FERNANDA VENTURATO ROQUIM

INVESTIGAÇÃO DA CAPACIDADE PREDITIVA DE MODELOS COM EFEITOS ALEATÓRIOS EM **GAMLSS** : UM ESTUDO EM DADOS DE SEGUROS DE AUTOMÓVEIS

LAVRAS – MG

2022

FERNANDA VENTURATO ROQUIM

INVESTIGAÇÃO DA CAPACIDADE PREDITIVA DE MODELOS COM EFEITOS ALEATÓRIOS EM GAMLSS:
UM ESTUDO EM DADOS DE SEGUROS DE AUTOMÓVEIS

Tese apresentada à Universidade Federal de Lavras como parte dos requisitos do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária para obtenção do título de doutor. Área de concentração: Análise de Regressão.

Prof. Dr. Renato Ribeiro de Lima
Orientador

Prof. Dr. Luiz Ricardo Nakamura
Coorientador

LAVRAS – MG
2022

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Roquim, Fernanda Venturato

Investigação da capacidade preditiva de modelos com efeitos aleatórios em GAMLSS : um estudo em dados de seguros de automóveis / Fernanda Venturato Roquim. – 2022.

111 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2022.
Orientador: Prof. Dr. Renato Ribeiro de Lima.
Coorientador: Prof. Dr. Luiz Ricardo Nakamura.
Bibliografia.

1. Precificação atuarial. 2. Distribuições ajustadas em zero. 3. Classificação por experiência. I. de Lima, Renato Ribeiro. II. Nakamura, Luiz Ricardo. III. Título.

FERNANDA VENTURATO ROQUIM

INVESTIGAÇÃO DA CAPACIDADE PREDITIVA DE MODELOS COM EFEITOS ALEATÓRIOS EM GAMLSS: UM ESTUDO EM DADOS DE SEGUROS DE AUTOMÓVEIS

INVESTIGATION OF THE PREDICTIVE CAPACITY OF MODELS WITH RANDOM EFFECTS IN GAMLSS: A STUDY ON AUTO INSURANCE DATA

Tese apresentada à Universidade Federal de Lavras como parte dos requisitos do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária para obtenção do título de doutor. Área de concentração: Análise de Regressão.

APROVADA em 25 de Outubro de 2022.

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho	UFLA
Prof. Dr. Danilo Machado Pires	UNIFAL
Prof. Dr. Thiago Gentil Ramires	UTFPR

Prof. Dr. Renato Ribeiro de Lima
Orientador

Prof. Dr. Luiz Ricardo Nakamura
Co-Orientador

LAVRAS – MG
2022

Ao meu amor, Rossi Chaves
e à minha avó adotiva, Ruth Chaves (in memoriam).

AGRADECIMENTOS

Agradeço aos meus orientadores, professores Renato e Nakamura, a estarem comigo há quase 7 anos nessa jornada. Obrigada por confiarem no meu potencial. Vocês o fizeram muitas vezes mais do que eu mesma. Obrigada por todos os ensinamentos e por não me deixarem desistir mesmo diante dos inúmeros obstáculos que se colocaram. Agradeço a todos os membros da banca, professores Júlio, Danilo e Thiago, todas as contribuições feitas e proporcionarem um ótimo momento durante a defesa. Vocês fizeram propostas essenciais para a melhoria deste trabalho. Obrigada por todas as contribuições extremamente pertinentes. Espero um dia me tornar uma excelente profissional, assim como vocês. Vocês me inspiram.

Agradeço ao meu companheiro de vida, Rossi. Sem você este trabalho não seria possível. Você foi o meu principal alicerce. Obrigada por todo auxílio na discussão aplicada a este trabalho, que você fez com maestria. Obrigada por ser caminho quando eu não sabia para onde ir.

Agradeço aos meus pais e familiares que sempre estiveram ao meu lado, incentivando-me e apoiando incondicionalmente. Agradeço à minha segunda família, meus sogros e meus cunhados, que sempre me lembraram de que eu seria capaz de concluir este trabalho e tornaram a minha trajetória mais leve e cheia de alegria. Peço desculpas pelas vezes em que não pude estar junto de vocês mesmo quando a saudade existia.

Agradeço aos meus colegas de doutorado, em especial meus amigos, Jorge e Ariana, as deliciosas conversas e pelo compartilhamento de angústias com a pesquisa. Ao meu amigo Matheus Felipe, o acreditar em mim e estar sempre presente, muitas vezes transformando minha casa em Lavras em um lar. Ao meu amigo Victor, todo apoio e companheirismo durante o *lockdown* em Governador Valadares. À minha amiga Isabela, que me recebeu de braços abertos em Vitória. Sou muito feliz de ter vocês na minha vida. Obrigada pelo alento, confiança e motivação.

Agradeço a todos os colegas da Secretaria de Avaliação Institucional da Universidade Federal do Espírito Santo a excelente e calorosa recepção logo que cheguei. Obrigada por toda gentileza e acolhimento. O apoio de vocês também foi fundamental para que eu conseguisse concluir esta pesquisa.

Por último, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES a financiar parcialmente esta pesquisa e à Universidade Federal de Lavras a oportunidade de estar adquirindo o título de doutora.

"Talvez seja bom partir do final
Afinal, é um ano todo só de sexta-feira treze
Cê também podia me ligar de vez em quando
Eu ando igual lagarta, triste, sem poder sair
...
E as fotos amarelam, como os dentes
As plantas, a gente, a chama, a febre intermitente
Vazia estrada, cheia a caixa de entrada
E, de repente, uma luz quadrada quente, diz que

Viver é partir, voltar e repartir
Partir, voltar e repartir"

(Emicida; É tudo pra ontem, 2020)

RESUMO

Os veículos automotivos são máquinas de grande relevância porque possibilitam, não só, mobilidade para os indivíduos, mas também diversos outros benefícios. Independentemente de sua serventia, a quantidade exorbitante de veículos que há circulando cotidianamente trazem alguns prejuízos, como o aumento no número de acidentes. As seguradoras se inseriram no mercado de seguros veicular como resposta a essa necessidade de asseguramento financeiro dos proprietários. A precificação deste tipo de seguro pode ser complicada, porque diferentes proprietários terão distintas características, que chamamos de classes de risco, e também diferentes comportamentos de condução, que são avaliadas através do histórico do segurado. Além disso, as próprias características dos valores das indenizações são de difícil estimação, devido ao excesso de valores nulos e ocorrências de valores extremos. Por isso, quanto mais adaptável e robusto é um modelo, melhor serão as predições. Nesta ocasião, o objetivo principal deste trabalho foi propor um modelo para precificação de sinistros que consiga abarcar essa complexidade. Utilizamos a classe de modelos de regressão, mais especificamente, modelos aditivos generalizados mistos para locação, escala e forma (GAMMLSS). O conjunto de dados utilizado é longitudinal e refere-se a clientes de uma empresa seguradora espanhola, trazendo diversas informações de apólices de seguros de automóveis, que foram acompanhadas ao longo de cinco anos. Foram testadas duas distribuições para variável resposta com diversas combinações de preditores, de covariáveis e de termos aditivos. Os principais achados apontam que o modelo que considerou o histórico do segurado gerou predições mais precisas e mais acuradas. Também, este modelo apresentou um comportamento que representa mais fidedignamente o que ocorreu na realidade. A metodologia proposta pode ser facilmente expandida para outros tipos de seguros.

Palavras-chave: Classes de risco. Classificação por experiência. Gama ajustada em zero. Modelo misto. Normal inversa ajustada em zero. Precificação.

ABSTRACT

Automotive vehicles are machines of high relevance as they enable not only mobility for individuals, but also have several other benefits. Regardless of their use, the exorbitant amount of vehicles circulating daily brings some complications, such as the increase in the number of traffic accidents. Insurers joined the vehicle insurance market as a response to the vehicle owners' necessity for financial insurance. Pricing for this type of insurance can be a difficult matter, since different owners will have different characteristics - which are called risk classes - and will also have different driving behaviors - which are evaluated through the policyholder's experience. In addition, the characteristics of claim values are difficult to estimate, due to the excess of null values and the occurrence of extreme values. Therefore, the more adaptable and robust a model is, the better the predictions will be. At this occasion, the main objective of this work was to propose a model for the pricing of claims that can encompass this complexity. We use the class of regression models, more specifically, generalized additive mixed models for location, scale and shape (GAMMLSS). The data is longitudinal and refers to customers of a Spanish insurance company, containing some information from auto insurance policies, which were monitored for five years. Two distributions were tested for the response variable with different combinations of predictors, covariates and additive terms. The main findings indicate that the model that considered the experience of the insured generated more precise and more accurate estimates. Also, this model presented a behavior in the predictions that more faithfully represents what happened in reality. The proposed methodology can be easily expanded to other types of insurance.

Keywords: Mixed model. Pricing. Rating experience. Risk classes. Zero adjusted Gamma. Zero adjusted inverse gaussian.

SUMÁRIO

1 INTRODUÇÃO

A história das pessoas e seus automóveis é antiga, datando do século XVIII, com o surgimento dos primeiros veículos autopropulsados capazes de transportar seres humanos (ECKERMANN, 2001). Com o desenvolvimento industrial, a produção automotiva se transformou em um mercado importante para a economia, com crescimentos expressivos. Por exemplo, o número de veículos automotores, na década de 90, era cerca de 400 milhões. Em 2022, o número estimado é de quase 1,5 bilhão de automóveis em todo o mundo (CHEN et al., 2022). Isto porque essas máquinas têm importância fundamental na sociedade, proporcionando conforto, mobilidade, oportunidades de trabalho e possibilitando pontes entre pessoas. Apesar das grandes vantagens, a rápida ascensão do número de veículos circulando também trouxe uma série de adversidades.

Uma consequência lamentável foi o crescimento paralelo do número de acidentes automobilísticos, com significativas perdas humanas e de capital financeiro. Mesmo com o investimento em equipamentos de segurança nos veículos, o número ainda é elevado, tendo como causa majoritária, ocorrências por falhas humanas. Em 2021, estima-se que ocorreram mais de 1,3 milhão de mortes causadas por incidentes com veículos (United Nations, 2022). A Organização Mundial de Saúde contabilizou que, em 2015, as lesões causadas por esses acidentes foram a principal causa de morte entre jovens de 15 a 29 anos e está entre as três principais em pessoas de 5 a 44 anos (WHO, 2015).

Neste contexto, as entidades seguradoras viram uma oportunidade de mercado, a partir do desenvolvimento de seguros veiculares como resposta à necessidade de segurança financeira às pessoas. Estes seguros podem ser do tipo facultativos, em que o indivíduo contrata o seguro por livre vontade, ou compulsório. Em diversos países, os proprietários de veículos são obrigados a ter pelo menos uma cobertura mínima de danos a terceiros, como política de segurança pública. Além disso, a forma como o seguro funcionará pode ser capaz de produzir diferentes comportamentos na maneira como a população conduzirá seus veículos, podendo causar um aumento ou redução da prudência no trânsito, de emissão de poluentes no ar e de congestionamentos. Dada a sua importância econômica e social, este tipo de seguro desafia os profissionais da área a projetarem estruturas ta-

rifárias que distribuam de forma mais adequada o ônus dos acidentes e que prezem pelo aumento da segurança dos indivíduos (OHLSSON; JOHANSSON, 2010).

O mercado de seguros também está acompanhando a era tecnológica e de ciências de dados. Atualmente diversos procedimentos relacionados a este mercado estão sendo reinventados e atualizados. Atendimento ao cliente, reclames de sinistros e coleta de dados estão acontecendo de forma mais rápida e com menor custo através de dispositivos inteligentes, internet, redes sociais, sensores de telemática e inteligência artificial (KLAPKIV; KLAPKIV, 2017). Além disso, está se apresentando uma tendência mundial à adesão de sistemas de dados abertos, em que as companhias seguradoras poderão disponibilizar seus dados e, em contrapartida, ter acesso aos dados de outras companhias (SUSEP, 2022; OZCAN et al., 2021). Esses sistemas de dados abertos possivelmente trará benefícios para os bons motoristas e seguradoras, que poderão estimar os prêmios com base no histórico de sinistro dos segurados, que tradicionalmente, só era possível através da fidelização do cliente. De um lado, o mercado segurador tem se adaptado a essas mudanças, do outro, os estatísticos e atuários estão pensando em melhores formas de tratar esses dados.

A ciência atuarial é a área do conhecimento responsável por calcular e predizer estes tipos de fenômenos e, seguidamente, precificar o risco (IBA, 2020b). A capacidade de modelagem e previsão dos cálculos atuariais melhoraram rapidamente com os avanços computacionais e metodologias estatísticas, como cita Belli, Medeiros e Prado (2018) e Ferreira, Carlos e Siqueira (2018). As ciências estatísticas e atuariais caminham paralelamente porque o aprimoramento de modelos estatísticos podem gerar inovações para ciências atuariais e vice-versa. Como exemplo dessa complementação das duas áreas, de um lado temos dois matemáticos britânicos com participação notável nas ciências atuariais, Nelder e Wedderburn (1972), a quem são atribuídos o mérito do desenvolvimento da classe de modelos lineares generalizados (GLM), uma classe de modelos importantíssima para a Estatística, do outro, inúmeras técnicas atuariais foram identificadas como casos particulares de determinados modelos estatísticos. Quanto mais flexível e adaptável é um modelo, melhores serão as predições e, logo, mais as ciências atuariais irão se desenvolver, trazendo maior rentabilidade para as companhias e preços mais justos para os usuários.

Em 2005, Rigby e Stasinopoulos (2005), propuseram os modelos aditivos generalizados para locação, escala e forma (*generalized additive models for location, scale and shape - GAMLSS*), que talvez seja a classe de modelos de regressão mais versátil disponível atu-

almente. Modelos da classe de regressão linear e não linear, GLM (NELDER; WEDDERBURN, 1972), modelos aditivos generalizados (GAM) (HASTIE; TIBSHIRANI, 1990) e até modelos mistos (FISHER, 1918) foram unificados, tornando-se casos particulares de um GAMLSS. Além de proporcionar estas diversas análises, também amplificou o escopo, permitindo várias distribuições para a variável resposta, além da modelagem de outros parâmetros da distribuição. Nesse contexto, pode ser extremamente vantajoso utilizar esta classe de modelos em conjunto com técnicas atuariais buscando encontrar modelos cada vez mais robustos, acurados e precisos, para obter estimativas justas de prêmios.

Diante deste quadro, definimos como objetivo principal desta pesquisa propor um modelo para a severidade de sinistros em seguros de automóveis, que leve em consideração a complexidade desse tipo de dado, com presença de excesso de zeros e valores extremos, mas que também contemple informações históricas dos indivíduos, e também, as classes de risco. O intuito é, não só, incentivar o uso dos GAMLSS dentro das ciências atuariais, porque o prêmio de risco pode ser obtido diretamente do modelo final proposto, mas também tencionar a adaptabilidade desta classe de modelos para determinadas situações, a fim de contribuir com as próprias discussões relacionadas ao avanço dos GAMLSS. Do ponto de vista estatístico, compararemos os ganhos de acurácia e precisão nas predições de modelos semiparamétricos e modelos semiparamétricos mistos. Do ponto de vista atuarial, analisaremos a capacidade preditiva do prêmio de risco em acertar os verdadeiros sinistros quando se utiliza apenas classes de risco ou uma combinação de classes de riscos e experiências.

Como utilizaremos modelos semiparamétricos e modelos semiparamétricos mistos, ou seja, quando há efeitos fixos e aleatórios, que serão melhores descritos na Seção 2.2, dentro da estrutura dos GAMLSS, optamos por formalizar uma subclasse de modelos que denominamos de GAMMLSS (*generalized additive mixed models for location, scale and shape*). Queremos investigar a performance da capacidade preditiva de ambos os modelos, GAMLSS e GAMMLSS, e mostrar a utilidade de cada um. Em outras palavras, responderemos às perguntas: quando há informação histórica do indivíduo em seguros, é mais vantajoso tomar como pressuposto que as observações são independentes ou lidar com essa dependência? O que acontece com as predições de um modelo quando se utiliza um efeito aleatório? É possível realizar tais análises utilizando-se modelos mais complexos, como GAMLSS?

Estas perguntas se mostram relevantes porque se referem à necessidade inerente de modelos que sejam suficientemente flexíveis, para lidar com características naturais de dados de seguros que são bem desafiadoras de serem modeladas. Elas surgem também pelas transformações do mercado, que precisam de boas técnicas para o tratamento de dados e melhor precificação, dado o próprio cenário recente de elevado incremento tecnológico, para proporcionar preços mais justos para todos. Para além das justificativas mercadológicas e sociais, no contexto estatístico, já vem sendo discutido na literatura sobre o imprescindível aprimoramento entre as teorias de GAMLSS e modelos mistos, como apontado por Welsh (2019).

2 REFERENCIAL TEÓRICO

2.1 Modelagem de dados de seguros

Esta seção trata de alguns aspectos históricos apenas com o objetivo de introduzir certos conceitos ao leitor. Aqui não temos o interesse em esgotar todo o conteúdo referente a seguros, que é extenso.

2.1.1 Conceitos iniciais

O papel fundamental da comercialização de seguros consiste em proporcionar uma proteção financeira, oferecendo um método de transferências de riscos em troca de um aporte monetário, chamado de prêmio, que é pago pelo segurado à entidade seguradora. Trabalharemos com a definição de risco como sendo um evento futuro e incerto, de natureza súbita e imprevista, independente da vontade do segurado, cuja ocorrência pode provocar prejuízos (SUSEP, 2020). Em outras palavras, é o incidente que acarreta a indenização e que, muitas vezes, tem sua probabilidade calculada.

Levando em consideração que o risco não é idêntico para todos os indivíduos, uma vez que cada um tem diferentes características comportamentais individuais, é sensato pensar que este aporte deva corresponder às características do risco assegurado. Se todos os segurados, com diferentes riscos, pagarem uma mesma tarifa, acontece o fenômeno que é chamado de antisseleção, em que grandes riscos são cobertos a preços baixos, gerando grandes indenizações sem a equivalente entrada de prêmios, enquanto baixos riscos são cobertos com alto custo, o que, frequentemente, levará o segurado a não contratar o seguro (OHLSSON; JOHANSSON, 2010). Toda esta situação, pode, inclusive, levar a seguradora à falência. Portanto, há também, diretamente, o interesse da seguradora na melhor divisão dos prêmios possível. Isto é, tanto do ponto de vista do segurado, quanto da seguradora, é importante o aperfeiçoamento da precificação dos riscos. Esta necessidade é inerente a qualquer produto de seguro comercializado e em diferentes segmentos do mercado. Por exemplo, uma maneira de melhor especificar o risco individual pode ser a partir da divisão de toda uma carteira de apólices em subgrupos que tenham em comum determinadas características ou um mesmo fator de influência (MEYERS; HOYWEGHEN, 2018), que abordaremos com mais detalhes na Seção 2.1.2.

Este processo de modelar algum grau de incerteza – risco – objetivando determinar um preço ótimo para o seguro, ou para uma reserva, é uma das funções realizadas pelos cientistas atuariais (MARTINS, 2020). As ciências atuariais são um campo relativamente novo, tendo sua consolidação na metade do século XIX (IBA, 2020a), apesar de já existirem diversas formas mais elementares de cálculo de seguros anteriormente, mas que não eram formalizadas enquanto uma ciência. Pode-se dizer que a história dos seguros remonta do século XIII, quando surgiram as primeiras apólices para assegurar embarcações contra riscos marítimos (HICKMAN, 2006). Os processos de cálculos de seguros foram se desenvolvendo à medida que os modelos estatísticos e processamento computacional avançaram, até se tornar a ciência atuarial que conhecemos hoje. Esta é uma ciência que tem forte relação com os campos da matemática, da probabilidade e da estatística, justamente pela característica de incerteza que é inerente aos produtos de seguros. Mas esta ciência não trabalha apenas com seguros, atuando também em outras modalidades que envolvam alguma incerteza atrelada a produtos financeiros, como aposentadorias, pensões, planos de saúde, risco de crédito e muitos outros produtos atuariais (IBA, 2020a).

Os produtos atuariais são classificados em duas grandes áreas, sendo elas o ramo vida, que envolvem a segurança direta da vida de pessoas, e ramo não vida, que trata de seguros para bens materiais, objetos, tangíveis ou intangíveis, trazendo uma segurança indireta aos indivíduos proprietários (TEUGELS; RAMSEY, 2006). A classe do ramo vida abarca seguros de vida, planos de saúde, aposentadorias, pensões, previdências, dentre outros e, em geral, dependem fortemente de processos populacionais e demografia. O ramo não vida trata, por exemplo, de seguros de automóveis, imóveis, rurais, marítimos, aeronáuticos e até cotações de preços de *commodities*. Na presente análise focaremos na estimação do valor de seguros do ramo não vida, mais especificamente, seguro de automóveis.

Os cálculos para este ramo se assemelham consideravelmente com as metodologias de análise de regressão, uma vez que estaremos interessados em investigar determinado fenômeno, como ocorrência e severidade de sinistros, em função de variáveis explicativas, que são chamadas de classes de risco pela ciência atuarial. Historicamente, estes cálculos eram feitos de forma determinística e, posteriormente, limitados a modelos gaussianos (DAVID, 2015). Diversas técnicas de matemática atuarial surgiram, sendo que algumas delas serão abordados na Seção 2.1.3. Os atuários e estatísticos perceberam grande

vantagem em trabalhar conjuntamente, até porque ambas ciências se relacionam dialeticamente. A evolução e ou revisão de modelos estatísticos podem gerar novos paradigmas para ciências atuariais e vice-versa. Muitos modelos atuariais foram reconhecidos como casos particulares de determinados modelos de regressão. E observar que tais análises podem ser feitas dentro da estrutura de regressão coloca à disposição dos atuários todo o ferramental estatístico de inferência, adequação e predição na busca de melhor calcular os riscos. Maiores detalhes sobre ciências atuariais podem ser encontrados em MacGinnitie (1980), Bühlmann (1997), Boland (2007), IFA (2020) e IBA (2020b).

Também faz-se necessário definir alguns conceitos inerentes da área que são importantes para a análise. Utilizamos como fonte as definições apresentadas pela SUSEP (2020), IAIS (2022), EIOPA (2022) e Insurance Europe (2022), que são apresentadas no Quadro 2.1.

Quadro 2.1 – Conceitos e suas definições para a área de ciências atuariais

Conceito	**Definição**
Apólice	Documento emitido pela entidade seguradora formalizando a aceitação da cobertura solicitada pelo segurado e dos termos de funcionamento do seguro.
Carteira	Conjunto dos contratos de seguro de um mesmo ramo ou ramos afins, emitidos por uma seguradora.
Classe de risco	Forma de designar um grupo de segurados que apresentem riscos aproximadamente equivalentes.
Indenização	Valor a ser pago pela seguradora ao segurado ou beneficiário na ocorrência do evento coberto.
Prêmio	Importância paga pelo segurado à seguradora para que esta assuma o risco a que o segurado está exposto.
Reclamação	Ato de apresentação, pelo segurado, ao segurador, do seu pedido de indenização.
Severidade	Está associada ao valor do sinistro e representa a gravidade do impacto econômico do acidente.
Sinistro	Ocorrência do risco coberto, durante o período de vigência do plano de seguro.

Fonte: Da autora (2022)

Dentro da área de seguros, existem alguns tipos de prêmios. O cálculo do prêmio se inicia com o chamado prêmio de risco, ou prêmio estatístico, que é o valor precificado diretamente da classe de risco. Ele é dado por

$$\text{Prêmio de risco} = \text{Probabilidade da ocorrência} \times \text{Valor médio do sinistro},$$

para determinada classe de risco. Tendo em mãos o prêmio de risco, podemos obter todos os outros, com a adição direta de alguns carregamentos financeiros. O prêmio puro é dado pelo prêmio de risco adicionado de um carregamento de segurança. O prêmio comercial é dado pelo prêmio puro, adicionado de carregamentos comerciais, como lucro e comissionamento de venda, por exemplo. O prêmio bruto é dado pelo prêmio comercial adicionado de encargos tributários e custos de emissão da apólice. Por fim, obtém-se o prêmio final cobrado, que é o próprio prêmio bruto, multiplicado pelo número de objetos a serem assegurados na apólice (CUMMINS, 1988). No Quadro 2.2 é possível visualizar um esquema da construção dos diferentes prêmios. Na presente pesquisa focaremos apenas na obtenção do prêmio de risco.

Quadro 2.2 – Tipos de prêmios e carregamentos aplicados

	Prêmio de risco	Carregamentos de segurança	Carregamentos comerciais	Tributos	Nº objetos segurados
Prêmio puro	×	×			
Prêmio comercial	×	×	×		
Prêmio bruto	×	×	×	×	
Prêmio final	×	×	×	×	×

Fonte: Da autora (2022)

As seguradoras, ainda, costumam criar alguns tipos específicos de produtos de seguros, ou condições na apólice que alteram essa relação prêmio-indenização. Existem alguns processos que são bastante burocráticos nos seguros que geram certa demora entre o ato da reclamação do sinistro até o pagamento da indenização ao beneficiário, dentre eles, os processos de vistoria e em alguns casos, processos judiciais. Para aumentar a eficiência e redução do tempo destes procedimentos surgiu o seguro *sem culpa*, cujo nome se origina do fato de o segurado não ser obrigado a comprovar culpa no acidente. Este seguro, na maioria dos países, tem três principais características: o valor da indenização é pré-fixado, é um seguro compulsório para danos pessoais e restringe o direito de processo por acidente automobilístico. Cummins, Phillips e Weiss (2001) alertam que, mesmo sendo um procedimento mais ágil, foi observado um aumento da fatalidade e severidade dos acidentes após a implementação deste tipo de seguro, que é desaconselhada pelos autores.

2.1.2 Classificação por riscos

No que tange às tarifas das classes de riscos, alguns comportamentos em comum são detectados. Em geral, motoristas do sexo feminino apresentam menor risco que do sexo masculino. Bergdahl (2005) aponta que mulheres sentem mais necessidade de estarem asseguradas e são motoristas mais prudentes, respeitando mais as regras de trânsito. Além disso, membros do sexo masculino mostraram menor preocupação em relação às mulheres, quando dirigindo em situações adversas, como à noite, durante uma tempestade, em locais desconhecidos, ou até mesmo sob o efeito de álcool ou muito cansaço. Estatisticamente, acidentes de automóveis acontecem com mais frequência e maior gravidade para motoristas do sexo masculino, apesar de serem o público com maior exposição (VEEVERS; GEE, 1986). Conclusões análogas também foram constatadas por Lucas, Mendes-Da-Silva e Lyons (2019), Farrow e Brissing (1990), Harré, Field e Kirkwood (1996), Veevers (1982), Mannering (1993), dentre diversas outras pesquisas. Entretanto, Bergdahl e Norris (2002) dizem que o comportamento de direção feminino e masculino tendem a se tornarem mais semelhantes, como os aumentos da quantidade de motoristas mulheres, do tempo e da quilometragem que estas dirigem. Apesar de ser uma tarifagem ampla e historicamente utilizada pelas seguradoras, o mercado parece estar mudando. Desde 2012, a utilização do sexo para a precificação do prêmio é proibida na União Europeia (ASEERVATHAM; LEX; SPINDLER, 2016). Medders, Parson e Thomas-Reid (2021) criticam a adoção de precificação baseada em sexo e argumentam que é uma forma potencialmente injusta e discriminatória de classificação, que pode ter efeitos negativos no próprio mercado segurador.

A idade do motorista também é um fator de risco muito utilizado pelas seguradoras na hora de precificar. As justificativas são diversas e perpassam por questões fisiológicas, psicológicas e comportamentais. Em geral, motoristas mais novos são considerados mais imprudentes, irresponsáveis e menos experientes para conduzir veículos, estando associados a um maior risco para a seguradora. David (2015) constatou um decréscimo no valor médio dos sinistros à medida que avança a idade do cliente, isto é, os acidentes com maior severidade estavam prioritariamente associados a motoristas mais novos. Resultados semelhantes também foram encontrados por Islam e Mannering (2006), Abdel-Aty, Chen e Radwan (1999), Stamatiadis e Deacon (1995), Kim et al. (1998), Franceschi et al. (2022) e outros. No Brasil e na Espanha, a idade mínima para poder conduzir automóveis é de

18 anos, mas há países, como os Estados Unidos e Canadá, em que a idade mínima é de 16 anos e em alguns estados, 15 anos. Laberge-Nadeau, Maag e Bourbeau (1992) apresentam uma discussão crítica sobre essas idades mínimas para licenciamento e apontam para a correlação entre o excesso de mortes entre jovens associados a acidentes automobilísticos.

A região de trânsito do veículo também costuma ser um aspecto importante. Em geral, as seguradoras classificam as regiões em zona urbana e zona rural. Às vezes, também subdividem territórios em faixas associadas a índices de violência e criminalidade ou em densidade populacional, sendo esta última bastante comum na União Europeia (ABDEL-ATY, 2003). As divisões são diversas e feitas para tentar identificar possíveis fatores de influência nos acidentes ou roubos. Khorashadi et al. (2005) fazem uma análise em busca de verificar diferenças de risco entre zonas urbanas e rurais. Eles concluíram que ambas regiões apresentam suas particularidades mas nenhuma diferença significativa no risco entre estas. Em geral, a frequência de acidentes nas regiões rurais é menor que nas urbanas, mas com severidade e mortalidade maior que as registradas nos centros. Isso também foi constatado por Sherafati et al. (2017), Modarres et al. (2014) e Yazdani-Charati, Siamian e Ahmadi-Basiri (2014). As definições de zonas urbanas e rurais podem variar entre os diversos países e algumas delas são apresentadas em IBGE (2017).

A presença de um segundo motorista envolve naturalmente o registro de dois perfis de motorista, logo, maior variabilidade e maior risco. Se a classe de risco do segundo condutor for mais elevada que a do principal, a tendência é que o preço do prêmio aumente consideravelmente (EDLIN; KARACA-MANDIC, 2006). Isto também foi constatado por Saito, Kato e Shimane (2010), Hultkrantz, Nilsson e Arvidsson (2012) e muitos outros.

Além de coletar características do segurado, geralmente são coletadas informações do veículo, como seu valor de mercado, idade, quilometragem rodada, potência do motor, tipo, dentre outros. O valor do bem assegurado é ligado diretamente ao valor do prêmio, isto é, quanto maior o valor do veículo, maior pode ser o gasto que a seguradora terá em indenização no caso de sinistro (OHLSSON; JOHANSSON, 2010). A idade do veículo costuma apresentar diversas informações sobre o tipo e grandeza do risco. Em geral, veículos novos apresentam maior risco porque estão mais suscetíveis a roubos e a maiores indenizações por parte da seguradora devido ao seu valor. Já os mais antigos, costumam acionar o seguro mais frequentemente por falhas mecânicas, configurando indenizações mais baratas. Em geral, seguros para carros mais novos são mais caros do que para carros

mais antigos (FREES; VALDEZ, 2008). Em relação à potência do motor, esta costuma ter uma relação direta com o risco. Veículos mais potentes se envolvem em acidentes com maior gravidade e logo, maior indenização. Este efeito foi observado por Wen, Wang e Lawrence (2007), Aseervatham, Lex e Spindler (2016), Abdelhadi, Elbahnasy e Abdelsalam (2020) e Campbell (1986). A classificação em tipos podem ser as mais diversas, como por quantidade de eixos, por modelo, por marca, por categoria, e outros.

No contexto de seguros, todas as variáveis exemplificadas estão associadas ao que chamamos de classificação por risco. No contexto de regressão, podemos pensar nisso como uma modelagem da distribuição dos sinistros em termos de variáveis explicativas.

2.1.3 Classificação por experiência

Uma das principais atividades de um atuário é delinear uma estrutura de tarifas para que os prêmios sejam o mais justos possíveis para determinado segurado. Isto, pode ser feito através da partição da carteira em classes de riscos, como apresentado anteriormente na Seção 2.1.2, em que todos dentro daquela determinada classe, pagam o mesmo valor. Em teoria, estas classes são consideradas homogêneas, o que não se aplica muito bem na prática, uma vez que não levam em consideração algumas características individuais dos motoristas, como rapidez de reflexo, agressividade ao volante, domínio de condução do veículo, conhecimento de ruas e rodovias e até o uso de álcool ou substâncias ilícitas enquanto dirige. Estes fatores são extremamente difíceis, se não impossíveis, de serem mensurados de forma econômica e causam heterogeneidade dentro da classe de risco (LEMAIRE, 1995). Uma das formas de contornar este problema é através da classificação, também, por experiência ou mérito.

Para tal, a fidelização dos clientes dentro de uma carteira de seguros é um aspecto fundamental. Olhando pelo lado econômico, é de interesse da seguradora manter clientes rentáveis que irão renovar ou contratar novos tipos de seguros. Para isso, a companhia seguradora pode adotar diversos modelos de *marketing* e modelos para retenção. Serrano (2016) e Frees et al. (2021) apontam que estratégias de venda de diferentes produtos de seguro de forma conjunta aumentam a fidelidade. Meeboonsalang e Chaveesuk (2020) afirmam que, do ponto de vista da seguradora, é mais interessante investir em políticas de fidelização de clientes ativos, do que investir na obtenção de novos clientes. A confiança do segurado no serviço prestado pelo segurador e o preço do prêmio final cobrado tam-

bém são essenciais (DAMTEW; PAGIDIMARRI, 2013). Do ponto de vista estatístico, carteiras com informações históricas dos clientes proporcionam estimativas mais acuradas e mais precisas, baseadas no próprio comportamento do indivíduo. Em outras palavras, a seguradora vai aprendendo sobre o comportamento de risco daquele cliente ao longo dos anos e proporcionando estimativas, inclusive, mais justas, uma que vez que estes podem estar sendo penalizados por características externas que não podem controlar, como sexo e idade, por exemplo. Os resultados de Arvidsson (2011) indicam que novos clientes têm riscos desproporcionalmente mais altos que o de clientes leais.

O acesso ao histórico de sinistralidade de clientes, em geral, é obtido pela fidelização e, por ser uma informação valiosa, não é disponibilizado facilmente. Entretanto, como já mencionado anteriormente, as tecnologias estão alterando a forma de funcionamento e cálculo de seguros. Autoridades da União Europeia e dos Estados Unidos, em 2015, se organizaram para a criação e implementação de um sistema financeiro de dados abertos, chamado de *Open Banking*, que surgiu com o intuito de trazer maior equilíbrio e competitividade para este mercado (OZCAN et al., 2021). Desde sua implementação, diversas entidades financeiras poderiam compartilhar seus dados através deste sistema que é regulamentado por diversos órgãos regulatórios, prezando pela segurança das informações. Assim, outras companhias também poderiam ter acesso a este grande conjunto de dados gerais (BOTTA et al., 2018). No Brasil, este sistema foi implementado no começo de 2021 (CNN Brasil, 2022), trazendo consigo o Sistema de Seguros Abertos, que possibilita as entidade seguradoras a compartilharem entre si suas informações e registros sobre clientes. Só poderão acessar estes, aquelas sociedades que também disponibilizaram os seus, sendo que o sistema também garante meios de privacidade e proteção destes dados (SUSEP, 2022). É um movimento que está ocorrendo em todo o mercado mundial e beneficiará os bons motoristas e as próprias entidades seguradoras, que poderão estimar prêmios mais competitivos, com maior informação histórica dos indivíduos.

Os métodos de classificação por experiência podem ser aplicados tanto de forma retrospectiva quanto prospectiva. Nos métodos retrospectivos, é feito o reembolso de parte do valor pago no prêmio, na ocorrência de um evento favorável do ponto de vista da seguradora. Este é um método mais comum para seguros do ramo vida. Para o ramo não vida, o mais usual são métodos prospectivos, em que, na ocorrência de um evento favorável, o cliente é recompensado com um menor preço de prêmio para a renovação

(FREES, 2009). Há duas maneiras de classificação por experiência que são amplamente utilizadas pelo atuários, sendo elas, os sistemas *bonus-malus* e a teoria de credibilidade.

Um dos métodos mais simples, mas que ainda é amplamente utilizado em seguros de automóveis, é o sistema *bonus-malus*, que consiste na adição de um peso no valor do prêmio puro, sendo que cada país tem sua forma específica de cálculo. Este sistema funciona através da aplicação de um desconto no prêmio do ano seguinte (*bonus*), caso o segurado não tenha reclamado seguro no ano anterior ou um acréscimo (*malus*), caso tenha utilizado. Quanto mais anos consecutivos sem utilizar o seguro, maior o desconto. O valor do coeficiente do sistema de *bonus-malus* costuma ser um valor inferior a 1, para desconto, e maior que 1, quando acréscimo. David (2015) aponta que tarifas que levam em consideração o histórico de sinistros do indivíduo tendem a criar um sistema em que os segurados são encorajados a dirigem com mais prudência. Lemaire (1995) elenca diversas técnicas de se obter pesos ótimos para esse coeficiente.

2.1.4 Teoria de credibilidade

A teoria de credibilidade, também chamada de credibilidade atuarial, é outra técnica que visa a melhor forma de obter prêmios de seguros para um determinado risco, levando em conta a sua sinistralidade média e futura (RINCON, 2012). O princípio geral de credibilidade aplicado a um prêmio de risco é dado por uma média ponderada da forma

$$\text{Próximo prêmio} = \zeta \times \text{Experiência de sinistros} + (1 - \zeta) \times \text{Prêmio anterior},$$

em que ζ é o fator de credibilidade ($0 \leq \zeta \leq 1$), sendo que se $\zeta = 1$ tem-se o chamado fator de credibilidade completo, em que somente a experiência é utilizada para precificação e, se $\zeta = 0$ tem-se fator sem credibilidade, em que toda a experiência é ignorada e apenas as classes de risco (Prêmio anterior) são utilizadas (FREES, 2009). O desafio está em obter um valor ótimo para ζ.

Mowbray (1914) propôs uma alternativa para se saber o tamanho amostral mínimo para que se fosse possível utilizar credibilidade completa. No entanto, havia o pressuposto de distribuição Normal dos valores dos sinistros e, frequentemente, grandes amostras eram demandadas, o que nem sempre estava disponível. Whitney (1918) propôs uma forma de se obter ζ para esses casos, denominada de credibilidade parcial, que é dada por

$$\zeta = \min\left\{1, \sqrt{n/n_F}\right\},$$

em que n é o tamanho amostral e n_F é o tamanho amostral mínimo estimado pelo método de Mowbray.

Esses métodos são consideravelmente elementares e foram utilizados por décadas, até que Bühlmann (1967) propôs a credibilidade de acurácia ótima. Bühlmann hipotetizou que existem características não observadas, que chamaremos de α, que são comuns a todas as observações de um mesmo grupo. Embora não observáveis, é possível obter alguma informação sobre elas, a partir de repetições das observações, no caso, de sinistros. Sendo Y uma variável aleatória i.i.d. dos valores de sinistros, provindas de uma população desconhecida, a esperança e variância condicional sendo $E(Y|\alpha)$ e $Var(Y|\alpha)$, respectivamente, o estimador de credibilidade de Bühlmann é dado por

$$\zeta = \frac{T}{T + \frac{E[Var(Y|\alpha)]}{Var[E(Y|\alpha)]}},$$

em que T é o número de observações dentro de um grupo, ou repetições. Esse estimador é linear de mínima variância não viciado. Por depender apenas das funções esperança e variância, não precisa do pressuposto de normalidade das observações e pode facilmente ser estendido para outras distribuições (BÜHLMANN, 1967). Porém, como obter informações sobre α?

Ao expressar este problema de credibilidade em termos de um esquema de amostra baseado em regressão, podemos utilizar técnicas de regressão para estimar parâmetros e predizer quantidades desconhecidas. A forma usual de se predizer quantidades desconhecidas pode ser por meio do uso de efeitos aleatórios no modelo (BRAZAUSKAS; DORNHEIM; RATNAM, 2013), conforme será apresentado na Seção 2.2.2. Olhar para o fator de credibilidade de Bühlmann na forma de um efeito aleatório implicará que

$$\zeta = \frac{T}{T + \frac{\hat{\sigma}^2}{\hat{\sigma}^2_\alpha}},$$

em que a razão $\frac{\hat{\sigma}^2}{\hat{\sigma}^2_\alpha}$ é associada ao chamado estimador de encolhimento. Observe que se $\zeta \to 1$ significa que ou $T \to \infty$ ou $\frac{\hat{\sigma}^2}{\hat{\sigma}^2_\alpha} \to 0$. Isto é, se o número de observações dentro do grupo é muito grande, ou se a variabilidade dentro do grupo se torna muito superior à

variabilidade da variável resposta, em termos atuariais, temos que a informação do grupo, no caso longitudinal, a experiência do individuo, se torna mais crível (FREES, 2009).

A credibilidade de Bühlmann, da forma como foi apresentada até o momento, requer que todos os grupos tenham o mesmo número de repetições. Por isso, este fator de credibilidade também aparece pelo nome de modelos de Bühlmann balanceado. Existem outros fatores de credibilidade que foram se desenvolvendo desde então, para lidar com dados mais realistas. A credibilidade de Bühlmann-Straub é uma generalização que permite incluir um coeficiente de peso, para diferentes contagens de repetições ou exposições da apólice (BÜHLMANN; STRAUB; BROOKS, 1970). Outros modelos de credibilidade podem ser encontrados em Antonio e Beirlant (2007), Hachemeister (1975) e Dutang, Goulet e Pigeon (2008).

Nelder e Verrall (1997) mencionam que todos os modelos de credibilidade tradicionalmente utilizados são equivalentes a casos particulares de modelos completamente aleatórios, em que não há efeitos fixos (efeitos das classes de risco). Por isso, invariavelmente, os modelos de credibilidade podem ser analisados dentro de uma estrutura estatística mais ampla e com maior potencial de análises, envolvendo simultaneamente os efeitos fixos e efeitos aleatórios, por meio de modelos generalizados ou semiparamétricos, por exemplo, que ainda proporcionam ferramentas de diagnósticos para verificação de pressuposições.

2.1.5 O mercado segurador espanhol

Quando se analisa dados de seguros é importante também entender sobre a região que está sendo estudada, uma vez que diferentes países têm legislações e estruturas populacionais diferenciadas e características que muitas vezes são particulares, que contribuem para entender alguns fenômenos econômicos do mercado segurador. No presente trabalho são analisados dados de seguros de automóveis de uma sociedade seguradora espanhola, cujos maiores detalhes serão apresentados na Seção 3.1.

Em relação à expectativa de vida da sua população, a Espanha é um dos países com maior expectativa no nascimento, que, em 2021, era cerca de 83 anos de idade. García, García e Rodríguez (2017) ressaltam, por exemplo, que em primeiro de janeiro de 2016 haviam mais de 8,5 milhões de pessoas com 65 anos ou mais, cerca de 18,4% da população, sendo esperado por algumas projeções demográficas que em 2066 esta

porcentagem alcance o patamar de 34% da população. Cabe ressaltar que no período do início de 2010 ao final de 2015, a população da Espanha se manteve em torno de 46,5 milhões de pessoas. Atualmente o país possui cerca de 47,4 milhões de habitantes (INE, 2022).

Quando observada a composição de sexo da população espanhola, segundo dados do Banco Mundial, entre 2010 e 2021, 49% da população são do sexo masculino e 51% do feminino. Entre 2010 e 2015, a população urbana passou de cerca de 36,4 para 36,9 milhões de pessoas, sendo que, atualmente, esse contingente já ultrapassou 38,3 milhões de pessoas. Por outro lado, a população rural reduziu neste mesmo período, de 10 milhões para 9,4 milhões de pessoas. Atualmente, este conjunto compreende aproximadamente 8,9 milhões de pessoas (BANCO MUNDIAL, 2022).

No que diz respeito ao mercado automotivo espanhol, segundo o relatório divulgado pela ACEA (2022), o país possuía, em 2016, 23,3 milhões de veículos de passageiros, alcançando os 25 milhões no ano de 2020. Quando considerado o total de veículos motorizados, o mesmo aponta aumento de 27,3 milhões em 2016 para 29,7 milhões em 2020, configurando o quinto maior contingente quando considerados os países que compõem a União Europeia. Este relatório ainda traz outra informação importante em relação aos anos dos carros de passageiros. Ele revela que, em 2020, a média de idade dos veículos era de 13 anos, tal que mais de 16 milhões destes possuem mais de 10 anos de idade, o que é considerado bastante envelhecido. Em relação à taxa de motorização, dada pela relação entre a quantidade de veículos e a população, há um aumento de 502, em 2016, para 532 veículos por mil habitantes em 2020, ou seja, mais de um automóvel para cada duas pessoas (ACEA, 2022).

2.2 Modelos de regressão

Neste tópico será apresentada uma introdução sobre os modelos de regressão, apresentando alguns dos principais desenvolvimentos ao longo dos anos. Enfatiza-se que esta seção não tem por objetivo esgotar toda a teoria e história sobre essa classe de modelos, que é muito extensa.

2.2.1 Modelo de regressão linear

As técnicas de análise de regressão surgem com o advento dos modelos lineares e do método de mínimos quadrados, tendo como objetivo central estabelecer um relacionamento quantificável entre variáveis, sendo os modelos compostos, principalmente, de efeitos fixos e efeitos aleatórios. O termo *regressão* se originou do fato de que as observações tendem a regredir para um valor médio, e foi primeiramente definido por Sir Francis Galton por volta de 1870, ao observar este comportamento quando estudava fenômenos biométricos hereditários (RODGERS; NICEWANDER, 1988).

O modelo de regressão linear (MRL), ou modelo de Gauss-Markov, é um dos mais simples e que ainda é bastante utilizado, podendo ser representado na forma matricial como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

em que $\boldsymbol{Y}$ é o vetor de valores observados da variável resposta, $\boldsymbol{X}$ é a matriz de valores da(s) variável(is) explicativa(s), também chamada de matriz de incidência dos efeitos fixos, $\boldsymbol{\beta}$ é o vetor de parâmetros e $\boldsymbol{\epsilon} \sim N(\mathbf{0}; \boldsymbol{I}\sigma^2)$ é o vetor dos erros (RENCHER; SCHAALJE, 2008). Nessa definição tem-se que $\boldsymbol{\beta}$ é o vetor de efeitos fixos a serem estimados e os erros, $\boldsymbol{\epsilon}$, uma fonte de variação aleatória, sendo, nessa classe, o único efeito aleatório do modelo. Quando o modelo apresenta apenas os erros como efeito aleatório, ele é denominado como modelo fixo.

Também é usual descrever o modelo por meio da esperança de $\boldsymbol{Y}$, que é dada por

$$\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta},$$

já que a média dos erros é pressuposta ser nula. Neste caso, $\boldsymbol{\mu}$ refere-se aos valores esperados da variável resposta. O processo de estimação para esta classe de modelos é, usualmente, feito pelo método de mínimos quadrados (CHARNET et al., 2008).

Pelo fato dos erros serem pressupostos $\boldsymbol{\epsilon} \sim N(\mathbf{0}; \boldsymbol{I}\sigma^2)$, esta classe apresenta três principais restrições: a normalidade para distribuição da variável resposta como consequência dessa pressuposição; a necessidade da independência das observações; e homoscedasticidade, que invariavelmente limitam o leque de possibilidades para análises. Por volta do início do século XX, já se discutia que esta técnica precisava de mais flexibi-

lidade e, desde então, várias extensões foram surgindo para proporcionar modelos mais adequados para determinadas situações. Para contornar este problema, algumas alternativas foram criadas na época, como, por exemplo, a aplicação de transformações lineares diretamente na variável resposta em busca de normalizá-la e/ou estabilizar a variância. Porém, nem sempre esse método pode ser aplicado, podendo ainda comprometer a interpretação do modelo final. Posteriormente, também, outras alternativas foram pensadas, sendo algumas delas apresentadas nas próximas seções.

2.2.2 Efeitos aleatórios

Ronald A. Fisher pode ser considerado um dos primeiros estudiosos a consolidar uma teoria que estendia os modelos de regressão, quando apresentou os modelos mistos (FISHER, 1918), cuja nomenclatura advém do fato de existirem tanto efeitos fixos como aleatórios – além do erro – no modelo. Esta generalização permitiu adicionar novas fontes de variação ao mesmo, e é recomendada para situações em que o efeito populacional não é observável (SINGER; NOBRE; ROCHA, 2018) ou quando há correlação entre observações, como características de um mesmo grupo de indivíduos, dados longitudinais, experimentos em blocos, dentre outros. Além da nomenclatura *modelo misto*, esta classe também aparece na literatura pelos nomes de modelos para medidas repetidas, dados em painel, modelos hierárquicos e modelos multinível (DEMIDENKO, 2013).

Os modelos mistos podem ser definidos, matricialmente, como

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon} \tag{2.2}$$

com esperança da variável resposta dada por

$$\boldsymbol{\mu} = E(\boldsymbol{Y}|\boldsymbol{\gamma}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \tag{2.3}$$

em que se tem os mesmos componentes do modelo de regressão linear, apresentado na Equação 2.1 com a adição de $\boldsymbol{Z}$, que é a matriz de incidência dos efeitos aleatórios, $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$ que representa o vetor de efeitos aleatórios e $\boldsymbol{\Psi}$ é a matriz de variâncias e covariâncias que pode assumir diferentes estruturas (MCCULLOCH; SEARLE, 2001). Os efeitos fixos são utilizados para modelar a esperança do parâmetro, enquanto que os

efeitos aleatórios alteram a estrutura da matriz de variâncias e covariâncias, introduzindo mais de uma fonte de variação nos dados (CAMARINHA FILHO, 2002).

A seleção da estrutura da matriz de variâncias e covariâncias mais apropriada irá depender do tipo de análise desejada e do próprio conjunto de dados. Uma estrutura que pode ser citada, a título de exemplificação, é de uma matriz simétrica positiva definida, que pode ser reduzidamente representada por

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21}^2 & \sigma_{31}^2 \\ \sigma_{21}^2 & \sigma_2^2 & \sigma_{32}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_3^2 \end{bmatrix}.$$

Também é comum o uso de matrizes diagonais, bloco diagonais, simétricas compostas e identidade (PINHEIRO; BATES, 2006).

Para estimação dos parâmetros e predição dos efeitos aleatórios, geralmente, são utilizadas as equações de Henderson (1973), ou também chamadas de equações de modelos mistos (MME), que advêm da maximização da distribuição conjunta de $\boldsymbol{Y}$ ($f(\boldsymbol{Y},\boldsymbol{\gamma}) = f(\boldsymbol{Y}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})$). Quando a distribuição dos erros é normal, a solução das MME levam aos *best linear unbiased estimators* (BLUEs) e aos *best linear unbiased predictors* (BLUPs) dos vetores $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, respectivamente (HENDERSON, 1973).

Os BLUE e os BLUP são semelhantes em relação à sua função. O termo *estimar* é utilizado para os efeitos fixos e o termo *predizer* geralmente está associado aos efeitos aleatórios, que compõem uma amostra de uma população que se comporta conforme uma distribuição de probabilidade. Apesar de semelhantes, as equações possuem suas diferenças. Estes estimadores e preditores têm uma importância vital no que tange à análise de modelos mistos, principalmente os BLUPs, não só por suas excelentes propriedades, mas porque revolucionaram diversas teorias dentro e fora das ciências estatísticas. Historicamente eles foram amplamente utilizados para melhoramento genético e também dentro da estrutura de teoria de credibilidade em seguros (ROBINSON, 1991).

Para estimar os componentes de variância, habitualmente, é utilizado o método da máxima verossimilhança restrita (REML) (ROBINSON, 1991) que se trata de uma aproximação em que apenas a porção da verossimilhança que não depende de efeitos fixos é maximizada, foi proposta por Patterson e Thompson (1971) de forma a remover o viés nas estimativas de componentes de variância, gerado no método tradicional. A

estimação por este método leva em consideração a perda de graus de liberdade resultante da estimação dos efeitos fixos (HARVILLE, 1977) e não possui fórmula fechada.

Para encontrar os valores estimados para os efeitos fixos e preditos para os efeitos aleatórios, $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$, métodos iterativos são necessários, sendo comum a utilização do algoritmo *expectation-maximization* (EM) (DEMPSTER; LAIRD; RUBIN, 1977), em que os componentes de variância são tratados como parâmetros de perturbação não observados (variáveis latentes) (LINDSTROM; BATES, 1988). Casella e Berger (2002) afirmam que o algoritmo EM convergirá também para o estimador de máxima verossimilhança. O método consiste na alternância entre uma etapa de esperança, em que é criada uma função para a verossimilhança com base nos parâmetros atuais; e uma etapa de maximização, em que são encontrados novos valores de parâmetros que maximizem a função dada na etapa anterior, gerando novas estimativas para os parâmetros para a etapa de determinação da esperança matemática (DEMPSTER; LAIRD; RUBIN, 1977).

Um caso particular dessa classe de modelos ocorre quando só há efeitos aleatórios para explicar o valor esperado da variável resposta, apresentando como efeito fixo apenas a média geral. Nesse caso estaremos utilizando os chamados modelos completamente aleatórios ($\boldsymbol{\mu} = E(\boldsymbol{Y}|\boldsymbol{\gamma}) = \boldsymbol{Z}\boldsymbol{\gamma}$) que mencionamos na Seção 2.1.4 sobre teoria de credibilidade atuarial.

2.2.3 Funções de ligação

Ainda ao final do século XX ocorreu outro grande avanço na metodologia de modelos de regressão com o surgimento dos modelos lineares generalizados (GLM), propostos por Nelder e Wedderburn (1972). Esta teoria, que unificou uma série de modelos pré-existentes, como a regressão log-linear e a regressão logística, surgiu com o objetivo de flexibilizar, por meio do uso de uma função, a relação entre o parâmetro – frequentemente a média – e seu preditor linear, permitindo o uso de outras distribuições que não somente a normal para a resposta. A função de ligação é utilizada para garantir que as estimativas para a média, por exemplo, estejam dentro da sua amplitude de valores possíveis e, em alguns casos, garantir boas propriedade estatísticas, como é o caso das funções de ligação canônicas (DEMÉTRIO, 2001).

Os GLM podem ser definidos, matricialmente, por

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta},$$

em que $\boldsymbol{\eta}$ é o preditor linear, $\boldsymbol{\mu} = E(\boldsymbol{Y})$ e $g(\cdot)$ é uma função de ligação (NELDER; WEDDERBURN, 1972). O processo de estimação dessa classe de modelos é, frequentemente, feito também pelo método da máxima verossimilhança. Em GLM, a distribuição a ser utilizada para a variável resposta precisa pertencer à família exponencial, para garantir boas propriedades teóricas e simplicidade dos cálculos. Uma distribuição pertence à família exponencial se sua função de (densidade) probabilidade (f.d.p.) puder ser escrita na forma $f(y;\theta,\phi) = exp\left(\frac{1}{a(\phi)}[y\theta - b(\theta)] + c(y;\phi)\right)\mathbb{I}_A(y)$ (MCCULLAGH; NELDER, 1989). Atualmente, com o avanço de processamento computacional e ferramental estatístico, isto não é mais, necessariamente, uma restrição, como abordaremos nas seções subsequentes.

A Tabela 2.1 apresenta, matematicamente, algumas das principais funções de ligação utilizadas para modelos generalizados. A função de ligação identidade não altera a relação da média e seu preditor. A ligação logarítmica é bastante utilizada para modelos de contagem, com variável resposta discreta positiva, como modelo Poisson. A logit, probit e complemento log-log para modelos de proporção, como um modelo binomial. E as inversas para variáveis contínuas estritamente positivas, por exemplo, modelos gama ou normal inversa.

Tabela 2.1 – Principais funções de ligação.

Nome	**Função de ligação**
Identidade	$\eta = \mu$
Logarítmica	$\eta = ln(\mu)$
Logit	$\eta = ln\left(\frac{\mu}{1-\mu}\right)$
Inversa	$\eta = \frac{1}{\mu}$
Inversa quadrática	$\eta = \frac{1}{\mu^2}$
Raiz quadrática	$\eta = \sqrt{\mu}$
Probit	$\eta = \phi^{-1}(\mu)$
Log-Log	$\eta = -ln[-ln(\mu)]$
Complemento Log-Log	$\eta = ln[-ln(1-\mu)]$

Fonte: Adaptado de McCullagh e Nelder (1989)

É possível utilizar qualquer outra função, desde que ela seja contínua, diferenciável, monótona e que garanta a amplitude de valores possíveis para o parâmetro. Também é desejável que proporcione interpretações simples para os coeficientes (DEMÉTRIO, 2001).

Em 1993, Breslow e Clayton (1993) propuseram modelos com função de ligação e efeitos aleatórios, dando origem aos modelos lineares generalizados mistos (GLMM). Para todas a classes de modelos de regressão apresentadas até o momento, a relação entre o valor esperado da variável resposta, ou uma função dele, e suas variáveis explicativas, é linear. Para determinados tipos de dados, considerar uma relação linear pode não ser o mais apropriado, se na prática o relacionamento se mostrar outro. Forçar este relacionamento pode, por consequência, prejudicar a obtenção de um bom ajuste, o que levará a inferências incorretas, impactando a tomada de decisões.

2.2.4 Suavizadores e termos não paramétricos

Concomitantemente aos GLMM, estavam também sendo desenvolvidos os modelos aditivos generalizados (GAM), propostos por Hastie e Tibshirani (1990), que consistem na adição de termos não paramétricos ao preditor linear, originando os modelos semi-paramétricos. O objetivo é descrever de maneira mais apropriada o relacionamento das variáveis explicativas com o parâmetro através da utilização de uma função de suavização, permitindo, assim, que os próprios valores das observações conduzam a relação com o parâmetro. Ademais, essa extensão permite, em determinados casos, corrigir eventuais problemas de assimetria e heterocedasticidade dos erros.

Esta classe de modelos pode ser representada por

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + s_1(\boldsymbol{x}_1) + \ldots + s_J(\boldsymbol{x}_J),$$

em que $s_j(\cdot)$, $j = 1, \ldots, J$, é uma função de suavização não paramétrica aplicada à variável $\boldsymbol{x}_j$ (RUPPERT; WAND; CARROLL, 2003).

As técnicas de suavização podem ser utilizadas em diversos contextos, como na realização de análises exploratórias, construção de modelos e verificação da qualidade de ajuste, por exemplo. Elas visam capturar padrões importantes nos dados, desconsiderando pequenas variações, atribuídas a ruídos, proporcionando análises mais flexíveis e robustas. E não demandam nenhuma pressuposição, como ocorre em análises paramétricas, sendo baseadas quase que puramente em processamento computacional (WOOD; PYA; SÄFKEN, 2016).

O conceito de suavização não deve ser confundido com ajuste de modelos. No ajuste estamos interessados em uma fórmula explícita e interpretável dos parâmetros, enquanto

que, na suavização, em geral, não há uma fórmula fechada funcional. Na suavização, a ideia central é desconsiderar as pequenas variações e nos modelos, queremos a melhor curva possível para representar os dados. Ainda, nos modelos é comum existirem vários parâmetros para descrever o comportamento. Em contrapartida, na suavização, costuma existir apenas um parâmetro, que irá regular o grau de suavização, normalmente chamado de $\boldsymbol{\lambda}$ ou hiperparâmetro (SIMONOFF, 2012).

Existem diversos tipos de termos aditivos não paramétricos que podem ser utilizados. Alguns deles, cada um com sua própria utilidade para determinada situação, estão mencionados no Quadro 2.3.

Quadro 2.3 – Alguns exemplos de funções de suavização ou termos não paramétricos aditivos

Splines	**Filtros**	**Suavizador**	**Outros**
Cúbicos	Butterworth	Aditivo	Árvore de decisão
Cíclicos	Chebyshev	Exponencial	Polinômios fracionais
Monótonos	Digital	Kernel	Regressão local (loess)
Encolhidos em zero	Elíptico	Laplaciano	Média móvel
Coeficientes variantes	Kalman		Redes neurais
Penalizados	Kolmogorov–Zurbenko		

Fonte: Adaptado de Stasinopoulos et al. (2017) e Simonoff (2012)

A verificação da necessidade de uma função de suavização, enquanto modificador da relação da variável com o parâmetro, pode ser feita por meio de gráficos de dispersão da variável resposta contra a variável explicativa. Se o comportamento observado for linear, então, consequentemente, a relação é linear e não há necessidade de suavizar, caso contrário, suaviza-se. O processo de estimação dessa classe de modelos é, em geral, dado por penalizações no método de mínimos quadrados ou no método de máxima verossimilhança, sendo o algoritmo *backfitting* o mais comumente utilizado (BREIMAN; FRIEDMAN, 1985). Após o ajuste, também é possível obter alguns diagnósticos para verificar a adequação da suavização, que pode ser feito visualmente por meio de gráficos dos termos de regressão e numericamente por meio da análise dos graus de liberdade efetivos do suavizador com o parâmetro ajustado (STASINOPOULOS et al., 2017). Em especial, destacam-se os suavizadores penalizados porque talvez sejam os suavizadores mais importantes da família de funções de suavização, devido à sua flexibilidade e ao fato de poderem ser escritos na estrutura de um modelo misto (STASINOPOULOS et al., 2017).

2.2.4.1 Suavizadores penalizados e efeitos aleatórios

É possível demonstrar que grande parte dos suavizadores penalizados podem ser escritos na forma de um efeito aleatório. Suponha um modelo completamente aleatório representado por

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

que pressupõe duas fontes de variação, a dos erros e a dos efeitos aleatórios, em que $\boldsymbol{y}$ é a variável de interesse, $\boldsymbol{Z}$ é uma matriz de incidência, ou de base, $\boldsymbol{\gamma}$ é o vetor de efeitos aleatórios, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{W}^{-1})$, $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \sigma_b^2 \boldsymbol{G}^{-1})$, $\boldsymbol{W}$ é uma matriz de pesos e $\boldsymbol{G}$ é uma matriz de penalidades (STASINOPOULOS et al., 2017). A obtenção de um suavizador penalizado acontece por meio da solução do seguinte problema de mínimos quadrados em relação a $\boldsymbol{\gamma}$, em que uma penalidade quadrática é aplicada à equação

$$Q = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})^\top \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \boldsymbol{\lambda}\boldsymbol{\gamma}^\top \boldsymbol{G}\boldsymbol{\gamma}, \tag{2.4}$$

λ é o parâmetro que regula o grau de suavização. A solução é dada por

$$\hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\lambda}\boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{y},$$

e os valores ajustados por

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\lambda}\boldsymbol{G})^{-1} \boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{y}, \\ \hat{\boldsymbol{y}} &= \boldsymbol{S}\boldsymbol{y}, \end{aligned} \tag{2.5}$$

em que $\boldsymbol{S}$ é chamada de matriz de suavização. O grau de liberdade efetivo é obtido ao se calcular o traço de $\boldsymbol{S}$. A matriz de penalidades $\boldsymbol{G}$ também é de grande interesse. Em geral ela é dada por $\boldsymbol{G} = \boldsymbol{D}_k^\top \boldsymbol{D}_k$ em que $\boldsymbol{D}$ é uma matriz de diferenças de alguma ordem (STASINOPOULOS et al., 2017). As matrizes de diferenças são utilizadas para fazer subtração de termos consecutivos em um vetor.

As matrizes $\boldsymbol{W}$ e $\boldsymbol{G}$ são conhecidas, entretanto, é preciso estimar σ_e^2 e σ_b^2. Ao considerar o hiperparâmetro como $\boldsymbol{\lambda} = \sigma_e^2/\sigma_b^2$, os métodos de estimação por máxima verossimilhança hierárquica coincidirá com o método dos mínimos quadrados penalizados. Além disso, também é possível utilizar o método da máxima verossimilhança restrita para

estimar σ_e^2 e σ_b^2 e, consequentemente, obter o hiperparâmetro $\boldsymbol{\lambda}$ (STASINOPOULOS et al., 2017). O parâmetro de suavização $\boldsymbol{\lambda}$ pode ser visto como de encolhimento. Este parâmetro também é utilizado em modelos de credibilidade, como foi apresentado na Seção 2.1.4.

Próximo ao fim do milênio, surgem os modelos aditivos generalizados mistos (GAMM), que possibilitaram a utilização de modelos com funções de suavização, funções de ligação e efeitos aleatórios (LIN; ZHANG, 1999). Apesar dos notáveis progressos, essas técnicas ainda são limitadas por dois principais motivos. O primeiro, é que elas modelam apenas o parâmetro de média da distribuição da variável resposta e, o segundo, é que ainda não seria possível utilizar uma distribuição que não pudesse ser escrita na forma da família exponencial. Para várias situações, como por exemplo, variáveis com excesso de zeros, pode ser necessário modelar outros parâmetros da distribuição. A depender da complexidade dos dados, também é imprescindível utilizar distribuições mais complexas, que se adequem a dados extremamente assimétricos ou com problemas de curtose, por exemplo.

2.2.5 Modelos aditivos generalizados para locação, escala e forma

Rigby e Stasinopoulos (2005) propuseram os modelos aditivos generalizados para locação, escala e forma (GAMLSS), que também aparecem na literatura como *distributional regression models* (HELLER; ROBLEDO; MARSCHNER, 2022), que permitem este nível de flexibilização: todos os parâmetros da distribuição podem ter seu respectivo preditor em função das variáveis explicativas, funções não paramétricas de suavização, efeitos aleatórios, ou outros termos aditivos. Por isso são considerados uma classe de modelos de regressão além da média (KNEIB, 2013). Os GAMLSS possuem todos os modelos supracitados como casos particulares e ainda se enquadram em uma nova classe de modelos.

Esta classe pode ser definida por

$$\boldsymbol{\eta}_k = g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k + s_{k1}(\boldsymbol{x}_{k1}) + \ldots + s_{kJ_k}(\boldsymbol{x}_{kJ_k}), \tag{2.6}$$

sendo $\boldsymbol{\theta}_k$ o vetor de parâmetros a serem modelados e $k = 1, \ldots, p$ com p sendo o número de parâmetros da distribuição a ser utilizada (RIGBY; STASINOPOULOS, 2005). Para melhor visualização dos diversos preditores que podem existir nessa classe de modelos,

podemos tomar como exemplo um modelo cuja distribuição da variável resposta possua quatro parâmetros ($k = 1, 2, 3, 4$). Então, o modelo (2.6) pode ser reescrito de forma expandida como

$$\begin{aligned}
\boldsymbol{Y} &\overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + s_{11}(\boldsymbol{x}_{11}) + \ldots + s_{1J_1}(\boldsymbol{x}_{1J_1}) \\
\boldsymbol{\eta}_2 = g_2(\boldsymbol{\sigma}) &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + s_{21}(\boldsymbol{x}_{21}) + \ldots + s_{2J_2}(\boldsymbol{x}_{2J_2}) \\
\boldsymbol{\eta}_3 = g_3(\boldsymbol{\nu}) &= \boldsymbol{X}_3\boldsymbol{\beta}_3 + s_{31}(\boldsymbol{x}_{31}) + \ldots + s_{3J_3}(\boldsymbol{x}_{3J_3}) \\
\boldsymbol{\eta}_4 = g_4(\boldsymbol{\tau}) &= \boldsymbol{X}_4\boldsymbol{\beta}_4 + s_{41}(\boldsymbol{x}_{41}) + \ldots + s_{4J_4}(\boldsymbol{x}_{4J_4}),
\end{aligned}$$

em que $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ é uma distribuição com quatro parâmetros, $\boldsymbol{\mu}$ normalmente é um parâmetro de locação, $\boldsymbol{\sigma}$ é frequentemente um parâmetro de escala, e $\boldsymbol{\nu}$ e $\boldsymbol{\tau}$ são, em geral, os parâmetros de forma da distribuição. As matrizes $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ e $\boldsymbol{X}_4$ em geral, são diferentes, isto é, o preditor de cada parâmetro da distribuição pode receber diferentes variáveis explicativas. Esta classe permite o ajuste de modelos com $k = 1, 2, 3, \ldots, K$ preditores lineares (RIGBY; STASINOPOULOS, 2005).

Diferentes funções de ligação (Tabela 2.1) são utilizadas para garantir que o parâmetro seja estimado dentro de sua amplitude de valores. Se o parâmetro pode assumir qualquer valor real, então a ligação *identidade*, por exemplo, pode ser utilizada. Se pode assumir valores entre $(0, 1)$ então a ligação *logit* é recomendada. Se o parâmetro pode assumir valores entre $(0, \infty)$ então a ligação *logarítmica* é apropriada. Estas são as funções de ligação mais comuns em GAMLSS, entretanto, qualquer outra função de ligação pode ser utilizada, desde que adequada para os valores possíveis do parâmetro (STASINOPOULOS et al., 2017).

Em relação às distribuições possíveis em GAMLSS, o leque é, de fato, extenso. Existem inúmeras distribuições possíveis, dentre discretas, contínuas e de mistura, incluindo distribuições apropriadas para dados extremamente complexos em sua forma, que controlam assimetria e curtose. Ainda, podem-se construir versões truncadas ou censuradas das distribuições. É possível também criar distribuições de mistura, muito úteis para modelar dados que possuem inflação de algum valor, como zero, por exemplo, sendo alguns detalhes apresentados no Capítulo 3. Em relação às distribuições contínuas, estas também podem ser discretizadas (RIGBY et al., 2019). Os GAMLSS estão implementados no

software R (R Core Team, 2020), que trataremos na Seção 3.3, sendo também permitido a inclusão e criação de novas distribuições, cujo procedimento pode ser encontrado em Stasinopoulos et al. (2017) e Roquim et al. (2021).

As variáveis explicativas podem ser incluídas de diversas maneiras no modelo, seja de forma paramétrica linear ou não linear, por meio de efeitos aleatórios, ou ainda, de forma não paramétrica por meio do uso de suavizadores. Dentre os diversos termos aditivos que são possíveis incluir em GAMLSS, citam-se: árvores de decisão, *loess*, redes neurais e diversos tipos de *splines*, como *P-splines* (EILERS; MARX, 2010).

2.2.5.1 Estimação e inferência

Com relação à estimação e inferência em GAMLSS, é interessante utilizarmos o conceito de que grande parte das funções de suavização, em especial as penalizadas, podem ser escritas na forma de um efeito aleatório, como apresentado na Seção 2.2.4.1. Portanto, denotaremos a Equação (2.6) na forma

$$\boldsymbol{\eta}_k = g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.7}$$

que chamaremos de GAMLSS semiparamétrico, em que $\boldsymbol{s}_{jk}(\boldsymbol{x}) = \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$, com $J = 1, \ldots, J_k$ sendo a quantidade de termos não paramétricos ou suavizadores no modelo.

A estimação do modelo apresentado na Equação 2.7, em geral, é realizada por meio de métodos de máxima verossimilhança, em especial, por meio do logaritmo da função de verossimilhança penalizada, definido por

$$l_p = l - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^{\top}\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})\boldsymbol{\gamma}_{kj},$$

em que $l = \sum_{i=1}^{n} ln[f(y_i|\theta_i)]$ é o logaritmo da função de verossimilhança da amostra e o hiperparâmetro $\boldsymbol{\lambda}$ pressuposto constante (STASINOPOULOS et al., 2017).

Este método, como tantos outros, gera um sistemas de equações que não possui solução analítica, e, portanto, métodos iterativos devem ser utilizados. Na estrutura dos GAMLSS, destacam-se os métodos Cole e Green (CG) e Rigby e Stasinopoulos (RS). O primeiro, algoritmo CG, é uma generalização do algoritmo de Cole e Green (1992), que utiliza as primeiras derivadas e os valores exatos ou aproximados das derivadas de segunda ordem e derivadas cruzadas da função de verossimilhança dos dados. Entretanto, para

muitas distribuições, os parâmetros possuem informação ortogonal, ou seja, os valores das derivadas cruzadas são iguais a zero. Neste caso, utiliza-se o algoritmo RS, proposto por Rigby e Stasinopoulos (2005), que não utiliza as derivadas cruzadas. A principal diferença é que o algoritmo RS maximiza a função de verossimilhança em cada parâmetro por vez até atingir a convergência, enquanto que o algoritmo CG tem a capacidade de atualizar todos os parâmetros conjuntamente a cada iteração (STASINOPOULOS et al., 2017).

O valor para o hiperparâmetro pode ser obtido por meio de métodos baseados em otimização numérica, máxima verossimilhança internamente nos algoritmos CG ou RS, que Rigby e Stasinopoulos (2014) mostram coincidir com a quasi-verossimilhança penalizada, ou até pela própria relação com efeitos aleatórios, dado por $\boldsymbol{\lambda} = \sigma_e^2/\sigma_b^2$.

Para diferentes estruturas para $\boldsymbol{Z}$ e $\boldsymbol{\gamma}$, obtêm-se diferentes suavizadores ou, ainda, diferentes tipos de modelos mistos. Por isso, estaremos interessados nas estimativas de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$. Para tal, é pressuposto que a variável possua uma função (densidade) de probabilidade, $f(\boldsymbol{\gamma}|\boldsymbol{\lambda})$, em que $\boldsymbol{\lambda}$ é o vetor de hiperparâmetros a ser estimado e $\boldsymbol{\gamma}$ sendo um vetor de variáveis latentes ou variáveis não observadas na amostra, mas que existem no modelo e explicam a interdependência das observações (STASINOPOULOS et al., 2017).

Sob o modelo da Equação (2.7), dado os efeitos aleatórios $\boldsymbol{\gamma}$, a variável resposta $Y_1, Y_2, ..., Y_n$ são pressupostas independentes com função (densidade) de probabilidade $f(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma})$. Sob estas suposições, a função (densidade) de probabilidade marginal de $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^\top$ é dada por

$$f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) = \int_{\boldsymbol{\gamma}} f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) \quad f(\boldsymbol{\gamma}|\boldsymbol{\lambda}) \quad d\boldsymbol{\gamma},$$

em que, dado $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$, $f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda})$ denota a distribuição marginal de $\boldsymbol{Y}$, $f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$ é a distribuição condicional de $\boldsymbol{Y}$ dado $\boldsymbol{\gamma}$, e $f(\boldsymbol{\gamma}|\boldsymbol{\lambda})$ é a distribuição marginal dos efeitos aleatórios γ (STASINOPOULOS et al., 2017).

Devido à dificuldade de se encontrar as integrais, que geralmente não possuem solução explícita, também é difícil definir a distribuição marginal de $\boldsymbol{Y}$. Quando a distribuição condicional de $\boldsymbol{Y}$ e a marginal de $\boldsymbol{\gamma}$ são normais, consequentemente a marginal de $\boldsymbol{Y}$ também será normal, mas esta situação não ocorre para o caso dos GAMLSS, porque a condicional pode assumir qualquer distribuição. Algumas alternativas para encontrar ou aproximar as integrais são: algoritmo EM, aproximação de Laplace, cadeias de Markov de Monte Carlo (MCMC) ou quadratura gaussiana (STASINOPOULOS et al., 2017).

Stasinopoulos et al. (2017) mostram que, se o modelo está bem ajustado, então os estimadores de máxima verossimilhança serão assintoticamente consistentes, levando a inferências corretas. As inferências para a porção paramétrica geralmente se concentram em estimar os parâmetros, testar sua significância, obter intervalos de confiança e interpretar a contribuição de cada coeficiente ao modelo.

Para a porção semiparamétrica são necessárias ferramentas extras para realizar inferências. O formato das curvas ajustadas e seus referidos graus de liberdade são muito importantes neste momento. Para os suavizadores penalizados, os erros padrões das curvas ajustadas são função do grau de liberdade efetivo, que é dado pelo traço da matriz de suavização $\boldsymbol{S}$, definida na Equação 2.5. O valor de $\boldsymbol{\lambda}$ também é de suma importância, porque determinará o formato da curva de suavização. É raro o interesse em erros padrões desse parâmetro, entretanto, eles podem ser aproximados por técnicas de simulação, como *bootstrap* (STASINOPOULOS et al., 2017).

Também, há interesse em predizer o valor de qualquer parâmetro da distribuição da variável resposta, ou ainda obter a própria distribuição ajustada como um todo, que pode ser pela substituição de todos os parâmetros pelos valores estimados. Os valores esperados para a variável resposta normalmente são funções dos parâmetros e também podem ser obtidos da mesma maneira. É preciso ressaltar que o ato de substituir um parâmetro por uma estimativa deve ser interpretado com cautela, uma vez que não leva em consideração as possibilidades de erros de estimação.

Quando há o interesse em realizar predições, geralmente o modelo tem seu desempenho testado em um outro conjunto de dados, chamados de conjunto teste. Quando queremos saber a acurácia e precisão do modelo, podemos utilizar algumas medidas que comparam os valores esperados preditos pelo modelo com o que foi realmente observado no conjunto de dados teste.

Para além de obtermos boas inferências e interpretações, estaremos também interessados em testar diversos modelos para selecionar um que seja útil para determinado objetivo, seja explicativo (boas interpretações) ou preditivo. Dado a infinidade de modelos que a classe GAMLSS pode proporcionar, serão definidas algumas metodologias norteadoras do processo de seleção.

2.2.5.2 Seleção de modelos

As técnicas de seleção de modelos de regressão buscam resolver o problema de se selecionar preditores adequados dentre uma infinidade de possíveis preditores em potencial. A seleção de modelos em GAMLSS envolve a escolha da melhor distribuição para a variável resposta, dos preditores adequados para os parâmetros da distribuição selecionada, das funções de ligação e dos hiperparâmetros. Em geral, buscam-se modelos parcimoniosos, que não sejam sobreajustados, isto é, que possuam muitos parâmetros com interpretações complexas e de difícil generalização, e que não sejam subajustados, oferecendo um modelo pobre em adequação (RAMIRES et al., 2021a).

Com relação à função de ligação, a escolha é relativamente simples. Ela deve garantir que o parâmetro esteja em sua amplitude e é desejável possibilitar fáceis interpretações. Diferentes funções de ligação podem afetar consideravelmente o ajuste do modelo. A escolha do valor do hiperparâmetro foi discutido anteriormente e Rigby e Stasinopoulos (2014) propõem um algoritmo para a escolha do melhor valor automaticamente. O valor de hiperparâmetro também pode ser fixado de modo arbitrário pelo pesquisador. No que tange à seleção de distribuições e seleção de variáveis, é importante definir algumas medidas. Destaca-se o desvio global, dado por

$$GDEV = -2l(\hat{\boldsymbol{\theta}}),$$

em que $l(\hat{\boldsymbol{\theta}})$ é o logaritmo da função de verossimilhança ajustada (STASINOPOULOS et al., 2017). Esta quantidade é utilizada na definição do critério de Akaike generalizado (GAIC) (VOUDOURIS et al., 2012), dado por

$$GAIC(\kappa) = GDEV + (\kappa \times df),$$

em que df denota o total efetivo de graus de liberdade do modelo e κ é a penalidade para cada grau de liberdade utilizado. Se $\kappa = 2$, o critério coincide com o critério de Akaike (AIC) (AKAIKE, 1974). Se $\kappa = ln(n)$, o critério coincide com o critério de informação bayesiano (BIC)(SCHWARZ, 1978). O $GAIC(\kappa)$ penaliza modelos com muitos parâmetros, de forma que, para algum κ escolhido, quanto menor o valor de $GAIC(\kappa)$, mais parcimonioso é considerado o modelo (STASINOPOULOS et al., 2017). Ademais,

diferentes escolhas para κ podem ser realizadas, implicando na flexibilização de seleção das variáveis explanatórias, como apresentado por Ramires et al. (2021c).

A seleção da melhor distribuição pode ocorrer em dois estágios, o de ajuste e o de diagnóstico. O primeiro envolve a comparação de modelos ajustados utilizando diferentes distribuições, que usa o $GAIC(\kappa)$. O estágio de diagnóstico envolve a análise da adequação daquela distribuição. De nada adianta aquele modelo ter o menor $GAIC(\kappa)$, se a distribuição não se adequou bem aos dados. Na Seção 2.2.5.3 serão apresentados tais procedimentos.

A seleção de variáveis explicativas é, na prática, um dos assuntos mais importantes no ajuste do modelo estatístico. Usualmente, $\boldsymbol{X}_k$ conterá fatores e variáveis quantitativas que podem entrar no modelo de forma linear ou por meio de funções não paramétricas. Na literatura existem diversas formas de se selecionar variáveis. Enfatizaremos os métodos baseados em critérios, porque são a forma mais utilizada em seleção de variáveis em GAMLSS. Os procedimentos para seleção de variáveis, via critério de informação em GAMLSS, são bem semelhantes aos tradicionais da classe de regressão, podendo-se citar os métodos *forward*, *backward* e *stepwise*. A principal diferença é que não estamos selecionando variáveis para apenas um parâmetro, e sim, para vários. A inclusão ou não de uma variável, em um determinado parâmetro, afeta a disposição das variáveis para os demais parâmetros. Por isso, Rigby e Stasinopoulos (2005) propõem inúmeras estratégias de seleção, sendo uma delas melhor apresentada na Seção 3.3. A quantidade de variáveis disponíveis afeta consideravelmente o tempo de processamento para seleção, já que diversas combinações são testadas. Por este motivo, é fortemente recomendado que seja feita uma análise de colinearidade das variáveis ou utilizar alguma técnica multivariada para redução ou classificação do conjunto de dados, como componentes principais (STASINOPOULOS et al., 2022) ou, agrupamento (RAMIRES et al., 2021b), respectivamente.

Vários modelos podem ser adequados para explicar e predizer algum fenômeno. Não existe um *melhor modelo* único. As técnicas matemáticas são utilizadas para auxiliar no processo de seleção. Entretanto, o próprio pesquisador pode fixar determinadas disposições para o modelo, caso haja interesse. Isso significa que diferentes problemas podem requerer diferentes estratégias de seleção. Independente de qual seja o modelo final escolhido, este só não pode estar pobremente ajustado aos dados. Inadequações levam à não veracidade nas inferências e resultados. Os diagnósticos de adequação do modelo podem

ser feitos por meio de uma análise de resíduos, procedimento muito comum da classe de modelos de regressão.

2.2.5.3 Análise de resíduos

Os resíduos ordinários são definidos diretamente pela diferença entre os valores observados e estimados, frequentemente pressupostos segundo uma distribuição normal, com média nula e variância constante. Nessa forma mais simples, acabam sendo limitados no sentido de que são difíceis de se generalizar para outras distribuições de probabilidade. Neste contexto, Dunn e Smyth (1996) propuseram os resíduos quantílicos (aleatorizados) normalizados que são amplamente utilizados nos diagnósticos de GAMLSS e são definidos por

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i),$$

em que Φ^{-1} é o inverso da distribuição acumulada de uma normal padrão e $\hat{u}_i$ são os resíduo quantílicos (DUNN; SMYTH, 1996). Se a distribuição é contínua, então $u = F(y|\boldsymbol{\theta})$ e $\hat{u} = F(y|\hat{\boldsymbol{\theta}})$ são os valores da função na distribuição acumulada do modelo e do ajuste, respectivamente. Se o modelo está bem especificado, então u terá distribuição uniforme entre zero e um. Se a distribuição for discreta, ou de mistura do tipo discreta-contínua, então $F(y|\boldsymbol{\theta})$ tem probabilidades não nulas, de forma que se torna necessário definir u e $\hat{u}$ como um valor aleatório de uma distribuição uniforme, por isso o termo *aleatorizado* aparece entre parênteses (STASINOPOULOS et al., 2017). Para este último caso, o usual é realizar várias aleatorizações e avaliar o valor mediano para cada resíduo. A principal vantagem dos resíduos quantílicos (aleatorizados) normalizados é que, para qualquer que seja a distribuição da variável resposta, os resíduos sempre terão uma distribuição normal padrão quando o modelo assumido está adequado.

Uma das principais técnicas utilizadas para analisar visualmente estes resíduos em GAMLSS são os *worm plots*, traduzidos para o português como gráficos de minhoca (VAN BUUREN; FREDRIKS, 2001). Esse tipo de gráfico é uma ferramenta para verificar resíduos, considerando diferentes intervalos. O gráfico de minhoca é semelhante a um gráfico normal quantil-quantil (*Q-Q plot*), porém, sem a tendência crescente, sendo que o nome advém do formato que os pontos geralmente possui. Quanto mais perto da linha de origem os pontos estiverem, mais os resíduos se aproximam de uma distribuição normal

padrão. É uma excelente ferramenta uma vez que, a depender do formato e posição dos pontos, é possível diagnosticar os problemas de ajuste. Se mais de 5% dos pontos extrapolam os limites das bandas de confiança e:

a) o nível está localizado acima ou abaixo da reta de origem, é o indicativo de que o parâmetro de locação foi subestimado ou superestimado, respectivamente;

b) aparecem como uma reta, isto é, um comportamento linear, seja com inclinação positiva ou negativa, assinala que o parâmetro de escala foi subestimado ou superestimado, respectivamente;

c) assemelham a um formato de parábola, isto é, um comportamento quadrático, seja com concavidade voltada para cima ou para baixo, demonstra que o parâmetro de assimetria foi subestimado ou superestimado, respectivamente;

d) apresentam um comportamento cúbico, revela que certamente houve problema na estimação do parâmetro de curtose.

Da mesma forma que ocorre para os resíduos quantílicos (aleatorizados) normalizados, para os casos em que a distribuição é discreta, ou de mistura do tipo discreta-contínua, é necessário repetir o processo de aleatorização e gerar vários gráficos de minhoca, avaliando o comportamento dos pontos em todos eles.

2.2.5.4 Modelos mistos na estrutura GAMLSS

As possibilidades de uso de efeitos aleatórios em GAMLSS são extensas, incluindo os modelos apresentados na Seção 2.2.2, que trata dos modelos mistos, e também dos modelos não paramétricos, apresentados na Seção 2.2.4. Em comparação ao modelo mostrado na Equação 2.2, em GAMLSS há algumas diferenças, dada a infinidade de distribuições da variável resposta, com seus respectivos parâmetros e preditores. Existem diferentes possibilidades de onde incluir um efeito aleatório. Nesta subseção faremos apenas algumas distinções, para evitar o equívoco entre estas duas classes.

Para o caso dos modelos com suavizadores, estamos tratando dos GAMLSS semiparamétricos, definido na Equação (2.7), que chamaremos apenas de GAMLSS ao longo do texto. Quando houver o uso de efeitos aleatórios tradicionais, aqueles utilizados para lidar com dados agrupados, em que as observações individuais naturalmente se classificam

em grupos, ocasionando a correlação, denotaremos de GAMMLSS (*Generalized Additive Mixed Models for Location, Scale and Shape*), aproveitando a mesma terminologia já utilizada para este tipo de modelo. Note que em GAMMLSS também pode ou não conter termos não paramétricos suavizadores. A principal diferença está na presença ou ausência de um efeito aleatório, que seja o próprio efeito da variável.

Nas suavizações, em geral, não temos interesse em investigar o vetor de efeitos aleatórios, enquanto que nos GAMMLSS estaremos muitíssimos interessados nas predições dos efeitos aleatórios, e não só, mas também nas estimativas dos componentes de variância. Tanto para os GAMLSS semiparamétrico, quanto para GAMMLSS, a estimação pode ser feita por uma aproximação normal local na verossimilhança, conhecida como quasi-verossimilhança penalizada (PQL) (BRESLOW; CLAYTON, 1993) ou por máxima verossimilhança restrita. Os valores ajustados são encontrados por meio da função de verossimilhança conjunta, enquanto que, para inferências, é utilizada a função de verossimilhança condicional, dado os efeitos aleatórios (STASINOPOULOS et al., 2017).

Ao maximizar o logaritmo da função de verossimilhança penalizada, os estimadores gerados, $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$, são máximos a posteriori (MAP), máxima verossimilhança hierárquica e também, máxima verossimilhança penalizada (RIGBY; STASINOPOULOS, 2005). Também, os estimadores MAP são BLUP's (ROBINSON, 1991). Stasinopoulos et al. (2017), entretanto, destacam que este método gera estimativas aproximadas, sendo alguns dos problemas do método apontado por Nelder (2005).

As técnicas de seleção de variáveis e análise de resíduos são as mesmas dos GAMLSS, com exceção que para GAMMLSS, estaremos, também, interessados em avaliar a distribuição de $\boldsymbol{\gamma}$.

3 MATERIAIS E MÉTODOS

3.1 O conjunto de dados

O conjunto de dados é longitudinal e foi primeiramente analisado por Frees et al. (2021), em que os autores apresentaram um método para análise das desistências ao longo dos anos em uma carteira de seguros. Os dados se referem a clientes de uma empresa seguradora espanhola ao longo de cinco anos, entre início de 2010 ao fim de 2014, trazendo informações daqueles que possuem tanto seguros de automóveis quanto de imóveis concomitantemente. Além disso, o conjunto menciona apenas a apólices vendidas para pessoas físicas.

São 40.284 apólices acompanhadas, gerando um conjunto com 122.935 observações. Inicialmente foi realizada uma análise preliminar e uma filtragem dos dados. Primeiro, filtramos apenas as apólices referentes a seguro de automóveis que permaneceram na carteira durante todo o período de 5 anos. Também foram corrigidas as variáveis de tempo, que apareciam com valores constantes ao longo dos anos. Ao final, restaram 66.405 observações, de 13.281 apólices. O último ano foi separado para validação da capacidade preditiva do modelo.

O conjunto de dados possui informações sobre o segurado, características do veículo e do imóvel. Analisamos apenas as apólices e variáveis referentes aos seguros veiculares. Na Tabela 3.1 é possível visualizar a descrição de cada uma das variáveis que foram utilizadas na análise. Os dados originais podem ser obtidos em <https://www.sciencedirect.com/science/article/pii/S2352340921009148>.

3.2 Análises estatísticas

Em GAMLSS, existem dois tipos de efeitos aleatórios, não paramétricos e normais multivariados. Na presente pesquisa foi utilizado o efeito aleatório normal, tal qual apresentado na Seção 2.2.2, do tipo intercepto aleatório, ou seja, cada indivíduo teve seu próprio intercepto. A estrutura da matriz de variâncias e covariâncias utilizada foi do tipo pesos fixos, isto é, a variabilidade é considerada constante dentro das observações de um mesmo indivíduo.

Tabela 3.1 – Descritores das variáveis do conjunto de dados

Variável	**Valores**
Sexo	0 para masculino, 1 para feminino
Idade do cliente (anos)	20-98
Região de residência	0 para rural, 1 para urbano
Presença de segundo motorista	0 para não, 1 para sim
Método de pagamento	0 para mensal, 1 para anual
Fidelidade (anos)	5-47,71
Apólice (nominal)	Identificador do segurado
Idade do automóvel (anos)	0-35
Potência (hp)	5,4-560,00
Valor do sinistro (€)	0,00-27.841,14

Fonte: Da autora (2022)

As distribuições ajustadas em zero são baseadas no conceito de distribuições de mistura do tipo discreta-contínua, uma vez que compreendem um valor de $Y = 0$ com uma determinada probabilidade p_0 e $Y = Y_1$ com probabilidade $(1 - p_0)$, em que Y_1 tem uma distribuição contínua (STASINOPOULOS; ENEA; RIGBY, 2017). Essas distribuições são úteis para casos em que há um excesso de zeros em uma distribuição contínua, como o caso do valor de sinistros em um ano. A maioria das apólices não vão ter nenhum sinistro e por isso há uma alta probabilidade do valor do sinistro ser zero, mas para as apólices que acionaram o seguro a distribuição dos valores será definida na reta dos reais positivos (RIGBY et al., 2019).

As distribuições gama ajustada em zero (ZAGA) e normal inversa ajustada em zero (ZAIG) podem ser apropriadas quando a variável Y assume valores maiores ou iguais a zero, isto é, $[0, \infty)$. A função de probabilidade de mistura da distribuição gama ajustada em zero, denotada por ZAGA(μ, σ, ν), é dada por

$$f_Y(y|\mu,\sigma,\nu) = \begin{cases} \nu, & se \quad y = 0 \\ (1-\nu)\left[\frac{y^{1/\sigma^2-1}e^{-y/(\sigma^2\mu)}}{(\sigma^2\mu)^{1/\sigma^2}\Gamma(1/\sigma^2)}\right], & se \quad y > 0 \end{cases}$$

e a função de probabilidade de mistura da distribuição normal inversa ajustada em zero, denotada por ZAIG(μ, σ, ν), é dada por

$$f_Y(y|\mu,\sigma,\nu) = \begin{cases} \nu, & se \quad y = 0 \\ (1-\nu)\frac{1}{\sqrt{2\pi\sigma^2 y^3}}\exp\left[-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right], & se \quad y > 0 \end{cases}$$

em que $\mu > 0$ e $\sigma > 0$ são a média e um parâmetro de dispersão, respectivamente, e $0 < \nu < 1$ é a probabilidade do zero (RIGBY et al., 2019), para ambas as distribuições.

A média para ambas as distribuições é dada por $(1 - \nu)\mu$. As variâncias são iguais a $(1 - \nu)\mu^2(\sigma^2 + \nu)$ e $(1 - \nu)\mu^2(\nu + \mu\sigma^2)$, para ZAGA e ZAIG, respectivamente. Outras características da distribuição, como moda, assimetria, curtose, função geradora de momentos e função de distribuição acumulada podem ser encontrados em Rigby et al. (2019).

As funções de ligação utilizadas foram a logarítmica para μ e σ e logit para ν, ou seja, funções de ligação usuais que garantem que o parâmetro será estimado dentro de sua amplitude e proporcionam boas interpretações (Seção 2.2.3).

A função de suavização utilizada foi a *P-spline*. Uma *spline* é uma curva definida matematicamente por dois ou mais pontos de controle, com um suporte mínimo, em relação a determinados graus de liberdade, suavização e domínio (EILERS; MARX; DURBÁN, 2015). A *P-spline* se refere a *B-spline* penalizada, isto é, utiliza-se a representação *B-spline* em que os coeficientes são determinados parte pelos próprios dados e parte por um parâmetro de penalização, que força determinado grau de suavização, evitando o sobre ajuste (EILERS; MARX, 1996).

3.3 Análise via *Software* R

Todas as análises, resultados e gráficos apresentados no presente trabalho foram obtidos por meio da linguagem de programação R (R Core Team, 2020). A linguagem R é gratuita e amplamente utilizada para análises estatísticas no meio acadêmico e também corporativo. Todas as funções necessárias para ajuste dos GAMLSS e GAMMLSS estão implementadas neste ambiente por meio de uma série de pacotes e seus respectivos manuais, que estão disponíveis gratuitamente no próprio repositório do R. No Apêndice A deste trabalho, ou clicando aqui , é possível visualizar a rotina em R utilizada para as devidas consultas de interesse.

Foram utilizados os pacotes:

a) `readr` (WICKHAM; HESTER, 2021), que proporciona a leitura de arquivos CSV pela função `read_csv()`;

b) `gamlss` (RIGBY; STASINOPOULOS, 2005), que carrega automaticamente os pacotes `gamlss.data` (STASINOPOULOS; RIGBY; DE BASTIANI, 2021) e `gamlss.dist` (STASINOPOULOS; RIGBY, 2021), que proporcionam suporte de conjunto de dados e de distribuições em GAMLSS, respectivamente;

c) `lme4` (BATES et al., 2015), que possibilita o ajuste de modelos mistos;

d) `psych` (REVELLE, 2021), para uso da função `describe()` proporcionando mais informações que a tradicional função `summary()`;

e) `epiDisplay` (CHONGSUVIVATWONG, 2018), para uso da função `tab1()`, proporcionando melhor informação sobre frequências de categorias e melhores gráficos se comparado a função `table()`; e,

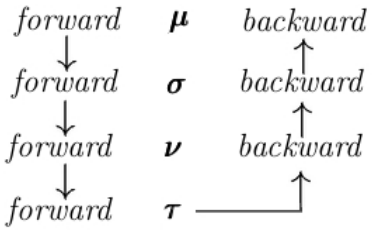
f) `tdr` (LAMIGUEIRO, 2018), para uso da função `tdStats()`, que proporciona a análise da qualidade de predições com base em um conjunto de dados teste, a partir de diversas medidas de acurácia e precisão.

Ainda, no que tange ao ajuste dos GAMLSS e GAMMLSS, utilizando o pacote `gamlss`, destacam-se algumas funções. Foi utilizado a função `fitDist()`, que ajusta modelos marginais para todas as distribuições implementadas em determinado tipo de variável e sua amplitude.

Para a seleção de variáveis foi utilizado a função `stepGAIC()` para modelos com apenas um preditor e `stepGAICAll.A()` para modelos com mais de um preditor. A seleção é baseada em procedimentos *stepwise* e feita através de algum determinado critério de parcimoniosidade, como AIC ou BIC (RAMIRES et al., 2021c). Para o caso da `stepGAICAll.A()` é utilizado a estratégia A que acontece da seguinte maneira. Primeiro, é ajustado um modelo para μ por meio do procedimento *forward* para algum GAIC selecionado, considerando os demais parâmetros constantes. Em seguida, é ajustado para σ, pelo mesmo procedimento, considerando o primeiro modelo para μ e demais constantes. Se houverem mais parâmetros, então ν é ajustado, considerando os modelos de μ e σ. Este procedimento ocorre até que o último parâmetro da distribuição seja modelado. A seleção de modelos para o último parâmetro da distribuição ocorre uma única vez, de forma que, após todos os parâmetros ajustados por *forward*, o algoritmo começa a retroceder, reajustando o penúltimo parâmetro, pelo procedimento *backward*, dado todos os

demais já ajustados, até retornar no reajuste de μ (NAKAMURA et al., 2017). Ao fim deste processo, o algoritmo para e os modelos para cada parâmetro compõem o GAMLSS final, considerado o mais adequado pelo critério escolhido. Para o presente trabalho, foi utilizado $\kappa = 2$ que leva ao critério de Akaike. Uma síntese desse processo pode ser visualizado na Figura 3.1.

Figura 3.1 – Esquema do processo de seleção de modelos via Estratégia A para um modelo com quatro parâmetros em GAMLSS



Fonte: Da autora (2022)

A análise da contribuição e efeito das funções de suavização foi feita por meio das funções `edfAll()` e `term.plot()`. A função `edfAll()` mostra os graus de liberdade efetivos de diferentes termos não paramétricos. Quando o valor é próximo de dois é um indicativo de que não é necessário utilizar uma função de suavização. A função `term.plot()` mostra o gráfico do relacionamento da variável com determinado parâmetro, importante para realizar as devidas interpretações.

Para verificar a adequação do modelo, para variáveis estritamente contínuas e sem efeitos aleatórios, utiliza-se a função `wp()` que gera os gráficos de minhoca. Quando a distribuição é discreta, de mistura do tipo discreta-contínua, ou possui efeitos aleatórios normais, utiliza-se a função `rqres.plot()`, que também gera gráficos de minhoca, com a diferença de que várias repetições são feitas para gerar os resíduos quantílicos aleatorizados, conforme mencionado na Seção 2.2.5.3. Esta última também fornece a mediana dos resíduos para todas as repetições.

Em GAMLSS também existem diversas formas de se inserir um efeito aleatório ao modelo. A forma utilizada na presente pesquisa foi por meio da função `re()` do pacote `gamlss` que faz a interligação com a função `lme()` do pacote `lme4` para gerar os modelos que denominamos de GAMMLSS. Para obter os valores estimados para os parâmetros, utilizou-se a função `predictAll()` e, com ela, calculou-se o valor esperado da variável resposta por meio da fórmula $(1 - \nu) \times \mu$, que é a média para as distribuições ZAGA e

ZAIG. A avaliação dos efeitos aleatórios pode ser feita por meio das funções `getSmo()` e `ranef()`.

No pacote `gamlss.demo` (STASINOPOULOS et al., 2015) é possível visualizar várias demonstrações do comportamento de distribuições e termos aditivos não paramétricos. O pacote `gamlss.inf` (ENEA et al., 2019) possibilita que qualquer distribuição no intervalo dos reais positivos seja zero-ajustada por meio da função `gamlssZadj()`, mas que ainda não funciona com a interface do pacote `lme4` para uso de efeitos aleatórios normais. Foram testadas outras distribuições para os valores positivos de sinistros, em especial, que modelassem também a assimetria e/ou curtose, como, por exemplo, a distribuição Box-Cox Cole e Green, Box-Cox exponencial potência, Box-Cox t e as versões generalizadas da gama e normal inversa. Todavia, não foi verificada melhora na adequação e ajuste em relação às distribuições gama e normal inversa tradicionais. Ainda, está em desenvolvimento o pacote `gamlss.rsm`, que ainda não foi oficialmente lançado, mas que proporciona ajuste de modelos do tipo contagem-contínuo, útil quando se deseja analisar determinado o valor e sua respectiva frequência de ocorrência, simultaneamente (HELLER; STASINOPOULOS; RIGBY, 2007).

4 RESULTADOS E DISCUSSÃO

4.1 Análise exploratória

A variável resposta analisada se refere ao valor do sinistro, que é a quantia indenizatória que a seguradora paga ao segurado. Essa variável é conhecida pela grande quantidade de valores nulos, chamados excesso de zeros, e presença de alguns valores extremos, sinistros muito caros, que também são características inerentes ao fenômeno e não podem ser desconsideradas da análise. A carteira, após a filtragem especificada na Seção 3.1 possui, de um total de 66.405 observações, apenas 573 ocorrências de sinistro, com 99,13% de valores nulos. O máximo desta variável é 27.841,14 €, com média global no valor de 12,56 € e com desvio padrão 256,33 €. O valor do sinistro é extremamente assimétrico à direita, com coeficiente de assimetria igual a 52,09 e leptocúrtico, com valor de coeficiente de curtose igual a 36.846,07. Na Tabela 4.1 é possível visualizar mais algumas estatísticas dos dados, separadas por ano. Não há grandes variações ano a ano, porém, destaca-se o ano de 2012, que teve a menor ocorrência de sinistros, ao passo que obteve o maior valor observado. O ano de 2013 foi o que apresentou o maior número de ocorrências dentro do período observado.

Tabela 4.1 – Estatísticas das apólices com sinistros

	2010	**2011**	**2012**	**2013**	**2014**	**Global**
Nº de ocorrências	108	109	99	133	124	573
Média dos sinistros (€)	1.253,68	1.648,82	1.471,51	1.664,95	1.222,13	1.455,12
Valor máximo (€)	14.698,86	19.794,65	27.841,14	13.837,72	9.009,20	27.841,14
% de zeros	99,19	99,18	99,25	98,99	99,06	99,13

Fonte: Da autora (2022)

Na Figura 4.1 é possível observar a distribuição de frequências do valor das ocorrências positivas. Por questões visuais, o gráfico está truncado entre valores maiores que zero até 5.000 €, sendo que 22 observações com valores extremos foram omitidas. Um resumo dos valores omitidos pode ser visualizado na Tabela 4.2. O pico observado ocorre no valor 882 €, e é referente ao valor do seguro *sem culpa*, que foi descrito brevemente na Seção 2.1.1. É uma categoria de seguros em que o valor pago pela seguradora é fixado previamente e o segurado não precisa comprovar a culpa no acidente. Por ser menos burocrático e mais rápido, essa categoria é muito utilizada pelos segurados. Para esta carteira, cerca

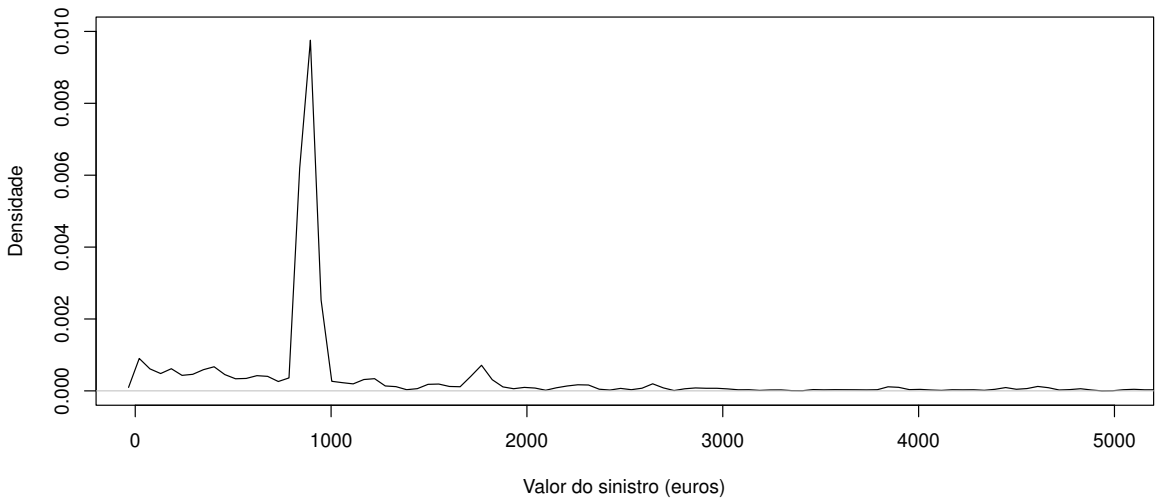
de metade dos segurados que tiveram acidente utilizaram esse tipo de cobertura. Sinistros que custam mais de 2.000 € ocorrem com baixa frequência. Entretanto, a seguradora precisa ter as devidas reservas para pagamento dessas indenizações, caso ocorram.

Tabela 4.2 – Estatísticas descritivas das observações omitidas (€)

Variável	**Mín.**	**Máx.**	**Média**	**Mediana**	**Desvio Padrão**
Valor do sinistro	5.064,74	27.841,14	10.664,63	7.860,66	5.817,50

Fonte: Da autora (2022)

Figura 4.1 – Distribuição de densidade das ocorrências



Fonte: Da autora (2022)

Em relação às variáveis explicativas contínuas é possível observar, na Tabela 4.3, algumas estatísticas descritivas, com medidas de locação, escala e forma. Esta carteira de apólices é bastante envelhecida, com idade média dos segurados igual a 63 anos, que é reflexo da própria característica populacional dos espanhóis, como apontado na Seção 2.1.5. A distribuição desta variável é simétrica. A idade do veículo é, em média, igual a 11 anos, também com distribuição simétrica, e similar ao valor apontado também na Seção 2.1.5. A potência média dos veículos segurados é de 111 cavalos, valor que equivale a potência do motor de um carro popular. A distribuição do valor de potência é assimétrico, contendo algumas observações discrepantes, podendo chegar a veículos com potência até 560 cavalos. A fidelidade dos clientes dessa seguradora é bastante alta, com média de 12 anos. Isto pode ocorrer por diversos motivos. Talvez a seguradora invista em manter

seus clientes, através de *marketing* e melhores preços, por exemplo, para poder fazer análises históricas e reduzir o risco na carteira, como apontou Arvidsson (2011). Talvez também seja consequência de um mercado de seguros estagnado, mas consolidado (Swiss Re Institute, 2022). A fidelidade observada chega a atingir até cerca de 50 anos.

Tabela 4.3 – Estatísticas descritivas das variáveis quantitativas

Variável	**Mín.**	**Máx.**	**Média**	**Mediana**	**Desvio Padrão**	**Assimetria**	**Curtose**
Idade do cliente	20,00	98,00	63,29	63,00	12,74	-0,06	-0,78
Idade do veículo	0,00	35,00	11,19	12,00	6,45	0,08	-0,36
Potência do veículo	5,40	560,00	111,35	105,00	44,16	1,94	8,28
Fidelidade	5,00	47,71	12,02	10,02	5,46	1,43	2,58

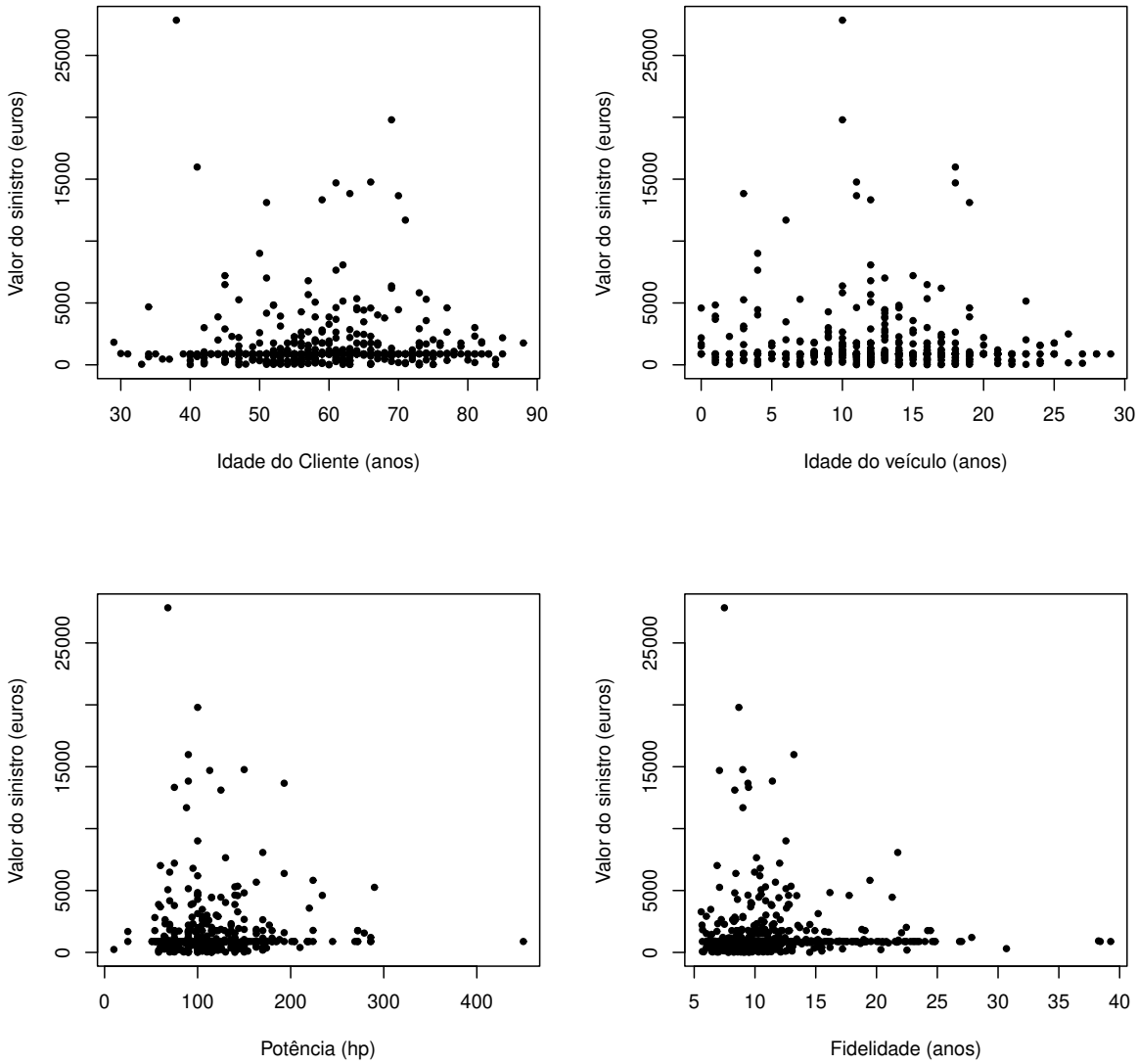
Fonte: Da autora (2022)

Na Figura 4.2 são apresentados gráficos de dispersão, da variável resposta contra as variáveis quantitativas para as ocorrências, possibilitando análise visual da relação entre as variáveis. Os acidentes mais caros ocorreram mais frequentemente entre clientes de 40 a 80 anos, amplitude bastante grande. Observaram-se ocorrências extremas desde veículos novos até cerca de 25 anos. Em ambos os gráficos que tratam de idade, é possível notar um ligeiro formato de sino nos pontos. A maioria do reclames de seguro parece ocorrer para veículos entre 50 a 250 cavalos de potência, com maior frequência no valor médio de 100 cavalos. O gráfico da fidelidade é bastante interessante. É possível concluir que a maioria das ocorrências ocorrem para novos clientes, com média próxima ao valor de 10 anos consecutivos. Após este período, observa-se uma tendência decrescente do valor do sinistro, tendendo a valores ínfimos para clientes com muitos anos na seguradora. Apesar de não observamos um relacionamento explícito entre as variáveis, isto não significa que não sejam significativas para explicar o comportamento do valor dos sinistros, uma vez que elas podem explicar não só o valor esperado, mas também a variância e formato das distribuição, ou, ainda, ter uma relação não linear.

Analisou-se, também, a correlação entre as covariáveis, que é apresentada na Tabela 4.4. Pela própria característica das variáveis já se esperava pouca ou nenhuma correlação, o que é confirmado pelos valores observados. Portanto, nenhum problema de colinearidade foi detectado e todas estas variáveis foram consideradas no processo de seleção de modelos com potencial explicativo da variável resposta.

Na Tabela 4.5 é apresentada a distribuição de frequência dos fatores. Esta carteira é predominante e excessivamente feminina, tendo quase 80% dos segurados nesta catego-

Figura 4.2 – Dispersão das covariáveis quantitativas contra a variável resposta



Fonte: Da autora (2022)

Tabela 4.4 – Correlação de Person para as covariáveis

	Idade do cliente	Idade do veículo	Potência	Fidelidade
Idade do cliente	1			
Idade do veículo	-0,0397	1		
Potência	-0.0912	0.0011	1	
Fidelidade	0.2416	-0.2129	-0.1357	1

Fonte: Da autora (2022)

ria. Há, talvez, dois motivos para isso. O primeiro, é que as mulheres, em geral, sentem maior necessidade de estarem asseguradas (BERGDAHL, 2005). O segundo, é pelo fato de os seguros, normalmente, serem mais baratos para esta classe, por apresentarem me-

nor risco, como apontado na Seção 2.1.2, gerando essa tendência. A divisão de sexos da população espanhola é aproximadamente equilibrada (BANCO MUNDIAL, 2022). Também, a distribuição da região é bastante discrepante, com cerca de 85% do veículos com trânsito prioritário em zonas rurais. Isso pode ser efeito da maior necessidade de se sentir seguro por parte do motorista, quando trafegando em rodovias e estradas, dada a maior severidade nos acidentes nessas regiões, causada pela alta velocidade (SHERAFATI et al., 2017). Com relação à presença de segundo motorista, observa-se que isto ocorre apenas para 12% das apólices, comportamento esperado que pode ser explicado pela própria tendência das pessoas terem seus veículos individuais para atenderem suas necessidades particulares, como é o caso da Espanha, dado que são um país extremamente motorizado, como mencionado na Seção 2.1.5. Por fim, destaca-se que o método de pagamento predominante é o anual.

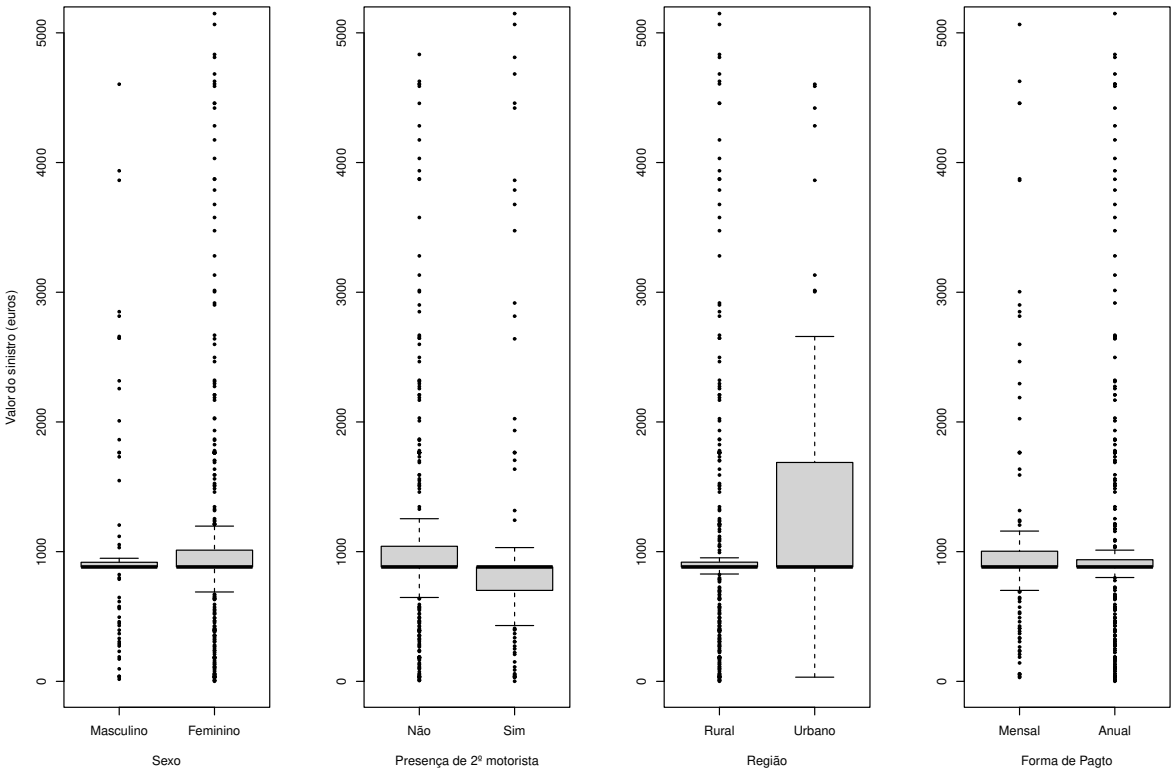
Tabela 4.5 – Tabela de distribuição de frequências dos fatores

Variável	**Nível**	**Frequência absoluta**	**%**
Sexo	Masculino	11.740	22,1
	Feminino	41.384	77,9
Região de residência	Rural	45.020	84,7
	Urbana	8.104	15,3
2º Motorista	Não	46.732	88,0
	Sim	6.392	12,0
Método de Pagto	Mensal	10.292	19,4
	Anual	42.832	80,6

Fonte: Da autora (2022)

Na Figura 4.3 são apresentados os gráficos de caixa da distribuição do valor do sinistro em cada nível dos fatores, restringido ao intervalo maior que 0 € até 5.000 €, para melhor visualização. É possível perceber uma maior dispersão do valor do sinistro entre as pessoas do sexo feminino em relação ao masculino. Ao sexo feminino há mais observações discrepantes, mas não se pode afirmar se é causal, uma vez que esta classe tem o triplo de observações que a masculina. Em relação à presença de segundo motorista, vemos valores mais elevados de sinistro para as apólices que possuem motorista único. A dispersão da região urbana é pequena, mas com muitas observações discrepantes. Enquanto que, para a rural acontece exatamente o contrário, maior dispersão e menos valores extremos. As diferenças nas formas de pagamento são discretas, porém, pode-se afirmar que a variabilidade do valor dos sinistros é menor no pagamento anual, em relação ao mensal.

Figura 4.3 – Gráficos de caixas da variável resposta em cada nível dos fatores



Fonte: Da autora (2022)

4.2 Ajuste e comparação de modelos

Para o ajuste da severidade dos sinistros foram utilizados apenas os anos de 2010 a 2013. O ano 2014 foi separado como um conjunto de dados teste, para posterior avaliação da capacidade de predição dos modelos. Pelas próprias características dos dados, percebe-se claramente a necessidade de distribuições ajustadas em zero para proporcionar um bom modelo. Com relação à distribuição para os valores positivos, foram testadas diversas distribuições, desde as mais simples, com dois parâmetros, até mais complexas, com quatro parâmetros. Não foi observado uma melhora notável na qualidade do ajuste para distribuições com mais parâmetros. Por este motivo, foram escolhidas duas potenciais distribuições em sua versão ajustada em zero: a gama (ZAGA) e a normal inversa (ZAIG). Ambas já são amplamente utilizadas para esse tipo de análise e aparecem com frequência na literatura, como em Heller et al. (2006), Bortoluzzo et al. (2011) e Resti, Ismail e Jamaan (2013), devido à fácil e intuitiva interpretação que proporciona, além de ter uma forma de cálculo simples para o valor esperado da variável resposta que coincide

com a própria fórmula do cálculo de prêmio de risco, apresentada na Seção 2.1.1. Essas distribuições foram testadas com diversas possibilidades de preditores, dos modelos mais simples aos mais complexos, de forma a possibilitar a comparação de ganhos e necessidade de uso de cada classe. Foram consideradas as seguintes classes de modelos:

a) GLM: apenas para μ, sem efeitos aleatórios e funções de suavização;

b) GLMM: apenas para μ, com efeitos aleatórios, mas sem funções de suavização;

c) GAM: apenas para μ, sem efeitos aleatórios, mas com funções de suavização;

d) GAMM: apenas para μ, com efeitos aleatórios e funções de suavização;

e) GAMLSS: todos os parâmetros da distribuição, sem efeitos aleatórios, mas com funções de suavização;

f) GAMMLSS: todos os parâmetros da distribuição, com efeitos aleatórios e funções de suavização.

Os modelos serão descritos ao longo do trabalho utilizando siglas conforme a configuração CLASSE_DISTRIBUIÇÃO. Por exemplo, o modelo GLM com distribuição ZAGA aparecerá como GLM_ZAGA.

De 12 potenciais classes de modelos, destaca-se que, para quatro casos, não foi possível obter um ajuste. Duas delas não convergiram, mesmo depois de inúmeras tentativas com diferentes métodos iterativos e valores iniciais, sendo elas o GLMM_ZAIG e GAMM_ZAGA. Os motivos pelos quais um modelo não converge são complexos e dependem da estrutura do modelo e do conjunto de dados. Stasinopoulos et al. (2017) sugerem que isso pode ocorrer, talvez, pela dificuldade de se encontrar o máximo de uma soma, que ocorre dentro do logaritmo da função de verossimilhança para os casos de distribuição de mistura. Em outros dois casos, o modelo final selecionado por critério não se encaixou na classe desejada. Estes correspondem aos modelos GAM_ZAIG e GAMM_ZAIG, os quais não selecionaram nenhuma variável com suavização e, portanto, recaem à classe GLM.

Em relação às variáveis explicativas, diferentes distribuições e estruturas para preditores levam a diferentes formas de relacionamento das covariáveis com o parâmetro e, portanto, diferentes variáveis podem ser selecionadas. No Quadro 4.1 é possível visualizar as variáveis significativas, ao nível de 90% de confiança, nos diferentes modelos avaliados.

Quadro 4.1 – Presença e forma de relacionamento das variáveis em cada modelo

Modelo	Preditor	Idade do cliente	2º Motorista	Método de pagamento	Fidelidade	Idade do automóvel	Potência	Id. da apólice
GLM_ZAGA	μ		linear	linear		linear		
GLM_ZAIG	μ						linear	
GLMM_ZAGA	μ		linear	linear		linear		re()
GAM_ZAGA	μ	pb()						
GAMLSS_ZAGA	μ	pb()						
	σ							
	ν	pb()	linear	linear	pb()	linear		
GAMMLSS_ZAGA	μ	pb()						
	σ							
	ν	pb()	linear	linear	pb()	linear		re()
GAMLSS_ZAIG	μ		linear	linear				
	σ	pb()	linear	linear	pb()	pb()	pb()	
	ν	pb()	linear	linear	pb()	linear		
GAMMLSS_ZAIG	μ		linear	linear				
	σ	pb()	linear	linear	pb()	pb()	pb()	
	ν	pb()	linear	linear	pb()	linear		re()

Fonte: Da autora (2022)

Destaca-se que região e sexo não foram selecionadas em nenhum modelo. Estas são variáveis que normalmente são significativas para explicar fenômenos de seguros, como apresentado na Seção 2.1.2. Em relação ao sexo, uma possível explicação é a tendência do comportamento na condução masculina e feminina estar se assemelhando, como apontou Medders, Parson e Thomas-Reid (2021). Outra possível explicação é que, a proporção de mulheres que foi observada no conjunto de dados é muito discrepante da proporção estimada da população espanhola, gerando o confundimento do efeito. Em relação à região, em geral, as urbanas estão associadas a maior frequência, enquanto que as rurais, à maior severidade (SHERAFATI et al., 2017). Possivelmente ocorreu uma situação semelhante ao efeito de sexo: confundimento do efeito de região devido a uma proporção observada muito discrepante da proporção encontrada na realidade territorial espanhola (BANCO MUNDIAL, 2022). Para essas variáveis, seria necessário obter informações mais específicas sobre o produto de seguro que foi comercializado para composição desta carteira, a fim de melhor entender de onde se originam esses comportamentos atípicos.

Ainda em relação ao Quadro 4.1, observa-se que, os modelos para média também foram pouco capazes de captar os possíveis efeitos das covariáveis. Os modelos GLM_ZAIG e GAM_ZAGA selecionaram apenas uma variável, sendo a potência do veículo e idade do cliente, respectivamente. Para os modelos GLM_ZAGA e GLMM_ZAGA foram selecionadas três variáveis explicativas, sendo elas, a presença de segundo motorista, método de pagamento e idade do automóvel. Cada modelo teve um déficit diferente no que tange à capacidade explicativa, pois deixaram de considerar variáveis importantes e informati-

vas para a precificação de seguros, como apresentado na Seção 2.1.2. O principal a se destacar aqui é que nenhum dos modelos para média captou a informação de fidelidade, variável cuja relevância já foi apresentada na Seção 2.1.3. Isto pode ser analisado como um indicativo de como estas classes de modelos já não são mais suficientes para explicar os fenômenos de sinistros e tampouco acompanhar os avanços tecnológicos deste mercado.

O preditor de ν apresentou as mesmas variáveis nos modelos GAMLSS e GAMMLSS. Nota-se que múltiplas variáveis foram selecionadas para explicar a probabilidade do zero e a dispersão do valor dos sinistros ocorridos, mostrando a importância de também modelar estes parâmetros. A possibilidade de captar efeitos, que podem ser relacionar de forma não linear, nos diversos parâmetros da distribuição, faz do GAMLSS e GAMMLSS uma excelente ferramenta para a precificação de seguros, trazendo um leque de informações e resultados vastos que auxiliam os atuários e estatísticos a compreender este evento.

Na Tabela 4.6 é possível verificar os valores de desvio global, AIC e BIC para os modelos ajustados. Para todos os casos, quanto menor o valor, mais parcimonioso é o modelo, sendo estes destacados em negrito. O modelo GAMMLSS_ZAGA é o que apresenta menor desvio global, com valor $8.107,56$, seguido do modelo GAMMLSS_ZAIG, com valor $9.155,78$. O modelo GLMM_ZAGA foi o que apresentou menor AIC ($12.310,25$) e o modelo GLM_ZAGA foi o que apresentou menor BIC ($12.713,12$). No que tange aos valores de AIC calculados, as diferenças entre os modelos são relativamente pequenas. Já para o valor de BIC, percebe-se que os modelos que possuem efeitos aleatórios, receberam maior penalização. Em geral, os critérios AIC e BIC penalizam os modelos GAMLSS e modelos mistos, devido à maior quantidade de parâmetros. O cálculo dessas medidas é bastante comum quando se utiliza modelos estatísticos. Neste trabalho, essas medidas foram obtidas apenas para avaliar a parcimônia dos diversos modelos avaliados. Entretanto, estes valores não trazem nenhuma informação da adequação do modelo e, principalmente, da capacidade preditiva, que é o nosso objetivo central.

Para avaliação da adequação dos modelos, foram utilizados os gráficos de minhoca, introduzidos na Seção 2.2.5.3, que são apresentados na Figura 4.4. Os pontos em cinza representam as repetições da aleatorização e os pontos em preto são a mediana do valor de todas as repetições para determinado resíduo. Para os quatro primeiros gráficos, que representam os modelos com preditor apenas em μ, pode-se perceber grandes desvios das

Tabela 4.6 – Diferentes métricas de qualidade de ajuste para os modelos

Modelo	Desvio Global	AIC	BIC
GLM_ZAGA	12.647,83	12.659,83	**12.713,12**
GLM_ZAIG	13.189,94	13.197,94	13.233,46
GLMM_ZAGA	11.524,53	**12.310,25**	15.799,01
GAM_ZAGA	12.613,25	12.652,88	12.828,80
GAMLSS_ZAGA	12.437,94	12.500,60	12.778,80
GAMLSS_ZAIG	12.357,73	12.510,66	13.189,70
GAMMLSS_ZAGA	**8.107,56**	14.118,64	40.808,95
GAMMLSS_ZAIG	9.115,78	14.476,90	38.281,29

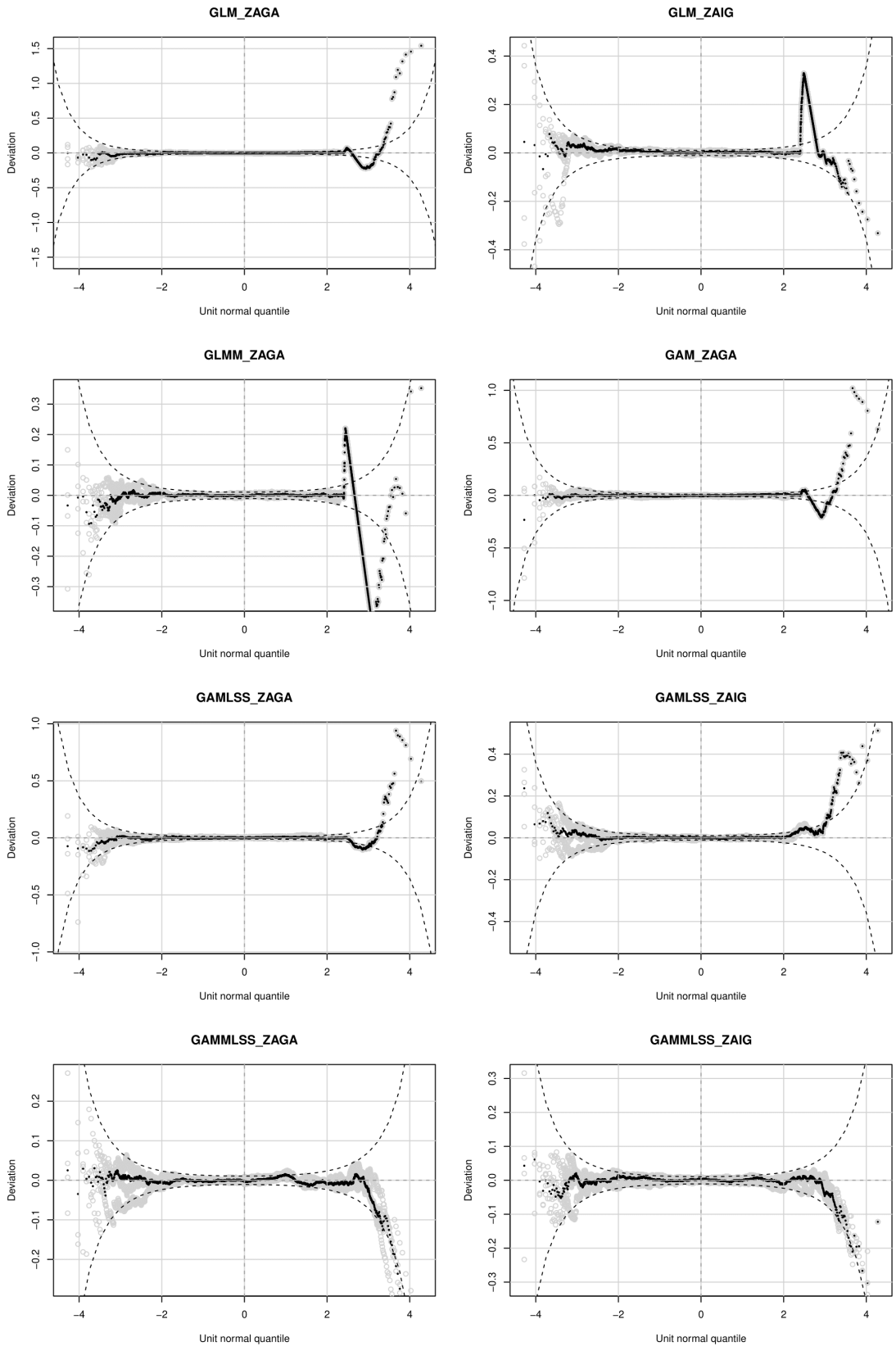
Fonte: Da autora (2022)

bandas, nas caudas superiores dos gráficos. Os desvios apresentados nos modelos para média indicam que o modelo não está ajustado adequadamente, sugerindo que outros parâmetros da distribuição devem ser modelados, o que justifica a necessidade dos GAMLSS. Os GAMLSS, apesar de apresentarem melhor adequação do que os modelos para a média, também mostraram inadequações na mesma região. Nos dois últimos gráficos podemos observar a adequação dos GAMMLSS, que apresentam qualidade superior a todos os demais modelos. A inclusão do efeito aleatório do histórico do segurado – classificação por experiência – melhorou a adequação dentro da estrutura GAMLSS para GAMMLSS. A correta adequação de um modelo é crucial para garantir que as pressuposições do modelo foram atendidas e, portanto, que as predições obtidas sejam confiáveis.

Um maior acerto nas predições é o objetivo principal e o que mais interessa à seguradora. Modelos que apresentam melhor performance calculam de forma mais acurada o prêmio de risco. Logo, o risco do seguro é dividido mais apropriadamente entre os segurados, assim como as seguradoras podem apresentar preços mais competitivos no mercado. Por este motivo, foi investigada a capacidade preditiva de todos eles, utilizando medidas de acurácia, precisão e percentual de acerto da perda agregada da seguradora para o último ano de avaliação. A perda agregada é uma quantia muito importante do ponto de vista da seguradora, porque representa o montante mínimo necessário para cumprir com seu compromisso perante os segurados. Na Tabela 4.7 é possível ver tais medidas, sendo que os melhores valores estão apresentados em negrito.

Para o ano de 2014, a perda agregada da carteira, que é dada pela soma de todas as indenização no período, foi igual a 151.545,50 €. Pode-se observar que o modelo GAMMLSS_ZAIG foi o modelo que mais se aproximou na predição do valor que realmente aconteceu no ano de 2014, subestimando esse valor em apenas 2,48%. Este é o

Figura 4.4 – Gráficos de minhoca para cada modelo



Fonte: Da autora (2022)

Tabela 4.7 – Medidas de acurácia e precisão para os modelos

Modelo	**Viés médio do erro**	**Média dos erros absoluta**	**Raiz quadrada média do erro**	**Perda predita agregada**	**% do real**
GLM_ZAGA	1,18	23,77	167,61	167.258,60	10,36
GLM_ZAIG	1,41	24,01	167,62	170.400,90	12,44
GLMM_ZAGA	-2,34	**20,29**	167,70	120.401,00	-20,55
GAM_ZAGA	1,51	24,09	167,69	171.686,20	13,29
GAMLSS_ZAGA	1,38	23,89	167,64	169.884,00	12,10
GAMLSS_ZAIG	-1,86	20,72	**167,55**	126.741,70	-16,36
GAMMLSS_ZAGA	5,02	25,82	236,85	218.293,10	44,04
GAMMLSS_ZAIG	**-0,28**	20,39	174,19	147.779,90	**-2,48**

Fonte: Da autora (2022)

modelo mais acurado dentre todos, avaliando pelo viés médio do erro $(-0,28)$ e segundo mais acurado, avaliando a média dos erros absoluta $(20,29)$, aproximando-se bastante da primeira posição ocupada pelo GLMM_ZAGA com diferença de apenas $0,10$ pontos.

A precisão de todos os modelos foi bastante próxima, com exceção do modelo GAMMLSS_ZAGA, que foi mais impreciso dentre os avaliados, por meio da raiz quadrada média do erro $(25,82)$. Como a ênfase desse trabalho está no desempenho de predição, escolheu-se como melhores potenciais, os modelos GAMLSS e GAMMLSS, com distribuição ZAIG, sendo que o modelo mais acurado foi o GAMMLSS_ZAIG e o mais preciso o GAMLSS_ZAIG.

4.2.1 GAMLSS *versus* GAMMLSS com distribuição ZAIG

Os modelos finais selecionados ao considerar o GAMLSS_ZAIG e o GAMMLSS_ZAIG são muito semelhantes. Eles foram obtidos por meio da Estratégia A de seleção, apresentada na Seção 3.3, e podem ser representados por:

GAMLSS_ZAIG

$$
\begin{aligned}
log(\mu) &= \text{2oMotorista} + \text{FormaPagto} \\
log(\sigma) &= pb(\text{Fidelidade}) + pb(\text{IdadeCliente}) + \text{2oMotorista} \\
&\quad + \text{FormaPagto} + pb(\text{IdadeVeic}) + pb(\text{Potencia}) \\
logit(\nu) &= pb(\text{IdadeCliente}) + pb(\text{Fidelidade}) + \text{2oMotorista} \\
&\quad + \text{IdadeVeic} + \text{FormaPagto},
\end{aligned}
$$

GAMMLSS_ZAIG

$$
\begin{aligned}
log(\mu) &= \text{2oMotorista} + \text{FormaPagto} \\
log(\sigma) &= pb(\text{Fidelidade}) + pb(\text{IdadeCliente}) + \text{2oMotorista} \\
&\quad + \text{FormaPagto} + pb(\text{IdadeVeic}) + pb(\text{Potencia}) \\
logit(\nu) &= pb(\text{IdadeCliente}) + pb(\text{Fidelidade}) + \text{2oMotorista} \\
&\quad + \text{IdadeVeic} + \text{FormaPagto} + re(\text{ApoliceID}).
\end{aligned}
$$

O GAMMLSS_ZAIG se diferencia do GAMLSS_ZAIG apenas pelo fato de possuir o efeito aleatório da classificação por experiência do segurado (re(ApoliceID)), incluído aditivamente em ν.

Na Tabela 4.8 é possível visualizar os coeficientes dos modelos e seu respectivo valor-p. Nota-se que um nível das covariáveis qualitativas está oculto, isto ocorre porque só é possível se obter estimativas para fatores, considerando alguma restrição (RENCHER; SCHAALJE, 2008). A restrição utilizada foi considerar que o primeiro nível tem efeito nulo – efeito confundido com o intercepto – de forma que apresentamos os efeitos dos fatores estimáveis. Algumas variáveis não foram significativas, destacadas na cor cinza claríssimo, mas foram mantidas no modelo. Variáveis que foram selecionadas de acordo com determinado critério, mas que não foram significativas, observando o valor-p não devem ser retiradas manualmente do modelo porque causará sobreajuste, prejudicando interpretações e resultados, como alertam Lee et al. (2016). Com exceção da idade do veículo em ν, todas as demais variáveis contínuas apresentaram um relacionamento não linear com seu respectivo parâmetro e foram suavizadas com a função *P-spline*. Os coeficientes das variáveis suavizadas também foram destacados, aparecendo em cinza escuro, uma vez que não devem ser interpretados como os demais efeitos fixos. Para estes casos, interpreta-se o comportamento dos valores da variável ao longo dos valores dos parâmetros (RAMIRES et al., 2019), como apresentado na Figura 4.5. Estes gráficos são idênticos tanto para o GAMLSS quanto para GAMMLSS.

Se a forma de pagamento adotada for a anual, então, espera-se que o valor da média das ocorrências de sinistro seja 0,65 vezes menor e que o parâmetro de dispersão desses valores tenha uma redução, quando comparada ao pagamento mensal. Ainda, para pagamentos anuais, o modelo aponta que a probabilidade de não ocorrência é aumentada em relação à mensal. Há pouca ou nenhuma literatura que afirme existir maior ou menor risco envolvendo diferentes formas de pagamento. Entretanto, há análises do comportamento do risco em apólices com vigências fixadas em períodos de tempo e fixadas no uso, como em Soleymanian, Weinberg e Zhu (2019) e Tselentis, Yannis e Vlahogianni (2016). Nosso modelo aponta, porém, que existe maior risco quando o pagamento é mensal. Uma possível justificativa seria inferir que, em períodos de pagamentos mais curtos, haja uma mudança comportamental da condução dos veículos por parte dos segurados, se tornando mais imprudentes. Todavia, esta seria apenas uma suposição. Novamente, seria necessário

Tabela 4.8 – Coeficientes dos modelos GAMLSS_ZAIG e GAMMLSS_ZAIG

		GAMLSS		**GAMMLSS**	
		Estimativa	**Valor-p**	**Estimativa**	**Valor-p**
$log(\mu)$	(Intercepto)	6,8884	0,0000	6,8879	0,0000
	2oMotorista_Sim	-0,0844	0,4231	-0,0841	0,4246
	FormaPagto_Anual	0,2241	0,0002	0,2240	0,0002
$log(\sigma)$	(Intercepto)	-3,2902	0,0000	-3,2894	0,0000
	pb(Fidelidade)	-0,0439	0,0000	-0,0439	0,0000
	pb(IdadeCliente)	0,0002	0,9409	0,0002	0,9406
	2oMotorista_Sim	0,8189	0,0000	0,8085	0,0000
	FormaPagto_Anual	0,5370	0,0000	0,5374	0,0000
	pb(IdadeVeic)	0,0089	0,1585	0,0089	0,1596
	pb(Potencia)	-0,0026	0,0005	-0,0026	0,0005
$logit(\nu)$	(Intercepto)	2,7718	0,0000	4,1570	0,0000
	pb(IdadeCliente)	0,0269	0,0000	0,02386	0,0000
	pb(Fidelidade)	0,0436	0,0001	0,0253	0,0835
	2oMotorista_Sim	-0,5971	0,0000	-0,7661	0,0000
	IdadeVeic	-0,0182	0,0264	-0,0495	0,0000
	FormaPagto_Anual	0,2467	0,0233	0,3374	0,0107

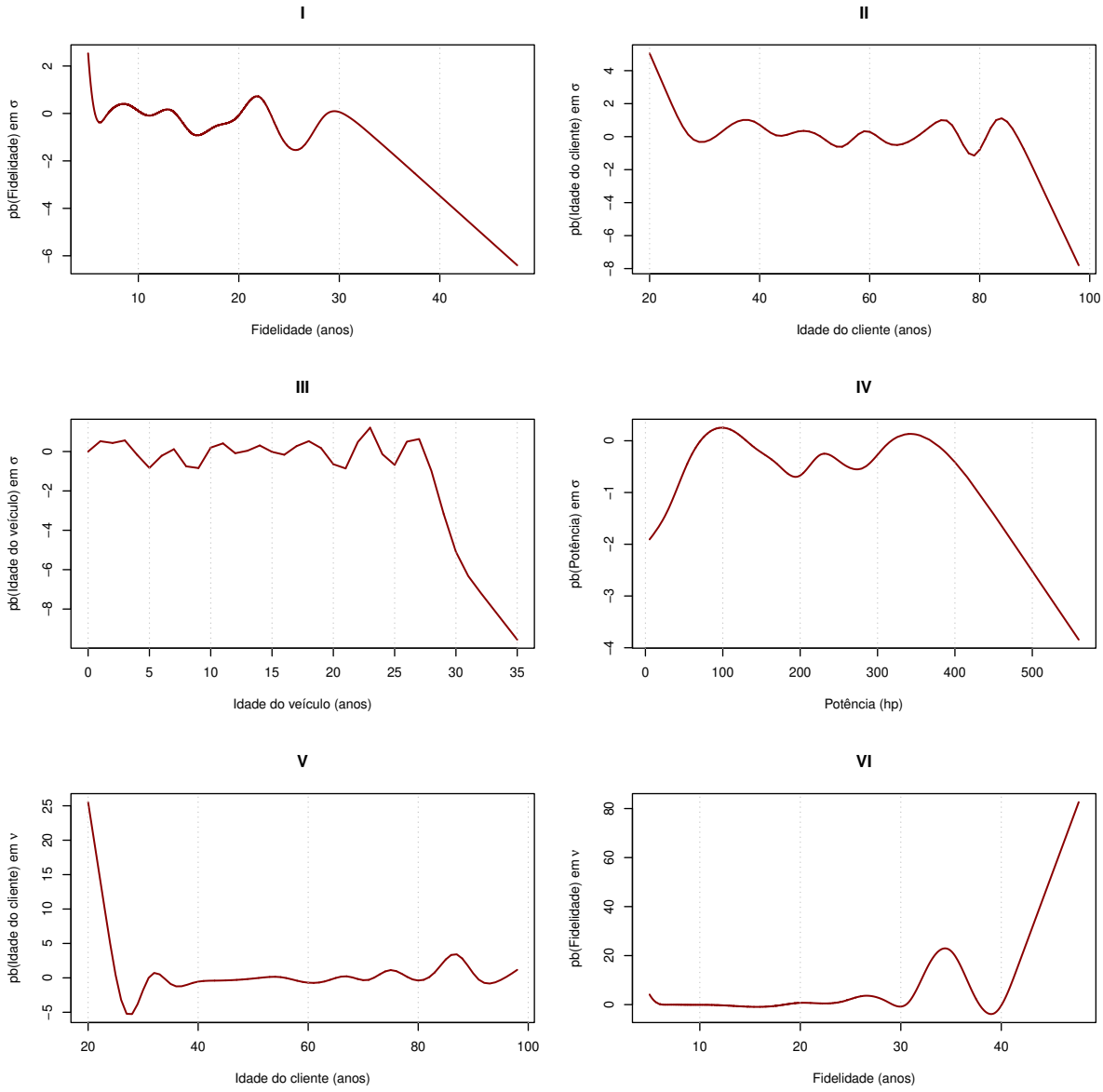
Fonte: Da autora (2022)

obter informações mais específicas sobre o produto de seguro que foi comercializado para composição dessa carteira, para entender melhor o efeito e comportamento da variável forma de pagamento.

Ambos os modelos apontam para uma redução no parâmetro de dispersão do valor da ocorrência para apólices que têm um segundo motorista, em comparação às que não possuem. Quanto maior a variabilidade, maior será a probabilidade esperada de ocorrência de valores extremos (CARDELL, 1997). O resultado relativo à presença de segundo motorista contrasta com o que é encontrado em outras análises. Em geral, a presença de dois motoristas implica naturalmente em maior variabilidade, como apontam Edlin e Karaca-Mandic (2006) e Saito, Kato e Shimane (2010), o que não ocorreu para essa carteira de apólices, cujos sinistros são mais constantes para essa categoria. Com relação à probabilidade de não ocorrência de reclame do sinistro, espera-se uma menor probabilidade para as apólices que possuem segundo motorista cadastrado. Isto é, maior frequência para esta categoria, logo, maior risco, como já apontado na Seção 2.1.1.

Além disso, quanto mais antigo é o veículo, menor será a probabilidade da não ocorrência. Resultado coerente com o que atestamos na Seção 2.1.2. Por um lado, veículos mais novos são associados à maior severidade, e, por outro, veículos antigos à maior frequência, devido à maior probabilidade de falhas mecânicas (FREES; VALDEZ, 2008).

Figura 4.5 – Gráfico de termos da regressão para os parâmetros



Fonte: Da autora (2022)

Pela Figura 4.5 no gráfico III, é apresentado o relacionamento da idade do veículo com o parâmetro σ. Desde carros recém-comprados até veículos com cerca de 27 anos há pequenas oscilações no valor de σ. Após esta idade, a tendência do valor do parâmetro é decrescente;

Ainda, observando a Figura 4.5, temos que:

a) Para novos clientes, a dispersão dos valores tende a ser mais elevada, apresentando uma queda após cerca de 5 anos e com pequenas oscilações até o valor de 30 anos de fidelidade. Após este valor, o parâmetro apresenta tendência decrescente

linear (gráfico I). A probabilidade de não ocorrer reclame de sinistro é praticamente constante para novos clientes, apresenta um pico entre 30 e 40 anos, e, após este valor, a tendência da probabilidade do zero é crescente (gráfico VI);

b) Para clientes jovens a dispersão do valor dos sinistros tem início com valor elevado, apresentando queda até a faixa de 30 anos. Entre 30 e 85 anos apresenta pequenas oscilações e, após esta idade, o valor apresenta tendência decrescente linear (gráfico II). A probabilidade da não ocorrência é decrescente na faixa de 20 a 30 anos, com um leve aumento no final da faixa, e quase constante para as demais idades (gráfico V);

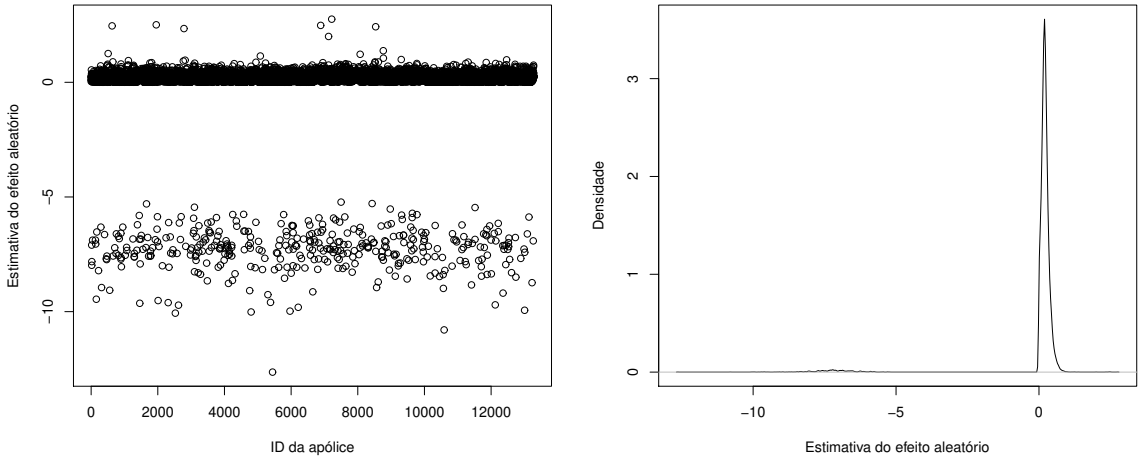
c) A tendência da dispersão do valor dos sinistros ocorridos é crescente para veículos com até 100 hp de potência, com pequenas oscilações até o 350 hp e após este valor, a tendência do valor do parâmetro é decrescente (gráfico IV).

Ao avaliar a Figura 4.5, a principal observação a se fazer é que o valor do sinistro tende a variar menos para clientes mais velhos, veículos mais antigos, veículos mais potentes e para clientes que estão na seguradora por mais tempo. Portanto, nosso modelo aponta que estas classes têm menor risco de ocorrerem valores extremos. Já a probabilidade da não ocorrência é diminuída para clientes jovens. Este é um resultado que concorda com o que é observado na realidade, como aponta, por exemplo, David (2015) e várias outras referências, sendo algumas, mencionadas na Seção 2.1.2. Outro resultado geral importante é que quanto mais anos um cliente tenha de fidelidade à seguradora, maior é a sua probabilidade da não ocorrência. Este comportamento também foi encontrado por Arvidsson (2011).

Para os GAMMLSS, o efeito aleatório de apólice foi incluído na probabilidade dos zeros ($\boldsymbol{\nu}$), alterando os interceptos no preditor deste parâmetro. Foi adotada uma estrutura de pesos fixos para a matriz de covariâncias, ou seja, cada indivíduo, que no caso é visto como um grupo contendo suas próprias observações ao longo do tempo, recebe sua respectiva variância que é constante dentro do grupo. Na Figura 4.6 é possível observar os gráficos de dispersão e distribuição de densidade dos efeitos aleatórios. Os valores observados abaixo da origem representam as apólices que, em algum momento, apresentaram histórico positivo de sinistro, aplicando uma diminuição no valor do intercepto da probabilidade de zero, ou seja, uma apólice com maior risco para o ano seguinte. O contrário

também se aplica. Para os valores acima da origem, o intercepto é aumentado, de forma a aumentar a probabilidade do zero, dado um histórico de nunca haver reclamado sinistro.

Figura 4.6 – Gráfico de dispersão e densidade dos valores ajustados para os efeitos aleatórios

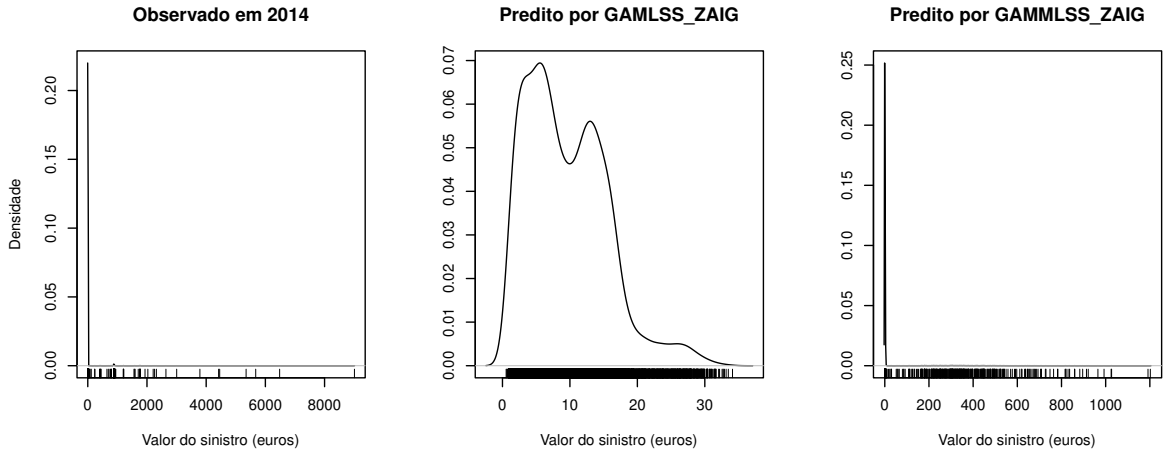


Fonte: Da autora (2022)

A inclusão do efeito aleatório alterou substancialmente as predições para o ano de 2014. Na Figura 4.7 é possível perceber o comportamento destas para cada modelo e, também, para os dados reais de 2014. Note que as escalas dos eixos de cada gráfico são bastante diferentes, pois não foi possível padronizar os eixos, devido à discrepância dos valores das escalas, o que impedia uma melhor visualização. O modelo GAMLSS, localizado ao centro, possui menor amplitude dos valores de ocorrência, estimando um valor médio para o prêmio de risco para todas as apólices, e considera apenas os efeitos das classes de risco. Já o modelo GAMMLSS, localizado à direita, leva em consideração a classificação por experiência e por risco do segurado, apresentando um formato de densidade mais semelhante à densidade observada verdadeiramente no ano de 2014. Para o GAMMLSS, aqueles que nunca acionaram a seguradora, possuem uma predição com valor bem baixo, quase nulo; e aqueles que já acionaram alguma vez são penalizados com uma predição mais alta, dado o histórico de maior risco. O modelo misto reproduziu mais fidedignamente o comportamento que realmente ocorreu, enquanto que o modelo GAMLSS diluiu o risco entre todos os segurados, desconsiderando a informação histórica.

Na Tabela 4.9 é possível analisar o comportamento de ambos os modelos em relação à sua acurácia e precisão em predizer a ocorrência, ou não, dos sinistros para o ano de

Figura 4.7 – Gráficos de densidade dos valores preditos para cada modelo e para os valores observados no ano de 2014



Note que os eixos dos gráficos são diferentes. Considerar este fato ao fazer comparações.
Fonte: Da autora (2022)

2014. O modelo com efeito aleatório foi mais preciso e mais acurado em predizer o valor da ocorrência onde realmente houve sinistro. Em relação à predição da não ocorrência, observou-se pouca diferença na acurácia pela adição ou não do efeito aleatório. Entretanto, o GAMLSS foi mais preciso nesta situação.

Tabela 4.9 – Medidas de acurácia e precisão para os modelos subdividido em onde houve ou não ocorrência

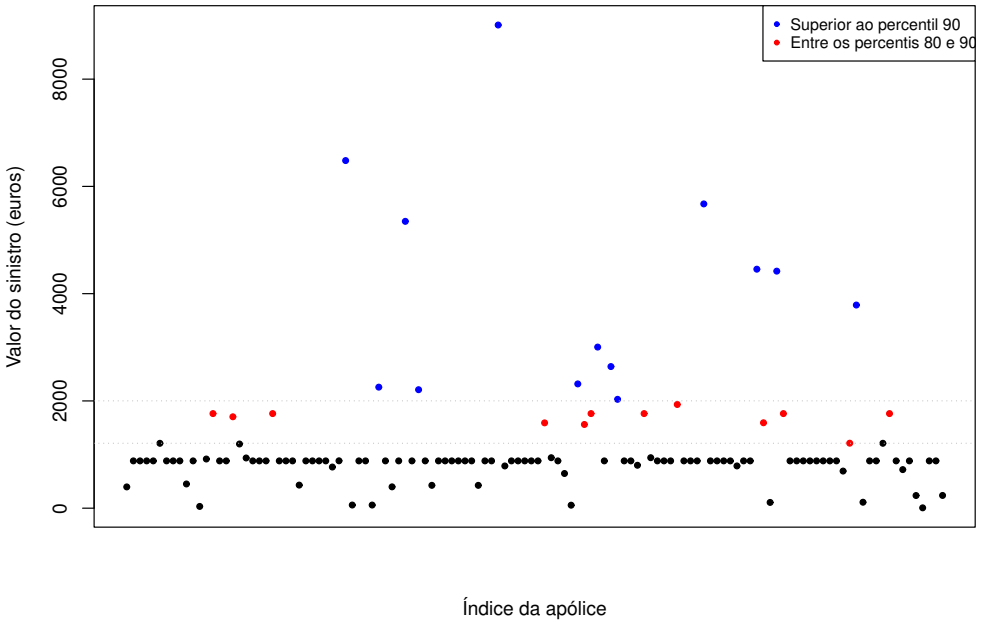
	Modelo	**Viés médio do erro**	**Média dos erros absoluta**	**Raiz quadrada média do erro**
Houve	GAMLSS_ZAIG	-1.209,98	1.209,98	1.730,19
ocorrência	GAMMLSS_ZAIG	-1.091,05	1.123,79	1666,33
Não houve	GAMLSS_ZAIG	9,52	9,52	11,24
ocorrência	GAMMLSS_ZAIG	9,99	9,99	66,79

Fonte: Da autora (2022)

Também foi feita análise de alguns grupos de apólices para identificar como os modelos funcionam. Com base no que ocorreu no ano de 2014, foram considerados os percentis 90 e 80 como bandas de corte, para definir os grupos a serem analisados. O valor do percentil 90 foi igual a 2001,18 €; e, do percentil 80, igual a 1210,98 €. Essa definição teve o objetivo de analisar o comportamento do modelos para observações discrepantes, ou seja, aquelas superiores ao percentil 90 e observações que consideramos altas, que estão entre o percentil 80 e 90. Além disso, também foi considerada uma amostra para avaliar o comportamento de ambos modelos para segurados que nunca reclamaram sinistro. A Figura 4.8 mostra as ocorrências positivas no ano de 2014. Foram detectadas 13 observa-

ções discrepantes e 12 observações elevadas, de um total de 124 ocorrências positivas nesse ano. Na referida figura é possível visualizar também o grande número de observações no valor 882,00 €, que é referente ao seguro *sem culpa*, já explicado anteriormente na Seção 2.1.1.

Figura 4.8 – Gráfico de dispersão dos valores positivos observados no ano de 2014 e percentis 80 e 90



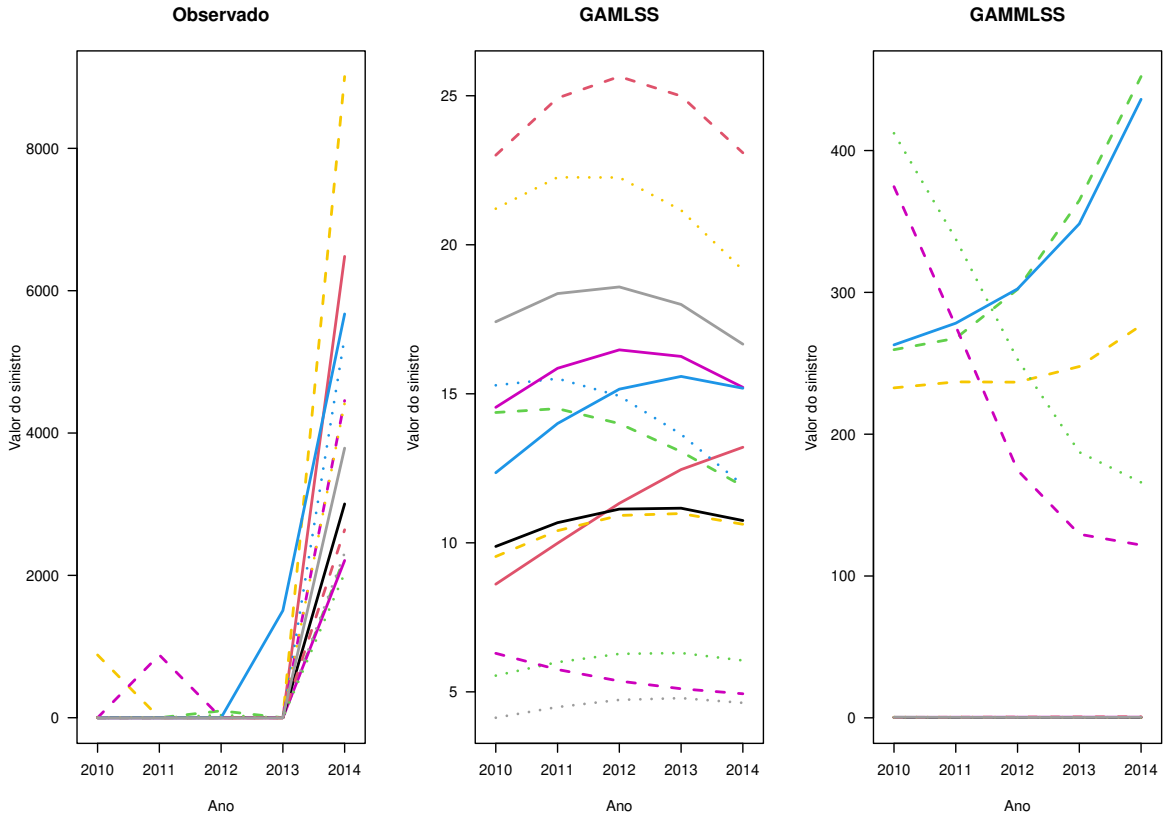
Fonte: Da autora (2022)

Primeiramente analisaremos o histórico do grupo das observações discrepantes. Na Figura 4.9 é possível visualizar este perfil de sinistros das 13 apólices discrepantes ao longo dos cinco anos. Cada linha, com sua respectiva cor e tipo de traço, representa o perfil histórico de um segurado. No gráfico dos valores observados é possível notar como o comportamento do valor foi atípico, com valores extremos em 2014 e valores relativamente baixos ou nulos para os anos anteriores, sendo este um comportamento inesperado e difícil de ser predito. Para estas apólices, a seguradora receberia individualmente valores entre 5 € e 25 € pelo modelo GAMLSS, enquanto que, na realidade, ela teria gastos individuais com indenizações que superam 8000 €.

No gráfico do GAMMLSS são apresentadas as predições para este modelo. Para as apólices com histórico nulo de sinistro, o valor predito também é bem próximo do zero. No entanto, para aquelas que possuem algum histórico de ocorrência, têm-se valores

mais altos, sendo que há estimativas de prêmios de risco de até 450 €, valor até 18 vezes maior do que o estimado pelo GAMLSS. Nota-se que os eixos das ordenadas de cada gráfico são consideravelmente diferentes. O GAMMLSS consegue captar e levar em consideração o comportamento individual, penalizando os poucos indivíduos que têm um histórico de risco maior e bonificando aqueles muitos segurados que possuem histórico de risco individual nulo. Todas as predições com valores maiores que zero obtidas com o GAMMLSS coincidem com as curvas que tiveram ao menos um sinistro em anos anteriores. Para estas apólices, o modelo misto (GAMMLSS) foi mais acurado e mais preciso que o modelo semiparamétrico (GAMLSS).

Figura 4.9 – Gráfico de perfis dos valores observados e preditos para as apólices discrepantes no ano de 2014



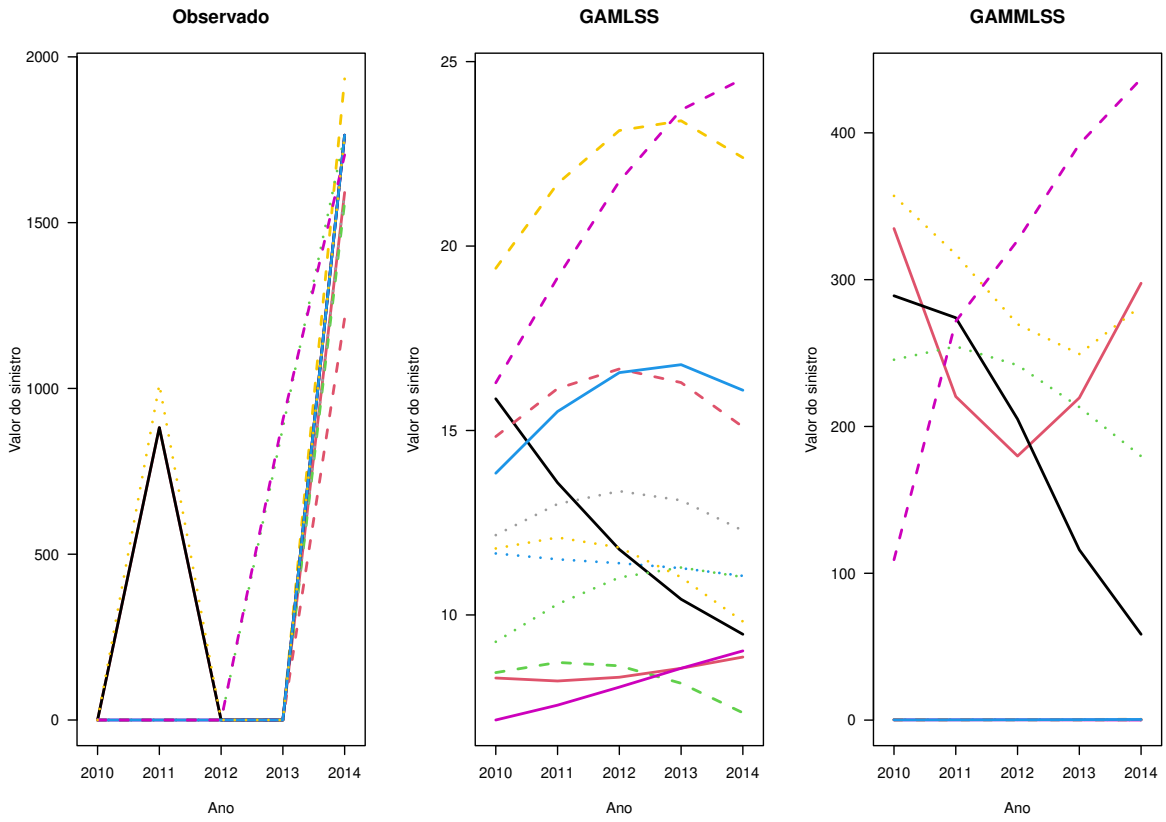
Note que os eixos das ordenadas dos gráficos são diferentes. Considerar este fato ao fazer comparações.
Fonte: Da autora (2022)

Para este grupo de indivíduos, a média da idade dos clientes foi entre 60,76 anos; a dos veículos, de 12 anos; e, a do tempo de fidelidade, de 12,48 anos. Além disso, cerca de 23% possuem um segundo motorista cadastrado e 37,5% faziam o pagamento mensalmente. É difícil atribuir essas classes de risco como causas para estes valores extremos no quinto ano, uma vez que o histórico é muito inferior nos anos que antecedem 2014. Em

especial, parece não haver nenhum padrão para identificar essas ocorrências, que, para o modelo, são dadas por infortúnio do acaso e de difícil predição, como já mencionado anteriormente. Mesmo para estes casos, GAMMLSS contribuiu para minimizar o erro de predição.

Na Figura 4.10 são mostrados gráficos similares aos da figura anterior, só que referente a 12 observações que foram classificadas como elevadas, ou seja, aquelas que ocorreram entre os valores dos percentis 80 e 90. Nestes gráficos, cada linha representa uma apólice e o padrão observado na Figura 4.9 se repete, ou seja, o GAMLSS ajusta valores médios para todas as apólices, enquanto que o GAMMLSS leva em consideração a característica individual do segurado. Novamente, as estimativas maiores que zero em GAMMLSS coincidem com as mesmas apólices que possuem histórico positivo. GAMLSS estima valores até 25 € e GAMMLSS até 450 €. Para os casos atípicos, isto é, os casos em que o primeiro pedido de reclame de sinistro se deu no último ano e não há histórico positivo, houve pouca diferença entre a precisão e acurácia de ambos modelos.

Figura 4.10 – Gráfico de perfis dos valores observados e preditos para as apólices com valor elevado no ano de 2014



Note que os eixos das ordenadas dos gráficos são diferentes. Considerar este fato ao fazer comparações.
Fonte: Da autora (2022)

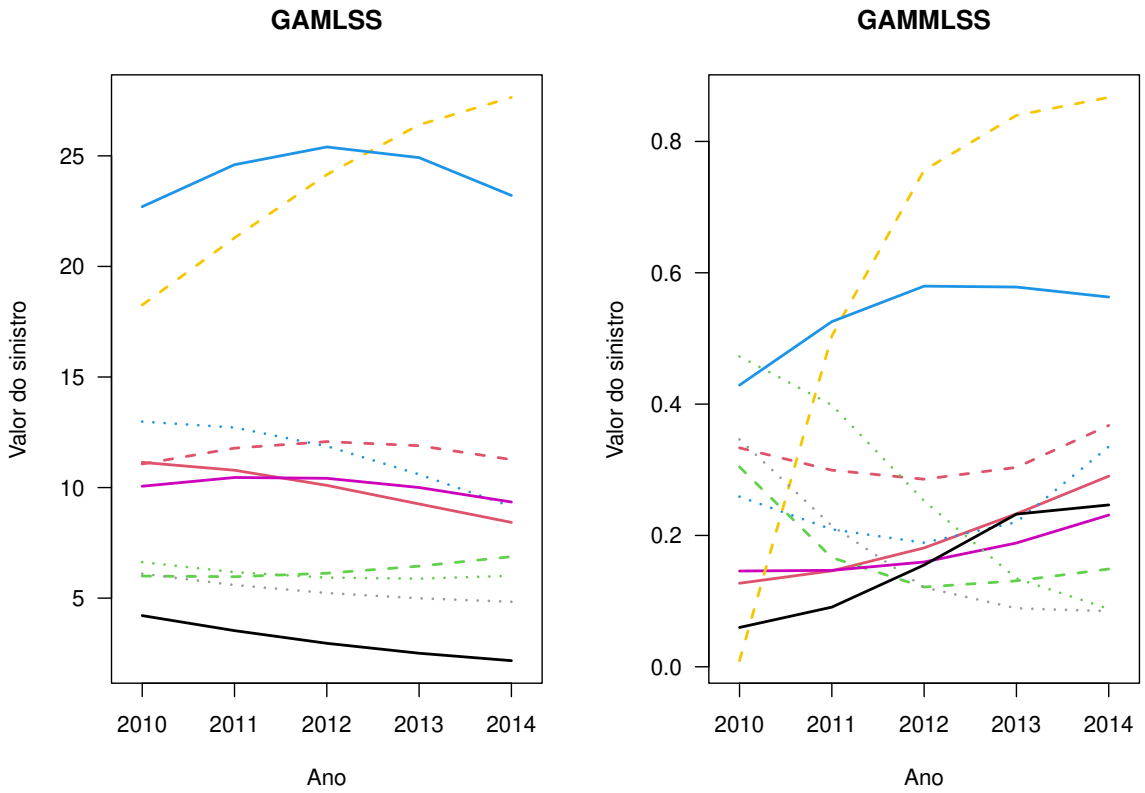
Para este grupo de apólices, a média da idade do cliente foi de 63 anos, semelhante à média geral, e a média da idade do veículo é igual a 16,25 anos, superior à geral. A média da fidelidade foi de 15 anos, dois anos a mais da média de toda a carteira. Ainda, em 33% delas há um segundo motorista e, em 33%, o pagamento é mensal, percentuais também mais elevados que o observado em toda carteira.

A Figura 4.11 é também similar às duas anteriores, com exceção de não haver um gráfico para os valores observados, dado que as apólices analisadas nunca acionaram o seguro. Portanto, o histórico dessa apólices seria uma linha constante na origem. Foram amostradas 10 apólices, de um conjunto de 13.157 segurados que nunca acionaram o seguro. Nesta Figura é possível observar uma inversão nas predições, ou seja, o GAMLSS estima valores superiores aos obtidos pelo GAMMLSS. Para todos os casos, GAMLSS estimou um valor próximo da média para todos, não importando o fato de aquela apólice já ter usufruído do seguro em algum momento ou não. Neste caso, GAMMLSS estimou valores bem pequenos para essa categoria que nunca acionou, bonificando com um custo mais barato. Para o caso geral, onde não houve ocorrência, GAMMLSS é mais preciso e mais acurado também.

Para a amostra de apólices, considerando na Figura 4.11, a idade média do segurado foi de 56,6 anos; da idade do veículo, de 9,6 anos; e, a fidelidade de 13,64 anos, sendo que apenas 20% das apólices tinham um segundo motorista e 20% pagavam mensalmente. Valores estes que são bem próximos do observado na média global e substancialmente inferiores aos detectados nas observações discrepantes e elevadas.

O fato de haver um histórico positivo de sinistros anteriormente à contratação, pode ser um indicativo de maior risco para aquele indivíduo e, logo, para seguradora (ARVIDSSON, 2011), mas essa não é uma regra estrita. Também ocorrerá casos em que o indivíduo segurado tenha passado todos os anos sem nunca haver reclamado uma indenização e ocorrer um valor extremo no ano seguinte. Para estes casos, a seguradora deve manter reservas monetárias e aplicar carregamentos de segurança no valor dos prêmios. Haverá situações, entretanto, em que ocorrerá justamente o contrário. Um segurado precisou acionar o seguro, por exemplo, logo no primeiro ano em que ingressou à carteira, e devido a este fato, teve prêmios de renovação mais caros. Porém, suponha-se que ele nunca mais precisou reclamar sinistro. Neste caso, a tendência é que o segurado evada da carteira, migrando para outra sociedade seguradora, ou até mesmo fique sem

Figura 4.11 – Gráfico de perfis dos valores observados e preditos para uma amostra das apólices sem ocorrências



Note que os eixos das ordenadas dos gráficos são diferentes. Considerar este fato ao fazer comparações.
Fonte: Da autora (2022)

seguro. O principal desafio enfrentado pelos atuários é exatamente esse. Qual a melhor maneira de diluir o custo das indenizações entre todos os segurados? Desconsiderando o aspecto econômico, monetário, comercial e de comportamento do consumidor, o modelo GAMMLSS_ZAIG se mostra promissor na obtenção das estimativas dos sinistros a ocorrerem no ano seguinte, auxiliando no cálculo dos prêmios com maior precisão e maior acurácia.

5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo propor uma metologia para cálculo do prêmio de risco em seguros de automóveis que levasse em consideração, não só as classes de riscos e a complexidade de tal variável, como também, o histórico de ocorrências do usuário. O objetivo foi alcançado na medida em que foram comparados diversos modelos em relação à sua capacidade preditiva e adequação, analisando-se as respectivas medidas de acurácia e precisão, relativas ao que realmente aconteceu no último ano, para um caso particular de uma seguradora espanhola. Este trabalho, também, mostrou como a utilização de modelos flexíveis podem ser de grande vantagem na precificação de seguros. Ambos os modelos apresentados na Seção 4.2.1, os GAMLSS e os GAMMLSS, podem ser considerados bons. Apesar da pouca diferença ao analisar a acurácia e precisão, pode-se afirmar que GAMMLSS apresentou uma performance superior, uma vez que este foi o modelo que mais se aproximou do valor real da perda agregada para o último ano avaliado. Apesar da estrutura de ambos os modelos serem semelhantes, as predições foram bem diferentes para cada modelo. Além disso, os GAMLSS semiparamétricos ou mistos, seja com distribuição ZAGA ou ZAIG, apresentaram adequação superior, quando comparados com os modelos ajustados apenas para a média.

Para determinados casos reais, como o considerado no presente estudo e em diversos ramos de seguros, principalmente do ramo não vida, é preciso um modelo que seja suficientemente flexível, que lide com excessos de zeros e valores extremos, garantindo boas predições para o prêmio de risco. De um lado, os GAMLSS se colocam como uma excelente opção e que pode ser melhor explorada pelos atuários, principalmente por oferecer diversas distribuições para a variável resposta, diversos termos aditivos para os preditores, a possibilidade de modelar todos os parâmetros da distribuição em função das classes de risco e oferecer um ferramental intuitivo de diagnóstico da qualidade do modelo. Do outro, o mercado de seguros tem mudado o seu funcionamento, inclusive, pressionado pelas restrições impostas nos períodos de *lockdown* da pandemia do coronavírus, por meio da adaptação de tecnologias e compartilhamento de dados. Isto certamente trará mudanças nas formas tradicionais de análise e de composição de dados das carteiras de apólices, as quais apesentarão dados longitudinais com maior frequência, trazendo novas discussões e

possibilidades para o avanço da classe de modelos GAMLSS, em conjunto com a classe de modelos mistos. Além disso, a teoria de credibilidade se relacionou com o modelo GAMMLSS proposto, à medida que os componentes de variância do efeito aleatório funcionam como uma espécie de fator de credibilidade dentro do modelo, como apontado na Seção 2.1.4.

O modelo semiparamétrico (GAMLSS) é aquele que tradicionalmente seria utilizado pelas seguradoras, uma vez que a presença da informação histórica é bastante dispendioso de se obter. Entretanto, com as mudanças tecnológicas supracitadas, o modelo misto (GAMMLSS) se mostra como um grande potencial para aperfeiçoar a obtenção de predições, e logo, proporcionar precificações e estabelecimento de reservas de segurança de forma mais acurada e mais precisa, principalmente porque incorpora o histórico do usuário. Este resultado traz inúmeros benefícios, tanto para o segurado, quanto para a seguradora. Do ponto de vista do motorista, este seria beneficiado com valor de prêmio pequeno, caso nunca houvesse utilizado o sinistro, incentivando o cuidado e prudência no trânsito para manter um bom preço no seguro. Do ponto de vista da companhia seguradora, esta pode obter preços mais competitivos no mercado, além da possibilidade aumentar seus lucros com estimativas cada vez mais corretas.

Como pesquisas futuras, é possível apontar alguns caminhos. O primeiro é que esta é apenas uma aplicação do método. Este deve ser avaliado novamente com diferentes bancos de dados de seguros, dos mais diversos países e das mais diversas categorias do ramo não vida, sendo um possível ferramental para identificação de tendências. O segundo é que realizamos uma filtragem dos dados de forma que todas as apólices tivessem cinco observações, ou seja, dados nos cinco anos estudados. Isto foi feito para que fosse possível obter as medidas de desempenho da predição ao compararmos com o que realmente ocorreu no último ano. Por isso, também faz-se necessário investigar qual seria o comportamento das predições para os casos em que houvessem diferentes exposições das apólices dentro da carteira. Além disso, também no desenvolvimento dessa pesquisa, surgiram novos desafios que precisam ser trabalhados, como propor um método de inflação de um valor específico, como ocorre para o 882,00 € neste conjunto de dados, ou ainda, aprofundar sobre os diagnósticos através dos gráficos de minhoca, investigando quais seriam os possíveis formatos quando aplicado em distribuições do tipo discreta-contínuas, como distribuições ajustadas em zero.

REFERÊNCIAS

ABDEL-ATY, M. Analysis of driver injury severity levels at multiple locations using ordered probit models. **Journal of safety research**, Elsevier, v. 34, n. 5, p. 597–603, 2003.

ABDEL-ATY, M. A.; CHEN, C. L.; RADWAN, A. E. Using conditional probability to find driver age effect in crashes. **Journal of transportation engineering**, American Society of Civil Engineers, v. 125, n. 6, p. 502–507, 1999.

ABDELHADI, S.; ELBAHNASY, K.; ABDELSALAM, M. A proposed model to predict auto insurance claims using machine learning techniques. **Journal of Theoretical and Applied Information Technology**, v. 98, n. 22, 2020.

ACEA, E. A. M. A. Acea report vehicles in use europe 2022. **European Automobile Manufacturers Association: Southfield, MI, USA**, 2022.

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974.

ANTONIO, K.; BEIRLANT, J. Actuarial statistics with generalized linear mixed models. **Insurance: Mathematics and Economics**, Elsevier, v. 40, n. 1, p. 58–76, 2007.

ARVIDSSON, S. **Predictors of customer loyalty in automobile insurance: the role of private information in risky driving behavior and claim history**. [S.l.]: Statens väg-och transportforskningsinstitut, 2011.

ASEERVATHAM, V.; LEX, C.; SPINDLER, M. How do unisex rating regulations affect gender differences in insurance premiums? **The Geneva Papers on Risk and Insurance-Issues and Practice**, Springer, v. 41, n. 1, p. 128–160, 2016.

BANCO MUNDIAL, B. M. **Datos del Banco Mundial**. 2022. Disponível em: <https://datos.bancomundial.org/indicator/SP.POP.TOTL.FE.ZS?end=2021&locations=ES&start=2010>. Acesso em: 12 de março de 2022.

BATES, D. et al. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015.

BELLI, V.; MEDEIROS, L.; PRADO, T. Substituição de pessoas por máquinas e o uso de inteligência artificial pelo mercado segurador. **Revista Brasileira de Risco e Seguro**, v. 14, n. 24, 2018.

BERGDAHL, J. Sex differences in attitudes toward driving: A survey. **The Social Science Journal**, Elsevier, v. 42, n. 4, p. 595–601, 2005.

BERGDAHL, J.; NORRIS, M. R. Sex differences in single vehicle fatal crashes: a research note. **The Social Science Journal**, Taylor & Francis, v. 39, n. 2, p. 287–293, 2002.

BOLAND, P. J. **Statistical and probabilistic methods in actuarial science**. [S.l.]: CRC Press, 2007.

BORTOLUZZO, A. B. et al. Estimating total claim size in the auto insurance industry: a comparison between tweedie and zero-adjusted inverse gaussian distribution. **BAR-Brazilian Administration Review**, SciELO Brasil, v. 8, n. 1, p. 37–47, 2011.

BOTTA, A. et al. Psd2: Taking advantage of open-banking disruption. **McKinsey and Company**, 2018.

BRAZAUSKAS, V.; DORNHEIM, H.; RATNAM, P. Credibility and regression-type modeling. 2013.

BREIMAN, L.; FRIEDMAN, J. H. Estimating optimal transformations for multiple regression and correlation. **Journal of the American statistical Association**, Taylor & Francis Group, v. 80, n. 391, p. 580–598, 1985.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American statistical Association**, Taylor & Francis Group, v. 88, n. 421, p. 9–25, 1993.

BÜHLMANN, H. Experience rating and credibility. **ASTIN Bulletin: The Journal of the IAA**, Cambridge University Press, v. 4, n. 3, p. 199–207, 1967.

BÜHLMANN, H.; STRAUB, E.; BROOKS, C. Glaubwürdigkeit für schadensätze, mitteilungen der vereinigung schweizerischer versicherungsmathematiker. 1970.

BüHLMANN, H. The actuary: the role and limitations of the profession since the mid-19th century. ASTIN Bulletin, 1997.

CAMARINHA FILHO, J. A. **Modelos lineares mistos: estruturas de matrizes de variâncias e covariâncias e seleção de modelos.** Tese (Doutorado) — Universidade de São Paulo, 2002.

CAMPBELL, M. An integrated system for estimating the risk premium of individual car models in motor insurance. **ASTIN Bulletin: The Journal of the IAA**, Cambridge University Press, v. 16, n. 2, p. 165–183, 1986.

CARDELL, N. S. Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. **Econometric Theory**, Cambridge University Press, v. 13, n. 2, p. 185–213, 1997.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Duxbury Pacific Grove, CA, 2002. v. 2.

CHARNET, R. et al. **Análise de Regressão Linear: com aplicações**. [S.l.]: Editora da UNICAMP, 2008. v. 2ª edição.

CHEN, X. et al. Used car prices in india: What about future? In: ATLANTIS PRESS. **2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)**. [S.l.], 2022. p. 831–840.

CHONGSUVIVATWONG, V. **epiDisplay: Epidemiological Data Display Package**. [S.l.], 2018. R package version 3.5.0.1. Disponível em: <https://CRAN.R-project.org/package=epiDisplay>.

CNN Brasil. **Open Banking faz 1 ano: veja a trajetória do sistema financeiro aberto no Brasil**. 2022. Disponível em: <https://www.cnnbrasil.com.br/business/open-banking-faz-1-ano-veja-a-trajetoria-do-sistema-financeiro-aberto-no-brasil>. Acesso em: 15 de julho de 2022.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. **Statistics in medicine**, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992.

CUMMINS, J. D. Risk-based premiums for insurance guaranty funds. **The journal of Finance**, Wiley Online Library, v. 43, n. 4, p. 823–839, 1988.

CUMMINS, J. D.; PHILLIPS, R. D.; WEISS, M. A. The incentive effects of no-fault automobile insurance. **The Journal of Law and Economics**, The University of Chicago Press, v. 44, n. 2, p. 427–464, 2001.

DAMTEW, K.; PAGIDIMARRI, V. The role of trust in building customer loyalty in insurance sector: A study. **IOSR journal of business and management**, v. 14, n. 4, p. 82–93, 2013.

DAVID, M. Auto insurance premium calculation using generalized linear models. **Procedia Economics and Finance**, Elsevier, v. 20, p. 147–156, 2015.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. [S.l.]: USP/ESALQ, 2001.

DEMIDENKO, E. **Mixed models: theory and applications with R**. [S.l.]: John Wiley & Sons, 2013. v. 2.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

DUTANG, C.; GOULET, V.; PIGEON, M. actuar: An r package for actuarial science. **Journal of Statistical Software**, v. 25, n. 7, p. 38, 2008. Disponível em: <http://www.jstatsoft.org/v25/i07>.

ECKERMANN, E. **World history of the automobile**. [S.l.]: SAE, 2001.

EDLIN, A. S.; KARACA-MANDIC, P. The accident externality from driving. **Journal of Political Economy**, The University of Chicago Press, v. 114, n. 5, p. 931–955, 2006.

EILERS, P. H.; MARX, B. D. Flexible smoothing with b-splines and penalties. **Statistical science**, Institute of Mathematical Statistics, v. 11, n. 2, p. 89–121, 1996.

EILERS, P. H.; MARX, B. D. Splines, knots, and penalties. **Wiley interdisciplinary reviews: Computational statistics**, Wiley Online Library, v. 2, n. 6, p. 637–653, 2010.

EILERS, P. H.; MARX, B. D.; DURBÁN, M. Twenty years of p-splines. **SORT: statistics and operations research transactions**, v. 39, n. 2, p. 0149–186, 2015.

EIOPA, E. I. . O. P. A. **European Insurance and Occupational Pensions Authority**. 2022. Disponível em: <https://www.eiopa.europa.eu/>. Acesso em: 15 de julho de 2022.

ENEA, M. et al. **gamlss.inf: Fitting Mixed (Inflated and Adjusted) Distributions**. [S.l.], 2019. R package version 1.0-1. Disponível em: <https://CRAN.R-project.org/package=gamlss.inf>.

FARROW, J. A.; BRISSING, P. Risk for dwi: A new look at gender differences in drinking and driving influences, experiences, and attitudes among new adolescent drivers. **Health Education Quarterly**, Sage Publications Sage CA: Thousand Oaks, CA, v. 17, n. 2, p. 213–221, 1990.

FERREIRA, L. Q.; CARLOS, F.; SIQUEIRA, É. S. e-insurance ou seguros digitais: As tecnologias de informação e comunicação utilizadas pelas principais empresas seguradoras do brasil. **Journal of Perspective in Management**, v. 2, n. 2, p. 51–65, 2018.

FISHER, R. A. The correlation between relatives on the supposition of mendelian inheritance. **Earth and Environmental Science Transactions of the Royal Society of Edinburgh**, Royal Society of Edinburgh Scotland Foundation, v. 52, n. 2, p. 399–433, 1918.

FRANCESCHI, L. et al. Factors related to highway crash severity in brazil through a multinomial logistic regression model. **TRANSPORTES**, v. 30, n. 1, p. 2566–2566, 2022.

FREES, E. W. **Regression modeling with actuarial and financial applications**. [S.l.]: Cambridge University Press, 2009.

FREES, E. W. et al. Dependence modeling of multivariate longitudinal hybrid insurance data with dropout. **Expert Systems with Applications**, Elsevier, v. 185, p. 115552, 2021.

FREES, E. W.; VALDEZ, E. A. Hierarchical insurance claims modeling. **Journal of the American Statistical Association**, Taylor & Francis, v. 103, n. 484, p. 1457–1469, 2008.

GARCÍA, A. A.; GARCÍA, A. A.; RODRÍGUEZ, R. P. Un perfil de las personas mayores en españa, 2017. indicadores estadísticos básicos. CSIC-Instituto de Economía, Geografía y Demografía (IEGD), 2017.

HACHEMEISTER, C. Credibility for regression models with application to trend (reprint). **Credibility: Theory and Applications. Edited by P. Kahn. New York: Academic Press, Inc**, p. 307–48, 1975.

HARRÉ, N.; FIELD, J.; KIRKWOOD, B. Gender differences and areas of common concern in the driving behaviors and attitudes of adolescents. **Journal of Safety Research**, Elsevier, v. 27, n. 3, p. 163–173, 1996.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. **Journal of the American statistical association**, Taylor & Francis Group, v. 72, n. 358, p. 320–338, 1977.

HASTIE, T.; TIBSHIRANI, R. **Generalized additive models**. [S.l.]: Wiley Online Library, 1990.

HELLER, G. et al. The zero-adjusted inverse gaussian distribution as a model for insurance claims. In: GALWAY. **Proceedings of the 21th International Workshop on Statistical Modelling**. [S.l.], 2006. v. 226233.

HELLER, G.; STASINOPOULOS, M.; RIGBY, R. Randomly stopped models. In: L'INSTITUT D'ESTADISTICA DE CATALUNYA, IDESCAT. **International Workshop on Statistical Modelling (22nd: 2007)**. [S.l.], 2007. p. 323–328.

HELLER, G. Z.; ROBLEDO, K. P.; MARSCHNER, I. C. Distributional regression in clinical trials: treatment effects on parameters other than the mean. **BMC medical research methodology**, BioMed Central, v. 22, n. 1, p. 1–12, 2022.

HENDERSON, C. R. Sire evaluation and genetic trends. **Journal of Animal Science**, Oxford University Press, v. 1973, n. Symposium, p. 10–41, 1973.

HICKMAN, J. History of actuarial profession. **Encyclopedia of actuarial science**, Wiley Online Library, v. 2, 2006.

HULTKRANTZ, L.; NILSSON, J.-E.; ARVIDSSON, S. Voluntary internalization of speeding externalities with vehicle insurance. **Transportation research part A: policy and practice**, Elsevier, v. 46, n. 6, p. 926–937, 2012.

IAIS, I. A. of I. S. **International Association of Insurance Supervisors**. 2022. Disponível em: <https://www.iaisweb.org/>. Acesso em: 15 de julho de 2022.

IBA, I. B. de A. **Instituto Brasileiro de Atuária, IBA**. 2020. Disponível em: <https://www.atuarios.org.br/>. Acesso em: 21 de agosto de 2020.

IBA, I. B. de A. **O atuário**. 2020. Disponível em: <https://www.atuarios.org.br/o-atuario>. Acesso em: 15 de julho de 2020.

IBGE, I. B. de Geografia e E. Classificação e caracterização dos espaços rurais e urbanos do brasil: uma primeira aproximação. **Rio de Janeiro**, 2017.

IFA, I. F. of A. **What is an actuary?** 2020. Disponível em: <https://www.actuaries.org.uk/become-actuary/what-actuary>. Acesso em: 15 de julho de 2020.

INE, I. N. de E. **Instituto Nacional de Estadística (INE)**. 2022. Disponível em: <https://www.ine.es/>. Acesso em: 12 de março de 2022.

Insurance Europe. **Insurance Europe**. 2022. Disponível em: <https://www.insuranceeurope.eu/>. Acesso em: 15 de julho de 2022.

ISLAM, S.; MANNERING, F. Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence. **Journal of safety Research**, Elsevier, v. 37, n. 3, p. 267–276, 2006.

KHORASHADI, A. et al. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. **Accident Analysis & Prevention**, Elsevier, v. 37, n. 5, p. 910–921, 2005.

KIM, K. et al. Drivers at fault: influences of age, sex, and vehicle type. **Journal of Safety Research**, Elsevier, v. 29, n. 3, p. 171–179, 1998.

KLAPKIV, L.; KLAPKIV, J. Technological innovations in the insurance industry. Rzecznik Finansowy, Fundacja Edukacji Ubezpieczeniowej, 2017.

KNEIB, T. Beyond mean regression. **Statistical Modelling**, Sage Publications Sage India: New Delhi, India, v. 13, n. 4, p. 275–303, 2013.

LABERGE-NADEAU, C.; MAAG, U.; BOURBEAU, R. The effects of age and experience on accidents with injuries: should the licensing age be raised? **Accident Analysis & Prevention**, Elsevier, v. 24, n. 2, p. 107–116, 1992.

LAMIGUEIRO, O. P. **tdr: Target Diagram**. [S.l.], 2018. R package version 0.13. Disponível em: <https://CRAN.R-project.org/package=tdr>.

LEE, J. D. et al. Exact post-selection inference, with application to the lasso. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 44, n. 3, p. 907–927, 2016.

LEMAIRE, J. **Bonus-malus systems in automobile insurance**. [S.l.]: Springer science & business media, 1995. v. 19.

LIN, X.; ZHANG, D. Inference in generalized additive mixed models by using smoothing splines. **Journal of the royal statistical society: Series b (statistical methodology)**, Wiley Online Library, v. 61, n. 2, p. 381–400, 1999.

LINDSTROM, M. J.; BATES, D. M. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. **Journal of the American Statistical Association**, Taylor & Francis, v. 83, n. 404, p. 1014–1022, 1988.

LUCAS, E. C.; MENDES-DA-SILVA, W.; LYONS, A. C. Gender differences and automobile insurance acquisition. In: **Individual Behaviors and Technologies for Financial Innovations**. [S.l.]: Springer, 2019. p. 25–45.

MacGINNITIE, J. The actuary and his profession: growth, development, promise. Proceedings of the Casualty Actuarial Society, 1980.

MANNERING, F. L. Male/female driver characteristics and accident risk: some new evidence. **Accident Analysis & Prevention**, Elsevier, v. 25, n. 1, p. 77–84, 1993.

MARTINS, A. Profissão atuarial e seguridade social no brasil da primeira república à era vargas. **Revista Contabilidade & Finanças**, SciELO Brasil, v. 31, p. 364–377, 2020.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. [S.l.]: CRC press, 1989. v. 37.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, linear, and mixed models**. [S.l.]: Wiley series in probability and statistics, 2001. v. 1.

MEDDERS, L. A.; PARSON, J. A.; THOMAS-REID, M. Gender x and auto insurance: Is gender rating unfairly discriminatory? **Journal of Insurance Regulation**, v. 40, n. 7, 2021.

MEEBOONSALANG, W.; CHAVEESUK, S. Factors affecting customer loyalty in the automobile insurance industry in thailand. **Editorial Board**, p. 65, 2020.

MEYERS, G.; HOYWEGHEN, I. V. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. **Science as Culture**, Taylor & Francis, v. 27, n. 4, p. 413–438, 2018.

MODARRES, S. R. et al. Epidemiological characteristics of fatal traumatic accidents in babol, iran: A hospital-based survey. **Bulletin of emergency & trauma**, Trauma Research Center of Shiraz University of Medical Sciences, v. 2, n. 4, p. 146, 2014.

MOWBRAY, A. H. How extensive a payroll exposure is necessary to give a dependable pure premium. In: **Proceedings of the Casualty Actuarial society**. [S.l.: s.n.], 1914. v. 1, n. 1, p. 24–30.

NAKAMURA, L. R. et al. Modelling location, scale and shape parameters of the birnbaum-saunders generalized t distribution. **Journal of Data Science**, v. 15, n. 2, p. 221–237, 2017.

NELDER, J. Contribution to the discussion of "generalized additive models for location, scale and shape", by ra rigby and dm stasinopoulos. **Applied Statistics**, v. 54, p. 547, 2005.

NELDER, J. A.; VERRALL, R. J. Credibility theory and generalized linear models. **ASTIN Bulletin: The Journal of the IAA**, Cambridge University Press, v. 27, n. 1, p. 71–82, 1997.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, v. 135, n. 3, p. 370–384, 1972.

OHLSSON, E.; JOHANSSON, B. **Non-life insurance pricing with generalized linear models**. [S.l.]: Springer, 2010. v. 174.

OZCAN, P. et al. Open banking as a catalyst for industry transformation: Lessons learned from implementing psd2 in europe. **Swift Institute**, 2021.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, Oxford University Press, v. 58, n. 3, p. 545–554, 1971.

PINHEIRO, J.; BATES, D. **Mixed-effects models in S and S-PLUS**. [S.l.]: Springer Science & Business Media, 2006.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

RAMIRES, T. G. et al. A new semiparametric weibull cure rate model: fitting different behaviors within gamlss. **Journal of Applied Statistics**, Taylor & Francis, v. 46, n. 15, p. 2744–2760, 2019.

RAMIRES, T. G. et al. Comparison between highly complex location models and gamlss. **Entropy**, MDPI, v. 23, n. 4, p. 469, 2021.

RAMIRES, T. G. et al. Incorporating clustering techniques into gamlss. **Stats**, MDPI, v. 4, n. 4, p. 916–930, 2021.

RAMIRES, T. G. et al. Validation of stepwise-based procedure in gamlss. **Journal of Data Science**, v. 19, n. 1, p. 96–110, 2021.

RENCHER, A. C.; SCHAALJE, G. B. **Linear models in statistics**. [S.l.]: John Wiley & Sons, 2008.

RESTI, Y.; ISMAIL, N.; JAMAAN, S. H. Estimation of claim cost data using zero adjusted gamma and inverse gaussian regression models. **Journal of Mathematics and Statistics**, v. 9, n. 3, p. 186–192, 2013.

REVELLE, W. **psych: Procedures for Psychological, Psychometric, and Personality Research**. Evanston, Illinois, 2021. R package version 2.1.9. Disponível em: <https://CRAN.R-project.org/package=psych>.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005.

RIGBY, R. A.; STASINOPOULOS, D. M. Automatic smoothing parameter selection in gamlss with an application to centile estimation. **Statistical methods in medical research**, Sage Publications Sage UK: London, England, v. 23, n. 4, p. 318–332, 2014.

RIGBY, R. A. et al. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. [S.l.]: CRC press, 2019.

RINCON, L. **Introducción a la teoría del riesgo**. [S.l.]: Departamento de Matemáticas da Facultad de Ciencias UNAM Circuito Exterior de CU, 2012.

ROBINSON, G. K. That blup is a good thing: the estimation of random effects. **Statistical science**, Institute of Mathematical Statistics, v. 6, n. 1, p. 15–32, 1991.

RODGERS, J. L.; NICEWANDER, W. A. Thirteen ways to look at the correlation coefficient. **The American Statistician**, Taylor & Francis, v. 42, n. 1, p. 59–66, 1988.

ROQUIM, F. V. et al. Building flexible regression models: including the birnbaum-saunders distribution in the gamlss package. **Semina: Exact and Technological Sciences**, v. 42, n. 2, p. 163–168, 2021.

RUPPERT, D.; WAND, M. P.; CARROLL, R. J. **Semiparametric regression**. [S.l.]: Cambridge university press, 2003.

SAITO, K.; KATO, T.; SHIMANE, T. Traffic congestion and accident externality: A japan-us comparison. **The BE Journal of Economic Analysis & Policy**, De Gruyter, v. 10, n. 1, 2010.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978.

SERRANO, C. G. C. Fidelización y rentabilización de usuarios de seguros todo riesgo de vehículos por medio de la venta cruzada y la venta escalonada. un enfoque promocional para la industria aseguradora. **Universidad & Empresa**, Universidad del Rosario, v. 18, n. 30, p. 143–157, 2016.

SHERAFATI, F. et al. Risk factors of road traffic accidents associated mortality in northern iran; a single center experience utilizing oaxaca blinder decomposition. **Bulletin of Emergency & Trauma**, Trauma Research Center of Shiraz University of Medical Sciences, v. 5, n. 2, p. 116, 2017.

SIMONOFF, J. S. **Smoothing methods in statistics**. [S.l.]: Springer Science & Business Media, 2012.

SINGER, J. M.; NOBRE, J. S.; ROCHA, F. M. M. **Análise de dados longitudinais**. [S.l.]: Departamento de Estatistica da Universidade de São Paulo, 2018.

SOLEYMANIAN, M.; WEINBERG, C. B.; ZHU, T. Sensor data and behavioral tracking: Does usage-based auto insurance benefit drivers? **Marketing Science**, INFORMS, v. 38, n. 1, p. 21–43, 2019.

STAMATIADIS, N.; DEACON, J. A. Trends in highway safety: effects of an aging population on accident propensity. **Accident Analysis & Prevention**, Elsevier, v. 27, n. 4, p. 443–459, 1995.

STASINOPOULOS, M.; ENEA, M.; RIGBY, R. A. Zero adjusted distributions on the positive real line. **URL http://www. gamlss. com/wp-content/uploads/2018/01/ZeroAdjustedDistributions. pdf**, 2017.

STASINOPOULOS, M.; RIGBY, B.; DE BASTIANI, F. **gamlss.data: Data for Generalised Additive Models for Location Scale and Shape**. [S.l.], 2021. R package version 6.0-1. Disponível em: <https://CRAN.R-project.org/package=gamlss.data>.

STASINOPOULOS, M. et al. **gamlss.demo: Demos for GAMLSS**. [S.l.], 2015. R package version 4.3-3. Disponível em: <https://CRAN.R-project.org/package=gamlss.demo>.

STASINOPOULOS, M.; RIGBY, R. **gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape**. [S.l.], 2021. R package version 5.3-2. Disponível em: <https://CRAN.R-project.org/package=gamlss.dist>.

STASINOPOULOS, M. et al. Principal component regression in gamlss applied to greek–german government bond yield spreads. **Statistical Modelling**, SAGE Publications Sage India: New Delhi, India, v. 22, n. 1-2, p. 127–145, 2022.

STASINOPOULOS, M. D. et al. **Flexible Regression and Smoothing: Using GAMLSS in R**. [S.l.]: CRC Press, 2017.

SUSEP, S. de S. P. **Glossário de termos em seguros**. 2020. Disponível em: <http://www.susep.gov.br/menu/informacoes-ao-publico/glossario>. Acesso em: 15 de julho de 2020.

SUSEP, S. de S. P. **Open Insurance - Sistema de Seguros Aberto**. 2022. Disponível em: <https://openinsurance.susep.gov.br/>. Acesso em: 10 de fevereiro de 2022.

Swiss Re Institute. **sigma explorer**. 2022. Disponível em: <https://www.sigma-explorer.com/>. Acesso em: 22 de junho de 2022.

TEUGELS, J. L.; RAMSEY, H. The encyclopedia of actuarial science. Institute of Mathematics of the National Academy of Sciences of Ukraine, 2006.

TSELENTIS, D. I.; YANNIS, G.; VLAHOGIANNI, E. I. Innovative insurance schemes: pay as/how you drive. **Transportation Research Procedia**, Elsevier, v. 14, p. 362–371, 2016.

United Nations. **With 1.3 million annual road deaths, UN wants to halve number by 2030**. 2022. Disponível em: <https://news.un.org/en/story/2021/12/1107152>. Acesso em: 22 de junho de 2022.

VAN BUUREN, S.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in medicine**, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

VEEVERS, J. Women in the driver's seat: Trends in sex differences in driving and death. **Population Research and Policy Review**, Springer, v. 1, n. 2, p. 171–182, 1982.

VEEVERS, J. E.; GEE, E. M. Playing it safe: Accident mortality and gender roles. **Sociological Focus**, JSTOR, p. 349–360, 1986.

VOUDOURIS, V. et al. Modelling skewness and kurtosis with the bcpe density in gamlss. **Journal of Applied Statistics**, Taylor & Francis, v. 39, n. 6, p. 1279–1293, 2012.

WELSH, A. Book review. **Australian & New Zealand Journal of Statistics**, v. 61, n. 3, p. 392–395, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/anzs.12272>.

WEN, C.-H.; WANG, M.-J.; LAWRENCE, W. Modeling repeated choice behaviors of physical damage coverage for new car owners. In: EASTERN ASIA SOCIETY FOR TRANSPORTATION STUDIES. **Proceedings of the Eastern Asia Society for Transportation Studies Vol. 6 (The 7th International Conference of Eastern Asia Society for Transportation Studies, 2007)**. [S.l.], 2007. p. 114–114.

WHITNEY, A. W. **Theory of experience rating**. [S.l.: s.n.], 1918.

WHO, W. H. O. **Global status report on road safety 2015: Summary**. [S.l.], 2015.

WICKHAM, H.; HESTER, J. **readr: Read Rectangular Text Data**. [S.l.], 2021. R package version 2.0.1. Disponível em: <https://CRAN.R-project.org/package=readr>.

WOOD, S. N.; PYA, N.; SÄFKEN, B. Smoothing parameter and model selection for general smooth models. **Journal of the American Statistical Association**, Taylor & Francis, v. 111, n. 516, p. 1548–1563, 2016.

YAZDANI-CHARATI, J.; SIAMIAN, H.; AHMADI-BASIRI, E. Spatial analysis and geographic variation of fatal and injury crashes in mazandaran province from 2006 to 2010. **Materia socio-medica**, The Academy of Medical Sciences of Bosnia and Herzegovina, v. 26, n. 3, p. 177, 2014.

Código R

```r
# Fernanda Venturato Roquim
# AGO/2022
# Análise dos dados -Severidade do sinistro de veículos

rm(list=ls())

# Set Working Directory!!!
setwd("COLOQUE AQUI O SEU CAMINHO PARA ACESSO AOS DADOS")

# Chamando pacotes necessarios
library(readr) # para ler CSV
library(gamlss)
library(lme4)
library(psych) #funcao describe()
library(epiDisplay) #funcao tab1()
library(tdr) #função tdStats()

# Lendo os dados
dados <-read_csv("data.csv", col_types =cols(...1 =col_skip(),Car_2ndDriver_M =
    col_factor(levels =c("0","1")), Claims2 =col_skip(), Insuredcapital_content_
    re =col_skip(), Insuredcapital_continent_re =col_skip(), NClaims2 =col_skip(
    ), PolID =col_character(), Policy_PaymentMethodA =col_factor(levels =c("0",
    "1")), Policy_PaymentMethodH =col_skip(), Retention =col_skip(), Types =col_
    skip(), appartment =col_skip(), gender =col_factor(levels =c("0", "1")),
    metro_code =col_factor(levels =c("0", "1")), num_policiesC =col_skip(),
    NClaims1 =col_integer()))

# Excluindo os carros outliers
dados <-subset(dados, age_of_car_M <36)

# Balanceando apenas para as apolices que tem 5 observacoes
```

```
26 dados_b <-data.frame(subset(dados, year ==5)$PolID)
27 vetor <-dados_b$subset.dados..year....5..PolID
28 dados <-subset(dados, PolID %in% vetor)
29
30 # Corrigindo variaveis de tempo ao longo do periodo
31 vetor1s <-rep(1, length(dados$PolID))
32 dados$Age_client <-(dados$Age_client -vetor1s) +as.numeric(dados$year)
33 dados$age_of_car_M <-(dados$age_of_car_M -vetor1s) +as.numeric(dados$year)
34 dados$Client_Seniority <-(dados$Client_Seniority -vetor1s) +as.numeric(dados$
       year)
35
36 #Corrigindo as ocorrencias gratuitas (Substitui em NClaims com base no criterio
       de Claims)
37 dados$NClaims1[dados$Claims1 ==0] <-0
38
39 dados_brutos <-dados
40
41 # Separando os anos
42 dados1 <-subset(dados, year ==1)
43 dados2 <-subset(dados, year ==2)
44 dados3 <-subset(dados, year ==3)
45 dados4 <-subset(dados, year ==4)
46 dados5 <-subset(dados, year ==5)
47
48 #Separando ocorrencia de nao ocorrencia
49 dados_oc <-subset(dados, Claims1 > 0)
50 dados_ze <-subset(dados, Claims1 ==0)
51
52 # Excluindo o 5 ano dos dados principais
53 dados <-subset(dados, year <5)
54
55 # Ocorrencias ano a ano
56 dados_oc1 <-subset(dados1, Claims1 > 0)
57 dados_oc2 <-subset(dados2, Claims1 > 0)
```

```
58 dados_oc3 <-subset(dados3, Claims1 > 0)
59 dados_oc4 <-subset(dados4, Claims1 > 0)
60 dados_oc5 <-subset(dados5, Claims1 > 0)
61 #============================================================================
62 #Analise Exploratoria
63
64 describe(dados) # anos de 1 a 4
65 describe(dados_brutos) #dados brutos inclui o ano 5
66
67 #Ano a ano
68 describe(dados_oc1$Claims1)
69 describe(dados_oc2$Claims1)
70 describe(dados_oc3$Claims1)
71 describe(dados_oc4$Claims1)
72 describe(dados_oc5$Claims1)
73
74
75 #Variavel Resposta
76 hist(dados$Claims1) #a densidade de zeros atrapalha muito a visualizacao
77 boxplot(dados$Claims1)
78 plot(density(dados$Claims1))
79
80 #Fazendo alguns cortes para entender melhor as ocorrencias
81 hist(dados_oc$Claims1, breaks =300, xlim =c(0,5000))
82
83 plot(density(dados_oc$Claims1), xlim =c(0,5000), xlab ="Valor do sinistro", ylab
       =
84       "Densidade", main ="", ylim=c(0,0.014))
85
86 nrow(dados[dados$Claims1>5000, ]) #quantas obs foram omitidas
87
88 table(dados_oc$Claims1) #contagem da freq em 882,00
89 boxplot(dados_oc$Claims1, las =1)
90
```

```
91  #---------
92  # Covariaveis Continuas
93
94  #Idade do cliente
95  par(mfrow=c(3,1))
96  hist(dados$Age_client, xlim =range(20,100), breaks =30)
97  hist(dados_oc$Age_client, xlim =range(20,100), breaks =30, main =)
98  hist(dados_ze$Age_client, xlim =range(20,100), breaks =30)
99
100 par(mfrow=c(1,3))
101 boxplot(dados$Age_client, main ="Idade do cliente total", ylim =c(20,100))
102 boxplot(dados_oc$Age_client, main ="Idade do cliente -obs com ocorrência", ylim
        =c(20,100))
103 boxplot(dados_ze$Age_client, main ="Idade do cliente -obs sem ocorrência", ylim
        =c(20,100))
104
105 #Idade do veiculo
106 par(mfrow=c(3,1))
107 hist(dados$age_of_car_M, xlim =range(0,35), breaks =30)
108 hist(dados_oc$age_of_car_M, xlim =range(0,35), breaks =30)
109 hist(dados_ze$age_of_car_M, xlim =range(0,35), breaks =30)
110
111 par(mfrow=c(1,3))
112 boxplot(dados$age_of_car_M, main ="Idade do veículo total", ylim =c(0,35))
113 boxplot(dados_oc$age_of_car_M, main ="Idade do veiculo -obs com ocorrência",
        ylim =c(0,35))
114 boxplot(dados_ze$age_of_car_M, main ="Idade do veículo -obs sem ocorrência",
        ylim =c(0,35))
115
116 #Potencia do veículo
117 par(mfrow=c(3,1))
118 hist(dados$Car_power_M, xlim =range(5,560), breaks =30)
119 hist(dados_oc$Car_power_M, xlim =range(5,560), breaks =20)
120 hist(dados_ze$Car_power_M, xlim =range(5,560), breaks =30)
```

```
121
122 par(mfrow=c(1,3))
123 boxplot(dados$Car_power_M, main ="Potencia do veículo total", ylim =c(5,560))
124 boxplot(dados_oc$Car_power_M, main ="Potencia do veiculo -obs com ocorrência",
        ylim =c(5,560))
125 boxplot(dados_ze$Car_power_M, main ="Potencia do veículo -obs sem ocorrência",
        ylim =c(5,560))
126
127 #Fidelidade
128 par(mfrow=c(3,1))
129 hist(dados$Client_Seniority, xlim =range(5,50), breaks =30)
130 hist(dados_oc$Client_Seniority, xlim =range(5,50), breaks =20)
131 hist(dados_ze$Client_Seniority, xlim =range(5,50), breaks =30)
132
133 par(mfrow=c(1,3))
134 boxplot(dados$Client_Seniority, main ="Fidelidade total", ylim =c(5,50))
135 boxplot(dados_oc$Client_Seniority, main ="Fidelidade -obs com ocorrência", ylim
        =c(5,50))
136 boxplot(dados_ze$Client_Seniority, main ="Fidelidade -obs sem ocorrência", ylim
        =c(5,50))
137
138
139 #--------------------
140 #Covariaveis discretas ou qualitativas
141
142 #Sexo: 0=homem, 1=mulher
143 tab1(dados$gender)
144 tab1(dados_oc$gender)
145 tab1(dados_ze$gender)
146
147 # Ano
148 tab1(dados$year)
149 tab1(dados_oc$year)
150 tab1(dados_ze$year)
```

```
151
152 # Presenca de segundo motorista
153 tab1(dados$Car_2ndDriver_M)
154 tab1(dados_oc$Car_2ndDriver_M)
155 tab1(dados_ze$Car_2ndDriver_M)
156
157 # Regiao: 0=rural, 1=urbano
158 tab1(dados$metro_code)
159 tab1(dados_oc$metro_code)
160 tab1(dados_ze$metro_code)
161
162 #Forma de Pagto: 0=Mensal, 1=Anual
163 tab1(dados$Policy_PaymentMethodA)
164 tab1(dados_oc$Policy_PaymentMethodA)
165 tab1(dados_ze$Policy_PaymentMethodA)
166
167 #--------
168 #Relacionamento entre variaveis
169 par(mfrow=c(2,2))
170 plot(dados_oc$Age_client, dados_oc$Claims1, pch=20, ylab="Valor do sinistro",
        xlab ="Idade do Cliente")
171 plot(dados_oc$age_of_car_M, dados_oc$Claims1, pch=20, ylab="", xlab="Idade do ve
        ículo")
172 plot(dados_oc$Car_power_M, dados_oc$Claims1, pch=20, ylab="Valor do sinistro",
        xlab="Potência")
173 plot(dados_oc$Client_Seniority, dados_oc$Claims1, pch=20, ylab="", xlab="
        Fidelidade")
174
175 par(mfrow=c(1,4))
176 plot(dados_oc$gender, dados_oc$Claims1, pch =20, ylab="Valor do sinistro", xlab
        ="Sexo", ylim=c(0,5000))
177 plot(dados_oc$Car_2ndDriver_M, dados_oc$Claims1, pch=20, ylab="", xlab="Presença
         de 2° motorista", ylim=c(0,5000))
```

```
178 plot(dados_oc$metro_code, dados_oc$Claims1, pch=20, ylab="", xlab="Região", ylim
    =c(0,5000))
179 plot(dados_oc$Policy_PaymentMethodA, dados_oc$Claims1, pch=20, ylab="", xlab="
    Forma de Pagto", ylim=c(0,5000))
180
181 covariaveis <-data.frame(dados$Age_client, dados$age_of_car_M, dados$Car_power_M
    , dados$Client_Seniority)
182
183 cor(covariaveis, method ="pearson")
184
185 #============================================================================
186 #Potenciais distribuicoes
187
188 marginal <-fitDist(dados_oc$Claims1, type ="realplus")
189 marginal$fits
190
191 #Testei em um script parelelo e nao houve ganho de adequacao para uso de uma
    dist mais complexa
192 #============================================================================
193
194 #Modelos marginais
195 ZAGAfixo <-gamlss(Claims1~1, family=ZAGA, data =dados, method =RS(500))
196 ZAIGfixo <-gamlss(Claims1~1, family=ZAIG, data =dados, method =RS(500))
197
198 #GLM -ZAGA
199 GLM_ZAGAselec <-stepGAIC(ZAGAfixo, scope=list(lower=~1, upper= ~gender +Age_
    client +
200                                               +age_of_car_M +Car_power_M
201                                             +Car_2ndDriver_M
202                                             +metro_code +Policy_PaymentMethodA
203                                             +Client_Seniority),
204                          trace =T, data=dados, what ="mu")
205 summary(GLM_ZAGAselec)
206 GLM_ZAGA <-gamlss(formula =Claims1 ~age_of_car_M +Policy_PaymentMethodA +
```

```
207                 Car_2ndDriver_M, family =ZAGA, data =dados, trace =T)
208
209 #GLM -ZAIG
210 GLM_ZAIGselec <-stepGAIC(ZAIGfixo, scope=list(lower=~1, upper= ~gender +Age_
        client +
211                                             +age_of_car_M +Car_power_M
212                                           +Car_2ndDriver_M
213                                           +metro_code +Policy_PaymentMethodA
214                                           +Client_Seniority),
215                         trace =T, data=dados, what ="mu")
216 summary(GLM_ZAIGselec)
217 GLM_ZAIG <-gamlss(formula =Claims1 ~Car_power_M, family =ZAIG,
218                 data =dados, trace =T)
219
220 #GLMM -ZAGA
221 GLMM_ZAGA <-gamlss(formula =Claims1 ~age_of_car_M +Policy_PaymentMethodA +
222                   Car_2ndDriver_M +re(random=~1|PolID), family =ZAGA, data =
                        dados, trace =T)
223
224 #GLMM -ZAIG (Nao converge)
225 # GLMM_ZAIG <-gamlss(formula =Claims1 ~Car_power_M +re(random=~1|PolID), family
        =ZAIG,
226 # data =dados, trace =T, gd.tol=Inf, method =CG(50),
227 # n.cycs =100)
228
229 #GAM -ZAGA
230 GAM_ZAGAselec <-stepGAIC(ZAGAfixo, scope=list(lower=~1, upper= ~gender +pb(Age_
        client) +
231                                             +pb(age_of_car_M) +pb(Car_power_M)
232                                           +Car_2ndDriver_M
233                                           +metro_code +Policy_PaymentMethodA
234                                           +pb(Client_Seniority)),
235                         trace =T, data=dados, what ="mu")
236 summary(GAM_ZAGAselec)
```

```
237 term.plot(GAM_ZAGAselec)
238 edfAll(GAM_ZAGAselec) #Manter pb()
239 GAM_ZAGA <-gamlss(formula =Claims1 ~pb(Age_client), family =ZAGA,
240                  data =dados, method =RS(500), trace =T)
241
242 # GAM -ZAIG
243 # GAM_ZAIGselec <-stepGAIC(ZAIGfixo, scope=list(lower=~1, upper= ~gender +pb(
        Age_client) +
244 # +pb(age_of_car_M) +pb(Car_power_M)
245 # +Car_2ndDriver_M
246 # +metro_code +Policy_PaymentMethodA
247 # +pb(Client_Seniority)),
248 # trace =T, data=dados, what ="mu", method =RS(500))
249 #
250 # summary(GAM_ZAIGselec) #Nao eh GAM
251
252 # GAMM -ZAGA (Nao converge)
253 # GAMM_ZAGA <-gamlss(formula =Claims1 ~pb(Age_client) +re(random=~1|PolID),
254 # family =ZAGA, data =dados, trace =T, method =RS(100), mu.start=1000)
255
256 # GAMM -ZAIG (Nao eh GAM)
257
258 #GAMLSS -ZAGA
259 GAMLSS_ZAGAselec <-stepGAICAll.A(ZAGAfixo, scope=list(lower=~1, upper= ~gender
        +pb(Age_client) +
260                                                  +pb(age_of_car_M) +pb(Car_power_M)
261                                               +Car_2ndDriver_M
262                                               +metro_code +Policy_PaymentMethodA
263                                               +pb(Client_Seniority)),
264                          trace =T, data=dados)
265 summary(GAMLSS_ZAGAselec)
266 term.plot(GAMLSS_ZAGAselec, what ="mu")
267 term.plot(GAMLSS_ZAGAselec, what ="nu")
268 edfAll(GAMLSS_ZAGAselec) #proximo de 2, pb() desnecessario
```

```
GAMLSS_ZAGA <-gamlss(formula =Claims1 ~pb(Age_client), sigma.formula =~1,
               nu.formula =~pb(Age_client) +pb(Client_Seniority) +
                 Car_2ndDriver_M +age_of_car_M +Policy_PaymentMethodA,
               family =ZAGA, data =dados, trace =T, method =RS(100))


#GAMLSS -ZAIG
GAMLSS_ZAIGselec <-stepGAICAll.A(ZAIGfixo, scope=list(lower=~1, upper= ~gender
    +pb(Age_client) +
                                          +pb(age_of_car_M) +pb(Car_power_M)
                                        +Car_2ndDriver_M
                                        +metro_code +Policy_PaymentMethodA
                                        +pb(Client_Seniority)),
                         trace =2, data=dados, method =RS(500))
summary(GAMLSS_ZAIGselec)
term.plot(GAMLSS_ZAIGselec, what ="sigma")
term.plot(GAMLSS_ZAIGselec, what ="nu")
edfAll(GAMLSS_ZAIGselec)
GAMLSS_ZAIG <-gamlss(formula =Claims1 ~Car_2ndDriver_M +Policy_PaymentMethodA,
                     sigma.formula =~pb(Client_Seniority) +pb(Age_client) +
                       Car_2ndDriver_M +Policy_PaymentMethodA +pb(age_of_car_M) +
                       pb(Car_power_M), nu.formula =~pb(Age_client) +
                       pb(Client_Seniority) +Car_2ndDriver_M +age_of_car_M +
                       Policy_PaymentMethodA, family =ZAIG, data =dados, trace =T
                           )

#GAMMLSS -ZAGA
GAMMLSS_ZAGA <-gamlss(formula =Claims1 ~Age_client +re(random=~1|PolID), sigma.
    formula =~1,
                      nu.formula =~pb(Age_client) +pb(Client_Seniority) +
                        Car_2ndDriver_M +age_of_car_M +Policy_PaymentMethodA +re(
                            random=~1|PolID),
                      family =ZAGA, data =dados, trace =T, method =RS(500))

```

```
#GAMMLSS -ZAIG
GAMMLSS_ZAIG <-gamlss(formula =Claims1 ~Car_2ndDriver_M +Policy_PaymentMethodA,
                      sigma.formula =~pb(Client_Seniority) +pb(Age_client) +
                        Car_2ndDriver_M +Policy_PaymentMethodA +pb(age_of_car_M)
                            +
                        pb(Car_power_M), nu.formula =~pb(Age_client) +
                        pb(Client_Seniority) +Car_2ndDriver_M +age_of_car_M +
                        Policy_PaymentMethodA +re(random=~1|PolID)
                      , family =ZAIG, data =dados, trace =T,
                      method =RS(100), gd.tol=Inf, mu.start =1000)

# Total de 8 modelos de 12 idealizados
#-----------------------------------------------------------------------
#Resumos

summary(GLM_ZAGA)
summary(GLM_ZAIG)
summary(GLMM_ZAGA)
summary(GAM_ZAGA)
summary(GAMLSS_ZAGA)
summary(GAMLSS_ZAIG)
summary(GAMMLSS_ZAGA)
summary(GAMMLSS_ZAIG)

#-----------------------------------------------------------------------
#Adequacao
par(mfrow=c(4,2))
rqres.plot(GLM_ZAGA, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
          plot.type ="all"); title(main ="GLM_ZAGA")
rqres.plot(GLM_ZAIG, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
          plot.type ="all"); title(main ="GLM_ZAIG")
rqres.plot(GLMM_ZAGA, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
          plot.type ="all"); title(main ="GLMM_ZAGA")
rqres.plot(GAM_ZAGA, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
```

```
332           plot.type ="all"); title(main ="GAM_ZAGA")
333 rqres.plot(GAMLSS_ZAGA, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
334           plot.type ="all"); title(main ="GAMLSS_ZAGA")
335 rqres.plot(GAMLSS_ZAIG, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
336           plot.type ="all"); title(main ="GAMLSS_ZAIG")
337 rqres.plot(GAMMLSS_ZAGA, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
338           plot.type ="all"); title(main ="GAMMLSS_ZAGA")
339 rqres.plot(GAMMLSS_ZAIG, howmany =4, cex=.5, pch=20, ylim.all=0.7, xlim.all =5,
340           plot.type ="all"); title(main ="GAMMLSS_ZAIG")
341
342 #-------------------------------------------------------------------
343 # Capacidade preditiva
344
345 # Rearranjando o quinto ano para leitura em Predict
346 dados_pred <-dados5
347 observados <-dados_pred$Claims1
348 dados_pred$Claims1 <-NULL
349 sum(observados) #perda agregada real
350
351 # Obtendo as predicoes
352 predGLM_ZAGA <-data.frame(predictAll(GLM_ZAGA, newdata =dados_pred, type ="
        response"))
353 predGLM_ZAIG <-data.frame(predictAll(GLM_ZAIG, newdata =dados_pred, type ="
        response"))
354 predGLMM_ZAGA <-data.frame(predictAll(GLMM_ZAGA, newdata =dados_pred, type ="
        response"))
355 predGAM_ZAGA <-data.frame(predictAll(GAM_ZAGA, newdata =dados_pred, type ="
        response"))
356 predGAMLSS_ZAGA <-data.frame(predictAll(GAMLSS_ZAGA, newdata =dados_pred, type =
        "response"))
357 predGAMLSS_ZAIG <-data.frame(predictAll(GAMLSS_ZAIG, newdata =dados_pred, type =
        "response"))
358 predGAMMLSS_ZAGA <-data.frame(predictAll(GAMMLSS_ZAGA, newdata =dados_pred, type
         ="response"))
```

```
predGAMMLSS_ZAIG <-data.frame(predictAll(GAMMLSS_ZAIG, newdata =dados_pred, type
    ="response"))

# Obtendo o valor esperado (formula das zero ajustadas)
preditoGLM_ZAGA <-(1 -predGLM_ZAGA$nu) * predGLM_ZAGA$mu
preditoGLM_ZAIG <-(1 -predGLM_ZAIG$nu) * predGLM_ZAIG$mu
preditoGLMM_ZAGA <-(1 -predGLMM_ZAGA$nu) * predGLMM_ZAGA$mu
preditoGAM_ZAGA <-(1 -predGAM_ZAGA$nu) * predGAM_ZAGA$mu
preditoGAMLSS_ZAGA <-(1 -predGAMLSS_ZAGA$nu) * predGAMLSS_ZAGA$mu
preditoGAMLSS_ZAIG <-(1 -predGAMLSS_ZAIG$nu) * predGAMLSS_ZAIG$mu
preditoGAMMLSS_ZAGA <-(1 -predGAMMLSS_ZAGA$nu) * predGAMMLSS_ZAGA$mu
preditoGAMMLSS_ZAIG <-(1 -predGAMMLSS_ZAIG$nu) * predGAMMLSS_ZAIG$mu

# Perda agregada predita
sum(preditoGLM_ZAGA)
sum(preditoGLM_ZAIG)
sum(preditoGLMM_ZAGA)
sum(preditoGAM_ZAGA)
sum(preditoGAMLSS_ZAGA)
sum(preditoGAMLSS_ZAIG)
sum(preditoGAMMLSS_ZAGA)
sum(preditoGAMMLSS_ZAIG)

#Acuracia (mbe =vies medio do erro, mae =media do erro absoluto)
#Precisao (rmse =raiz quadrada media do erro)

tdStats(preditoGLM_ZAGA, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGLM_ZAIG, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGLMM_ZAGA, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGAM_ZAGA, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGAMLSS_ZAGA, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGAMLSS_ZAIG, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGAMMLSS_ZAGA, observados, functions =c("mbe", "mae", "rmse"))
tdStats(preditoGAMMLSS_ZAIG, observados, functions =c("mbe", "mae", "rmse"))
```

```
392
393 #-----------------------------------------------------------------------
394 # Mais detalhes dos modelos GAMLSS ZAIG
395
396 plot(GAMMLSS_ZAIG) #leva em consideracao apenas mu
397
398 #Avaliacao de pb
399 par(mfrow=c(3,2))
400
401 term.plot(GAMMLSS_ZAIG, what ="sigma", ylim ="free", terms =1, scheme ="lines",
        se=F,
402           xlab ="Fidelidade", ylab =expression(paste("pb(Fidelidade) em ", sigma
                ,)))
403 grid(nx =NULL, ny =NA, lty =3, col ="gray", lwd =1)
404 title(main ="I")
405
406 term.plot(GAMMLSS_ZAIG, what ="sigma", ylim ="free", terms =2, scheme ="lines",
        se=F,
407           xlab ="Idade do cliente", ylab =expression(paste("pb(Idade do cliente)
                em ", sigma,)))
408 grid(nx =NULL, ny =NA, lty =3, col ="gray", lwd =1)
409 title(main ="II")
410
411 term.plot(GAMMLSS_ZAIG, what ="sigma", ylim ="free", terms =5, scheme ="lines",
        se=F,
412           xlab ="Idade do veículo", ylab =expression(paste("pb(Idade do veículo)
                em ", sigma,)))
413 grid(nx =NULL, ny =NA, lty =3, col ="gray", lwd =1)
414 title(main ="III")
415
416 term.plot(GAMMLSS_ZAIG, what ="sigma", ylim ="free", terms =6, scheme ="lines",
        se=F,
417           xlab ="Potência", ylab =expression(paste("pb(Potência) em ", sigma,)))
418 grid(nx =NULL, ny =NA, lty =3, col ="gray", lwd =1)
```

```
419 title(main ="IV")
420
421 term.plot(GAMMLSS_ZAIG, what ="nu", ylim ="free", terms =1, scheme ="lines", se=
        F,
422          xlab ="Idade do cliente", ylab =expression(paste("pb(Idade do cliente)
                em ", nu,)))
423 grid(nx =NULL, ny =NA, lty =3, col ="gray", lwd =1)
424 title(main ="V")
425
426 term.plot(GAMMLSS_ZAIG, what ="nu", ylim ="free", terms =2, scheme ="lines", se=
        F,
427          xlab ="Fidelidade", ylab =expression(paste("pb(Fidelidade) em ", nu,)))
428 grid(nx =NULL, ny =NA, lty =3, col ="gray", lwd =1)
429 title(main ="VI")
430
431 #Avaliando os efeitos aleatorios
432 summary(getSmo(GAMMLSS_ZAIG, what ="nu", which =3))
433 plot(getSmo(GAMMLSS_ZAIG, what ="nu", which =3))
434 ranef <-ranef(getSmo(GAMMLSS_ZAIG, what ="nu", which =3))$`(Intercept)`
435 plot(ranef, xlab ="ID da apólice", ylab ="Estimativa do efeito aleatório", main
        ="")
436 plot(density(ranef), ylab ="Densidade", xlab ="Estimativa do efeito aleatório",
        main="")
437 ks.test(ranef, 'pnorm') #eh normal
438 summary(ranef)
439 sd(ranef)
440
441 #Densidade das predicoes
442 par(mfrow=c(1,3))
443 plot(density(observados, bw=1), main ="Observado no ano 5", ylab ="Densidade",
        xlab ="Valor do sinistro")
444 rug(jitter(observados))
445 plot(density(preditoGAMLSS_ZAIG, bw=1), main ="Predito por GAMLSS_ZAIG", ylab ="
        ", xlab ="Valor do sinistro")
```

```
446 rug(jitter(preditoGAMLSS_ZAIG))
447 plot(density(preditoGAMMLSS_ZAIG, bw=1), main ="Predito por GAMMLSS_ZAIG", ylab
    =" ", xlab ="Valor do sinistro")
448 rug(jitter(preditoGAMMLSS_ZAIG))
449
450 #Comportamento do modelo onde houve e onde nao houve sinistro no ano 5
451 dados_sim <-dados_oc5
452 obs_sim <-dados_sim$Claims1
453 dados_sim$Claims1 <-NULL
454
455 predGAMLSS_ZAIG_sim <-data.frame(predictAll(GAMLSS_ZAIG, newdata =dados_sim,
    type ="response", output ="matrix"))
456 preditoGAMLSS_ZAIG_sim <-(1 -predGAMLSS_ZAIG_sim$nu) * predGAMLSS_ZAIG_sim$mu
457 predGAMMLSS_ZAIG_sim <-data.frame(predictAll(GAMMLSS_ZAIG, newdata =dados_sim,
    type ="response", output ="matrix"))
458 preditoGAMMLSS_ZAIG_sim <-(1 -predGAMMLSS_ZAIG_sim$nu) * predGAMMLSS_ZAIG_sim$mu
459
460 dados_nao <-subset(dados5, Claims1 ==0)
461 obs_nao <-dados_nao$Claims1
462 dados_nao$Claims1 <-NULL
463
464 predGAMLSS_ZAIG_nao <-data.frame(predictAll(GAMLSS_ZAIG, newdata =dados_nao,
    type ="response", output ="matrix"))
465 preditoGAMLSS_ZAIG_nao <-(1 -predGAMLSS_ZAIG_nao$nu) * predGAMLSS_ZAIG_nao$mu
466 predGAMMLSS_ZAIG_nao <-data.frame(predictAll(GAMMLSS_ZAIG, newdata =dados_nao,
    type ="response", output ="matrix"))
467 preditoGAMMLSS_ZAIG_nao <-(1 -predGAMMLSS_ZAIG_nao$nu) * predGAMMLSS_ZAIG_nao$mu
468
469 tdStats(preditoGAMLSS_ZAIG_sim, obs_sim, functions =c("mbe", "mae", "rmse"))
470 tdStats(preditoGAMMLSS_ZAIG_sim, obs_sim, functions =c("mbe", "mae", "rmse"))
471 tdStats(preditoGAMLSS_ZAIG_nao, obs_nao, functions =c("mbe", "mae", "rmse"))
472 tdStats(preditoGAMMLSS_ZAIG_nao, obs_nao, functions =c("mbe", "mae", "rmse"))
473
474 #-----------------------------------------------------------------------
```

```
475 # Estudando alguns casos particulares
476
477 #Grafico ano 5
478 par(mfrow=c(1,1))
479 percentil90 <-quantile(dados_oc5$Claims1, 0.90)
480 percentil80 <-quantile(dados_oc5$Claims1, 0.80)
481 plot(obs_sim, xlab="Índice da apólice", ylab="Valor observado", xaxt="n", pch=
        20,
482      col =ifelse(obs_sim <percentil80, "black", ifelse(obs_sim > percentil90, "
            blue", "red")))
483 abline(percentil90, 0, col="gray", lty=3)
484 abline(percentil80, 0, col="gray", lty=3)
485 legend("topright", legend=c("Superior ao percentil 90", "Entre os percentis 80 e
         90"),
486        col=c("blue", "red"), pch=20, cex=0.8)
487
488 #-----------------------------------------------------------------------
489 #Analise de outliers (acima do percentil 90) no ano 5.
490 corte90 <-subset(dados5, Claims1 > percentil90) #13 observacoes
491 dados_corte90 <-subset(dados_brutos, PolID %in% corte90$PolID)
492 plot(dados_corte90$Claims1)
493 dados90 <-dados_corte90
494 dados90$Claims1 <-NULL
495
496 mean(corte90$Age_client)
497 mean(corte90$age_of_car_M)
498 mean(corte90$Client_Seniority)
499 tab1(corte90$Car_2ndDriver_M)
500 tab1(corte90$Policy_PaymentMethodA)
501
502 predGAMLSS_ZAIG_90 <-data.frame(predictAll(GAMLSS_ZAIG, newdata =dados90, type =
        "response", output ="matrix"))
503 preditoGAMLSS_ZAIG_90 <-(1 -predGAMLSS_ZAIG_90$nu) * predGAMLSS_ZAIG_90$mu
```

```
504 predGAMMLSS_ZAIG_90 <-data.frame(predictAll(GAMMLSS_ZAIG, newdata =dados90, type
        ="response", output ="matrix"))
505 preditoGAMMLSS_ZAIG_90 <-(1 -predGAMMLSS_ZAIG_90$nu) * predGAMMLSS_ZAIG_90$mu
506
507 dados_corte90$GAMLSS <-preditoGAMLSS_ZAIG_90
508 dados_corte90$GAMMLSS <-preditoGAMMLSS_ZAIG_90
509
510 par(mfrow=c(1,3)) #spaghetti plot
511 interaction.plot(dados_corte90$year, dados_corte90$PolID, main ="Observado", las
        =1, col=c(1,2,3,4,6),
512                  lty =1:3, dados_corte90$Claims1, xlab="Ano", ylab="Valor do
                         sinistro", legend=F)
513 interaction.plot(dados_corte90$year, dados_corte90$PolID, main ="GAMLSS", las=1,
         col=c(1,2,3,4,6),
514                  lty =1:3, dados_corte90$GAMLSS, xlab="Ano", ylab="Valor do
                         sinistro", legend=F)
515 interaction.plot(dados_corte90$year, dados_corte90$PolID, main ="GAMMLSS", las=
        1, col=c(1,2,3,4,6),
516                  lty =1:3, dados_corte90$GAMMLSS, xlab="Ano", ylab="Valor do
                         sinistro", legend=F)
517
518 tdStats(preditoGAMLSS_ZAIG_90, corte90$Claims1, functions =c("mbe", "mae", "rmse
        "))
519 tdStats(preditoGAMMLSS_ZAIG_90, corte90$Claims1, functions =c("mbe", "mae", "
        rmse")) #melhor*
520
521 #-------------------------------------------------------------------------
522 #Analise dentre percentil 80-90 no ano 5.
523 corte8090 <-subset(dados5, Claims1 >= percentil80 & Claims1 <percentil90) #12
        observacoes
524 dados_corte8090 <-subset(dados_brutos, PolID %in% corte8090$PolID)
525 plot(dados_corte8090$Claims1)
526 dados8090 <-dados_corte8090
527 dados8090$Claims1 <-NULL
```

```
528
529 mean(corte8090$Age_client)
530 mean(corte8090$age_of_car_M)
531 mean(corte8090$Client_Seniority)
532 tab1(corte8090$Car_2ndDriver_M)
533 tab1(corte8090$Policy_PaymentMethodA)
534
535
536 predGAMLSS_ZAIG_8090 <-data.frame(predictAll(GAMLSS_ZAIG, newdata =dados8090,
        type ="response", output ="matrix"))
537 preditoGAMLSS_ZAIG_8090 <-(1 -predGAMLSS_ZAIG_8090$nu) * predGAMLSS_ZAIG_8090$mu
538 predGAMMLSS_ZAIG_8090 <-data.frame(predictAll(GAMMLSS_ZAIG, newdata =dados8090,
        type ="response", output ="matrix"))
539 preditoGAMMLSS_ZAIG_8090 <-(1 -predGAMMLSS_ZAIG_8090$nu) * predGAMMLSS_ZAIG_8090
        $mu
540
541 dados_corte8090$GAMLSS <-preditoGAMLSS_ZAIG_8090
542 dados_corte8090$GAMMLSS <-preditoGAMMLSS_ZAIG_8090
543
544 par(mfrow=c(1,3)) #spaghetti plot
545 interaction.plot(dados_corte8090$year, dados_corte8090$PolID, main ="Observado",
         las=1, col=c(1,2,3,4,6),
546                  lty =1:3, dados_corte8090$Claims1, xlab="Ano", ylab="Valor do
                       sinistro", legend=F)
547 interaction.plot(dados_corte8090$year, dados_corte8090$PolID, main ="GAMLSS",
        las=1, col=c(1,2,3,4,6),
548                  lty =1:3, dados_corte8090$GAMLSS, xlab="Ano", ylab="Valor do
                       sinistro", legend=F)
549 interaction.plot(dados_corte8090$year, dados_corte8090$PolID, main ="GAMMLSS",
        las=1, col=c(1,2,3,4,6),
550                  lty =1:3, dados_corte8090$GAMMLSS, xlab="Ano", ylab="Valor do
                       sinistro", legend=F)
551
```

```
552 tdStats(preditoGAMLSS_ZAIG_8090, dados_corte8090$Claims1, functions =c("mbe", "
    mae", "rmse"))
553 tdStats(preditoGAMMLSS_ZAIG_8090, dados_corte8090$Claims1, functions =c("mbe", "
    mae", "rmse")) #melhor*
554
555
556 #-------------------------------------------------------------------------
557 #Amostrando 10 apolices que nunca acionaram seguros
558
559 dados_zeb <-data.frame(subset(dados_ze, year ==5)$PolID)
560 vetor_ze <-dados_zeb$subset.dados_ze..year....5..PolID
561 set.seed(50)
562 amostra <-sample(vetor_ze, 10, replace =FALSE)
563 dados0 <-subset(dados_ze, PolID %in% amostra)
564
565 mean(dados0$Age_client)
566 mean(dados0$age_of_car_M)
567 mean(dados0$Client_Seniority)
568 tab1(dados0$Car_2ndDriver_M)
569 tab1(dados0$Policy_PaymentMethodA)
570
571 predGAMLSS_ZAIG_0 <-data.frame(predictAll(GAMLSS_ZAIG, newdata =dados0, type ="
    response"))
572 preditoGAMLSS_ZAIG_0 <-(1 -predGAMLSS_ZAIG_0$nu) * predGAMLSS_ZAIG_0$mu
573 predGAMMLSS_ZAIG_0 <-data.frame(predictAll(GAMMLSS_ZAIG, newdata =dados0, type =
    "response"))
574 preditoGAMMLSS_ZAIG_0 <-(1 -predGAMMLSS_ZAIG_0$nu) * predGAMMLSS_ZAIG_0$mu
575
576 dados0$GAMLSS <-preditoGAMLSS_ZAIG_0
577 dados0$GAMMLSS <-preditoGAMMLSS_ZAIG_0
578
579 par(mfrow=c(1,2)) #spaghetti plot
580 interaction.plot(dados0$year, dados0$PolID, main ="GAMLSS", las=1, col=c(
    1,2,3,4,6),
```

```
581                 lty =1:3, dados0$GAMLSS, xlab="Ano", ylab="Valor do sinistro",
                        legend=F)
582 interaction.plot(dados0$year, dados0$PolID, main ="GAMMLSS", las=1, col=c(
    1,2,3,4,6),
583                 lty =1:3, dados0$GAMMLSS, xlab="Ano", ylab="Valor do sinistro",
                        legend=F)
584
585 tdStats(preditoGAMLSS_ZAIG_0, dados0$Claims1, functions =c("mbe", "mae", "rmse")
    )
586 tdStats(preditoGAMMLSS_ZAIG_0, dados0$Claims1, functions =c("mbe", "mae", "rmse"
    )) #melhor*
```

ScriptTese_V8.R