NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

SARA LOPES DE OLIVEIRA

Licenciada em Matemática

ÁRVORES DE REGRESSÃO NO CONTEXTO DA TARIFAÇÃO *A PRIORI* DO SEGURO AUTOMÓVEL

MESTRADO EM MATEMÁTICA E APLICAÇÕES
RAMO ATUARIADO, ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Universidade NOVA de Lisboa
Novembro, 2021

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

ÁRVORES DE REGRESSÃO NO CONTEXTO DA TARIFAÇÃO *A PRIORI* DO SEGURO AUTOMÓVEL

SARA LOPES DE OLIVEIRA

Licenciada em Matemática

Orientadora: Gracinda Rita Guerreiro
Professora Associada, Faculdade de Ciências e Tecnologia

Coorientadora: Regina Bispo
Professora Associada, Faculdade de Ciências e Tecnologia

Júri:

Presidente: Maria de Lourdes Afonso
Professora Associada, Faculdade de Ciências e Tecnologia

Arguente: Pedro Alexandre Sousa
Professor Associado, Faculdade de Ciências e Tecnologia

Orientadora: Gracinda Rita Guerreiro
Professora Associada, Faculdade de Ciências e Tecnologia

MESTRADO EM MATEMÁTICA E APLICAÇÕES
RAMO ATUARIADO, ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Universidade NOVA de Lisboa
Novembro, 2021

Árvores de Regressão no contexto da Tarifação *a priori* do Seguro Automóvel

Aos meus.

Agradecimentos

Atípico foi o ambiente para o desenvolvimento desta dissertação, consequência da pandemia provocada pela Covid-19. Ainda assim, o processo de elaboração da mesma foi pautado pelo apoio e pela constante superação individual.

Às minhas orientadoras, a quem dirijo uma palavra de apreço pela incansável orientação que me concederam. À Professora Regina Bispo pelo voto de confiança nas minhas capacidades para me sugerir o tema da dissertação. À Professora Gracinda Rita Guerreiro pela amabilidade e disponibilidade demonstradas para clarificar as minhas dúvidas. A simpatia, o rigor e o encorajamento revelados pelas Professoras levaram à conclusão deste trabalho. Um obrigada não chega para expressar a gratidão que sinto.

Aos meus familiares, especialmente aos meus pais e ao meu irmão, agradeço a paciência, o carinho e o entusiasmo imprescindíveis durante a elaboração deste projeto. Sem eles, estou certa que teria sido mais difícil.

Resumo

A construção de uma tarifa tem sido abordada de uma perspetiva mais conservadora, recorrendo a Modelos Lineares Generalizados. No entanto, o crescente interesse pela análise de dados tem motivado a procura de abordagens alternativas para o problema da tarifação, nomeadamente métodos de *Machine Learning*. Estes métodos estão na vanguarda da modelação e apresentam-se como um forte competidor aos Modelos Lineares Generalizados em termos de resultados.

Nesta dissertação apresenta-se uma abordagem alternativa à modelação da tarifa do Seguro Automóvel, recorrendo a Árvores de Regressão. Para efeitos comparativos, considerou-se também a modelação através de Modelos Lineares Generalizados. Os fatores de risco contínuos (idade do condutor e idade do veículo) foram categorizados usando Árvores de Regressão para posterior integração nos Modelos Lineares Generalizados.

Ambas as abordagens conduziram a Prémios Puros semelhantes, no entanto verificou-se que os Modelos Lineares Generalizados conseguem ser mais diferenciadores dos riscos em carteira.

Palavras-chave: Modelos Lineares Generalizados, *Machine Learning*, Árvores de Regressão, Tarifação *a priori*, Seguro Automóvel.

Abstract

The construction of a tariff has been approached from a more conservative perspective, using Generalized Linear Models. However, the growing interest in data analysis has motivated the search for alternative approaches to the pricing problem, namely *Machine Learning* methods. These methods are at the forefront of modeling and are a strong competitor to Generalized Linear Models in terms of results.

This dissertation presents an alternative approach to modeling automobile insurance premiums, using Regression Trees. For comparative purposes, modeling using Generalized Linear Models was also considered. The continuous risk factors (age of the driver and age of the vehicle) were categorized using Regression Trees for further integration into Generalized Linear Models.

Both approaches led to similar Pure Premiums, however it was found that the Generalized Linear Models are able to be more discriminating of the risks in the portfolio.

Keywords: Generalized Linear Models, *Machine Learning*, Regression Trees, Pricing, Automobile Insurance.

Índice

Índice de Figuras

Índice de Tabelas

Siglas

AIC *Akaike Information Criterion* 15, 17, 41, 45, 47
AUC *Area Under the Curve* x, 48

BIC *Bayesian Information Criterion* 15, 17, 41, 45, 47

ML *Machine Learning* 2, 57
MLG Modelos Lineares Generalizados 2, 4, 7, 8, 12, 28, 31, 50, 54, 55, 57

ROC *Receiver Operating Characteristic* x, 47, 48

1

Introdução

A imprevisibilidade da ocorrência de um acontecimento danoso motivou a necessidade da existência de um modelo de negócio que assegurasse o pagamento dos prejuízos associados ao mesmo. A atividade seguradora surge como resposta a essa necessidade e consiste na transferência de risco de um segurado para a Seguradora mediante um acordo prévio: o contrato de seguro. Neste contrato, a Seguradora compromete-se a indemnizar o cliente por sinistros ocorridos durante um período de tempo, geralmente um ano, em troca do pagamento de um prémio: o prémio de seguro.

A atribuição de um prémio a determinado segurado deverá fazer jus ao risco que este representa para a Seguradora. Entre outras características, pretende-se que o prémio de seguro seja justo e competitivo face aos prémios praticados pelo mercado segurador. Se não, veja-se o seguinte: a aplicação de um único prémio transversal a todos os segurados implicaria que os segurados de menor (maior) risco estariam a pagar um prémio muito superior (inferior) ao devido. Por conseguinte, os segurados de menor risco tenderiam a anular o seu contrato e os de maior risco a permanecer em carteira. A curto prazo, observar-se-ia a perda de prémio e aumento de risco na carteira da Seguradora, uma vez que os prémios atribuídos aos perfis de maior risco serão insuficientes para salvaguardar possíveis indemnizações futuras.

Pelas razões acima enunciadas, sublinha-se a necessidade de cobrar um prémio de seguro adequado para cada apólice subscrita. Sabendo que o risco varia de apólice para apólice e sendo impossível distinguir individualmente as apólices que originarão sinistros, recorre-se à Lei dos Grandes Números para estimar a perda esperada da carteira. De acordo com esta lei, verifica-se que o valor da perda agregada tende para o seu valor esperado.

Ao contrário das restantes atividades comerciais, a atividade seguradora vende um produto para o qual desconhece, à partida, os possíveis custos futuramente associados ao mesmo. Assim, a Seguradora rege-se pelo seu histórico de sinistralidade para estipular os prémios da sua carteira.

O Seguro Automóvel é um seguro de massas e vigora nas modalidades de Responsabilidade Civil e de Danos Próprios. A aquisição de um seguro de Responsabilidade

Civil tornou-se obrigatória por lei para qualquer veículo em circulação na via pública. Ao abrigo do *Decreto-Lei 291/2007 de 21/08*, a cobertura obrigatória deste seguro contempla a indemnização de "danos corporais ou materiais causados a terceiros por um veículo terrestre a motor [...]". Os montantes mínimos, em vigor a partir de 1 de junho de 2012, a cargo da Seguradora por acidente, são 5 000 000€ para danos corporais e 1 000 000€ para danos materiais. Ao contrário do seguro automóvel de Responsabilidade Civil, um seguro de Danos Próprios é de caráter facultativo e vem complementar o primeiro, na medida em que os prejuízos causados no veículo do próprio segurado, mesmo que da sua responsabilidade, também estarão cobertos.

As carteiras de seguros de massas são compostas por uma panóplia de perfis de risco, expressando a heterogeneidade dos riscos característica de carteiras deste tipo de seguros. O elevado número de apólices inviabiliza a atribuição de um prémio distinto a cada uma, procedendo-se à divisão da carteira em subgrupos de risco homogéneos – *escalões tarifários* –, de forma a que o prémio a cobrar aos segurados de cada subgrupo seja semelhante.

Analisar a heterogeneidade dos riscos em carteira, identificar as causas que influenciam o risco – *fatores tarifários* – e quantificar o risco inerente a cada apólice de seguro constituem as principais tarefas que permitem proceder à divisão da carteira em subgrupos de risco homogéneos. Consequentemente, a tarifa resulta da combinação dos modelos para Frequência de Sinistralidade e para a Severidade dos Sinistros.

O teor estatístico das tarefas implicadas na modelação de uma tarifa torna os Modelos Lineares Generalizados (MLG) uma metodologia apelativa, sendo que a grande parte das Seguradoras opta por estimar a sua tarifa recorrendo à referida abordagem.

Apesar da simplicidade de interpretação dos modelos produzidos por MLG, esta metodologia possui algumas exigências que podem ser limitantes na modelação. Neste sentido, tem-se assistido à procura de abordagens igualmente robustas, mas que libertem os modelos de muitas das limitações impostas. Os métodos de *Machine Learning* (ML) são apontados como a vanguarda das técnicas de modelação e têm sido conduzidos estudos para a sua integração na atividade seguradora, como Staudt e Wagner (2019). As Árvores de Regressão Breiman, Friedman, Olshen e Stone (1984) são um dos métodos a considerar, uma vez que subentendem a divisão do espaço dos preditores em regiões e será o método adotado nesta dissertação, como alternativa aos MLG.

A dissertação encontra-se organizada em cinco capítulos. No capítulo um, introduzem-se os principais conceitos referentes à tarifação *a priori*. Segue-se o capítulo dois, onde se apresenta uma breve revisão de literatura sobre as abordagens endereçadas a esta temática. Nos capítulos três e quatro, descrevem-se, respetivamente, os fundamentos teóricos relativos aos Modelos Lineares Generalizados e às Árvores de Regressão, motivando a sua aplicação a uma carteira de Seguro Automóvel de Responsabilidade Civil, no capítulo cinco. Ainda neste capítulo, procede-se à comparação dos prémios estimados por cada uma das abordagens. Por fim, no capítulo seis sumariam-se as principais conclusões. Em anexo encontram-se os fatores de risco da carteira de Seguro Automóvel.

2

Tarifação *a priori*

2.1 Generalidades

A variabilidade do risco reforça a necessidade de recorrer a modelos para estimar, fidedignamente, as perdas que poderão ocorrer.

O Prémio Puro a cobrar durante um período de tempo $[0,t)$, $PP(t)$, obtém-se pelo produto da Frequência de Sinistralidade e da Severidade dos Sinistros. A Frequência de Sinistralidade, $Freq(t)$, respeita ao Número de Sinistros ocorridos durante o período de tempo referido. Por sua vez, a Severidade dos Sinistros, $CM(t)$, define-se pelo quociente entre o Montante Total de sinistros pagos pela Seguradora e o Número de Sinistros reportados que originaram indemnização. Desta forma, tem-se

$$PP(t) = Freq(t) \times CM(t). \tag{2.1}$$

Ao segurado será cobrado o Prémio Puro acrescido de cargas relacionadas com a margem de segurança e com os encargos e impostos que advêm da celebração do contrato de seguro.

Equivalentemente, o conceito de Prémio Puro pode ser definido como o valor esperado do montante agregado de perda originada por uma carteira. Desta forma, e denotando por S o referido montante, tem-se o seguinte:

$$S(t) = \sum_{i=0}^{N(t)} C_i, \tag{2.2}$$

sendo $N(t)$ o Número de Sinistros ocorridos durante o período de tempo $[0,t)$ e C_i o Custo de cada sinistro, $i = \{1,2,\ldots,N\}$, em que $C_0 = 0$. Pela Lei dos Grandes Números, a perda associada a uma carteira de apólices tende para o seu valor esperado. Assim, a modelação do Prémio Puro passa por "explicar" $\mathbb{E}[S]$.

A construção de uma tarifa permite diferenciar os segurados em carteira agrupando-os em subgrupos de risco homogéneos. A segurados que pertençam ao mesmo subgrupo ser-lhes-á aplicado o mesmo Prémio Puro.

2.2 Fatores de Risco e Variáveis de Interesse

Identificar as causas que influenciam o risco deve ser o ponto de partida para a construção de uma estrutura tarifária. Dependendo do fenómeno, haverão fatores mais impactantes do que outros. No Seguro Automóvel, por exemplo, observa-se que as causas que influenciam a Frequência de Sinistralidade podem diferir das que influenciam a Severidade dos Sinistros e apresentam-se em maior número.

A maioria dos fatores de risco implicados na tarifação *a priori* são do tipo categórico, sendo que cada categoria se designa por *nível tarifário*. Apesar de poderem existir fatores do tipo discreto ou contínuo, observa-se uma tendência para a categorização dos mesmos, aquando da utilização de MLG. Essencialmente, dois aspetos deverão ser considerados aquando da conversão dos fatores de risco contínuos em categóricos. Um, prende-se com a amplitude excessiva dos intervalos que poderá "mascarar" diferenças existentes na classificação do risco. O outro relaciona-se com o número suficiente de apólices presentes num nível tarifário que assegure a robustez das estimativas do modelo.

As variáveis de interesse são a Frequência de Sinistralidade e a Severidade dos Sinistros, conforme se esteja a modelar o Número de Sinistros ou o Montante dos Sinistros, respetivamente. A tabela 2.1 sintetiza as variáveis de interesse implicadas na modelação e o modo como se obtêm.

Tabela 2.1: Variáveis de Interesse.

Variável Resposta, X	**Exposição**, $expo$	**Variável de Interesse**, $Y = X/expo$
Número de Sinistros	Número de Apólices	Frequência de Sinistralidade
Custo dos Sinistros	Número de Sinistros	Severidade dos Sinistros

Identificados os fatores mais impactantes sobre o fenómeno de estudo, pretende-se estimar o valor esperado da variável de interesse condicionado pelos fatores de risco determinados. Como a natureza das variáveis aleatórias Número de Sinistros, N, e Montante dos Sinistros, C, são distintas, opta-se por modelar cada uma individualmente. De facto, considerando que as variáveis C_i, $i \in \{1,2,\ldots,N\}$, são identicamente distribuídas e assumindo a independência entre as variáveis aleatórias N e C, o valor esperado da perda esperada pode decompor-se da seguinte forma:

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=0}^{N} C_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=0}^{N} C_i | N = n\right]\right] = \mathbb{E}[N]\,\mathbb{E}[C], \tag{2.3}$$

com C uma variável aleatória com a mesma distribuição das variáveis aleatórias C_i, $i \in \{1,2,\ldots,N\}$.

2.3 Pressupostos

Durante a modelação de uma tarifa são considerados alguns pressupostos, nomeadamente relativos à Independência entre Apólices, a Incrementos Independentes e à Homogeneidade, ver Ohlsson e Johansson (2010). Seguidamente, apresentam-se os referidos pressupostos e as implicações que trazem do ponto de vista do Seguro Automóvel.

Por simplificação de escrita, designe-se por Y a variável aleatória de interesse.

Pressuposto 1 (Independência entre Apólices)**.** *: Seja Y_i a variável aleatória correspondente à apólice i, $i \in \{1, 2, \ldots, n\}$.*

As variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$ consideram-se independentes.

Apesar da sua importância, podem ocorrer situações em que o Pressuposto 1 não se verifique. Suponha-se, por exemplo, que os condutores implicados num acidente de viação estão cobertos pela mesma Seguradora. Naturalmente, o pressuposto de independência na frequência é violado. Apesar disso, a não verificação deste pressuposto tem um impacto reduzido sobre a modelação, dada a raridade do evento.

Pressuposto 2 (Incrementos Independentes)**.** *: Considerem-se n intervalos de tempo disjuntos e Y_i a variável aleatória correspondente ao intervalo i, $i \in \{1, 2, \ldots, n\}$.*

As variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$ consideram-se independentes.

À luz deste Pressuposto, o Número e o Custo de Sinistros consideram-se independentes em períodos de tempo disjuntos. No entanto, a ocorrência de acidente de viação pode despoletar uma condução cuidadosa, influenciando o Número de Sinistros futuros.

A assunção dos Pressupostos 1 e 2 vem simplificar a modelação e, embora haja situações em que possam ser violados, a sua não verificação é pouco impactante.

Pressuposto 3 (Homogeneidade)**.** *: Considerem-se n apólices no mesmo escalão tarifário e Y_i a variável aleatória correspondente à apólice i, $i \in \{1, 2, \ldots, n\}$.*

As variáveis aleatórias $Y_1, Y_2 \ldots, Y_n$ partilham a mesma distribuição de probabilidade.

Este Pressuposto ignora a existência de tendências de propensão à sinistralidade, prevalecendo a duração da apólice sobre o momento em que esta se inicia. Com um Sistema de *Bonus Malus* Lemaire (2012) é possível corrigir o Prémio Puro atribuído *a priori* com base na sinistralidade apresentada pelo segurado.

Lema 1 (Média e Variância de X e Y)**.** *Sob os Pressupostos 1, 2 e 3, o valor esperado e a variância das variáveis X e Y, presentes na tabela 2.1, obtêm-se da seguinte forma:*

$$\begin{aligned} \mathbb{E}[X] &= expo \cdot \mu \qquad e \qquad \mathbb{V}[X] = expo \cdot \sigma^2 \\ \mathbb{E}[Y] &= \mu \qquad e \qquad \mathbb{V}[Y] = \frac{\sigma^2}{expo} \end{aligned} \quad ,$$

com $expo > 0$.

A ideia geral da prova do Lema 1 consiste em definir X como uma soma de m variáveis aleatórias independentes e identicamente distribuídas $Z_k, k \in \{1, 2, \ldots m\}$, que resultam da divisão da exposição *expo* em m intervalos de igual amplitude, $1/n$. Os detalhes da prova podem ser consultados em Ohlsson e Johansson (2010).

3

Revisão de Literatura

Múltiplas têm sido as abordagens dirigidas à temática da tarifação *a priori*, desde as mais conservadoras às mais inovadoras. De facto, existem estudos que promovem a incorporação de novas técnicas de modelação de uma tarifa no Mercado Segurador.

Os Modelos Lineares Generalizados, introduzidos por Nelder e Wedderburn (1972), são amplamente utilizados em tarefas de modelação. Estes modelos primam pela versatilidade de problemas que abrangem e pela facilidade de interpretação dos modelos por eles produzidos, tornando-os a forte aposta na área da tarifação.

De uma maneira geral, a estimação de uma tarifa assenta na construção de modelos para a Frequência de Sinistralidade e para o Custo dos Sinistros. Em Ohlsson e Johansson (2010), por exemplo, apresentam-se os fundamentos teóricos para a construção de uma tarifa com MLG. Para aplicar esta metodologia é necessário determinar as distribuições de probabilidade das variáveis resposta, ou seja, do Número de Sinistros e do Custo dos Sinistros. Sendo que o Número de Sinistros representa uma contagem ao longo de um período de tempo, habitualmente, consideram-se os Processos de *Poisson* para descrever tal fenómeno. Por sua vez, e tendo em conta que o Custo dos Sinistros possui, tipicamente, uma assimetria positiva e cauda pesada, distribuições como a Gama, a Lognormal ou a Inversa Gaussiana revelam-se adequadas para a referida variável.

Outro dos pontos que merece a devida atenção quando se modela uma tarifa com MLG é a definição de um escalão tarifário de base, identificando-o como Segurado Padrão. Idealmente, deverão ser escolhidos os níveis tarifários com maior exposição ao risco para caracterizar o Segurado Padrão. Os Prémios Puros dos restantes segurados ou de novos serão determinados a partir do Prémio Puro do Segurado Padrão, aplicando-lhe descontos ou agravamentos consoante possuam níveis tarifários menos ou mais gravosos do que os que estão presentes no Segurado Padrão, respetivamente.

A razão pela qual se opta por construir modelos individuais para a Frequência de Sinistralidade e para a Severidade dos Sinistros deve-se, principalmente, às diferentes naturezas que as variáveis resposta apresentam. Os autores Garrido, Genest e Schulz (2016) sugerem a modelação da Frequência e da Severidade por MLG assumindo a dependência entre ambas. Esta dependência é introduzida considerando o Número de Sinistros como

variável explicativa no modelo da Severidade. Apesar de não se terem verificado melhorias significativas por adotar esta abordagem, os autores recomendam que a modelação de uma tarifa contemple a dependência entre Frequência e da Severidade sempre que existir uma associação significativa entre as variáveis de interesse.

A segmentação da carteira em subgrupos de risco homogéneos justifica a categorização dos fatores de risco que não sejam categóricos. Uma das questões abordadas em Henckaerts, Antonio, Clijsters e Verbelen (2018) é precisamente a categorização de fatores de risco contínuos e geográficos no contexto da construção de uma tarifa automóvel numa Seguradora Belga. Neste artigo, ajustaram-se Modelos Aditivos Generalizados Hastie e Tibshirani (1990) às variáveis de interesse integrando fatores de risco do tipo categórico, contínuo e geográfico. Com os modelos obtidos, procedeu-se à categorização dos fatores de risco contínuos e geográficos com recurso a *Evolutionary Trees* Grubinger, Zeileis e Pfeiffer (2014) e ao Algoritmo da Partição Natural de *Fisher* Fisher (1958), respetivamente. Com as categorias criadas, construíram-se MLG para a Frequência de Sinistralidade e para a Severidade dos Sinistros. Pôde constatar-se que os Prémios Puros obtidos pelos MLG aproximam os que foram obtidos pelos Modelos Aditivos Generalizados.

Identificar os fatores de risco com maior influência sobre as variáveis resposta e uma distribuição de probabilidade para as mesmas pode tornar-se um processo moroso no primeiro caso e difícil no segundo. Os algoritmos de *Machine Learning* têm-se afirmado em múltiplas áreas pela sua capacidade de oferecer resultados competitivos em relação às abordagens tradicionalmente utilizadas.

Neste sentido, os autores Henckaerts, Côté, Antonio e Verbelen (2020) propõem metodologias alternativas aos MLG para a modelação da Frequência de Sinistralidade e da Severidade dos Sinistros, baseadas em Árvores de Decisão. Árvores de Regressão, *Random Forests* e *Gradient Boosting Machines* foram as metodologias elegidas para modelar as variáveis de interesse. Para além disso, os autores sugerem a minimização da função de erro *deviance* ao invés da habitual medida, o Erro Quadrático Médio. Estas alterações levaram à criação de um *package* que permitisse englobar essa adaptação: `distRforest` Henckaerts (2021). A metodologia *Gradient Boosting Machines* é a que produz melhores resultados, mas peca pela falta de interpretação, tornando-a pouco apelativa para a modelação de uma tarifa.

A interpretação dos modelos obtidos por algoritmos *Machine Learning* é a principal dificuldade à medida que a complexidade destes cresce. Desta forma, é necessário definir estratégias que permitam decifrar tais modelos. Em Kuo e Lupton (2020) apresenta-se um conjunto de técnicas que permitem interpretar os modelos mais fechados como as *Boosted Trees*. Entre elas, destacam-se as Importâncias Relativas dos Preditores e os Gráficos de Dependências Parciais.

A aplicação de novas técnicas para a construção de estruturas tarifárias é um assunto que se encontra atualmente em plena expressão, sendo diversos os estudos desenvolvidos nesse sentido. Pretende-se, com esta dissertação, contribuir para este tópico de particular interesse para a atividade seguradora.

4

Modelos Lineares Generalizados

Os Modelos Lineares Generalizados permitem, tal como a Regressão Linear Múltipla, modelar a relação existente entre uma variável de interesse e um conjunto de preditores ou variáveis explicativas. A generalização reside na possibilidade de incluir outras distribuições de probabilidade para além da Normal e, se desejável, contemplar a transformação do valor médio da variável de interesse assegurando a relação linear com os preditores.

Seguem-se os fundamentos teóricos relativos a esta metodologia, apresentando os conceitos relevantes para a aplicação da mesma no contexto da tarifação *a priori*.

4.1 Descrição

Seguindo a notação de Turkman e Silva (2000), seja Y a variável de interesse e $\boldsymbol{X} = \{X_1, X_2, \ldots, X_p\}$ o conjunto de preditores. As observações de $(Y, \boldsymbol{X})$ são da forma $(y_i, \boldsymbol{x}_i)$, $i \in \{1, 2, \ldots, n\}$, podendo ser esquematizadas através de uma matriz (4.1).

$$\begin{array}{ccccc} y_1 & x_{11} & x_{12} & \ldots & x_{1p} \\ y_2 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ y_n & x_{n1} & x_{n2} & \ldots & x_{np} \end{array} \tag{4.1}$$

Os Modelos Lineares Generalizados definem-se por duas componentes: aleatória e sistemática. A primeira diz respeito à proporção de variabilidade da variável de interesse que se deve à aleatoriedade e a segunda à variabilidade observada na variável de interesse que é passível de "explicação" através dos preditores. De facto,

- a componente aleatória estipula que as variáveis $Y_i|\boldsymbol{X}_i$, $i \in \{1, 2, \ldots, n\}$, são independentes e com distribuição pertencente à Família Exponencial, (ver secção 4.1.1).
- a componente sistemática estabelece uma relação entre $\mathbb{E}[Y_i|\boldsymbol{X}_i] = \mu_i$ e um preditor linear $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ através de uma função monótona, diferenciável e invertível, h. Assim, tem-se:

  $$\eta_i = h(\mu_i),$$

  sendo $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ um vetor de parâmetros de dimensão p.

4.1.1 Família Exponencial

A aplicação dos Modelos Lineares Generalizados requer que a distribuição da variável de interesse pertença à Família Exponencial. Deste modo, segue-se a definição deste conceito, sob a validade do pressuposto de independência das variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$.

Definição 1 (Família Exponencial)**.** *A distribuição de uma variável aleatória Y diz-se pertencer à Família Exponencial se a sua função densidade de probabilidade (caso contínuo) ou função de probabilidade (caso discreto), f_Y, se puder escrever como*

$$f_Y(y) = f_Y(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{4.2}$$

sendo θ e ϕ parâmetros escalares e $a(\cdot)$, $b(\cdot)$ e $c(\cdot,\cdot)$ funções reais conhecidas.

Os parâmetros θ e ϕ designam-se por parâmetros *canónico* e de *dispersão*, respetivamente, e determinam univocamente a distribuição em causa. Admite-se, sem perda de generalidade, que a função $a(\phi)$ é positiva e contínua, $b(\theta)$ é diferenciável de segunda ordem e $c(y,\phi)$ é independente de θ.

A tabela 4.1 concretiza os valores dos parâmetros canónico e de dispersão e da função $b(\cdot)$ para algumas distribuições da Família Exponencial.

Tabela 4.1: Distribuições da Família Exponencial.

Distribuição		θ	$b(\theta)$	ϕ
Poisson	$Poi(\lambda)$	$\log(\lambda)$	e^θ	1
Binomial	$B(n,\pi)$	$\log\left(\frac{\pi}{1-\pi}\right)$	$n\log(1+e^\theta)$	1
Binomial Negativa	$BN(\mu,k)$	$\log\left(\frac{k\mu}{1+k\mu}\right)$	$-\frac{1}{k}\log(1-ke^\theta)$	1
Gama	$G(\mu,\nu)$	$-\frac{1}{\nu}$	$-\log(-\theta)$	$\frac{1}{\nu}$
Inversa Gaussiana	$IG(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2

O Lema 2 enuncia a relação existente entre a média e variância da variável de interesse e os parâmetros θ e ϕ. A prova deste Lema pode ser consultada em Turkman e Silva (2000).

Lema 2 (Média e Variância)**.** *Seja Y uma variável aleatória com distribuição pertencente à Família Exponencial. Então*

$$\mathbb{E}[Y] = \mu = b'(\theta) \quad e \quad \mathbb{V}[Y] = a(\phi)b''(\theta).$$

Pela tabela 2.1 verifica-se que as variáveis de interesse têm a forma de um rácio entre a variável resposta e a exposição ao risco, ou seja, $Y_i = X_i/w_i$, $i \in \{1,2,\ldots,n\}$. Assim, a expressão (4.2) reescreve-se da seguinte forma:

$$f_{Y_i}(y_i|\theta_i,\phi,w_i) = \exp\left\{\frac{w_i}{\phi}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi,w_i)\right\}. \tag{4.3}$$

A média e a variância de cada variável aleatória $Y_i, i \in \{1, 2, \ldots, n\}$, obtêm-se por:

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i) \quad e \quad \mathbb{V}[Y_i] = \frac{\phi}{w_i} b''(\theta_i).$$

4.1.2 Função de Ligação

A função de ligação é parte integrante de um Modelo Linear Generalizado e permite estabelecer uma relação entre o valor médio da variável de interesse e uma combinação linear dos preditores, como referido anteriormente.

Ora, considerando uma função monótona, diferenciável e invertível, h, tem-se

$$h(\mu_i) = \eta_i = \sum_{v=1}^{p} x_{iv}\beta_v, \quad i \in \{1, 2, \ldots, n\}.$$

Se a função de ligação h satisfizer $h(\cdot) = (b'(\cdot))^{-1}$, verifica-se que

$$\theta_i = \eta_i = \sum_{v=1}^{p} x_{iv}\beta_v, \quad i \in \{1, 2, \ldots, n\},$$

e, por isso, o preditor linear η_i modela diretamente o parâmetro canónico θ_i. Nestas situações a função de ligação designa-se por canónica.

A tabela 4.2 apresenta as funções de ligação canónica correspondentes a algumas distribuições de probabilidade.

Tabela 4.2: Funções de Ligação Canónica.

Distribuição	**Função de Ligação Canónica**	
Poisson	Logarítmica	$\eta = \log(\mu)$
Binomial	Logística	$\eta = \log\left(\frac{\mu}{1-\mu}\right)$
Gama	Inversa	$\eta = 1/\mu$
Inversa Gaussiana	Inversa Quadrática	$\eta = 1/\mu^2$

A escolha da função de ligação tem repercussões no tipo de estrutura tarifária que se obterá. Habitualmente, as funções de ligação utilizadas na tarifação *a priori* são a identidade e a logarítmica, já que se admitem respostas com distribuição *Poisson* e Gama.

Se se optar pela função de ligação identidade, isto é, $h(\mu) = \mu$, obtém-se uma Tarifa Aditiva pois

$$\mu_i = \eta_i = \sum_{v=1}^{p} x_{iv}\beta_v, \quad i \in \{1, 2, \ldots, n\}.$$

Se, por outro lado, se considerar a função de ligação logarítmica, isto é, $h(\mu) = \log(\mu)$, obtém-se uma Tarifa Multiplicativa pois

$$\log(\mu_i) = \eta_i = \sum_{v=1}^{p} x_{iv}\beta_v \Leftrightarrow \mu_i = \exp\left\{\sum_{v=1}^{p} x_{iv}\beta_v\right\} = \prod_{v=1}^{p} \exp(x_{iv}\beta_v), \quad i \in \{1, 2, \ldots, n\}.$$

Numa Tarifa Aditiva, a relação entre os prémios é definida através da soma de uma quantidade monetária que reflete as diferenças de risco entre segurados. À medida que se adicionam descontos ao prémio do Segurado Padrão, corre-se o risco de obter prémios negativos ou irrisórios. Desta forma, a Tarifa Multiplicativa é uma alternativa mais consistente para a determinação de prémios, uma vez que as diferenças de risco são contempladas através de uma proporção do prémio do Segurado Padrão.

4.2 Estimação dos Parâmetros

As componentes do parâmetro $\boldsymbol{\beta}$ são estimadas pelo Método da Máxima Verosimilhança. Considere-se uma amostra de n observações independentes como definida em (4.1), um MLG definido como em (4.3) e uma função de ligação $h(\cdot)$ tal que $h(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$.

A função log-verosimilhança, l, define-se da seguinte forma:

$$\begin{aligned} l(\boldsymbol{\theta};\phi,\boldsymbol{y}) &= \log\left(\prod_{i=1}^{n} f_{Y_i}(y_i|\theta_i,\phi,\omega_i)\right) = \sum_{i=1}^{n}\left(\frac{\omega_i}{\phi}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi,\omega_i)\right) \\ &= \sum_{i=1}^{n} l_i(\theta_i;\phi,y_i). \end{aligned} \tag{4.4}$$

Os estimadores da máxima verosimilhança de $\boldsymbol{\beta}$ obtêm-se pelo Sistema de Equações de Verosimilhança:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_v} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_v} = 0, \quad v \in \{1,2,\ldots,p\}. \tag{4.5}$$

Pelas regras de derivação composta, verifica-se:

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_v} = \frac{\partial l_i(\theta_i)}{\partial \theta_i} \cdot \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \cdot \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \cdot \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_v},$$

pelo que a equação (4.5) se resume a

$$\sum_{i=1}^{n} \frac{1}{\mathbb{V}[Y_i]}(y_i - \mu_i)x_{iv}\frac{\partial \mu_i}{\partial \eta_i} = 0, \quad v \in \{1,2,\ldots,p\}. \tag{4.6}$$

As equações do sistema (4.6) não têm, muitas vezes, uma solução analítica, induzindo a necessidade de recorrer a Métodos Numéricos para determinar uma solução do mesmo. Em Turkman e Silva (2000) detalha-se ao pormenor o processo de estimação dos parâmetros do modelo.

Seguidamente, destacam-se algumas propriedades assintóticas dos estimadores da máxima verosimilhança de $\boldsymbol{\beta}$:

- O estimador $\hat{\boldsymbol{\beta}}$ é assintoticamente centrado,

$$\lim_{n\to\infty} \mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$

- A matriz de covariâncias de $\hat{\boldsymbol{\beta}}$ é aproximadamente igual ao inverso da matriz de informação de *Fisher*[1]

$$\lim_{n\to\infty} \text{cov}(\hat{\boldsymbol{\beta}}) = I^{-1}(\boldsymbol{\beta})).$$

- O estimador $\hat{\boldsymbol{\beta}}$ tem distribuição assintótica Normal p-variada:

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N_p(\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})).$$

4.3 Inferência sobre o Modelo

Perante um Modelo Linear Generalizado que contenha todos os preditores, é necessário selecionar aqueles que expressam um maior impacto sobre a variável de interesse. Noutras palavras, é necessário inferir sobre a significância dos parâmetros do modelo. Para esse efeito, recorre-se a testes de hipóteses.

Os testes mais comuns testam a nulidade de um parâmetro (4.7) ou de um conjunto de parâmetros em simultâneo (4.8) e as suas hipóteses definem-se, respetivamente, como:

- Teste à nulidade de um parâmetro:

$$\mathcal{H}_0 : \beta_v = 0 \quad vs \quad \mathcal{H}_1 : \beta_v \neq 0, \quad v \in \{1, 2, \ldots, p\} \tag{4.7}$$

- Teste à nulidade de um conjunto de parâmetros:

$$\mathcal{H}_0 : \boldsymbol{\beta}_r = 0 \quad vs \quad \mathcal{H}_1 : \boldsymbol{\beta}_r \neq 0, \quad r < v, \quad v \in \{1, 2, \ldots, p\} \tag{4.8}$$

sendo $\boldsymbol{\beta}_r$ um subvetor de $\boldsymbol{\beta}$.

As hipóteses acima apresentadas podem ser escritas de forma condensada, considerando uma matriz $[C]_{q\times p}$, com $q \leq p$ e característica máxima q, e um vetor $\boldsymbol{\xi}$ de dimensão q previamente definido:

$$\mathcal{H}_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad vs \quad \mathcal{H}_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi} \tag{4.9}$$

Por exemplo, para (4.7), a matriz C definir-se-á por $C = (I_v \; O_{v\times(p-v)})$, sendo I_v a matriz identidade de dimensão v e $O_{v\times(p-v)}$ a matriz de zeros de dimensão $v \times (p-v)$.

As estatísticas de teste utilizadas nos testes (4.7) e (4.8) são as estatísticas de *Wald* e da Razão de Verosimilhanças. De seguida, encontra-se a explanação de cada uma delas, considerando que $\hat{\boldsymbol{\beta}}$ é o estimador da máxima verosimilhança de $\boldsymbol{\beta}$.

4.3.1 Teste de *Wald*

Suponha-se na presença de um teste de hipóteses definido como em (4.9). Das propriedades da distribuição Normal Multivariada e admitindo a validade de $I(\boldsymbol{\beta}) \approx I(\hat{\boldsymbol{\beta}})$ para grandes amostras, tem-se

$$C\hat{\boldsymbol{\beta}} \overset{a}{\sim} N_q(C\boldsymbol{\beta}, CI^{-1}(\hat{\boldsymbol{\beta}})C').$$

[1] $I(\boldsymbol{\beta}) = -H(\boldsymbol{\beta})$, sendo $H(\cdot)$ a matriz hessiana.

Sob a hipótese nula, a estatística de *Wald*, W, define-se por

$$W = (C\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})^T [CI^{-1}(\hat{\boldsymbol{\beta}})C^T]^{-1}(C\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}),$$

e tem distribuição assintótica de um Qui-Quadrado com q graus de liberdade, $W \overset{a}{\sim} \chi^2_q$.

Ao nível de significância α, rejeita-se a hipótese nula se o valor da estatística de teste de *Wald* pertencer à Região Crítica, RC_W.

$$RC_W = \left]\chi^2_{q,(1-\alpha)}; +\infty\right[$$

A rejeição da hipótese nula implica que o parâmetro em teste é significativo para o modelo e, por isso, o preditor associado ao mesmo não deve ser removido do modelo.

4.3.2 Teste da Razão de Verosimilhanças

Este teste estatístico tem particular interesse na comparação de modelos encaixados, ou seja, modelos em que um deles é submodelo do outro.

A estatística da Razão de Verosimilhanças ou *Estatística de Wilks*, RV, baseia-se na distribuição assintótica do máximo das verosimilhanças sob as hipóteses $\mathcal{H}_0$ e $\mathcal{H}_0 \cup \mathcal{H}_1$, definindo-se por:

$$RV = -2\log \frac{\max_{\mathcal{H}_0} L(\boldsymbol{\beta})}{\max_{\mathcal{H}_0 \cup \mathcal{H}_1} L(\boldsymbol{\beta})} = -2\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\},$$

sendo $\tilde{\boldsymbol{\beta}}$ o estimador da máxima verosimilhança restrito à hipótese $C\boldsymbol{\beta} = \boldsymbol{\xi}$.

Sob a hipótese nula, a distribuição assintótica desta estatística de teste é um Qui-Quadrado com q graus de liberdade, $RV \overset{a}{\sim} \chi^2_q$. O número de graus de liberdade resulta da diferença entre o número de parâmetros a estimar sob $\mathcal{H}_0 \cup \mathcal{H}_1$ (p) e o número de parâmetros a estimar sob $\mathcal{H}_0$ $(p-q)$.

Ao nível de significância α, a hipótese nula é rejeitada se o valor observado da estatística de teste pertencer à Região Crítica, RC_{RV}, definida por:

$$RC_{RV} = \left]\chi^2_{q,(1-\alpha)}; +\infty\right[$$

A rejeição da hipótese nula implica que o conjunto de parâmetros em teste é significativo para o modelo e, por isso, os preditores associados não deverão ser removidos do modelo.

4.3.3 Teste de *Tukey - HSD*

Além de determinar os preditores mais significativos, importa perceber se os grupos definidos pelo modelo são suficientemente distintos entre si que justifiquem uma diferenciação mais detalhada. Este tópico vê-se enfatizado no contexto da tarifação *a priori*, uma vez que se pretende estruturar subgrupos de risco diferentes entre si.

O teste de *Tukey - HSD*, *Tukey's Honestly Significant Difference test*, permite proceder à comparação múltipla de grupos, dois a dois, através das suas médias.

Seguindo Abdi e Williams (2010), denote-se por A o número de grupos existentes e N o número total de observações.

As hipóteses de teste para a comparação dos grupos $a \in A$ e $a' \in A$ são:

$$\mathcal{H}_0 : \mu_a = \mu_{a'} \quad vs \quad \mathcal{H}_1 : \mu_a \neq \mu_{a'},$$

com μ_a o valor médio do grupo a e $\mu_{a'}$ o valor médio do grupo a'.

Sob a hipótese nula, a estatística q define-se por

$$q = \frac{\mu_a - \mu_{a'}}{\sqrt{\frac{1}{2} MS_{S(A)} \left(\frac{1}{n_a} + \frac{1}{n_{a'}} \right)}}, \tag{4.10}$$

sendo $MS_{S(A)}$ o erro quadrático médio calculado dentro dos grupos e n_a e $n_{a'}$ o número de observações dos grupos a e a', respetivamente.

Sabendo que q segue uma distribuição *Studentized Range*[2] com amplitude A e $N - A$ graus de liberdade, $q_{A,N-A}$, da equação (4.10) resulta a estatística de teste, HSD:

$$HSD = q_{A,N-A;\alpha} \sqrt{\frac{1}{2} MS_{S(A)} \left(\frac{1}{n_a} + \frac{1}{n_{a'}} \right)}.$$

Assim, considera-se que a diferença entre dois grupos é significativa, ao nível de significância α, se se verificar

$$|\mu_a - \mu_{a'}| \geq HSD.$$

O número de comparações a testar é $\frac{A(A-1)}{2}$, repetindo-se o procedimento para cada par de grupos.

4.4 Medidas de Seleção e Avaliação

A qualidade de ajustamento de um modelo depreende-se pela sua capacidade de retratar fidedignamente a variável de interesse recorrendo a determinados preditores. Assim, um "bom" modelo oferece uma boa interpretação do problema com o menor número de preditores possível e, ainda, um bom ajustamento aos dados.

Medidas como a Função Desvio e os critérios *Akaike Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC) são frequentemente usadas para a seleção e avaliação de modelos. A primeira para aferir a qualidade do ajustamento e as duas últimas para proceder à comparação e posterior seleção de modelos.

Para além das medidas anteriormente referidas, a análise de resíduos do modelo é elucidativa sobre qualidade do ajustamento, uma vez que permite detetar observações mal ajustadas.

[2] Ramseyer e Tcheng (1973)

4.4.1 Qualidade do Ajustamento

A função desvio define-se através das funções log-verosimilhança do modelo ajustado, M, e do modelo saturado, S. O modelo saturado contém tantos parâmetros como observações, revelando-se pouco interessante na interpretação do problema. De facto, este tipo de modelo serve de termo de comparação com outros modelos.

A função desvio $D(\boldsymbol{y}, \boldsymbol{\mu})$ permite comparar o modelo saturado com o modelo ajustado como sugere a expressão (4.11).

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = 2 \times \left[l(\hat{\boldsymbol{\beta}}_S) - l(\hat{\boldsymbol{\beta}}_M) \right] \tag{4.11}$$

As funções desvio, D, e desvio reduzido, D^*, do modelo ajustado obtêm-se por:

$$\begin{aligned} D^*(\boldsymbol{y}, \boldsymbol{\mu}) &= 2 \times \left(\sum_{i=1}^{n} \frac{\omega_i}{\phi} (y_i \widehat{\theta_{iS}} - b(\widehat{\theta_{iS}})) + c(y_i; \phi, \omega_i) - \sum_{i=1}^{n} \frac{\omega_i}{\phi} (y_i \widehat{\theta_{iM}} - b(\widehat{\theta_{iM}})) + c(y_i; \phi, \omega_i) \right) \\ &= \frac{1}{\phi} \times 2 \sum_{i=1}^{n} \omega_i \left\{ y_i (\widehat{\theta_{iS}} - \widehat{\theta_{iM}}) - [b(\widehat{\theta_{iS}}) - b(\widehat{\theta_{iM}})] \right\} = \frac{1}{\phi} D(\boldsymbol{y}, \boldsymbol{\mu}), \end{aligned} \tag{4.12}$$

sendo $(\widehat{\theta_{iS}}, \widehat{\theta_{iM}})$ os estimadores de θ_i obtidos pelo modelo saturado e ajustado, respetivamente.

A distribuição assintótica da função desvio é

$$\frac{D(\boldsymbol{y}, \boldsymbol{\mu})}{\phi} \overset{a}{\sim} \chi^2_{n-p}. \tag{4.13}$$

Quando o valor de ϕ é desconhecido e precisa de ser estimado, a distribuição assintótica fica comprometida De Jong e Heller (2008). Assim, é aconselhável complementar a análise feita à função desvio com outras medidas, nomeadamente uma análise de resíduos ao modelo ajustado.

Um resíduo R_i respeita à discrepância entre os valores observado, y_i e ajustado pelo modelo, $\hat{y}_i$, $i = \{1, 2, \ldots, n\}$.

Os Resíduos de *Pearson* são uma das opções e obtêm-se da seguinte forma:

$$R_i^P = w_i \cdot \frac{y_i - \hat{y}_i}{\sqrt{\mathbb{V}[\hat{y}_i]}}, \quad i \in \{1, 2, \ldots, n\}.$$

Por outro lado, os Resíduos do Desvio exprimem a contribuição de cada observação para a função desvio e definem-se por:

$$R_i^D = \text{sinal}(y_i - \hat{y}_i) \sqrt{w_i \cdot d(y_i, \hat{y}_i)}, \quad i \in \{1, 2, \ldots, n\},$$

Tipicamente, os resíduos são analisados através de uma representação gráfica dos seus valores contra os valores ajustados, pretendendo-se que os resíduos estejam distribuídos em torno de zero.

4.4.2 Seleção dos Modelos

Por outro lado, os critérios de informação AIC e BIC são medidas que permitem comparar e escolher modelos. Estes critérios definem-se em função da função log-verosimilhança de um modelo e de um termo penalizador, tendo-se

$$\text{AIC} = -2l(\boldsymbol{\beta};\boldsymbol{\phi},\boldsymbol{y}) + 2p$$
$$\text{BIC} = -2l(\boldsymbol{\beta};\boldsymbol{\phi},\boldsymbol{y}) + p\log(n).$$

Os valores de p e n respeitam ao número de parâmetros do modelo e ao número de observações, respetivamente. O critério BIC é mais penalizador do que o AIC, conduzindo a modelos mais simples.

Estas medidas têm especial interesse no processo de seleção de variáveis explicativas para o modelo, na medida em que é possível percecionar as melhorias que advêm da inclusão ou exclusão de determinados preditores.

4.5 Modelos para o Número de Sinistros

Seguindo a exposição feita em Guerreiro (2001), apresentam-se dois modelos que permitem ajustar uma distribuição de probabilidade ao Número de Sinistros reportados à Seguradora durante uma anuidade, sob as perspetivas de homogeneidade e heterogeneidade de riscos em carteira.

Os Modelos de *Poisson* são amplamente utilizados para modelar o número de acontecimentos ocorridos num intervalo de tempo, sendo uma opção natural para descrever o Número de Sinistros participados numa anuidade. Distinguem-se os Modelos de *Poisson* Homogéneo e Misto.

Num Modelo de *Poisson* Homogéneo assume-se a homogeneidade do risco, ou seja, considera-se que todas as apólices representam o mesmo risco para a Seguradora e, consequentemente, possuem a mesma propensão à sinistralidade.

Seja N a variável aleatória que representa o número de sinistros participados durante uma anuidade e suponha-se que N tem distribuição de *Poisson* com parâmetro λ, isto é:

$$Pr[N = n] = e^{-\lambda}\frac{\lambda^n}{n!}, \quad n = 0, 1, 2, \ldots \tag{4.14}$$

O parâmetro λ representa o número médio de sinistros participados durante uma anuidade, ou equivalentemente, a frequência de sinistralidade.

O valor médio e a variância de N são ambos iguais a λ:

$$\mathbb{E}[N] = \lambda = \mathbb{V}[N].$$

A aceitação deste modelo pressupõe que a sinistralidade dos segurados é consequência do acaso, ignorando a existência de fatores que influenciam a ocorrência de sinistro. No entanto, é recorrente observar heterogeneidade de riscos numa carteira, sendo prudente ajustar um modelo que permita incorporar tal fenómeno nas estimativas dos parâmetros.

O Modelo de *Poisson* Misto permite que a intensidade do número de sinistros, λ, varie de segurado para segurado. Neste sentido, considere-se que λ resulta da observação de uma variável aleatória, Λ, não negativa. A função de distribuição de Λ designa-se por distribuição de estrutura e define-se por:

$$U(\lambda) = Pr[\Lambda \leq \lambda].$$

Comummente, considera-se o modelo *Poisson*-Gama para modelar o número de sinistros. Neste modelo, a variável Λ tem distribuição de estrutura $Gama(\alpha, \beta)$, com função densidade dada por:

$$u(\lambda) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} e^{-\beta\lambda} \lambda^{\alpha-1}, \quad \lambda > 0, \ \ \alpha > 0, \ \ \beta > 0. \tag{4.15}$$

O valor médio e a variância da variável de estrutura, Λ, correspondem a

$$\mathbb{E}[\Lambda] = \frac{\alpha}{\beta} \quad e \quad \mathbb{V}[\Lambda] = \frac{\alpha}{\beta^2}\,.$$

Ora, sabendo que $N \sim Poi(\Lambda)$, com $\Lambda \sim Gama(\alpha, \beta)$, a função de probabilidade de N é

$$Pr[N = n] = \binom{\alpha + n - 1}{n} \left(\frac{1}{\beta+1}\right)^n \left(\frac{\beta}{\beta+1}\right)^{\alpha},$$

que corresponde à distribuição Binomial Negativa de parâmetros α e $p = \frac{1}{\beta+1}$.

O valor médio e a variância de N são:

$$\mathbb{E}[N] = \frac{\alpha}{\beta} \quad e \quad \mathbb{V}[N] = \frac{\alpha}{\beta}\left(\frac{\beta+1}{\beta}\right),$$

observando-se que $\mathbb{V}[N] > \mathbb{E}[N]$, como pretendido.

Os estimadores da máxima verosimilhança para β é $\hat{\beta} = \frac{\alpha}{\bar{n}}$ e para α é a solução da equação

$$\sum_{k=0}^{m} n_k \left(\frac{1}{\hat{\alpha}} + \ldots + \frac{1}{\hat{\alpha}+k-1}\right) = \sum_{k=0}^{m} n_k \log\left(1 + \frac{\bar{n}}{\hat{\alpha}}\right)$$

Num qualquer Modelo de *Poisson*, a exposição ao risco é incorporada através de uma variável designada por *offset*. Na verdade, é outra variável explicativa, cujo coeficiente iguala um. Desta forma e supondo que a função de ligação escolhida é a logarítmica, obtém-se

$$\boldsymbol{N}|\boldsymbol{X} \sim Poi(\exp\{\boldsymbol{X\beta} + \log \boldsymbol{expo}\}),$$

com $\boldsymbol{expo}$ a exposição ao risco.

4.6 Modelos para o Custo dos Sinistros

Nesta secção apresentam-se algumas distribuições de probabilidade para ajustar aos Custos dos Sinistros participados numa anuidade. Tendo em conta que os custos são não negativos e possuem um enviesamento à direita, distribuições como a Gama, Exponencial e Inversa Gaussiana revelam-se adequadas. A distribuição Lognormal é também uma opção válida, mesmo não pertencendo à Família Exponencial. De facto, procede-se ao ajustamento de uma Regressão Linear Múltipla ao logaritmo dos Custos dos Sinistros.

A existência de custos muito discrepantes em relação aos restantes pode condicionar o ajustamento de uma distribuição de probabilidade. Neste sentido, dividem-se os sinistros em dois grupos, classificando-os em "Regulares" e "Grandes". Esta distinção processa-se pela especificação de um valor $s > 0$ a partir do qual se situam os custos relativos aos "Grandes" sinistros. A escolha de s deve atender simultaneamente a dois aspetos: o valor deve ser elevado para representar um "Grande" sinistro e baixo para salvaguardar a suficiência de observações para a modelação dos custos de "Grandes" sinistros.

Seja C a variável aleatória indicativa do Custo dos Sinistros reportados à Seguradora durante uma anuidade. Atendendo à divisão definida e ao Teorema da Probabilidade Total, tem-se:

$$\mathbb{E}[C|\boldsymbol{X}] = \underbrace{\mathbb{E}[C|\boldsymbol{X}, C \leq s]}_{(a)} \underbrace{Pr[C \leq s|\boldsymbol{X}]}_{(b)} + \underbrace{\mathbb{E}[C|\boldsymbol{X}, C > s]}_{(c)} \underbrace{Pr[C > s|\boldsymbol{X}]}_{(d)}, \tag{4.16}$$

sendo que (a) representa custo médio por sinistros "Regulares", (b) a probabilidade de ocorrência de um sinistro "Regular", (c) o custo médio por "Grandes" sinistros e (d) probabilidade de ocorrência de um "Grande" sinistro.

A distribuição de probabilidade para os "Grandes" sinistros é, habitualmente, diferente da dos sinistros "Regulares" e, por vezes, é difícil ajustar uma que pertença à Família Exponencial. Destaca-se a distribuição Pareto Generalizada – Brazauskas e Kleefeld (2009) – cuja função de distribuição é dada por:

$$F(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-u}{\sigma}\right)^{-1/\xi}, & \text{se } \xi \neq 0 \\ 1 - \exp\left(-\frac{x-u}{\sigma}\right), & \text{se } \xi = 0 \end{cases}, \tag{4.17}$$

sendo $u \in \mathbb{R}$, $\sigma > 0$ e $\xi \in \mathbb{R}$ os parâmetros da distribuição. Os parâmetros σ e ξ precisam de ser estimados e u corresponde ao menor valor observado.

Supondo que a variável aleatória G tem distribuição GPD(u, σ, ξ), o valor médio e a variância de G obtêm-se por

$$\mathbb{E}[G] = u + \frac{\sigma}{1-\xi}, \quad \text{se } \xi < 1. \tag{4.18}$$

$$\mathbb{V}[G] = \frac{\sigma^2}{(1-2\xi)(1-\xi)^2}, \quad \text{se } \xi < \frac{1}{2}. \tag{4.19}$$

A distribuição Pareto Generalizada reduz-se à distribuição Exponencial, quando $\xi = 0$, ou à distribuição Uniforme $U(u, \sigma)$, quando $\xi = -1$. Para $\xi < 0$, obtém-se a distribuição Pareto.

4.6.1 Modelo de Regressão Logística

O Modelo de Regressão Logística permite modelar a probabilidade de ocorrência de um evento, π, condicionada pelos preditores.

Seja Y a variável dependente que toma dois valores: um, se ocorrer evento, e zero, caso contrário. A distribuição de probabilidade mais adequada para Y é a *Bernoulli* de parâmetro π:

$$f(y) = \pi^y (1-\pi)^{1-y}, \quad y = 0, 1. \tag{4.20}$$

O valor médio e a variância de Y são:

$$\mathbb{E}[Y] = \pi \quad e \quad \mathbb{V}[Y] = \pi(1-\pi),$$

Utilizando a função de ligação logística para a modelação, a probabilidade de sucesso é dada pela expressão:

$$\pi = \mathbb{P}[Y = 1|\boldsymbol{X}] = \frac{\exp(\boldsymbol{X}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}\boldsymbol{\beta})}.$$

A interpretação do Modelo de Regressão Logística é sugerida pelos *odd ratios*, OR:

$$OR = \frac{\mathbb{P}[Y = 1|\boldsymbol{X}]}{\mathbb{P}[Y = 0|\boldsymbol{X}]} = \exp(\boldsymbol{X}\boldsymbol{\beta}).$$

Os *odd ratios* permitem comparar a probabilidade de ocorrência de um evento entre grupos, podendo ocorrer três situações:

1. Se $OR = 1$, a probabilidade de ocorrência de evento num grupo é igual no grupo de base.

2. Se $OR > 1$, a probabilidade de ocorrência de evento num grupo é superior no grupo de base.

3. Se $OR < 1$, a probabilidade de ocorrência de evento num grupo é inferior no grupo de base.

No caso particular da tarifação, a probabilidade de ocorrer um "Grande" sinistro, representada por (d), na equação (4.16), pode também ser estimada pelo Modelo de Regressão Logística.

5

Métodos de *Machine Learning*

O conceito de *Machine Learning* foi introduzido pela primeira vez em 1959 pelo americano Arthur Samuel, um pioneiro na área da Inteligência Artificial. De acordo com o autor, *Machine Learning* define-se como a área de estudos que capacita os computadores de ferramentas para que estes "aprendam" sem terem de ser explicitamente programados.

De uma forma simplista e desprovida de rigor científico, os algoritmos de *Machine Learning* permitem identificar padrões relacionados com o fenómeno em estudo num conjunto de dados. Ou, equivalentemente, "prever" resultados de determinado fenómeno de acordo com um conjunto de observações. À semelhança dos Modelos Lineares Generalizados, os métodos de *Machine Learning* necessitam de três componentes para a modelação: um conjunto de dados, um conjunto de variáveis e um algoritmo, sendo que o último deverá ser escolhido de forma a endereçar o estudo em mãos.

5.1 Etapas da Modelação

A modelação de um fenómeno de estudo através de métodos de *Machine Learning* faz-se seguindo um conjunto de passos iterativos, designado por algoritmo. Neste, inclui-se uma função de erro, definida de acordo com o tipo de modelação que se pretende otimizar.

A divisão do conjunto de dados em treino e teste, a determinação dos hiperparâmetros e a verificação da *performance* do modelo proposto pelo algoritmo são momentos-chave no processo de modelação por *Machine Learning*. Cada uma destas fases serão seguidamente abordadas, seguindo Boehmke e Greenwell (2019).

5.1.1 Divisão de Dados

Um modelo constrói-se sobre um conjunto de dados e pretende-se que descreva o fenómeno em estudo com a maior precisão possível. No entanto, pode acontecer que o modelo não se comporte de igual forma com um novo conjunto de dados. A capacidade de generalização de um modelo é uma das principais preocupações em modelação.

De forma a antever a capacidade de generalização de um modelo procede-se à divisão dos dados em dois conjuntos disjuntos tipicamente designados por treino e teste. No

conjunto de treino são executadas todas as atividades relacionadas com a formalização do modelo, nomeadamente, a implementação de algoritmos e o ajuste dos respetivos hiperparâmetros. Por sua vez, o conjunto de teste apenas é utilizado na fase final para aferição da capacidade de generalização do modelo proposto.

O conjunto de treino deverá conter um maior número de observações do que o de teste, uma vez que as atividades de formalização do modelo estão nele centradas. As proporções recomendadas para a divisão dos dados são (60%, 40%), (70%, 30%) ou (80%, 20%), sendo que o primeiro valor respeita à percentagem de dados transportada para o conjunto de treino e o segundo para o conjunto de teste.

A separação dos dados em dois conjuntos, de treino e de teste, pode ser efetuada por amostragem aleatória simples ou amostragem estratificada. A amostragem aleatória simples é a forma mais direta de divisão dos dados: considerando um par de percentagens, (%treino, %teste), o conjunto de treino define-se através de uma amostra de dimensão $n \times$ %treino e o conjunto de teste pelas observações remanescentes. Por outro lado, e se a variável resposta for do tipo contínuo, a amostragem estratificada procede à segmentação da mesma em quantis e depois à seleção aleatória de uma realização de cada quantil.

Um dos critérios para a divisão dos dados é a representatividade do fenómeno em estudo nos conjuntos definidos, isto é, pretende-se que em ambos os conjuntos, o fenómeno em estudo se distribua de forma semelhante. Quando os dados são desequilibrados, a amostragem aleatória simples não assegura que o critério anteriormente enunciado seja cumprido, colocando em causa a validade do modelo a propor. Assim, a amostragem estratificada é preferível à simples para dados desequilibrados, uma vez que os conjuntos de treino e de teste por ela produzidos serão igualmente representativos do fenómeno em estudo.

5.1.2 Métodos de Reamostragem

Como referido na secção 5.1.1, no conjunto de treino são incorporadas a maioria das atividades que se relacionam com a formalização e o aperfeiçoamento do modelo, nomeadamente a determinação do conjunto de hiperparâmetros ótimos específicos do algoritmo a aplicar. O conjunto de teste encontra-se reservado durante este processo para que se possa perceber de que forma o modelo proposto reage a um novo conjunto de dados.

Um senão à utilização do conjunto de teste para aferição da qualidade do modelo proposto prende-se com a tardia perceção sobre a qualidade do mesmo, podendo colocar em xeque todo o processo de treino desenvolvido. Assim, é imperativo definir estratégias que permitam aceder à *performance* do modelo durante a sua fase de treino.

A avaliação do modelo pode ser feita no conjunto de treino, escolhendo uma métrica para aferir a mesma. No entanto, é expectável que a *performance* do modelo no conjunto de treino seja bastante satisfatória, podendo não espelhar os mesmos resultados no conjunto de teste. Desta forma, os métodos de reamostragem constituem uma ferramenta de real importância para aferir a qualidade do modelo durante a fase de treino e antecipar a sua

capacidade de generalização em novos dados. A ideia base por detrás destes métodos é a divisão do conjunto de treino em subconjuntos tais que o algoritmo é aplicado nuns e avaliado noutros.

De entre alguns métodos de reamostragem destacam-se os métodos de *Cross-Validation*, nomeadamente o de *k-Fold Cross-Validation*. Estes métodos caracterizam-se pela sucessiva divisão do conjunto de treino em subconjuntos de treino e de teste e pela avaliação da função de erro em cada subconjunto de teste. Por fim, o erro global é reportado como a média dos erros observados em cada subconjunto de teste.

No caso particular do método *k-Fold Cross-Validation*, esquematizado na figura 5.1 para o caso $k = 5$, o conjunto de dados é dividido em k grupos de tamanho e estrutura semelhantes. O modelo é ajustado em $k - 1$ subgrupos de treino e a sua *performance* é aferida no subgrupo restante através da função de erro. Este procedimento é repetido k vezes alternando o subgrupo de teste em cada grupo. No final, obtém-se um vetor de erros $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k)$ cujas componentes respeitam aos erros de cada subgrupo de teste. A média dos erros observados permite quantificar o erro que se espera observar num novo conjunto de dados.

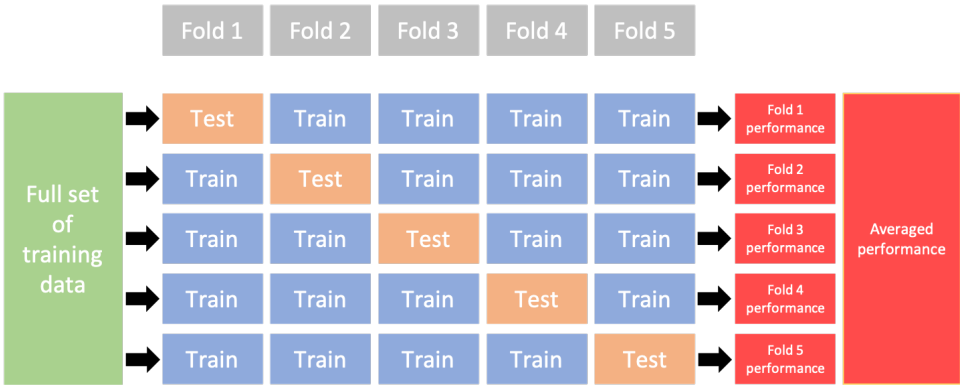


Figura 5.1: Esquema de *5-Fold Cross-Validation*.

O reduzido número de observações é também uma razão válida para adotar este tipo de estratégia, uma vez que as estimativas de erro serão mais precisas implicando uma noção mais exata sobre a qualidade do modelo ajustado.

5.1.3 Hiperparâmetros

A diferença entre os parâmetros e os hiperparâmetros de um modelo reside no momento em que ambos são definidos, isto é, os primeiros resultam do ajustamento do modelo e os segundos são especificados antes do ajustamento, ditando a forma como o modelo deve ser ajustado. Nem todos os algoritmos de *Machine Learning* possuem hiperparâmetros, mas existindo alguns que os têm coloca-se o desafio de determinar a combinação ótima de hiperparâmetros que se reflita numa boa *performance* do modelo ajustado e interpretação do problema.

As Pesquisas de Grelha Cartesiana e Aleatória são as estratégias mais comuns para determinar a combinação ótima de hiperparâmetros. A Pesquisa de Grelha Cartesiana pressupõe a definição de múltiplas combinações de hiperparâmetros para posterior ajustamento. No entanto, o número de hiperparâmetros aumenta com a complexidade do algoritmo a aplicar e esta abordagem pode tornar-se ineficiente em termos computacionais. Por outro lado, uma Pesquisa de Grelha Aleatória, seleciona aleatoriamente valores para os hiperparâmetros dentro de uma gama de possíveis valores para os mesmos.

As pesquisas mencionadas revelam-se pouco eficientes e demoradas, uma vez que é necessário testar todas as combinações de hiperparâmetros para se escolher a ótima. Kuhn (2014) apresenta um novo conceito de otimização de hiperparâmetros: a Reamostragem Adaptativa. Nesta, apenas se consideram as combinações de hiperparâmetros que induzam uma melhoria significativa na *performance* do modelo. Algoritmos complexos e com inúmeros hiperparâmetros a otimizar beneficiam deste tipo de abordagem, na medida em que o tempo computacional despendido vê-se reduzido.

O Algoritmo 1 refere-se à determinação do erro de validação de cada combinação de hiperparâmetros numa grelha por *5-Fold Cross-Validation*. A combinação ótima de hiperparâmetros corresponde àquela com menor erro de validação.

Algoritmo 1: Esquema de *5-Fold Cross-Validation*.

Input: Grelha de Pesquisa (tuneGrid).

Dividir o conjunto de treino, D, em 5 subconjuntos $(D_1, \ldots, D_5)$ por reamostragem estratificada;

foreach *combinação de hiperparâmetros em* tuneGrid **do**

  for $k = 1, \ldots, 5$ **do**

    ajustar o modelo f_k no conjunto $D \setminus D_k$;
    avaliar a performance do modelo em D_k pela função de erro $\mathscr{L}(\cdot, \cdot)$;
    $\text{cv\_comb}_k \leftarrow \mathscr{L}(y_i, f_k(\boldsymbol{x_i}))$;

  $\text{valid\_comb}_l \leftarrow \frac{1}{5} \sum_{i=1}^{5} \text{cv\_comb}_i$;

Output: Erro de Validação de cada Combinação de Hiperparâmetros na Rede de Pesquisa (tuneGrid).

5.1.4 Compromisso Viés - Variância

Os conceitos de viés e de variância estão ligados aos erros das estimativas obtidas pelo modelo ajustado. O viés define-se como o erro que advém de estimar incorretamente o fenómeno em estudo e permite perceber se o modelo capta a estrutura subjacente aos dados. Por sua vez, a variância diz respeito à variabilidade das estimativas obtidas, ou seja, mede a volatilidade das estimativas aquando da mudança do conjunto de treino.

Sendo y_0 o valor observado do fenómeno de estudo e $\hat{f}(x_0)$ a estimativa obtida para um determinado valor x_0 pelo modelo ajustado, o erro quadrático médio (esperado) pode

decompor-se em

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \mathbb{V}[\hat{f}(x_0)] + \text{Viés}[\hat{f}(x_0)]^2 + \mathbb{V}[\varepsilon], \tag{5.1}$$

com ε representativo do erro, James, Witten, Hastie e Tibshirani (2013).

As quantidades apresentadas no lado direito da equação (5.1) são estritamente positivas e, por isso, a minimização do erro quadrático médio implica a minimização do viés e da variância. No entanto, a minimização conjunta dos dois conceitos nem sempre é possível, sendo necessário estabelecer um compromisso entre ambos para que o modelo ajustado não apresente resultados absolutamente discrepantes quando sujeito a um novo conjunto de dados.

5.2 Árvores de Regressão

As Árvores de Decisão são um algoritmo de *Machine Learning* bastante apelativo uma vez que permitem modelar uma série de problemas, sejam eles de Regressão ou de Classificação. O modelo produzido é de fácil análise, sendo necessário seguir os ramos da árvores respondendo a simples questões de "Sim" ou "Não".

Seguem-se os detalhes teóricos desta metodologia, de acordo com James et al. (2013).

5.2.1 Noções Básicas

As Árvores de Decisão segmentam um conjunto de dados em subgrupos homogéneos de acordo com determinadas regras. O algoritmo `CART`, *Classification And Regression Trees*, introduzido por Breiman et al. (1984) permite construir Árvores de Decisão.

Considere-se um conjunto de dados composto por p preditores, $\boldsymbol{X} \in \mathbb{R}^p$, e por uma variável de interesse Y de tal forma que $(y_i, \boldsymbol{x}_i), i \in \{1, 2, \ldots, n\}$, sintetiza as n observações de $(Y, \boldsymbol{X})$. Sem perda de generalidade, denote-se o espaço dos preditores por R.

De uma forma geral, as Árvores de Decisão particionam o espaço dos preditores em J regiões que satisfazem

$$\forall_{q,t \in \{1,2,\ldots,J\}, q \neq t} : R_q \cap R_t = \varnothing, \quad R = R_1 \cup R_2 \cup \ldots \cup R_J. \tag{5.2}$$

A estimativa de uma nova observação $\boldsymbol{x}$ obtém-se da equação (5.3). A árvore estima o valor $\hat{y}_{R_j}$ para observações que partilhem a região $R_j, j \in \{1, 2, \ldots, J\}$.

$$f(\boldsymbol{x}) = \sum_{j=1}^{J} \hat{y}_{R_j} \mathbb{1}(\boldsymbol{x} \in R_j). \tag{5.3}$$

Sendo computacionalmente impossível considerar toda e qualquer partição de R, o algoritmo `CART` opera segundo uma heurística de *Greedy* denominada Partição Recursiva Binária. A cada passo iterativo, determina-se a melhor partição do espaço dos preditores, sendo incerto que a Árvore de Decisão reportada seja ótima em termos globais.

5.2.2 Construção de uma Árvore de Regressão

No contexto de um problema de regressão, o algoritmo de CART inicia-se com a escolha de um preditor $X_v, v \in \{1, 2, \ldots, p\}$, e de um ponto de corte s tais que o espaço dos preditores R se obtém da união das regiões R_1 e R_2, definidas por

$$R_1(v,s) = \{R|X_v < s\} \quad \text{e} \quad R_2(v,s) = \{R|X_v \geq s\}. \tag{5.4}$$

As observações presentes nas regiões R_1 e R_2 satisfazem as condições $X_v < s$ e $X_v \geq s$, respetivamente.

O preditor e o ponto de corte são escolhidos pela maior redução provocada na função de erro, $\mathscr{L}(\cdot,\cdot)$. Assim, pretende-se determinar os índices v e s que minimizem a expressão (5.5).

$$\sum_{i:\boldsymbol{x}_i \in R_1(v,s)} \mathscr{L}(y_i, \hat{y}_{R_1}) + \sum_{i:\boldsymbol{x}_i \in R_2(v,s)} \mathscr{L}(y_i, \hat{y}_{R_2}). \tag{5.5}$$

O algoritmo prossegue com a partição das regiões R_1 e/ou R_2 até que um critério de paragem seja atingido. Este critério controla o tamanho da árvore e pode definir-se como o número mínimo de observações num nó terminal ou como o tamanho máximo que a árvore pode tomar, por exemplo.

Os preditores do tipo categórico são tratados da mesma forma que os contínuos, após a conversão das suas categorias no valor médio da variável resposta.

5.2.3 *Cost-Complexity Pruning*

O critério de paragem regula o crescimento da árvore podendo conduzir ao sobreajustamento ou ao subajustamento dos dados, se definido de forma inapropriada. Árvores complexas são suscetíveis de sobreajustamento dos dados conduzindo a um modelo de fraca *performance* quando exposto a um novo conjunto de dados. Por outro lado, a precoce paragem do processo de divisão do espaço dos preditores implica que a estrutura subjacente aos dados não seja captada, resultando em subajustamento.

O compromisso viés-variância referido na secção 5.1.4 deve ser considerado durante o processo de construção de um modelo por Árvores de Regressão. Uma árvore mais (menos) complexa possui um viés baixo (alto), mas variância elevada (baixa).

Estabelecer o tamanho da árvore *a priori* pode ser algo redutor e, por isso, opta-se por uma abordagem mais robusta designada por *Cost-Complexity Pruning*. Esta consiste na construção de uma árvore complexa, T_0, e com base na definição de um parâmetro de complexidade cp obter uma sequência de sub-árvores indexadas pelo mesmo. Cada sub-árvore T é construída sob a perspetiva de minimização da expressão (5.6).

$$\sum_{m=1}^{|T|} \sum_{i:\boldsymbol{x}_i \in R_m} \mathscr{L}(y_i, \hat{y}_{R_m}) + cp \cdot |T|, \tag{5.6}$$

sendo $|T|$ o número de nós terminais da sub-árvore T.

Quanto maior for o valor de cp, mais penalizador será o termo $cp \cdot |T|$ e, consequentemente, menor será a árvore. O valor do hiperparâmetro cp deve ser determinado através de um método de reamostragem como *Cross-Validation*, por exemplo.

5.2.4 Funções de Erro

Os algoritmos de *Machine Learning* pressupõem a definição de uma função de erro a minimizar durante a fase de treino do modelo. Habitualmente, em problemas de regressão, essa função de erro é o Erro Quadrático:

$$\mathscr{L}(y_i, f(\boldsymbol{x}_i)) \propto \{y_i - f(\boldsymbol{x}_i)\}^2, \tag{5.7}$$

sendo y_i o valor observado da variável resposta e $f(\boldsymbol{x}_i)$ o valor estimado pelo modelo para $\boldsymbol{x}_i$.

Henckaerts et al. (2020), demonstraram que a utilização do Erro Quadrático como função de erro é adequada quando se trata de uma distribuição contínua e simétrica em torno da sua média e com variância constante. Perante dados desequilibrados e com uma assimetria positiva, característicos do Número e do Custos de Sinistros, respetivamente, os autores sugerem a utilização da função desvio como função de erro para efeitos de modelação.

A função desvio define-se como o quociente entre a função de verosimilhança do modelo ajustado, $L(f(\boldsymbol{x}))$, e a função de verosimilhança do modelo saturado, $L(y)$, isto é,

$$D(y, f(\boldsymbol{x})) = -2\log\frac{L(f(\boldsymbol{x}))}{L(y)} = -2\log L(f(\boldsymbol{x})) + 2\log L(y). \tag{5.8}$$

Tipicamente, observa-se que o Número de Sinistros segue uma distribuição de *Poisson*, motivando a adoção da função desvio da *Poisson* como função de erro.

$$\begin{aligned} D_{Poisson}(y, f(\boldsymbol{x})) &= 2\log\left(\prod_{i=1}^{n}\exp(-y_i)\frac{y_i^{y_i}}{y_i!}\right) - 2\log\left(\prod_{i=1}^{n}\exp(-f(\boldsymbol{x}_i))\frac{f(\boldsymbol{x}_i)^{y_i}}{y_i!}\right) \\ &= 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{f(\boldsymbol{x}_i)}\right) - (y_i - f(\boldsymbol{x}_i))\right]. \end{aligned} \tag{5.9}$$

A exposição ao risco, $expo_i$, é incorporada substituindo $f(\boldsymbol{x}_i)$ por $expo_i \cdot f(\boldsymbol{x}_i)$ na expressão (5.9). A estimativa da Frequência de Sinistralidade obtém-se de (5.3) fazendo

$$\hat{y}_{R_j} = \frac{\sum_{i\in I_j} N_i}{\sum_{i\in I_j} expo_i}, \quad I_j = \{i : \boldsymbol{x}_i \in R_j\}, \tag{5.10}$$

com N_i o Número de Sinistros da apólice i e $expo_i$ a respetiva exposição ao risco.

Segundo Therneau, Atkinson et al. (1997) e atendendo à possibilidade da existência de nós terminais sem eventos, introduz-se um termo à expressão (5.10) que previne a divisão por zero na função desvio, $D_{Poisson}$. Assumindo que os eventos existentes num nó

terminal seguem uma distribuição $Gama(\mu,\sigma)$, o parâmetro γ define-se como μ/σ. Assim, a estimativa para a Frequência de Sinistralidade reescreve-se da seguinte forma:

$$\hat{y}_{R_j} = \frac{\alpha + \sum_{i\in I_j} N_i}{\beta + \sum_{i\in I_j} expo_i}, \quad I_j = \{i : \boldsymbol{x}_i \in R_j\}, \tag{5.11}$$

sendo $\alpha = \gamma^{-2}$ e $\beta = \alpha/\hat{y}_{R_j}$.

Da mesma forma, a função desvio da Gama é uma alternativa ao Erro Quadrático para modelar o Custo dos Sinistros, definindo-se como:

$$\begin{aligned} D_{Gama}(y, f(\boldsymbol{x})) &= 2\log\left[\prod_{i=1}^{n} \frac{1}{y_i\Gamma(\alpha)}\left(\frac{\alpha y_i}{y_i}\right)^{\alpha}\exp\left(-\frac{\alpha y_i}{y_i}\right)\right] \\ &\quad - 2\log\left[\prod_{i=1}^{n} \frac{1}{y_i\Gamma(\alpha)}\left(\frac{\alpha y_i}{f(\boldsymbol{x}_i)}\right)^{\alpha}\exp\left(-\frac{\alpha y_i}{f(\boldsymbol{x}_i)}\right)\right] \\ &= 2\sum_{i=1}^{n}\alpha\left[\frac{y_i - f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} - \log\left(\frac{y_i}{f(\boldsymbol{x}_i)}\right)\right]. \end{aligned} \tag{5.12}$$

A estimativa para o Custo dos Sinistros numa árvore de regressão corresponde à média dos custos observados num determinado nó terminal:

$$\hat{y}_{R_j} = \frac{1}{|I_j|}\sum_{i\in I_j} C_i, \quad I_j = \{i : \boldsymbol{x}_i \in R_j\}, \tag{5.13}$$

com C_i o custo com a apólice i.

5.3 Interpretação do Modelo

Apesar do modelo produzido por Árvores de Decisão ser de fácil interpretação, apresentam-se duas medidas que complementam o modelo na medida em que permitem discernir sobre os modelos obtidos, seguindo Friedman, Hastie, Tibshirani et al. (2001).

5.3.1 Importância Relativa dos Preditores

Naturalmente, existem preditores que exercem uma maior influência sobre a variável resposta do que outros. Nos MLG, a seleção dos preditores mais significativos faz-se de acordo com a relevância estatística de um determinado parâmetro para o modelo. Infelizmente, não é possível transpor a mesma estratégia para os algoritmos de *Machine Learning*, particularmente para as Árvores de Decisão.

Neste sentido, introduz-se o conceito de Importância Relativa de um Preditor que se mede pela redução provocada na função de erro. Ou seja, a importância de um preditor x_l, $l \in \{1,2,\ldots,p\}$, no modelo f define-se como a soma de melhorias induzidas na função de erro quando esta é escolhida para dividir o espaço dos preditores:

$$I_l(f) = \sum_{j=1}^{J-1} \mathbb{1}(v(j) = l)(\Delta\mathscr{L})_j. \tag{5.14}$$

A soma é efetuada nos $J-1$ nós terminais da árvore, mas apenas são considerados aqueles em que a variável x_l é implicada. O termo $(\Delta\mathscr{L})_j$, representa a diferença de valor observado na função de erro antes e depois da divisão j, $j \in \{1,2,\ldots,J-1\}$, ser efetuada.

Eventualmente, os valores obtidos podem ser normalizados para que somem 100%.

5.3.2 Gráficos de Dependências Parciais

Outro complemento interessante para acompanhar o modelo é a relação existente entre determinados preditores e a variável resposta. Os Gráficos de Dependências Parciais permitem estabelecer tal relação.

Considerando um subconjunto de preditores, X_S, $S \subset \{1,2,\ldots,p\}$, e o subconjunto complementar $C = \{1,2,\ldots,p\} \setminus S$, as dependências parciais de f com respeito aos preditores X_S definem-se como

$$f_{X_S}(x_S) = \mathbb{E}_{X_C}[f(X_S, X_C)]. \tag{5.15}$$

As funções de dependências parciais podem ser estimadas por

$$\tilde{f}_{X_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_s, x_{iC}), \tag{5.16}$$

sendo, $\{x_{1C}, x_{2C}, \ldots, x_{nC}\}$ as observações de X_C.

Como se pôde verificar, os Modelos Lineares Generalizados são uma ferramenta que se caracteriza por uma forte componente estatística, transpondo resultados teóricos para a modelação. Ao invés, as Árvores de Regressão retiram o maior proveito do conjunto de dados, utilizando uma medida de erro para modelar o fenómeno em estudo.

6

Aplicação

O presente capítulo destina-se à aplicação das metodologias enunciadas nos capítulos precedentes (4 e 5) para a construção de uma tarifa automóvel. Os dados utilizados respeitam à carteira da Linha de Negócios Automóvel e foram disponibilizados pela *Seguradora T*, nome fictício por motivos de confidencialidade.

6.1 Descrição da Carteira Automóvel

A carteira automóvel de Responsabilidade Civil da *Seguradora T* possui 50 000 apólices distintas com informações relativas às variáveis descritas na tabela 6.1 e aos fatores de risco apresentados na tabela A.1, em anexo. As variáveis `nclaims` e `cost` representam as variáveis resposta para efeitos de modelação. Os valores dos fatores de risco são recolhidos aquando da subscrição do contrato de seguro e permitem traçar o perfil de risco do segurado.

Tabela 6.1: Descrição das Variáveis.

Variável	Descrição
`nclaims`	Número de sinistros reportados.
`exposition`	Fração do ano correspondente à exposição ao risco de cada apólice.
`cost`	Indemnização originada pela ocorrência de sinistro, em euros.

A figura 6.1 contém a representação gráfica das variáveis presentes na tabela 6.1. Observa-se que a maioria das apólices (95.02%) não reporta sinistros durante o período subscrito e uma pequena porção das mesmas reporta um (4.63%) ou mais sinistros (0.35%). Quanto à exposição verifica-se que a maioria dos segurados (75.21%) está há menos de uma anuidade na *Seguradora T*. Os custos dos sinistros foram truncados para melhor interpretação do gráfico, evidenciando-se uma assimetria positiva.

Por sua vez, a figura 6.2 apresenta uma visão geral dos fatores de risco. Todos os fatores de risco são do tipo categórico, à exceção de `agedriver`, `agevehicle` e `bonus`.

No que respeita aos fatores de risco do tipo categórico, observa-se que a maioria dos segurados em carteira habitam nas zonas de residência C (28.21%), D (22.32%) e E

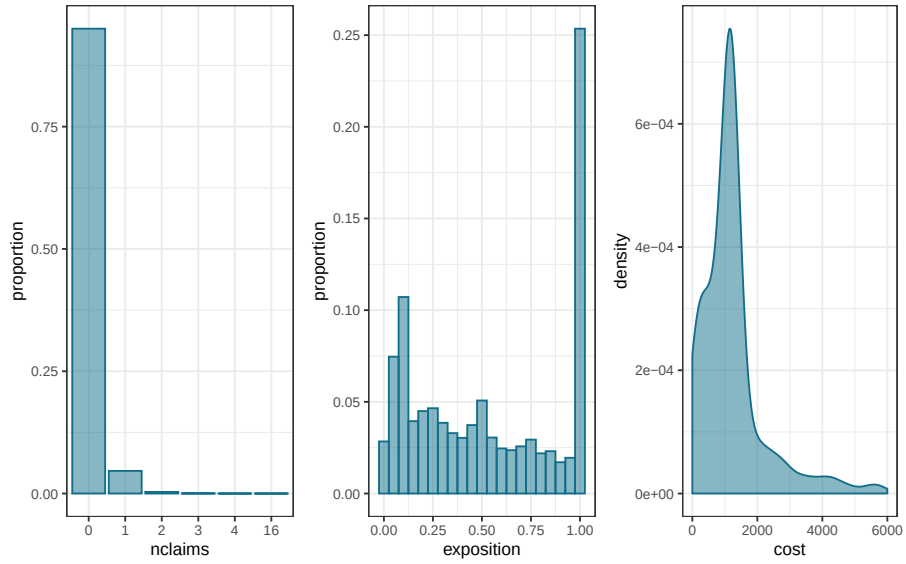


Figura 6.1: Distribuições Empíricas de `nclaims`, `exposition` e `cost`.

(20.26%). Dos doze níveis possíveis para a classificação da potência do veículo, os níveis 6 (21.96%) e 7 (21.21%) são os mais comuns. As marcas 1 (24.04%), 2 (23.36%) e 12 (24.65%) são característica da maioria das apólices. Finalmente, verifica-se um equilíbrio entre as proporções de apólices que possuem um veículo com combustível Gasolina (E) (51.17%) e as que possuem um veículo com combustível Diesel (D) (48.83%).

Por outro lado e para os fatores de risco do tipo contínuo, observa-se uma predominância de apólices cuja idade do veículo seguro é inferior ou igual a 15 anos (92.44%). A idade dos segurados varia, essencialmente, entre os 25 e os 60 anos (80.51%). A proporção de apólices nas classes de desconto de um Sistema de *Bonus Malus* supera aquelas que se situam numa classe de agravamento, observando-se uma concentração das mesmas na classe com 50% de desconto (56.67%). Apesar do fator de risco `bonus` ser, muitas vezes, excluído da tarifação *a priori*, este fornece informação sobre os segurados em carteira. Desejavelmente, pretende-se que exista um equilíbrio entre os riscos menos e mais gravosos.

6.2 Categorização dos Fatores de Risco `agedriver` e `agevehicle`

A segmentação da carteira em subgrupos de risco homogéneos constitui um dos principais desafios aquando da elaboração de uma estrutura tarifária. A tipologia dos fatores de risco a inserir num modelo é determinante quanto à diferenciação de tais subgrupos.

A inserção dos fatores de risco `agedriver` e `agevehicle` como contínuos num MLG pode revelar-se desajustada na medida em que os prémios estimados não transparecerão o real risco de sinistralidade que diferentes idades de segurados e de veículos seguros representam para a Seguradora. Assim, é frequente proceder-se à categorização dos fatores mencionados previamente à sua incorporação num MLG.

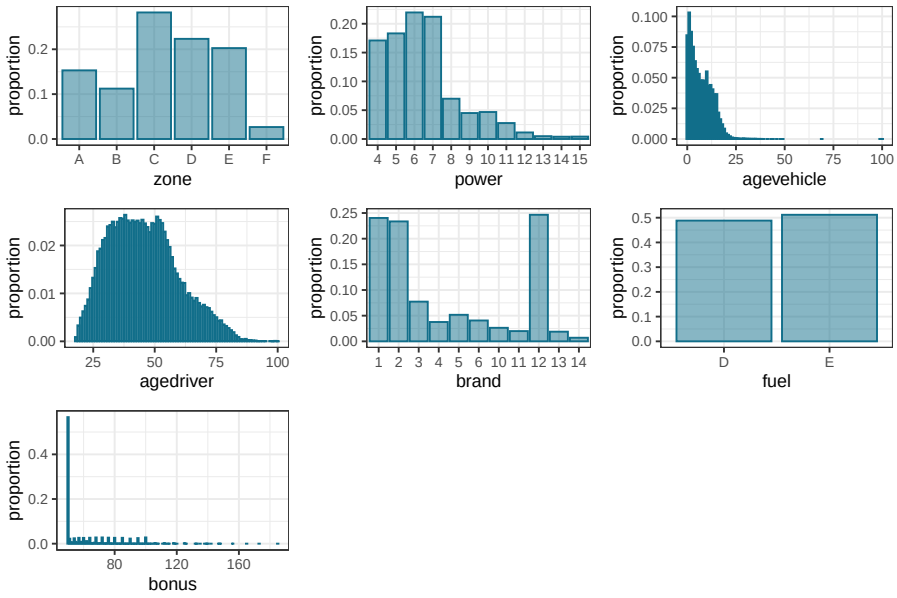


Figura 6.2: Distribuições Empíricas dos Fatores de Risco `zone`, `power`, `agevehicle`, `agedriver`, `brand`, `fuel` e `bonus`.

Uma possível abordagem para a definição dos níveis tarifários de cada fator de risco contínuo passa por considerar a divisão dos seus valores em intervalos de amplitudes semelhantes. No entanto, esta abordagem suscita questões relacionadas com o número de intervalos a considerar e, posteriormente, com a necessidade de justificar se tais intervalos são suficientemente distintos entre si e se existe homogeneidade em cada escalão, em termos de risco de sinistralidade.

Desta forma, propõe-se uma abordagem alternativa à anterior que permite determinar os níveis tarifários recorrendo a Árvores de Regressão de *Poisson*. Para tal, definiu-se o conjunto de treino com 80% das observações por amostragem estratificada. As tabelas 6.2 e 6.3 demonstram que ambos os conjuntos de treino e de teste apresentam a mesma distribuição do Número de Sinistros, como pretendido.

Tabela 6.2: Proporção do Número de Sinistros no Conjunto de Treino.

$N = n$	0	1	2	3	4	16
Proporção	0.95108	0.04548	0.00315	0.00025	0.00002	0.00002

Tabela 6.3: Proporção do Número de Sinistros no Conjunto de Teste.

$N = n$	0	1	2	3	4
Proporção	0.9467	0.0494	0.0036	0.0002	0.0001

De seguida, procedeu-se à otimização da Árvore de Regressão considerando uma Rede Cartesiana contendo possíveis valores de hiperparâmetros característicos desta metodologia, ilustrados na tabela 6.4.

Tabela 6.4: Rede Cartesiana - Árvore de Regressão de *Poisson*.

$cp \in \{1.0 \times 10^{-5}, 1.1 \times 10^{-4}, \ldots, 9.91 \times 10^{-3}\}$
$\gamma \in \{2^{-3}, 2^{-2}, 2^{-1}, 2^{0}\}$

A combinação ótima de hiperparâmetros foi determinada recorrendo ao método de *5-fold Cross-Validation* e corresponde àquela com o menor valor da função de erro. A tabela 6.5 evidencia o equilíbrio existente entre os subconjuntos utilizados em *5-fold Cross-Validation*. De facto, a Frequência de Sinistralidade é semelhante em cada um dos subconjuntos D_1 a D_5.

Tabela 6.5: Frequência de Sinistralidade em cada Subconjunto do Conjunto de Treino.

	D_1	D_2	D_3	D_4	D_5
$\sum_i N_i / \sum_i expo_i$	0.1015185	0.0997074	0.09840521	0.1020947	0.1006079

Além da Rede Cartesiana, estipulou-se um critério de paragem que permite assegurar a representatividade dos subgrupos produzidos pela árvore final: cada nó terminal deverá conter pelo menos 5% das apólices presentes no conjunto de treino.

6.2.1 Fator de Risco `agedriver`

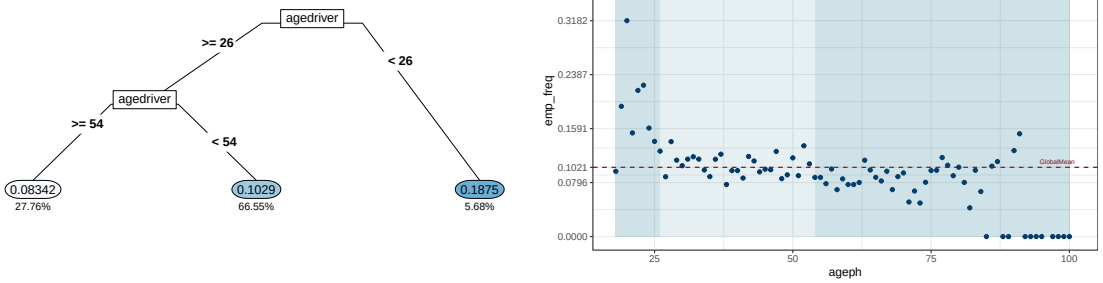
Considerando apenas o preditor `agedriver` para a aplicação da Árvore de Regressão e a Rede Cartesiana definida (6.4), a combinação ótima de hiperparâmetros

$$cp = 6.1 \times 10^{-4}, \quad \gamma = 1$$

conduz aos níveis tarifários $[18, 26)$, $[26, 54)$ e $[54, 101)$.

A figura 6.3 esquematiza a Árvore de Regressão obtida e as frequências empíricas determinadas para cada segurado em carteira. A figura 6.3a confirma as suspeitas de que o nível tarifário $[18, 26)$ é o mais propício a sinistralidade, com a Frequência de Sinistralidade estimada mais elevada. De facto, pela figura 6.3b observa-se que as frequências empíricas de tal nível se encontram muito acima da frequência média de sinistralidade.

No conjunto de teste, que contém 20% do total das observações, a Árvore de Regressão gera um erro de 0.3264, não muito distinto do valor obtido para a combinação ótima, 0.3115. Sendo que a função desvio mede o desvio entre os valores observados e estimados, considera-se que o erro obtido é aceitável e, por isso, as categorias definidas são plausíveis.



(a) Árvore de Regressão. (b) Categorias obtidas pela Árvore.

Figura 6.3: Representação Gráfica da Árvore e das Categorias obtidas para o Fator de Risco `agedriver`.

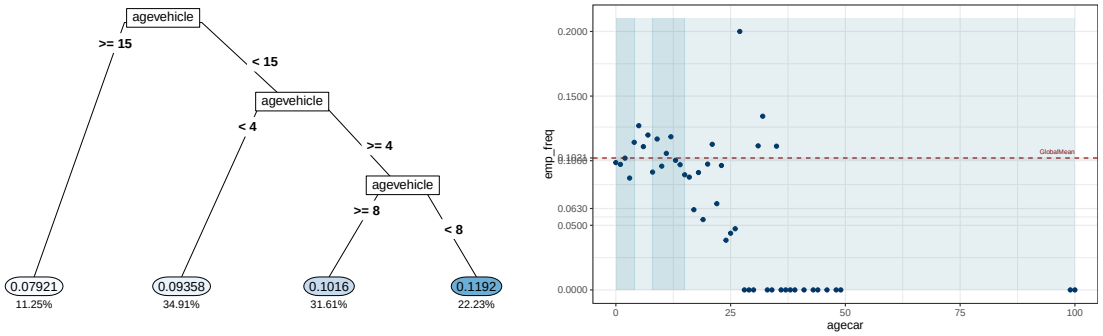
6.2.2 Fator de Risco `agevehicle`

Desta vez, considerando apenas o preditor `agevehicle` para a aplicação da Árvore de Regressão e a Rede Cartesiana definida (6.4), a combinação ótima de hiperparâmetros

$$cp = 3.1 \times 10^{-4}, \quad \gamma = 2^{-3}$$

codifica o fator em causa nos níveis tarifários $[0,4)$, $[4,8)$, $[8,15)$ e $[15,101)$.

A figura 6.4 esquematiza a Árvore de Regressão obtida e as frequências empíricas determinadas para cada idade de veículo seguro. Pela Árvore de Regressão (6.4a), verifica-se que os veículos com idades compreendidas entre os 4 e os 8 anos apresentam a Frequência de Sinistralidade mais elevada, comprovando-se tal facto pela figura 6.4b.



(a) Árvore de Regressão. (b) Categorias obtidas pela Árvore.

Figura 6.4: Representação Gráfica da Árvore e das Categorias obtidas para o Fator de Risco `agevehicle`.

Quando aplicada ao conjunto de teste, a Árvore de Regressão obtida produz um erro de 0.3289, semelhante ao obtido para a combinação ótima determinada, com um valor de 0.3126. Mais uma vez, este valor é indicativo de um modelo razoável, tornando níveis tarifários estabelecidos satisfatórios.

As categorias que codificam os fatores de risco `agedriver` e `agevehicle` serão posteriormente utilizadas para a construção de modelos para a Frequência de Sinistralidade e para a Severidade dos Sinistros, através de Modelos Lineares Generalizados.

6.3 Análise Exploratória

Os gráficos das figuras 6.5 a 6.10 contemplam a Frequência de Sinistralidade e o Custo Médio empíricos para cada nível tarifário dos fatores de risco, bem como os intervalos de confiança construídos com recurso ao Teorema do Limite Central.

As barras azuis representam a exposição ao risco, a vermelho encontram-se representadas as frequências de sinistralidade empíricas e a verde os custos médios empíricos.

Análises preliminares deste tipo permitem percecionar as diferenças existentes entre os níveis tarifários relativamente à Frequência de Sinistralidade e ao Custo Médio empíricos. Como nível de referência considerou-se a Frequência de Sinistralidade Global (linha horizontal tracejada, a vermelho) e o Custo Médio Global (linha horizontal tracejada, a verde) da carteira cujos valores são, 0.1021 e 1 715.51 €, respetivamente.

Fator de Risco: `agedriver`

A faixa etária $[18, 26)$ é a faixa mais propícia a sinistralidade, pois apresenta a maior Frequência de Sinistralidade, cerca de 0.1853. De notar que é a faixa etária que possui menos apólices, justificando a maior amplitude do intervalo de confiança.

No que respeita ao Custo dos Sinistros, o padrão mantém-se: a faixa etária $[18, 26)$ originou os custos médios mais elevados (1 945.21 €).

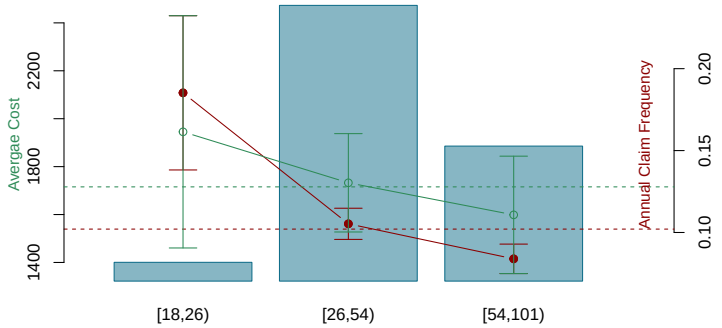


Figura 6.5: Frequência e Severidade de Sinistralidade Empíricas pelo Fator de Risco `agedriver`.

Fator de Risco: `agevehicle`

O número de unidades expostas ao risco é semelhante em todos os níveis tarifários, à exceção do nível $[15, 101)$. Tomando como referência a Frequência Global da carteira, observa-se que os níveis $[4, 8)$ e $[8, 15)$ são os que apresentam a maior Frequência de Sinistralidade.

Em média, observa-se que os custos com sinistros são mais elevados no nível tarifário $[0, 4)$, cerca de 1 874.11 €, e mais baixos no nível $[15, 101)$, cerca de 1 460.29 €. Por outro lado, os níveis $[4, 8)$ e $[8, 15)$ apresentam um custo médio por sinistro semelhante ao Custo Médio Global.

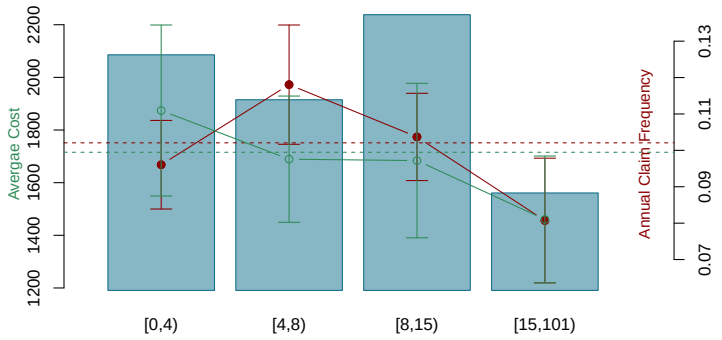


Figura 6.6: Frequência e Severidade de Sinistralidade Empíricas pelo Fator de Risco `agevehicle`.

Fator de Risco: `zone`

As zonas de residência menos populosas - A, B, C - detêm a Frequência de Sinistralidade mais baixa, enquanto que as restantes se encontram acima da média global.

Por outro lado, os custos mais elevados com sinistros foram registados nas zonas D, cerca de 1 844.31 €, e E, cerca de 1 881.85 €, e os mais baixos correspondem a segurados residentes na zona F, cerca de 1 324.69 €.

Fator de Risco: `power`

De uma forma geral, tanto a Frequência de Sinistralidade como o Custo Médio por Sinistro apresentam um comportamento semelhante em todos os níveis tarifários, sendo que os seus valores oscilam em torno dos valores globais.

Fator de Risco: `brand`

À semelhança do que se constatou no fator de risco `power`, verifica-se que a Frequência de Sinistralidade e o Custo Médio por sinistro não distam bruscamente entre os níveis

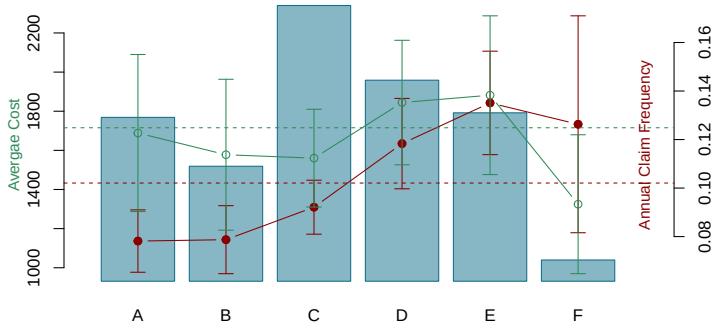


Figura 6.7: Frequência e Severidade de Sinistralidade Empíricas pelo Fator de Risco zone.

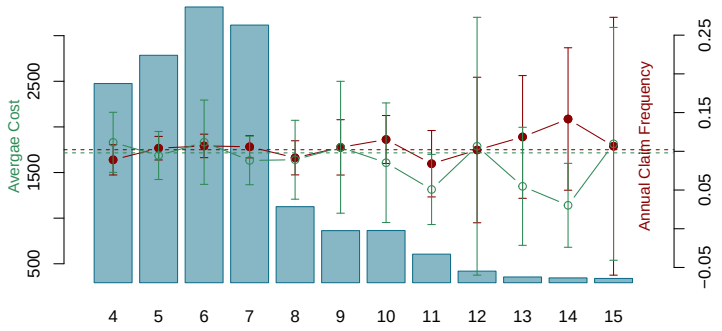


Figura 6.8: Frequência e Severidade de Sinistralidade Empíricas pelo Fator de Risco power.

tarifários do fator brand. As Frequências de Sinistralidade mais elevadas registam-se em veículos com as marcas 10 (0.1317), 11 (0.1314) e 13 (0.1285) e o Custo Médio mais alto ocorreu na marca 11 (2 738.48 €).

Fator de Risco: fuel

O tipo de combustível Diesel (D) supera o tipo Gasolina (E) em termos de Frequência de Sinistralidade e de Custo Médio por Sinistro, com valores superiores às respetivas médias globais.

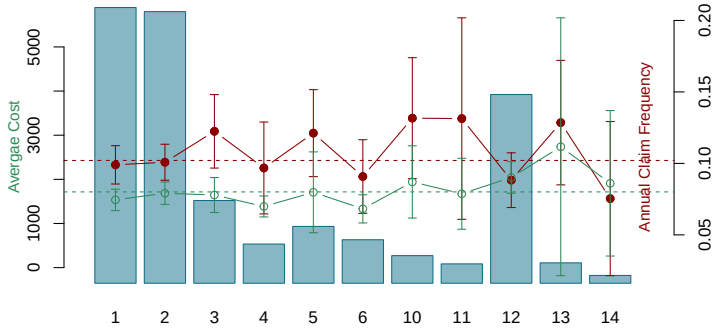


Figura 6.9: Frequência e Severidade de Sinistralidade Empíricas pelo Fator de Risco `brand`.

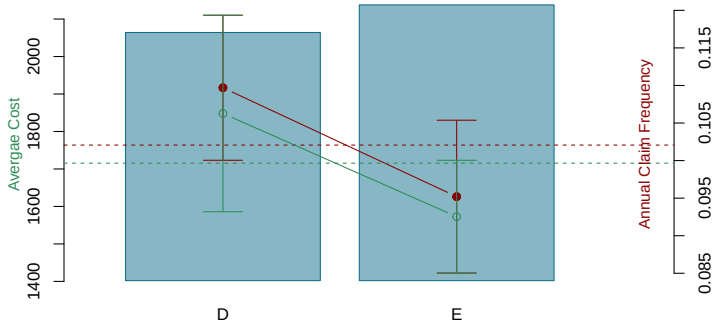


Figura 6.10: Frequência e Severidade de Sinistralidade Empíricas pelo Fator de Risco `fuel`.

6.4 Modelos Lineares Generalizados

A aplicação de Modelos Lineares Generalizados para modelação da Frequência de Sinistralidade e da Severidade dos Sinistros requer o ajustamento prévio de distribuições de probabilidade ao Número dos Sinistros e aos seus Custos.

Para efeitos de modelação das variáveis de interesse, definiu-se um Segurado Padrão com os níveis tarifários que apresentam a maior exposição ao risco. À luz da análise feita na secção 6.3, considerou-se um indivíduo na faixa etária $[26, 54)$ anos, que habita na zona C e cujo veículo seguro se caracteriza por ter idade no intervalo $[8, 15)$ anos, potência 6, marca 1 (Renault, Nissan) e combustível Gasolina (E).

O aperfeiçoamento dos modelos foi regido por dois critérios:

i) Níveis tarifários que apresentem um p-valor igual ou superior a 0.05 devem ser agregados ao nível base.

ii) Níveis tarifários (de um mesmo fator tarifário) cujas estimativas difiram, em valor

absoluto, até 0.05, devem ser testados com vista à sua junção. Se se observar um p-valor igual ou superior a 0.05, deve proceder-se à sua agregação.

A função de ligação escolhida para a modelação foi a logarítmica, uma vez que se pretende obter uma Tarifa Multiplicativa.

6.4.1 Modelação da Frequência de Sinistralidade

Seja N_i a variável aleatória que representa o Número de Sinistros participados pela apólice i e $expo_i$ a exposição ao risco da mesma apólice, $i \in \{1, 2, \ldots, n\}$.

Inicialmente, fez-se o ajuste de uma distribuição de probabilidade ao Número de Sinistros, necessário para modelar o fenómeno da Frequência de Sinistralidade. Para tal consideraram-se os Modelos de *Poisson* Homogéneo e Misto.

Os Dados

A tabela 6.6 apresenta a distribuição observada do Número de Sinistros da carteira de Responsabilidade Civil automóvel da *Seguradora T.* [1]

Tabela 6.6: Distribuição Observada do Número de Sinistros

Número de Sinistros	Número de Apólices
0	47 510
1	2 313
2	162
3	12
4	2
16	1
Total	**50 000**

Da análise da tabela 6.6, verifica-se que a média observada para o número de sinistros é 0.05394 e a variância é 0.06423. Sendo que estes valores são distintos, pode suspeitar-se que os dados não seguirão uma distribuição de *Poisson*.

A ocorrência de 16 sinistros numa só anuidade é um fenómeno extremo que pode enviesar os testes de ajustamento realizados. Assim, a apólice que continha esse número de sinistros foi removida, prosseguindo-se o ajustamento da distribuição de probabilidade com os dados remanescentes.

O Ajustamento

Para o Modelo de *Poisson* Simples, a estimativa da máxima verosimilhança obtida para o parâmetro λ foi $\hat{\lambda} = 0.05362107$.

[1] A possibilidade de existirem sinistros não reportados à Seguradora foi descartada.

Pelo teste do Chi-Quadrado de *Pearson*[2] (tabela 6.7) verificou-se que o valor observado para a estatística de teste χ^2 foi $\chi^2_{obs} = 184.7286$.

Tabela 6.7: Número de Sinistros Observados e Ajustados - Modelo de *Poisson* Simples.

k	n_k	p_k	np_k
0	47 510	0.947791183	47 388.6113
1	2 313	0.050821580	2 541.0282
≥ 2	176	0.001387238	69.3605
Total	**49 999**	**1.00000**	**49 999**

Como $\chi^2_{obs} > \chi^2_{2;0.95} = 5.991465$, rejeita-se a hipótese de que os dados sejam provenientes de uma distribuição de *Poisson*, confirmando-se a heterogeneidade dos riscos em carteira.

Neste sentido, testou-se o Modelo de *Poisson* Misto com a Gama como distribuição de estrutura, isto é, o Modelo *Poisson*-Gama. As estimativas da máxima verosimilhança para os parâmetros da distribuição Binomial Negativa são $\hat{n} = 0.5202953$ e $\hat{p} = 0.906569$.

Da aplicação do teste do Chi-Quadrado de *Pearson* obtiveram-se os resultados apresentados na tabela 6.8.

Tabela 6.8: Número de Sinistros Observados e Ajustados - Modelo da Binomial Negativa.

k	n_k	p_k	np_k
0	47 510	0.9502455816	47 511.32884
1	2 313	0.0461930846	2 309.60804
2	162	0.0032806970	164.03157
≥ 3	14	0.0002806367	14.03156
Total	**49 999**	**1.00000**	**49 999**

O valor observado para a estatística de teste χ^2 foi $\chi^2_{obs} = 0.030251$.

Como $\chi^2_{obs} < \chi^2_{3;0.95} = 7.814728$, não se rejeita a hipótese de que os dados sejam provenientes de uma distribuição Binomial Negativa.

Desta forma, pode concluir-se que o Número de Sinistros da carteira segue um Modelo de *Poisson*-Gama.

O Modelo

Determinada a distribuição de probabilidade para o Número de Sinistros, seguiu-se a modelação da Frequência de Sinistralidade, incluindo a exposição ao risco como *offset* no modelo. O modelo obtido encontra-se resumido na tabela 6.9.

[2]Sob $\mathcal{H}_0$, a estatística de teste é $\chi^2 = \sum_{k=0}^{m} \frac{(n_k - np_k)^2}{np_k} \overset{a}{\sim} \chi^2_{m-p}$, sendo m o número de classes e p o número de parâmetros da distribuição em teste.

Tabela 6.9: Estimativas dos Parâmetros - Modelo da Frequência de Sinistralidade.

	Estimate	*Std. Error*	*z-value*	*p-value*
(Intercept)	-2.42251	0.05192	-46.655	$< 2 \times 10^{-16}$
zone A + B	-0.16562	0.05790	-2.860	4.23×10^{-03}
zone D	0.26262	0.05700	4.608	4.07×10^{-06}
zone E + F	0.41122	0.05665	7.259	3.90×10^{-13}
fuel D	0.16538	0.04153	3.983	6.82×10^{-05}
brand 12	-0.22680	0.05623	-4.033	5.50×10^{-05}
ageph [18,26)	0.57052	0.07802	7.312	2.63×10^{-13}
ageph [54,101)	-0.19692	0.04714	-4.177	2.95×10^{-05}
agecar [4,8)	0.14843	0.04698	3.160	1.58×10^{-03}
agecar [15,101)	-0.18499	0.07093	-2.608	9.10×10^{-03}

Os fatores mais influentes sobre o fenómeno da Frequência de Sinistralidade são a zona de residência e idade do segurado, o combustível, a marca e a idade do veículo. De acordo com este modelo, o perfil de menor risco corresponde a um segurado com idade no intervalo $[54, 101)$ que habita nas zonas A ou B e possui um veículo com idade no intervalo $[15, 101)$, de marca 12 (Japoneses e Coreanos) e combustível Gasolina (E). Ao invés, o segurado com maior propensão à sinistralidade respeita a um indivíduo de idade no intervalo $[18, 26)$ que habita nas zonas E ou F e possui um veículo com idade no intervalo $[4, 8)$, de marca 1 (Renault e Nissan) e combustível D (Diesel).

De notar que os níveis tarifários em falta, que se referem às restantes marcas, exercem o mesmo impacto sobre a Frequência de Sinistralidade que a marca 1.

A tabela 6.10 apresenta as Frequências de Sinistralidade estimadas pelo modelo para o Segurado Padrão e para os perfis de menor e maior risco, considerando uma anuidade como período de exposição ao risco.

Tabela 6.10: Frequência de Sinistralidade Estimada.

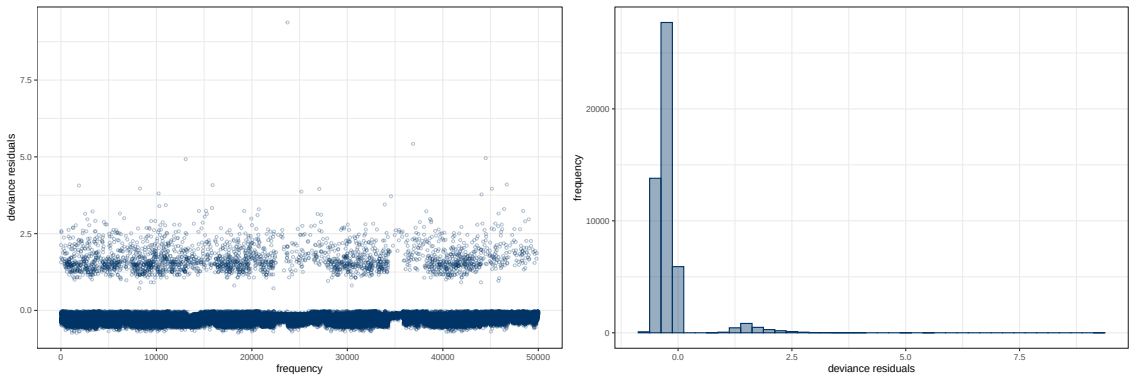
	Segurado Padrão	**Perfil de Menor Risco**	**Perfil de Maior Risco**
$\mathbb{E}[N]$	0.08869873	0.0408914	0.324017

A tabela 6.11 demonstra que o aperfeiçoamento do modelo saturado, com a seleção dos níveis tarifários mais impactantes e a junção dos que possuíam estimativas semelhantes, melhorou os valores da função desvio e dos critérios de informação AIC e BIC.

Tabela 6.11: Qualidade do Ajustamento - Modelo da Frequência de Sinistralidade.

	Função Desvio	**AIC**	**BIC**
Modelo Saturado	12 657.86	20 508.64	20 808.52
Modelo Ajustado	12 649.12	20 488.81	20 585.83

Por outro lado e analisando a figura 6.11, verifica-se que os resíduos do desvio se situam em torno do valor zero e não possuem uma particular distribuição, indiciando um bom ajustamento por parte do modelo.



(a) Gráfico de Dispersão dos Resíduos de Desvio. (b) Histograma dos Resíduos do Desvio.

Figura 6.11: Resíduos do Desvio para o Modelo da Frequência de Sinistralidade.

6.4.2 Modelação da Severidade dos Sinistros

Seja C_i a variável aleatória que representa o Custo da indemnização i, $i \in \{1, 2, \ldots, n\}$.

Dada a necessidade de repartir os Custos dos Sinistros em dois conjuntos, apresentam-se os resultados do ajustamento de distribuição aos custos dos sinistros "Regulares" e aos custos dos "Grandes" sinistros, bem como os respetivos modelos.

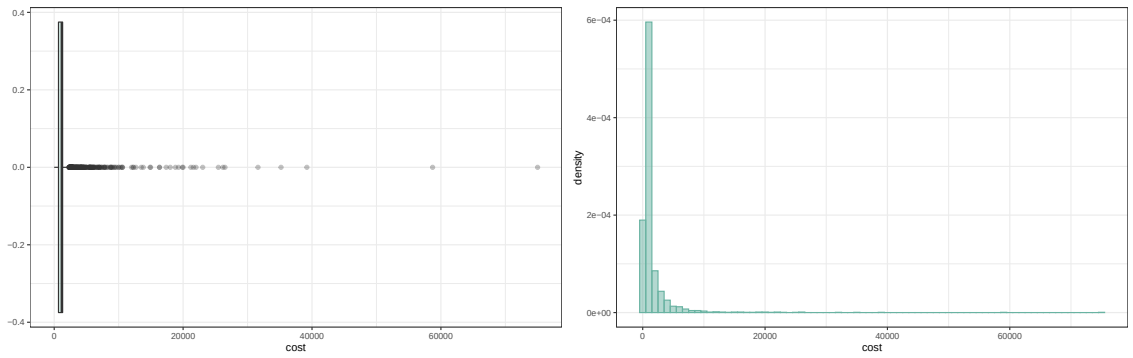
Os Dados

Foram 1924 os sinistros que originaram indemnização. As indemnizações possuem uma variabilidade acentuada, evidenciada na figura 6.12. Os valores encontram-se situados entre 0.01 €[3] e 75 000.00 €. A média dos custos é 1 715.51 € e o desvio padrão é 3 449.71 €, verificando-se que o desvio padrão é cerca de duas vezes superior à média.

Posto isto, os custos dos sinistros foram divididos em dois grupos pela imposição de um limite, $s > 0$. Este foi definido de tal forma que 96% das indemnizações surgem como sinistros ditos "Regulares" e as restantes como "Grandes" sinistros. Por conseguinte, das 1924 indemnizações geradas, 1847 possuem custos inferiores a 5 929.67 € e 77 registam custos superiores.

Feita a divisão dos custos, efetuou-se o ajustamento de distribuições de probabilidade para os custos com sinistros "Regulares" e para os custos com "Grandes" sinistros.

[3] Embora sejam poucos os custos de magnitude tão baixa, importa ressalvar que, muito provavelmente, respeitam a um ajuste de contas e não a despesas vindas de um sinistro propriamente ditas.



(a) *Boxplot* dos Custos de Sinistros.

(b) Histograma dos Custos de Sinistros.

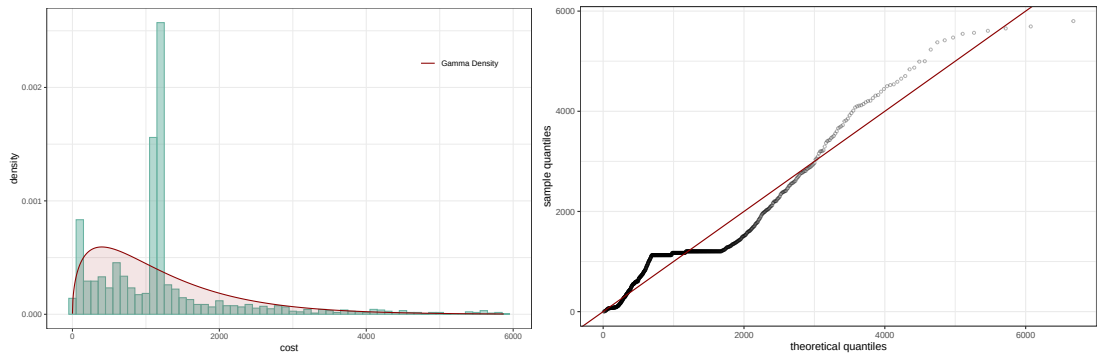
Figura 6.12: *Boxplot* e Histograma dos Custos de Sinistros.

Ajustamento dos Custos com Sinistros "Regulares"

A distribuição para o custos dos sinistros "Regulares" deve transparecer devidamente as observações e, por isso, deve possuir um suporte em $\mathbb{R}_0^+$, um enviesamento à direita e uma cauda pesada. A distribuição Gama encaixa no perfil, sendo frequentemente utilizada no ajustamento dos custos dos sinistros.

De facto, verificou-se que, ao nível de significância 0.05, os custos dos sinistros classificados como "Regulares" seguem uma distribuição Gama(α, β). As estimativas dos parâmetros da distribuição são $\hat{\alpha} = 1.484084$ e $\hat{\beta} = 0.001212398$ e foram obtidas pelos estimadores de Villaseñor e González-Estrada (2015).

A figura 6.13 demonstra o ajustamento da distribuição Gama aos custos com sinistros "Regulares". Observando-se que os pontos presentes no QQ-Plot (figura 6.13b) se encontram em torno da reta $y = x$, pode concluir-se que o ajustamento da distribuição Gama é favorável aos custos com sinistros "Regulares".



(a) Histograma dos Custos e Densidade Teórica.

(b) QQ-plot do Ajustamento.

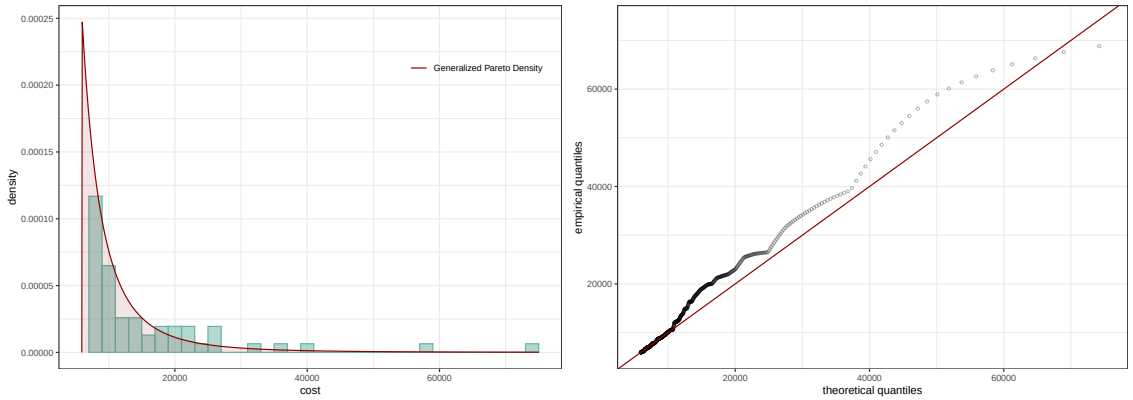
Figura 6.13: Ajustamento de Distribuição Teórica aos Custos com Sinistros "Regulares".

Ajustamento dos Custos com "Grandes" Sinistros

À semelhança do que foi feito para os custos com sinistros "Regulares", procedeu-se ao ajustamento de uma distribuição aos custos com "Grandes" sinistros.

Considerando o mesmo nível de significância (0.05), a distribuição Pareto Generalizada revelou ser a mais adequada para estes custos. As estimativas dos parâmetros desta distribuição são $\hat{\sigma} = 0.3837$ e $\hat{\xi} = 3949.0469$ e foram obtidas por máxima verosimilhança – Villaseñor-Alva e González-Estrada (2009). O valor de u corresponde ao menor custo registado dos "Grandes" sinistros, tomando o valor de 5 929.67 €.

A figura 6.14 reflete o bom ajustamento oferecido pela distribuição Pareto Generalizada aos custos dos "Grandes" sinistros, evidenciado pelo QQ-Plot, principalmente.



(a) Histograma dos Custos e Densidade Teórica. (b) QQ-plot do Ajustamento.

Figura 6.14: Ajustamento de Distribuição Teórica aos Custos com "Grandes" Sinistros.

Estão, portanto, reunidas as condições para iniciar a modelação individual dos custos dos sinistros "Regulares" e dos "Grandes" sinistros. Seguem-se os modelos obtidos para estimar as parcelas (a), (c) e (d) da equação (4.16).

Regra geral, os fatores de risco que influenciam os custos de potenciais sinistros são em número mais reduzido em relação aos que influenciam a frequência dos mesmos, uma vez que a condução do segurado não se relaciona diretamente com as indemnizações que este poderá gerar.

Modelo dos Custos com Sinistros "Regulares"

O modelo obtido para estimar os custos com sinistros "Regulares" encontra-se na tabela 6.12. Apesar do nível tarifário referente à marca de veículo 2 possuir um p-valor ligeiramente superior a 0.05, optou-se por não o agregar ao nível base.

De acordo com o modelo apresentado, o perfil de menor risco corresponde a um segurado com idade compreendida entre os 26 e os 53 anos, que habita na zona C e possui um veículo de marca 1 (Renault e Nissan), coincidindo com o perfil do Segurado Padrão. Por outro lado, o perfil de maior risco corresponde a segurados mais jovens, com idades

Tabela 6.12: Estimativas dos Parâmetros - Modelo dos Custos com Sinistros "Regulares".

	Estimate	*Std. Error*	*z-value*	*p-value*
(Intercept)	7.00873	0.02752	254.657	$< 2 \times 10^{-16}$
zone D	0.09440	0.04349	2.171	3.01×10^{-02}
brand 2	0.08637	0.04454	1.939	5.26×10^{-02}
brand 10+12	0.20830	0.04812	4.329	1.58×10^{-05}
ageph [18,26)	0.13553	0.06644	2.040	4.15×10^{-02}

compreendidas entre os 18 e os 25 anos, que habitam na zona D e conduzem um veículo de marca 10 (Mercedes e Chrysler) ou 12 (Japoneses e Coreanos).

Os níveis tarifários ausentes na tabela 6.12, têm a mesma influência sobre o fenómeno em estudo que os níveis que definem o Segurado Padrão.

Os custos médios estimados de um sinistro "Regular" para cada perfil de risco encontram-se na tabela 6.13. O custo médio estimado de um sinistro "Regular" para o perfil de maior risco é muito baixo, tendo em conta que o maior valor de um sinistro "Regular" supera o valor de 5 900 €. Deste modo, sugere-se a aplicação de um princípio de cálculo de prémios que reflita a variabilidade dos custos, como o Princípio do Desvio Padrão ou da Variância.

Tabela 6.13: Severidade dos Sinistros "Regulares" Estimada.

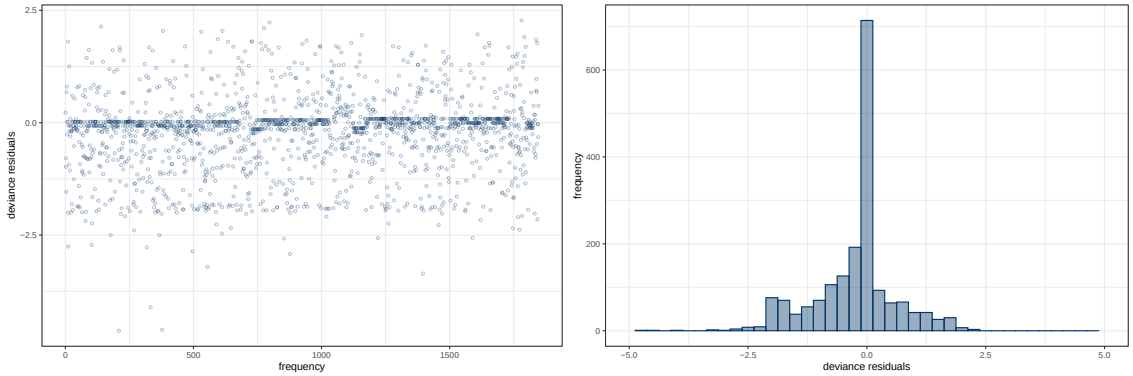
	Segurado Padrão	**Perfil de Menor Risco**	**Perfil de Maior Risco**
$\mathbb{E}[C\|C \leq s]$	1 106.25 €	1 106.25 €	1 714.64 €

Pela análise da tabela 6.14, conclui-se que os ajustamentos efetuados ao modelo saturado se revelaram efetivos, uma vez que os critérios de informação AIC e BIC se viram reduzidos. O ligeiro aumento da função desvio, advém de uma decisão da junção de dois níveis tarifários.

Tabela 6.14: Qualidade do Ajustamento - Modelo dos Custos com Sinistros "Regulares".

	Função Desvio	**AIC**	**BIC**
Modelo Saturado	1 423.394	29 874.71	30 062.43
Modelo Ajustado	1 434.003	29 834.10	29 867.23

Mais uma vez, a dispersão do pontos contemplada na representação gráfica 6.15a e a concentração de valores em torno de zero no histograma 6.15b são indicativos de um bom ajustamento do modelo aos dados.



(a) Gráfico de Dispersão dos Resíduos de Desvio. (b) Histograma dos Resíduos do Desvio.

Figura 6.15: Resíduos do Desvio para o Modelo dos Custos com "Sinistros Regulares".

Modelo dos Custos com "Grandes" Sinistros

No âmbito da modelação, pretende-se que as estimativas do modelo proposto sejam robustas, para que este "explique" fidedignamente o fenómeno em estudo. A robustez das estimativas depende, em grande parte, do número de observações utilizadas para ajustar o modelo.

Neste caso específico, o número de indemnizações que excederam o limite definido, s, é bastante reduzido, contemplando apenas 77 observações. Assim, optou-se por atribuir o valor médio da Pareto Generalizada ao custo dos "Grandes" sinistros, obtendo-se

$$\mathbb{E}[C|C > s] = 12\ 337.34 €$$

A integração dos "Grandes" sinistros numa tarifa deve ser devidamente ponderada, dependendo da experiência do atuário. No entanto, o valor médio determinado não é representativo do risco inerente a um "Grande" sinistro. Assim, sugere-se novamente a aplicação de um Princípio de Cálculo de Prémios que permita incorporar a variabilidade dos custos observada.

Modelo para a Probabilidade de Reportar um "Grande" Sinistro

O modelo para a probabilidade de reportar um "Grande" sinistro foi obtido pela aplicação do Modelo de Regressão Logística, definindo-se uma variável auxiliar que contém zeros e uns, conforme se observe um sinistro "Regular" ou "Grande", respetivamente. O modelo obtido encontra-se na tabela 6.15.

Tabela 6.15: Estimativas dos Parâmetros - Modelo de Regressão Logística.

	Estimate	***Std. Error***	***z-value***	***p-value***
(Intercept)	-3.3329	0.1372	-24.287	$< 2 \times 10^{-16}$
brand 12	0.7003	0.2599	2.694	7.05×10^{-03}

Tal como na modelação dos custos, na modelação da participação de um "Grande" sinistro são poucos os fatores de risco relevantes para o modelo. De facto, o nível tarifário relativo à marca 12 é o único que expressa influência sobre a participação de um "Grande" sinistro, reforçando a tese de que o fenómeno é difícil de "explicar" pelos fatores de risco.

O perfil de menor risco coincide com o perfil do Segurado Padrão e o perfil de maior risco contempla os segurados que possuem um veículo de marca 12 (Japoneses e Coreanos). Na verdade, os segurados que possuem um veículo com a referida marca, têm cerca do dobro da probabilidade de participar um "Grande" sinistro do que os restantes segurados.

A tabela 6.16 apresenta as probabilidades estimadas a partir do modelo para os diferentes perfis de risco.

Tabela 6.16: Probabilidade Estimada de Reportar um "Grande" Sinistro.

	Segurado Padrão	**Perfil de Menor Risco**	**Perfil de Maior Risco**
$Pr[C\|C > s]$	0.03446115	0.03446115	0.06707317

Estimada a probabilidade de reportar um "Grande" sinistro, a probabilidade de reportar um sinistro "Regular" obtém-se pelas propriedades das probabilidades:

$$Pr[C|C \leq s] = 1 - Pr[C|C > s]$$

A tabela 6.17 ilustra os valores das medidas da qualidade do ajustamento. Contrariamente ao aumento do valor da função desvio, os valores dos critérios de informação AIC e BIC diminuíram, indicando uma melhoria do modelo saturado para o ajustado, resultante das alterações feitas ao mesmo.

Tabela 6.17: Qualidade do Ajustamento - Modelo de Regressão Logística.

	Função Desvio	**AIC**	**BIC**
Modelo Saturado	613.0676	679.0676	862.6189
Modelo Ajustado	639.9309	643.9309	655.0553

A qualidade de um modelo obtido por Regressão Logística mede-se pela sua capacidade de distinguir corretamente a presença ou ausência de determinada característica numa amostra.

A curva ROC é uma representação gráfica que permite concluir sobre a qualidade do modelo. Nesta, apresentam-se os conceitos de *True Positive Rate* e de *False Positive Rate*. O primeiro diz respeito ao número de observações corretamente classificadas pelo modelo e o segundo ao número de observações incorretamente classificadas.

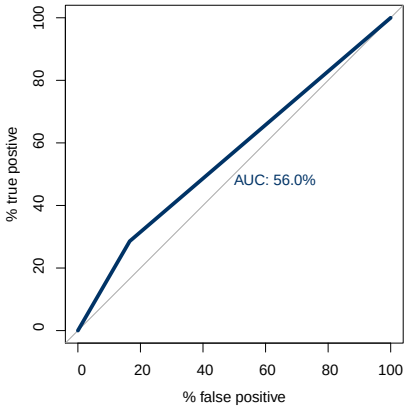
Quanto maior for a distância entre a curva ROC e a reta $y = x$, melhor é a capacidade do modelo na distinção de observações. Se, por ventura, a curva ROC se sobrepuser à

referida reta, pode dizer-se que o modelo não é efetivo na identificação da presença ou ausência de característica.

A área AUC complementa a informação presente no gráfico e corresponde à área abaixo da curva ROC, pretendendo-se que seja tão grande quanto possível.

A curva ROC e a respetiva área AUC foram utilizadas para aferir a qualidade do modelo de Regressão Logística para modelar a ocorrência de um "Grande" sinistro. Pela figura 6.16, percebe-se que a curva se encontra muito perto da reta, indicando que o modelo ajustado não consegue distinguir efetivamente um "Grande" sinistro de um sinistro "Regular".

Figura 6.16: Curva ROC e AUC - Modelo de Regressão Logística.



A carência de fatores de risco que influenciem a participação de um "Grande" sinistro, leva a adoção de outras estratégias para estimar a probabilidade. Pode, por exemplo, considerar-se que todos os segurados em carteira possuem a mesma probabilidade de reportar um "Grande" sinistro e que, deste modo, todos devem contribuir de igual forma para a eventualidade de ocorrência de um sinistro de elevada magnitude.

Por fim, conjugando os modelos obtidos para a Frequência de Sinistralidade e para a Severidade dos Sinistros, obtêm-se as estruturas tarifárias 6.18a ou 6.18b, conforme se considere o modelo para a probabilidade de reportar um "Grande" sinistro (6.15) ou não, respetivamente.

Tabela 6.18: Estruturas Tarifárias - Modelos Lineares Generalizados.

(a) Probabilidade de Reportar um "Grande" Sinistro obtida pelo Modelo.

Nível Tarifário	Tarifa
Segurado Padrão	133.68 €
zone A	84.7%
zone B	84.7%
zone C	100.0%
zone D	139.1%
zone E	150.9%
zone F	150.9%
ageph [18,26)	195.1%
ageph [26,54)	100%
ageph [54,101)	82.1%
agecar [0,4)	100.0%
agecar [4,8)	116.0%
agecar [8,15)	100%
agecar [15,101)	83.1%
power 4 - 15	100.0%
brand 1	100.0%
brand 2	106.4%
brand 3 - 6	100.0%
brand 10	116.4%
brand 11	100.0%
brand 12	113.8%
brand 13 - 14	100.0%
fuel E	100.0%
fuel D	118.0%

(b) Igual Probabilidade de Reportar um "Grande" Sinistro por todas as apólices.

Nível Tarifário	Tarifa
Segurado Padrão	137.97 €
zone A	84.7%
zone B	84.7%
zone C	100.0%
zone D	138.8%
zone E	150.9%
zone F	150.9%
ageph [18,26)	194.4%
ageph [26,54)	100.0%
ageph [54,101)	82.1%
agecar [0,4)	100.0%
agecar [4,8)	116.0%
agecar [8,15)	100.0%
agecar [15,101)	83.1%
power 4 - 15	100.0%
brand 1	100.0%
brand 2	106.2%
brand 3 - 6	100.0%
brand 10	115.8%
brand 11	100.0%
brand 12	92.3%
brand 13 - 14	100.0%
fuel E	100.0%
fuel D	118.0%

O Prémio Puro de cada perfil de risco em carteira, obtém-se pela aplicação de descontos ou agravamentos ao prémio do Segurado Padrão. Os Prémios Puros do Segurado Padrão e dos perfis de menor e maior risco encontram-se na tabela 6.19.

Tabela 6.19: Prémios Puros dos Perfis de Risco - Modelos Lineares Generalizados.

	Estrutura 6.18a	**Estrutura** 6.18b
Segurado Padrão	133.68 €	137.97 €
Perfil Menor Risco	77.32 €	73.66 €
Perfil Maior Risco	626.77 €	641.54 €

A título de exemplo, calcule-se o Prémio Puro do perfil de menor risco, de acordo com a estrutura 6.18a. Ora, o perfil de menor risco corresponde a um segurado que habite na zona A ou B, tenha idade na faixa etária 54 – 100 anos e possua um veículo com idade entre 15 e 100 anos, de marca 1, 3, 4, 5, 6, 11, 13 ou 14, potência qualquer e combustível

Gasolina (E). Assim, o seu Prémio Puro obtém-se da seguinte forma:

$$\begin{aligned} \text{PP} &= \text{PP}_{Seg.Pad.} \times \text{Tarifa}_{zona_A} \times \text{Tarifa}_{ageph_{[54,101)}} \times \text{Tarifa}_{agecar_{[15,101)}} \\ &= 133.68 \times 0.847 \times 0.821 \times 0.831 = 77.32 \text{€} \end{aligned}$$

6.5 Árvores de Regressão

As Árvores de Regressão foram a abordagem alternativa aos Modelos Lineares Generalizados escolhida para modelar uma estrutura tarifária. A ordem de trabalhos assemelha-se à dos MLG, modelando-se a Frequência de Sinistralidade e a Severidade dos Sinistros individualmente. Numa primeira instância, considera-se uma partição do conjunto de dados em conjunto de treino e de teste, seguindo-se o ajustamento das Árvores de Regressão no conjunto de treino com a determinação da combinação de hiperparâmetros ótima. Findadas as referidas tarefas, avalia-se a qualidade do modelo no conjunto de teste e procede-se à interpretação do modelo através das medidas enunciadas na secção 5.3.

6.5.1 Modelação da Frequência de Sinistralidade

Novamente, seja N_i a variável aleatória que representa o Número de Sinistros participados pela apólice i e $expo_i$ a exposição ao risco da mesma apólice, $i \in \{1, 2, \ldots, n\}$.

Divisão dos Dados

Os conjuntos de treino e de teste utilizados para a modelação da Frequência de Sinistralidade foram os mesmos da categorização dos fatores de risco contínuos `agedriver` e `agevehicle` (secção 6.2).

Da mesma maneira, foram usados os subconjuntos $D_1, D_2, \ldots, D_5$ para determinação dos hiperparâmetros ótimos (tabela 6.5) e ainda o critério de paragem que impõe a existência de pelo menos 5% de apólices em cada nó terminal.

O Modelo

O modelo para a Frequência de Sinistralidade foi obtido pela aplicação de Árvores de Regressão de *Poisson* e encontra-se esquematizado na figura 6.17. Das combinações de hiperparâmetros presentes na tabela 6.4, a combinação

$$cp = 4.1 \times 10^{-4}, \quad \gamma = 2^{-3},$$

é ótima e foi determinada pela aplicação do método de *5-Fold Cross-Validation*.

De acordo com este modelo, o perfil de menor risco corresponde a um segurado de idade igual ou superior a 29 anos, que habita nas zonas A, B ou C e possui um veículo de marca 1, 4, 12 ou 14 e combustível E. Já o perfil de maior risco, corresponde a um segurado de idade inferior a 34 anos, que habita nas zonas D, E ou F e possui um veículo com marca 1, 2, 3, 4, 5, 10, 11 ou 13. Acrescenta-se ainda que a maioria das apólices, cerca

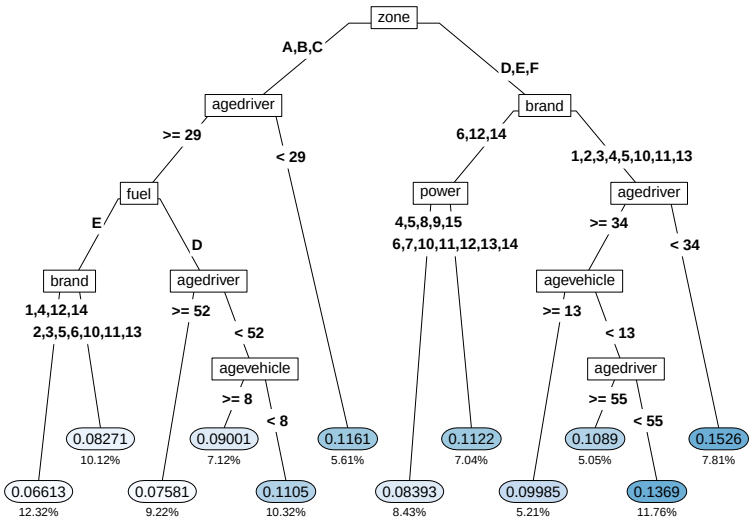


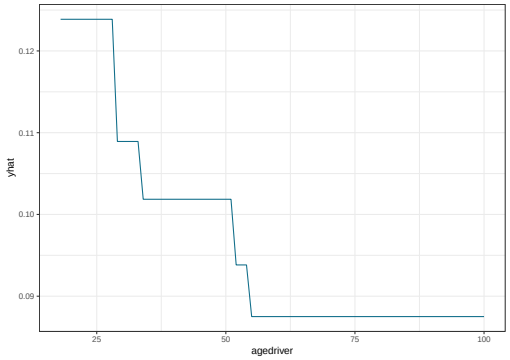
Figura 6.17: Árvore de Regressão - Modelo da Frequência de Sinistralidade.

de 12%, possui uma frequência de sinistralidade estimada baixa (0.06613) e que apenas uma pequena porção das mesmas, cerca de 7.8%, detém a frequência de sinistralidade estimada mais elevada (0.1526).

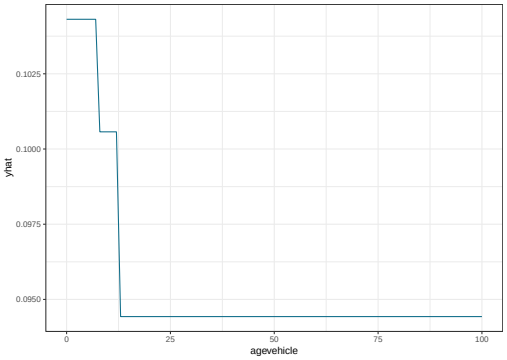
No que respeita à *performance* do modelo no conjunto de teste, observa-se que quando sujeito a um novo conjunto de dados, produz um valor de 0.325697 para a função desvio da *Poisson*, não muito diferente do valor da mesma função obtido para a combinação ótima encontrada, 0.31078.

Se se colocarem os fatores de risco num *ranking* de fatores mais relevantes, as primeiras posições serão ocupadas pela zona de residência, a idade do segurado e a marca do veículo. Os restantes fatores são de importância reduzida, mas ainda assim impactantes sobre o fenómeno em estudo.

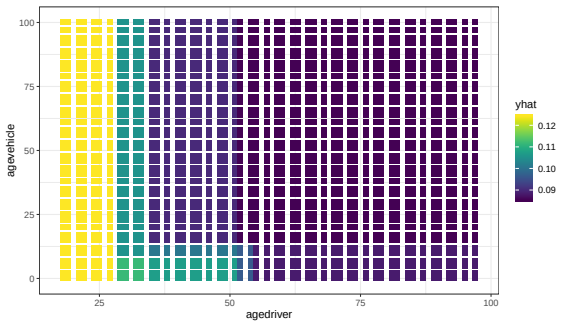
Apesar dos modelos obtidos por Árvores de Regressão serem de fácil leitura e interpretação, seguem-se os Gráficos de Dependências Parciais para os fatores de risco contínuos `agedriver` e `agevehicle`. A partir destes gráficos pode percecionar-se quais as idades mais propícias a sinistralidade. De facto, pelo gráfico 6.18a percebe-se que as camadas mais jovens de segurados possuem, em média, uma frequência de sinistralidade estimada mais elevada do que as restantes idades. Relativamente à idade do veículo seguro, figura 6.18b, o padrão repete-se, fazendo-se notar apenas a distinção de três categorias. Por outro lado, o gráfico da figura 6.18c permite perceber a influência conjunta que os fatores de risco `agedriver` e `agevehicle` exercem sobre o fenómeno em estudo. Segurados com idades compreendidas entre 18 e 30 anos, sensivelmente, que possuam um veículo com idade qualquer, detêm a maior frequência de sinistralidade estimada.



(a) Fator de Risco `agedriver`.



(b) Fator de Risco `agevehicle`.



(c) Fatores de Risco `agedriver` e `agevehicle`.

Figura 6.18: Gráficos de Dependências Parciais - Modelo da Frequência de Sinistralidade.

6.5.2 Modelação da Severidade dos Sinistros

Seja C_i o custo de um sinistro participado à Seguradora, $i \in \{1, 2, \ldots, n\}$.

Divisão dos Dados

O reduzido número de apólices que participaram sinistros à Seguradora, levou à alteração das percentagens anteriormente utilizadas para definir os conjuntos de treino e de teste. Definiu-se o conjunto de treino com 70% das apólices por amostragem estratificada, ficando o conjunto de teste com as restantes.

A otimização da Árvore de Regressão Gama passa pela aplicação do método *5-Fold Cross-Validation*, sendo necessário proceder à divisão do conjunto de treino em cinco subconjuntos D_1 a D_5, tais que as indemnizações presentes em cada um deles se distribuam de forma idêntica (tabela 6.20).

Tabela 6.20: Custo Médio dos Sinistros em cada Subconjunto do Conjunto de Treino.

	D_1	D_2	D_3	D_4	D_5
$\frac{1}{\lvert D_k \rvert}\sum_i C_i$	1 690.026 €	1 697.836 €	1 780.693 €	1 735.974 €	1 680.479 €

Novamente e atendendo ao reduzido número de indemnizações geradas, estipulou-se um mínimo de 10% de apólices em cada nó terminal.

O Modelo

A aplicação da Árvore de Regressão Gama pressupõe a otimização do único hiperparâmetro deste algoritmo, *cp*. A par com os subconjuntos D_1 a D_5, definiu-se o seguinte conjunto de possíveis valores para o hiperparâmetro *cp*, de forma a determinar o ótimo.

$$cp \in \{1.0 \times 10^{-5}, 2.0 \times 10^{-5}, \ldots, 1.0 \times 10^{-2}\}$$

O valor ótimo para o hiperparâmetro *cp* corresponde ao valor 8.09×10^{-3} e conduz ao modelo esquematizado na figura 6.19.

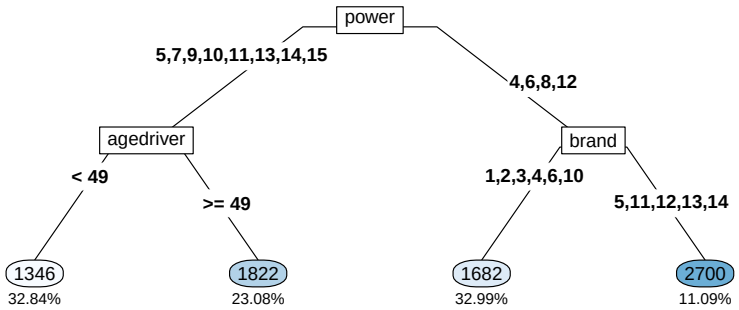


Figura 6.19: Árvore de Regressão - Modelo da Severidade dos Sinistros.

O perfil de menor risco respeita a um segurado com idade inferior a 49 anos cujo veículo possui uma potência 5, 7, 9, 10, 11, 13, 14 ou 15. Por sua vez, o perfil de maior risco pertence a veículos seguros que possuam potência 4, 6, 8 ou 12 e marca 5, 11, 12, 13 ou 14. De acordo com o modelo, estima-se que 32.84% das apólices originem indemnizações no valor médio de 1 346 € e apenas uma minoria, cerca de 11%, originem sinistros com um custo médio de 2 700 €.

A avaliação da *performance* do modelo obtido fez-se no conjunto de teste, verificando-se que o valor da função desvio *Gama* nesse mesmo conjunto foi 1.32559. O valor da mesma função para o valor ótimo do hiperparâmetro foi 1.361261, não distando significativamente do obtido para o conjunto de teste.

Mais uma vez, evidencia-se a carência de fatores de risco que tenham influência sobre a Severidade dos Sinistros, uma vez que os prejuízos a ressarcir pela Seguradora são influenciados também por variáveis e fatores extrínsecos não mensuráveis na tarifa. Assim sendo, os Gráficos de Dependências Parciais não têm especial interesse dada a simplicidade do modelo obtido.

6.6 Comparação das Abordagens

A comparação das abordagens seguidamente apresentada assenta nos Prémios Puros obtidos por cada uma delas. Desta forma, procedeu-se ao cálculo dos Prémios Puros de cada apólice da carteira considerando a Frequência de Sinistralidade e a Severidade dos Sinistros estimadas pelos modelos obtidos tanto por Modelos Lineares Generalizados como por Árvores de Regressão através da expressão:

$$PP_i = \mathbb{E}[N_i] \times \mathbb{E}[C_i], \quad i = 1, 2, \ldots, n.$$

No caso particular dos MLG, os Prémios Puros foram obtidos considerando a anuidade como período de exposição ao risco e sob o pressuposto de igual probabilidade de participar um "Grande" sinistro para todas as apólices em carteira.

A figura 6.20 ilustra a relação existente entre os Prémios Puros determinados por cada uma das abordagens estudadas. Os prémios encontram-se ordenados de forma ascendente.

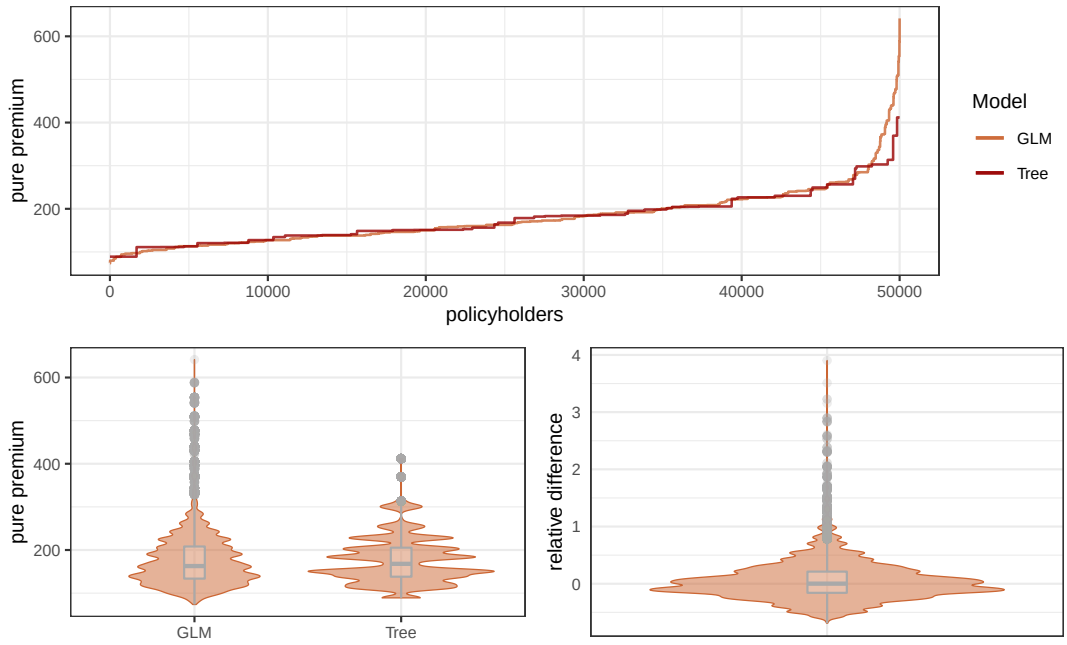


Figura 6.20: Comparação de Prémios Puros obtidos por ambas as Abordagens.

Pelo gráfico apresentado no painel superior verifica-se que os prémios obtidos por ambas as abordagens seguem a mesma tendência crescente. No painel inferior apresentam-se os Gráficos de Violino relativamente à dispersão dos Prémios Puros em carteira (esquerda) e às diferenças relativas entre os Prémios Puros (direita) obtidos pelas duas abordagens, definidas pela expressão (6.1).

$$RD_i = \frac{PP_i^{MLG} - PP_i^{Tree}}{PP_i^{Tree}}, \quad i = 1, 2, \ldots, n. \tag{6.1}$$

De notar que se $RD_i > 0$, $i = 1, 2, \ldots, n$, o Prémio Puro estimado para a apólice i por Modelos Lineares Generalizados é superior ao estimado por Árvores de Regressão. Da mesma forma se verifica o reverso da situação, quando $RD_i < 0$.

O gráfico da esquerda evidencia que os Modelos Lineares Generalizados são mais diferenciadores do risco do que as Árvores de Regressão na medida em que apresentam uma gama de prémios mais ampla. Apesar disso, e pelo gráfico da direita, verifica-se que as diferenças relativas entre os Prémios Puros obtidos pelas duas abordagens se situam em torno do zero, indicando que ambas as abordagens estimam Prémios Puros semelhantes para as apólices em carteira. As diferenças relativas mínima e máxima correspondem a -0.69 e 3.9, respetivamente. Em média, os prémios diferem cerca de 4.68 pontos percentuais entre si.

As tabelas 6.21a e 6.21b apresentam uma comparação entre os escalões tarifários de menor e maior risco, bem como os respetivos Prémios Puros obtidos por ambas as abordagens. Regra geral, observa-se que os escalões tarifários de menor e maior risco determinados pelas Árvores de Regressão estão alinhados com os determinados pelos Modelos Lineares Generalizados. A exceção verifica-se no escalão de menor risco, mais concretamente no nível tarifário correspondente à idade do segurado. Ao passo que os MLG identificam as idades do segurado 54 – 100 como as menos gravosas, as Árvores de Regressão caracterizam as idades no intervalo 29 – 48 como sendo as de menor risco.

Tabela 6.21: Comparação de Prémios Puros dos Perfis de Menor e Maior Risco.

(a) Perfil de Menor Risco.

	MLG	**Árvores de Regressão**
agedriver	54 – 100	29 – 48
agevehicle	15 – 100	Qualquer
zone	A,B	A,B,C
fuel	E	E
brand	12	1,4,12,14
power	Qualquer	5,7,9,10,11,13,14,15
Prémio Puro	73.66€	89.02€

(b) Perfil de Maior Risco.

	MLG	**Árvores de Regressão**
agedriver	18 – 25	18 – 33
agevehicle	4 – 7	Qualquer
zone	E	D,E,F
fuel	D	Qualquer
brand	10	5,11,13
power	Qualquer	4,6,8,12
Prémio Puro	641.54€	411.97€

7

Conclusão

A atividade seguradora subsiste da venda de contratos de seguro, nos quais visa a compensação económica do segurado no caso de participação de sinistro. A atribuição de um Prémio Puro a uma apólice deve refletir o risco que esta representa para a Seguradora e tem sido alvo de estudo nos últimos anos.

Em seguros de massas, particularmente o Seguro Automóvel, opta-se pela construção de uma estrutura tarifária, a partir da qual se determinam os Prémios Puros de todas as apólices em carteira. A estimação de tal estrutura passa por modelar a Frequência de Sinistralidade e a Severidade dos Sinistros, recorrendo ao histórico da Seguradora.

A construção de uma estrutura tarifária para o Seguro Automóvel de Responsabilidade Civil foi o principal foco do presente trabalho. Numa perspetiva comparativa, consideraram-se duas abordagens para a modelação: Modelos Lineares Generalizados e Árvores de Regressão.

Os Modelos Lineares Generalizados são utilizados em larga escala para modelar a tarifa das Seguradoras. A escolha desta abordagem em detrimento de outras, deve-se à sua abrangência para modelar fenómenos de variadas naturezas. Sucintamente, para a modelação de uma tarifa com Modelos Lineares Generalizados, há que definir um Segurado Padrão, ajustar uma distribuição de probabilidade ao Número de Sinistros e aos seus Custos e identificar os fatores de risco mais influentes sobre cada um dos fenómenos referidos. As contrariedades desta abordagem situam-se entre o ajustamento de uma distribuição e a modelação dos fenómenos propriamente dita. De facto, encontrar uma distribuição que se adeqúe ao Número de Sinistros e aos seus Custos pode não ser acessível uma vez que a aplicação de Modelos Lineares Generalizados requer que a mesma pertença à Família Exponencial. Por outro lado, é frequente efetuar-se a divisão dos Custos em dois conjuntos respeitantes aos sinistros ditos "Regulares" e "Grandes", modelando-se cada um de forma individual. Na maioria das vezes é necessário recorrer a Teoria de Valores Extremos para modelar os custos relativos ao segundo conjunto de sinistros, dado o reduzido número de apólices nessas condições e a dificuldade de enquadrar uma distribuição de probabilidade que pertença à Família Exponencial.

Os fatores tarifários que integram uma estrutura tarifária são, habitualmente, categóricos. Existindo fatores tarifários que não sejam, procede-se à categorização dos mesmos para que os riscos inerentes a diferentes categorias possam ser refletidos na aplicação de MLG. Na iminência de definir excessivos níveis tarifários para os fatores de risco contínuos, utilizaram-se Árvores de Regressão de *Poisson* para a categorização de tais fatores, modelando a Frequência de Sinistralidade em função de cada um dos fatores de risco. Os níveis tarifários obtidos foram os esperados, evidenciando a menor ou maior propensão a sinistralidade em cada nível.

A metodologia de *Machine Learning* escolhida para a estimação de uma tarifa foi Árvores de Regressão. Esta escolha alinha-se com o objetivo principal de uma tarifa: segmentar a carteira em subgrupos de risco homogéneos. Apesar da aplicação desta abordagem resultar em Prémios Puros bastante similares aos obtidos pelos Modelos Lineares Generalizados, é necessário acautelar que possíveis mudanças no conjunto de dados podem implicar drásticas alterações no modelo proposto. As Árvores de Regressão são, por isso, o método de *Machine Learning* mais instável em termos de resultados.

Os métodos de *Machine Learning* têm despertado o interesse de muitos autores, tendo-se assistido a tentativas de os incorporar no ramo Segurador. Estes métodos libertam a modelação dos detalhes estatísticos como o ajuste de distribuição teórica e a divisão dos sinistros em duas classes, essenciais nos Modelos Lineares Generalizados. No entanto, à medida que aumenta o grau de complexidade do algoritmo de *Machine Learning*, diminui a possibilidade de interpretar o modelo produzido. Querendo manter a interpretabilidade do modelo, os modelos de *Machine Learning* mais robustos como *Random Forests* e *Gradient Boosting Machines* vêem-se impedidos de operar na estimação de uma tarifa.

O trabalho apresentado pretende contribuir para a incorporação de métodos de *Machine Learning* no ramo Segurador. Tentou estabelecer-se uma comparação entre a abordagem tradicional, por Modelos Lineares Generalizados, e a abordagem alternativa, por Árvores de Regressão. Do leque de algoritmos de ML, este último é o que produz os modelos mais interpretáveis, contudo peca pela volatilidade dos mesmos.

Futuramente, poderão considerar-se os métodos de *Machine Learning* mais complexos mencionados para identificar os fatores de risco mais influentes sobre a Frequência de Sinistralidade e a Severidade dos Sinistros, para posterior integração dos mesmos em Modelos Lineares Generalizados.

Bibliografia

Abdi, H. & Williams, L. J. (2010). Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design*, *3*(1), 1–5.

Boehmke, B. & Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC Press.

Brazauskas, V. & Kleefeld, A. (2009). Robust and efficient fitting of the generalized pareto distribution with actuarial applications in view. *Insurance: Mathematics and Economics*, *45*(3), 424–435.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. Chapman e Hall/CRC.

De Jong, P. & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American statistical Association*, *53*(284), 789–798.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2001). *The Elements of Statistical Learning*. Springer series in statistics New York.

Garrido, J., Genest, C. & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, *70*, 205–215.

Grubinger, T., Zeileis, A. & Pfeiffer, K.-P. (2014). Evtree: Evolutionary learning of globally optimal classification and regression trees in r. *Journal of statistical software*, *61*(1), 1–29.

Guerreiro, G. R. (2001). *Uma Abordagem Alternativa para Bonus Malus* (tese de mestrado, FCT-NOVA).

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. CRC press.

Henckaerts, R. (2021). *Distrforest: Distribution-based random forest*. https://github.com/henckr/distRforest.

Henckaerts, R., Antonio, K., Clijsters, M. & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(8), 681–705.

Henckaerts, R., Côté, M.-P., Antonio, K. & Verbelen, R. (2020). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 1–31.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kuhn, M. (2014). Futility analysis in the cross-validation of machine learning models. *arXiv preprint arXiv:1405.6974*.

Kuo, K. & Lupton, D. (2020). Towards Explainability of Machine Learning Models in Insurance Pricing. *arXiv preprint arXiv:2003.10674*.

Lemaire, J. (2012). *Bonus-malus systems in automobile insurance*. Springer science & business media.

Nelder, J. A. & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Ohlsson, E. & Johansson, B. (2010). *Non-life Insurance Pricing with Generalized Linear Models*. Springer.

Ramseyer, G. C. & Tcheng, T.-K. (1973). The robustness of the studentized range statistic to violations of the normality and homogeneity of variance assumptions. *American Educational Research Journal*, *10*(3), 235–240.

Staudt, Y. & Wagner, J. (2019). *Comparison of Machine Learning and Traditional Severity-Frequency Regression Models for Car Insurance Pricing*. Working Paper. Lausanne: University of Lausanne.

Therneau, T. M., Atkinson, E. J. et al. (1997). *An introduction to recursive partitioning using the rpart routines*. Technical report Mayo Foundation.

Turkman, M. A. A. & Silva, G. L. (2000). *Modelos Lineares Generalizados – da teoria à prática*. Sociedade Portuguesa de Estatística.

Villaseñor, J. A. & González-Estrada, E. (2015). A variance ratio test of fit for gamma distributions. *Statistics & Probability Letters*, *96*, 281–286.

Villaseñor-Alva, J. A. & González-Estrada, E. (2009). A bootstrap goodness of fit test for the generalized pareto distribution. *Computational statistics & data analysis*, *53*(11), 3835–3841.

A

Fatores de Risco

Tabela A.1: Descrição dos Fatores de Risco - Carteira de Seguro Automóvel.

Fator de Risco	**Descrição**
`coverage`	Tipo de cobertura subscrita. 1RC (Responsabilidade Civil), 2DO, 3VI, 4BG, 5CO, 6CL.
`zone`	Zona de residência do segurado, de acordo com N.º Hab./km^2. A (1 a 50), B (51 a 100), C (101 a 500), D (501 a 2 000), E (2 001 a 10 000), F (Superior a 10 000).
`fuel`	Tipo de combustível do veículo seguro. D (Diesel), E (Gasolina).
`brand`	Marca do veículo seguro. 1 (Renault, Nissan), 2 (Peugeout, Citroën), 3 (Volkswagen, Audi, Scoda, Seat), 4 (Opel, GM), 5 (Ford), 6 (Fiat), 10 (Mercedes, Chrysler), 11 (BMW, Mini), 12 (Japoneses e Coreanos), 13 (Outros Europeus), 14 (Outras Marcas).
`power`	Potência do veículo seguro. Categorizada de 4 a 15.
`agedriver`	Idade do segurado, em anos.
`agevehicle`	Idade do veículo, em anos.
`bonus`	Percentagem de desconto ou agravamento correspondente à classe do sistema de *Bonus-Malus* integrada.

B

Packages

O *software* utilizado para a implementação das abordagens exploradas neste documento foi o R (versão 3.6.1), tendo-se recorrido aos *packages* que se seguem:

- `tidyverse` (versão 1.3.0);
- `vcd` (versão 1.4.8);
- `goft` (versão 1.3.6);
- `fitdistrplus` (versão 1.0.4);
- `multcomp` (versão 1.4.15);
- pROC (versão 1.17.0);
- `splitTools` (versão 0.3.1);
- `rpart` (versão 4.1.15);
- `distRforest` (versão 1.0.0);
- `pdp` (versão 0.7.0).