Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise dos Valores de Mercado e dos Prêmios de Seguro de Veículos em uma Região do Interior Paulista

Eloah Assine Brandini

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise dos Valores de Mercado e dos Prêmios de Seguro de Veículos em uma Região do Interior Paulista

Eloah Assine Brandini
Orientadora: Profª. Drª. Maria Sílvia de Assis Moura

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

An Analysis of Vehicle Market Prices and Insurance Premiums in an Inland Region of São Paulo State

Eloah Assine Brandini
Advisor: Profª. Drª. Maria Sílvia de Assis Moura

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
December 2025

Eloah Assine Brandini

Análise dos Valores de Mercado e dos Prêmios de Seguro de Veículos em uma Região do Interior Paulista

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Eloah Assine Brandini e aprovado pela banca examinadora.

Aprovado em 04 de dezembro de 2025.

Banca Examinadora:

- Profª. Drª. Maria Sílvia de Assis Moura
- Profº. Drº. Francisco Antonio Rojas Rojas
- Profª. Drª. Estela Maris Pereira Bereta

Resumo

Este trabalho investigou a relação entre os valores médios de mercado de veículos populares e os prêmios médios de seguro automotivo na região da Grande Campinas, no período de 2016.1 a 2021.2. A pesquisa teve início com a consolidação e padronização das bases de dados da SUSEP e da Tabela Fipe (via Kaggle), organizadas em ambiente Excel, etapa fundamental para garantir comparabilidade entre modelos e semestres. Foram analisados seis veículos amplamente presentes no mercado nacional: Ford Ka, Hyundai HB20, Nissan March, Renault Sandero, Toyota Etios e Volkswagen Gol.

A análise descritiva inicial mostrou variações relevantes tanto nos valores de mercado quanto nos prêmios médios entre os modelos e ao longo do tempo. Em alguns períodos, os veículos apresentaram valorização expressiva, enquanto os prêmios seguiram padrões distintos, sugerindo que essa relação não evolui de forma proporcional. Esses resultados motivaram a aplicação de duas abordagens estatísticas complementares. O modelo Beta, utilizado para o índice prêmio/valor, evidenciou tendência de redução desse índice em diversos semestres, além de diferenças importantes entre modelos e perfis de segurados. Já o modelo Gama, aplicado ao prêmio médio em valores absolutos, confirmou o papel central do valor do veículo, da indenização média e da exposição na formação do prêmio.

Os diagnósticos e análises de calibração indicaram ajuste adequado em ambos os modelos, reforçando a consistência das conclusões. De modo geral, os resultados mostram que o prêmio médio e o valor de mercado dos veículos não caminham de forma paralela ao longo do tempo. Enquanto alguns veículos apresentaram valorização significativa, o custo relativo do seguro diminuiu, indicando mudanças estruturais no comportamento da carteira e no mercado segurador. A combinação das abordagens relativa e absoluta proporcionou uma visão abrangente da dinâmica entre valor do veículo e prêmio de seguro automotivo.

Palavras-chave: *seguro automotivo, FIPE, prêmios médios, SUSEP, valores de mercado, veículos, análise descritiva, regressão Beta, regressão Gama.*

Abstract

This study examined the relationship between the average market values of popular vehicles and their corresponding automobile insurance premiums in the Greater Campinas region from 2016.1 to 2021.2. The research began with the consolidation and standardization of datasets from SUSEP and the Fipe Table (via Kaggle), organized in Excel to ensure consistency across vehicle models and semesters. Six widely marketed vehicles in Brazil were analyzed: Ford Ka, Hyundai HB20, Nissan March, Renault Sandero, Toyota Etios, and Volkswagen Gol.

The descriptive analysis revealed substantial variation in both market prices and insurance premiums across models and over time, indicating that these variables do not evolve proportionally. These findings supported the use of two complementary statistical approaches. The Beta regression model, applied to the premium-to-value index, showed a downward trend in several semesters and highlighted differences between vehicle models and policyholder profiles. The Gamma regression model, used to analyze premiums in absolute terms, confirmed the influence of vehicle value, average claim cost, and exposure on insurance pricing.

Diagnostic and calibration analyses indicated satisfactory fit for both models. Overall, the results show that insurance premiums and vehicle market values do not follow parallel trajectories over time. While some models exhibited significant appreciation, the relative cost of insurance decreased, suggesting structural changes in portfolio behavior and in the insurance market. The combined use of relative and absolute modeling approaches provided a comprehensive understanding of how multiple factors interact in the formation of automobile insurance premiums.

Keywords: *automobile insurance; FIPE price index; insurance premiums; SUSEP; vehicle market value; passenger vehicles; descriptive analysis; Beta regression; Gamma regression..*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O seguro automotivo desempenha um papel essencial na proteção financeira dos proprietários de veículos e no funcionamento do mercado como um todo. Além de oferecer cobertura diante de eventualidades, ele contribui para a estabilidade econômica das famílias e para a continuidade das atividades de transporte no país. O prêmio de seguro, que representa o valor pago pelo segurado para ter acesso à cobertura contratada, é influenciado por diversos fatores ligados ao perfil do condutor, às características do veículo e às condições do mercado. Entre esses elementos, o valor de mercado do automóvel se destaca por sua relevância, uma vez que tende a refletir diretamente o custo potencial de reposição ou reparo. Assim, compreender como o prêmio médio se relaciona com o valor médio dos veículos é fundamental para analisar o comportamento desse segmento e obter interpretações consistentes sobre sua dinâmica Dória e Gonzaga (2015).

O estudo do seguro automotivo envolve a compreensão simultânea do comportamento do mercado de veículos e da formação dos prêmios de seguro, já que alterações no valor de um automóvel podem influenciar diretamente o custo de sua proteção. Essa relação é especialmente relevante no caso de modelos populares, que representam grande parte da frota brasileira e concentram a maior demanda por seguro. Com esse pano de fundo, este trabalho analisa seis veículos amplamente comercializados no país — Ford Ka, Hyundai HB20, Nissan March, Renault Sandera, Toyota Etios e Volkswagen Gol — considerando seus valores médios de mercado e os prêmios médios de seguro ao longo dos semestres de 2016.1 a 2021.2, na região de Campinas, em São Paulo.

A primeira etapa do estudo envolveu a organização e análise descritiva das bases de dados provenientes da Tabela de Valores Médios de Veículos (Tabela Fipe FIPE - Fundação Instituto de Pesquisas Econômicas (2024)), cujas informações utilizadas neste trabalho

foram obtidas por meio de um conjunto público disponibilizado na plataforma Kaggle (2023), e da AutoSeg (2023) / SUSEP - Superintendência de Seguros Privados (2024). Essa etapa permitiu identificar padrões importantes, como períodos de valorização e desvalorização dos veículos, oscilações no prêmio médio de seguro e diferenças relevantes entre os modelos analisados. A partir dessas informações, foi possível construir um panorama claro da evolução dos valores de mercado e dos prêmios ao longo do tempo, fornecendo a base necessária para aprofundar a investigação.

Com esse contexto estabelecido, o trabalho avança para uma análise estatística voltada a quantificar a relação entre o valor médio dos veículos e o prêmio médio de seguro. O objetivo é avaliar de forma objetiva se o valor de mercado do veículo influencia o prêmio e qual é a intensidade dessa influência, analisando tanto comportamentos proporcionais quanto variações em valores absolutos. Para isso, são empregadas duas modelagens que se complementam. o primeiro é Modelo de Regressão Beta, aplicado à razão entre o prêmio médio e o valor médio do veículo, uma medida que se encontra no intervalo entre zero e um e que permite investigar o comportamento relativo do prêmio em função do valor do bem. A segunda é um Modelo Linear Generalizado com distribuição Gama, ajustado diretamente ao prêmio médio, levando em consideração que essa variável é contínua, positiva e assimétrica.

A combinação dessas abordagens possibilita examinar o fenômeno sob diferentes perspectivas: enquanto a regressão Beta permite avaliar a proporcionalidade entre prêmio e valor do veículo, o modelo Gama possibilita interpretar os efeitos em termos absolutos. Dessa forma, o estudo integra a análise exploratória inicial com técnicas inferenciais que ampliam a compreensão da relação entre essas variáveis.

Este trabalho está organizado da seguinte forma: no Capítulo 2 são apresentados os fundamentos teóricos necessários para a análise, incluindo os conceitos de seguro automotivo e as distribuições estatísticas utilizadas. O Capítulo 3 descreve a construção e padronização das bases de dados e os procedimentos metodológicos adotados. No Capítulo 4, são apresentados os resultados dos modelos ajustados sob as distribuições Beta e Gama, juntamente com as análises exploratórias e diagnósticos. Por fim, o Capítulo 5 reúne as conclusões gerais do estudo e as sugestões para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

O estudo da relação entre o valor de mercado dos veículos e o prêmio médio de seguro requer o uso de métodos estatísticos que permitam modelar variáveis com comportamentos distintos. No caso dos dados analisados neste trabalho, o prêmio médio apresenta valores contínuos e positivos, enquanto o índice formado pela razão entre o prêmio médio e o valor médio do veículo é uma variável contínua limitada entre zero e um. Para lidar com essas duas naturezas de variáveis, utilizam-se abordagens fundamentadas na estrutura teórica dos Modelos Lineares Generalizados, apresentados em Demétrio (2002) e Paula (2013). Além disso, a regressão Beta utilizada neste estudo segue a formulação proposta por Ferrari e Cribari-Neto (2004), referência fundamental para modelagem de proporções.

Ao longo desta seção, o índice $i = 1, 2, \ldots, n$ será utilizado para representar cada unidade de análise do banco de dados, isto é, cada combinação de modelo de veículo, semestre e região considerada. Assim, sempre que aparecer o índice i em uma expressão, ele estará se referindo a uma linha específica da base de dados utilizada neste trabalho.

2.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (GLM), conforme descritos em Paula (2013), ampliam os modelos lineares clássicos ao permitir que a variável resposta siga distribuições pertencentes à família exponencial. Essa família inclui distribuições como Normal, Poisson, Binomial e Gama, o que possibilita ajustar o modelo de acordo com o comportamento específico dos dados observados, mantendo uma estrutura comum para a ligação entre a média da variável resposta e as covariáveis.

A forma geral de uma distribuição pertencente à família exponencial pode ser escrita

como

$$f(y_i \mid \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{2.1}$$

em que:

- y_i denota o valor observado da variável resposta na unidade de análise i;
- $f(y_i \mid \theta_i, \phi)$ é a função densidade de probabilidade (ou de massa, no caso discreto) da família exponencial avaliada na i-ésima observação, condicionada aos parâmetros θ_i e ϕ;
- θ_i é o parâmetro natural (ou canônico) associado à unidade i, usualmente definido em função do preditor linear do modelo e responsável por conectar a média condicional da resposta à estrutura de regressão;
- $b(\theta_i)$ é a função cumulante (ou função de log-partição) da distribuição, que garante a normalização da densidade e determina propriedades importantes da família, como a média e a variância em termos do parâmetro natural;
- $a(\phi)$ é uma função estritamente positiva que depende apenas do parâmetro de dispersão ϕ, atuando como fator de escala na variância da distribuição;
- $c(y_i, \phi)$ é a função de normalização (ou função base) que depende dos dados observados e, eventualmente, do parâmetro de dispersão, completando a especificação da densidade na forma exponencial;
- ϕ é o parâmetro de dispersão comum a todas as unidades de análise, controlando a variabilidade da resposta em torno da média e permitindo acomodar diferentes graus de heterogeneidade;
- $i = 1, 2, \ldots, n$ identifica cada uma das n unidades de análise do conjunto de dados.

A relação entre a média condicional da variável resposta e as covariáveis é estabelecida por meio de uma função de ligação,

$$g(\mu_i) = \eta_i = X_i\beta, \tag{2.2}$$

em que:

- $g(\cdot)$ é a função de ligação escolhida para o modelo, responsável por transformar a média condicional μ_i para a escala do preditor linear η_i;
- $\mu_i = \mathbb{E}(Y_i)$ é a média condicional da variável resposta na unidade i, determinada pela combinação das covariáveis e dos parâmetros do modelo;
- η_i é o preditor linear correspondente à unidade de análise i, que sintetiza, em uma única quantidade escalar, o efeito das covariáveis sobre a média condicional da resposta;
- X_i é o vetor linha de covariáveis observado na unidade i, contendo os valores das variáveis explicativas (incluindo, quando apropriado, o termo constante associado ao intercepto do modelo);
- β é o vetor coluna de parâmetros desconhecidos a serem estimados, que quantificam a contribuição de cada covariável na explicação da média condicional da resposta;
- $i = 1, 2, \ldots, n$ identifica cada uma das n unidades de análise do conjunto de dados.

De acordo com Paula (2013), a estimação de β é feita pelo método da máxima verossimilhança, resolvida numericamente por meio de um algoritmo de mínimos quadrados ponderados iterativos. Em termos matriciais, a atualização pode ser escrita como

$$\beta^{(\mathbf{t+1})} = \left(\mathbf{X}^\top \mathbf{W}^{(\mathbf{t})} \mathbf{X}\right)^{-\mathbf{1}} \mathbf{X}^\top \mathbf{W}^{(\mathbf{t})} \mathbf{z}^{(\mathbf{t})}, \tag{2.3}$$

em que:

- $\boldsymbol{\beta}^{(t)}$ é o vetor de coeficientes estimado na iteração t, representando os efeitos das covariáveis sobre a resposta (prêmio médio ou índice prêmio/valor) no contexto deste estudo;
- X é a matriz de covariáveis do modelo, formada pelas características consideradas no trabalho — como grupo do veículo, sexo do condutor, faixa etária e semestre — sendo cada linha correspondente a uma unidade de análise agregada;
- $W^{(t)}$ é a matriz diagonal de pesos calculada na iteração t, ajustando a influência de cada unidade de análise no processo de estimação de acordo com a estrutura da distribuição utilizada (Gama ou Beta);

- $z^{(t)}$ é o vetor de respostas ajustadas na iteração t, construído para aproximar o problema não linear da máxima verossimilhança por um sistema de equações lineares ponderadas, facilitando a atualização de $\boldsymbol{\beta}$;

- $i = 1, 2, \ldots, n$ identifica cada uma das unidades de análise utilizadas no modelo, resultantes da agregação das informações de prêmio, exposição e características do segurado e do veículo.

Essa estrutura geral de Modelos Lineares Generalizados será utilizada como base para especificar o Modelos de Regressão Beta e o Modelos de Regressão Gama, de acordo com a natureza de cada variável resposta analisada neste trabalho.

2.2 Distribuição Beta

A regressão Beta é adequada para variáveis contínuas restritas ao intervalo aberto entre zero e um. Neste estudo, essa modelagem é aplicada ao índice formado pela razão entre o prêmio médio de seguro e o valor médio de mercado do veículo. A formulação adotada segue Ferrari e Cribari-Neto (2004), que propõem uma parametrização em termos de média e precisão para a distribuição Beta.

A densidade de probabilidade da distribuição Beta, na parametrização média–precisão, é dada por

$$f(y_i \mid \mu_i, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma\big((1-\mu_i)\phi\big)}\, y_i^{\mu_i\phi-1}(1-y_i)^{(1-\mu_i)\phi-1}, \quad 0 < y_i < 1, \tag{2.4}$$

em que:

- y_i é o valor observado do índice prêmio/valor na unidade de análise i, representando a proporção entre o prêmio médio e o valor médio do veículo no contexto deste estudo;

- μ_i é a média condicional de Y_i na unidade i, interpretada como a proporção média esperada do índice prêmio/valor, determinada pelas características do segurado e do veículo consideradas no modelo;

- $\phi > 0$ é o parâmetro de precisão da distribuição Beta, controlando o grau de concentração dos valores de Y_i em torno de μ_i; valores maiores de ϕ implicam menor variabilidade e maior aderência dos dados à média estimada;

- $\Gamma(\cdot)$ denota a função Gama, utilizada na expressão da densidade para garantir a normalização da distribuição;

- $i = 1, 2, \ldots, n$ identifica cada unidade de análise utilizada no modelo, correspondendo às combinações agregadas de região, grupo do veículo, sexo do condutor, faixa etária e semestre.

Na parametrização clássica, os parâmetros de forma α e β podem ser escritos em função de μ e ϕ como

$$\alpha = \mu\phi, \qquad \beta = (1 - \mu)\phi, \tag{2.5}$$

em que:

- μ representa a média da distribuição Beta na parametrização média–precisão, interpretada neste trabalho como a proporção média esperada do índice prêmio/valor;

- $\phi > 0$ é o parâmetro de precisão da distribuição, controlando o grau de concentração dos valores em torno de μ; na parametrização clássica, ele pode ser visto como o "tamanho total" $\alpha + \beta$;

- α é o parâmetro de forma associado ao comportamento da densidade na vizinhança de zero, influenciando a inclinação e a curvatura da função densidade quando y assume valores próximos de 0;

- β é o parâmetro de forma associado ao comportamento da densidade na vizinhança de um, determinando como a distribuição se comporta quando y assume valores próximos de 1.

A média e a variância da variável resposta Y_i são dadas por Ferrari e Cribari-Neto (2004) por

$$\mathbb{E}(Y_i) = \mu_i, \tag{2.6}$$

e

$$\mathrm{Var}(Y_i) = \frac{\mu_i(1 - \mu_i)}{1 + \phi}, \tag{2.7}$$

em que:

- $\mathbb{E}(Y_i)$ representa o valor esperado da variável resposta na unidade de análise i, indicando a proporção média do índice prêmio/valor prevista pelo modelo;

- $\text{Var}(Y_i)$ corresponde à variância condicional de Y_i, expressando o grau de dispersão dos valores observados em torno da média μ_i para a unidade i;

- μ_i é a média condicional do índice prêmio/valor na unidade i, determinada pela estrutura de regressão do modelo e refletindo a proporção média esperada para aquela combinação específica de características dos segurados e dos veículos;

- ϕ é o parâmetro de precisão da distribuição Beta, responsável por controlar a concentração dos valores em torno da média; valores maiores de ϕ implicam menor variabilidade e maior precisão do modelo;

- $i = 1, 2, \ldots, n$ identifica cada unidade de análise considerada após a etapa de agregação dos dados.

A função de ligação utilizada neste trabalho para a regressão Beta é a logito,

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i, \tag{2.8}$$

em que:

- $g(\cdot)$ denota a função de ligação logito, definida por $g(\mu_i) = \log\big(\mu_i/(1-\mu_i)\big)$, a qual transforma a média $\mu_i \in (0,1)$ para a escala real;

- η_i é o preditor linear correspondente à unidade de análise i, obtido a partir da combinação das covariáveis por meio dos coeficientes de regressão do modelo;

- μ_i é a média condicional do índice prêmio/valor na unidade i, interpretada como a proporção média esperada do prêmio em relação ao valor do veículo, condicionada às características consideradas no modelo;

- $i = 1, 2, \ldots, n$ identifica cada unidade de análise do conjunto de dados.

Essa modelagem permite estudar o comportamento relativo do prêmio médio em relação ao valor médio do veículo, analisando se o índice prêmio/valor permanece aproximadamente constante ou se apresenta variações sistemáticas entre os modelos de veículos e ao longo dos semestres.

2.3 Distribuição Gama

A regressão Gama é apropriada para variáveis contínuas, estritamente positivas e assimétricas, como os prêmios médios de seguro. Neste trabalho, ela é utilizada para modelar diretamente o prêmio médio, permitindo avaliar a relação em termos absolutos entre o valor médio do veículo e o valor pago pelo seguro. A formulação adotada segue a apresentação da distribuição Gama no contexto de Modelos Lineares Generalizados em Paula (2013).

A densidade de probabilidade da distribuição Gama pode ser escrita como

$$f(y_i \mid k, \theta) = \frac{1}{\Gamma(k)\,\theta^k}\, y_i^{k-1} \exp\left(-\frac{y_i}{\theta}\right), \qquad y_i > 0, \tag{2.9}$$

em que:

- y_i é o prêmio médio observado na unidade de análise i, representando o valor positivo a ser modelado pela distribuição Gama no contexto deste estudo;
- $k > 0$ é o parâmetro de forma da distribuição Gama, responsável por determinar a assimetria e a curvatura da densidade; valores menores de k resultam em maior assimetria à direita, enquanto valores maiores tornam a distribuição mais simétrica;
- $\theta > 0$ é o parâmetro de escala, controlando a dispersão da distribuição; em conjunto com k, define a média ($k\theta$) e a variância ($k\theta^2$) da resposta;
- $\Gamma(k)$ denota a função Gama avaliada em k, aparecendo na expressão da densidade para garantir a normalização da distribuição;
- $i = 1, 2, \ldots, n$ identifica cada uma das unidades de análise utilizadas no modelo, resultantes da agregação das informações de prêmio médio, características do condutor e especificações do veículo.

A média e a variância de Y_i na parametrização forma–escala são dadas por Paula (2013) por

$$\mu_i = k\theta, \tag{2.10}$$

e

$$\mathrm{Var}(Y_i) = k\theta^2, \tag{2.11}$$

em que:

- μ_i é a média condicional de Y_i na unidade de análise i, interpretada neste trabalho como o valor esperado do prêmio médio de seguro para aquela combinação específica de características do segurado e do veículo;
- $\text{Var}(Y_i)$ representa a variância condicional de Y_i na unidade i, indicando o grau de dispersão dos valores de prêmio médio em torno da média μ_i; essa variabilidade é diretamente proporcional aos parâmetros k e θ;
- k e θ são, respectivamente, os parâmetros de forma e de escala da distribuição Gama, determinando suas propriedades fundamentais: a média é dada por $k\theta$ e a variância por $k\theta^2$, de modo que valores maiores de θ aumentam a dispersão, enquanto k controla a assimetria e o formato da distribuição.

No contexto de Modelos Lineares Generalizados, é comum utilizar uma parametrização em termos de média μ_i e parâmetro de dispersão $\phi > 0$, na qual a variância pode ser escrita como

$$\text{Var}(Y_i) = \phi\, \mu_i^2, \tag{2.12}$$

em que:

- $\text{Var}(Y_i)$ é a variância condicional de Y_i na unidade de análise i, indicando o grau de dispersão dos valores de prêmio médio em torno da média esperada;
- μ_i é a média condicional de Y_i na unidade i, interpretada neste trabalho como o valor médio esperado do prêmio de seguro para cada combinação de características do segurado e do veículo;
- $\phi > 0$ é o parâmetro de dispersão do modelo Gama, responsável por ajustar a variabilidade da resposta em relação ao valor da média; nessa parametrização, valores maiores de ϕ implicam maior dispersão dos prêmios médios em torno de μ_i;
- $i = 1, 2, \ldots, n$ identifica as unidades de análise utilizadas no modelo, resultantes da agregação das informações de prêmio médio, exposição e perfis dos segurados.

A função de ligação utilizada para a regressão Gama neste trabalho é a logarítmica,

$$g(\mu_i) = \log(\mu_i) = \eta_i, \tag{2.13}$$

em que:

- $g(\cdot)$ é a função de ligação logarítmica, responsável por transformar a média condicional μ_i (sempre positiva no caso da regressão Gama) para a escala real, permitindo modelar o prêmio médio por meio de um preditor linear;

- η_i é o preditor linear associado à unidade de análise i, resultante da combinação das covariáveis consideradas no estudo — como grupo do veículo, sexo do condutor, faixa etária e semestre — e dos coeficientes estimados pelo modelo;

- μ_i é a média condicional do prêmio médio na unidade i, interpretada como o valor médio esperado do prêmio de seguro para a combinação específica de características do segurado e do veículo presentes naquela unidade;

- $i = 1, 2, \ldots, n$ identifica cada uma das unidades de análise utilizadas no modelo, originadas após a agregação dos dados de prêmio, exposição e perfis dos segurados.

No contexto deste trabalho, a regressão Gama é utilizada sem termo de exposição (off-set), uma vez que a variável resposta corresponde ao prêmio médio de seguro, já expresso em valores monetários absolutos. Como não se trata de uma taxa associada a uma quantidade de exposição, mas de um valor médio consolidado para cada combinação de modelo de veículo e semestre, a inclusão de um off-set não é necessária. Assim, o preditor linear é formado exclusivamente pelas covariáveis selecionadas, seguindo a formulação padrão da regressão Gama apresentada em Paula (2013).

Essa modelagem permite interpretar os coeficientes em termos multiplicativos sobre a média do prêmio, o que é conveniente em aplicações nas quais variações relativas são mais informativas do que variações absolutas em unidades monetárias. Dessa forma, a regressão Gama complementa a regressão Beta ao fornecer uma visão em termos absolutos do prêmio médio de seguro.

2.4 Diagnósticos

A verificação do ajuste dos modelos constitui uma etapa fundamental para assegurar a validade das conclusões obtidas a partir das regressões Beta e Gama. Conforme discutido em Paula (2013), diferentes medidas e ferramentas gráficas podem ser utilizadas para avaliar a qualidade do ajuste em Modelos Lineares Generalizados, permitindo identificar possíveis inadequações e comparar alternativas de modelagem.

Entre os critérios utilizados neste trabalho, destaca-se o critério de informação de Akaike (AIC), que possibilita comparar modelos ajustados a partir de diferentes especificações e selecionar aquele que apresenta melhor equilíbrio entre qualidade de ajuste e complexidade. Em geral, valores menores de AIC indicam modelos mais parcimoniosos e com melhor desempenho relativo.

Além do AIC, foram analisados resíduos de Pearson, resíduos de desvio, gráficos de resíduos versus valores ajustados e gráficos quantil–quantil. Esses elementos auxiliam na detecção de padrões sistemáticos que possam indicar má especificação do modelo, violação de pressupostos ou inadequação da função de ligação utilizada. A análise residual é especialmente útil para avaliar se os valores ajustados capturam adequadamente a estrutura dos dados ou se há indícios de heterogeneidade não modelada.

Medidas de influência também foram avaliadas, com destaque para a distância de Cook, que permite identificar unidades de análise com impacto potencialmente elevado sobre os parâmetros estimados. A identificação dessas observações é importante para compreender se resultados específicos podem estar sendo influenciados por valores extremos ou por combinações atípicas das covariáveis.

Os procedimentos diagnósticos empregados neste estudo foram realizados no software R Core Team (2024), complementados por representações gráficas que auxiliam na interpretação dos resultados. A combinação entre critérios de ajuste, análise residual e medidas de influência forneceu uma base consistente para validar as modelagens adotadas, contribuindo para a robustez e a confiabilidade das conclusões obtidas.

Capítulo 3

Materias e Métodos

3.1 Construção do banco de dados

A construção do banco de dados constituiu uma etapa fundamental deste estudo, uma vez que as informações necessárias à análise não estavam disponíveis em uma base única, pronta para uso. Os dados utilizados foram obtidos a partir de fontes distintas, com estruturas próprias, diferentes níveis de agregação e referências temporais específicas, o que exigiu um processo cuidadoso de coleta, filtragem, padronização e integração antes da realização das análises estatísticas.

Os dados referentes ao valor de mercado dos veículos foram extraídos da Tabela FIPE FIPE - Fundação Instituto de Pesquisas Econômicas (2024), por meio de uma base disponibilizada na plataforma Kaggle Kaggle (2023). Essa base contém registros individuais de preços, organizados por modelo, ano e período de referência, e não valores médios previamente consolidados. A partir desses dados, realizou-se inicialmente a filtragem dos modelos de veículos com informações disponíveis de forma consistente ao longo do período de interesse. Essa etapa foi decisiva para a definição do conjunto final de modelos analisados no estudo, garantindo a continuidade das séries temporais e a comparabilidade entre os veículos selecionados.

Após a definição dos modelos, os dados da FIPE foram organizados de modo a permitir a construção de valores médios semestrais para cada veículo. Para isso, os registros individuais de preços foram agregados por modelo e semestre, resultando em uma medida representativa do valor médio de mercado em cada período considerado. Essa consolidação foi necessária para adequar a base da FIPE ao formato requerido pela análise e possibilitar sua integração com os dados provenientes da SUSEP, cuja estrutura já se

encontra organizada em base semestral.

Os dados relativos ao seguro automotivo foram obtidos diretamente no site da Superintendência de Seguros Privados (SUSEP), por meio do sistema AutoSeg AutoSeg (2023). Nessa etapa, procedeu-se à filtragem da categoria de automóveis e da região de interesse, restringindo os dados ao município de Campinas/SP. Além disso, os dados foram filtrados e baixados semestre a semestre, procedimento que permitiu definir explicitamente a dimensão temporal do estudo e assegurar a correspondência entre os períodos analisados nas duas bases.

Com ambas as fontes organizadas de forma compatível, realizou-se a padronização das estruturas, contemplando a harmonização dos nomes dos modelos, a organização das referências temporais e a verificação da coerência interna das variáveis. Foram conduzidas etapas de limpeza dos dados, incluindo a identificação e eliminação de registros duplicados, bem como a inspeção de valores inconsistentes ou incompatíveis com os critérios definidos para o estudo. Esse processo foi conduzido de forma conservadora, de modo a preservar a integridade das informações e evitar a introdução de distorções artificiais nos dados.

Na sequência, as bases da FIPE e da SUSEP foram integradas, associando, para cada modelo de veículo e semestre de referência, o valor médio de mercado e o respectivo prêmio médio de seguro. Essa integração resultou na construção de um banco de dados único e consolidado, no qual cada observação representa simultaneamente as duas dimensões centrais do estudo. Ressalta-se que a principal variável analisada sob a abordagem proporcional — o índice prêmio/valor — não estava disponível de forma direta em nenhuma das fontes originais, sendo construída especificamente a partir dessa etapa de integração.

Por fim, o banco de dados consolidado passou por verificações finais de consistência, assegurando a ausência de duplicidades remanescentes, a coerência dos valores obtidos e a adequação da base às análises estatísticas subsequentes. Dessa forma, a etapa de construção do banco de dados configurou-se como parte essencial do método adotado, viabilizando a aplicação dos modelos estatísticos propostos e garantindo que os resultados obtidos refletissem efetivamente o comportamento conjunto do valor de mercado dos veículos e do prêmio médio de seguro ao longo do período analisado.

3.2 Bases de dados e metodologia

A metodologia empregada neste trabalho foi estruturada de forma a integrar informações provenientes de diferentes fontes e aplicar modelos estatísticos compatíveis com a natureza das variáveis analisadas. O objetivo central consistiu em investigar a relação entre o valor de mercado dos veículos e o prêmio médio de seguro automotivo, tanto sob uma perspectiva proporcional, por meio do índice prêmio/valor, quanto sob uma perspectiva absoluta, baseada no valor monetário do prêmio.

O estudo utilizou duas bases públicas amplamente reconhecidas pela qualidade e abrangência das informações. A primeira delas é a Tabela FIPE FIPE - Fundação Instituto de Pesquisas Econômicas (2024), cujos valores foram obtidos a partir de um conjunto de dados disponibilizado na plataforma Kaggle (2023). O banco disponibilizado no Kaggle contém registros individuais de preços e não valores médios previamente consolidados. Assim, para adequar essa base ao formato requerido pela análise e torná-la comparável aos dados da SUSEP, foi necessário calcular, para cada modelo de veículo, o valor médio correspondente a cada semestre do período estudado. Esses valores consolidados representam o valor de mercado de referência utilizado em todas as etapas da modelagem.

A segunda base corresponde aos prêmios médios de seguro disponibilizados pela SUSEP - Superintendência de Seguros Privados (2024), extraídos dos painéis públicos da instituição AutoSeg (2023) e filtrados especificamente para a categoria de automóveis na região de Campinas/SP. Como a SUSEP já apresenta os prêmios em formato semestral, a construção dos valores médios da FIPE permitiu alinhar a granularidade temporal entre as duas bases. Após esse alinhamento, cada observação do banco final passou a representar, simultaneamente, o valor médio FIPE (derivado dos dados do Kaggle) e o prêmio médio correspondente ao mesmo modelo de veículo e ao mesmo semestre.

Foram considerados seis modelos de grande representatividade no mercado nacional — Ford Ka, Hyundai HB20, Nissan March, Renault Sandero, Toyota Etios e Volkswagen Gol — avaliados na condição de veículos zero quilômetro. Para cada modelo, foram coletados valores médios semestrais correspondentes ao período de 2016.1 a 2021.2. A utilização de semestres como unidade temporal permitiu reduzir oscilações pontuais e obter estimativas mais estáveis, facilitando a comparação entre os valores de mercado e os prêmios de seguro ao longo do tempo.

Antes da modelagem, as bases passaram por uma etapa de padronização, que in-

cluiu a harmonização dos nomes dos modelos, a reorganização dos períodos e a filtragem para manter somente observações referentes à região de Campinas/SP. Procedimentos de limpeza foram aplicados para eliminar registros duplicados e verificar a consistência dos valores fornecidos pelas duas fontes. Valores extremamente discrepantes foram avaliados e tratados de forma conservadora, evitando que observações atípicas tivessem influência excessiva sobre os modelos ajustados. Após essa etapa, foi calculado o índice prêmio/valor, definido como a razão entre o prêmio médio do seguro e o valor médio do veículo, o qual serviu como variável resposta na regressão Beta. O prêmio médio foi utilizado diretamente como variável resposta no modelo Gama.

Foram ajustados dois modelos estatísticos principais: (i) um modelo Beta, adequado a variáveis contínuas restritas ao intervalo entre zero e um, aplicado ao índice prêmio/valor; e (ii) um modelo Gama com função de ligação logarítmica, apropriado para variáveis contínuas, positivas e assimétricas, aplicado diretamente ao prêmio médio de seguro. Ambos os modelos foram estimados por máxima verossimilhança com o uso do software R Core Team (2024). Para o ajuste do modelo Beta, empregou-se a função `betareg()` do pacote `betareg`; para o modelo Gama, utilizou-se a função `glm()` com família Gama e ligação logarítmica, conforme a formulação apresentada em Paula (2013).

Cabe ressaltar que a especificação dos modelos não incluiu termo de offset. Essa escolha decorre da própria estrutura dos dados: tanto os valores médios dos veículos quanto os prêmios dos seguros correspondem a médias consolidadas por semestre, de modo que cada observação representa um valor agregado e não uma medida dependente de exposição individual, como quantidade de apólices, tempo em risco ou número de segurados. Dessa forma, não há variável estrutural que justifique a inclusão de um offset, e os modelos ajustados refletem diretamente o efeito das covariáveis sobre a média da variável resposta.

A seleção dos modelos foi baseada no critério de informação de Akaike (AIC), que permitiu comparar especificações alternativas e escolher aquelas com melhor equilíbrio entre ajuste e parcimônia. Medidas complementares, como o pseudo-R^2, foram utilizadas para avaliar a proporção da variabilidade explicada. Também foram empregados critérios gráficos obtidos por meio da análise dos resíduos padronizados e da distância de Cook, seguindo as recomendações apresentadas em Paula (2013).

As análises foram conduzidas com apoio dos pacotes `dplyr`, `ggplot2`, `betareg`, `car`, `performance` e `MASS`, que ofereceram suporte para manipulação dos dados, estimação dos

modelos, visualização gráfica e geração dos diagnósticos. Os resultados foram organizados em tabelas e figuras que sintetizam as estimativas obtidas, permitindo interpretar os efeitos associados ao valor médio dos carros e aos semestres para cada um dos modelos de veículos.

A estrutura metodológica adotada permitiu combinar duas perspectivas complementares: uma abordagem relativa, baseada no índice prêmio/valor, e uma abordagem absoluta, centrada no valor monetário dos prêmios. Essa integração garantiu uma análise abrangente da relação entre valor de mercado e custo do seguro automotivo, fornecendo bases estatísticas sólidas para as discussões e conclusões apresentadas nas seções seguintes.

Capítulo 4

Resultados

Nesta seção são apresentados e discutidos os principais resultados obtidos a partir dos modelos ajustados. Inicialmente, analisa-se o comportamento do índice prêmio/valor, definido como a razão entre o prêmio médio de seguro e o valor médio dos veículos, utilizando um modelo de regressão Beta. Essa abordagem permite avaliar o custo relativo do seguro em comparação ao valor de mercado de cada modelo.

Em seguida, discute-se o ajuste do modelo de regressão Gama para o prêmio médio de seguro, o que possibilita examinar essa relação em escala absoluta, considerando diretamente os valores monetários dos prêmios. As duas modelagens são complementares e contribuem para uma compreensão mais ampla de como o valor médio dos veículos se relaciona com o prêmio de seguro ao longo do período de 2016.1 a 2021.2, abrangendo os seis modelos estudados.

4.1 Modelo de Regressão Beta para o Índice Prêmio/Valor

4.1.1 Formulação do Modelo de Regressão Beta

A modelagem do índice prêmio/valor foi conduzida por meio de um Modelo de Regressão Beta com função de ligação logito para a média, adequado para variáveis restritas ao intervalo (0, 1). Seja Y_i o índice prêmio/valor da i-ésima observação. Assume-se que

$$Y_i \sim \text{Beta}(\mu_i, \phi_i),$$

em que μ_i representa a média condicional do índice e ϕ_i é o parâmetro de precisão. A relação entre a média e as covariáveis é especificada pelo preditor linear no domínio da

função logito. A formulação reduzida do modelo utilizado neste estudo é dada por:

$$\eta_i = \beta_0 + \boldsymbol{\beta}_{\text{sem}}^{\top}\mathbf{Z}_{\text{sem},i} + \boldsymbol{\beta}_{\text{grupo}}^{\top}\mathbf{Z}_{\text{grupo},i} + \boldsymbol{\beta}_{\text{sexo}}^{\top}\mathbf{Z}_{\text{sexo},i} + \boldsymbol{\beta}_{\text{faixa}}^{\top}\mathbf{Z}_{\text{faixa},i} +$$

$$+ \boldsymbol{\beta}_{\text{sexo}\times\text{faixa}}^{\top}\mathbf{Z}_{\text{sexo}\times\text{faixa},i} + \beta_{\text{IS}}\log(\text{IS}_i) + \beta_{\text{valor}}\log(\text{Valor}_i) + \beta_{\text{exp}}\log(\text{Expostos}_i), \quad (4.1)$$

em que $\eta_i = \text{logit}(\mu_i)$.

Além do submodelo da média, o modelo Beta permite que o parâmetro de precisão também varie segundo covariáveis. Neste estudo, conforme ajustado no software R, a precisão é modelada como

$$\log(\phi_i) = \gamma_0 + \gamma_{\text{exp}}\ \log(\text{Expostos}_i), \quad (4.2)$$

o que possibilita capturar mudanças na variabilidade do índice prêmio/valor associadas ao nível de exposição da carteira.

Os componentes do preditor linear e do submodelo da precisão podem ser descritos da seguinte forma:

- Y_i: índice prêmio/valor, definido como a razão entre o prêmio médio e o valor médio do veículo na observação i;
- μ_i: média condicional do índice prêmio/valor, representando a proporção média esperada para a unidade de análise i;
- ϕ_i: parâmetro de precisão da distribuição Beta, controlando o grau de concentração dos valores de Y_i em torno de μ_i;
- $\eta_i = \text{logit}(\mu_i)$: preditor linear da média, expresso na escala da função de ligação;
- β_0: intercepto associado à combinação de categorias de referência para todas as variáveis categóricas;
- $\mathbf{Z}_{\text{sem},i}$ e $\boldsymbol{\beta}_{\text{sem}}$: vetor de indicadores e coeficientes para os semestres de referência (2016.1 a 2021.2);
- $\mathbf{Z}_{\text{grupo},i}$ e $\boldsymbol{\beta}_{\text{grupo}}$: variáveis dummy e coeficientes correspondentes aos seis modelos de veículos considerados no estudo;
- $\mathbf{Z}_{\text{sexo},i}$ e $\boldsymbol{\beta}_{\text{sexo}}$: indicadores e coeficientes referentes aos níveis da variável sexo do condutor;

- $\mathbf{Z}_{\text{faixa},i}$ e $\boldsymbol{\beta}_{\text{faixa}}$: indicadores e coeficientes das faixas etárias utilizadas na análise;
- $\mathbf{Z}_{\text{sexo}\times\text{faixa},i}$ e $\boldsymbol{\beta}_{\text{sexo}\times\text{faixa}}$: termos de interação entre sexo e faixa etária, permitindo avaliar diferenças adicionais entre perfis combinados de condutores;
- $\log(\text{IS}_i)$ e β_{IS}: logaritmo da indenização média da unidade i e seu coeficiente associado, capturando o efeito relativo da severidade das indenizações sobre o índice prêmio/valor;
- $\log(\text{Valor}_i)$ e β_{valor}: logaritmo do valor médio do veículo e coeficiente associado, refletindo a associação entre o preço do automóvel e a proporção prêmio/valor;
- $\log(\text{Expostos}_i)$ e β_{exp}: logaritmo do número de expostos na unidade i, incorporado ao preditor da média para ajustar diferenças de porte da carteira;
- γ_0 e γ_{exp}: parâmetros do submodelo da precisão, em que γ_{exp} quantifica como o nível de exposição influencia a variabilidade do índice prêmio/valor.

Essa formulação resume de maneira compacta a estrutura utilizada na modelagem Beta, mantendo coerência com o procedimento de ajuste implementado no R e permitindo interpretar adequadamente os efeitos estimados sobre o índice prêmio/valor.

4.1.2 Análise Exploratória do Índice Prêmio/Valor

Antes do ajuste do Modelo de Regressão Beta, foi feita uma análise descritiva do índice prêmio/valor, definido como a razão entre o prêmio médio de seguro e o valor médio FIPE para cada combinação de modelo de veículo, semestre, sexo do condutor, faixa etária e tipo de pessoa. A Tabela 4.1 apresenta um resumo das principais medidas descritivas do índice no período de 2016.1 a 2021.2.

Tabela 4.1: Medidas descritivas do índice prêmio/valor no período de 2016.1 a 2021.2.

n	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
859	0,0106	0,0224	0,0264	0,0282	0,0320	0,0896

Observa-se que o índice prêmio/valor apresenta mediana em torno de 0,026 e média próxima de 0,028, o que corresponde, de forma aproximada, a uma razão entre 2,5% e 3% do valor do veículo. O valor máximo (cerca de 0,090) indica a presença de poucas combinações com índice relativamente elevado, enquanto o primeiro e o terceiro quartis

mostram que a maior parte das observações se concentra em um intervalo relativamente estreito, entre aproximadamente 0,022 e 0,032.

A distribuição global do índice pode ser visualizada por meio do histograma da Figura 4.1, que confirma a assimetria à direita, com concentração de observações entre 0,02 e 0,035 e poucos valores acima de 0,05. Esse comportamento é coerente com a escolha da distribuição Beta para modelagem do índice, conforme discutido na fundamentação teórica.

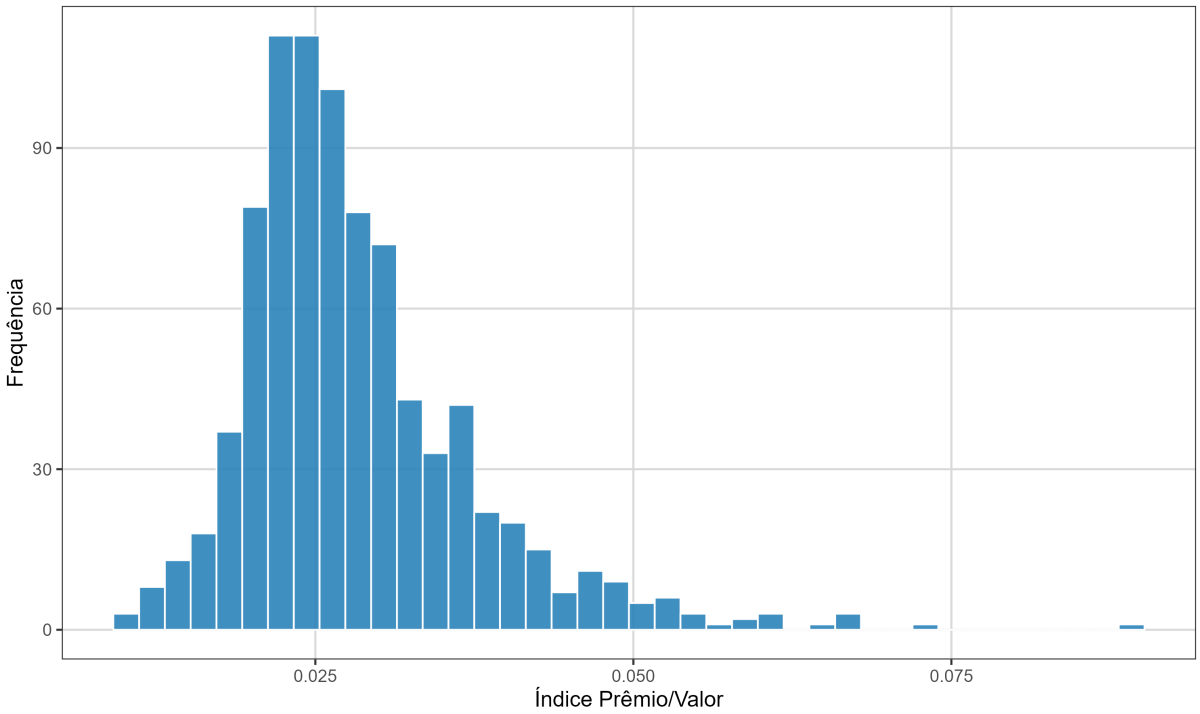

Figura 4.1: Distribuição do índice prêmio/valor no período de 2016.1 a 2021.2.

A Tabela 4.2 resume o comportamento do índice por modelo de veículo. Para cada grupo, são apresentadas o número de observações e as principais medidas descritivas.

Tabela 4.2: Medidas descritivas do índice prêmio/valor por modelo de veículo.

Grupo	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Ford Ka	0,0137	0,0221	0,0249	0,0259	0,0286	0,0466
Hyundai HB20	0,0121	0,0236	0,0283	0,0293	0,0340	0,0549
Nissan March	0,0133	0,0226	0,0257	0,0264	0,0287	0,0615
Renault Sandero	0,0127	0,0207	0,0234	0,0251	0,0273	0,0591
Toyota Etios	0,0106	0,0218	0,0251	0,0263	0,0295	0,0531
Volkswagen Gol	0,0109	0,0248	0,0315	0,0334	0,0389	0,0896

De modo geral, o Volkswagen Gol e o Hyundai HB20 apresentam as maiores médias do índice prêmio/valor, com valores próximos de 0,033 e 0,029, respectivamente. Esses

resultados sugerem que, em termos relativos, esses modelos tendem a apresentar prêmios de seguro mais altos em relação ao valor de mercado, quando comparados aos demais veículos analisados. Em contraste, o Ford Ka, Renault Sandero, Toyota Etios e Nissan March exibem médias mais próximas de 0,026, com dispersões um pouco menores, indicando um comportamento relativamente mais homogêneo do índice nesses grupos.

A evolução do índice ao longo do tempo é apresentada na Tabela 4.3 e nas Figuras 4.2 e 4.3. A tabela mostra, para cada semestre, o número de observações, a média e a mediana do índice.

Tabela 4.3: Medidas descritivas do índice prêmio/valor por semestre.

Semestre	n	Média	Mediana
2016.1	88	0,0326	0,0295
2016.2	92	0,0309	0,0278
2017.1	87	0,0354	0,0316
2017.2	91	0,0303	0,0279
2018.1	62	0,0284	0,0259
2018.2	88	0,0262	0,0240
2019.1	89	0,0269	0,0254
2019.2	99	0,0258	0,0239
2020.1	50	0,0233	0,0211
2020.2	75	0,0238	0,0215
2021.2	38	0,0195	0,0173

Os resultados indicam uma tendência clara de redução do índice prêmio/valor ao longo do período analisado. As maiores médias ocorrem entre 2016.1 e 2017.1, com valores em torno de 0,031 a 0,035. A partir de 2018.1, observa-se um movimento de queda gradual, com médias próximas de 0,026 em 2018.2 e 2019.2 e valores em torno de 0,023 nos semestres de 2020. No último semestre da série, 2021.2, a média do índice atinge aproximadamente 0,019, com mediana em torno de 0,017.

Essa tendência é reforçada pelo boxplot da Figura 4.2, que apresenta a distribuição do índice por semestre, e pela linha de tendência da Figura 4.3, que mostra o comportamento do índice médio ponderado ao longo do tempo.

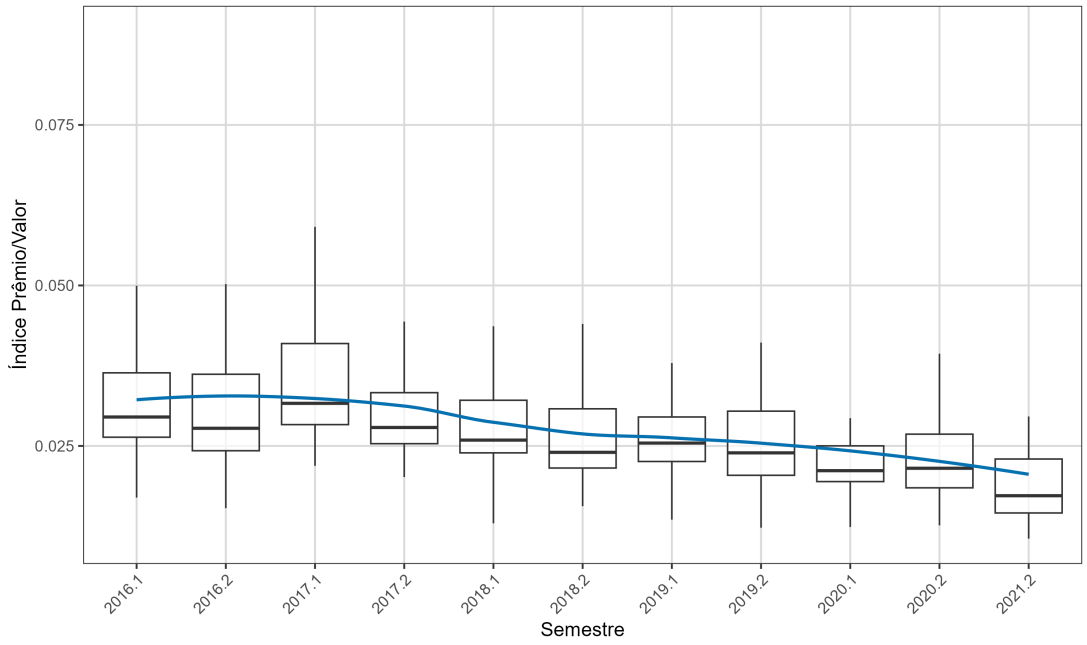

Figura 4.2: Distribuição do índice prêmio/valor por semestre.

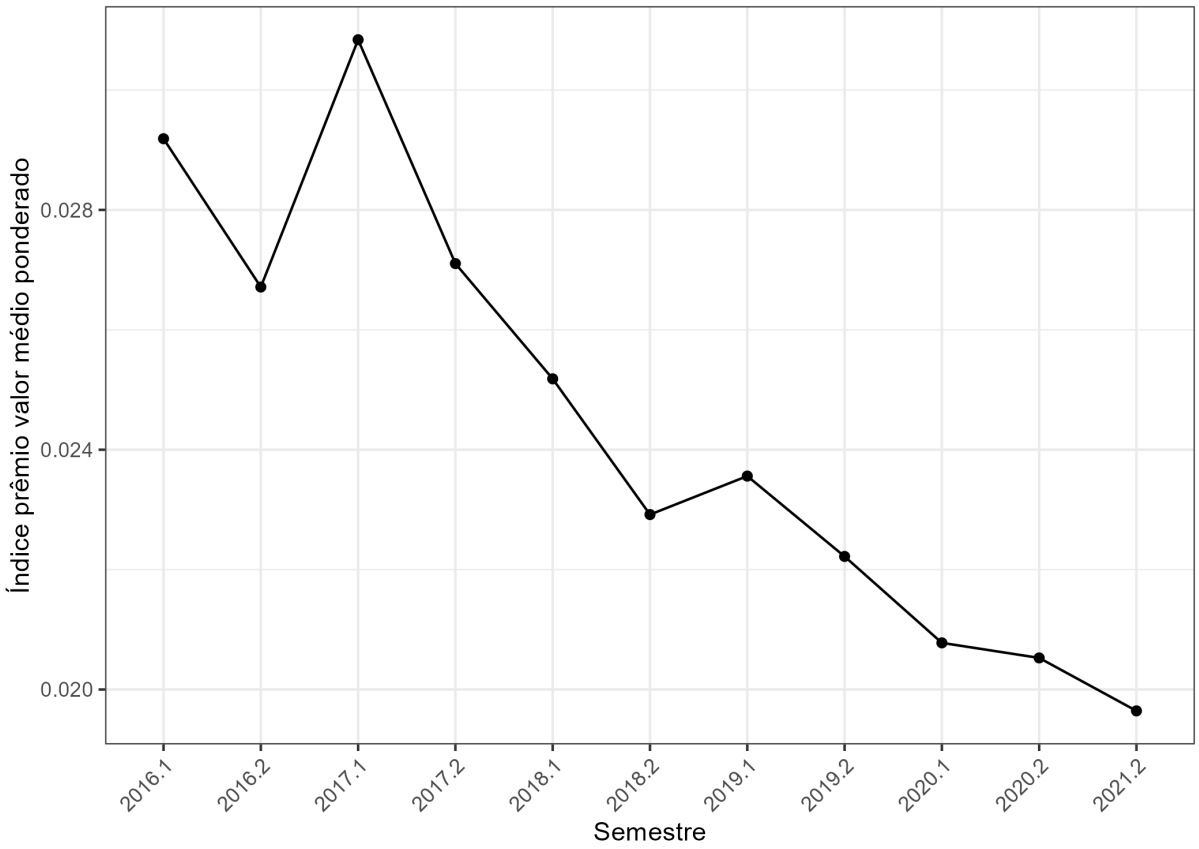

Figura 4.3: Evolução do índice prêmio/valor médio por semestre.

Os boxplots mostram deslocamentos sucessivos das medianas para níveis menores a partir de 2018.1, acompanhados de redução gradual da dispersão, especialmente nos quar-

tis superiores. A partir de 2019.2, a variabilidade entre as observações diminui de forma mais evidente, sugerindo um comportamento mais concentrado do índice em torno de valores menores. A Figura 4.3 sintetiza essa trajetória, evidenciando uma queda praticamente monotônica do índice médio após 2017.1.

Durante a etapa exploratória também foram identificadas observações com valores relativamente elevados do índice prêmio/valor. A Figura 4.4 destaca essas observações a partir do critério de distância interquartílica (IQR), isto é, valores acima do limite superior definido por 1,5 vez o IQR em cada semestre.

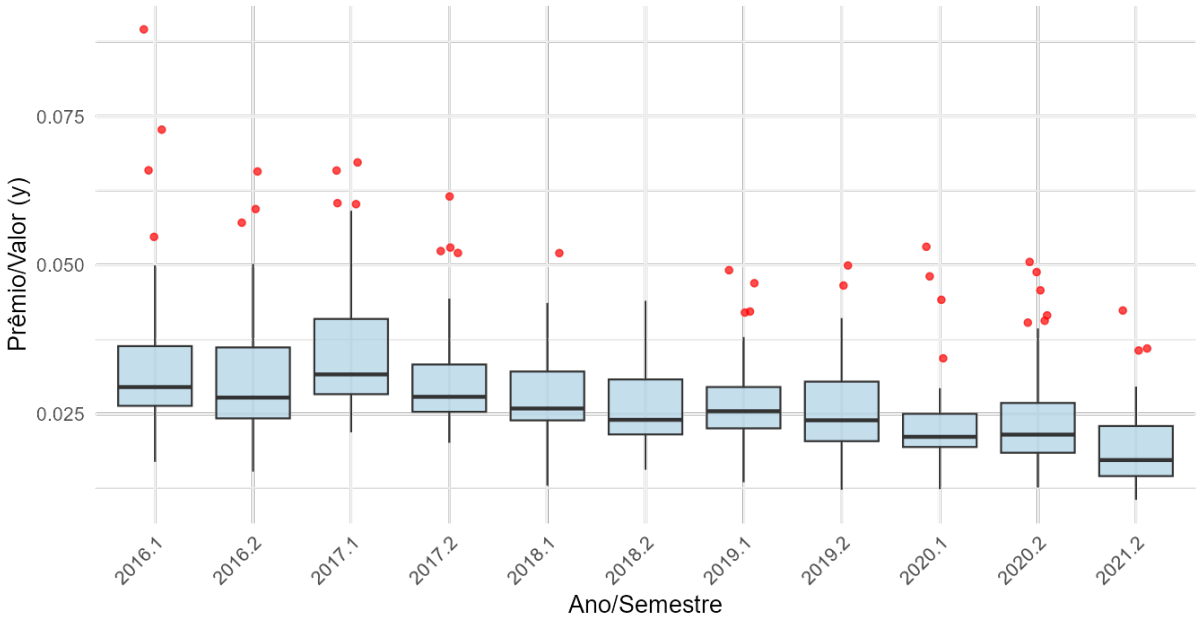

Figura 4.4: Identificação de possíveis outliers do índice prêmio/valor por semestre.

Os pontos em destaque indicam que o número de observações com índice muito acima da maioria é pequeno e espalhado ao longo dos semestres. Essas combinações correspondem, em geral, a casos específicos em que o prêmio médio é relativamente alto em relação ao valor do veículo. Como essas observações representam situações plausíveis do mercado e não foram identificadas inconsistências nos dados, optou-se por mantê-las na modelagem, verificando posteriormente sua influência por meio dos resíduos do modelo Beta.

4.1.3 Ajuste e Seleção do Modelo Beta

O índice prêmio/valor foi modelado por meio de Regressão Beta com função de ligação logito, conforme apresentado na Fundamentação Teórica. As covariáveis consideradas fo-

ram: semestre (efeito categórico, tendo 2016.1 como categoria de referência), modelo de veículo (referência Ford Ka), sexo do condutor ou pessoa segurada (referência pessoa física do sexo feminino), faixa etária (referência até 25 anos), logaritmo do índice de sinistralidade (logIS), logaritmo do valor médio do veículo (logValor) e logaritmo da exposição (logExp). Além dos efeitos principais, foram incluídas interações entre sexo e faixa etária, com o objetivo de capturar possíveis diferenças na relação prêmio/valor entre perfis de segurados.

A escolha do modelo final foi baseada na comparação entre um modelo Beta inicial, contendo apenas os efeitos principais, e um modelo estendido com as interações entre sexo e faixa etária. A Tabela 4.4 apresenta os valores do critério de informação de Akaike (AIC) e do critério de informação bayesiano (BIC) para os dois ajustes.

Tabela 4.4: Comparação de modelos Beta para o índice prêmio/valor.

Modelo	AIC	BIC
Base (efeitos principais)	-131823,6	-131699,9
Com interação sexo × faixa etária	-132129,4	-131967,7

O modelo com interação entre sexo e faixa etária apresenta valores menores de AIC e BIC em relação ao modelo base, indicando melhor equilíbrio entre qualidade de ajuste e parcimônia. Assim, esse modelo foi adotado como referência para as análises seguintes.

4.1.4 Resultados do Modelo de Regressão Beta

Efeitos principais

A Tabela 4.5 apresenta os coeficientes estimados para o componente de média do Modelo Beta, considerando os efeitos principais. São mostradas a estimativa na escala do logito, o erro padrão, a estatística de teste aproximada (z), o valor p e a razão de chances (OR), obtida pela exponenciação da estimativa.

Tabela 4.5: Coeficientes do modelo Beta para o índice prêmio/valor (efeitos principais).

Termo	Estimativa	Erro padrão	z	Valor p	OR
Intercepto	0,051	0,227	0,23	0,822	1,052
Semestre 2016.2	-0,047	0,004	-12,78	$< 0,001$	0,954
Semestre 2017.1	0,083	0,004	21,05	$< 0,001$	1,086
Semestre 2017.2	0,008	0,005	1,70	0,088	1,008
Semestre 2018.1	-0,064	0,005	-13,37	$< 0,001$	0,938
Semestre 2018.2	-0,152	0,005	-30,64	$< 0,001$	0,859
Semestre 2019.1	-0,171	0,005	-32,55	$< 0,001$	0,843
Semestre 2019.2	-0,204	0,005	-37,73	$< 0,001$	0,816
Semestre 2020.1	-0,276	0,008	-36,23	$< 0,001$	0,759
Semestre 2020.2	-0,270	0,008	-32,74	$< 0,001$	0,764
Semestre 2021.2	-0,302	0,018	-16,49	$< 0,001$	0,739
Hyundai HB20	0,148	0,005	27,31	$< 0,001$	1,159
Nissan March	0,009	0,005	1,99	0,047	1,009
Renault Sandero	0,018	0,004	4,02	$< 0,001$	1,018
Toyota Etios	0,069	0,005	14,69	$< 0,001$	1,071
Volkswagen Gol	0,128	0,006	21,37	$< 0,001$	1,137
Pessoa jurídica	0,057	0,014	4,15	$< 0,001$	1,059
Sexo masculino	0,085	0,006	14,25	$< 0,001$	1,089
Faixa 26–35 anos	-0,231	0,005	-43,84	$< 0,001$	0,793
Faixa 36–45 anos	-0,307	0,005	-57,85	$< 0,001$	0,736
Faixa 46–55 anos	-0,297	0,005	-56,53	$< 0,001$	0,743
Faixa acima de 55 anos	-0,332	0,005	-61,98	$< 0,001$	0,718
log(IS)	0,347	0,012	28,16	$< 0,001$	1,415
log(Valor)	-0,650	0,021	-31,00	$< 0,001$	0,522
log(Expostos)	-0,030	0,002	-14,43	$< 0,001$	0,971

No que se refere ao efeito do tempo, os semestres são interpretados em relação ao semestre de referência (2016.1). As estimativas negativas a partir de 2018.1 indicam redução sistemática do índice prêmio/valor em comparação com 2016.1, mesmo após o controle pelos demais fatores. Por exemplo, o coeficiente estimado para 2021.2 é aproximadamente -0,302, o que corresponde a uma razão de chances em torno de 0,739. Isso significa que, mantendo fixos o modelo de veículo, o perfil do segurado e as demais covariáveis, a chance de observar um índice prêmio/valor maior é cerca de 26% menor em 2021.2 do que em 2016.1. Essa conclusão é coerente com a análise descritiva por semestre, que já sugeria uma queda gradual do índice ao longo do período.

Os coeficientes associados aos modelos de veículos têm como referência o Ford Ka. As estimativas positivas para Hyundai HB20, Toyota Etios, Renault Sandero, Nissan March e Volkswagen Gol indicam que esses modelos, em média, apresentam índices prêmio/valor ligeiramente superiores ao do Ford Ka. Entre eles, destacam-se o Hyundai HB20 e o Volkswagen Gol, com razões de chances em torno de 1,16 e 1,14, respectivamente. Isso

sugere que, controlando pelos demais fatores, esses veículos tendem a ter prêmios de seguro relativamente mais altos em relação ao seu valor de mercado.

Em relação ao perfil do segurado, a categoria de referência é a pessoa física do sexo feminino com até 25 anos. Os coeficientes para pessoa jurídica e sexo masculino são positivos, indicando que, em linhas gerais, esses perfis estão associados a índices prêmio/valor mais elevados. Para pessoa jurídica, a razão de chances estimada é próxima de 1,06, enquanto para segurados do sexo masculino é cerca de 1,09, o que aponta para diferenças modestas, mas estatisticamente significativas, em comparação com o grupo de referência.

As faixas etárias apresentam coeficientes negativos em relação ao grupo até 25 anos, com razões de chances variando entre aproximadamente 0,79 e 0,72. Isso indica que, à medida que a idade aumenta, o índice prêmio/valor tende a ser menor, o que é compatível com a ideia de que perfis mais jovens estão associados a maior risco relativo e, consequentemente, a prêmios proporcionalmente mais altos.

Os efeitos contínuos de logIS, logValor e logExp também são importantes. O coeficiente positivo de logIS, com razão de chances em torno de 1,42, indica que, para aumentos no índice de sinistralidade, o índice prêmio/valor cresce, refletindo a sensibilidade do prêmio ao nível de risco observado. Em contraste, o coeficiente negativo de logValor, com OR próxima de 0,52, sugere que veículos de maior valor tendem a apresentar prêmios relativamente menores em proporção ao valor de mercado, apontando para uma relação subproporcional entre prêmio e valor do veículo. Por fim, o coeficiente negativo de logExp mostra que, em contextos com maior exposição, o índice prêmio/valor tende a ser ligeiramente menor, o que pode estar relacionado a efeitos de carteira e diluição do risco.

Interações entre sexo e faixa etária

A Tabela 4.6 apresenta os coeficientes estimados para as interações entre sexo (pessoa jurídica e sexo masculino) e faixas etárias. Esses termos refinam a interpretação dos efeitos principais, permitindo avaliar se as diferenças entre grupos de sexo variam de acordo com a idade.

Tabela 4.6: Coeficientes do modelo Beta para o índice prêmio/valor (interações sexo × faixa etária).

Termo	Estimativa	Erro padrão	z	Valor p	OR
Pessoa jurídica × 26–35 anos	0,118	0,016	7,30	$< 0,001$	1,125
Sexo masculino × 26–35 anos	-0,032	0,007	-4,67	$< 0,001$	0,969
Pessoa jurídica × 36–45 anos	0,091	0,016	5,74	$< 0,001$	1,095
Sexo masculino × 36–45 anos	-0,051	0,007	-7,31	$< 0,001$	0,950
Pessoa jurídica × 46–55 anos	0,042	0,019	2,25	0,025	1,043
Sexo masculino × 46–55 anos	-0,085	0,007	-12,04	$< 0,001$	0,919
Pessoa jurídica × acima de 55 anos	0,027	0,019	1,41	0,160	1,027
Sexo masculino × acima de 55 anos	-0,077	0,007	-11,37	$< 0,001$	0,926

Os resultados mostram que, para pessoas jurídicas, as faixas etárias acima de 25 anos tendem a apresentar índices prêmio/valor um pouco mais altos do que o esperado apenas pelos efeitos principais de faixa etária e tipo de pessoa. As razões de chances associadas à interação entre pessoa jurídica e as faixas de 26–35, 36–45 e 46–55 anos são ligeiramente superiores a 1, indicando um acréscimo moderado no índice prêmio/valor para esses perfis, em comparação com o padrão definido pelos efeitos principais.

Para o sexo masculino, as interações com as faixas mais altas apresentam estimativas negativas, com razões de chances entre aproximadamente 0,97 e 0,93. Isso indica que, embora o efeito principal de sexo masculino esteja associado a índices proporcionalmente maiores em relação ao grupo de referência, essa diferença tende a diminuir nas faixas etárias mais elevadas. Em outras palavras, as diferenças entre homens e mulheres em termos de índice prêmio/valor são mais acentuadas nas idades mais jovens e se reduzem à medida que a idade aumenta.

Em conjunto, os termos de interação mostram que a combinação entre sexo e faixa etária tem papel relevante na explicação do índice prêmio/valor, o que justifica a adoção do modelo com interação em vez do modelo apenas com efeitos principais.

4.1.5 Diagnósticos do Modelo Beta

A qualidade do ajuste do Modelo Beta foi avaliada por meio de diferentes diagnósticos, incluindo análise de resíduos e verificação de possíveis observações influentes. A Figura 4.5 apresenta o gráfico quantil–quantil dos resíduos quantílicos, enquanto as Figuras 4.6 e 4.7 mostram os gráficos de resíduos de Pearson e de resíduos quantílicos em função dos valores ajustados.

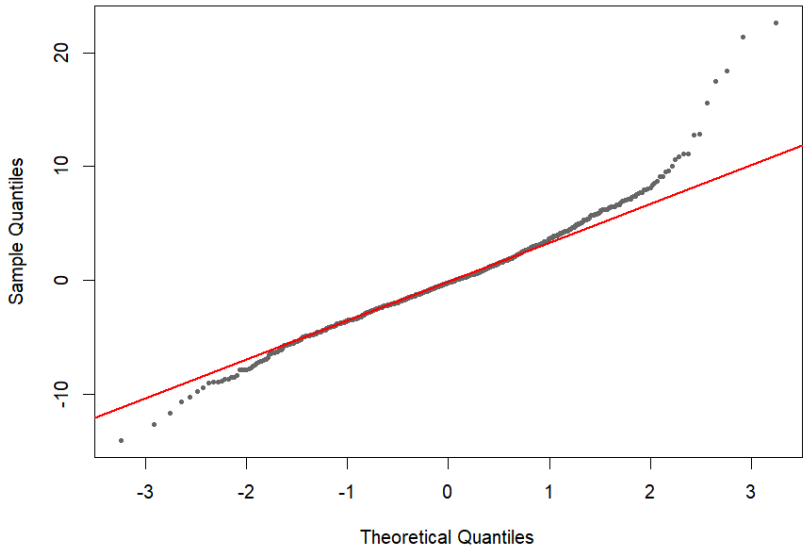

Figura 4.5: Gráfico quantil–quantil dos resíduos quantílicos do modelo Beta.

O gráfico quantil–quantil da Figura 4.5 mostra que, na maior parte do intervalo, os resíduos quantílicos seguem de forma razoável a reta de referência, indicando aderência geral à suposição de distribuição aproximada normal desses resíduos. Pequenas discrepâncias aparecem nas caudas, especialmente para resíduos positivos mais extremos, o que é esperado em presença de algumas observações com índices relativamente altos. No entanto, essas discrepâncias não são suficientes para caracterizar uma falta de ajuste grave do modelo.

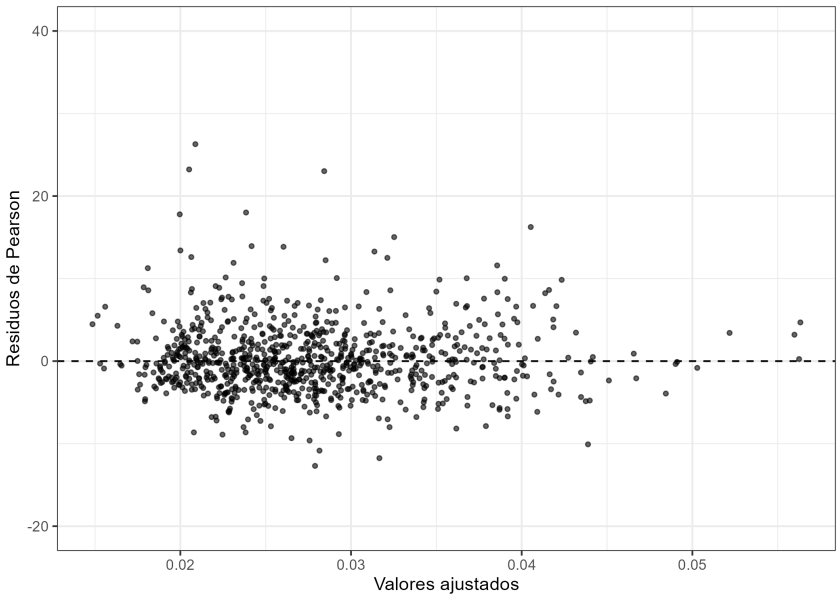

Figura 4.6: Resíduos de Pearson em função dos valores ajustados do modelo Beta.

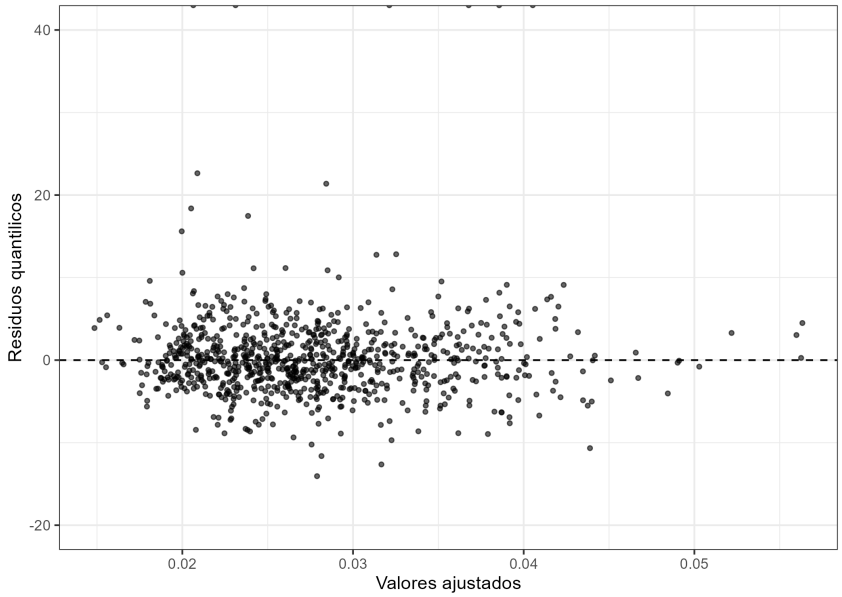

Figura 4.7: Resíduos quantílicos em função dos valores ajustados do modelo Beta.

Os gráficos de resíduos de Pearson e quantílicos em função dos valores ajustados (Figuras 4.6 e 4.7) não exibem padrões estruturados ou tendências claras, sugerindo que o preditor linear captura adequadamente a relação entre o índice prêmio/valor e as covariáveis incluídas. A dispersão dos pontos em torno da linha horizontal em zero é relativamente homogênea, sem indicação evidente de heterocedasticidade forte ou de regiões do espaço de preditores mal ajustadas.

Foi também realizada uma inspeção mais detalhada das observações com resíduos quantílicos de maior magnitude. A Figura 4.8 destaca essas observações, associando-as às combinações específicas de semestre, modelo de veículo, sexo e faixa etária.

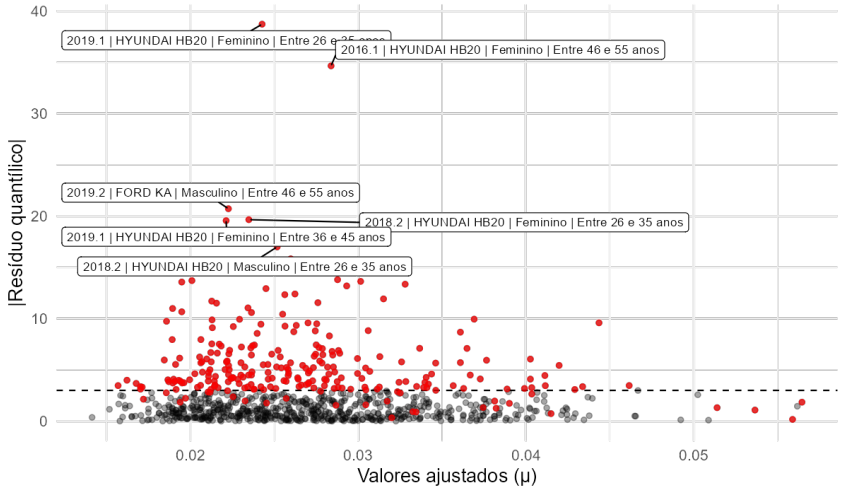

Figura 4.8: Observações com resíduos quantílicos elevados no modelo Beta.

As observações mais discrepantes estão associadas, em geral, a combinações específicas envolvendo principalmente o Hyundai HB20 e o Ford Ka em determinados semestres e faixas etárias, muitas vezes para perfis com prêmios relativamente altos em relação ao valor do veículo. Apesar de apresentarem resíduos elevados, essas unidades representam cenários plausíveis dentro do contexto de seguros automotivos e não foram identificadas evidências de erro de registro. Por esse motivo, optou-se por mantê-las na base, entendendo que elas contribuem para representar adequadamente a variabilidade real do mercado.

De forma complementar aos gráficos diagnósticos apresentados, a adequação do modelo final foi avaliada por meio de critérios formais. A inclusão da interação entre sexo e faixa etária foi testada por meio do teste da razão de verossimilhança, o qual indicou melhora estatisticamente significativa do ajuste em relação ao modelo base (X^2 = 321,83 ; gl = 8; $p < 0,001$). Como medida adicional de qualidade do ajuste, o modelo apresentou pseudo-R^2 igual a 0,76, indicando elevada capacidade explicativa do índice prêmio/valor. Por fim, a avaliação da dispersão de Pearson, em conjunto com a análise dos resíduos, não indicou evidências relevantes de má especificação, corroborando a adequação do modelo Beta ajustado.

4.1.6 Síntese e Considerações da Modelagem Beta

Os resultados da Modelagem Beta do Índice Prêmio/Valor indicam alguns padrões consistentes. Em primeiro lugar, há uma tendência clara de redução do índice ao longo do período de 2016.1 a 2021.2, com efeitos de semestre significativamente negativos a partir de 2018.1. Essa queda permanece mesmo após o controle pelos modelos de veículos, pelo perfil dos segurados e pelas variáveis contínuas logIS, logValor e logExp, o que sugere uma mudança estrutural na relação entre prêmios e valores de mercado ao longo do tempo.

Em segundo lugar, observam-se diferenças relevantes entre os modelos de veículos. O Hyundai HB20 e o Volkswagen Gol apresentam, em média, índices prêmio/valor mais altos que o Ford Ka, enquanto os demais modelos têm níveis mais próximos entre si. Isso indica que, para alguns veículos, o prêmio de seguro é proporcionalmente mais elevado em relação ao valor de mercado, possivelmente refletindo características específicas de risco ou de posicionamento desses modelos no mercado segurador.

O perfil dos segurados também desempenha papel importante. A combinação entre sexo, tipo de pessoa e faixa etária mostra que perfis mais jovens tendem a apresentar

índices prêmio/valor relativamente maiores, enquanto o aumento da idade está associado a redução desse índice. As interações entre sexo e faixa etária indicam que as diferenças entre homens e mulheres são mais acentuadas nas idades mais baixas e se tornam menores nas faixas etárias mais altas, o que é coerente com a ideia de que o risco relativo se aproxima entre os grupos com o aumento da idade.

Por fim, os efeitos de logIS, logValor e logExp fornecem evidências de que o índice prêmio/valor responde tanto ao nível de sinistralidade observado quanto às características econômicas da carteira. Veículos com maior índice de sinistralidade tendem a ter prêmios relativamente maiores, enquanto veículos de maior valor apresentam, em média, prêmios proporcionalmente menores, caracterizando uma relação subproporcional entre prêmio e valor de mercado. A maior exposição está associada a uma leve redução do índice, possivelmente refletindo efeitos de diversificação e diluição do risco em carteiras maiores.

De forma geral, os diagnósticos indicam que o modelo Beta ajustado descreve de maneira adequada o comportamento do índice prêmio/valor, sem evidências fortes de falta de ajuste ou de influência excessiva de poucas observações. Esses resultados fornecem uma base consistente para a etapa seguinte da análise, em que será considerado o prêmio médio de seguro em termos absolutos por meio de um modelo Gama.

4.2 Modelo de Regressão Gama para o Prêmio Médio de Seguro

Nesta subseção apresenta-se o ajuste do Modelo de Regressão Gama com função de ligação logarítmica para o Prêmio Médio de Seguro Automotivo, considerando especificamente a região de Campinas. O objetivo é avaliar de que forma o prêmio médio varia em função das características dos veículos, dos condutores e do histórico de exposição, permitindo compreender a estrutura de precificação praticada ao longo dos semestres analisados. Esta modelagem complementa a análise Beta desenvolvida anteriormente, oferecendo uma perspectiva em escala absoluta do prêmio, enquanto o modelo Beta descreve o comportamento proporcional do índice prêmio/valor.

4.2.1 Formulação do Modelo de Regressão Gama

A variável resposta considerada é o prêmio médio individual ($Premio_i$), assumido como distribuído segundo uma distribuição Gama parametrizada pela média, conforme apresentado por Paula (2013). Seja $\mu_i = \mathbb{E}(Premio_i)$ a média do prêmio para a observação i, adota-se o vínculo logarítmico:

$$g(\mu_i) = \log(\mu_i) = \eta_i,$$

onde o preditor linear η_i é definido como:

$$\eta_i = \beta_0 + \boldsymbol{\beta}_{\text{sem}}^{\top}\mathbf{Z}_{\text{sem},i} + \boldsymbol{\beta}_{\text{grupo}}^{\top}\mathbf{Z}_{\text{grupo},i} + \boldsymbol{\beta}_{\text{sexo}}^{\top}\mathbf{Z}_{\text{sexo},i} + \boldsymbol{\beta}_{\text{faixa}}^{\top}\mathbf{Z}_{\text{faixa},i} +$$

$$+ \beta_{\text{IS}}\log(\text{IS}_i) + \beta_{\text{exp}}\log(\text{Expostos}_i) + \beta_{\text{valor}}\log(\text{Valor}_i), \quad (4.3)$$

em que:

- $i = 1, \ldots, n$: índice das observações da base de dados, cada i correspondendo a uma combinação de semestre, modelo de veículo, tipo de pessoa, faixa etária, valor médio do veículo, importância segurada média e número de expostos na região de Campinas.

- η_i: preditor linear do Modelo Linear Generalizado para a observação i, definido na escala do logaritmo do prêmio médio.

- β_0: intercepto do modelo, associado ao grupo de referência definido pelas categorias base de semestre, grupo de veículo, sexo e faixa etária.

- $\boldsymbol{\beta}_{\text{sem}}$: vetor de coeficientes associado às categorias de semestre (por exemplo, 2016.2, 2017.1, ..., 2021.2), excluindo o semestre de referência.

- $\mathbf{Z}_{\text{sem},i}$: vetor de variáveis indicadoras (dummies) que codifica o semestre correspondente à observação i. Cada componente do vetor assume valor 1 quando a observação pertence a um semestre específico e 0 caso contrário.

- $\boldsymbol{\beta}_{\text{grupo}}$: vetor de coeficientes associado aos grupos de veículos (Ford Ka, Hyundai HB20, Nissan March, Renault Sandero, Toyota Etios e Volkswagen Gol), tomando um desses modelos como categoria de referência.

- $\mathbf{Z}_{\text{grupo},i}$: vetor de variáveis indicadoras que representa o grupo de veículo da observação i.

- $\boldsymbol{\beta}_{\text{sexo}}$: vetor de coeficientes relacionado às categorias da variável sexo (por exemplo, pessoa física do sexo feminino, pessoa física do sexo masculino e pessoa jurídica).

- $\mathbf{Z}_{\text{sexo},i}$: vetor de variáveis indicadoras que identifica o tipo de pessoa associado à observação i.

- $\boldsymbol{\beta}_{\text{faixa}}$: vetor de coeficientes correspondente às faixas etárias consideradas (por exemplo, entre 18 e 25 anos, entre 26 e 35 anos, entre 36 e 45 anos, entre 46 e 55 anos e maior que 55 anos), com uma faixa de referência.

- $\mathbf{Z}_{\text{faixa},i}$: vetor de variáveis indicadoras que codifica a faixa etária do condutor na observação i.

- β_{IS}: coeficiente associado ao logaritmo da importância segurada média.

- IS_i: importância segurada média, em reais, para a observação i, após os procedimentos de consolidação e winsorização descritos na metodologia.

- β_{exp}: coeficiente associado ao logaritmo do número de expostos.

- Expostos_i: número de expostos winsorizado na observação i, representando a quantidade de unidades seguradas utilizadas como base de cálculo do prêmio médio.

- β_{valor}: coeficiente associado ao logaritmo do valor médio do veículo.

- Valor_i: valor médio consolidado do veículo para a observação i, em reais, obtido a partir dos dados da Tabela FIPE organizados via Kaggle e compatibilizados com os dados da SUSEP.

4.2.2 Análise Exploratória do Prêmio Médio

Antes de apresentar o ajuste do Modelo Gama, é importante compreender o comportamento da variável resposta na base consolidada. Os resumos descritivos e os gráficos exploratórios auxiliam na identificação de padrões, diferenças entre grupos de veículos e possíveis assimetrias, o que justifica o uso de uma distribuição contínua, positiva e assimétrica para modelar o prêmio médio.

A Tabela 4.7 apresenta o resumo geral do prêmio médio na região de Campinas/SP. Os valores variam de aproximadamente R$ 120 a quase R$ 4900, indicando grande amplitude e assimetria à direita. A diferença entre a média (R$ 1451) e a mediana (R$ 1351) reforça esse comportamento assimétrico, frequentemente observado em variáveis monetárias.

Tabela 4.7: Resumo descritivo geral do prêmio médio (R$).

n	Mín.	Q1	Mediana	Média	Q3	Máx.
858	123	1183	1351	1451	1612	4890

A Tabela 4.8 detalha os valores por modelo de veículo. O Hyundai HB20 apresenta, em média, os maiores prêmios, seguido pelo Etios e pelo Gol. O Sandero e o March, por sua vez, possuem valores medianos mais baixos. Essa variação sugere que o modelo do veículo é uma covariável relevante na explicação do prêmio médio.

Tabela 4.8: Resumo do prêmio médio por grupo de veículo (R$).

Grupo	n	Mediana	Média	Q3	Máx.
Ford Ka	121	1273	1349	1467	2659
HB20	136	1558	1702	1917	4486
March	107	1246	1321	1398	3073
Sandero	135	1248	1371	1471	3302
Etios	153	1369	1481	1645	4890
Gol	206	1374	1443	1667	3124

A Figura 4.9 mostra a distribuição do prêmio por grupo de veículo. O comportamento é consistente com os resumos apresentados: o HB20 concentra valores mais altos, enquanto o March e o Sandero apresentam distribuições mais baixas. O Etios e o Gol ocupam posições intermediárias, mas ainda acima do Ford Ka.

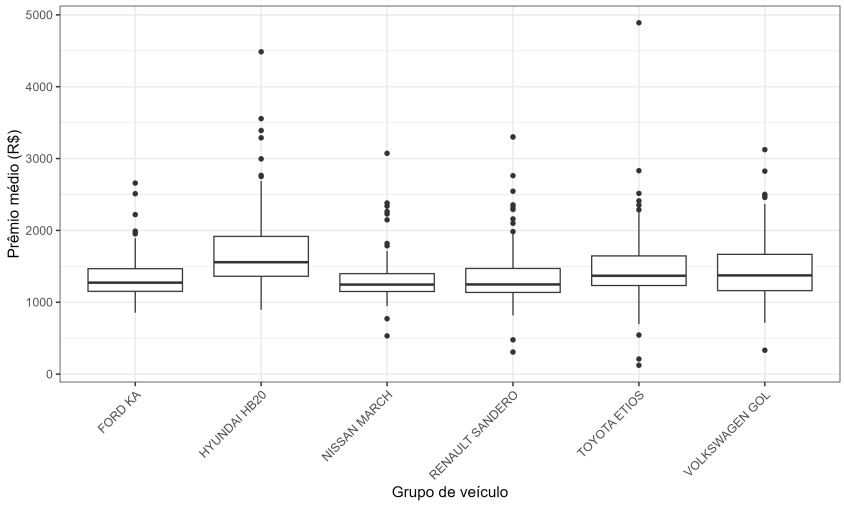

Figura 4.9: Distribuição do prêmio médio por grupo de veículo.

A Figura 4.10 apresenta a distribuição do prêmio médio por faixa etária do condutor. Observa-se que condutores mais jovens tendem a apresentar prêmios maiores, enquanto faixas etárias mais altas exibem valores medianos menores. Esse padrão é coerente com o comportamento esperado do mercado, em que perfis mais jovens costumam estar associados a maior risco.

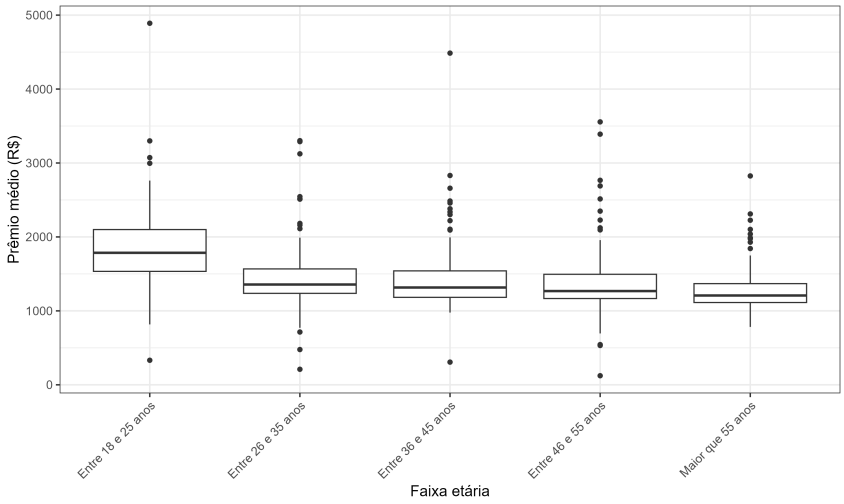

Figura 4.10: Distribuição do prêmio médio por faixa etária.

A Figura 4.11 apresenta a evolução temporal da razão entre prêmio e valor dos veículos.

Apesar de a análise principal desta seção se concentrar no prêmio absoluto, a trajetória descendente ao longo dos semestres ajuda a contextualizar a dinâmica observada no mercado e reforça a necessidade de modelagem conjunta.

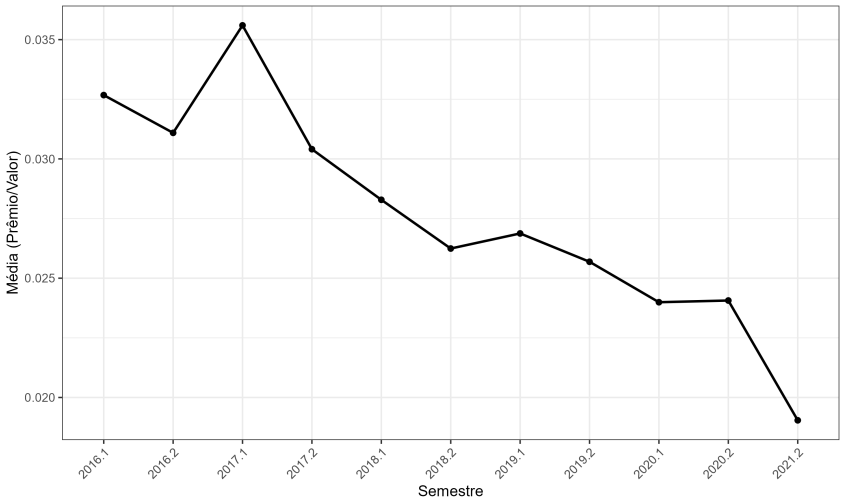

Figura 4.11: Evolução da razão prêmio/valor ao longo dos semestres.

No geral, os resultados descritivos apontam alta variabilidade do prêmio médio, assimetria à direita e diferenças marcantes entre grupos de veículos e faixas etárias. Essas características justificam a escolha da distribuição Gama com função de ligação logarítmica na etapa seguinte de modelagem e mostram que há estrutura suficiente nos dados para explicar parte relevante da variação no prêmio médio. A análise detalhada dos coeficientes permitirá compreender como cada covariável influencia esse comportamento.

4.2.3 Coeficientes e Interpretação do Modelo Gama

Após o ajuste do modelo Gamma com ligação logarítmica, analisam-se os principais coeficientes estimados, apresentados na Tabela 4.9. Os valores foram convertidos para multiplicadores, o que facilita a interpretação dos efeitos em termos proporcionais sobre o prêmio médio.

Tabela 4.9: Coeficientes principais do modelo Gamma (multiplicadores).

Termo	Multiplicador	IC95% Inf	IC95% Sup
log(Valor)	1.588	1.286	1.961
log(IS)	1.351	1.178	1.549
log(Expostos)	0.964	0.943	0.986
HB20	1.140	1.078	1.207
Gol	1.125	1.067	1.185
Faixa 46–55	0.729	0.697	0.761
2019.2	0.813	0.767	0.861
2020.1	0.755	0.705	0.809

De forma geral, os resultados obtidos refletem relações que são consistentes com o comportamento esperado do mercado de seguros. O valor do veículo e o custo médio de indenização (IS) surgem como fatores centrais na explicação do prêmio, enquanto características do condutor e a evolução temporal também desempenham papéis relevantes.

O efeito do valor médio do veículo se destaca pela magnitude: o multiplicador de aproximadamente 1,59 indica que, para um aumento proporcional no valor do bem segurado, o prêmio tende a aumentar cerca de 59%, mantendo constantes as demais variáveis. Trata-se de um resultado coerente com a literatura de tarifação, já que veículos mais caros implicam maior custo potencial de reposição ou reparo.

Outro componente importante é a indenização média de sinistro. O multiplicador de 1,35 mostra que valores mais altos de IS estão associados a prêmios maiores, reforçando a ideia de que seguradoras internalizam, no preço, o risco técnico associado ao histórico e ao potencial de perdas financeiras.

A variável log(Expostos) apresenta multiplicador inferior a 1, em torno de 0,96. Isso significa que, à medida que aumenta o número de exposições, o prêmio médio tende a diminuir levemente. Esse comportamento sugere possível efeito de credibilidade: bases com maior volume contribuem para reduzir a variabilidade e, consequentemente, o valor ajustado do prêmio.

Algumas diferenças importantes também aparecem entre os grupos de veículos. Os modelos HB20 e Gol, por exemplo, apresentam multiplicadores superiores a 1, indicando prêmios sistematicamente maiores quando comparados ao grupo de referência. Esse padrão está de acordo com os resumos descritivos e reforça que determinadas características de modelos específicos influenciam diretamente o custo de seguro.

Em relação à faixa etária, destaque para o grupo entre 46 e 55 anos, cujo multiplicador

é de 0,73. Isso significa que condutores dessa faixa possuem, em média, prêmios cerca de 27% menores do que o grupo-base, o que sugere perfil de menor risco. Esse comportamento também apareceu na análise exploratória e se mostra consistente na modelagem.

Por fim, os semestres mais recentes exibem multiplicadores abaixo de 1, como é o caso de 2019.2 (0,81) e 2020.1 (0,75). Esses valores indicam redução relativa no prêmio médio ao longo desses períodos, mesmo em um contexto em que o valor dos veículos nem sempre apresentou queda. Esse movimento pode refletir ajustes de mercado, mudanças na competitividade entre seguradoras ou variações no perfil agregado de risco da carteira.

De forma integrada, os coeficientes estimados mostram que a variação do prêmio médio decorre de um conjunto de fatores combinados, envolvendo tanto características do veículo quanto atributos do segurado e aspectos temporais. Os resultados são coerentes com os padrões observados nos resumos descritivos e contribuem para uma compreensão mais clara do comportamento do seguro automotivo na região analisada.

4.2.4 Diagnósticos e Calibração do Modelo Gama

A verificação dos diagnósticos do modelo é uma etapa essencial para garantir que as conclusões obtidas a partir dos coeficientes estimados sejam confiáveis. No caso do modelo Gama com ligação logarítmica, espera-se que a variância aumente com a média e que os resíduos sigam comportamento compatível com essa estrutura. Assim, a análise gráfica a seguir permite avaliar, de forma integrada, a adequação do modelo, a presença de possíveis padrões não explicados e a estabilidade das estimativas.

Resíduos deviance

A Figura 4.12 apresenta o gráfico dos resíduos deviance em função dos valores ajustados. Observa-se que a dispersão dos resíduos aumenta à medida que os valores previstos também crescem, o que é compatível com a suposição da família Gama, em que a variância é proporcional ao quadrado da média. Esse comportamento indica que o modelo está capturando adequadamente a heterocedasticidade natural do prêmio, cuja variabilidade tende a ser maior entre veículos de maior valor ou perfis com prêmios mais altos.

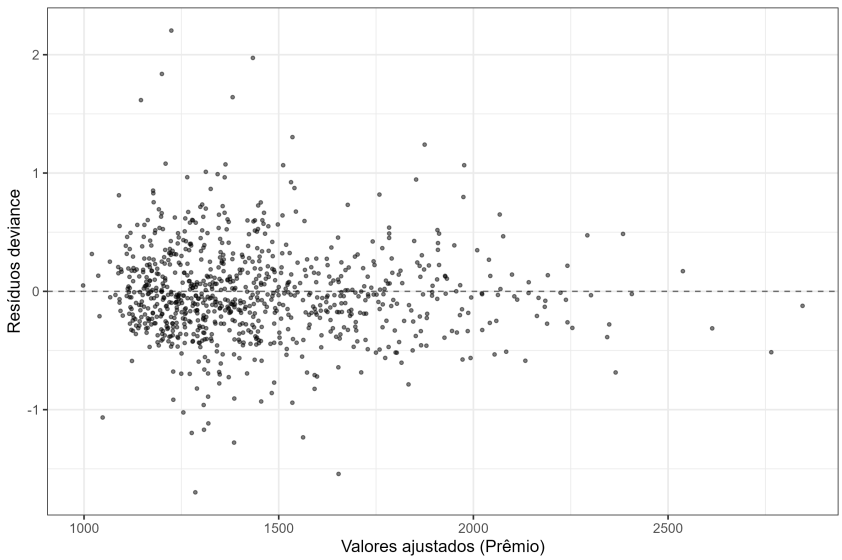

Figura 4.12: Resíduos deviance versus valores ajustados.

Além disso, não se observam padrões estruturados, como faixas horizontais, curvaturas sistemáticas ou agrupamentos que indiquem especificação inadequada da média. Os resíduos se distribuem de forma difusa ao redor de zero, sugerindo que o conjunto de covariáveis incluído no modelo, juntamente com as transformações aplicadas (logs e suavização dos valores extremos), está capturando bem a dinâmica do prêmio. A presença de alguns valores extremos é esperada, dado que a base contém perfis com pouca exposição e alguns registros com prêmios atipicamente altos, mas nenhum desses pontos aparece de maneira a comprometer o ajuste global.

Aderência dos resíduos à distribuição teórica

A Figura 4.13 apresenta o QQ-plot dos resíduos de Pearson. Esse gráfico compara os resíduos padronizados à distribuição normal teórica, permitindo avaliar a adequação do componente aleatório do modelo.

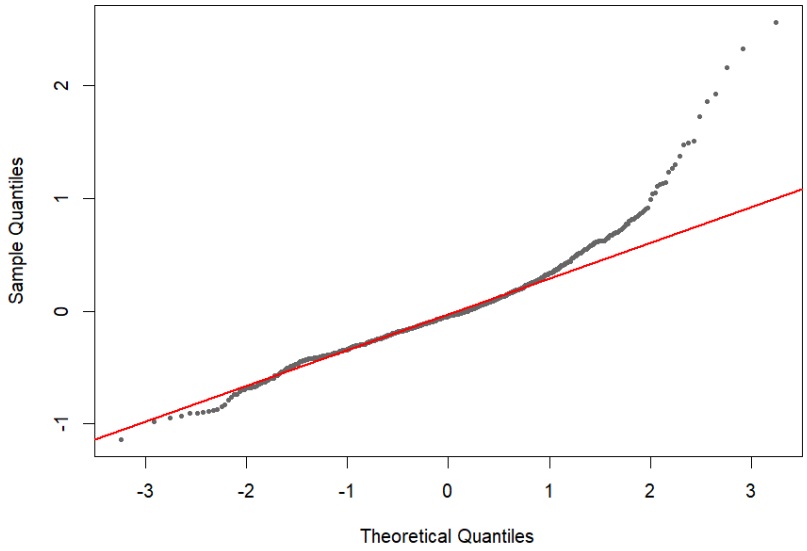

Figura 4.13: QQ-plot dos resíduos de Pearson.

De modo geral, os resíduos seguem a linha teórica ao longo de grande parte da distribuição, com pequenos desvios nas caudas. Esse comportamento é típico em modelos Gama ajustados a dados reais, pois a distribuição do prêmio tende a ser assimétrica e contém valores extremos, especialmente em segmentos com menor quantidade de segurados. O leve afastamento nas extremidades não caracteriza falha de ajuste; ao contrário, indica que o modelo está capturando bem a estrutura central dos dados, mantendo estabilidade nas regiões de maior densidade de observações.

Influência das observações

A Figura 4.14 apresenta o histograma da distância de Cook, utilizada para identificar observações potencialmente influentes. A maior parte dos casos apresenta valores muito baixos, próximos de zero, o que indica que nenhuma única observação exerce grande influência sobre as estimativas do modelo.

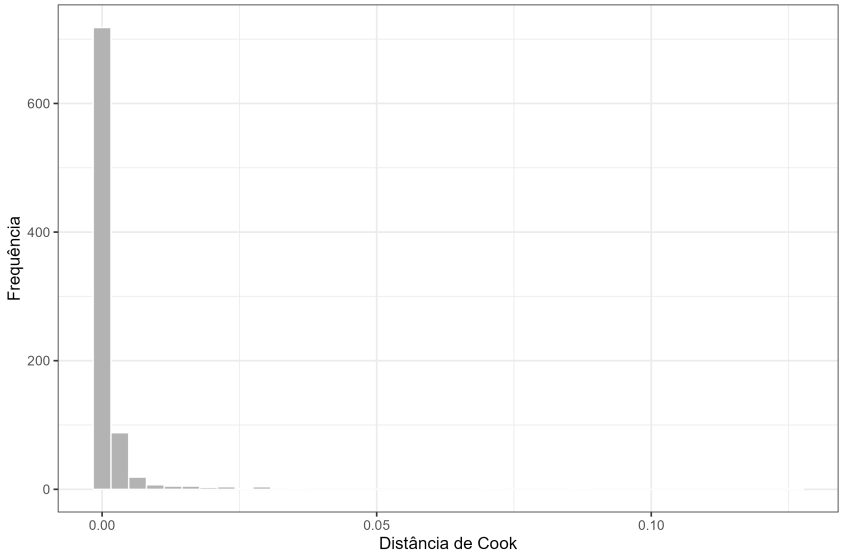

Figura 4.14: Distribuição da distância de Cook.

Observações com valores mais elevados de Cook aparecem com baixa frequência e não ultrapassam níveis considerados críticos. Isso significa que o ajuste está sendo conduzido de forma estável e que o modelo não depende de casos isolados para determinar os coeficientes. Esse resultado é particularmente importante nesta aplicação, pois a base contém perfis com diferentes valores de exposição (*expostos*), o que poderia potencialmente gerar efeitos desproporcionais. A suavização dos valores extremos aplicada às variáveis auxilia nesse controle, dando maior robustez ao processo de estimação.

Calibração Observado vs. Predito

A Figura 4.15 compara as médias observadas e preditas em decis da distribuição dos valores ajustados. A proximidade dos pontos com a linha de identidade mostra que o modelo está bem calibrado, ou seja, as previsões condizem com os valores efetivamente registrados na base.

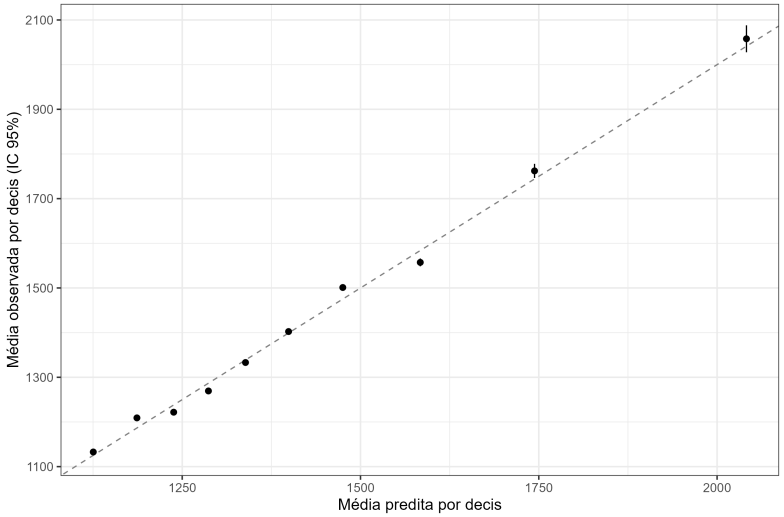

Figura 4.15: Calibração por decis: valores observados vs. preditos.

Nas primeiras faixas da distribuição, o modelo apresenta ajuste bastante preciso, com diferenças pequenas entre observado e previsto. Nos decis superiores, observa-se uma oscilação ligeiramente maior, associada ao fato de que veículos de maior valor e perfis com prêmios mais altos apresentam maior variabilidade e menor quantidade de observações. Ainda assim, as médias observadas permanecem dentro dos intervalos esperados, o que reforça a adequação do modelo mesmo em regiões mais extremas.

Razão Observado/Esperado (O/E)

A Figura 4.16 apresenta a razão Observado/Esperado (O/E) por decis. Um modelo bem calibrado deve produzir valores próximos de 1, especialmente nos decis centrais, que contêm maior quantidade de dados.

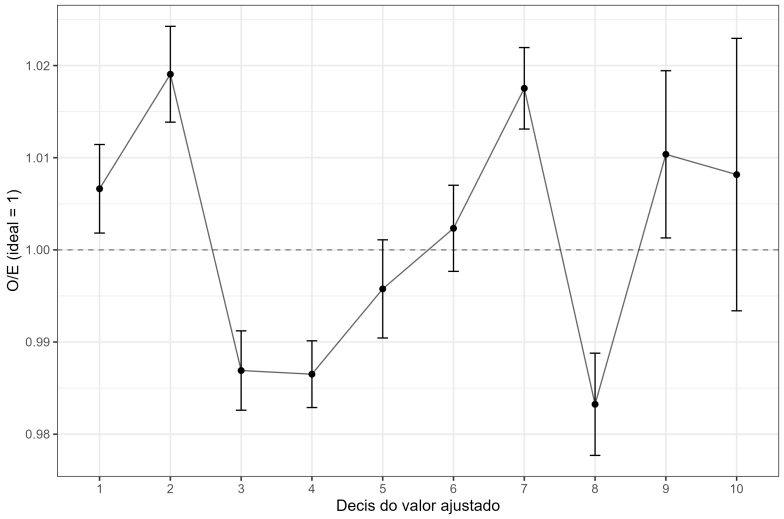

Figura 4.16: Razão Observado/Esperado (O/E) por decis.

A linha pontilhada horizontal representa o valor ideal de O/E = 1, indicando equilíbrio entre o prêmio previsto pelo modelo e o prêmio observado na base. Os decis centrais se mantêm muito próximos desse valor, o que mostra que o modelo realiza estimativas consistentes. As oscilações verificadas nos decis 9 e 10 são esperadas e refletem a maior variabilidade natural dos prêmios altos. Mesmo nessas faixas, as razões O/E permanecem próximas de 1, sem indicar super ou subestimações sistemáticas.

Em conjunto, os diagnósticos gráficos e numéricos indicam que o modelo Gama apresenta boa adequação aos dados. O teste da razão de verossimilhança, ao comparar o modelo completo com o modelo nulo, resultou em uma estatística Qui-quadrado elevada ($\chi^2 = 16115{,}74$, com 24 graus de liberdade e $p < 0{,}001$), evidenciando que o conjunto de covariáveis incluídas contribui de forma significativa para a explicação do prêmio médio de seguro. A análise da dispersão, baseada nos resíduos de Pearson, apresentou valor $\phi = 0{,}17$, indicando ausência de superdispersão e compatibilidade entre a variância observada e a estrutura assumida pelo modelo Gama. Os resíduos exibiram comportamento consistente com o esperado, sem indícios de problemas estruturais ou de observações excessivamente influentes, e a análise de calibração mostrou que o modelo reproduz adequadamente as médias observadas ao longo de toda a distribuição. De forma geral, essas evidências reforçam a confiabilidade das inferências e das interpretações apresentadas, indicando que o modelo Gama com ligação logarítmica fornece uma representação apropriada do comportamento do prêmio médio de seguro na região analisada.

4.2.5 Síntese e Considerações da Modelagem Gama

Os resultados do Modelo de Regressão Gama ajustado para o Prêmio Médio de Seguro permitem identificar um conjunto de padrões claros e coerentes com a estrutura dos dados analisados. Assim como na modelagem Beta, observa-se a presença de efeitos temporais bem definidos. A partir de 2018.1, a maior parte dos semestres apresenta multiplicadores significativamente inferiores a 1, indicando que, mesmo após o controle pelas demais covariáveis, os prêmios médios tendem a se reduzir de forma consistente ao longo do período. Essa trajetória sugere que fatores externos à composição da carteira — como ajustes operacionais, mudanças de mercado ou reorganização dos preços — influenciaram a evolução do prêmio médio em valores absolutos.

Além da tendência temporal, também se destacam diferenças importantes entre os grupos de veículos. O Hyundai HB20 e o Volkswagen Gol apresentam, em média, prêmios

mais altos em relação ao grupo de referência, padrão que está alinhado ao observado no índice prêmio/valor. Para os demais modelos, os níveis de prêmio tendem a ser mais próximos, o que indica que a precificação acompanha características próprias de cada veículo, mas sem discrepâncias tão acentuadas quanto as observadas nos dois modelos mencionados. Esses resultados reforçam que parte da variação no prêmio é explicada por diferenças estruturais associadas ao tipo de veículo.

O perfil dos segurados novamente desempenha papel relevante, com destaque para a variável faixa etária. As faixas a partir dos 26 anos apresentam multiplicadores significativamente menores, indicando que o prêmio médio diminui conforme aumenta a idade do condutor. O comportamento é uniforme e estatisticamente robusto, refletindo um risco relativo menor em faixas etárias mais altas. Além disso, o efeito associado ao sexo e ao tipo de pessoa mostra que segurados pessoas jurídicas apresentam prêmios médios maiores, enquanto homens apresentam pequeno aumento em relação às mulheres, coerente com o padrão encontrado na modelagem Beta.

Entre as variáveis contínuas, o valor médio dos veículos apresenta o efeito mais expressivo. O coeficiente associado ao logaritmo do valor indica que veículos mais caros tendem a ter prêmios proporcionalmente maiores, reforçando a relação direta entre o valor do bem e o custo do seguro. A indenização média (IS) também apresenta forte impacto positivo sobre o prêmio médio, mostrando que, quanto maior o custo potencial associado aos eventos cobertos, maior o valor cobrado. Por outro lado, a variável `logExpostos` mostra um efeito negativo moderado, sugerindo que exposições maiores resultam em prêmios médios ligeiramente menores, o que pode refletir efeitos de estabilidade, credibilidade ou diluição do risco.

Os diagnósticos do modelo indicam ajuste adequado. Apesar de uma variabilidade crescente nos resíduos — característica esperada de modelos Gamma —, não há padrões que indiquem falta de ajuste sistemática. O QQ-plot apresenta desvios suaves apenas nas extremidades da distribuição, o que não compromete a validade do modelo. A análise da distância de Cook também não revela observações com influência excessiva, reforçando a estabilidade das estimativas. Da mesma forma, os gráficos de calibração mostram boa aderência entre valores observados e preditos, com pequenas oscilações apenas nos decis superiores, onde o número de observações é menor.

De modo geral, o modelo Gamma ajustado resume adequadamente o comportamento do prêmio médio em valores absolutos. Os resultados obtidos são consistentes com os

padrões esperados e fornecem uma visão complementar à análise relativa obtida pelo modelo Beta. Em conjunto, as duas modelagens permitem compreender tanto a relação proporcional entre prêmio e valor do veículo quanto a evolução do prêmio médio propriamente dito, oferecendo uma base sólida para as discussões apresentadas nas seções seguintes.

Capítulo 5

Conclusão

O estudo desenvolvido possibilitou compreender, de maneira integrada, a relação entre os valores médios de mercado dos veículos e os prêmios médios de seguro praticados na região da Grande Campinas ao longo do período analisado. O trabalho iniciou-se com uma etapa extensa de consolidação das bases de dados provenientes da SUSEP SUSEP - Superintendência de Seguros Privados (2024) e da Tabela Fipe FIPE - Fundação Instituto de Pesquisas Econômicas (2024), obtida na plataforma Kaggle Kaggle (2023). Todo o processo de organização, limpeza e padronização foi conduzido em Excel Microsoft Corporation (2024), garantindo a compatibilidade entre períodos, modelos e variáveis. Essa etapa exigiu atenção especial, principalmente porque foi necessário calcular os valores médios consolidados de mercado e de prêmio para que as variáveis ficassem comparáveis entre os semestres e entre os seis modelos selecionados.

Além disso, é importante ressaltar que a etapa de construção do banco de dados constituiu uma contribuição metodológica central deste trabalho e foi conduzida integralmente pela autora do estudo. Diferentemente de estudos que utilizam bases previamente consolidadas, o presente estudo demandou a coleta direta dos dados, a definição dos recortes veicular, temporal e regional, bem como a padronização, agregação e integração das informações provenientes de fontes distintas. Esse processo envolveu decisões metodológicas fundamentais, desde a seleção dos modelos com séries temporais consistentes até a construção das variáveis analíticas utilizadas na modelagem, e foi determinante para a viabilidade das análises realizadas e para a consistência dos resultados apresentados ao longo do trabalho.

A análise descritiva revelou que tanto os valores dos veículos quanto os prêmios médios oscilaram ao longo do tempo, com variações significativas entre modelos. Alguns veículos

apresentaram valorização em períodos específicos, enquanto o comportamento dos prêmios não acompanhou essas mudanças de maneira uniforme. Esse contraste reforçou a necessidade de utilizar modelos estatísticos capazes de capturar tanto variações relativas quanto absolutas da relação entre prêmio e valor.

A regressão Beta, aplicada ao índice prêmio/valor, permitiu analisar o seguro de forma proporcional. Os resultados mostraram uma redução constante do índice ao longo dos semestres, mesmo após o controle pelas características dos veículos, do perfil dos segurados e das variáveis contínuas de exposição, indenização média e valor. Observou-se também que HB20 e Gol apresentam índices proporcionalmente maiores. Em relação ao perfil do segurado, faixas etárias mais jovens exibiram índices mais elevados, sinalizando maior custo relativo do seguro nesses grupos.

A seguir, a modelagem Gama avaliou o prêmio médio em valores absolutos. Confirmou-se a relação positiva entre o valor médio do veículo e o prêmio, assim como o efeito expressivo da indenização média. A variável de exposição apresentou efeito negativo, sugerindo que carteiras maiores tendem a ter prêmios médios mais baixos devido à diluição natural do risco. As diferenças entre modelos também foram consistentes com os resultados obtidos na abordagem relativa, na qual o prêmio do seguro é analisado em relação ao valor do veículo.

Os ajustes de ambos os modelos apontaram diagnósticos satisfatórios. Os resíduos exibiram comportamento dentro do esperado para os tipos de distribuição utilizados, sem indicação de falta de ajuste relevante ou observações demasiadamente influentes. Os gráficos de calibração reforçaram a proximidade entre valores preditos e observados, principalmente nos decis intermediários.

De forma geral, os resultados mostram que a evolução do prêmio médio e do valor de mercado dos veículos não ocorre de maneira proporcional ao longo do tempo. Enquanto alguns veículos apresentaram aumentos expressivos no valor de mercado, o prêmio médio não acompanhou esse movimento na mesma medida. Essa divergência fica evidente ao observar o índice prêmio/valor, que apresentou tendência de redução mesmo em períodos de valorização dos veículos. Esse comportamento sugere ajustes estruturais no setor segurador, possivelmente associados à dinâmica da carteira, ao perfil dos segurados e às condições de mercado naquele período. A combinação da modelagem Beta e da modelagem Gama permitiu visualizar essa dualidade: por um lado, o seguro tornou-se relativamente mais barato; por outro, o prêmio absoluto continuou respondendo de forma

clara às características do veículo, da carteira e da sinistralidade.

Além dos resultados obtidos, o processo de construção do banco de dados mostrou-se fundamental. A necessidade de consolidação manual das bases e de cálculo das médias reforçou a importância de ferramentas de organização, como o Excel, e evidenciou que análises mais profundas dependem diretamente da qualidade e granularidade das informações disponíveis. Essa etapa exigiu persistência e motivação, especialmente pela ausência de bases detalhadas com observações individuais de apólices.

Para trabalhos futuros, algumas direções surgem naturalmente. Uma delas é ampliar o conjunto de veículos analisados, incluindo modelos como o Chevrolet Onix, que permitiria comparações mais amplas e consistentes com o segmento de carros populares. Outro avanço seria trabalhar com bases mais detalhadas, contendo informações individuais de apólices, frequência e severidade, quilometragem, local de circulação e características completas do perfil do condutor. Essa granularidade possibilitaria a construção de modelos mais realistas, reduzindo a dependência de médias e de agregações.

Em síntese, o estudo mostrou que a relação entre o valor de mercado dos veículos e o prêmio de seguro é influenciada tanto por características do veículo quanto por fatores de carteira e de mercado. Os resultados obtidos oferecem uma base sólida para investigações futuras e reforçam a importância de métodos estatísticos bem estruturados e de bases de dados consolidadas para compreender a dinâmica do setor de seguros automotivos no Brasil.

Referências Bibliográficas

AutoSeg (2023). Autoseg - sistema de estatísticas de automóveis - susep. Acesso em: jun. 2025.

Demétrio, C. G. B. (2002). *Modelos Lineares Generalizados*. ESALQ-USP, Piracicaba.

Dória, F. e Gonzaga, L. (2015). *Seguro e risco: teoria e prática*. Escola Nacional de Seguros, Rio de Janeiro.

Ferrari, S. L. P. e Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

FIPE - Fundação Instituto de Pesquisas Econômicas (2024). Tabela de preços de veículos fipe.

Kaggle (2023). Tabela fipe - dados históricos. Acesso em: jun. 2025.

Microsoft Corporation (2024). Microsoft excel. Versão 2024.

Paula, G. A. (2013). *Modelos de Regressão com Apoio Computacional*. IME-USP, São Paulo.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

SUSEP - Superintendência de Seguros Privados (2024). Relatórios e estatísticas do setor. https://www.susep.gov.br. Acesso em: 25 jun. 2025.

Apêndice A

Análise Exploratória

A.1 Análise dos valores de mercado (FIPE)

A.1.1 Toyota Etios

Tabela A.1: Resumo estatístico por semestre do valor de mercado do Toyota Etios (Tabela FIPE)

Ano - Semestre	Média (R$)	Mediana	Desvio	Mínimo	Máximo	n
2016 - 1º	49.054	49.811	5.467	39.888	60.514	38
2016 - 2º	53.109	52.894	5.710	43.020	62.810	61
2017 - 1º	54.809	54.710	5.882	43.270	63.406	64
2017 - 2º	55.916	55.648	6.114	45.094	63.945	54
2018 - 1º	56.677	56.019	5.963	45.037	65.880	57
2018 - 2º	57.292	57.235	5.478	47.252	65.900	60
2019 - 1º	57.129	56.205	5.354	47.586	65.900	48
2019 - 2º	57.573	57.506	4.148	50.072	63.858	35
2020 - 1º	57.049	57.243	4.074	51.100	63.134	24
2020 - 2º	59.930	60.238	5.006	51.551	68.947	24
2021 - 1º	64.269	63.995	5.426	55.200	74.316	24
2021 - 2º	66.806	66.586	5.278	58.740	74.965	24

A Tabela A.1 apresenta o comportamento estatístico semestral dos valores de mercado do Toyota Etios, conforme dados da Tabela FIPE. Observa-se uma evolução consistente dos valores médios ao longo do período analisado, com aumentos graduais entre 2016 e 2021. A dispersão dos dados, refletida pelo desvio padrão, se manteve relativamente estável, o que sugere controle na variação dos preços. Os valores máximos e mínimos também acompanham a tendência de alta, indicando um reposicionamento do modelo no mercado ao longo dos anos.

A.1.2 Volkswagen Gol

Tabela A.2: Resumo estatístico por semestre do valor de mercado do Volkswagen Gol (Tabela FIPE)

Ano - Semestre	Média (R$)	Mediana	Desvio	Mínimo	Máximo	n
2017 - 2º	47.782	46.812	6.795	36.054	58.190	54
2018 - 1º	47.436	46.355	5.875	35.932	57.704	46
2018 - 2º	48.479	47.023	4.340	41.425	56.543	46
2019 - 1º	47.396	45.710	4.700	39.896	56.212	41
2019 - 2º	49.777	50.926	4.992	40.918	57.195	22
2020 - 1º	52.654	53.491	4.587	45.420	59.162	18
2020 - 2º	55.907	57.076	5.540	46.616	63.718	18
2021 - 1º	62.416	64.122	5.998	52.113	72.836	18
2021 - 2º	71.589	70.840	7.159	60.234	85.367	18

A Tabela A.2 apresenta as estatísticas descritivas dos valores de mercado do Volkswagen Gol por semestre, entre 2017 e 2021. É possível observar uma tendência clara de valorização ao longo do período, com crescimento progressivo da média semestral, especialmente a partir do segundo semestre de 2019.

A.1.3 Hyundai HB20

Tabela A.3: Resumo estatístico por semestre do valor de mercado do Hyundai HB20 (Tabela FIPE)

Ano - Semestre	Média (R$)	Mediana	Desvio	Mínimo	Máximo	n
2016 - 1º	50.960	50.199	7.137	39.202	65.917	104
2016 - 2º	52.941	51.932	7.639	38.810	66.040	108
2017 - 1º	55.187	55.815	7.058	38.760	67.900	97
2017 - 2º	57.196	58.006	6.874	44.711	69.477	74
2018 - 1º	57.596	59.889	7.229	44.647	70.185	79
2018 - 2º	57.983	60.362	7.765	43.069	70.154	94
2019 - 1º	57.748	59.854	7.811	43.029	70.885	96
2019 - 2º	59.203	59.450	8.907	41.832	75.810	87
2020 - 1º	62.946	64.800	9.851	41.255	79.269	118
2020 - 2º	66.171	68.075	8.434	47.151	79.845	90
2021 - 1º	73.091	74.110	9.875	49.601	92.490	92
2021 - 2º	82.805	84.123	9.564	60.496	102.120	123

A Tabela A.3 apresenta a evolução dos valores de mercado do Hyundai HB20 0km ao longo dos semestres entre 2016 e 2021. Nota-se uma valorização significativa do modelo, com aumento progressivo das médias semestrais. A média do 1º semestre de 2016 era

de aproximadamente R$50 mil, enquanto no 2º semestre de 2021 ultrapassa os R$82 mil, representando um crescimento expressivo ao longo da série.

Observa-se também um crescimento gradual da mediana e um aumento nos valores máximos registrados, o que indica que, além da valorização geral, o modelo passou a contar com versões mais completas ou sofisticadas ao longo do tempo. O desvio padrão também se elevou nos últimos anos, o que reflete uma maior dispersão nos preços praticados, possivelmente devido à ampliação da oferta de versões.

A.1.4 Ford Ka

Tabela A.4: Resumo estatístico por semestre do valor de mercado do Ford Ka (Tabela FIPE)

Ano - Semestre	Média (R$)	Mediana	Desvio	Mínimo	Máximo	n
2016 - 1º	45.550	45.413	3.182	40.270	50.517	20
2016 - 2º	46.113	46.277	3.427	40.862	51.089	24
2017 - 1º	46.867	46.501	3.362	40.939	52.042	30
2017 - 2º	47.962	48.037	3.322	41.685	53.036	37
2018 - 1º	47.010	47.921	3.731	39.905	52.854	48
2018 - 2º	52.453	50.861	7.652	40.109	68.926	78
2019 - 1º	53.835	53.126	7.828	41.612	67.353	72
2019 - 2º	57.092	56.788	7.800	43.721	68.548	62
2020 - 1º	56.845	57.551	7.470	43.750	68.010	54
2020 - 2º	58.443	59.110	7.074	43.740	69.654	54
2021 - 1º	60.744	62.116	6.368	49.206	70.282	47
2021 - 2º	62.265	62.385	6.111	51.225	72.002	42

A Tabela A.4 apresenta a evolução do valor de mercado do Ford Ka entre os anos de 2016 e 2021, segmentada por semestre. Os dados mostram um padrão de valorização contínua ao longo do período, com a média saindo de cerca de R$45 mil no primeiro semestre de 2016 e atingindo R$62 mil no segundo semestre de 2021.

A.1.5 Nissan March

Tabela A.5: Resumo estatístico por semestre do valor de mercado do Nissan March (Tabela FIPE)

Ano - Semestre	Média (R$)	Mediana	Desvio	Mínimo	Máximo	n
2016 - 1º	40.718	39.120	5.076	34.086	55.143	55
2016 - 2º	44.993	45.168	7.089	33.802	57.440	88
2017 - 1º	45.736	45.900	7.414	33.590	59.560	85
2017 - 2º	49.215	50.442	6.941	36.933	60.960	56
2018 - 1º	49.889	50.214	6.891	38.683	61.431	45
2018 - 2º	51.405	51.242	6.242	41.723	62.189	38
2019 - 1º	52.221	52.125	7.107	42.788	64.839	33
2019 - 2º	56.882	56.753	7.536	43.487	66.361	27
2020 - 1º	57.897	57.793	5.371	49.720	65.385	24
2020 - 2º	59.024	60.108	7.316	48.514	70.092	24
2021 - 1º	66.778	70.270	7.098	48.545	72.451	20
2021 - 2º	68.707	71.489	3.990	63.449	71.982	16

A Tabela A.5 apresenta a evolução semestral do valor de mercado do Nissan March entre 2016 e 2021, evidenciando um padrão claro de valorização ao longo do período. A média do primeiro semestre de 2016, que era próxima de R$40 mil, sobe gradualmente até atingir valores próximos de R$69 mil no segundo semestre de 2021.

A.1.6 Renault Sandero

Tabela A.6: Resumo estatístico por semestre do valor de mercado do Renault Sandero (Tabela FIPE)

Ano - Semestre	Média (R$)	Mediana	Desvio	Mínimo	Máximo	n
2016 - 1º	49.259	49.680	7.391	36.525	60.272	75
2016 - 2º	50.694	50.407	7.522	36.856	61.551	87
2017 - 1º	51.214	51.477	7.127	36.723	65.842	145
2017 - 2º	53.076	53.410	7.608	36.668	67.050	149
2018 - 1º	54.540	55.874	7.699	39.495	66.283	122
2018 - 2º	56.780	58.330	7.111	41.518	68.524	92
2019 - 1º	55.822	56.145	7.629	42.376	67.892	79
2019 - 2º	55.534	54.503	7.215	43.166	67.120	82
2020 - 1º	56.565	57.355	7.958	41.880	69.440	56
2020 - 2º	62.024	61.341	9.293	47.355	76.330	53
2021 - 1º	70.899	71.132	9.867	54.139	85.238	50
2021 - 2º	76.844	77.366	8.364	62.978	96.246	49

A Tabela A.6 evidencia uma trajetória de valorização constante do Renault Sandero entre 2016 e 2021. A média dos valores segue crescendo ao longo dos semestres, com destaque para os aumentos expressivos observados a partir de 2020.2. O desvio padrão também aumenta gradualmente, indicando maior dispersão dos preços. Os valores mínimos e máximos sobem significativamente no fim da série, reforçando a elevação dos patamares de preço do modelo. Além disso, a quantidade de observações por semestre garante boa representatividade para a análise descritiva.

A.1.7 Evolução dos Valores Médios dos Veículos

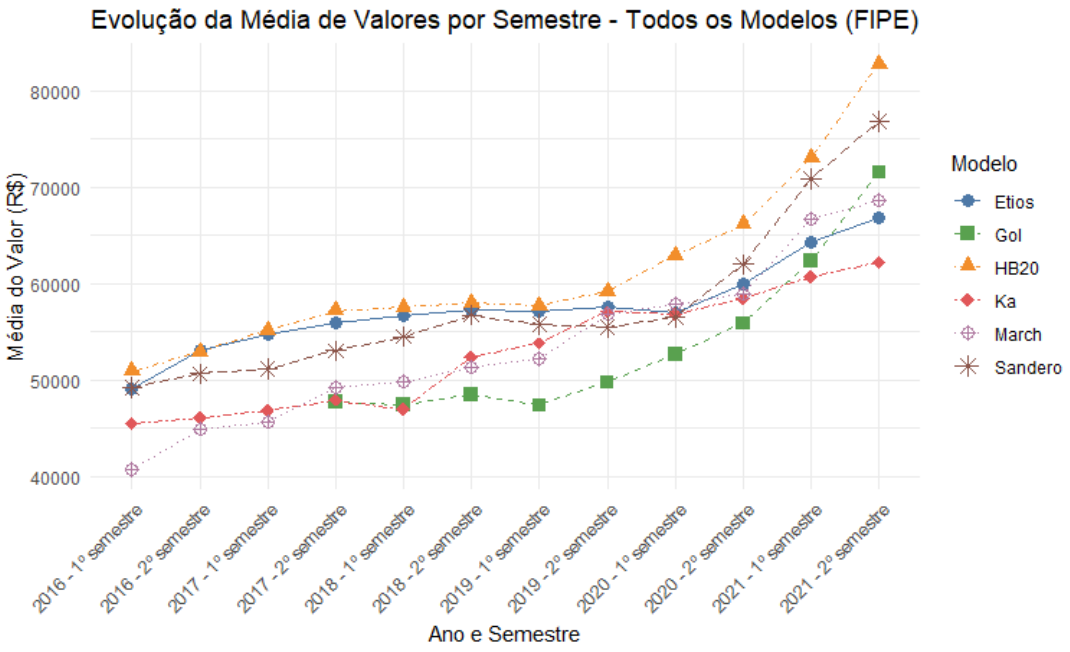

Figura A.1: Evolução da média dos valores de mercado (FIPE) por semestre para os modelos Etios, Gol, HB20, Ka, March e Sandero no período de 2016.1 a 2021.2.

A.2 Análise dos valores médios do prêmio do seguro (SUSEP)

A.2.1 Toyota Etios

Tabela A.7: Estatísticas descritivas por semestre do prêmio médio do seguro para o Toyota Etios (SUSEP)

Ano - Semestre	Média (R$)	Mediana (R$)	Desvio (R$)	Mínimo (R$)	Máximo (R$)	n
2016 - 1º	1393	1364	133	1231	1705	15
2016 - 2º	1480	1314	378	1159	2515	15
2017 - 1º	1693	1593	282	1274	2287	15
2017 - 2º	1608	1419	384	1276	2353	15
2018 - 1º	1609	1465	316	1223	2249	15
2018 - 2º	1543	1365	306	1202	2019	15
2019 - 1º	1505	1375	362	1171	2411	15
2019 - 2º	1394	1167	420	695	2133	15
2020 - 1º	1476	1210	984	544	4890	15
2020 - 2º	1207	1122	140	1060	1464	13
2021 - 1º	1557	1557	NA	1557	1557	1
2021 - 2º	1310	1412	751	123	2831	11

A Tabela A.7 apresenta as estatísticas descritivas do prêmio médio do seguro para o modelo Toyota Etios ao longo dos semestres entre 2016 e 2021. Observa-se que os maiores valores de prêmio médio ocorreram entre os anos de 2017 e 2018, especialmente no primeiro semestre de 2017, com média de R$ 1693. A partir de 2019, verifica-se tendência de redução, atingindo o menor valor médio no segundo semestre de 2020 (R$ 1207).

A.2.2 Volkswagen Gol

Tabela A.8: Estatísticas descritivas por semestre do prêmio médio do seguro para o Volkswagen Gol (SUSEP)

Ano - Semestre	Média (R$)	Mediana (R$)	Desvio (R$)	Mínimo (R$)	Máximo (R$)	n
2016 - 1º	1481	1353	518	953	3124	27
2016 - 2º	1463	1281	662	812	4353	27
2017 - 1º	1623	1581	414	772	2586	29
2017 - 2º	1541	1482	314	963	2502	27
2018 - 2º	1410	1397	388	405	2416	29
2019 - 1º	1463	1426	359	714	2329	28
2019 - 2º	1396	1314	367	829	2486	30
2020 - 2º	1413	1324	516	332	2825	24
2021 - 1º	1354	1178	589	769	2629	10
2021 - 2º	1292	1214	401	783	2338	20

A Tabela A.8 apresenta a evolução estatística dos prêmios médios do seguro para o modelo Gol ao longo dos semestres. Os maiores valores de prêmio foram registrados em 2017, especialmente no primeiro semestre, com média de R$ 1.623. A partir de 2018, nota-se tendência de redução, com médias inferiores a R$ 1.500 nos períodos subsequentes.

A.2.3 Hyundai HB20

Tabela A.9: Estatísticas descritivas por semestre do prêmio médio do seguro para o Hyundai HB20 (SUSEP)

Ano - Semestre	Média (R$)	Mediana (R$)	Desvio (R$)	Mínimo (R$)	Máximo (R$)	n
2016 - 1º	1683	1641	264	1364	2336	15
2016 - 2º	1727	1533	364	1350	2293	15
2017 - 1º	1851	1655	533	1461	3556	15
2017 - 2º	1863	1724	500	1441	3289	15
2018 - 1º	1761	1496	530	1314	2996	15
2018 - 2º	1472	1283	334	1015	2140	15
2019 - 1º	1497	1537	215	1195	1875	15
2019 - 2º	1490	1400	286	1137	1929	15
2020 - 1º	1510	1078	1158	260	3390	8
2020 - 2º	2248	2088	963	895	4486	10
2021 - 2º	3084	1538	3895	932	13662	10

A Tabela A.9 detalha as estatísticas semestrais do prêmio médio do seguro do HB20. Entre 2016 e 2019, as médias mantêm-se relativamente estáveis, com oscilações leves. A partir de 2020, os valores começam a subir, com saltos marcantes em 2020.2 e 2021.2. O último semestre da série (2021.2) apresenta média de R$ 3.084, impulsionada por valores extremos, incluindo um prêmio máximo de R$ 13.662.

A variabilidade é particularmente alta nos semestres de 2020 e 2021, com desvios padrão superiores a R$ 900 e, no caso de 2021.2, alcançando R$ 3.895. Esse comportamento indica forte dispersão nos dados.

A.2.4 Ford Ka

Tabela A.10: Estatísticas descritivas por semestre do prêmio médio do seguro para o Ford Ka (SUSEP)

Ano - Semestre	Média (R$)	Mediana (R$)	Desvio (R$)	Mínimo (R$)	Máximo (R$)	n
2016 - 1º	1101	1110	311	211	1657	25
2016 - 2º	1097	1048	327	84	1980	26
2017 - 1º	1341	1255	269	1057	2173	25
2017 - 2º	1246	1174	226	844	2064	27
2018 - 1º	1272	1247	300	278	1880	28
2018 - 2º	1275	1124	277	909	1942	29
2019 - 1º	1378	1336	292	1040	1990	29
2019 - 2º	1400	1292	369	971	2659	29
2020 - 1º	1373	1224	510	935	3157	28
2020 - 2º	1262	1216	190	975	1634	26
2021 - 1º	1612	1612	NA	1612	1612	1
2021 - 2º	1388	1372	385	749	2388	22

A Tabela A.10 reúne os dados descritivos por semestre para o Ford Ka. Observa-se crescimento gradual nos valores médios ao longo dos anos, com destaque para o primeiro semestre de 2021, cuja média foi de R$ 1.612. No entanto, o valor representa apenas uma observação, o que impede a análise de dispersão para o período.

O semestre com maior desvio padrão foi o primeiro de 2020 (R$ 510), refletindo elevada heterogeneidade nos prêmios. O menor desvio foi registrado no segundo semestre de 2020 (R$ 190), sugerindo maior homogeneidade nesse período. A presença de valores mínimos tão baixos quanto R$ 84 e máximos superiores a R$ 3.100 ao longo da série indica que o comportamento do prêmio médio do Ka é influenciado por variáveis que compõe o valor do prêmio do seguro.

A.2.5 Nissan March

Tabela A.11: Estatísticas descritivas por semestre do prêmio médio do seguro para o Nissan March (SUSEP)

Ano - Semestre	Média (R$)	Mediana (R$)	Desvio (R$)	Mínimo (R$)	Máximo (R$)	n
2016 - 1º	1219	1166	329	836	2229	14
2016 - 2º	1176	1150	192	898	1580	14
2017 - 1º	1679	1368	843	1226	4217	13
2017 - 2º	1457	1308	518	1142	3073	12
2018 - 1º	1443	1407	187	1226	1818	12
2018 - 2º	1402	1279	317	1093	2262	13
2019 - 1º	1308	1237	400	532	2116	13
2019 - 2º	1441	1255	405	1074	2338	14
2020 - 1º	1215	1259	245	771	1597	11
2020 - 2º	1327	1170	462	935	2381	11
2021 - 2º	1590	1590	NA	1590	1590	1

A Tabela A.11 apresenta a evolução dos prêmios médios do March por semestre. O maior valor foi registrado no primeiro semestre de 2017 (R$ 1.679), influenciado por alta dispersão (desvio padrão de R$ 843) e valor máximo superior a R$ 4.200. A partir de então, a média se mantém em patamares mais baixos, com oscilações moderadas.

Os menores prêmios aparecem em 2016 e 2020, com médias próximas de R$ 1.200. A retomada do crescimento ocorre no final da série, com R$ 1.590 no segundo semestre de 2021, embora esse dado corresponda a uma única observação. A variabilidade elevada em semestres como 2017.2 e 2020.2 destaca a heterogeneidade entre segurados.

A.2.6 Renault Sandero

Tabela A.12: Estatísticas descritivas por semestre do prêmio médio do seguro para o Renault Sandero (SUSEP)

Ano - Semestre	Média (R$)	Mediana (R$)	Desvio (R$)	Mínimo (R$)	Máximo (R$)	n
2016 - 1º	1312	1211	380	835	2308	15
2016 - 2º	1300	1184	263	1014	1984	15
2017 - 1º	1674	1491	589	796	3299	15
2017 - 2º	1590	1435	425	1208	2762	15
2018 - 1º	1242	1262	484	276	2292	13
2018 - 2º	1266	1225	228	887	1855	15
2019 - 1º	1302	1239	310	1070	2346	14
2019 - 2º	1161	1127	316	307	1786	14
2020 - 1º	1238	1158	186	1059	1650	12
2020 - 2º	1366	1182	664	817	3302	12
2021 - 1º	1963	1963	NA	1963	1963	1
2021 - 2º	1552	1686	374	1101	2160	8

A Tabela A.12 traz a evolução dos prêmios médios por semestre. O maior valor foi registrado no primeiro semestre de 2021 (R$ 1.963), embora represente apenas uma observação. O segundo semestre do mesmo ano manteve o patamar elevado, com média de R$ 1.552. A dispersão mais acentuada foi verificada em 2020.2 (desvio padrão de R$ 664), o que indica forte heterogeneidade dos prêmios. Já o menor valor médio da série aparece em 2019.2 (R$ 1.161), período de menor variação entre os contratos. A alternância de picos e quedas ao longo dos anos evidencia a sensibilidade do valor dos prêmios.

A.2.7 Evolução dos Prêmios Médios de Seguro

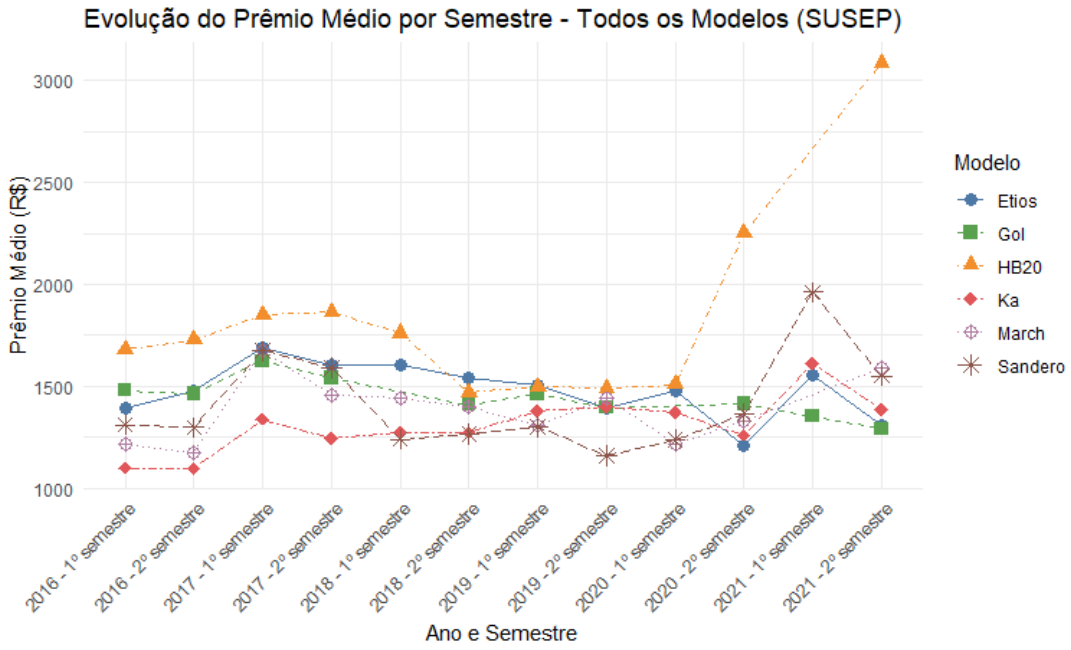

Figura A.2: Evolução do prêmio médio do seguro por semestre para todos os modelos analisados (SUSEP)

Apêndice B

Código em R

```
#########################################################
# TG B
# Plano 1
# Modelo Beta para indice premio/valor
# Regiao SP Grande Campinas
#########################################################

suppressPackageStartupMessages({
  library(readxl)
  library(writexl)
  library(dplyr)
  library(tidyr)
  library(ggplot2)
  library(betareg)
  library(broom)
  library(forcats)
  library(stringr)
})

options(contrasts = c("contr.treatment",
"contr.treatment"))

#########################################################
```

```
# 1 Leitura dos dados e selecao da base
############################################################

ARQ_BASE <- "C:/Users/Cliente/Documents/TCC/TG -
B/Consolidação dos Dados/Dados Consolidados.xlsx"
OUT_DIR  <- "C:/Users/Cliente/Documents/TCC/TG -
B/.Plano 1/Final"
if (!dir.exists(OUT_DIR)) dir.create(OUT_DIR,
recursive = TRUE, showWarnings = FALSE)

# ler planilha principal
dados_raw <- read_excel(ARQ_BASE, sheet = "Planilha1")

# padronizar nomes internos
dados <- dados_raw %>%
  rename(
    regiao    = `Regiao`,
    grupo     = Grupo,
    sexo      = `Sexo Condutor`,
    faixa     = `Faixa Etaria`,
    semestre  = `AnoModelo/Semestre`,
    premio_br = `Premio Medio (R$)`,
    valor_br  = `Valor Medio do Veiculo (R$)`,
    is_br     = `IS Media (R$)`,
    exp_br    = Expostos
  )

############################################################
# 2 Filtros de consistencia e recorte SP Grande Campinas
############################################################

dados_gc <- dados %>%
  filter(
```

```
    regiao == "SP - Grande Campinas",
    !is.na(grupo),  grupo  != "",
    !is.na(sexo),   sexo   != "",
    !is.na(faixa),  faixa  != "",
    !is.na(semestre), semestre != "",
    !is.na(valor_br), valor_br > 0,
    !is.na(premio_br), premio_br >= 0,
    !is.na(is_br),  is_br >= 0,
    !is.na(exp_br), exp_br > 0
  ) %>%
  mutate(
    grupo    = droplevels(factor(grupo)),
    sexo     = droplevels(factor(sexo)),
    faixa    = droplevels(factor(faixa)),
    semestre = factor(semestre)
  )

# ordenar semestres
niv_sem <- sort(unique(dados_gc$semestre))
dados_gc <- dados_gc %>%
  mutate(
    semestre = factor(semestre, levels = niv_sem)
  )

##########################################################
# 3 Tratamento de outliers e construcao do indice resposta
##########################################################

# funcao winsor 1 e 99
winsor <- function(x, probs = c(0.01, 0.99)) {
  qs <- quantile(x, probs = probs, na.rm = TRUE)
  pmin(pmax(x, qs[1]), qs[2])
}
```

```
# aplicar winsor e construir variaveis modeladas
dados_gc <- dados_gc %>%
  mutate(
    Premio   = winsor(premio_br),
    Valor    = winsor(valor_br),
    IS       = winsor(is_br),
    Expostos = winsor(exp_br),
    y        = Premio / Valor
  )

# correcao Smithson Verkuilen se houver 0 ou 1
n  <- nrow(dados_gc)
n0 <- sum(dados_gc$y <= 0, na.rm = TRUE)
n1 <- sum(dados_gc$y >= 1, na.rm = TRUE)
if ((n0 + n1) > 0) {
  dados_gc <- dados_gc %>%
    mutate(
      y = (y * (n - 1) + 0.5) / n
    )
}

# variaveis log
dados_gc <- dados_gc %>%
  mutate(
    logIS    = log(pmax(IS,    1e-8)),
    logValor = log(pmax(Valor, 1e-8)),
    logExp   = log1p(Expostos)
  )

##########################################################
# 4 Analise descritiva do indice e covariaveis
##########################################################
```

```
resumo_y <- dados_gc %>%
  summarise(
    n_obs  = n(),
    min_y  = min(y),
    q1_y   = quantile(y, 0.25),
    med_y  = median(y),
    mean_y = mean(y),
    q3_y   = quantile(y, 0.75),
    max_y  = max(y)
  )

resumo_y_grupo <- dados_gc %>%
  group_by(grupo) %>%
  summarise(
    n_obs   = n(),
    exp_tot = sum(Expostos),
    min_y   = min(y),
    q1_y    = quantile(y, 0.25),
    med_y   = median(y),
    mean_y  = mean(y),
    q3_y    = quantile(y, 0.75),
    max_y   = max(y),
    .groups = "drop"
  )

resumo_y_semestre <- dados_gc %>%
  group_by(semestre) %>%
  summarise(
    n_obs   = n(),
    exp_tot = sum(Expostos),
    mean_y  = mean(y),
    med_y   = median(y),
```

```
    .groups = "drop"
  )

write_xlsx(
  list(
    resumo_geral        = resumo_y,
    resumo_por_grupo    = resumo_y_grupo,
    resumo_por_semestre = resumo_y_semestre
  ),
  path = file.path(OUT_DIR,
  "resumos_indice_premio_valor.xlsx")
)

# grafico historico media ponderada por semestre
g_sem <- dados_gc %>%
  group_by(semestre) %>%
  summarise(
    y_med_w = weighted.mean(y, w = Expostos),
    .groups = "drop"
  ) %>%
  ggplot(aes(x = semestre, y = y_med_w, group = 1)) +
  geom_line() +
  geom_point() +
  labs(
    x = "Semestre",
    y = "Índice prêmio valor médio ponderado",
    title = "Evolução do índice prêmio valor por
    semestre"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

```
ggsave(file.path(OUT_DIR,
"indice_premio_valor_por_semestre.png"),
       g_sem, width = 7, height = 5, dpi = 300)

# boxplot por grupo
g_box_grupo <- ggplot(dados_gc, aes(x = grupo, y = y))
+
  geom_boxplot() +
  labs(
    x = "Grupo de veículo",
    y = "Índice prêmio valor",
    title = "Distribuição do índice prêmio valor por
    grupo"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(file.path(OUT_DIR,
"boxplot_indice_por_grupo.png"),
       g_box_grupo, width = 7, height = 5, dpi = 300)

# boxplot por faixa etaria
g_box_faixa <- ggplot(dados_gc, aes(x = faixa, y = y))
+
  geom_boxplot() +
  labs(
    x = "Faixa etária",
    y = "Índice prêmio valor",
    title = "Distribuição do índice prêmio valor por
    faixa etária"
```

```
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(file.path(OUT_DIR,
"boxplot_indice_por_faixa.png"),
       g_box_faixa, width = 7, height = 5, dpi = 300)

#########################################################
# Boxplot do índice Prêmio/Valor por semestre (padrão ABNT)
#########################################################

g_box_semestre <- ggplot(dados_gc, aes(x = semestre, y
= y)) +
  geom_boxplot(
    #fill = "#92c5de",
    alpha = 0.7,
    outlier.shape = NA
  ) +
  geom_smooth(
    aes(group = 1),
    method = "loess",
    se = FALSE,
    color = "#0571b0",
    linewidth = 1
  ) +
  labs(
    title = "Distribuição do índice Prêmio/Valor por
    semestre",
    x = "Semestre",
    y = "Índice Prêmio/Valor"
```

```
  ) +
  theme_bw(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid.major = element_line(color = "gray85"),
    panel.grid.minor = element_blank()
  )

ggsave(
  filename = file.path(OUT_DIR,
  "boxplot_indice_por_semestre.png"),
  plot = g_box_semestre,
  width = 10,
  height = 6,
  dpi = 300
)



############################################################
# 5 Ajuste dos modelos beta
############################################################

dados_beta <- dados_gc %>%
  mutate(
    grupo = droplevels(factor(grupo)),
    sexo  = droplevels(factor(sexo)),
    faixa = droplevels(factor(faixa))
  )

form_media_base <- y ~ semestre + grupo + sexo + faixa
+ logIS + logValor + logExp
form_phi_base   <- ~ logExp
```

```
mod_base <- betareg(
  formula  = form_media_base,
  data     = dados_beta,
  weights  = Expostos,
  link     = "logit",
  link.phi = "log"
)

form_media_int <- y ~ semestre + grupo + sexo * faixa
+ logIS + logValor + logExp
form_phi_int   <- ~ logExp

mod_int <- betareg(
  formula  = form_media_int,
  data     = dados_beta,
  weights  = Expostos,
  link     = "logit",
  link.phi = "log"
)

############################################################
# 6 Comparacao de modelos por AIC e BIC
############################################################

comp_mod <- tibble(
  modelo = c("base", "sexo_faixa"),
  AIC    = c(AIC(mod_base), AIC(mod_int)),
  BIC    = c(BIC(mod_base), BIC(mod_int))
)

write_xlsx(
  list(
```

```
    comparacao_modelos = comp_mod
  ),
  path = file.path(OUT_DIR,
  "comparacao_modelos_beta.xlsx")
)

##########################################################
# 7 Resumo do modelo final e tabelas de coeficientes
##########################################################

sum_int   <- summary(mod_int)
pseudo_R2 <- sum_int$pseudo.r.squared

tab_media <- tidy(mod_int, component = "mean") %>%
  mutate(
    OR = exp(estimate)
  )

tab_phi <- tidy(mod_int, component = "precision") %>%
  mutate(
    OR = exp(estimate)
  )

write_xlsx(
  list(
    coef_media    = tab_media,
    coef_precisao = tab_phi
  ),
  path = file.path(OUT_DIR,
  "coeficientes_modelo_beta.xlsx")
)

##########################################################
```

```
# 8 Diagnosticos do modelo beta final
##########################################################

dados_diag <- dados_beta %>%
  mutate(
    fitted      = fitted(mod_int, type = "response"),
    res_quant   = residuals(mod_int, type =
    "quantile"),
    res_pearson = residuals(mod_int, type = "pearson")
  )

# residuos quantilicos vs ajuste
g_rq <- ggplot(dados_diag, aes(x = fitted, y =
res_quant)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_point(alpha = 0.6, size = 1) + ylim(-20, 40) +
  labs(
    x = "Valores ajustados",
    y = "Residuos quantilicos",
    title = "Residuos quantilicos versus valores
    ajustados"
  ) +
  theme_bw()

ggsave(file.path(OUT_DIR,
"residuos_quantil_vs_ajuste.png"),
       g_rq, width = 7, height = 5, dpi = 300)

# residuos de pearson vs ajuste
g_rp <- ggplot(dados_diag, aes(x = fitted, y =
res_pearson)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_point(alpha = 0.6, size = 1) + ylim(-20, 40) +
```

```
  labs(
    x = "Valores ajustados",
    y = "Residuos de Pearson",
    title = "Residuos de Pearson versus valores
    ajustados"
  ) +
  theme_bw()

ggsave(file.path(OUT_DIR,
"residuos_pearson_vs_ajuste.png"),
       g_rp, width = 7, height = 5, dpi = 300)


##########################################################
# QQ plot com qqnorm e qqline (com limpeza de NAs e
infinitos)
##########################################################

# vetor de residuos quantilicos
rq <- dados_diag$res_quant

# manter apenas valores finitos
rq_f <- rq[is.finite(rq)]

# gerar apenas se houver dados suficientes
if (length(rq_f) > 5) {

  png(file.path(OUT_DIR,
  "qqplot_residuos_quantilicos.png"),
      width = 900, height = 700, res = 120)

  qqnorm(
    rq_f,
```

```
    main = "QQ plot dos residuos quantilicos",
    pch  = 19,
    cex  = 0.6,
    col  = "gray40"
  )
  qqline(rq_f, col = "red", lwd = 2)

  dev.off()

} else {
  warning("Residuos quantilicos insuficientes para QQ
  plot")
}

############################################################
# Histograma do índice Prêmio/Valor (padrão ABNT)
############################################################

g_hist <- ggplot(dados_gc, aes(x = y)) +
  geom_histogram(
    bins = 40,
    fill = "#2b83ba",
    color = "white",
    alpha = 0.9
  ) +
  labs(
    title = "Distribuição do índice Prêmio/Valor",
    x = "Índice Prêmio/Valor",
    y = "Frequência"
  ) +
  theme_bw(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold"),
```

```
    panel.grid.major = element_line(color = "gray85"),
    panel.grid.minor = element_blank()
  )

ggsave(
  filename = file.path(OUT_DIR,
  "histograma_indice_premio_valor.png"),
  plot = g_hist,
  width = 10,
  height = 6,
  dpi = 300
)


############################################################
# 9 Medias ajustadas para interpretacao
############################################################

med_logIS    <- median(dados_beta$logIS)
med_logValor <- median(dados_beta$logValor)
med_logExp   <- median(dados_beta$logExp)

# perfil de referencia para semestre
new_semestre <- expand_grid(
  semestre = levels(dados_beta$semestre)
) %>%
  mutate(
    grupo    = fct_relevel(dados_beta$grupo,
    sort(levels(dados_beta$grupo)))[1],
    sexo     = fct_relevel(dados_beta$sexo,
    sort(levels(dados_beta$sexo)))[1],
    faixa    = fct_relevel(dados_beta$faixa,
    sort(levels(dados_beta$faixa)))[1],
```

```
    logIS    = med_logIS,
    logValor = med_logValor,
    logExp   = med_logExp
  ) %>%
  mutate(
    y_hat = predict(mod_int, newdata = ., type =
    "response")
  )

g_sem_hat <- ggplot(new_semestre, aes(x = semestre, y
= y_hat, group = 1)) +
  geom_line() +
  geom_point() +
  labs(
    x = "Semestre",
    y = "Indice premio valor ajustado",
    title = "Indice ajustado por semestre para perfil
    de referencia"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(file.path(OUT_DIR,
"indice_ajustado_por_semestre.png"),
       g_sem_hat, width = 7, height = 5, dpi = 300)

# medias ajustadas por grupo
new_grupo <- expand_grid(
  grupo = levels(dados_beta$grupo)
) %>%
  mutate(
```

```
    semestre = levels(dados_beta$semestre)[1],
    sexo     = fct_relevel(dados_beta$sexo,
    sort(levels(dados_beta$sexo)))[1],
    faixa    = fct_relevel(dados_beta$faixa,
    sort(levels(dados_beta$faixa)))[1],
    logIS    = med_logIS,
    logValor = med_logValor,
    logExp   = med_logExp
  ) %>%
  mutate(
    y_hat = predict(mod_int, newdata = ., type =
    "response")
  )

g_grupo_hat <- ggplot(new_grupo, aes(x = grupo, y =
y_hat)) +
  geom_col() +
  labs(
    x = "Grupo de veiculo",
    y = "Indice premio valor ajustado",
    title = "Indice ajustado por grupo de veiculo"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(file.path(OUT_DIR,
"indice_ajustado_por_grupo.png"),
       g_grupo_hat, width = 7, height = 5, dpi = 300)

############################################################
# 10 Salvando objetos importantes
```

```
##########################################################

saveRDS(
  list(
    dados_filtrados    = dados_gc,
    dados_modelo       = dados_beta,
    modelo_base        = mod_base,
    modelo_final       = mod_int,
    pseudo_R2_final    = pseudo_R2,
    comparacao_modelos = comp_mod
  ),
  file = file.path(OUT_DIR,
  "objetos_plano1_beta_gc.rds")
)



##########################################################
# TG B - PLANO 2
# Modelo Gamma (link log) para prêmio médio (R$)
# Região: SP - Grande Campinas
# Resposta: premio
# Preditores: semestre + grupo + sexo + faixa + log_is
+ log_exp + log_valor
##########################################################

suppressPackageStartupMessages({
  library(readxl)
  library(writexl)
  library(dplyr)
  library(tidyr)
  library(ggplot2)
  library(scales)
  library(lmtest)
```

```
  library(sandwich)
  library(tibble)
  library(forcats)
  library(readr)
})

options(contrasts = c("contr.treatment",
"contr.treatment"))

############################################################
# 1 Leitura dos dados e seleção da base
############################################################

ARQ_BASE <- "C:/Users/Cliente/Documents/TCC/TG -
B/Consolidação dos Dados/Dados Consolidados.xlsx"
SHEET    <- "Planilha1"

OUT_DIR  <- "C:/Users/Cliente/Documents/TCC/TG -
B/.Plano 2_GLM_SEM_OFFSET/Final"
if (!dir.exists(OUT_DIR)) dir.create(OUT_DIR,
recursive = TRUE, showWarnings = FALSE)

if (!file.exists(ARQ_BASE)) stop("Arquivo não
encontrado: ", ARQ_BASE)
if (!(SHEET %in% readxl::excel_sheets(ARQ_BASE)))
stop("Aba não encontrada: ", SHEET)

# ler planilha principal
dados_raw <- read_excel(ARQ_BASE, sheet = SHEET)

# checar colunas necessárias
cols_needed <- c(
  "Regiao", "Grupo", "Sexo Condutor", "Faixa Etaria",
```

```
  "AnoModelo/Semestre",
  "Valor Medio do Veiculo (R$)", "Premio Medio (R$)",
  "IS Media (R$)", "Expostos"
)
miss <- setdiff(cols_needed, names(dados_raw))
if (length(miss) > 0) stop("Faltam colunas: ",
paste(miss, collapse = ", "))

# padronizar nomes internos
dados <- dados_raw %>%
  rename(
    regiao    = `Regiao`,
    grupo     = Grupo,
    sexo      = `Sexo Condutor`,
    faixa     = `Faixa Etaria`,
    semestre  = `AnoModelo/Semestre`,
    premio_br = `Premio Medio (R$)`,
    valor_br  = `Valor Medio do Veiculo (R$)`,
    is_br     = `IS Media (R$)`,
    exp_br    = Expostos
  )

############################################################
# 2 Filtros de consistência e recorte SP - Grande
Campinas
############################################################

# ordenar níveis de semestre a partir da própria
região GC
niv_sem_gc <- dados %>%
  filter(regiao == "SP - Grande Campinas") %>%
  pull(semestre) %>%
  unique() %>%
```

```
  sort()

dados_gc <- dados %>%
  filter(
    regiao == "SP - Grande Campinas",
    !is.na(grupo),  grupo  != "",
    !is.na(sexo),   sexo   != "",
    !is.na(faixa),  faixa  != "",
    !is.na(semestre), semestre != "",
    !is.na(valor_br),  valor_br  > 0,
    !is.na(premio_br), premio_br > 0,
    !is.na(is_br),     is_br     > 0,
    !is.na(exp_br),    exp_br    > 0
  ) %>%
  mutate(
    grupo    = droplevels(factor(grupo)),
    sexo     = droplevels(factor(sexo)),
    faixa    = droplevels(factor(faixa)),
    semestre = factor(semestre, levels = niv_sem_gc),
    premio   = as.numeric(premio_br),
    valor    = as.numeric(valor_br),
    is_media = as.numeric(is_br),
    exp_br   = as.numeric(exp_br)
  ) %>%
  droplevels()

stopifnot(nrow(dados_gc) > 0)

##########################################################
# 3 Tratamento de outliers e variáveis transformadas
(winsor + logs)
##########################################################
```

```
# função winsor 1 e 99
winsor <- function(x, probs = c(0.01, 0.99)) {
  qs <- quantile(x, probs = probs, na.rm = TRUE)
  pmin(pmax(x, qs[1]), qs[2])
}

dados_gc <- dados_gc %>%
  mutate(
    # NÃO winsorizar a resposta (premio)
    valor_w = winsor(valor),
    is_w    = winsor(is_media),
    exp_w   = winsor(exp_br),

    log_valor = log(pmax(valor_w, 1e-8)),
    log_is    = log(pmax(is_w,    1e-8)),
    log_exp   = log1p(pmax(exp_w, 0))
  )

# poda de faixas etárias com pouca sustentação
(remover faixas muito raras)
faixas_ok <- dados_gc %>%
  group_by(faixa) %>%
  summarise(
    n_faixa  = n(),
    exp_tot  = sum(exp_w),
    .groups  = "drop"
  ) %>%
  filter(n_faixa >= 5, exp_tot >= 50) %>%
  pull(faixa)

dados_gc <- dados_gc %>%
  filter(faixa %in% faixas_ok) %>%
  droplevels()
```

```
############################################################
# 4 Análise descritiva do prêmio e covariáveis
principais
############################################################

# resumo geral do prêmio
resumo_premio <- dados_gc %>%
  summarise(
    n_obs     = n(),
    min_prem  = min(premio),
    q1_prem   = quantile(premio, 0.25),
    med_prem  = median(premio),
    mean_prem = mean(premio),
    q3_prem   = quantile(premio, 0.75),
    max_prem  = max(premio)
  )

# resumo por grupo
resumo_premio_grupo <- dados_gc %>%
  group_by(grupo) %>%
  summarise(
    n_obs    = n(),
    exp_tot  = sum(exp_w),
    min_prem = min(premio),
    q1_prem  = quantile(premio, 0.25),
    med_prem = median(premio),
    mean_prem= mean(premio),
    q3_prem  = quantile(premio, 0.75),
    max_prem = max(premio),
    .groups  = "drop"
  )
```

```
# resumo por semestre
resumo_premio_semestre <- dados_gc %>%
  group_by(semestre) %>%
  summarise(
    n_obs    = n(),
    exp_tot  = sum(exp_w),
    med_prem = median(premio),
    mean_prem= mean(premio),
    .groups  = "drop"
  )

write_xlsx(
  list(
    resumo_geral        = resumo_premio,
    resumo_por_grupo    = resumo_premio_grupo,
    resumo_por_semestre = resumo_premio_semestre
  ),
  path = file.path(OUT_DIR,
  "resumos_premio_gamma_plano2.xlsx")
)

# boxplot do prêmio por grupo
g_box_premio_grupo <- ggplot(dados_gc, aes(x = grupo,
y = premio)) +
  geom_boxplot() +
  labs(
    x = "Grupo de veículo",
    y = "Prêmio médio (R$)",
    title = "Distribuição do prêmio médio por grupo de
    veículo"
  ) +
  theme_bw(base_size = 13) +
  theme(
```

```
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(
  filename = file.path(OUT_DIR,
  "boxplot_premio_por_grupo.png"),
  plot = g_box_premio_grupo,
  width = 10,
  height = 6,
  dpi = 300
)

# boxplot do prêmio por faixa etária
g_box_premio_faixa <- ggplot(dados_gc, aes(x = faixa,
y = premio)) +
  geom_boxplot() +
  labs(
    x = "Faixa etária",
    y = "Prêmio médio (R$)",
    title = "Distribuição do prêmio médio por faixa
    etária"
  ) +
  theme_bw(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(
  filename = file.path(OUT_DIR,
  "boxplot_premio_por_faixa.png"),
  plot = g_box_premio_faixa,
  width = 10,
  height = 6,
```

```
  dpi = 300
)

# série descritiva prêmio/valor por semestre
linha_ratio <- dados_gc %>%
  group_by(semestre) %>%
  summarise(
    premio_sobre_valor = mean(premio / valor_w, na.rm
    = TRUE),
    .groups = "drop"
  ) %>%
  arrange(semestre)

g_linha_ratio <- ggplot(linha_ratio, aes(x = semestre,
y = premio_sobre_valor, group = 1)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Evolução do prêmio/valor por semestre",
    x = "Semestre",
    y = "Média (Prêmio/Valor)"
  ) +
  theme_bw(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggsave(
  filename = file.path(OUT_DIR,
  "linha_media_premio_sobre_valor_plano2.png"),
  plot = g_linha_ratio,
  width = 10,
  height = 6,
```

```
  dpi = 300
)

############################################################
# 5 Ajuste do modelo Gamma (log) { SEM OFFSET
############################################################

# fórmula final alinhada ao resumo do Plano 2
form_gamma <- premio ~ semestre + grupo + sexo + faixa
+
  log_is + log_exp + log_valor

mod_gamma <- glm(
  formula = form_gamma,
  data    = dados_gc,
  family  = Gamma(link = "log"),
  weights = exp_w,
  control = glm.control(maxit = 300)
)

############################################################
# 6 Coeficientes com erros padrão robustos HC3 e
tabelas
############################################################

vc_HC3  <- sandwich::vcovHC(mod_gamma, type = "HC3")
mat_HC3 <- as.matrix(lmtest::coeftest(mod_gamma, vcov.
= vc_HC3))

ct <- tibble(
  Termo     = rownames(mat_HC3),
  Estimate  = mat_HC3[, 1],
  Std.Error = mat_HC3[, 2],
```

```
  Stat      = mat_HC3[, 3],
  p.value   = mat_HC3[, 4]
) %>%
  mutate(
    Multiplicador = exp(Estimate),
    IC95_inf      = exp(Estimate - qnorm(0.975) *
    Std.Error),
    IC95_sup      = exp(Estimate + qnorm(0.975) *
    Std.Error)
  )

rownames(ct) <- NULL

# salvar coeficientes completos
write_xlsx(
  list(
    coeficientes_gamma_sem_offset = ct
  ),
  path = file.path(OUT_DIR,
  "coeficientes_gamma_sem_offset.xlsx")
)

# tabela específica de semestre (para interpretação
temporal)
tab_sem <- ct %>%
  filter(grepl("^semestre", Termo)) %>%
  mutate(
    Semestre = sub("^semestre", "", Termo)
  ) %>%
  select(Semestre, Estimate, Std.Error,
         Multiplicador, IC95_inf, IC95_sup, p.value)
         %>%
  arrange(Semestre)
```

```
write_csv(tab_sem, file.path(OUT_DIR,
"tabela_semestre_gamma_sem_offset.csv"))

# coeficiente isolado de log_valor (efeito do valor do
veículo)
log_valor_row <- ct %>%
  filter(Termo == "log_valor")

write_csv(log_valor_row, file.path(OUT_DIR,
"coef_log_valor_sem_offset.csv"))

##########################################################
# 7 Diagnósticos do modelo (resíduos, Cook, QQ-plot)
##########################################################

dados_diag <- dados_gc %>%
  mutate(
    fitted   = fitted(mod_gamma, type = "response"),
    res_dev  = residuals(mod_gamma, type = "deviance"),
    res_pea  = residuals(mod_gamma, type = "pearson"),
    cook     = cooks.distance(mod_gamma)
  )

# (a) Resíduo deviance vs valores ajustados
g_res_dev <- ggplot(dados_diag, aes(x = fitted, y =
res_dev)) +
  geom_hline(yintercept = 0, linetype = "dashed",
  color = "gray40") +
  geom_point(alpha = 0.5, size = 1) +
  labs(
    title = "Resíduos deviance versus valores
    ajustados | Gamma (log) sem offset",
```

```
    x = "Valores ajustados (Prêmio)",
    y = "Resíduos deviance"
  ) +
  theme_bw(base_size = 13)

ggsave(
  filename = file.path(OUT_DIR,
  "residuos_deviance_vs_ajustados_sem_offset.png"),
  plot = g_res_dev,
  width = 9,
  height = 6,
  dpi = 300
)

# (b) QQ-plot dos resíduos de Pearson (filtrando não
finitos)
res_pea <- dados_diag$res_pea
res_pea_f <- res_pea[is.finite(res_pea)]

if (length(res_pea_f) > 5) {
  png(
    filename = file.path(OUT_DIR,
    "qqplot_residuos_pearson_sem_offset.png"),
    width = 900, height = 700, res = 120
  )
  qqnorm(
    res_pea_f,
  main = "QQ-plot dos resíduos de Pearson | Gamma
  (log) sem offset",
    pch  = 19,
    cex  = 0.6,
    col  = "gray40"
  )
```

```
  qqline(res_pea_f, col = "red", lwd = 2)
  dev.off()
} else {
  warning("Resíduos de Pearson insuficientes para QQ-
  plot")
}

# (c) Histograma da distância de Cook
g_cook <- ggplot(dados_diag, aes(x = cook)) +
  geom_histogram(bins = 40, color = "white", fill =
  "gray70") +
  labs(
    title = "Distribuição da distância de Cook | Gamma
    (log) sem offset",
    x = "Distância de Cook",
    y = "Frequência"
  ) +
  theme_bw(base_size = 13)

ggsave(
  filename = file.path(OUT_DIR,
  "histograma_cook_sem_offset.png"),
  plot = g_cook,
  width = 9,
  height = 6,
  dpi = 300
)

# Top 20 mais influentes (Cook)
top20_cook <- dados_diag %>%
  mutate(id = row_number()) %>%
  arrange(desc(cook)) %>%
  slice(1:20)
```

```
write_csv(top20_cook, file.path(OUT_DIR,
"top20_influentes_cook_sem_offset.csv"))

############################################################
# 8 Calibração por decis (Observed vs Predicted) e
razão O/E
############################################################

dados_plot <- dados_diag %>%
  transmute(
    premio,
    exp_w,
    fitted,
    res_pea
  ) %>%
  mutate(
    peso = ifelse(is.finite(exp_w) & exp_w > 0, exp_w,
    1)
  )

n_groups <- 10

# 8.1 Calibração por decis
dados_cal <- dados_plot %>%
  mutate(decis = dplyr::ntile(fitted, n_groups)) %>%
  group_by(decis) %>%
  summarise(
    pred_mean = weighted.mean(fitted, w = peso, na.rm
    = TRUE),
    obs_mean  = weighted.mean(premio, w = peso, na.rm
    = TRUE),
    w_sum     = sum(peso),
```

```
    w_var     = sum(peso * (premio - obs_mean)^2) /
    pmax(w_sum - 1, 1),
    se_obs    = sqrt(w_var / pmax(w_sum, 1)),
    ic_low    = obs_mean - qnorm(0.975) * se_obs,
    ic_high   = obs_mean + qnorm(0.975) * se_obs,
    .groups   = "drop"
  )

write_csv(dados_cal, file.path(OUT_DIR,
"tabela_calibracao_decis_sem_offset.csv"))

g_calib <- ggplot(dados_cal, aes(x = pred_mean, y =
obs_mean)) +
  geom_abline(slope = 1, intercept = 0, linetype = 2,
  color = "grey50") +
  geom_errorbar(aes(ymin = ic_low, ymax = ic_high),
  width = 0) +
  geom_point(size = 2) +
  labs(
    title = "Calibração por decis | GLM Gamma (log)
    sem offset",
    x = "Média predita por decis",
    y = "Média observada por decis (IC 95%)"
  ) +
  theme_bw(base_size = 13)

ggsave(
  filename = file.path(OUT_DIR,
  "calibracao_decis_obs_vs_pred_sem_offset.png"),
  plot = g_calib,
  width = 9,
  height = 6,
  dpi = 300
```

```
)

# 8.2 Razão Observado/Esperado (O/E)
dados_oe <- dados_cal %>%
  mutate(
    OE      = obs_mean / pmax(pred_mean, 1e-9),
    se_OE   = se_obs   / pmax(pred_mean, 1e-9),
    OE_low  = OE - qnorm(0.975) * se_OE,
    OE_high = OE + qnorm(0.975) * se_OE
  )

write_csv(dados_oe, file.path(OUT_DIR,
"tabela_OE_decis_sem_offset.csv"))

g_oe <- ggplot(dados_oe, aes(x = factor(decis), y =
OE, group = 1)) +
  geom_hline(yintercept = 1, linetype = 2, color =
  "grey50") +
  geom_errorbar(aes(ymin = OE_low, ymax = OE_high),
  width = 0.15) +
  geom_point(size = 2) +
  geom_line(alpha = 0.6) +
  labs(
    title = "Razão Observado/Esperado (O/E) por decis
    | GLM Gamma sem offset",
    x = "Decis do valor ajustado",
    y = "O/E (ideal = 1)"
  ) +
  theme_bw(base_size = 13)

ggsave(
  filename = file.path(OUT_DIR,
  "calibracao_OE_por_decis_sem_offset.png"),
```

```
  plot = g_oe,
  width = 9,
  height = 6,
  dpi = 300
)

# 8.3 Resíduo de Pearson vs ajustado (com LOESS)
g_res_pearson <- ggplot(dados_diag, aes(x = fitted, y
= res_pea)) +
  geom_hline(yintercept = 0, linetype = 2, color =
  "grey50") +
  geom_point(alpha = 0.25, size = 1) +
  geom_smooth(method = "loess", se = TRUE) +
  labs(
    title = "Resíduos de Pearson versus valores
    ajustados | Gamma (log) sem offset",
    x = "Valores ajustados (Prêmio)",
    y = "Resíduos de Pearson"
  ) +
  theme_bw(base_size = 13)

ggsave(
  filename = file.path(OUT_DIR,
  "residuos_pearson_vs_ajustados_loess_sem_offset.png",
  plot = g_res_pearson,
  width = 9,
  height = 6,
  dpi = 300
)

##########################################################
# 9 Medidas de ajuste: AIC e pseudo-R²
##########################################################
```

```
comparacao_mod <- tibble(
  modelo   = "Gamma_log_sem_offset",
  AIC      = AIC(mod_gamma),
  pseudoR2 = 1 - mod_gamma$deviance /
  mod_gamma$null.deviance
)

write_csv(comparacao_mod, file.path(OUT_DIR,
"comparacao_modelo_gamma_sem_offset.csv"))

############################################################
# 10 Salvando objetos importantes
############################################################

saveRDS(
  list(
    dados_filtrados    = dados_gc,
    modelo_gamma       = mod_gamma,
    coeficientes_HC3   = ct,
    tabela_semestre    = tab_sem,
    coef_log_valor     = log_valor_row,
    diagnosticos       = dados_diag,
    calibracao_decis   = dados_cal,
    calibracao_OE      = dados_oe,
    medidas_ajuste     = comparacao_mod
  ),
  file = file.path(OUT_DIR,
  "objetos_plano2_gamma_gc_sem_offset.rds")
)
```