NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

SOFIA RAMOS DE MOURA RIBEIRO

Licenciada em Matemática

ESTATÍSTICA DE VALORES EXTREMOS NA GESTÃO DO RISCO

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Universidade NOVA de Lisboa

Setembro, 2022

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTAMENTO DE
MATEMÁTICA

ESTATÍSTICA DE VALORES EXTREMOS NA GESTÃO DO RISCO

SOFIA RAMOS DE MOURA RIBEIRO

Licenciada em Matemática

Orientador: Frederico Almeida Gião Gonçalves Caeiro
Associate Professor, NOVA University Lisbon

Coorientadora: Ayana Maria Xavier Furtado Mateus
Auxiliar Professor, NOVA University Lisbon

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

Universidade NOVA de Lisboa
Setembro, 2022

Estatística de Valores Extremos na Gestão do Risco

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo NOVAthesis (v6.9.5) [1].

Aos meus.

Agradecimentos

Esta dissertação consistiu num percurso longo e desafiante, com alguns precalços, que sem as palavras certas nos momentos certos, não teria sido possível concluir com sucesso.

Em primeiro lugar, gostava de agradecer aos meus orientadores, Professor Frederico Caeiro e Professora Ayana Mateus, pela competência e rigor, pela dedicação e disponibilidade demonstradas para responder às minhas dúvidas. A orientação que me concederam foi incansável, e por isso, o meu mais sincero obrigado.

À minha família, especialmente aos meus pais e irmãos, agradeço a paciência e o carinho, sem vocês este percurso não seria de todo possível. Obrigado por sempre me apoiarem e motivarem, nunca me deixando duvidar das minhas capacidades, foi o vosso entusiasmo e encorajamento que me levaram à conclusão deste trabalho. São os melhores do mundo !

Aos meus amigos de sempre, Joana, Kika, Zé e Bea, obrigada pelo forte apoio que sempre me transmitiram e por animarem os meus dias, principalmente os mais difíceis. Aos que fui encontrando pelo caminho graças à Matemática e ficaram amigos para a vida, gostava de referir alguns, Palhinhas, Kika, Fátima, Patrícia, Diana, Catarina, Ricardo, Nuno, Marcelino, Rita, Joana e André, obrigada por terem sido elementos chave no meu percurso académico e nunca me deixarem desamparada. A vossa ajuda foi essencial para chegar aqui e sem ela, estou certa que teria sido muito mais difícil.

Por último, queria agradecer aos meus avós, Mimi, Chico e Adelaide, embora já não seja possível festejarmos juntos esta vitória, foram vocês que me incutiram os valores e os princípios que me tornaram na pessoa que sou hoje. Em tudo, foram exemplos na minha vida, obrigada.

Resumo

Os eventos extremos são acontecimentos raros mas que podem causar grandes catástrofes em diversos contextos, como por exemplo no contexto ambiental ou no setor financeiro. Por isso, é importante estimá-los com alguma precisão. Assim, surge a necessidade de recorrer à Teoria de Valores Extremos, teoria esta que se debruça na análise e estimação destes eventos de caráter excecional.

Neste trabalho, com o intuito de realçar a importância da Teoria de Valores Extremos na gestão do risco, procedeu-se à análise de um conjunto de dados referentes a apólices de seguro no ramo automóvel, fornecidos por um pacote do software R studio. Desta forma, após um breve estudo da base de dados que serviu de suporte a este projeto, são apresentados sucintamente os resultados principais da Teoria de Valores Extremos. É de salientar a importância da modelação da cauda direita da distribuição subjacente aos dados, uma vez que é nesta que têm lugar os ditos eventos extremos. Nesse sentido, são apresentadas uma abordagem paramétrica e outra semi paramétrica da Teoria de Valores Extremos, sendo posteriormente aplicadas aos dados em estudo.

Palavras-chave: Teoria de Valores Extremos, Distribuição Generalizada de Valores Extremos, Excessos acima de um nível elevado, Índice de valores extremos, Seguro automóvel.

Abstract

Extreme events are rare but they can cause major catastrophes in several contexts, for example in the environmental context or in the financial sector. Therefore, it is important to estimate them with some precision. Thus, the need arises to resort to the Extreme Value Theory, a theory that focuses on the analysis and estimation of these exceptional events.

To highlight the importance of the Extreme Value Theory in risk management, in this work, a dataset of automobile insurance policies, provided by a package of the R studio software, was analysed. After a brief study of the database that supported this project, the main results of the Extreme Value Theory are presented. The importance of modelling the right tail of the distribution underlying the data should be highlighted, since it is in this tail that the so-called extreme events take place. In this regard, a parametric and a semi-parametric approach of the Extreme Value Theory are presented and then applied to the data under study.

Keywords: Extreme Value Theory, Generalized Extreme Value Distribution (GEV), Peaks Over Threshold, Extreme Value Index, Automobile insurance.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

A Teoria dos Valores Extremos também conhecida por *Extreme Value Theory (EVT)*, tem como objetivo a modelação de fenómenos aleatórios de ocorrência rara mas que têm um impacto significativo na sociedade. Desta forma, a EVT procura analisar e estimar a probabilidade de ocorrerem eventos extremos que se traduzem em acontecimentos cujos valores se desviam significativamente da média da amostra completa, tanto superior como inferiormente. Assim, o principal objetivo da Estatística de extremos passa pela previsão destes acontecimentos.

De forma a estudar as metodologias associadas à EVT foram analisados alguns artigos, sendo importante referir que apenas foram consideradas abordagens paramétricas ou semi paramétricas. A distinção destas, reside em que no primeiro caso, a distribuição a utilizar na modelação é conhecida, enquanto que no segundo caso, esta não é totalmente conhecida e o que se tem é apenas uma aproximação.

No artigo [2], é apresentada uma abordagem paramétrica da EVT conhecida por método dos excessos acima de um limiar que se baseia na análise de valores extremos de uma amostra que excedem um determinado limiar. Neste trabalho, é realizada uma aplicação do mesmo a um conjunto de dados de seguros pertencentes a uma seguradora dos Estados Unidos da América, utilizando várias formas de encontrar e estabelecer um limiar adequado, bem como analisar e modelar a amostra.

Por outro lado, o artigo [3], aborda uma metodologia semi paramétrica da EVT e mostra um exemplo desta aplicada a um conjunto de dados referentes a taxas de conversão diárias de dólares canadianos/americanos durante um período de 5 anos (1995-2000).

Assim, conclui-se que a EVT desempenha um papel importante em diversas áreas, sendo importante referir que a gestão do risco para seguros é uma delas, tal como se pode verificar no artigo [4] que se foca no estudo dos rácios de perdas anuais para seguros contra sismos na Califórnia entre 1971 e 1993. É de salientar a extrema importância de

equilibrar a probabilidade da ocorrência de acidentes extremos com as possíveis consequências que estes podem originar, de modo a que não haja perdas para a Seguradora. Assim, no setor dos seguros, um dos objetivos usualmente traçados consiste na modelação da frequência de acidentes com indemnizações elevadas em automóveis, casas ou outros investimentos de alto valor.

Esta dissertação tem início com uma breve descrição da base de dados em estudo na secção 2.2, onde é detalhado o percurso que é necessário fazer para chegar a este conjunto de dados, assim como são também apresentadas as variáveis presentes.

De seguida, na secção 2.3 é realizada uma análise exploratória dos dados com o intuito de compreender o comportamento de cada uma das variáveis em estudo, assim como as relações entre estas, de forma a analisar a sua independência. Neste capítulo, é ainda realizado o ajuste de algumas distribuições às variáveis que quantificam o número de sinistros e o montante médio das indemnizações. No primeiro caso, uma vez que o número de sinistros se trata de uma variável discreta, foi feito o ajuste das distribuições de Poisson, Binomial, Geométrica e Binomial Negativa. Enquanto que, no segundo caso, procedeu-se ao ajuste das distribuições Gama e LogNormal, pois a respetiva variável é contínua. Contudo, uma vez que, neste último caso, não se conseguiu encontrar um bom modelo para a amostra completa, pretendeu-se ajustar um modelo apenas à cauda da distribuição.

Desta forma, recorreu-se à Teoria dos Valores Extremos e a alguns resultados teóricos associados tais como o Teorema de Gnedenko e o Teorema de Pickands e Balkema-De Haan, que são abordados no capítulo 3. Neste capítulo são ainda apresentadas detalhadamente uma abordagem paramétrica, nomeada por *Excessos acima de um limiar*, e outra, semi paramétrica, da EVT, que, por fim, foram utilizadas no exemplo de aplicação referente à gestão de risco no setor segurado em ramo automóvel, apresentado no capítulo 4.

Em ambos os casos, o primeiro desafio passa pela escolha do limiar. Uma vez que, na abordagem paramétrica, se adota a distribuição Generalizada de Pareto como modelo limite para a função de distribuição em estudo, para a seleção do limiar recorreu-se à Regra de Ouro e a métodos gráficos, mais especificamente, ao Gráfico de Vida Residual Média e ao Gráfico da Estabilidade das Estimativas dos Parâmetros. Após isto, procedeu-se à estimação dos parâmetros de forma e escala da distribuição (ξ e σ, respetivamente) por máxima verosimilhança, calcularam-se os níveis de retorno a 5, 10 e 15 anos e terminou-se a aplicação com a verificação do modelo através de gráficos de probabilidade, de quantis, de níveis de retorno e de densidade. Por outro lado, na abordagem semi paramétrica utilizou-se o Gráfico de Hill para a escolha do limiar, visto que apenas se tem uma aproximação da distribuição a utilizar na modelação. De seguida, recorrendo ao estimador de Hill, foi calculada uma estimativa para o parâmetro de forma, assim como outros parâmetros de interesse e de grande utilidade na análise de valores extremos, tais como a estimativa de uma probabilidade de excedência de um nível elevado e a estimativa de quantis extremos.

Capítulo 2

Descrição e análise preliminar dos dados

2.1 Introdução

Ao longo deste capítulo é feita uma breve descrição dos dados na secção 2.2 seguida de uma análise exploratória dos dados que pode ser consultada na secção 2.3, cujo objetivo consiste em compreender o comportamento de cada uma das variáveis presentes, tanto a nível individual como analisar possíveis relações entre estas, pois um ponto a ter em conta é saber se as variáveis em estudo são independentes. Para finalizar, nas secções 2.3.1 e 2.3.2, são apresentados os resultados do estudo realizado do ajuste das distribuições ao número de sinistros e montante médio das indemnizações, respetivamente.

2.2 Descrição da base de dados

O pacote *CASdatasets 1.0-11* [5] do R é constituído por um conjunto variado de bases de dados atuariais, que são utilizadas no livro *Computational Actuarial Science with R* [6]. Uma vez que este pacote não se encontra no *Comprehensive R Archive Network(CRAN)*, não se instala como a maioria dos pacotes usuais. Assim, neste caso, é necessário aceder ao link *http://cas.uqam.ca/*, instalar os pacotes *xts*, *sp* e *zoo* e inserir a seguinte linha de código no R: *install.packages("CASdatasets", repos = "http://cas.uqam.ca/pub/", type="source")*. Após isto, já é então possível aceder às bases de dados deste pacote. A base de dados em estudo, cujo nome é *norauto*, pode ser encontrada neste pacote do R e refere-se à sinistralidade do ramo automóvel de uma companhia de seguros da Noruega.

O conjunto de dados em análise é constituído por 183 999 apólices de seguro automóvel relativos a um período de um ano, sendo que para cada uma destas observações foram registadas 7 características distintas:

1. *Male:* variável binária, em que 1 corresponde ao tomador do seguro ser homem e 0

caso contrário;

2. *Young:* variável binária, em que 1 significa que a idade do tomador do seguro é inferior a 26 anos e 0 caso contrário;

3. *Distlimit:* variável discreta, referente ao limite da distância como indicado no contrato de seguro, ou seja, a distância percorrida em quilómetros que é coberta pelo seguro;

4. *GeoRegion:* variável discreta que refere a densidade da região geográfica de onde o tomador do seguro pertence, sendo *High+* a opção mais alta e *Low-* a opção mais leve;

5. *NbClaim:* variável discreta que representa o número de sinistros por apólice, neste caso, para uma apólice pode haver 0,1,2 ou 3 acidentes;

6. *Expo:* variável contínua que é uma medida base do risco, a partir da qual se pretende mensurar o volume de risco assumido pela Seguradora, medido como fração do ano;

7. *ClaimAmount:* variável contínua que assume o valor 0 ou o montante médio da indemnização se $NbClaim > 0$, ou seja, o valor do montante pago pela seguradora.

2.3 Análise exploratória dos dados

O estudo das variáveis da base de dados utilizada é bastante importante para clarificar os objetivos a cumprir, por isso, começou-se por analisar cada uma destas.

De forma a tentar compreender o comportamento da variável *Male*, foi construído o gráfico que pode ser consultado na figura seguinte.

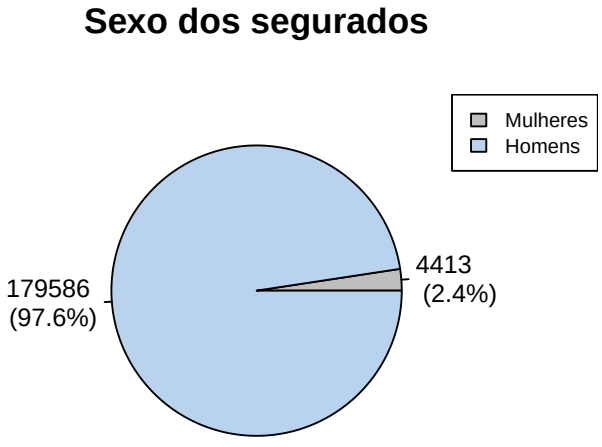


Figura 2.1: Gráfico circular do sexo dos segurados

No gráfico da figura 2.1 verifica-se que a maioria dos segurados são homens, sendo a percentagem de 97.6% correspondente a 179 586 apólices. A percentagem de mulheres é de 2.4%, o que significa que dos segurados, 4 413 são do sexo feminino.

De forma análoga, para a variável *Young*, com o intuito de perceber como se distribui a idade dos segurados neste estudo, elaborou-se a figura seguinte.

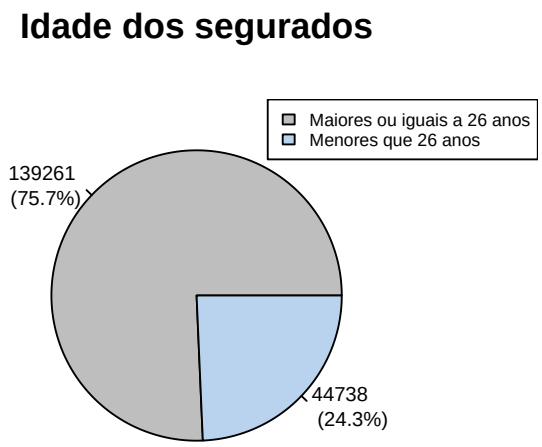


Figura 2.2: Gráfico circular da idade dos segurados

Ao observar o gráfico da figura 2.2, é possível concluir que 24.3%, ou seja, quase um quarto dos segurados deste estudo, correspondem a 44 738 segurados que têm menos de 26 anos de idade. Desta forma, os restantes 75.7% dizem respeito aos 139 261 tomadores de seguro que têm 26 anos de idade ou mais.

Quanto à variável *Distlimit*, foi determinado o número de observações para cada um dos limites de distâncias seguradas e, posteriormente, elaborado o gráfico de barras com as respetivas frequências absolutas que pode ser consultado na figura 2.3.

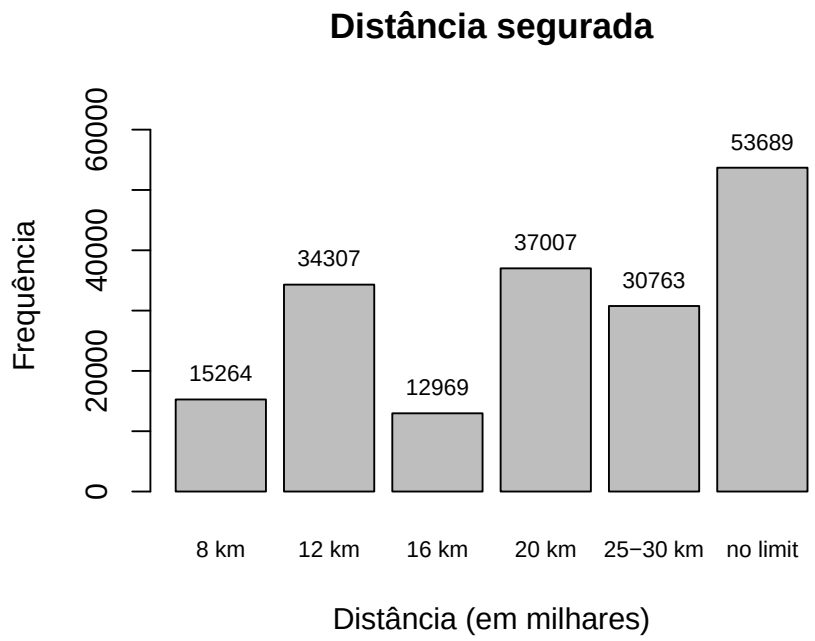


Figura 2.3: Gráfico de barras da distância segurada (em milhares)

Ao analisar o gráfico da figura 2.3, conclui-se que a classe predominante neste estudo é aquela que não tem limite de distância segurada, o que significa que a maioria das apólices presentes não considera uma distância limite no contrato de seguro efetuado. Por outro lado, a que tem o menor número de observações em estudo, é a que corresponde a um limite de distância segurada de 16 000 km.

Para a variável *GeoRegion*, procedeu-se de forma análoga ao estudo da variável *DistLimit*. Assim, na figura seguinte, é possível observar as diferentes densidades das regiões dos tomadores de seguro em estudo, com as respetivas frequências.

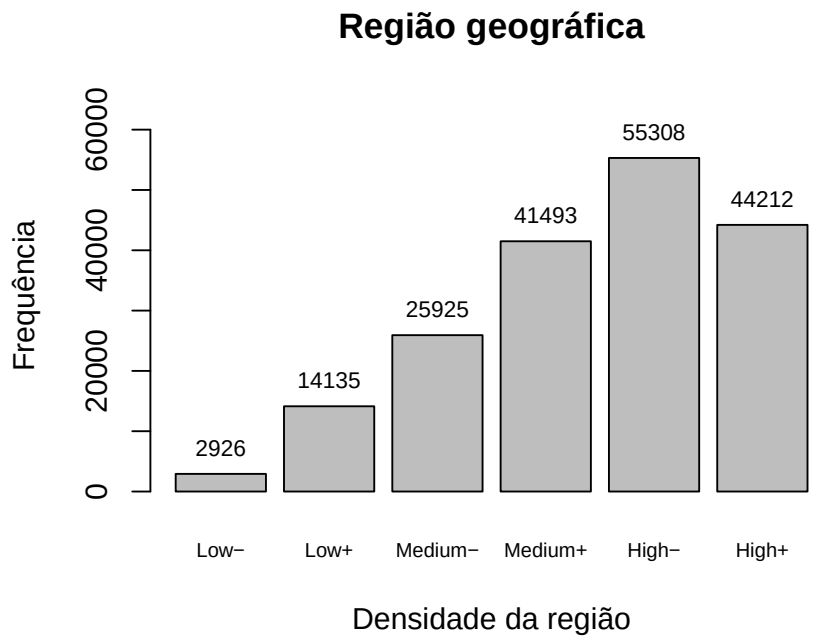


Figura 2.4: Gráfico de barras da densidade das regiões dos segurados

Ao observar o gráfico da figura 2.4, é possível afirmar que a maior parte dos segurados pertence a uma região geográfica cuja densidade é *High-*, zona esta de densidade alta, tal como o nome indica. Contudo, é importante relembrar que a classe que considera as regiões de densidades mais elevadas é nomeada por *High+* e corresponde à segunda classe mais numerosa nesta análise. Em contrapartida, a classe que apresenta menor número de observações é a que se refere à densidade geográfica mais baixa, a *Low-*. Assim, é possível concluir que neste estudo, a maioria dos tomadores de seguro vivem em zonas cuja densidade geográfica é mais densa.

Com o objetivo de compreender o comportamento da variável *NbClaim*, construiu-se o gráfico de barras apresentado na figura 2.5.

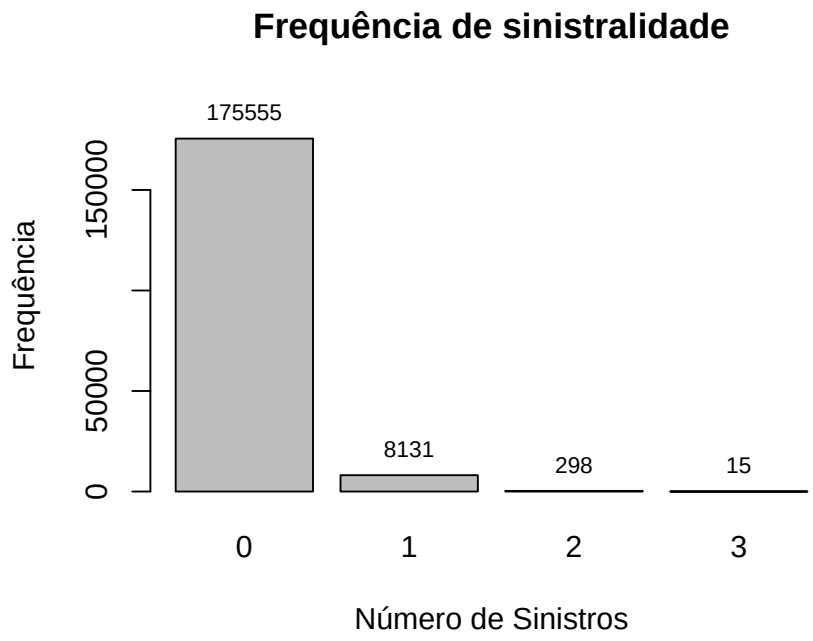


Figura 2.5: Gráfico de barras da frequência de sinistralidade

Analisando o gráfico da figura 2.5, é possível observar o número de apólices que reportaram 0, 1 ,2 ou 3 sinistros, respetivamente. Verifica-se que a grande maioria das apólices, cerca de 95%, não reportaram qualquer tipo de sinistro e as que o fizeram, reportaram na maioria apenas 1 sinistro. Por outro lado, apenas 15 das 183 999 apólices em estudo, tiveram 3 sinistros, tornando assim esta, a categoria com menos observações.

Relativamente à variável quantitativa *Expo*, que representa o potencial para um sinistro, calcularam-se algumas estatísticas que podem ser consultadas na tabela 2.1.

Tabela 2.1: Algumas estatísticas da Exposição ao Risco

Mínimo	1ºQuartil	Desvio Padrão	Mediana	3ºQuartil	Média	Máximo
0.101	0.581	0.911	0.956	1.003	1.006	25.849

Através da tabela 2.1, verifica-se que embora a variável em estudo apresente valores no intervalo $[0.101, 25.849]$, estes encontram-se concentrados perto do valor 1, sugerindo que o valor do extremo superior do intervalo tenha sido um acontecimento esporádico.

De forma a compreender a dispersão dos dados no que diz respeito a esta variável, foi também elaborado o diagrama de extremos e quartis que pode ser observado na figura 2.6.

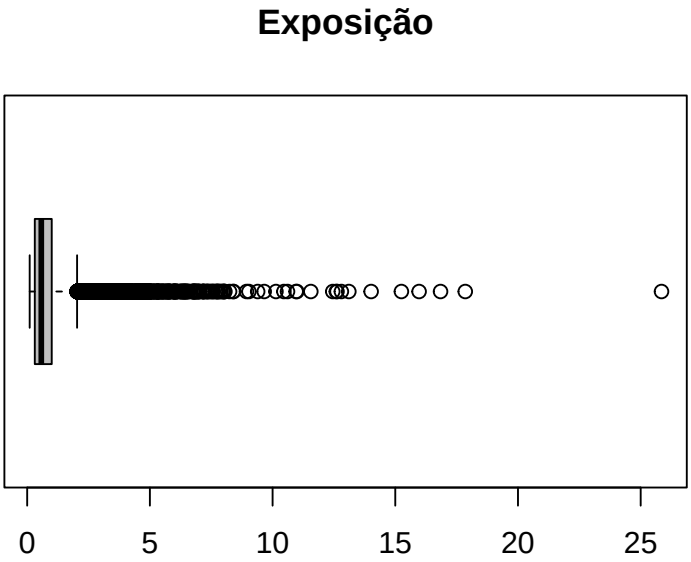


Figura 2.6: Diagrama de extremos e quartis da exposição dos segurados ao risco

Ao analisar o diagrama da figura 2.6, verifica-se que tanto os valores mínimo e máximo, como o primeiro, segundo e terceiro quartil são valores baixos, sendo visível uma ligeira maior dispersão dos dados acima da mediana. É também possível observar a existência de vários outliers.

Semelhante ao que foi feito para a variável quantitativa *Expo*, também para a variável *ClaimAmount* foram determinadas as respetivas estatísticas para as indemnizações superiores a 0 e apresentadas na tabela 2.2.

Tabela 2.2: Algumas estatísticas do montante médio das indemnizações

Mínimo	1ºQuartil	Mediana	Média	3ºQuartil	Desvio Padrão	Máximo
1	8 554	16 632	23 588	27 387	28 829	700 000

Observando a tabela 2.2, conclui-se que os valores que a variável em estudo apresenta estão presentes no intervalo $[1, 700000]$. Contudo, tal como se sucede para a variável *Expo*, é também possível afirmar que a maioria destes valores são relativamente baixos, quando comparados com o valor do extremo superior do intervalo.

Posteriormente, foi construído o diagrama de extremos e quartis com o intuito de estudar a dispersão dos valores desta variável. Este pode ser observado na figura 2.7.

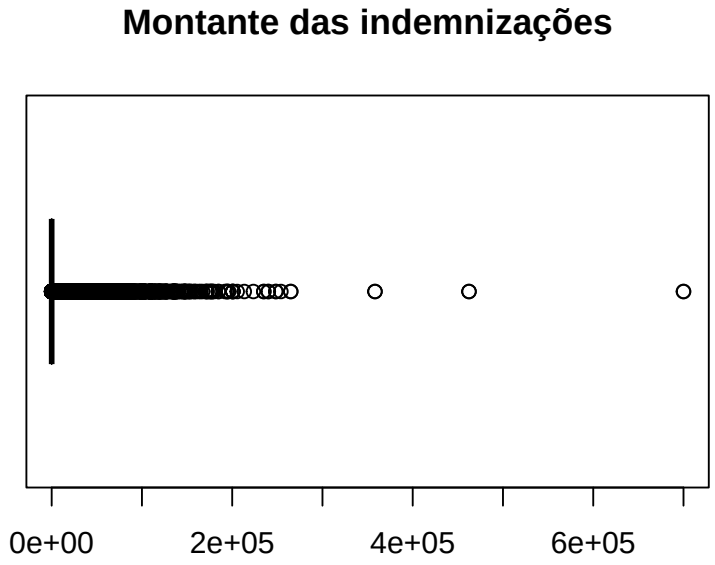


Figura 2.7: Diagrama de extremos e quartis do montante médio das indemnizações em estudo

A partir da figura 2.7, é possível concluir que as indemnizações são na sua grande maioria baixas, dentro dos mesmos valores. Por outro lado, o gráfico apresenta vários outliers no extremo superior, o que sugere a ocorrência de acidentes que resultaram em indemnizações bastante altas.

De forma a corroborar os resultados obtidos da figura 2.7, apresentam-se na tabela 2.3 os quantis amostrais do montante médio das indemnizações em estudo.

Tabela 2.3: Quantis amostrais dos custos dos sinistros

50%	90%	95%	99%	99.9%
16 632	49 000	72 096	134 846	251 494

Ao analisar a tabela 2.3, é possível concluir que a grande maioria das indemnizações (95%) se encontra entre 1 e 72 096, o que justifica a média e o desvio padrão observados na tabela 2.2.

Continuando o estudo da variável referente ao valor das indemnizações, construiu-se o histograma onde é apresentado o número de sinistros por cada intervalo de custos (ver figura 2.8).

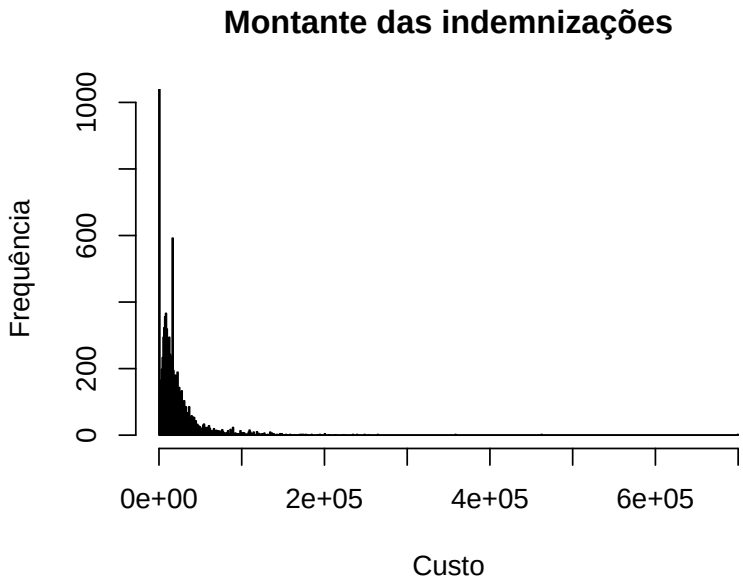


Figura 2.8: Histograma do montante médio das indemnizações (u.m.)

Ao observar a figura 2.8, é possível concluir que a grande maioria dos sinistros gerou um custo para a seguradora inferior a 100 000u.m., sendo notório que o acontecimento mais recorrente é um sinistro sem custos para a mesma. É também visível a existência de uma frequência que sobressai perante as restantes, trata-se do número de sinistros em estudo que deram origem a uma indemnização de 16 999u.m.. Note-se que, tanto o histograma da figura 2.8 como o gráfico da figura 2.7, evidenciam uma acentuada assimetria à direita como expectável.

É de salientar que apenas 8 444 dos 183 999 sinistros reportados, deram origem a indemnizações para a seguradora. Posto isto, na figura 2.9, observam-se os sinistros cuja

indemnização é superior a 0.

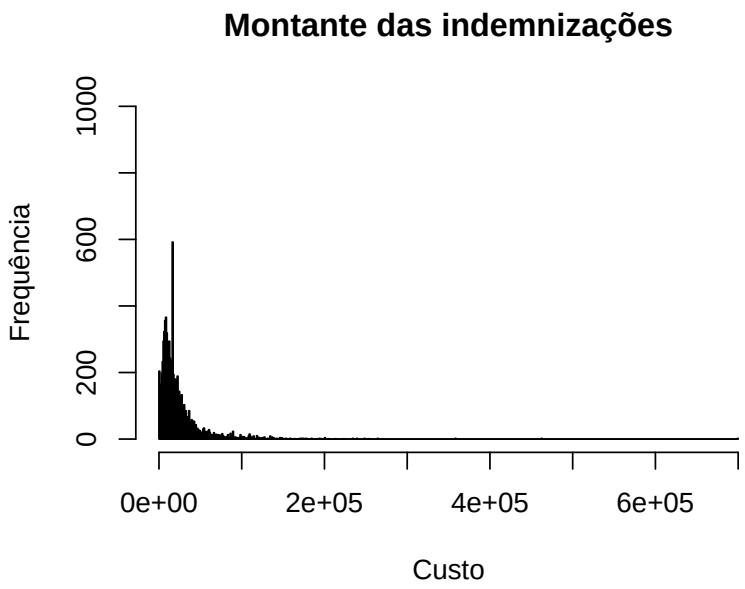


Figura 2.9: Histograma dos sinistros que deram origem a indemnizações

Após a análise de cada uma das variáveis em estudo, surge a necessidade de averiguar algumas possíveis relações entre estas. É de salientar que nos modelos atuariais é considerada a independência entre as variáveis aleatórias referentes ao número de sinistros e ao montante médio dos mesmos.

Desta forma, começou-se por estudar a possível relação linear entre as variáveis numéricas *Expo*, *ClaimAmount* e *NbClaim*. Para isto, calcularam-se os coeficientes de correlação entre cada par destas variáveis que podem ser consultados na Tabela 2.4. É de salientar que este coeficiente toma única e exclusivamente valores entre -1 e 1, sendo que só toma exatamente estes valores extremos do intervalo quando a relação é perfeita negativa ou positiva respetivamente, já quando a relação é difusa ou não linear o coeficiente é 0.

Tabela 2.4: Coeficientes de correlação entre variáveis numéricas

Coeficiente de correlação	Expo	ClaimAmount	NbClaim
Expo	1	-0.004	0.134
ClaimAmount	-0.004	1	-0.005
NbClaim	0.134	-0.005	1

Como é possível observar na tabela 2.4, os coeficientes de correlação entre cada par de variáveis numéricas distintas, estão perto de 0, o que significa que a correlação entre estas variáveis não é significativa. Para corroborar esta afirmação, elaborou-se o gráfico de dispersão, apresentado na figura 2.10.

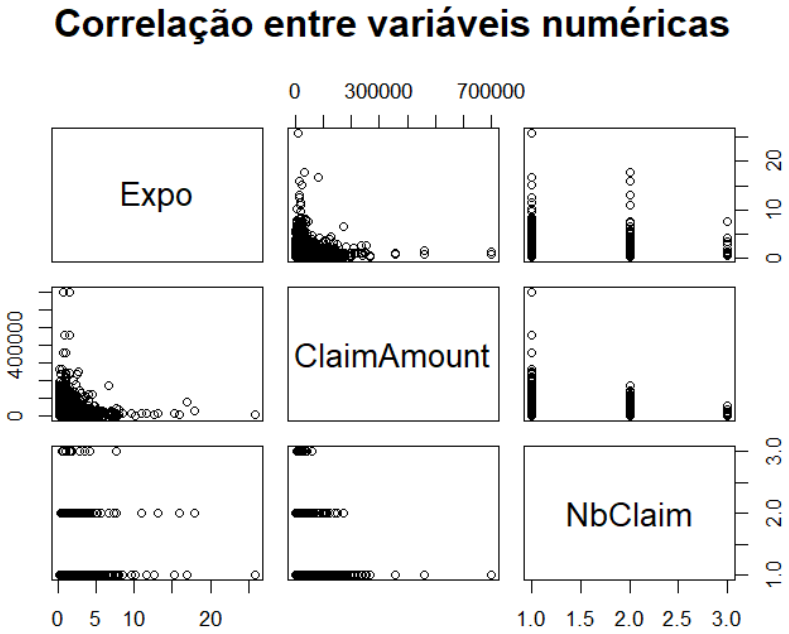
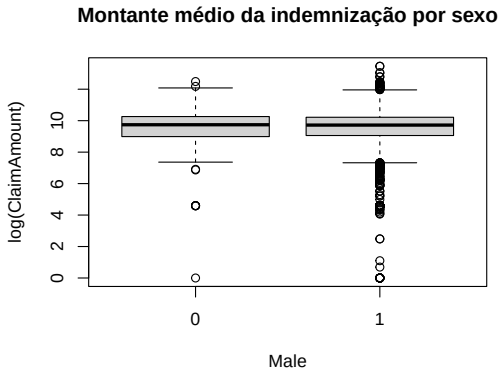


Figura 2.10: Gráficos de dispersão entre as variáveis numéricas em estudo

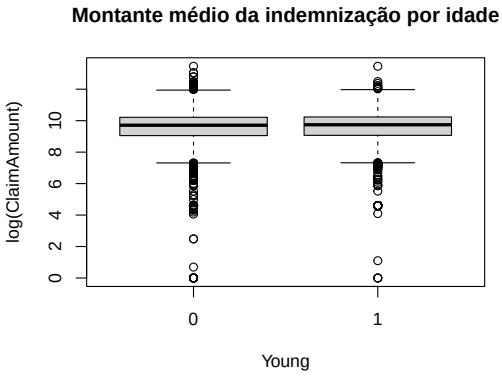
Ao observar o gráfico 2.10, este sugere a existência de um padrão entre as variáveis *Expo* e *ClaimAmount*, porém não linear. Desta forma, pode afirmar-se que não existe uma relação linear entre as três variáveis numéricas em estudo.

De seguida, para analisar o comportamento da variável numérica *ClaimAmount* relativamente às variáveis categóricas, ou seja, para perceber como esta variável se distribui dentro de cada uma das diferentes categorias, recorreu-se a gráficos de extremos e quartis. É importante referir a grande amplitude do intervalo de valores, que a variável que diz respeito ao montante médio das indemnizações apresenta. Isto, porque uma vez que a grande maioria destes valores são baixos e de forma a não serem comprimidos no fundo do gráfico, foi necessário recorrer à escala logarítmica, que torna mais visível e, consequentemente, mais informativa, a análise dos dados.

Para começar, estudou-se o valor da indemnização consoante o sexo e a idade do tomador do seguro, como se pode ver na figura 2.11.



(a) Montante médio das indemnizações consoante o sexo do segurado



(b) Montante médio das indemnizações consoante a idade do segurado

Figura 2.11: Gráficos de extremos e quartis do logaritmo do montante médio das indemnizações consoante as categorias das variáveis binárias *Male* e *Young*.

Ao analisar a figura 2.11, é possível concluir que o montante médio das indemnizações se porta de forma semelhante consoante o sexo e a idade.

A mediana do valor da indemnização, tal como os valores mínimo e máximo, caso o tomador de seguro seja homem, são muito semelhantes à de se este for mulher. A dispersão dos valores também é idêntica em ambos os sexos, embora no caso de o tomador ser homem estes apresentem uma ligeira menor dispersão. Relativamente aos outliers, é possível observar que surgem em maior número caso o segurado seja homem, o que significa que há mais valores discrepantes de indemnizações quando o segurado é homem do que quando é mulher, embora também existam alguns neste último caso.

Tal como acontece para o sexo do tomador do seguro, a mediana e os valores mínimo e máximo, se este for menor de 26 anos, são semelhantes para o caso contrário, bem como a dispersão dos dados. Em relação aos outliers, o montante médio da indemnização para segurados com idade igual ou superior a 26 anos, apresenta mais observações fora do comum do que aqueles com idade inferior a 26, embora neste último caso também surja uma quantidade significativa destes valores.

Após esta análise encontra-se na figura 2.12, o montante médio das indemnizações por distância segurada, em milhares.

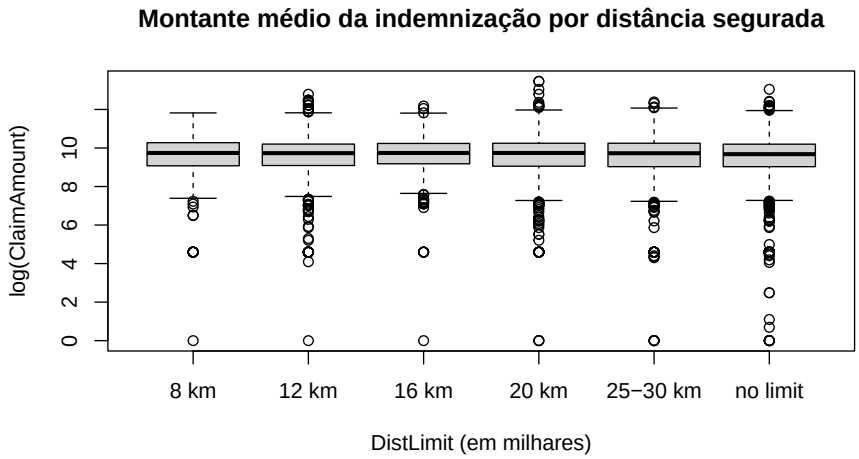


Figura 2.12: Gráfico de extremos e quartis do logaritmo do montante médio das indemnizações consoante a distância segurada

Ao observar a figura 2.12, verifica-se que, analogamente ao que acontece com o sexo e a idade, os valores da mediana, mínimo e máximo do montante médio das indemnizações por distância segurada e até a própria dispersão dos dados é bastante semelhante em todas as categorias. No que diz respeito aos outliers, é possível observá-los em qualquer que seja o limite de distância segurada, sendo que o valor da indemnização quando a distância segurada é 12 000 km ou 20 000km apresenta mais valores discrepantes mas mais concentrados, enquanto que quando não há limite de distância segurada os valores atípicos são muito mais dispersos.

Continuando o estudo do montante médio das indemnizações consoante as diferentes categorias, analisou-se o comportamento desta variável consoante a densidade da região geográfica de onde o tomador do seguro pertence, cujos resultados podem ser observados na figura 2.13.

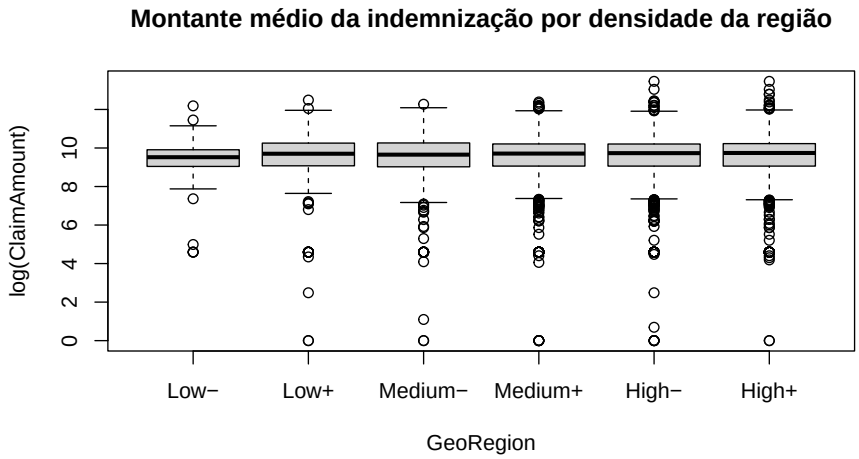


Figura 2.13: Gráfico de extremos e quartis do logaritmo do montante médio das indemnizações consoante a densidade da região geográfica do segurado

Na figura 2.13 verifica-se, tal como nos casos anteriores, uma grande semelhança no que diz respeito à dispersão dos dados. Quando a densidade da região é *Low-*, nota-se que tanto a mediana, como os valores mínimo e máximo do montante, são ligeiramente inferiores quando comparados aos respetivos valores das restantes categorias, assim como a dispersão é ligeiramente menor do que no resto dos casos. Existem valores discrepantes no valor das indemnizações mediante qualquer que seja a densidade da região geográfica do segurado, sendo que quando a densidade é *Low-* verifica-se um menor número de observações fora do comum.

Para terminar o estudo de possíveis relações lineares entre as variáveis presentes, avaliou-se a associação entre as duas variáveis binárias, *Male* e *Young*, através do teste do Qui-Quadrado de independência. Assim, começou-se por elaborar a tabela de contingência, apresentada na tabela 2.5.

Tabela 2.5: Tabela de contingência das variáveis sexo e idade do segurado

$Male \backslash Young$	0	1
0	212	74
1	6054	2104

De seguida, recorreu-se ao teste não paramétrico mencionado, com o intuito de analisar as seguintes hipóteses nula e alternativa, respetivamente:

H_0 : Não há associação entre o sexo e a idade dos segurados, ou seja, são independentes.

vs

H_1 : Há associação entre o sexo e a idade dos segurados, ou seja, são dependentes.

Após a definição das hipóteses, estabeleceu-se o nível de significância usual de 0.05 e realizou-se o teste do Qui-Quadrado através da função *chisq.test()* do R, obtendo os resultados que podem ser consultados na tabela 2.6.

Tabela 2.6: Tabela de contingência das variáveis sexo e idade do segurado

$X-squared$	df	$valor-p$
$1.5671e^{-29}$	1	1

Analisando a tabela 2.6, verifica-se que o valor da estatística de teste está próximo de zero e o número de graus de liberdade é igual a 1, bem como o valor do *valor-p* obtido. Assim, uma vez que o *valor-p* calculado com este teste foi superior ao nível de significância estipulado de 0.05, não se rejeita a hipótese nula. Conclui-se assim a independência das variáveis binárias em estudo *Male* e *Young*.

Por fim, estão assim reunidas as condições para se afirmar que as variáveis em estudo são independentes entre si.

2.3.1 Ajuste da distribuição do número de sinistros

Como referido no capítulo 2.2, a variável em estudo *NbClaim*, diz respeito ao número de sinistros por apólice, sendo por isso uma variável discreta, que neste caso só toma os valores 0, 1, 2 ou 3. Assim, com o intuito de compreender um pouco melhor o comportamento da distribuição desta variável, procedeu-se ao seu ajuste às distribuições discretas Poisson, Geométrica, Binomial e Binomial Negativa. Para este estudo, foi necessário recorrer a três funções do *R*, *fitdistr()* do pacote *MASS 7.3-54* para obter as estimativas dos parâmetros das respetivas distribuições por máxima verosimilhança, *goodfit()* do pacote *vcd 1.4-9* e *chisq.test()* para realizar o teste do Qui-Quadrado com o intuito de avaliar a qualidade do ajuste à distribuição em estudo.

2.3.1.1 Distribuição Poisson

Para se proceder ao ajuste da distribuição do número de sinistros a uma distribuição Poisson, começou-se por determinar a estimativa do parâmetro λ, cujo valor obtido é apresentado na tabela 2.7.

Tabela 2.7: Estimativa do parâmetro λ pelo método de máxima verosimilhança

Parâmetro	Estimativa
λ	0.0477

Após isto, com o intuito de perceber se a distribuição de Poisson era ou não um bom ajuste para a variável em estudo, realizou-se um teste do Qui-Quadrado de forma a testar as seguintes hipóteses:

$$H_0^1 : NbClaim \sim Poisson(\lambda) \quad vs \quad H_1^1 : NbClaim \nsim Poisson(\lambda).$$

Uma vez que o *valor-p* obtido neste teste foi aproximadamente 0, aos níveis de significância usuais, rejeitou-se a hipótese nula que afirma que os dados seguem uma distribuição de Poisson.

2.3.1.2 Distribuição Geométrica

Relativamente ao ajuste da distribuição Geométrica, procedeu-se de forma semelhante às anteriores, porém, é de salientar que a função *goodfit()* não contempla esta distribuição. Contudo, uma vez que a distribuição Geométrica é um caso particular da distribuição Binomial Negativa, que por sua vez tem como função de probabilidade

$$P(X = x) = \binom{r+x-1}{r-1} p^r (1-p)^x, \qquad r \in \mathbb{N}, \quad p \in]0,1[, \quad x \in \mathbb{N}_0.$$

Considerando o parâmetro r igual a 1 obtém-se a distribuição pretendida, contornando assim o obstáculo encontrado. Começou-se por determinar a estimativa do parâmetro p que se encontra na tabela 2.8.

Tabela 2.8: Estimativa do parâmetro p pelo método de máxima verosimilhança

Parâmetro	Estimativa
p	0.9545

Para concluir a qualidade do ajuste, foi realizado, tal como para os casos anteriores, um teste do Qui-Quadrado para testar as seguintes hipóteses:

$$H_0^2 : NbClaim \sim Geometrica(p) \quad vs \quad H_1^2 : NbClaim \nsim Geometrica(p).$$

O valor-p obtido foi 0.0022, aproximadamente, o que aos níveis de significância usuais leva à rejeição da hipótese nula. Assim, conclui-se que a distribuição Geométrica também não é um bom ajuste para a distribuição do número de sinistros em estudo.

2.3.1.3 Distribuição Binomial

De forma semelhante ao ajuste às distribuições anteriores, para a distribuição Binomial estimaram-se, pela máxima verosimilhança, os parâmetros $n \in \mathbb{N}$, $n > 0$, e $p \in \mathbb{R}$, $0 < p < 1$. Contudo, uma vez que a função *fitdistr()* utilizada para a estimação dos parâmetros das distribuições de Poisson e Geométrica, não suporta a distribuição Binomial, encontra-se no anexo I, como foi obtida a estimativa para o parâmetro p desta distribuição. Como o valor mais elevado que a variável $NbClaim$ pode tomar é 3, este foi o valor assumido para o parâmetro n. Assim, na tabela 2.9, apresentam-se as estimativas dos parâmetros da distribuição Binomial.

Tabela 2.9: Estimativa dos parâmetros n e p

Parâmetro	Estimativa
n	3
p	0.0159

De forma a concluir o ajuste à distribuição Binomial, realizou-se o teste do Qui-Quadrado para aferir qual das seguintes hipóteses seria a melhor opção:

$$H_0^3 : NbClaim \sim Binomial \quad vs \quad H_1^3 : NbClaim \nsim Binomial.$$

Tal como no ajuste à distribuição de Poisson, o *valor-p* obtido neste teste foi aproximadamente 0, o que leva à rejeição da hipótese nula aos níveis de significância usuais, sendo assim excluída a hipótese da distribuição Binomial ser um ajuste adequado à distribuição em estudo.

2.3.1.4 Distribuição Binomial Negativa

Por fim procedeu-se ao ajuste da distribuição em estudo a uma distribuição Binomial Negativa, $BN(n,p)$, com $n \in \mathbb{N}$ e $p \in]0,1[$. Contudo, esta Binomial Negativa resulta de considerar a variável aleatória relativa ao número de sinistros como uma Poisson, $NbClaim \sim Poisson(\Lambda)$ em que, por sua vez, o parâmetro desta, Λ, segue uma distribuição $Gama\left(n, \frac{1-p}{p}\right)$, estando assim perante uma distribuição de Poisson mista.

É de salientar que uma parametrização comum é dada pelo valor médio μ e n, sendo que $p = \frac{n}{n+\mu}$. Assim, de forma semelhante aos ajustes anteriores, começou por se determinar as estimativas dos parâmetros da Binomial Negativa n e p, recorrendo à parametrização enunciada, cujos resultados obtidos podem ser consultados na tabela 2.10.

Tabela 2.10: Estimativa dos parâmetros n, μ e p

Parâmetro	Estimativa
n	1.5613
μ	0.0477
p	0.9704

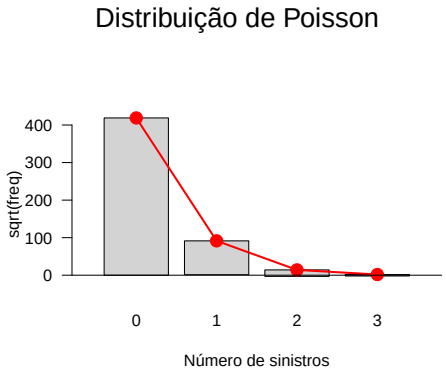
Para concluir o ajuste à distribuição Binomial Negativa, procedeu-se à realização do teste do Qui-Quadrado cujas hipóteses de teste são:

$$H_0^4 : NbClaim \sim Binomial\ Negativa \quad vs \quad H_1^4 : NbClaim \nsim Binomial\ Negativa.$$

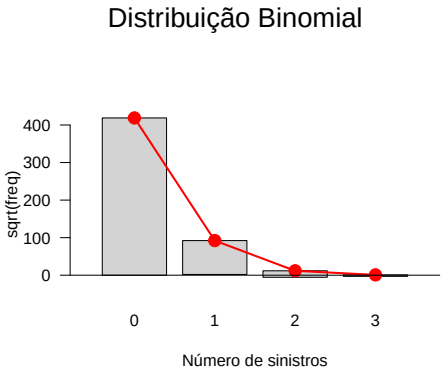
O *valor-p* obtido neste teste foi 0.6545, pelo que, aos níveis de significância usuais, não leva à rejeição da hipótese nula e, por isso, a distribuição Binomial Negativa pode ser considerada um bom ajuste à distribuição do número de sinistros.

2.3.1.5 Conclusão do ajuste

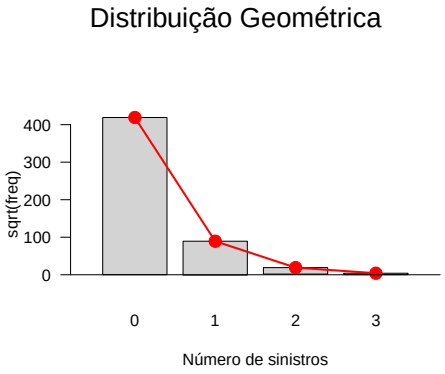
De forma a completar o estudo da qualidade do ajustamento da distribuição do número de sinistros, recorreu-se a uma ferramenta gráfica que pode ser consultada na figura 2.14. Estes gráficos comparam as frequências observadas com as frequências ajustadas de um modelo de probabilidade, em que as primeiras são apresentadas como barras e as segundas são representadas por uma linha. É de salientar, que por defeito, para a realização destes gráficos, são utilizadas as raízes quadradas das frequências com o objetivo de tornar mais visíveis as frequências mais pequenas.



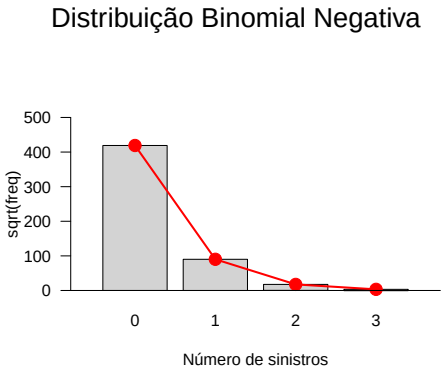
(a) Ajuste da distribuição Poisson



(b) Ajuste da distribuição Binomial



(c) Ajuste da distribuição Geométrica



(d) Ajuste da distribuição Binomial Negativa

Figura 2.14: Gráficos dos ajustes realizados à distribuição do número de sinistros

Perante os resultados obtidos, verifica-se que a distribuição Binomial Negativa apresenta um bom ajuste à distribuição do número de sinistros por apólice.

2.3.2 Ajuste da distribuição do montante médio das indemnizações

Relativamente à variável contínua que diz respeito ao montante médio das indemnizações, *ClaimAmount*, tentou ajustar-se as distribuições Gama e LogNormal. Note-se que para este ajuste não foram contemplados os zeros da amostra [7].

2.3.2.1 Distribuição Gama

Começou-se por recorrer à função *gamma_fit()* do pacote *goft 1.3.6* do R, que ajusta uma distribuição Gama à amostra aleatória de números reais positivos que constituem a variável em estudo, utilizando os estimadores de Villasenor e Gonzalez-Estrada(2015) [8]. Os resultados obtidos podem ser consultados na tabela seguinte.

Tabela 2.11: Estimativa dos parâmetros

Parâmetro	Estimativa
α	1.0566
λ	22325.3341

De seguida, para avaliar a qualidade do ajuste é feito um teste recorrendo à função *gamma_test()*, cujas hipóteses em análise são

$$H_0 : ClaimAmount \sim Gama(\alpha, \lambda) \quad vs \quad H_1 : ClaimAmount \nsim Gama(\alpha, \lambda).$$

Uma vez que o *valor-p* obtido neste teste foi, aproximadamente 0, aos níveis de significância usuais, rejeitou-se a hipótese nula, o que significa que a distribuição Gama não é um bom ajuste à distribuição do montante médio das indemnizações.

2.3.2.2 Distribuição LogNormal

Uma vez que o tamanho da amostra, cujo montante médio das indemnizações é superior a zero, é de 8444 observações, optou-se por recorrer ao teste de Kolmogorov-Smirnoff com correção de Lilliefors que testa a normalidade dos dados sem ser necessário especificar os parâmetros, ou seja, considera as seguintes hipóteses

$$H_0 : X \sim N(\mu, \sigma) \quad vs \quad H_1 : X \nsim N(\mu, \sigma).$$

Contudo, como o que pretendemos analisar é o ajuste a uma distribuição LogNormal e não a uma Normal, primeiro temos de normalizar os dados, ou seja,

$$\text{Se } X \sim LogNormal(\mu, \sigma^2) \text{ então } log(X) \sim N(a(\mu, \sigma^2), b(\mu, \sigma^2)),$$

com $a(\mu, \sigma^2) = ln(\mu) - \frac{1}{2} ln\left(1 + \frac{\sigma^2}{(\mu^2)}\right)$ e $b(\mu, \sigma^2) = ln\left(\frac{\sigma^2}{\mu^2} + 1\right)$.

Dado isto, recorrendo à função *lillie.test()* do pacote *nortest 1.0-4* do R, aplicou-se o teste de Kolmogorov-Smirnoff com correção de Lilliefors para testar a qualidade do ajuste do montante médio das indemnizações à distribuição Lognormal, ou seja, para testar o logaritmo do montante médio das indemnizações à distribuição Normal, sendo as hipóteses de teste as seguintes:

$$H_0 : log(ClaimAmount) \sim N(a(\mu, \sigma^2), b(\mu, \sigma^2)) \quad vs \quad H_1 : log(ClaimAmount) \nsim N(a(\mu, \sigma^2), b(\mu, \sigma^2)).$$

O valor-p obtido com este teste foi, aproximadamente, 0, o que leva à rejeição da hipótese nula aos níveis de significância usuais. Desta forma, conclui-se que tal como a distribuição Gama, a distribuição LogNormal não apresenta um bom ajuste para a distribuição do montante médio das indemnizações.

Uma vez que não foi encontrado um bom modelo para o valor das indemnizações em estudo, o foco, em diante, vai ser na modelação da cauda da distribuição.

Capítulo 3

Teoria de valores extremos

3.1 Introdução

Uma vez que os acontecimentos extremos estão associados a valores nas caudas das distribuições, as abordagens estatísticas comuns não são adequadas para a modelação destes eventos, pois a grande parte destas considera que a maioria dos pontos estão concentrados no centro da distribuição e a estimativa é difícil uma vez que as observações nas caudas são escassas. Assim, surge a necessidade de recorrer a modelos para modelar apenas a cauda.

Neste capítulo são apresentados os resultados principais da Teoria de Valores Extremos, tal como é o caso do Teorema de Gnedenko e do Teorema de Pickands e Balkema-de Haan, e são apresentadas metodologias paramétricas e semi paramétricas da EVT para a modelação da distribuição da cauda. É importante referir que os métodos abordados pressupõem uma sequência de variáveis aleatórias, independentes e identicamente distribuídas.

Uma das bases deste capítulo foi o livro *An Introduction to Statistical Modeling of Extreme Values* de Stuart Coles [9], cuja explicação sucinta de alguns resultados preliminares da EVT tal como a apresentação das metodologias paramétricas em estudo, se verificaram mais explícitas com as aplicações enunciadas e a respetiva interpretação de resultados. Por outro lado, relativamente à abordagem semi paramétrica, foram analisados alguns artigos, tais como [10] e [11], de forma a compreender a metodologia bem como a sua aplicação e a respetiva análise de resultados.

3.2 Alguns resultados teóricos

Existem várias metodologias associadas à Teoria de Valores Extremos para estimar as distribuições da cauda, sendo as mais comuns, a metodologia semi paramétrica e, no caso paramétrico, método dos blocos e o método dos excessos acima de um limiar. Enquanto que na abordagem semi paramétrica não se adota um modelo limite para a função de distribuição, F, nos casos paramétricos, a EVT modela os extremos utilizando a distribuição generalizada de valores extremos (subsecção 3.2.1) e a distribuição generalizada de Pareto (subsecção 3.2.2), respetivamente.

A técnica dos blocos máximos consiste em modelar e caracterizar o comportamento do máximo (ou mínimo, calculando o simétrico) de uma série, assumindo que as observações são independentes entre si e identicamente distribuídas. Começa-se por dividir os dados em m blocos, em que cada bloco tem n observações e seleciona-se o máximo de cada bloco, obtendo assim uma sequência de blocos máximos. É de salientar a necessidade da existência de um equilíbrio entre o número de blocos e o tamanho de cada bloco, pois ao ter blocos com um tamanho reduzido pode levar a resultados enviesados, enquanto que, perante um grande tamanho de cada bloco, há menos blocos e consequentemente, uma maior variabilidade.

Esta última abordagem referida é a mais adequada quando a única informação conhecida são os máximos de cada bloco. Contudo, quando há mais dados disponíveis, recorrer a uma técnica que apenas tem em consideração os máximos, leva a um desperdício de informação. Desta forma, surge a importância de recorrer a outras metodologias, como por exemplo, à abordagem dos excessos acima de um limiar que tem por base, tal como o nome indica, as excedências acima de um determinado limiar. Esta técnica também supõe um conjunto de variáveis independentes e identicamente distribuídas e como foi a técnica utilizada para a realização da parte prática deste projeto, é apresentada detalhadamente na secção 3.3.

3.2.1 Distribuição Generalizada de Valores Extremos

Seja $X_1, ..., X_n$ uma sequência de variáveis aleatórias reais (v.a.r.) independentes e identicamente distribuídas (i.i.d.), com função de distribuição comum F, em que $F(x) = P(X_i < x)$. Considere a variável aleatória real $M_n = max(X_1, ..., X_n)$, o máximo de $X_1, ..., X_n$. Consequentemente, a sua função de distribuição é dada por

$$P(M_n \leq x) = P(X_1 \leq x, ..., X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = [F(x)]^n, \tag{3.1}$$

o que só tem interesse prático quando F é conhecida. Quando esta não é conhecida, embora uma possibilidade para resolver o problema, passe por estimar F, recorrendo a técnicas estatísticas padrão, a partir de dados observados e substituir em (3.1), não é a mais adequada pois discrepâncias ainda que mínimas em F, podem influenciar significativamente F^n. Outra alternativa consiste em assumir que F é desconhecida e tentar

descobrir famílias de modelos que se ajustem aproximadamente a F^n, mas que possam ser estimados apenas tendo em conta os eventos extremos. Ainda assim, ao observar-se o comportamento de F^n com $n \to \infty$, conclui-se que isto não é suficiente pois a distribuição de M_n é degenerada com massa de probabilidade concentrada no limite superior do suporte de F. Desta forma, é necessário recorrer a uma normalização linear da variável M_n, ou seja,

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{3.2}$$

para sucessões de constantes $a_n > 0$ e b_n. Desta forma, pretendem-se encontrar distribuições limites para M_n^*, escolhendo apropriadamente a_n e b_n. No Teorema seguinte, também conhecido como *Teorema dos Modelos Extremos*, encontram-se as únicas distribuições de limites possíveis para a função do máximo normalizado, M_n^*.

Teorema 3.2.1 (Teorema de Gnedenko) *Se existem sucessões reais $a_n > 0$ e b_n tal que*

$$P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = F(a_n x + b_n)^n \to G(x), \qquad com\ n \to \infty, \tag{3.3}$$

onde G é uma função de distribuição não degenerada, então G pertence à família dos Valores Extremos Generalizados (GEV, Generalized Extreme Value), cuja função de distribuição é dada por

$$G(x) = exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{3.4}$$

definido no conjunto $\{x : 1 + \xi\,(x-\mu)/\sigma > 0\}$, em que $\mu \in \mathbb{R}$, $\sigma > 0$ e $\xi \in \mathbb{R}$ são os parâmetros de localização, escala e forma, respetivamente.

Note-se que, caso (3.3) se verifique, diz-se que a distribuição de X_i, $i = 1, ..., n$, F, está no domínio de atração para máximos de G. O parâmetro de forma, ξ, é também conhecido por índice de valor extremo ou índice de cauda, é fundamental na análise de valores extremos e é responsável pelo decaimento da cauda da distribuição. Existem três casos particulares da distribuição generalizada de valores extremos, consoante os valores tomados por este parâmetro:

I. quando $\xi \to 0$, tem-se a distribuição de Gumbel

$$G(x) = \left\{exp\left\{-exp\left[-\left(\tfrac{x-b}{a}\right)\right]\right\}, \quad x \in \mathbb{R}\right. \tag{3.5}$$

II. quando $\xi > 0$, tem-se a distribuição de Fréchet

$$G(x) = \begin{cases} 0, & x \leq b, \\ exp\left\{-\left(\tfrac{x-b}{a}\right)^{-\alpha}\right\}, & x > b; \end{cases} \tag{3.6}$$

III. quando $\xi < 0$, tem-se a distribuição de Weibull

$$G(x) = \begin{cases} exp\left\{-\left[-\left(\tfrac{x-b}{a}\right)^{-\alpha}\right]\right\}, & x < b, \\ 1, & x \geq b; \end{cases} \tag{3.7}$$

para os parâmetros $a > 0$, $b \in \mathbb{R}$ e, no caso das distribuições II. e III., $\alpha > 0$.

Os comportamentos distintos que estas distribuições apresentam entre si correspondem às diferentes formas do comportamento da cauda da função de distribuição F de X_i, como se pode verificar em [12]. No caso $\xi > 0$, fala-se em domínio de atração de Fréchet, em que a cauda é pesada e a cauda $1 - F(x)$ decai para o valor zero polinomialmente, sendo a extremidade direita infinita tal como acontece, por exemplo, na distribuição de Pareto. No domínio de atração de Gumbel, $\xi \to 0$ e a cauda da distribuição decresce exponencialmente para a distribuição de Gumbel, tal como acontece, por exemplo, nas distribuições exponencial, normal e gama, não sendo necessário que a extremidade seja infinita. Quando $\xi < 0$, as distribuições têm a cauda curta e apresentam um limite superior de suporte, pertencendo assim ao domínio de atração de Weibull, como é o caso da distribuição Beta. Embora todas estas distribuições tenham a cauda pesada, a distribuição de Fréchet lidera neste campo.

Em suma, o teorema 3.2.1 é semelhante ao teorema limite central para médias, sendo que este último se aplica à média de uma amostra de qualquer distribuição com variância finita, enquanto que o teorema de Gnedenko apenas afirma que se uma distribuição de um máximo normalizado converge, então o limite tem de pertencer a um grupo particular de distribuições.

3.2.2 Distribuição Generalizada de Pareto

Seja $X_1, ..., X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cuja função de distribuição marginal é F, e tal que $M_n = max\{X_1, ..., X_n\}$. Consideram-se eventos extremos os X_i cujos valores excedem um determinado limiar suficientemente alto u. Seja X um termo arbitrário de X_i, o comportamento estocástico de eventos extremos é dado por

$$P(X > u + y | X > u) = \frac{1 - F(u + y)}{1 - F(u)}, y > 0. \tag{3.8}$$

Considere-se o seguinte teorema que é frequentemente considerado como o segundo resultado na teoria de valores extremos.

Teorema 3.2.2 (Teorema de Pickands e Balkema-de Haan) *Assuma-se que*

$$P(M_n \leq x) = P(max(X_1, ..., X_n) \leq x) \approx G(x),$$

em que

$$G(x) = exp\left\{-\left[1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{3.9}$$

para determinados μ, σ e ξ. Assim, para u suficientemente grande, a distribuição de $X - u | X > u$ é aproximadamente uma distribuição Generalizada de Pareto (GP), cuja função de distribuição é dada por

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \tag{3.10}$$

definido em $\left\{y : y > 0, \left(1 + \frac{\xi y}{\tilde{\sigma}}\right) > 0\right\}$. *As distribuições GEV e GP partilham o mesmo parâmetro de forma e o parâmetro de escala satisfaz a relação*

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{3.11}$$

O Teorema3.2.2 mostra que se os máximos seguem uma distribuição generalizada de valores extremos, então as excedências acima de um dado limiar ajustam-se aproximadamente a uma distribuição generalizada de Pareto. Esta distribuição depende também de três parâmetros, o μ parâmetro de localização, σ parâmetro de escala e o ξ parâmetro de forma.

É de salientar que a distribuição GP satisfaz a propriedade de estabilidade de limiar que afirma que se esta distribuição é válida para os excessos de um dado limiar u_0, então deve ser igualmente válida para qualquer limiar u tal que $u > u_0$, com igual parâmetro de forma e a modificação do parâmetro de escala é dada por

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0). \tag{3.12}$$

.

De forma análoga à distribuição GEV, o parâmetro de forma ξ é dominante na determinação do comportamento qualitativo da distribuição GP. Quando o índice de cauda é negativo, $\xi < 0$, a cauda é finita, ou seja, há um limite superior de suporte, a distribuição de excessos tem um limiar superior de $u - \frac{\tilde{\sigma}}{\xi}$. Por outro lado, se $\xi > 0$, a cauda decai polinomialmente e a distribuição não tem limiar superior, enquanto que, no caso em que $\xi \to 0$, a cauda decresce exponencialmente, sendo assim a distribuição dada por

$$H(y) = 1 - exp\left(-\frac{y}{\tilde{\sigma}}\right), y > 0. \tag{3.13}$$

Naturalmente, o ajustamento desta distribuição depende da escolha do limiar.

3.3 Abordagem paramétrica: Excessos acima de um limiar

A metodologia dos excessos acima de um limiar mais conhecida por *POT (Peaks Over Threshold)* centra-se nas observações que excedem um determinado nível que se supõe elevado denominado limiar. O objetivo deste método passa pela criação de um modelo para estes valores extremos, modelando a cauda da distribuição de todos os valores que excedem o dito limiar.

3.3.1 Escolha do limiar

É necessário escolher um limiar u adequado, isto é, o menor u tal que a cauda da distribuição acima deste se comporte como uma generalizada de Pareto. É importante salientar a dificuldade desta escolha, que é semelhante à escolha do tamanho dos blocos na metodologia dos blocos máximos, pois tem de se encontrar um equilíbrio entre o viés e a

variância. Isto é, se o limiar escolhido for demasiado alto, consequentemente o número de observações para modelar corretamente a cauda será baixo, levando assim a uma variância elevada pois apenas se consideram valores muito extremos. Por outro lado, ao selecionar um limiar demasiado baixo, incluem-se demasiados valores o que se traduz num elevado enviesamento.

Existem várias técnicas para determinar o limiar, sendo as mais comuns a Regra de Ouro e a utilização de métodos gráficos.

Regra de Ouro

Tal como *Max Rydman* mencionou em [2], a Regra de Ouro é uma forma de estabelecer o limiar, escolhendo as k observações mais elevadas e modelando-as. Geralmente, neste caso, são utilizados o percentil de ordem 90, assim como $k = \sqrt{n}$ e $k = n^{\frac{2}{3}}$.

Embora estas técnicas sejam práticas e permitam estabelecer rapidamente um limiar, este método não é considerado uma forma viável de chegar a um bom limiar devido à inevitável diferença de comportamento entre diferentes tipos de dados.

Métodos gráficos

Outra forma de determinar um limiar adequado consiste em recorrer a ferramentas gráficas, como o Gráfico de Vida Residual Média e o Gráfico da Estabilidade das Estimativas dos Parâmetros, e analisar os resultados obtidos. A primeira ferramenta consiste numa técnica exploratória realizada antes da estimativa do modelo, enquanto que a segunda é referente a uma avaliação da estabilidade das estimativas dos parâmetros apresentando modelos com distintos limiares.

O Gráfico de Vida Residual Média também conhecido por *MRL-plot (Mean Residual Life Plot)*, é uma das técnicas mais utilizadas para a escolha do limiar. Este método baseia-se no valor esperado da distribuição generalizada de Pareto para os excessos, ou seja, considera-se um limiar u_0 e X um termo arbitrário dos excessos gerados por u_0 que seguem a distribuição (3.10). Assim, tem-se

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1 - \xi}, \tag{3.14}$$

em que σ_{u_0} e ξ são os parâmetros de escala e forma, respetivamente, e $\xi < 1$ de forma a garantir a existência da média, pois caso contrário esta é infinita. Uma vez que a distribuição generalizada de Pareto é válida para as excedências do limiar u_0, o mesmo deve acontecer para $\forall u$: $u > u_0$. Desta forma, e pela (3.12), o valor esperado dos excessos para qualquer limiar $u > u_0$, é dado por

$$E(X - u | X > u) = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u_0} + \xi\,(u - u_0)}{1 - \xi} \tag{3.15}$$

Assim, conclui-se que para $u > u_0$, $E(X - u|X > u)$ é uma função linear de u, ou seja, perante um modelo generalizado de Pareto apropriado, as estimativas fornecidas pela média dos excessos do limiar u mudam linearmente com u.

A interpretação de um gráfico de vida residual média nem sempre é simples. Geralmente, quando se verifica um comportamento linear nesta ferramenta gráfica, supõe-se que um limiar adequado pode ser estimado. Isto é, acima de um limiar u, no qual a distribuição generalizada de Pareto fornece uma aproximação válida à distribuição das excedências, o gráfico deve ser aproximadamente linear em u sendo assim este considerado um limiar adequado. Contudo, quando o limiar se torna demasiado grande, a variação dos poucos extremos que restam pode levar à perda do comportamento linear. Por outro lado, também pode acontecer não ser possível recolher nenhuma ou quase nenhuma informação utilizando este método, devido à constante presença de linearidade dos dados ou à ausência total da mesma. Além disto, nos gráficos de vida residual média é também possível observar os intervalos de confiança que têm por base a normalidade aproximada das médias da amostra.

O gráfico de estabilidade das estimativas dos parâmetros é conhecido como uma técnica complementar que auxilia o encontro de um limiar adequado através do ajuste da distribuição generalizada de Pareto a um intervalo de valores de u, e uma posterior análise das estimativas dos parâmetros de forma e escala.

Como foi referido na secção 3.2.2, se a distribuição generalizada de Pareto é válida para as excedências de um limiar u_0, a mesma também é válida para os excessos de um limiar superior u. No que diz respeito a estas duas distribuições, os parâmetros de forma são idênticos, enquanto que o de escala é dado por

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0), \tag{3.16}$$

sendo assim função linear de u, a não ser que $\xi = 0$. Para resolver esta dificuldade, reparametriza-se o parâmetro de escala da distribuição generalizada de Pareto da seguinte forma,

$$\sigma^* = \sigma_u - \xi u \tag{3.17}$$

tornando-se assim constante relativamente a u por (3.16).

Estão assim reunidas as condições para a elaboração dos gráficos de estabilidade para cada um dos parâmetros, em que são traçados estes estimadores no intervalo escolhido de u com o intervalo de confiança de 95% para os respetivos estimadores. Ao analisar estes gráficos, seleciona-se como limiar adequado o menor valor de u em que as estimativas permanecem aproximadamente constantes, pois devido à variabilidade da amostra estas não são exatamente constantes mas devem ser estáveis após a permissão para os seus erros de amostragem. Note-se que os intervalos de confiança de $\hat{\xi}$ são obtidos a partir da matriz de covariância V, enquanto que os de $\hat{\sigma}^*$ requerem o método delta, usando

$$Var(\hat{\sigma}^*) \approx \nabla\sigma^{*T} V \nabla\sigma^*, \tag{3.18}$$

em que

$$\nabla\sigma^{*T} = \left[\frac{\partial\sigma^*}{\partial\sigma_u}, \frac{\partial\sigma^*}{\partial\xi}\right] = [1, -u] \tag{3.19}$$

É importante referir que tal como acontece no gráfico de vida residual média, há casos em que não se consegue retirar informação deste gráfico da estabilidade das estimativas dos parâmetros devido ao comportamento dos dados.

Estes métodos gráficos são mais demorados do que a utilização de uma Regra de Ouro, contudo, estes permitem a identificação de um maior número de possíveis candidatos a limiar, alargando assim o leque de hipóteses para a escolha deste.

3.3.2 Estimação dos parâmetros

Com o limiar escolhido, segue-se agora a estimação dos parâmetros de forma e escala da distribuição generalizada de Pareto. Existem diversos métodos de estimação, sendo o mais comum o método da máxima verosimilhança, que foi o selecionado para abordar nesta secção. Este consiste em encontrar um estimador adequado para os parâmetros através da função de verosimilhança.

Método da máxima verosimilhança

O método da máxima verosimilhança é muito utilizado na estimação paramétrica dos parâmetros da distribuição generalizada de Pareto devido à sua simplicidade e às suas propriedades. Desta forma, derivando (3.10), obtém-se a função de densidade de probabilidade desta distribuição e a partir desta determina-se a função de log-verosimilhança para uma amostra de excessos de um limiar u. Sejam estas excedências designadas por $y_1, y_2, ..., y_k$, a função de log-verosimilhança, quando $\xi \neq 0$ é dada por

$$l(\sigma, \xi) = -klog\sigma - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{k} log\left(1 + \frac{\xi y_i}{\sigma}\right), \tag{3.20}$$

sendo que $\left(1 + \frac{\xi y_i}{\sigma}\right) > 0$ para $i = 1, ..., k$, caso contrário $l(\sigma, \xi) = -\infty$. Por outro lado, quando $\xi = 0$, pela equação (3.13), tem-se que a função de verosimilhança é dada por

$$l(\sigma) = -klog\sigma - \sigma^{-1}\sum_{i=1}^{k} y_i. \tag{3.21}$$

É de salientar que devido à impossibilidade de maximizar a log-verosimilhança analiticamente, não existem fórmulas específicas para os estimadores, o que torna inevitável a utilização de métodos numéricos sendo necessário especial atenção para evitar instabilidades numéricas quando $\xi \approx 0$ em (3.20). Uma vez feita a avaliação fora do espaço de parâmetros permitido é preciso verificar também que o algoritmo não falha.

3.3.3 Níveis de retorno

Uma quantidade que normalmente é obtida na Teoria de Valores Extremos são os níveis de retorno. Estes, são os valores que se espera que sejam excedidos, em média, uma vez a cada m observações, sendo a probabilidade deste acontecimento $\frac{1}{m}$. Sendo ζ uma estimativa da probabilidade de exceder o limiar u, no caso em que $\xi \neq 0$, o nível de retorno associado ao período de retorno m, é dado por

$$x_m = u + \frac{\sigma_u}{\xi}((m\zeta_u)^{\xi} - 1), \tag{3.22}$$

enquanto que, quando $\xi = 0$, este é obtido por

$$x_m = u + \sigma_u log(m\zeta_u), \tag{3.23}$$

para mais detalhes consultar a página 81 do livro [9].

Devido à simplicidade de interpretação e ao facto de a escolha de escala comprimir a cauda da distribuição, realçando assim o efeito de extrapolação, os gráficos de níveis de retorno são uma boa ferramenta a utilizar na apresentação e validação do modelo. Estes representam o nível de retorno x_m, contra o logaritmo do período de retorno m. A forma destes gráficos é determinada pelo valor de ξ, isto é, caso $\xi = 0$ o gráfico apresenta linearidade, enquanto que se $\xi > 0$ predomina a concavidade e, por outro lado, caso $\xi < 0$, a convexidade é a característica deste gráfico.

De um modo geral, o mais comum é os níveis de retorno serem dados numa escala anual, em que o nível de retorno do ano N é o nível que se espera exceder uma vez a cada N anos e n_y o número de observações por ano. Desta forma, o período de retorno m é tal que $m = N \times n_y$, em que o nível de retorno do ano N é dado por

$$z_N = u + \frac{\sigma}{\xi}\left[\left(Nn_y\zeta_u\right)^{\xi} - 1\right], \tag{3.24}$$

enquanto que, quando $\xi = 0$, este é obtido por

$$z_N = u + \sigma log(Nn_y\zeta_u). \tag{3.25}$$

3.3.4 Verificação do modelo

Para verificar a qualidade de um ajuste à distribuição generalizada de Pareto, é comum e conveniente recorrer-se a gráficos de probabilidade, de quantis, de níveis de retorno e de densidade. Considere-se u um determinado limiar, $y_{(1)} \leq ... \leq y_{(k)}$ as excedências desse limiar e $\hat{H}$ um modelo estimado.

O gráfico de probabilidade é representado pelos pares

$$\left\{\left(\frac{i}{k+1}, \hat{H}(y_{(i)})\right); i = 1, ..., k\right\},$$

em que, caso $\xi \neq 0$, por (3.10),

$$\hat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\xi}}}, \tag{3.26}$$

e, caso contrário, $\xi = 0$, por (3.13), o gráfico é construído utilizando

$$\hat{H}(y) = 1 - exp\left(-\frac{y}{\hat{\sigma}}\right). \tag{3.27}$$

De forma análoga, o gráfico de quantis é constituído pelos pares

$$\left\{\left(\hat{H}^{-1}\left(\frac{i}{k+1}\right), y_{(i)}\right); i = 1, ..., k\right\},$$

onde, considerando $\xi \neq 0$, por (3.26),

$$\hat{H}^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[y^{-\hat{\xi}} - 1\right]. \tag{3.28}$$

Caso estes dois gráficos sejam constituídos por pontos aproximadamente lineares, significa que a distribuição generalizada de Pareto se ajusta bem à modelação das excedências do limiar u.

Por outro lado, o gráfico de níveis de retorno consiste na localização dos pontos $\{(m, \hat{x}_m)\}$ para grandes valores de m, em que $\hat{x}_m$ é a estimativa do nível de retorno da m-ésima observação e, quando $\xi \neq 0$, por (3.22), esta é dada por

$$\hat{x}_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[(m\hat{\zeta}_u)^{\hat{\xi}} - 1\right], \tag{3.29}$$

caso contrário, por (3.23),

$$\hat{x}_m = u + \hat{\sigma} ln(m\hat{\zeta}_u). \tag{3.30}$$

Como foi referido na secção 3.3.3, estes gráficos traçam a curva de nível de retorno numa escala logarítmica, isto para evidenciar o efeito de extrapolação e acrescentar intervalos de confiança e estimativas empíricas dos níveis de retorno.

Já no gráfico de densidade, é comparada a função densidade da distribuição generalizada de Pareto e o histograma das excedências.

3.4 Abordagem semi paramétrica

O método dos blocos máximos e o método dos excessos de um nível elevado referidos anteriormente, são exemplos de abordagens paramétricas, pois em ambos é conhecida a distribuição a utilizar na modelação da cauda. No primeiro caso recorre-se à distribuição de valores extremos e no segundo à generalizada de Pareto, ou seja, o modelo está bem especificado. Por outro lado, existe também a abordagem semi paramétrica onde não se sabe ao certo qual a distribuição a utilizar e, por isso, o que se tem é uma aproximação para a mesma. Neste caso, é comum aproximar-se a distribuição à distribuição de Pareto, cujas funções de distribuição e de quantis são dadas, respetivamente, por

$$F(x) = 1 - \left(\frac{x}{c}\right)^{-\alpha}, x > c, c > 0, \alpha > 0, \tag{3.31}$$

e

$$Q(p) = F^{\leftarrow}(p) = c(1-p)^{-\frac{1}{\alpha}}, 0 < p < 1, c > 0, \alpha > 0, \tag{3.32}$$

em que c e α são os respetivos parâmetros de escala e forma, enquanto que, o p denota a menor probabilidade da cauda. O parâmetro de forma, α, é também denominado por índice de cauda uma vez que mede o peso da cauda direita da distribuição, sendo que $\xi = \frac{1}{\alpha}$ é o índice de valor extremo. Note-se que se a função de distribuição de X for dada por (3.31), a notação é dada por $X \sim P(c, \alpha)$.

3.4.1 Modelos do tipo Pareto

Tal como foi referido na subsecção 3.4, devem ser consideradas as distribuições do tipo Pareto ou também chamadas distribuições de caudas pesadas, ou seja, modelos em que o máximo de uma amostra independente e identicamente distribuída, após uma normalização linear adequada, tende para a distribuição de Fréchet para algum $\xi > 0$. Desta forma,

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\frac{1}{\xi}}, \quad \forall x > 0. \tag{3.33}$$

Consequentemente, os modelos do tipo Pareto satisfazem

$$1 - F(x) = x^{-\frac{1}{\xi}} L(x), \quad x \to \infty, \tag{3.34}$$

sabendo que $1 - F(x)$ é a função de sobrevivência, em que é possível observar uma parte paramétrica definida pelo $x^{-\alpha}$ e uma parte que não é conhecida representada pela função $L(x)$, sendo que L é uma função de variação lenta, isto é,

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1, \quad \forall x > 0. \tag{3.35}$$

Uma vez que se trata de uma classe que inclui diversos modelos, não é especificada a função de sobrevivência, pelo que para obter estimadores assume-se que a função L converge para uma constante, ou seja, $L(x) = c^{\frac{1}{\xi}}$, para valores elevados de x, com $c > 0$ uma constante. Desta forma, tem-se uma aproximação para a função de sobrevivência reunindo assim as condições necessárias para a determinação do estimador.

3.4.2 Estimação do índice de valor extremo

No contexto semi paramétrico, o facto de não se ter a função de distribuição bem definida faz com que não exista um estimador de máxima verosimilhança exato, e por isso, os que surgem são aproximadamente de máxima verosimilhança e alguns deles até têm uma expressão analítica, como é o caso do estimador de Hill [13].

Na estimação do índice de valor extremo, é derivado o estimador de Hill para ξ, não sendo necessário estimar a constante c.

Seja $X_1, X_2, ..., X_n$ uma amostra independente e identicamente distribuída de tamanho n, em que $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ é a amostra associada de estatísticas ordenadas de forma ascendente. A partir de (3.33), é possível escrever

$$\frac{1-F(tx)}{1-F(t)} = P(X > tx|X > t) = P\left(\frac{X}{t} > x|X > t\right) \approx x^{-\frac{1}{\xi}}, \forall x > 1.$$

Assim, tendo em conta que $Y = \frac{X}{t}|X > t$,

$$P(Y \leq x) \approx 1 - x^{-\frac{1}{\xi}}, \quad (modelo P(1, \frac{1}{\xi})).$$

Considerando $n_t = \#Y_i$, a função de log-verosimilhança é dada por

$$l\left(\xi|Y_1, Y_2, Y_{n_t}\right) = -n_t ln(\xi) - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{n_t} lnY_i.$$

Consequentemente, o estimador de máxima verosimilhança de ξ é

$$\hat{\xi} = \frac{1}{n_t} \sum_{i=1}^{n_t} lnY_i.$$

Se o limiar t for escolhido de uma forma aleatória, $t = X_{(n-k)}$, obtém-se o estimador de Hill [14] que é dado por

$$\hat{\xi}^H = \frac{1}{k} \sum_{i=1}^{k} logX_{(n-i+1)} - logX_{(n-k)}. \tag{3.36}$$

Note-se que na escolha de k reside um problema semelhante ao da determinação do limiar u, na metodologia POT, apresentada na secção 3.3.1.

3.4.3 Estimação de uma probabilidade de excedência

Note-se que

$$\frac{1-F(ty)}{1-F(t)} = \frac{P(X > ty)}{P(X > t)} \sim y^{-\frac{1}{\xi}}, \quad com \quad t \to \infty. \tag{3.37}$$

Posto isto, ao admitir $ty = x$ e $t = X_{(n-k)}$, tem-se que $P(X > t)$ pode ser substituída pelo estimador empírico $\frac{k}{n}$. Assim, a probabilidade de excedência pode ser estimada por

$$\hat{P}(X > x) = \frac{k}{n}\left(\frac{x}{X_{(n-k)}}\right)^{-\frac{1}{\hat{\xi}^H}(k)}. \tag{3.38}$$

É de salientar que o valor escolhido para x, convém ser um valor extremo, próximo ou superior ao máximo amostral.

3.4.4 Estimação de quantis extremos

Por outro lado, considere-se a estimação de quantis de ordem elevada da função de distribuição F, ou seja, valores que têm uma probabilidade muito pequena de ser excedidos. Estes são representados por q_{1-p}, valor que verifica $P(X > q_{1-p}) = p$, com p pequeno, próximo de zero. Desta forma, considerando $t = X_{(n-k)}$, $ty = q_{1-p}$ e $p = P(X > q_{1-p})$ em (3.37), obtém-se o seguinte estimador

$$\hat{q}_{1-p} = X_{(n-k)} \left(\frac{k}{np} \right)^{\hat{\xi}^H(k)}. \tag{3.39}$$

Note-se que um nível de retorno a T-anos é um exemplo de um quantil extremal de probabilidade $1 - \frac{1}{T}$, [15].

Capítulo 4

Análise de resultados

4.1 Introdução

Com o estudo exploratório inicialmente realizado, é possível concluir que existe uma grande dispersão dos dados, sendo impossível o ajuste dos mesmos como um só bloco como já foi referido. Desta forma, surge a necessidade da divisão dos custos em *Custos dos Sinistros Comuns* e *Custos dos Sinistros Elevados*, sendo que os primeiros se referem aos custos com indemnizações que representam a grande maioria das observações mas cujo limite é aplicado num valor razoável que permita o devido ajuste dos dados, enquanto que os restantes dizem respeito aos valores extremos da amostra.

De forma a provar a utilidade das metodologias desenvolvidas na EVT, este capítulo dedica-se à análise dos valores extremos dos montantes das indemnizações em estudo, primeiro recorrendo a uma abordagem paramétrica e, de seguida, segundo uma abordagem semi paramétrica.

4.2 Abordagem paramétrica

A abordagem paramétrica utilizada consiste na aplicação do método POT ou dos excessos acima de um limiar referido na secção 3.3.

4.2.1 Seleção do limiar

De forma a selecionar o melhor limiar u, foram utilizados os diferentes métodos mencionados na secção 3.3.1 e posteriormente comparados.

De acordo com a Regra de Ouro discutida na secção 3.3.1, calcularam-se os diferentes limiares que podem ser consultados na tabela seguinte.

Tabela 4.1: Resultados obtidos pela Regra de Ouro

	Percentil 90	k = $\sqrt{n}$	k = $n^{2/3}/log(log(n))$
Limiar	49 000	129 155	103 500
Número de excedências	844	92	189

Ao analisar a tabela 4.1, verifica-se que relativamente ao número de excedências, as diferenças são notórias. Ao utilizar o quantil de ordem 90, obtém-se um limiar de 49 000 e o número de excedências do mesmo é de 844, enquanto que recorrendo às regras $k = \sqrt{n}$ e $k = \frac{n^{\frac{2}{3}}}{log(log(n))}$ os limiares são 129 155 e 103 500 com 91 e 188 excedências, respetivamente.

Por outro lado, aplicando os métodos gráficos detalhados na subsecção 4.2, é possível observar nas figuras 4.1 e 4.2, os gráficos de vida residual média e o da estabilidade das estimativas dos parâmetros, respetivamente.

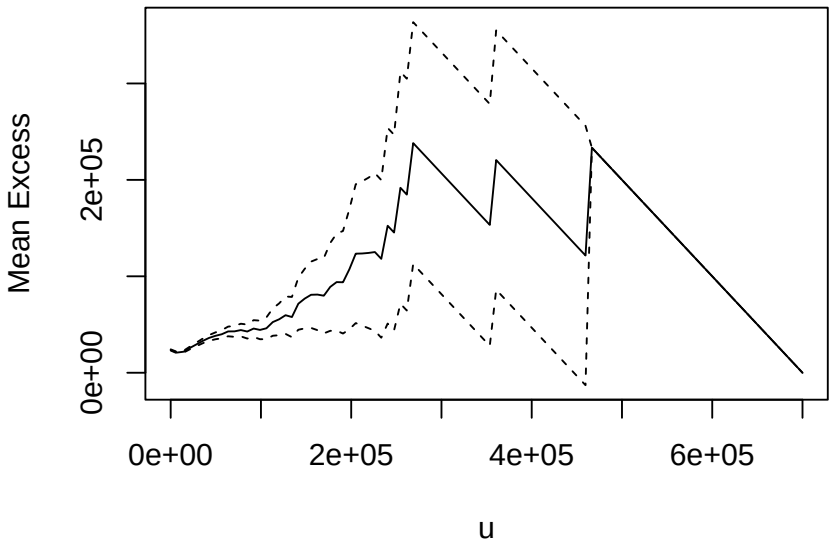


Figura 4.1: Gráfico de vida residual média

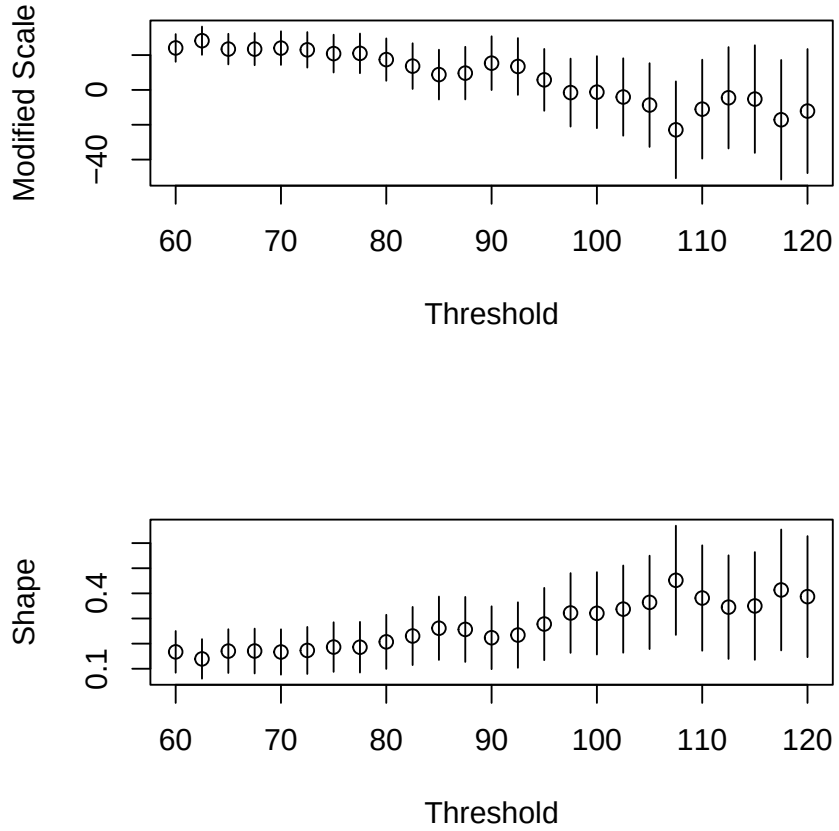


Figura 4.2: Gráfico da estabilidade das estimativas dos parâmetros

As figuras 4.1 e 4.2, sugerem que $u = 110\ 000$ pode ser uma boa escolha para o limiar, uma vez que o primeiro gráfico é, aproximadamente, linear neste valor e, por outro lado, no segundo gráfico, aparenta ser uma área estável, em que se verifica pequena variação das estimativas dos parâmetros.

Na tabela seguinte apresentam-se as informações relativamente a este limiar, tal como foi feito para os limiares obtidos pela Regra de Ouro.

Tabela 4.2: Resultados obtidos pela aproximação gráfica

	Aproximação Gráfica
Limiar	110 000
Número de excedências	158

4.2.2 Estimação dos parâmetros

Ao recorrer à Regra de Ouro e a métodos gráficos foram obtidos quatro potenciais limiares. A escolha de um modelo adequado passa não só pela escolha do limiar, mas também pela modelagem dos valores extremos existentes na amostra de dados. Para isto, é necessário

analisar mais detalhadamente as estimativas dos parâmetros associados aos diferentes limiares.

Na tabela 4.3 podem ser consultadas as estimativas dos parâmetros, assim como as estimativas de erro para os respetivos parâmetros de forma e de escala utilizando o pacote *ismev 1.42* do R. É de salientar que estas estimativas foram obtidas pelo estimador da máxima verosimilhança.

Tabela 4.3: Estimativas dos parâmetros de escala e de forma

	Regras gerais			Aproximação gráfica
	Percentil 90	k = $\sqrt{n}$	k = $n^{2/3}/log(log(n))$	
Limiar	49 000	129 155	103 500	110 000
Estimativa escala	31.308	35.680	32.275	30.917
Erro padrão escala	1.586	6.155	3.586	4.013
Estimativa forma	0.177	0.420	0.317	0.381
Erro padrão forma	0.038	0.144	0.087	0.107

Observando a tabela 4.3 verifica-se que as estimativas para o parâmetro de escala consoante os diferentes limiares obtidos não diferem muito entre si. Por outro lado, o mesmo não acontece para o parâmetro de forma, pois quando o limiar é dado pelo percentil 90, a estimativa obtida para este parâmetro diferencia-se das restantes.

4.2.3 Níveis de retorno

Os dados em estudo são referentes ao período de um ano, por isso, o nível de retorno do ano p tem como período de retorno $m = p \times n$, em que n é o número de observações por ano, que neste caso, corresponde à dimensão total da amostra. Assim, foram calculados os níveis de retorno a 5, 10 e 15 anos que podem ser consultados na tabela seguinte, respetivamente.

Tabela 4.4: Níveis de retorno a 5, 10 e 15 anos

m	5×n	10 ×n	15 ×n
Nível de retorno	1 061.547	1 374.003	1 598.944

Analisando a tabela 4.4, verifica-se que se espera que, em 5 anos, o montante 1 061.547 de indemnizações seja, em média, ultrapassado uma vez. Enquanto que, se o período de retorno for a 10 ou 15 anos, esse valor aumenta para 1 374.003 e 1 598.944, respetivamente.

4.2.4 Verificação do modelo

De forma a verificar a qualidade do ajuste à distribuição generalizada de Pareto, recorreu-se a gráficos de probabilidade, de quantis, de níveis de retorno e de densidade que podem ser consultados na figura seguinte.

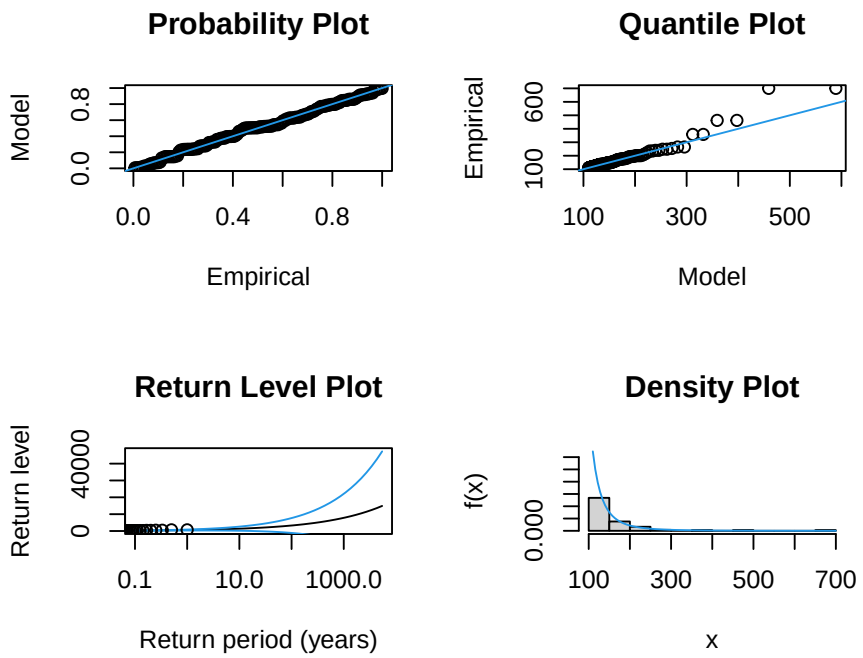


Figura 4.3: Gráficos de diagnóstico para o modelo de excessos acima de um limiar ajustado à sinistralidade do ramo automóvel existente na Noruega.

Ao observar a figura 4.3, verifica-se que tanto o gráfico de probabilidade como o de quantis, são constituídos por pontos aproximadamente lineares, embora no de quantis se veja uma minoria de observações a dispersar. Assim, conclui-se que os gráficos apresentados indicam que a distribuição generalizada de Pareto é um ajustamento razoável à modelação das excedências do limiar $u = 110\,000$.

Por curiosidade, foi testado o ajuste de duas funções convencionais à cauda da distribuição. Assim, através das funções *gamma_test()* e *lnorm_test()* do pacote *goft 1.3.6* do R, foram realizados os ajustes à distribuição Gama e à Lognormal, respetivamente. No primeiro caso, os parâmetros de escala e forma são desconhecidos e o teste é feito com base no rácio de dois estimadores da variância [8], enquanto que, no segundo caso, o teste consiste na transformação dos dados em observações normais. Considerando Z a cauda da distribuição, as hipóteses testadas em cada uma das respetivas situações descritas foram as seguintes:

$$H_0^1 : Z \sim Gama(\sigma, \xi) \quad vs \quad H_1^1 : Z \nsim Gama(\sigma, \xi).$$

$$H_0^2 : Z \sim Lognormal(\mu, \sigma) \quad vs \quad H_1^2 : Z \nsim Lognormal(\mu, \sigma).$$

É de salientar que os valor-p obtidos em ambos os testes foram aproximadamente zero, o que, aos níveis de significância usuais, leva à rejeição de cada uma das hipóteses nulas. Consequentemente, é possível concluir que nem a distribuição Gama nem a Lognormal são um bom ajuste para a cauda da distribuição em estudo.

4.3 Abordagem semi paramétrica

Nesta secção é aplicada uma abordagem semi paramétrica aos dados em estudo para analisar e modelar estes eventos extremos. Note-se que, em metodologias desta índole, a noção de ordem é fundamental.

4.3.1 Seleção do limiar

Analogamente à abordagem paramétrica, é necessário escolher um limiar. Embora se trate de uma metodologia subjetiva e nem sempre seja clara e fácil a decisão, a ferramenta utilizada foi o gráfico de Hill [16] apresentado na figura seguinte, através da função *hill()* do pacote *evir 1.7-4* do R.

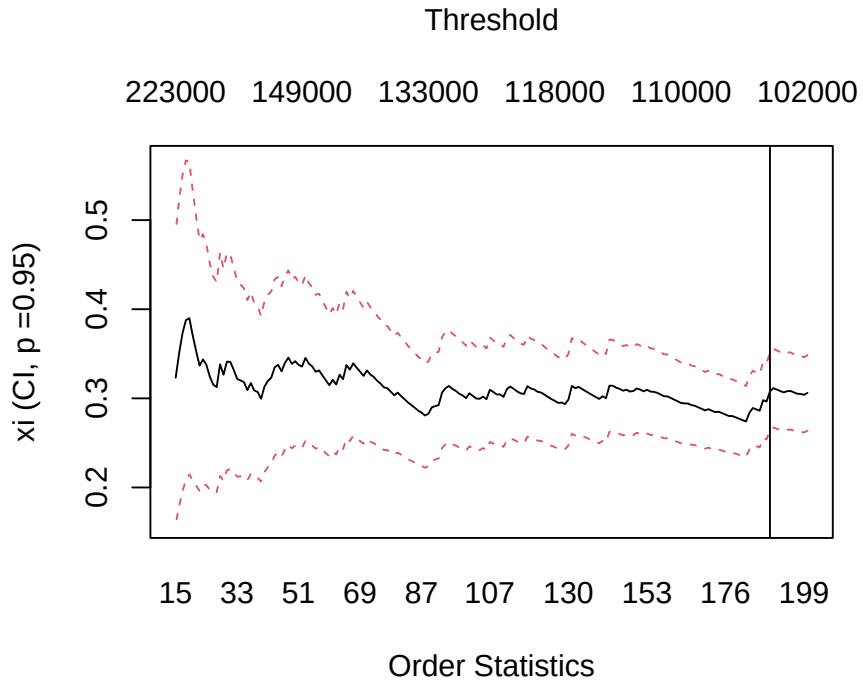


Figura 4.4: Gráfico de Hill para o cálculo de estimativas dos índices de valores extremos

Começou por se fazer o gráfico para todos os valores de k, contudo, a região de interesse ficava muito pequena, por isso estipulou-se um máximo de 200 para este valor. Ao observar o gráfico da Figura 4.4, verifica-se uma linha vertical preta sobre o valor de k escolhido, $k = 189$, uma vez que se situa numa área com alguma estabilidade, isto é, em que se verifica pequena variação nas estimativas dos parâmetros. Por outro lado, já se tinha chegado a este valor na abordagem paramétrica pela Regra de Ouro sugerindo assim ser uma boa opção, em que o limiar é dado por $u = 103\,500$.

4.3.2 Estimação dos parâmetros

Relativamente aos parâmetros, determinou-se uma estimativa para o índice de valores extremos, o estimador de Hill, cujo resultado obtido por (3.36) e considerando $k = 189$ foi, aproximadamente, 0.3131, o que corrobora os resultados do gráfico da Figura 4.4. Note-se que a estimativa do parâmetro de forma obtida, referente ao mesmo valor de k, na abordagem paramétrica, pelo método da máxima verosimilhança, foi de 0.317, muito próximo do estimador de Hill calculado.

Como mencionado na secção 3.4.2, a estimação do índice de valores extremos é essencial na abordagem semi paramétrica por determinar o comportamento da cauda direita da distribuição subjacente à amostra em estudo. Contudo, existem também outros parâmetros de interesse e de grande utilidade na análise de valores extremos, tais como a estimativa de uma probabilidade de excedência de um nível elevado e a estimativa de quantis extremos que foram abordados em 3.4.3 e 3.4.4, respetivamente.

Desta forma, calculou-se a estimativa de uma probabilidade de excedência de um nível elevado x, tendo sido considerados dois valores diferentes para este. Uma vez que o valor máximo amostral é 700 000 e como o valor de x deve ser um valor extremo, próximo ou superior ao valor máximo amostral, foram considerados os valores 700 000 e 800 000 para x, sendo que os resultados obtidos podem ser consultados na tabela seguinte.

Tabela 4.5: Estimativa de probabilidade de excedência do nível elevado x

x	700 000	800 000
$\hat{P}(X > x)$	0.0000023	0.0000015

Ao observar a tabela 4.5, conclui-se que a probabilidade de exceder o máximo amostral é, aproximadamente, 0.0000023, enquanto que há uma probabilidade de cerca de 0.0000015 de exceder o montante médio de indemnização de 800 000.

Dado que os quantis extremais são dos parâmetros mais importantes em eventos extremos, prosseguiu-se com o estudo dos mesmos. Através de (3.39) e considerando a probabilidade $p = 1/m$, em que, $m = N \times n_y$ com N o período de retorno (em anos) e n_y o número de observações por ano, foram calculadas as estimativas dos quantis extremais. Isto é, uma vez que os níveis de retorno a T anos são quantis extremais de probabilidade $p = 1 - \frac{1}{T}$, pode-se dizer que na tabela seguinte são apresentados os níveis de retorno a 5, 10 e 15 anos.

Tabela 4.6: Estimativa de quantis extremais

Período de retorno (em anos)	5	10	15
$\hat{q}_{1-p}$	884 451.6	1 098 852	1 247 616

Analisando a tabela 4.6 e relembrando que n refere a dimensão total da amostra, que neste caso diz respeito às observações de um ano e, por isso,$n_y = n$, conclui-se que se espera que, num período de 5 anos, o montante 884 451.6 das indemnizações seja, em

média, excedido uma vez. Já no caso do período de retorno ser a 10 ou 15 anos, este valor sofre um acréscimo para 1 098 852 e 1 247 616, respetivamente.

Capítulo 5

Conclusão

A atividade seguradora surge como forma de solução e prevenção dos indivíduos e famílias quando confrontadas com contingências inesperadas da vida, os sinistros, garantindo a sua compensação económica.

Esta dissertação pretende mostrar a utilidade e a consequente necessidade de incorporação da EVT na quantificação do risco de seguros, visto que esta se centra numa análise estatística de valores extremos em níveis de sinistralidade do ramo automóvel de uma companhia de seguros da Noruega num período de um ano.

Começou-se por tentar ajustar um modelo para os dados completos dos montantes médios das indemnizações, contudo, sem sucesso, uma vez que os eventos extremos se situam na cauda da distribuição e as distribuições comuns focam-se essencialmente nas observações centrais desta. Desta forma, o foco deste trabalho consistiu na modelação da cauda direita da distribuição segundo uma abordagem paramétrica e outra semi paramétrica, sendo que, em ambas, num primeiro momento procurou-se selecionar um limiar adequado e num segundo momento procedeu-se à estimação dos parâmetros do modelo. Além disto, foram também determinadas algumas quantidades fundamentais na previsão e prevenção de catástrofes associadas à ocorrência de sinistros com indemnizações elevadas, tais como os quantis extremais e os níveis de retorno de determinado número de anos.

Por outro lado, uma vez que apenas foi ajustado um modelo à cauda da distribuição, também seria interessante, futuramente, a procura de um modelo para as restantes observações.

Bibliografia

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (ver p. ii).

[2] M. Rydman. *Application of the Peaks-Over-Threshold Method on Insurance Data*. 2018 (ver pp. 1, 27).

[3] Y. Bensalah. «Steps in Applying Extreme Value Theory to Finance: A Review». Em: *Steps in Applying Extreme Value Theory to Finance: A Review* (2000). ISSN: 1192-5434 (ver p. 1).

[4] R. S. Embrechts P. e G. Samorodnitsky. «Extreme value theory as a risk management tool». Em: *North American Actuarial Journal* 3.2 (1999), pp. 30–41. ISSN: 1092-0277. DOI: 10.1080/10920277.1999.10595797 (ver p. 1).

[5] *Package CASdatasets*. URL: http://cas.uqam.ca/ (ver p. 3).

[6] A. Charpentier. *Computational actuarial science with R*. Taylor Francis Group, 2015. ISBN: 978-1-4665-9260-5 (ver p. 3).

[7] J. Eggers. *On Statistical Methods for Zero-Inflated Models*. 2015 (ver p. 20).

[8] E. Villaseñor J. A. andGonzález-Estrada. «A variance ratio test of fit for Gamma distributions». Em: 96 (2015), pp. 281–286. ISSN: 0167-7152. DOI: 10.1016/j.spl.2014.10.001 (ver pp. 20, 39).

[9] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. 1ª ed. Springer London, 2001. ISBN: 1-85233-459-2 (ver pp. 22, 30).

[10] H. K. Beirlant J. e T. J.L. «Estimation of the Extreme Value Index». Em: *Extreme Events in Finance: A Handbook of Extreme Value Theory and its Applications* (2016), pp. 97–115. DOI: 10.1002/9781118650318.ch5 (ver p. 22).

[11] F. Caeiro e M. Gomes. *Threshold Selection in Extreme Value Analysis*. 2014 (ver p. 22).

[12] R. L. Smith. *Risk Analysis and Extremes*. 2007 (ver p. 25).

[13] A. Scarrott C. MacDonald. «A review of extreme value threshold estimation and uncertainty quantification». Em: *REVSTAT - Statistical Journal* 10.1 (2012), pp. 33–60. DOI: https://doi.org/10.57805/revstat.v10i1.110 (ver p. 32).

[14] B. Hill. «A simple general approach to inference about the tail of a distribution». Em: *Annals of Statistics 13* (1975), pp. 331–341 (ver p. 33).

[15] A. M. I. F. N. C. Gomes M. I. «Análise de Valores Extremos: Uma Introdução». Em: (2013) (ver p. 34).

[16] S. Resnick e C. Starica. «Smoothing the Hill estimator». Em: *Advances in Applied Probability 29* (1997), pp. 271–293 (ver p. 40).

ANEXO I

Estimador da máxima verosimilhança de uma distribuição Binomial

Seja $(X_1, X_2, ..., X_k)$ uma amostra aleatória da distribuição Binomial, $B(n,p)$. A sua função de probabilidade é dada por

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, ..., n, 0 < p < 1.$$

Desta forma, tem-se a seguinte função de verosimilhança

$$L(n,p;x) = \prod_{i=1}^{k} P(X=x) = \prod_{i=1}^{k} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} = \left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^{\sum_{i=1}^{k} x_i} (1-p)^{kn-\sum_{i=1}^{k} x_i}.$$

Atendendo que a função logarítmica é uma função estritamente crescente, determinar um maximizante de $L = L_1$ é equivalente a determinar um maximizante de $lnL = lnL_1$, pelo que a função log-verosimilhança consiste em

$$\begin{aligned} l(n,p;x) = logL(n,p;x) &= log\left[\left[\prod_{i=1}^{k} \binom{n}{x_i}\right] p^{\sum_{i=1}^{k} x_i} (1-p)^{kn-\sum_{i=1}^{k} x_i}\right] \\ &= log\left[\prod_{i=1}^{k} \binom{n}{x_i}\right] + log\left(p^{\sum_{i=1}^{k} x_i}\right) + log\left[(1-p)^{kn-\sum_{i=1}^{k} x_i}\right] \\ &= \sum_{i=1}^{k} log\binom{n}{x_i} + \sum_{i=1}^{k} x_i log(p) + \left(kn - \sum_{i=1}^{k} x_i\right) log(1-p). \end{aligned}$$

Ao derivar a função de log-verosimilhança e igualando-a a zero, obtém-se

$$\begin{aligned}
\frac{\partial l}{\partial p}(n,p;x)=0 &\Leftrightarrow \frac{\sum_{i=1}^{k}x_i}{p}-\frac{\left(kn-\sum_{i=1}^{k}x_i\right)}{1-p}=0\\
&\Leftrightarrow \frac{\sum_{i=1}^{k}x_i-p\sum_{i=1}^{k}x_i-knp+p\sum_{i=1}^{k}x_i}{p(1-p)}=0\\
&\Leftrightarrow \sum_{i=1}^{k}x_i-knp=0\Leftrightarrow\sum_{i=1}^{k}x_i=knp\Leftrightarrow p=\frac{1}{kn}\sum_{i=1}^{k}x_i\Leftrightarrow p=\frac{1}{n}\bar{x}_k
\end{aligned}$$

Uma vez que

$$\frac{\partial^2 l}{\partial p^2}(n,p;x)=-\frac{\sum_{i=1}^{k}x_i}{p^2}+\frac{\left(kn-\sum_{i=1}^{k}x_i\right)(-1)}{(1-p)^2}=-\frac{\sum_{i=1}^{k}x_i}{p^2}-\frac{\left(kn-\sum_{i=1}^{k}x_i\right)}{(1-p)^2}<0,\forall p,$$

$\frac{1}{n}\bar{x_k}$ é um maximizante de $L=L_1$, pelo que, $\frac{1}{n}\bar{x_k}$ é o estimador da máxima verosimilhança de p, $\hat{p}=\frac{1}{n}\bar{x_k}$.