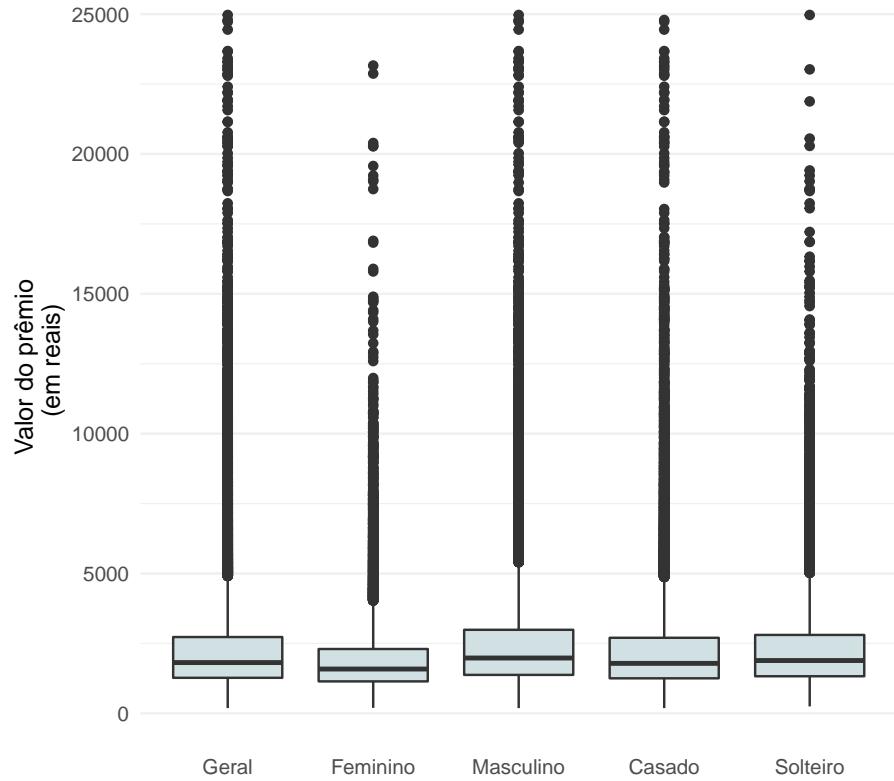


Figura 12: *Boxplot* do valor do prêmio na carteira de apólices (geral) e segregados de acordo com o sexo e o estado civil dos segurados.

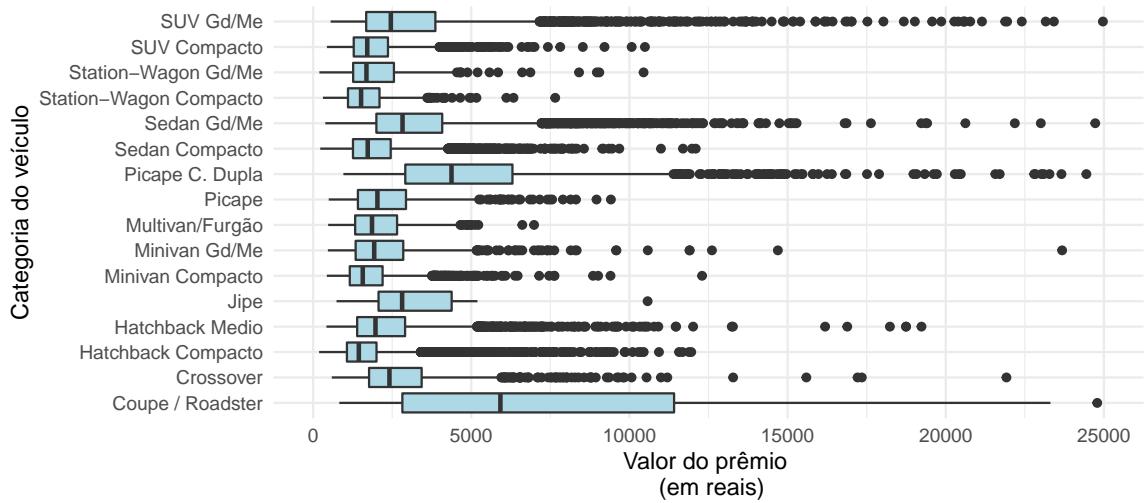


Figura 13: *Boxplot* do valor do prêmio de acordo com a categoria do veículo

apólices, quanto dentro das categorias das variáveis disponíveis. Isso é um indicativo que os modelos de regressão múltipla podem não ser adequados a esse conjunto de dados. Na próxima seção, serão apresentados os modelos ajustados para os valores do prêmio.

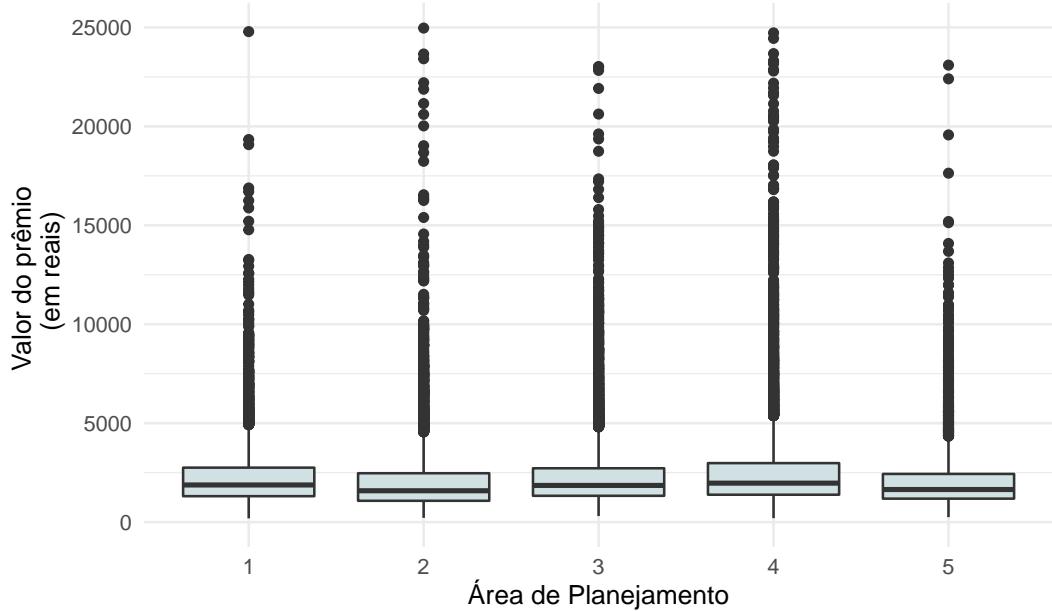


Figura 14: *Boxplot* do valor do prêmio de acordo com a área de planejamento

3.3 Modelagem do valor do prêmio

Propositalmente, os dados foram ajustados por um modelo de regressão múltipla, baseado em uma distribuição Normal, de acordo com a abordagem frequentista, vista na Seção 2.2. Neste contexto, a variável resposta Y_i representa o valor do prêmio. Portanto, a média dos prêmios será modelada por $E[Y_i] = \mu_i = \beta_0 + \beta_1$ Idade + β_2 Sexo Masculino + β_3 Estado Civil Solteiro + β_4 Categoria do Automóvel + β_5 Área de Planejamento, sendo as variáveis categoria do automóvel e área de planejamento varáveis *dummies*.

As estimativas dos coeficientes encontrados pelo modelo de regressão múltipla, assim como seus respectivos erros, podem ser encontrados no Apêndice E. Para determinar se existe alguma relação significativa entre o valor do prêmio e cada características do segurado e do veículo assegurado, foram realizados testes de hipóteses t, sendo H_0 a hipótese nula de que $\beta_j = 0$ e H_1 a hipótese alternativa de que $\beta_j \neq 0$ ($j = 0, \dots, 5$). Se o p-valor do teste for menor ou igual ao nível de significância α , então a hipótese nula é rejeitada e pode-se inferir que a variável X_j tem influência significativa na especificação do seguro. Os p-valores dos coeficientes do modelo também encontram-se no Apêndice E. De acordo com o teste t, considerando um nível de significância α igual a 5%, foi observado um p-valor menor para todas as variáveis, concluindo portanto que todas as características consideradas nesse trabalho são significativas.

Para verificar a adequação dos dados aos pressupostos do modelo de regressão múltipla, foram elaborados gráficos de resíduos padronizados *versus* os valores ajustados do prêmio e também um gráfico dos quantis amostrais *versus* os quantis teóricos da distribuição Normal, ambos apresentados na Figura 15. Observando o gráfico de resíduos, nota-se que a variância não é constante, pois à medida que o valor do prêmio ajustado aumenta, os resíduos ficam mais dispersos em torno do 0, ou seja, a suposição de homocedasticidade é violada. Além disso, é possível observar a presença de *outliers*, devido aos pontos distantes do restante da distribuição. Já pelo gráfico quantil-quantil (*Q-Q Plot*), apresentado à direita na Figura 15, percebe-se que os resíduos não seguem uma distribuição Normal e apresentam forte assimetria à direita. Portanto, como esperado, o modelo de regressão múltipla não é ideal para modelar o valor do prêmio.

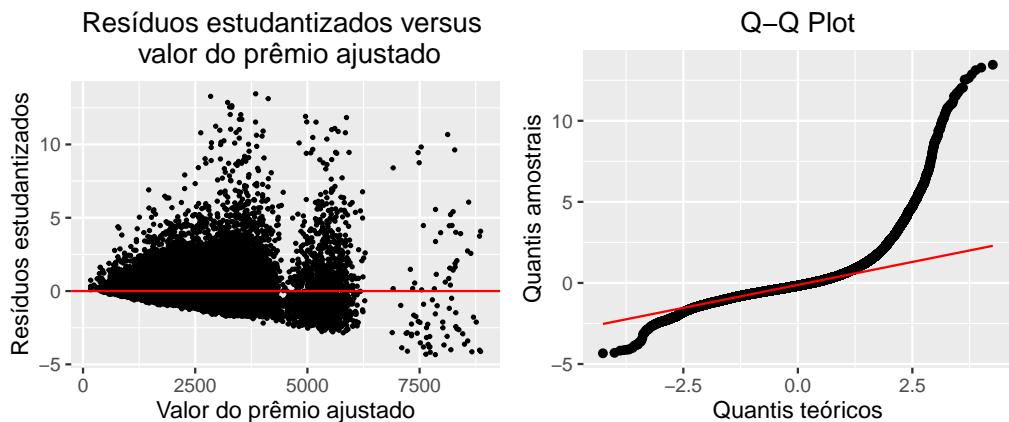


Figura 15: Gráfico de resíduos e Q-Q Plot para o modelo de regressão múltipla

Sendo assim, foram adotados os modelos lineares generalizados para modelar o valor do prêmio. Em especial, foram considerados os modelos Log-Normal e Gama, com função de ligação logarítmica, que são os mais apropriados para ajustar dados assimétricos à direita. Para fins de comparação, foram realizados os ajustes segundo as abordagens frequentista e Bayesiana. Os resultados serão apresentados em sequência.

3.3.1 Ajuste Frequentista

Os resultados dos ajustes frequentistas para o conjunto de dados fornecidos pela seguradora podem ser encontrados na Tabela 5. Nota-se que as estimativas pontuais e os intervalo de confiança são semelhantes entre os modelos Log-Normal e Gama. Inclusive, percebe-se que todos os coeficientes negativos estão relacionados às mesmas variáveis em

ambos modelos.

Ajuste Variável	Log-Normal Média (IC)	Gama Média (IC)
Intercepto	7.843 (7.819;7.867)	7.933 (7.904;7.961)
Idade	-0.011 (-0.012;-0.011)	-0.011 (-0.011;-0.010)
Sexo Masculino	0.175 (0.166;0.185)	0.179 (0.168;0.190)
Estado Civil Solteiro	0.087 (0.077;0.097)	0.082 (0.069;0.094)
Categoria do veículo		
Sedan Compacto	-	-
Coupe/Roadster	1.200 (1.096;1.302)	1.411 (1.289;1.537)
Crossover	0.411 (0.382;0.440))	0.422 (0.387;0.457)
Hatchback Compacto	-0.186 (-0.200;-0.173)	-0.188 (-0.203;-0.172)
Hatchback Medio	0.134 (0.114;0.153)	0.165 (0.142;0.188)
Jipe	0.554 (0.272;0.836)	0.672 (0.352;1.031)
Minivan Compacto	-0.044 (-0.063;-0.025)	-0.060 (-0.083;-0.037)
Minivan Médio ou Grande	0.191 (0.153;0.229)	0.239 (0.194;0.285)
Multivan/Furgão	0.107 (0.053;0.161)	0.108 (0.044;0.174)
Picape Cabine Dupla	0.941 (0.914;0.967)	0.994 (0.963;1.026)
Picape	0.179 (0.140;0.218)	0.201 (0.154;0.247)
Sedan Médio ou Grande	0.509 (0.501;0.518)	0.535 (0.515;0.555)
Station-Wagon Compacto	-0.072 (-0.119;-0.025)	-0.079 (-0.135;-0.023)
Station-Wagon Médio ou Grande	0.056 (0.003;0.110)	0.092 (0.028;0.157)
SUV Compacto	0.045 (0.021;0.070)	0.030 (0.000;0.059)
SUV Médio ou Grande	0.466 (0.447;0.484)	0.547 (0.525;0.569)
Área de Planejamento		
2	-0.198 (-0.215;-0.180)	-0.178 (-0.199;-0.157)
3	0.049 (0.033;0.066)	0.050 (0.030;0.069)
4	0.025 (0.009;0.042)	0.030 (0.010;0.050)
5	-0.050 (-0.070;-0.030)	-0.054 (-0.078;-0.030)

Tabela 5: Estimativas frequentistas (média e intervalo de confiança) dos parâmetros com ajuste pelas funções Log-Normal e Gama.

No contexto deste estudo, as variáveis que possuem coeficientes positivos, influenciam mais no aumento do prêmio comparado a sua variável de referência. Portanto, concluímos que, pelos modelos adotados, pessoas do sexo masculino tendem a pagar um valor maior pelo prêmio do que pessoas do sexo feminino, assim como os solteiros em relação aos casados. Ademais, as pessoas que possuem um veículo da categoria “Coupe/Roadster”, por exemplo, desembolsam um montante maior para o seguro do que as pessoas que possuem um “Sedan Compacto”. Por outro lado, pessoas mais velhas tendem a pagar um valor menor pelo seguro, assim como pessoas que moram nas Áreas de Planejamento 2 e

5 comparado às que moram na Área de Planejamento 1.

Para fins de comparação dos modelos ajustados, será utilizado o critério AIC, conforme exposto na Seção 2.6. O valor do AIC computado para os ajustes Log-Normal e Gama foram, respectivamente, 62.828 e 768.078, o que nos leva a concluir que, segundo a abordagem frequentista, o ajuste realizado através da função Log-Normal é o mais adequado na modelagem do valor do prêmio.

Sendo assim, foram obtidos os exponenciais das médias dos coeficientes de regressão ($\hat{\beta}_j$), a fim de esclarecer quais são os fatores mais importantes para determinar o valor do prêmio. Tais valores podem ser observados na Tabela 6.

Variável	$\hat{\beta}_j$	$\exp\{\hat{\beta}_j\}$
Intercepto	7.843	2547.837
Idade	-0.011	0.989
Sexo Masculino	0.175	1.192
Estado Civil Solteiro	0.087	1.091
Categoria do veículo		
Sedan Compacto	-	-
Coupe/Roadster	1.200	3.318
Crossover	0.411	1.508
Hatchback Compacto	-0.186	0.830
Hatchback Médio	0.134	1.143
Jipe	0.554	1.740
Minivan Compacto	-0.044	0.957
Minivan Médio ou Grande	0.191	1.210
Multivan/Furgão	0.107	1.113
Picape Cabine Dupla	0.941	2.561
Picape	0.179	1.196
Sedan Médio ou Grande	0.509	1.664
Station-Wagon Compacto	-0.072	0.930
Station-Wagon Médio ou Grande	0.056	1.058
SUV Compacto	0.045	1.046
SUV Médio ou Grande	0.466	1.593
Área de Planejamento		
2	-0.198	0.821
3	0.049	1.051
4	0.025	1.026
5	-0.050	0.951

Tabela 6: Exponenciais das médias dos coeficientes de regressão obtidas pela ajuste Log-Normal através da abordagem frequentista.

Ao observar o exponencial das médias dos coeficientes na Tabela 6, pode ser realizada uma interpretação mais clara sobre a influência de cada variável sobre o valor do prêmio.

Observa-se por exemplo que a cada unidade que a variável Idade aumenta, o valor do prêmio a ser pago decresce em 1.1% (1-0.989). Além disso, ao observar a informação relativa ao estado civil, percebe-se que um cliente solteiro tem um valor do prêmio 9,1% (1.091-1) mais alto em relação a um cliente casado. Outra conclusão possível é que, em relação às pessoas que possuem um Sedan compacto, pessoas que possuem uma Picape com cabine dupla pagam um valor do prêmio 156.1% (2.561-1) mais alto, enquanto que as que possuem um Hatchback compacto pagam um valor do prêmio 17% (1-0.83) mais baixo.

3.3.2 Ajuste Bayesiano

As distribuições *a priori* atribuídas para os parâmetros nos modelos Bayesiano foram vagas, similares às adotadas para os dados simulados, a saber: $\pi(\beta_j) \sim N(0, 100)$ para os coeficientes β_j ($j = 0, \dots, k$); $\pi(\tau) \sim Gama(0.01, 0.01)$, para o parâmetro de precisão $\tau = \frac{1}{\sigma^2}$ para o modelo Log-Normal; e $\pi(\alpha) \sim U(0, 100)$ para o parâmetro de forma α da distribuição Gama, conforme a parametrização definida na Equação 2.6.

Os resultados dos ajustes realizados através da abordagem Bayesiana encontram-se na Tabela 7, onde são apresentadas as médias *a posteriori* e os intervalos de 95% de credibilidade. Mais uma vez, observou-se que as estimativas para todos os coeficientes são semelhantes entre os dois modelos.

Em conformidade com os achados na modelagem frequentista, todas as variáveis consideradas nesse estudo mostram-se importantes para a especificação do seguro de automóveis, pois como pode-se observar na Tabela 7, o intervalo de credibilidade de nenhum coeficiente contém o valor zero, o que significa que todas as variáveis são relevantes para o modelo. Além disso, foi inferido que o prêmio será maior para os segurados do sexo masculino e os solteiros, assim como para aqueles cujo local de residência pertence às Áreas de Planejamento 3 e 4, em comparação àqueles que residem na Área de Planejamento 1. Já para a idade, pode-se dizer que pessoas mais novas tendem a pagar valores do prêmio mais elevados.

Ao compararmos as Tabelas 5 e 7, nota-se que as estimativas pontuais, bem como as estimativas intervalares, são semelhantes nas abordagens frequentista e Bayesiana, como já era esperado uma vez que foram adotadas distribuições *a priori* não informativas para os parâmetros nos ajustes Bayesianos.

Ajuste Variável	Log-Normal Média (IC)	Gama Média (IC)
Intercepto	7.580 (7.545;7.591)	7.672 (7.640;7.705)
Idade	-0.011 (-0.012;-0.011)	-0.011 (-0.011;-0.011)
Sexo Masculino	0.176 (0.166;0.186)	0.179 (0.170;0.188)
Estado Civil Solteiro	0.087 (0.078;0.097)	0.082 (0.071;0.092)
Categoria do veículo		
Sedan Compacto	-	-
Coupe/Roadster	1.199 (1.103;1.235)	1.413 (1.307;1.448)
Crossover	0.410 (0.382;0.439)	0.422 (0.390;0.433)
Hatchback Compacto	-0.186 (-0.199;-0.173)	-0.188 (-0.200;-0.175)
Hatchback Medio	0.134 (0.115;0.154)	0.165 (0.145;0.184)
Jipe	0.560 (0.286;0.844)	0.679 (0.394;0.983)
Minivan Compacto	-0.044 (-0.063;0.704)	-0.060 (-0.079;-0.041)
Minivan Médio ou Grande	0.191 (0.155;0.227)	0.240 (0.202;0.279)
Multivan/Furgão	0.107 (0.053;0.161)	0.110 (0.055;0.167)
Picape Cabine Dupla	0.941 (0.915;0.968)	0.994 (0.968;1.022)
Picape	0.179 (0.140;0.216)	0.202 (0.165;0.240)
Sedan Médio ou Grande	0.510 (0.493;0.526)	0.535 (0.518;0.552)
Station-Wagon Compacto	-0.073 (-0.120;-0.023)	-0.079 (-0.128;-0.033)
Station-Wagon Médio ou Grande	0.057 (0.004;0.110)	0.093 (0.038;0.149)
SUV Compacto	0.045 (0.019;0.071)	0.030 (0.005;0.054)
SUV Médio ou Grande	0.466 (0.447;0.485)	0.547 (0.528;0.566)
Área de Planejamento		
2	-0.198 (-0.207;-0.189)	-0.178 (-0.195;-0.160)
3	0.050 (0.041;0.059)	0.050 (0.033;0.066)
4	0.025 (0.016;0.034)	0.030 (0.014;0.047)
5	-0.050 (-0.060;-0.040)	-0.054 (-0.075;-0.035)

Tabela 7: Estimativas Bayesianas (média e intervalo de credibilidade) dos parâmetros com ajuste pelas funções Log-Normal e Gama.

Para fins de comparação dos modelos ajustados sob enfoque Bayesiano, será adotado o critério DIC, conforme exposto na Seção 2.6. O valor do DIC encontrado para o Log-Normal e Gama foram, respectivamente, 762.943 e 768.042. Como o ajuste realizado através da função Log-Normal possui o menor DIC, concluímos, portanto, que esse conjunto de dados se adequa melhor a uma distribuição Log-Normal.

Aqui, assim como feito na abordagem frequentista, também pode ser calculado o exponencial das médias do coeficiente do modelo escolhido a fim de facilitar a interpretação do modelo. Tais resultados podem ser observados na Tabela 8.

Variável	$\hat{\beta}_j$	$\exp\{\hat{\beta}_j\}$
Intercepto	7.580	1958.629
Idade	-0.011	0.989
Sexo Masculino	0.176	1.192
Estado Civil Solteiro	0.087	1.091
Categoria do veículo		
Sedan Compacto	-	-
Coupe/Roadster	1.199	3.320
Crossover	0.410	1.507
Hatchback Compacto	-0.186	0.830
Hatchback Médio	0.134	1.143
Jipe	0.560	1.751
Minivan Compacto	-0.044	0.957
Minivan Médio ou Grande	0.191	1.210
Multivan/Furgão	0.107	1.113
Picape Cabine Dupla	0.941	2.561
Picape	0.179	
Sedan Médio ou Grande	0.510	1.665
Station-Wagon Compacto	-0.073	0.930
Station-Wagon Médio ou Grande	0.057	1.059
SUV Compacto	0.045	1.046
SUV Médio ou Grande	0.466	1.593
Área de Planejamento		
2	-0.198	0.821
3	0.050	1.051
4	0.025	1.026
5	-0.050	0.951

Tabela 8: Exponenciais das médias dos coeficientes de regressão obtidas pela ajuste Log-Normal através da abordagem Bayesiana.

Consegue-se perceber que os resultados são similares aos da inferência frequentista. Uma interpretação possível ao observar a Tabela 8 é que, em relação às pessoas que moram na área de planejamento 1, pessoas que moram na área de planejamento 2 pagam um valor do prêmio 17.9% (1-0.821) menor, enquanto pessoas que moram na área de planejamento 3, pagam um valor do prêmio 5.1% (1.051-1) maior.

4 Conclusão

A motivação desse trabalho originou-se devido à importância de uma correta precificação de seguros, de forma que a seguradora possa assumir os riscos de possíveis sinistros, de acordo com as normas impostas pelo órgão regulador, assim como obter vantagens em relação aos concorrentes ao cobrar um valor competitivo no mercado de seguros.

Mais especificamente, o objetivo do trabalho foi avaliar diferentes modelos estatísticos a fim de inferir sobre o valor do prêmio de seguros de automóveis na cidade do Rio de Janeiro, identificando as variáveis que influenciam na precificação do mesmo, tais como as características do segurado e do automóvel assegurado.

Devido à natureza da variável resposta de interesse, a saber, o prêmio comercial (em reais), que apresenta forte tendência à assimetria, o modelo de regressão múltipla torna-se inapropriado, pois as suposições requeridas para aplicação dessa metodologia são violadas, conforme foi mostrado na Seção 3.3. Portanto, foram considerados os modelos lineares generalizados, que não necessitam das suposições de linearidade e normalidade, permitindo, assim, a modelagem de dados com distribuição assimétrica. Mais especificamente, foram analisados os modelos Log-normal e Gama com função de ligação logarítmica, considerando as abordagens frequentista e Bayesiana.

Através das Tabelas 5 e 7, nota-se que todas as características consideradas nesse trabalho foram importantes para a precificação do seguro de automóveis na Cidade do Rio de Janeiro. Ao realizar a comparação dos ajustes Log-Normal e Gama, através das medidas do AIC e do DIC para os modelos segundo a abordagem frequentista e Bayesiana, respectivamente, notou-se que o modelo Log-Normal se ajustou melhor aos dados analisados.

Tanto na modelagem dos dados simulados (Seção 3.1) quanto na modelagem dos dados reais (Seção 3.3), ficou evidente que as estimativas frequentistas foram similares às Bayesianas, o que se justifica devido ao fato de termos atribuído distribuições *a priori*

não-informativas para os parâmetros dos modelos, ou seja, distribuições com baixo grau de precisão acerca do conhecimento prévio sobre tais parâmetros. Considerando estas condições, para este caso, os resultados serão semelhantes ao utilizar o modelo clássico ou Bayesiano. Então pode-se adotar o modelo frequentista devido sua simplicidade e, realizar o ajuste através da distribuição Log-Normal, que foi a que se adequou melhor aos dados. Mas, é importante lembrar que não se sabe o que aconteceria caso existisse uma boa informação *a priori*.

Sendo assim, tem-se que o modelo de ajuste Log-Normal pode ser escrito como $Y = \exp(X^T\beta) + \epsilon$, desta forma, considerando como exemplo o resultado da abordagem frequentista, tem-se que o valor do prêmio comercial de uma pessoa solteira, do sexo masculino, que possui 27 anos, mora em algum dos bairros da área de planejamento 2 e possui um carro da categoria Hatchback Médio, seria estimado em aproximadamente R\$2.307, 00.

Portanto, fica claro que, na prática, este modelo oferece às seguradoras a oportunidade de prever o valor comercial do prêmio de um possível cliente, ou seja, de algum indivíduo que não foi considerado na construção deste modelo de precificação de seguros.

Vale destacar, ainda, que a base de dados utilizada nesse trabalho possui algumas limitações, implicando em resultados menos detalhados. Sabe-se que existem, ainda, outras variáveis importantes para a precificação do seguro que não foram consideradas nos modelos ajustados. De fato, o banco de dados disponibilizado não contém, por exemplo, determinadas informações sobre o veículo, que certamente influenciam no preço do seguro, como a marca e o ano do automóvel; ou ainda, o local pelo qual o segurado costuma trafegar; se o carro possui mais de um condutor e as características destes, entre outros fatores. Todas essas informações são relevantes na etapa de precificação do seguro de um automóvel e, se disponíveis, devem ser consideradas em trabalhos futuros.

Apesar de que neste trabalho, por serem consideradas prioris vagas, não haja diferença relevante entre o modelo frequentista e Bayesiano, é interessante que em trabalhos futuros, sejam analisadas distribuições *a priori* que sejam informativas, para que as mesmas possam ser analisadas como alternativas aos modelos encontrados neste trabalho. Pode-se, ainda, considerar a adoção de modelos lineares mistos, permitindo a inclusão de efeitos fixos e aleatórios. Outra possibilidade interessante é utilizar a base de dados de diversas seguradoras a fim de diversificar a amostra e ter uma conclusão mais abrangente.

APÊNDICE A – Tabela da relação de bairros e áreas de planejamento da cidade do Rio de Janeiro

Área de Planejamento	Bairro
1	Benfica, Caju, Catumbi, Centro, Cidade Nova, Estácio, Gamboa, Mangueira, Paquetá, Rio Comprido, Santa Teresa, Santo Cristo, São Cristóvão, Saúde e Vaco da Gama
2	Alto da Boa Vista, Andaraí, Botafogo, Catete, Copacabana, Cosme Velho, Flamengo, Gávea, Glória, Grajaú, Humaitá, Ipanema, Jardim Botânico, Lagoa, Laranjeiras, Leblon, Leme, Maracanã, Praça da Bandeira, Rocinha, São Conrado, Tijuca, Urca, Vidigal, Vila Isabel
3	Abolição, Acari, Água Santa, Anchieta, Bancários, Barros Filho, Bento Ribeiro, Bonsucesso, Brás de Pina, Cachambi, Cacuia, Campinho, Cascadura, Cavalcanti, Cidade Universitária, Cocotá, Coelho Neto, Colégio, Complexo do Alemão, Cordovil, Costa Barros, Del Castilho, Encantado, Engenheiro Leal, Engenho da Rainha, Engenho de Dentro, Engenho Novo, Freguesia, Galeão, Guadalupe, Higienópolis, Honório Gurgel, Inhaúma, Irajá, Jacaré, Jacarezinho, Jardim América, Jardim Carioca, Jardim Guanabara, Lins de Vasconcelos, Madureira, Manguinhos, Maré, Marechal Hermes, Maria da Graça, Méier, Moneró, Olaria, Osvaldo Cruz, Parada de Lucas, Parque Anchieta, Parque Colúmbia, Pavuna, Penha, Penha Circular, Piedade, Pilares, Pitangueiras, Portuguesa, Praia da Bandeira, Quintino Bocaiúva, Ramos Riachuelo, Ribeiro, Ricardo De Albuquerque, Rocha, Rocha Miranda, Sampaio, São Francisco Xavier, Tauá, Todos os Santos, Tomás Coelho, Turiaçú, Vaz Lobo, Vicente de Carvalho, Vigário Geral, Vila da Penha Vila Kosmos, Vista Alegre, Zumbi
4	Anil, Barra da Tijuca, Camorim, Cidade de Deus, Curicica, Freguesia Jacarepaguá, Gardênia Azul, Grumari, Itanhangá, Jacarepaguá, Joá, Pechincha, Praça Seca, Recreio dos Bandeirantes Tanque, Taquara, Vargem Grande, Vargem Pequena, Vila Valqueire
5	Bangu, Barra de Guaratiba, Campo dos Afonsos, Campo Grande, Cosmos, Deodoro, Gericinó, Guaratiba, Inhoaíba, Jardim Sulacap, Magalhães Bastos, Paciência, Padre Miguel, Pedra de Guaratiba, Realengo, Santa Cruz, Santíssimo, Senador Camará Senador Vasconcelos, Sepetiba, Vila Militar

Tabela 10: Relação de bairros e APs da cidade do Rio de Janeiro

APÊNDICE B – Estimador de β por mínimos quadrados

Considere Y uma variável aleatória com distribuição de probabilidade indexada pelo parâmetro de interesse β o vetor de parâmetro de interesse e $\hat{\beta}$ o vetor que minimiza a soma dos quadrados dos resíduos. Tal soma, apresentada na Equação 2.1 representa as distâncias quadráticas entre os valores observados de Y e seus valores ajustados pelo modelo, denotados por \hat{Y} .

$$SQE = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \quad (\text{B.1})$$

A demonstração da Equação B.1 acima é dada a seguir.

Derivando a Equação B.1, tem-se:

$$\begin{aligned} \frac{\partial QE}{\partial \beta} &= \frac{\partial (\mathbf{Y}^T \mathbf{Y} - 2\beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X} \beta)}{\partial \beta} \\ &= -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \beta. \end{aligned} \quad (\text{B.2})$$

Por fim, igualando a Equação B.2 a zero, tem-se:

$$\begin{aligned} -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \beta &= 0 \\ \Leftrightarrow \mathbf{X}^T \mathbf{X} \beta &= \mathbf{X}^T \mathbf{Y} \\ \Leftrightarrow (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ \Leftrightarrow \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \end{aligned}$$

APÊNDICE C – Propriedades da Família Exponencial

Considere uma variável aleatória Y com distribuição de probabilidade indexada pelo parâmetro de interesse θ . É dito que esta distribuição pertence à família exponencial caso possa ser escrita da seguinte forma:

$$f_y(y; \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}, \quad (\text{C.1})$$

onde a e d são funções de y independentes do parâmetro θ e b e c são funções apenas de θ .

Dentre as propriedades da família exponencial, pode-se mostrar que:

$$E[a(y)] = \frac{-c'(\theta)}{b'(\theta)}$$

e

$$\text{Var}(a(y)) = \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{[b'(\theta)]^3}.$$

Tais propriedades serão demonstradas a seguir:

Sabe-se que a derivada da função $f(y; \theta)$ com relação ao parâmetro θ é dada por:

$$\frac{d(f(y; \theta))}{d(\theta)} = (a(y)b'(\theta) + c'(\theta)) f(y; \theta) \quad (\text{C.2})$$

Logo, de acordo com as Equações C.1 e C.2:

$$\begin{aligned}
 \int \frac{d(f(y; \theta))}{d(\theta)} dy &= \int (a(y)b'(\theta) + c'(\theta)) f(y; \theta) dy \\
 &= \int a(y)b'(\theta)f(y; \theta)dy + \int c'(\theta)f(y; \theta)dy \\
 &= b'(\theta) \int a(y)f(y; \theta)dy + c'(\theta) \int f(y; \theta)dy \\
 &= b'(\theta)E[a(y)] + c'(\theta)
 \end{aligned}$$

Além disso, pela definição de densidade de probabilidade de C.1, sabe-se que

$$\begin{aligned}
 b'(\theta)E[a(y)] + c'(\theta) &= 0 \\
 \iff E[a(y)] &= \frac{-c'(\theta)}{b'(\theta)}
 \end{aligned} \tag{C.3}$$

Agora, para encontrar $Var(a(y))$, será resolvida a seguinte equação:

$$\int \frac{d^2(f(y; \theta))}{d(\theta)} dy = 0 \tag{C.4}$$

Sabe-se que

$$\begin{aligned}
 \frac{d^2(f(y; \theta))}{d(\theta)} &= \frac{d}{d(\theta)} [(a(y)b'(\theta) + c'(\theta))f(y; \theta)] \\
 &= \frac{d}{d(\theta)} [a(y)b'(\theta) + c'(\theta)]f(y; \theta) + [(a(y)b'(\theta) + c'(\theta))\frac{d}{d(\theta)}f(y; \theta)] \\
 &= a(y)b''(\theta) + c''(\theta)f(y; \theta) + [a(y)b'(\theta) + c'(\theta)]^2f(y; \theta) \\
 &= f(y; \theta)[(a(y)b''(\theta) + c''(\theta)) + [a(y)b'(\theta) + c'(\theta)]^2]
 \end{aligned}$$

Mas,

$$\begin{aligned}
 [a(y)b'(\theta) + c'(\theta)]^2 &= \left(b'(\theta)[a(y)\frac{c'(\theta)}{b'(\theta)}] \right)^2 \\
 &= (b'(\theta)[a(y) - E[a(y)])^2 \\
 &= [b'(\theta)]^2[a(y) - E[a(y)]]^2
 \end{aligned}$$

Logo, a Equação C.4 pode ser escrita como:

$$\begin{aligned} & \int (a(y)b''(\theta) + c''(\theta) + [b'(\theta)]^2[a(y)E[a(y)]]^2)f(y; \theta)dy = 0 \\ \iff & \int (a(y)b''(\theta))f(y; \theta)dy + \int c''(\theta)f(y; \theta)dy + \int [b'(\theta)]^2[a(y)E[a(y)]]^2f(y; \theta)dy = 0 \\ \iff & b''(\theta)E[a(y)] + c''(\theta) + [b'(\theta)]^2Var(a(y)) = 0 \end{aligned}$$

Ou seja,

$$Var(a(y)) = \frac{-b''(\theta)E[a(y)] - c''(\theta)}{[b'(\theta)]^2}$$

Portanto, substituindo $E[a(y)]$ de acordo com a Equação C.3, tem-se que:

$$\begin{aligned} Var(a(y)) &= \frac{-b''(\theta) \left(\frac{-c'(\theta)}{b'(\theta)} \right) - c''(\theta)}{[b'(\theta)]^2} \\ &= \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{[b'(\theta)]^3} \end{aligned}$$

Na forma canônica, tem-se que $a(y) = y$. Logo:

$$E[y] = \frac{-c'(\theta)}{b'(\theta)}$$

e

$$Var(y) = \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{[b'(\theta)]^3}.$$

APÊNDICE D – Propriedades da Função Escore

Considere uma variável aleatória Y com distribuição de probabilidade indexada pelo parâmetro de interesse θ . Suponha que U denota a função escore, onde o vetor de escores do j -ésimo coeficiente, na forma matricial, pode ser definido como

$$U_j(\beta) = X^T \cdot D \cdot V^{-1} (y - \mu),$$

onde X^T é a matriz transposta de X ; V^{-1} é a matriz inversa da variância e D é a diagonal da matriz de $\frac{\partial \mu}{\partial \eta}$. Algumas propriedades da função escore são:

- $E[U(\theta; y)] = 0;$
- $Var(U(\theta; y)) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta);$
- $Var(U) = E[U^2] = -E[U']$

Suas demonstrações são dadas a seguir.

- $E[U(\theta; y)] = 0$

Prova: Determinando a função escore por

$$U(\theta; y) = \frac{dl(\theta; y)}{d(\theta)} = a(y)b'(\theta) + c'(\theta),$$

tem-se que:

$$\begin{aligned} E[U(\theta; y)] &= b'(\theta)E[a(y)] + c'(\theta) \\ &= b'(\theta) \left(\frac{-c'(\theta)}{b'(\theta)} \right) + c'(\theta) \\ &= -c'(\theta) + c'(\theta) \\ &= 0 \end{aligned} \tag{D.1}$$

- $Var(U(\theta; y)) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$

Prova: Sabe-se que

$$\begin{aligned} Var(U(\theta; y)) &= Var(a(y)b'(\theta) + c'(\theta)) \\ &= Var(a(y)b'(\theta)) + Var(c'(\theta)) \\ &= [b'(\theta)]^2 Var(a(y)) \end{aligned}$$

que, de acordo com a Equação D.2, pode ser definida como:

$$\begin{aligned} Var(U(\theta; y)) &= \frac{[b'(\theta)]^2[b''(\theta)c'(\theta) - b'(\theta)c''(\theta)]}{[b'(\theta)]^3} \\ &= \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta) \end{aligned} \quad (\text{D.2})$$

- $Var(U) = E[U^2] = -E[U']$

Prova: De acordo com a Equação D.1, tem-se:

$$Var(U) = E[U^2].$$

Além disso,

$$\begin{aligned} U' &= \frac{dU}{d\theta} = \frac{d}{d\theta}(a(y)b'(\theta) + c'(\theta)) \\ &= a(y)b''(\theta) + c''(\theta). \end{aligned}$$

Logo,

$$\begin{aligned} E[U'] &= (E[a(y)])b''(\theta) + c''(\theta) \\ &= -\frac{c'(\theta)}{b'(\theta)}b''(\theta) + c''(\theta) \\ &= -\left[\frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)\right] \end{aligned}$$

que, de acordo com a Equação D.2

$$E[U'] = -Var(U)$$

APÊNDICE E – Resultados dos Dados Ajustados para uma Distribuição Normal

Os dados das características do segurados e seus veículos, fornecidos pela seguradora foram ajustados para uma distribuição Normal, de acordo com a abordagem frequentista, vista na Seção 2.2.

Neste contexto, a variável resposta Y_i representa o valor do prêmio. Portanto, a média dos prêmios foi modelada como $E[Y_i] = \mu_i = \beta_0 + \beta_1$ Idade + β_2 Sexo Masculino + β_3 Estado Civil Solteiro + β_4 Categoria do Automóvel + β_5 Área de Planejamento.

As estimativas dos coeficientes encontrados pelo modelo de regressão múltipla, assim como seus respectivos erros, podem ser encontrados na Tabela 11.

Tabela 11: Estimadores dos parâmetros para o modelo de regressão múltipla

Variável	$\hat{\beta}$	Erro	P-valor
Intercepto	2741.05	39.9	$< 2e^{-16}$
Idade	-24.22	0.58	$< 2e^{-16}$
Sexo Masculino	2404.60	15.31	$< 2e^{-16}$
Estado Civil Solteiro	211.26	16.8	$< 2e^{-16}$
Categoria do veículo			
Sedan Compacto	-	-	-
Coupe/Roadster	5938.53	169.99	$< 2e^{-16}$
Crossover	1024.70	48.22	$< 2e^{-16}$
Hatchback Compacto	-340.94	21.26	$< 2e^{-16}$
Hatchback Médio	376.39	32.08	$< 2e^{-16}$
Jipe	1728.50	464.72	0.0002
Minivan Compacto	-87.00	31.89	0.0064
Minivan Médio ou Grande	554.07	62.79	$< 2e^{-16}$
Multivan/Furgão	219.61	89.51	0.0141
Picape Cabine Dupla	3285.20	43.06	$< 2e^{-16}$
Picape	427.93	64.01	$2.34e^{-11}$
Sedan Médio ou Grande	1414.40	27.16	$< 2e^{-16}$
Station-Wagon Compacto	-118.37	77.08	0.1247
Station-Wagon Médio ou Grande	216.04	88.09	0.0142
SUV Compacto	74.86	40.17	0.0624
SUV Médio ou Grande	1388.20	30.31	$< 2e^{-16}$
Área de Planejamento			
2	-386.60	28.62	$< 2e^{-16}$
3	117.68	27.13	$1.44e^{-05}$
4	85.68	26.80	0.0014
5	-100.64	33.11	0.0024

APÊNDICE F – Código R utilizados no Estudo de Simulação

```
library(tidyverse)
library(sjPlot)
library(gtsummary)
library(flextable)
library(kableExtra)
library(lmtest)
library(stats)
library(MASS)
library(magrittr)
library(geobr)
library(rgdal)
library(leaflet)

#lendo a base de dados
base<- read_csv2("Dados_Rio.csv")

#excluindo variaveis que não interessam para análise
base<- base %>% select(-c(CODRA,CODBAIRRO,Regiao,UF,OBJECTID))
%>%
filter(Tipo_pesso == "F")

#renomeando variáveis
base<- base %>% rename(Bairro = NOME) %>%
rename('Estado civil' = Estado_civ) %>%
```

```

rename('Grupo do veículo' = Grupo_veic) %>%
  rename('Tipo de pessoa' = Tipo_pesso) %>%
  rename(Prêmio = premio_mod) %>%
  rename(Relatividade = Relativida) %>%
  rename('Região Adm.' = REGIAO_ADM) %>%
  rename('Área de Planejamento' = AREA_PLANE)

#transformando variáveis do tipo caracter para fatores
base <- mutate_if(base, is.character, as.factor)
base$'Área de Planejamento' <- as.factor(base$'Área de
Planejamento')

base$'Região Adm.' <- factor(base$'Região Adm.',
                                levels = c("ANCHIETA", "BANGU", "BARRA
DA TIJUCA", "BOTAFOGO", "CAMPO
GRANDE", "CENTRO", "CIDADE DE DEUS",
"COMPLEXO DA MARE", "COMPLEXO DO
ALEMÃO", "COPACABANA", "GUARATIBA", "
ILHA DO GOVERNADOR", "INHAUMA", "
IRAJA", "JACAREPAGUA", "JACAREZINHO"
,"LAGOA", "MADUREIRA", "MEIER", "
PAQUETA", "PAVUNA", "PENHA", "
PORTUARIA", "RAMOS", "REALENGO", "RIO
COMPRIDO", "ROCINHA", "SANTA CRUZ", "
SANTA TEREZA", "SAO CRISTOVAO", "
TIJUCA", "VIGARIO GERAL", "VILA
ISABEL"),
                                labels=c("Anchieta", "Bangu", "Barra da
Tijuca", "Botafogo", "Campo Grande",
"Centro", "Cidade de Deus", "Complexo
da Mare", "Complexo do Alemão", "
Copacabana", "Guaratiba", "Ilha do
Governador", "Inhauma", "Irajá", "
Jacarepagua", "Jacarezinho", "Lagoa",
"Madureira", "Meier", "Paqueta", "
Pavuna", "Penha", "Portuaria", "Ramos"
)

```

```

        , "Realengo", "Rio Comprido", "Rocinha"
        , "Santa Cruz", "Santa Tereza", "São
        Cristovao", "Tijuca", "Vigario Geral
        , "Vila Isabel"))

base <- base%>% mutate(Sexo = ifelse(
  Sexo == "001 - Masculino", "Masculino", "Feminino"))

base <- base%>% mutate('Estado civil' = ifelse(
  'Estado civil' == "001 - Solteiro", "Solteiro", "Casado"))

# retirando as observações que possuem NA
base<-na.omit(base)

#excluindo \textit{outlier} com valores de prêmio acima de 25.000
base <- base %>% filter(Prêmio<=25000

#verificando quais variáveis possuem observações faltantes
# table(is.na(base$Estado_civ)) - nao possui NA
# table(is.na(base$Sexo)) - nao possui NA
# table(is.na(base$Idade)) - nao possui NA
# table(is.na(base$premio_mod)) - nao possui NA
# table(is.na(base$Relativida)) - nao possui NA
# table(is.na(base$OBJECTID)) - nao possui NA
# table(is.na(base$NOME)) - nao possui NA
# table(is.na(base$REGIAO_ADM)) - nao possui NA
# table(is.na(base$AREA_PLANE)) - nao possui NA
# table(is.na(base$Tipo_pesso)) - nao possui NA
# table(is.na(base$sinistro)) - nao possui NA
# prop.table(table(is.na(base$'Grupo do veículo')))

#Tabela das variáveis qualitativas
TVEI<- flextable(
  base %>% group_by('Grupo do veículo') %>%
    summarise('Frequência absoluta' = n(),

```

```

`Frequência relativa (%)` = round((n())/length(base
$`Grupo do veículo`)*100, 2)) %>%
arrange(desc(`Frequência absoluta`)),
cwidth = 2) %>%
theme_vanilla()

TEC<- flextable(
na.omit(base) %>% group_by(`Estado civil`) %>%
summarise(`Frequência absoluta` = n(),
`Frequência relativa (%)` = round((n())/length(base
$`Estado civil`)*100, 2)) %>%
arrange(desc(`Frequência absoluta`)),
cwidth = 2) %>%
theme_vanilla()

TSEX<- flextable(
na.omit(base) %>% group_by(Sexo) %>%
summarise(`Frequência absoluta` = n(),
`Frequência relativa (%)` = round((n())/length(base$Sexo)*100, 2)) %>%
arrange(desc(`Frequência absoluta`)),
cwidth = 2) %>%
theme_vanilla()

TBAI<- flextable(
base %>% group_by(Bairro) %>%
summarise(`Frequência absoluta` = n(),
`Frequência relativa (%)` = round((n())/length(base$Bairro)*100, 2)) %>%
arrange(desc(`Frequência absoluta`)),
cwidth = 2) %>%
theme_vanilla()

TRA<- flextable(

```

```

na.omit(base) %>% group_by('Região Adm.') %>%
  summarise('Frequência absoluta' = n(),
            'Frequência relativa (%)' = round((n())/length(base
$'Região Adm.'))*100, 2)) %>%
  arrange(desc('Frequência absoluta')),
  cwidth = 2) %>%
theme_vanilla()

TREL<- flextable(
  na.omit(base) %>% group_by(Relatividade) %>%
    summarise('Frequência absoluta' = n(),
              'Frequência relativa (%)' = round((n())/length(base$Relatividade))*100, 2)) %>%
    arrange(desc('Frequência absoluta')),
    cwidth = 2) %>%
  theme_vanilla()

TAP<- flextable(
  na.omit(base) %>% group_by('Área de Planejamento') %>%
    summarise('Frequência absoluta' = n(),
              'Frequência relativa (%)' = round((n())/length(base$'Área de Planejamento'))*100, 2)) %>%
    arrange(desc('Frequência absoluta')),
    cwidth = 2) %>%
  theme_vanilla()

#Tabelas para as variáveis quantitativas
base %>%
  summarise('Média' = round(mean(Idade, na.rm = T), 2),
            'Desvio padrão' = round(sd(Idade, na.rm=T), 2),
            'Mínimo' = min(Idade, na.rm = T),
            'Mediana' = median(Idade, na.rm = T),
            'Máximo' = max(Idade, na.rm = T),
            )

```

```

`Q1` = quantile(base$Idade, 0.25),
`Q3` = quantile(base$Idade, 0.75)) %>%
kable(caption = "MEDIDAS DESCRIPTIVAS DA IDADE") %>% kable_paper()
() %>%
kable_styling(row_label_position = r)

base %>% filter(`Tipo de pessoa` == "F") %>%
summarise(`Média` = round(mean(Prêmio, na.rm = T), 2),
`Desvio padrão` = round(sd(Prêmio, na.rm=T), 2),
`Mínimo` = min(Prêmio, na.rm = T),
`Mediana` = median(Prêmio, na.rm = T),
`Máximo` = max(Prêmio, na.rm = T),
`Q1` = quantile(base$Prêmio, 0.25),
`Q3` = quantile(base$Prêmio, 0.75)) %>%
kable(caption = "MEDIDAS DESCRIPTIVAS DO PRÊMIO (em reais)") %>%
kable_paper()

#####

```

#Gráfico das variáveis quantitativas

```

# Idade

grafidade<- ggplot(data=base, aes(x=Idade)) +
  geom_histogram(col="grey",
                 fill="lightblue",
                 binwidth = 1) +
  labs(x="Idade", y="Frequência absoluta") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(20,76,5)) +
  geom_vline(aes(xintercept=mean(Idade)),
             color="yellow", linetype="dashed", size=1) +

```

```
ylim(c(0,2000))

grafidade

plot(base$Idade,base$Prêmio)

boxplotidade<- ggplot(data=base, aes(y=Idade)) +
  geom_boxplot(col="grey",
                fill="lightblue") +
  theme_minimal() +
  scale_y_continuous(breaks = seq(20,80,10))

boxplotidade

#Prêmio

boxplotPrêmio <- ggplot(data=base, aes(y=Prêmio)) +
  geom_boxplot(col="black",
                fill="#D0E0E3") + xlab(" ") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.background = element_blank(),
        panel.grid = element_line(colour = "#f0f0f0")) +
  ylim(c(0,25000)) +
  ylab("Valor do prêmio \n (em reais)")

#Prêmio x sexo

premio_sexo<- base %>%
  ggplot(aes(y=Prêmio,x=Sexo,fill=Sexo)) +
  geom_boxplot(fill="#D0E0E3") +
  theme(axis.ticks = element_blank(),
```

```

axis.text.y = element_blank(),
axis.title = element_blank(),
panel.grid.major.x = element_blank(),
panel.grid.minor.x = element_blank(),
panel.background = element_blank(),
panel.grid = element_line(colour = "#f0f0f0")) +
ylim(c(0,25000)) +
guides(fill=F)

#Prêmio x Estado Civil

premio_EC<- base %>%
ggplot(aes(y=Prêmio,x='Estado civil',fill='Estado civil')) +
geom_boxplot(fill="#DOE0E3") +
theme(axis.ticks = element_blank(),
      axis.text.y = element_blank(),
      axis.title = element_blank(),
      panel.grid.major.x = element_blank(),
      panel.grid.minor.x = element_blank(),
      panel.background = element_blank(),
      panel.grid = element_line(colour = "#f0f0f0")) +
ylim(c(0,25000)) +
ylab("Valor do prêmio \n (em reais)") + guides(fill=F)

#library(gridExtra)

grafs<- gridExtra::grid.arrange(boxplotPrêmio,premio_sexo,premio_EC, ncol=3, nrow=1)

#ggsave("bp_sex_ec.pdf", plot=grafs, width = 5.5, height = 5.0)

boxplot_AP<- ggplot(base, aes(x = 'Área de Planejamento',
                                y = Prêmio,
                                fill='Área de Planejamento')) +
geom_boxplot( fill="#DOE0E3") +

```

```

guides(fill=F) +
theme_minimal() +
ylim(c(0,25000)) +
ylab("Valor do prêmio \n (em reais)")

#ggsave("bp_AP.pdf")

#Premio X Região Administrativa

basenova<- base

basenova$'Região Adm.'<- factor(basenova$'Região Adm.',
c('Centro', 'Paqueta', 'Portuaria', 'Rio
Comprido', 'Santa Tereza', 'São Cristovao',
'Botafogo', 'Copacabana', 'Lagoa', 'Rocinha', ,
Tijuca', 'Vila Isabel',
'Anchieta', 'Complexo da Mare', 'Complexo do
Alemão', 'Ilha do Governador', 'Inhauma', 'Irajá',
,'Jacarezinho', 'Madureira', 'Meier', 'Pavuna', ,
Penha', 'Ramos', 'Vigario Geral',
'Barra da Tijuca', 'Cidade de Deus', 'Jacarepagua
',
'Bangu', 'Campo Grande', 'Guaratiba', 'Realengo', ,
Santa Cruz')))

basenova$'Região Adm.'<- fct_rev(basenova$'Região Adm.')

COR <- c("#DOE0E3", "#FFE5CC", "#CCFFCC", "#FFFFCC", "#E5CCFF")

boxplot_regadm <- basenova %>% ggplot(aes(y=Prêmio, x='Região Adm
.')) +
geom_boxplot(aes(fill= 'Área de Planejamento')) +
theme_minimal() +
ylim(c(0,25000)) +
ylab("Valor do prêmio \n (em reais)") + xlab("")+

```

```

ggtile("") + ggpubr::rotate_x_text() +
coord_flip() + scale_fill_manual(values=COL) + guides(fill=F)

boxplot_regadm

#ggsave("legenda_bp_regadm.pdf")

# Prêmio x Grupo de veículo

# CATEGORIZANDO OS CARROS

baseveiculo = base %>%
  mutate(`Grupo do veículo` = case_when(
    (`Grupo do veículo` == "Hatchback Compacto") |
      ~ "Hatchback Compacto",
    (`Grupo do veículo` == "Hatchback Médio") |
      ~ "Hatchback Médio",
    (`Grupo do veículo` == "Sedan Compacto") |
      ~ "Sedan Compacto",
    (`Grupo do veículo` == "Sedan Grande" | `Grupo do veículo` == "Sedan Médio") |
      ~ "Sedan Gd/Me",
    (`Grupo do veículo` == "Picape C. Dupla Pequena" | `Grupo do veículo` == "Picape C. Dupla Média" | `Grupo do veículo` == "Picape C. Dupla Grande") |
      ~ "Picape C. Dupla",
    (`Grupo do veículo` == "Picape Pequena" | `Grupo do veículo` == "Picape Grande" | `Grupo do veículo` == "Picape Média") |
      ~ "Picape",
    (`Grupo do veículo` == "SUV Grande" | `Grupo do veículo` == "SUV Médio") |
      ~ "SUV Gd/Me",
    (`Grupo do veículo` == "SUV Compacto"))
  )

```

```

~ "SUV Compacto" ,
('Grupo do veículo' == "Crossover") ~
"Crossover",
('Grupo do veículo' == "Coupe" |
 'Grupo do veículo' == "Coupe Conversivel" |
 'Grupo do veículo' == "Coupe Super Esportivo" |
 'Grupo do veículo' == "Roadster" ) ~
"Coupe / Roadster",
('Grupo do veículo' == "Station-Wagon Compacto") ~
"Station-Wagon Compacto",
('Grupo do veículo' == "Station-Wagon Grande" |
 'Grupo do veículo' == "Station-Wagon Medio") ~
"Station-Wagon Gd/Me",
('Grupo do veículo' == "Minivan Compacto" )
~ "Minivan Compacto",
('Grupo do veículo' == "Minivan Grande" |
 'Grupo do veículo' == "Minivan Medio") ~
"Minivan Gd/Me",
('Grupo do veículo' == "Jipe") ~
"Jipe",
('Grupo do veículo' == "Furgao Compacto" |
 'Grupo do veículo' == "Furgao Medio" |
 'Grupo do veículo' == "Multivan") ~
"Multivan/Furgão")
)

baseveiculo<- baseveiculo %>% rename(`Categoria do veículo` =
`Grupo do veículo`)

premio_vei<- na.omit(baseveiculo) %>%
ggplot(aes(y=Prêmio,x='Categoria do veículo',fill='Categoria do
veículo')) +
geom_boxplot(fill="lightblue") +
theme_minimal() +
ylim(c(0,25000)) +

```

```

ylab("Valor do prêmio \n (em reais)") +  

guides(fill=F) +  

coord_flip()  

premio_vei  
  

#ggsave("bp_catvei.pdf")  
  

##### modelos de regressao multipla  
  

baseveiculo <- mutate_if(baseveiculo, is.character, as.factor)  
  

base$Relatividade<- as.factor(base$Relatividade)  

baseveiculo$Relatividade<- as.factor((baseveiculo$Relatividade))  

base$'Área de Planejamento'<- as.factor(base$'Área de  

Planejamento')  

baseveiculo$'Área de Planejamento'<- as.factor(baseveiculo$'Área  

de Planejamento')  
  

#alterando variavel de referencia  
  

baseveiculo$'Região Adm.'<-relevel(baseveiculo$'Região Adm.', ref=  

"Centro")  

base$'Região Adm.'<- relevel(base$'Região Adm.', ref="Centro")  
  

baseveiculo$'Categoria do veículo' <- relevel(baseveiculo$'  

Categoria do veículo',  

ref = "Sedan  

Compacto")  
  

base$'Grupo do veículo' <- relevel(base$'Grupo do veículo',  

ref = "Sedan Compacto")  
  

# Modelo N1 - Normal + AP  

set.seed(100)

```

```
modelo_lm_1<- lm(data = baseveiculo,
                    formula = Prêmio ~ Idade + Sexo + 'Estado civil
                    ' +
                    'Categoria do veículo' + 'Área de Planejamento
                    ')
summary(modelo_lm_1)

# definindo os resíduos estudentizados e os valores ajustados
res.e<- stdres(modelo_lm_1)
y.chap<- fitted(modelo_lm_1)

#criando um tibble com os resíduos e os valores ajustados
valores<- tibble(y.chap,res.e)

#grafico resíduos x valores ajustados
graf_res1<- ggplot(valores, aes(x = y.chap,y = res.e)) +
  geom_point(size=0.5) +
  labs(x='Valor do prêmio ajustado',y='Resíduos estudentizados',
       title='Resíduos estudentizados versus \n valor do prêmio
       ajustado')+
  theme(plot.title=element_text(hjust=0.5)) +
  geom_hline(aes(yintercept = 0), col="red")

graf_res1

#grafico quantis amostrais x quantis dist. normal
graf_qq <- ggplot(modelo_lm_1,
                   aes(sample=res.e)) +
  stat_qq() +
  stat_qq_line(col="red", lwd=0.5) +
  labs(x='Quantis teóricos',y='Quantis amostrais',title='QQ Plot'
       ) +
  theme(plot.title=element_text(hjust=0.5))
graf_qq
```

```
gridExtra::grid.arrange(graf_res1,graf_qq,ncol=2,nrow=1)

# Modelo N2 - Normal + RA
set.seed(100)
modelo_lm_2<- lm(data = baseveiculo,
                   formula = Prêmio ~ Idade + Sexo + 'Estado civil'
                   +
                   'Categoria do veículo' + 'Região Adm.')
summary(modelo_lm_2)

# definindo os resíduos estudentizados e os valores ajustados
res.e2<- stdres(modelo_lm_2)
y.chap2<- fitted(modelo_lm_2)

#criando um tibble com os resíduos e os valores ajustados
valores2<- tibble(y.chap2,res.e2)

#gráfico resíduos x valores ajustados
grafres2<- ggplot(valores2, aes(x = y.chap2,y = res.e2)) +
  geom_point(size=0.5) +
  labs(x='Valor do prêmio ajustado',y='Resíduos estudentizados',
       title='Resíduos estudentizados versus \n valor do prêmio
       ajustado')+
  theme(plot.title=element_text(hjust=0.5)) +
  geom_hline(aes(yintercept = 0), col="red")

#gráfico quantis amostrais x quantis dist. normal
graf_qq2 <- ggplot(modelo_lm_2,
                    aes(sample=res.e2)) +
  stat_qq() +
  stat_qq_line(col="red", lwd=0.5) +
  labs(x='Quantis teóricos',y='Quantis amostrais',title='QQ Plot'
```

```
) +  
theme(plot.title=element_text(hjust=0.5))  
graf_qq2  
  
gridExtra::grid.arrange(grafres2,graf_qq2,ncol=2,nrow=1)  
  
##### MAPA  
  
# Plotando mapa com ggplot  
  
rio = rgdal::readOGR(dsn = ".", layer = "Limite_Bairro")  
  
## Primeiro é preciso converter o objeto criado pela função  
## readOGR para um objeto do tipo sf  
library(sf)  
rio_sf = as(rio, "sf")  
  
rio_sf$REGIAO_ADM[rio_sf$REGIAO_ADM == "COMPLEXO DO ALEMÃfO"] = "  
COMPLEXO DO ALEMÃo"  
rio_sf$REGIAO_ADM<-as.factor(rio_sf$REGIAO_ADM)  
  
##### regiao adm  
  
#write_csv2(base,"tabela_mapa_premio.csv")  
Tab<- read_csv2("tabela_mapa_premio.csv")  
Tab$REGIAO_ADM<- as.factor(Tab$REGIAO_ADM)  
  
merge<- inner_join(rio_sf, Tab, by="REGIAO_ADM")
```

```

#####
# bairro

#write_csv2(base , "tabela_mapa_premio_bairro.csv")
Tab2<- read_csv2("tabela_mapa_premio_bairro.csv")
Tab2$CODBAIRRO<- as.factor(Tab2$CODBAIRRO)

merge2<- inner_join(rio_sf, Tab2, by="CODBAIRRO")
merge2$CODBAIRRO<- as.factor(merge2$CODBAIRRO)

#####

## Criando função que simplifica todos os outros passos
# necessários para plotar o mapa com ggplot
library(ggplot2)
gg_area = function(shapefile, var, log.scale = FALSE, guidetitle
= 'Guide',
                    maptitle = 'Areal data'){

  if(log.scale){var = log(var + 1)}

  ggplot2::ggplot() +
    ggplot2::geom_sf(data = shapefile, ggplot2::aes(fill = var),
    col = "gray", lwd = 0.1) +
    ggplot2::xlab("Longitude") + ggplot2::ylab("Latitude") +
    ggplot2::scale_fill_distiller(name = guidetitle, direction =
    1) +
    ggplot2::ggtitle(maptitle) +
    theme_void() +
    theme(legend.title = ggplot2::element_text(size = 8, face =
    "italic"),
    panel.border = ggplot2::element_rect(fill = NA, color =
    "#bdbdbd", size = ggplot2::rel(1)),
    panel.grid = ggplot2::element_blank(),
    panel.background = ggplot2::element_blank(),
    
```

```
plot.background = ggplot2::element_rect(fill = "#ffffff",
                                         color = NA),
legend.background = ggplot2::element_rect(fill = "transparent",
                                            color = NA),
legend.key = ggplot2::element_rect(fill = "transparent",
                                      color = NA),
axis.text = ggplot2::element_blank(),
axis.title = ggplot2::element_blank(),
plot.title = ggplot2::element_text(size = 12),
legend.text = ggplot2::element_text(size = 8),
plot.subtitle = ggplot2::element_text(size = 7))
}

## Plotando mapa por área de planej
gg_area(shapefile = merge, var = merge$MED_PREMIO,
        maptitle = "", guidetitle = "Valor médio do prêmio")

## Plotando mapa por bairro
gg_area(shapefile = merge2, var = merge2$MED_PREMIO,
        maptitle = "", guidetitle = "Valor médio do prêmio")

#####
#####
#####

# SIMULAÇÃO DE DADOS

# SIMULAÇÃO DE DADOS

# 1) GERAÇÃO DE DADOS USANDO Y ~ LOGNORMAL

# Criando a base de dados
```

```
# variaveis explicativas e o erro:  
  
n<- 1000  
set.seed(123)  
x1<- rnorm(n, mean = 5, sd = 2)  
x2 <- rbinom(n, size = 1, prob = 0.3)  
e <- rnorm(n, mean = 0, sd = 1)  
  
# 2) Criando data frame  
b0 <- 0.5  
b1 <- 0.2  
b2 <- -0.1  
  
y <- exp(b0 + b1 * x1 + b2 * x2 + e)  
sim.dat <- data.frame(y, x1, x2)  
  
# 2) GERAÇÃO DE DADOS USANDO Y ~ Gama  
  
n<- 1000  
set.seed(123)  
x1<- rnorm(n, mean = 5, sd = 2)  
x2 <- rbinom(n, size = 1, prob = 0.3)  
  
# 2) Criando data frame  
a<-2  
b0 <- 0.5  
b1 <- 0.2  
b2 <- -0.1  
mu <- exp(b0 + b1 * x1 + b2 * x2) #fc de ligacao logaritimica  
y<- rgamma(n=n,shape = a, rate = a/mu)  
  
sim.dat <- data.frame(y, x1, x2)  
  
# 3) ESTIMAÇÃO FREQUENTISTA
```

```

# 3.1) AJUSTE LOGNORMAL

freq.mod <- lm(log(y) ~ x1 + x2, data = sim.dat)
summary(freq.mod)
AIC(freq.mod)
confint(freq.mod)

# 3.1) AJUSTE GAMMA

freq.mod <- glm(y ~ x1 + x2, data = sim.dat, family = Gamma(link
= "log"))
summary(freq.mod)
AIC(freq.mod)
confint(freq.mod)

# 4) ESTIMAÇÃO BAYESIANA

# 4.1) AJUSTE LOGNORMAL

# função para o JAGS
bayes.mod2 <- function() {
  for(i in 1:N){ #N é o tamanho da base de dados
    y[i] ~ dlnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * x1[i] + beta2 * x2[i]
  }
  beta0 ~ dnorm(0, .01) # segundo argumento é a precisão
  beta1 ~ dnorm(0, .01)
  beta2 ~ dnorm(0, .01)
  tau ~ dgamma(.01, .01) #priori vaga da gamma
}

# Definindo os vetores:

```

```

y <- sim.dat$y
x1 <- sim.dat$x1
x2 <- sim.dat$x2
N <- nrow(sim.dat)
sim.dat.jags <- as.list(sim.dat)
sim.dat.jags$N <- nrow(sim.dat)

# Definindo os parametros de interesse:
bayes.mod.params <- c("beta0", "beta1", "beta2")

library(R2jags)
set.seed(123)

bayes.mod.fit2 <- jags(data = sim.dat.jags,
                        parameters.to.save = bayes.mod.params, n.
                        chains = 2,
                        n.iter = 40000,
                        n.burnin = 10000, model.file = bayes.mod2,
                        n.thin=30)

print(bayes.mod.fit2)

# 4.2) AJUSTE GAMMA

# 4) Funcao para o JAGS

bayes.mod <- function() {
  for(i in 1:N){
    y[i] ~ dgamma(alfa, alfa/mu[i])
    mu[i] <- exp(beta0 + beta1 * x1[i] + beta2 * x2[i]) # funcao
    logaritimica
  }
}

```

```
beta0 ~ dnorm(0, .01) # media 200 para evitar mts valores
negativos
beta1 ~ dnorm(0, .01)
beta2 ~ dnorm(0, .01)
alfa ~ dunif(0,100)

}

# Definindo os vetores:

y <- sim.dat$y
x1 <- sim.dat$x1
x2 <- sim.dat$x2
N <- nrow(sim.dat)
sim.dat.jags <- as.list(sim.dat)
sim.dat.jags$N <- nrow(sim.dat)

# Definindo os parametros de interesse:

bayes.mod.params <- c("beta0", "beta1", "beta2", "alfa")

library(R2jags)
set.seed(123)

bayes.mod.fit <- jags(data = sim.dat.jags,
parameters.to.save = bayes.mod.params, n.
chains = 2,
n.iter = 40000,
n.burnin = 10000, model.file = bayes.mod.n.
thin = 30)

print(bayes.mod.fit.mcmc)
```

```
# 5) GRAFICOS

library(mcmcplots)
bayes.mod.fit2.mcmc <- as.mcmc(bayes.mod.fit2)

summary(bayes.mod.fit2.mcmc)
mcmcplot(bayes.mod.fit2.mcmc)

library(mcmcplots)
denplot(bayes.mod.fit2.mcmc)

denplot(bayes.mod.fit2.mcmc, parms = c("beta0"), ci=0.95, main = "")
)
abline(v=b0,col="black",lty=2)

denplot(bayes.mod.fit2.mcmc, parms = c("beta1"), ci=0.95, main = "")
)
abline(v=b1,lty=2)

denplot(bayes.mod.fit2.mcmc, parms = c("beta2"), ci=0.95, main = "")
)
abline(v=b2,lty=2)

#####
traplot(bayes.mod.fit2.mcmc, parms = c("beta0"),main = "")
abline(h=b0,lty=2)

traplot(bayes.mod.fit2.mcmc, parms = c("beta1"),main = "")
abline(h=b1,lty=2)
```


Referências

- AKAIKE, H. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, Oxford University Press, v. 60, n. 2, p. 255,265, 1973.
- BANDEIRA, M. do Carmo de O. R. M. Seguro de saúde: Custos de ambulatório - modelização linear generalizada. 2013. Disponível em: <https://repositorio.ul.pt/bitstream/10451/10311/1/ulfc106021_tm_Maria_Carmo_Bandeira.pdf>.
- BERGER, J. O. *Statistical Decision Theory and Bayesian Analysis*. 2. ed. [S.l.]: Springer-Verlag New York, 1985.
- BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *The Psychometric Society*, n. 3, p. 345,370, 1987.
- CASELLA, R. L. B. G. *Statistical Inference*. 2. ed. [S.l.]: Duxbury Press, 2001.
- CNSEG. Revista de seguros. *Revista de Seguros, ANO 92, No. 901, ABRIL/ MAIO/ JUNHO (2017)*, n. 901, p. 12,15, 2017.
- DOBSON, A.; BARNETT, A. *An introduction to generalized linear models*. 4. ed. [S.l.]: CRC Press, 2018.
- DUNCAN, A. *A Practitioner's Guide to Generalized Linear Models*. 3. ed. [S.l.]: TOWERS WATSON, 2007.
- FARIAS, T. A.; JESUS, J. C. de. Tarifação de seguros de automóveis no mercado do nordeste. 2020. Disponível em: <<https://sites.ufpe.br/ixsimpa/wp-content/uploads/sites/38/2020/03/Tarifa%C3%A7%C3%A3o-de-seguros-no-mercado-do-nordeste-revisado.pdf>>.
- FERREIRA, P. P. *Modelos de Precificação e Ruína para Seguros de Curto Prazo*. 1. ed. [S.l.]: FUNENSEG, 2002.
- GAMERMAN, D.; LOPES, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2. ed. [S.l.]: Chapman Hall/CRC, 2006.
- GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. PAMI-6, No. 6 (1984)*, pp. 721-741, PAMI-6,, n. 6, p. 721,741, 1984.
- HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, v. 57, p. 907,109, 1970.

- ISP. *Instituto de Segurança Pública*. 2021. Disponível em: <<http://www.ispvisualizacao.rj.gov.br/Monitoramento.html>>.
- METROPOLIS, N. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, v. 21, n. 6, p. 1087,1092, 1953.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, Vol. 135, No. 3 (1972), pp. 370-384, v. 135, n. 6, p. 370,384, 1972.
- PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B.; SILVA, G. L. *Estatística Bayesiana*. 2. ed. [S.l.]: Fundação Calouste Gulbenkian, 2018.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <<https://www.R-project.org/>>.
- RAO, C. R.; TOUTENBURG, H. *Linear Models: Least Squares and Alternatives*. 2. ed. [S.l.]: Springer, 1999.
- SANTOS, S. T. dos. Construção de uma tarifa de responsabilidade civil automóvel. 2008. Disponível em: <<https://core.ac.uk/download/pdf/303709491.pdf>>.
- SOUZA, S. de. *Seguros Contabilidade, Atuaria e Auditoria*. 2007. ed. [S.l.]: Editora Saraiva, 2018.
- SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit. *Journal Royal Statistical Society*, n. 64, p. 583,639, 2002.
- YAN, X.; SU, X. G. *Linear Regression Analysis: Theory and Computing*. 1. ed. [S.l.]: World Scientific Publishing, 2009.