Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Construção de uma Tarifa de Responsabilidade Civil Automóvel

Ana Isa Veríssimo Neves

Mestrado em Matemática Aplicada à Economia e à Gestão

Trabalho de Projeto orientado por:
Professora Doutora Marília Antunes

2017

Agradecimentos

É com enorme orgulho que chego ao fim desta etapa tão importante e marcante da minha vida.

Começo por agradecer a todos aqueles que de alguma forma contribuíram, amigos e família, em especial à minha mãe, por toda a paciência para me suportar nos momentos menos bons ao longo destes anos, por todos os ensinamentos sábios que só uma mãe sabe dar, por todo o carinho e acima de tudo, obrigada por todo o incentivo para continuar. Sem ela nunca teria conseguido chegar onde cheguei. Ao meu pai, um muito obrigada por permitir que tudo isto se tornasse possível. Ao meu irmão, por estar sempre ao meu lado e por ter sempre uma sugestão ou uma crítica construtiva para me dar.

Agradeço à minha orientadora, Professora Doutora Marília Antunes, por ter logo aceite este desafio, por todo o apoio incondicional, por me ter levantado sempre que precisei, por ter partilhado comigo a sua sabedoria e me ter ensinado tanto, por toda a sua dedicação, por toda a sua paciência comigo, por toda a motivação, por toda a disponibilidade e por todo o carinho que me deu sempre. Um simples obrigado não chega. O meu muito obrigado por tudo e acima de tudo, por me ter ajudado a tornar este projeto real sem nunca me deixar desistir. Muito mais que uma Professora, uma amiga para a vida.

À Companhia, obrigada pela oportunidade e confiança na partilha de informação.

Ao Atuário Responsável, Doutor Vasco Barros, obrigada por toda a amizade e partilha de conhecimentos.

Deixo um especial agradecimento aos colegas atuários, em particular ao Paulo Pontes e Luís Jesus, pela partilha de experiências e conhecimento, bastante úteis para iniciar e desenvolver este projeto. Sem eles nada disto teria sido possível.

Agradeço aos/às meus/minhas colegas de Licenciatura e Mestrado. Entre estes, um especial agradecimento ao meu amigo Sandro Matos, pela pessoa genuína que é, sempre pronto a ajudar, meu grande amigo, por nunca me deixar desistir, por todas aquelas noites e fins-de-semana de estudo na faculdade, sempre incansável comigo, não tenho palavras para agradecer.

Não podia deixar de agradecer ao Professor Fernando Sequeira, por todo o apoio e amizade ao longo destes anos. Sem ele, a paixão por Estatística teria sido de todo, muito menor.

Ana Isa

Conteúdo

Lista de Figuras

Lista de Tabelas

Resumo

A atividade seguradora oferece aos seus clientes a transferência de eventuais responsabilidades destes, assim como prejuízos por danos sofridos pelos seus bens, mediante o pagamento de um prémio de seguro, valor que deverá permitir a obtenção de resultados de exploração satisfatórios. Estas responsabilidades e danos não são conhecidos no momento em que o prémio é calculado, pelo que deverão ser estimados. No entanto, diferentes indivíduos enquadram-se em diferentes classes de risco, pelo que um dos desafios da atividade seguradora é a definição de uma tarifa tecnicamente equilibrada que permita à empresa assegurar o cumprimento das suas responsabilidades, mas que seja também justa e adaptada a cada cliente. Identificar, diferenciar e quantificar o grau de risco de cada segurado permite cobrar o prémio adequado.

Neste trabalho aborda-se a modelação da tarifa da garantia de Responsabilidade Civil do ramo Automóvel. O custo associado à sinistralidade depende de duas componentes: a frequência de sinistralidade (expressa em número de sinistros por ano) e o custo associado a cada sinistro. Assim sendo, faz todo o sentido combinar, para a estimação do custo associado à sinistralidade, dois modelos: um para a frequência e outro para o custo por sinistro. O custo esperado associado a cada apólice é então dado pelo produto da frequência esperada com o custo esperado por sinistro.

No ramo Automóvel, as tarifas são calculadas considerando as características do tomador de seguro e do veículo, como também alguma informação relevante (como por exemplo: localidade do segurado), de forma a que se consiga obter um valor adequado para o Prémio pago pelo cliente. Neste projeto, assume-se que as variáveis explicativas a considerar em ambos os modelos são as mesmas.

Os Modelos Lineares Generalizados, pelas suas características e flexibilidade, apresentam-se como uma boa escolha para a modelação da tarifa de Responsabilidade Civil Automóvel.

Palavras-chave: Regressão Linear, Modelos Lineares Generalizados, Regressão de Poisson, *Hurdle Models*, *Clustering*.

Abstract

Insurance offers its clients a transfer of their possible liabilities, as well as damage for injuries suffered to their goods and property by means of payment of an insurance premium, which should provide a satisfactory profit to the insurance activity. These damages and liabilities are not known the moment the insurance premium is calculated, and therefore is based on estimates. However, different people fall into different risk categories; one of the challenges of the insurance business is to define technically balanced rates which enable the insurance company to fulfil its obligations but which are also fair and adjusted to each client. By identifying, differentiating and quantifying the risk level for each insured individual, an appropriate premium can be charged.

The subject of this work is the rate modelation of the civil liability coverage in the automobile insurance. The cost associated with the accident depends on two components: the frequency (expressed in number of claims per year) and the cost associated with each claim. Therefore, it makes sense, to combine the estimation of costs associated with the accident, two models: one for frequency and one for the cost per claim. The expected cost associated with each policy is then given by the product of the expected frequency by the expected cost per claim.

In the Automobile Insurance, rates are calculated by considering the characteristics of the policyholder and the vehicle, as well as some relevant information (for example: location of the insured), so that we can not obtain an appropriate value for the premium paid by the customer. In this work, it's assumed that the explanatory variables to consider in both models are the same.

The Generalized Linear Models, by their characteristics and flexibility, then present themselves as a good choice for modeling the rate of Automobile Insurance.

Keywords: Linear Regression, Generalized Linear Models, Poisson Regression, Hurdle Models, Clustering.

Capítulo 1

Introdução

Os modelos de regressão linear gozam de grande popularidade em diversas áreas de aplicação, da Economia à Saúde, pela sua forma simples e pela facilidade de interpretação que proporcionam na descrição/predição de variáveis quantitativas. Nestes modelos, o valor esperado da variável resposta expressa-se como uma função linear de um conjunto de variáveis explicativas. Na prática, existem inúmeras situações em que estes requisitos não são cumpridos, sendo necessário procurar um modelo alternativo. Assim, no início dos anos 70, surgem os Modelos Lineares Generalizados (Turkman e Silva (2000)) que tal como o próprio nome indica, constituem uma generalização do modelo linear. Em comum, todos estes modelos apresentam uma estrutura de regressão linear e o facto da variável resposta pertencer a uma distribuição pertencente à família exponencial. São, portanto, mais flexíveis, permitindo-se que a variável resposta possa ser de qualquer natureza desde que dentro da família exponencial (ainda que com algumas restrições na parametrização) como, por exemplo, Bernoulli, multinomial, Poisson, exponencial, Gama e, naturalmente, a distribuição normal. Também a estrutura linear ganha flexibilidade pois liga-se ao valor esperado da variável resposta através de uma função convenientemente escolhida - a chamada função de ligação - que no modelo de regressão linear é a função identidade.

Estes modelos foram utilizados inicialmente em grupos de investigação restritos, mas o aumento da capacidade computacional e o alargamento da disponibilidade no mercado de *software* estatístico vieram permitir a utilização ampla destes modelos. Na área dos Seguros, os modelos lineares generalizados constituem uma abordagem importante na modelação da sinistralidade, nomeadamente no que se refere à tarifação (Santos (2008)).

O contrato de seguro garante a reparação ou pagamento, pela seguradora, dos danos decorrentes de um sinistro que se enquadre nas condições do referido contrato, mediante o pagamento de um prémio pelo tomador de seguro. Neste contexto, o problema que se coloca é o do cálculo do prémio a cobrar ao tomador de seguro, que deve ser suficiente para fazer face ao custo futuro de eventuais sinistros, mas também ser suportável por este. Uma vez que o prémio deverá fazer face à sinistralidade futura, sendo assim função do risco, deverá ter em conta não só o número de sinistros que uma apólice poderá gerar, como também o custo associado a estes.

A maior percentagem a cobrar ao segurado parte do valor de um Prémio Base, a que se dá o nome de Prémio de Risco. Uma abordagem para a determinação do Prémio de Risco consiste em fazê-lo corresponder ao valor esperado da indemnização que será devida associada à apólice. O valor total de indemnização associada a uma apólice corresponde ao somatório dos valores das indemnizações correspondentes aos sinistros ocorridos no período de tempo em que o contrato é válido. Estamos, portanto, perante um processo composto, uma vez que integra duas componentes - o número de sinistros e o custo associado a cada sinistro. Formalizando um pouco a questão, seja $N(t)$ o número de sinistros associados a determinada apólice, válida durante um período de tempo de duração igual a t. O custo total associado a estes sinistros (que corresponde

ao valor total de indemnização a pagar pela seguradora), é dado por

$$Y(t) = \sum_{j=1}^{N(t)} C_j, \tag{1.1}$$

onde C_j é o custo associado ao $j-$ésimo sinistro. Tratando-se de uma contagem num intervalo de tempo eventualmente variável, o mais frequente é considerar que $N(t)$ é bem descrito por um processo de Poisson, designando-se $Y(t)$ por processo de Poisson composto.

Considerando que sinistros sob uma mesma apólice ocorrem de forma independente e que o custo a cada um deles associado é a realização independente de uma variável aleatória C, o custo total esperado associado a uma apólice com duração t é

$$\begin{aligned} E[Y(t)] &= E_{N(t)}\left[E_{Y(t)}[Y(t)|N(t)]\right] \\ &= E_{N(t)}\left[N(t).E[C]\right] \\ &= E[N(t)].E[C]. \end{aligned} \tag{1.2}$$

Este resultado é válido mesmo que se considere para $N(t)$ um modelo diferente do processo de Poisson. Como o que se pretende é adequar o prémio ao que se espera de cada segurado, o caminho a seguir é incorporar informação relativa a este e ao veículo na modelação quer de $N(t)$ quer de C. De uma forma geral, dado um perfil $\mathbf{X}_i$ composto por características pessoais e do veículo ou outras, passamos a interessar-nos por modelar

$$Y_i(t|\mathbf{X}_i) = \sum_{j=1}^{N(t|\mathbf{X}_i)} C_{j|\mathbf{X}_i}. \tag{1.3}$$

Retomando a expressão do Prémio de Risco, que se apresenta como uma função de tipo multiplicativo entre a frequência esperada de sinistros e o custo esperado por sinistro,

$$\textit{Prémio de Risco} = E[\textit{custo}] = E[\textit{frequência}] \times E[\textit{custo}_{\textit{sinistro}}],$$

tem a forma do produto de duas quantidades que podem ser modeladas usando modelos da classe dos Modelos Lineares Generalizados:

$$E[Y(t|\mathbf{X})] = E[N(t|\mathbf{X})].E[C|\mathbf{X}]. \tag{1.4}$$

Como foi referido anteriormente, a seguradora não se limita a cobrar ao segurado o valor do Prémio de Risco, mas sim o mesmo agravado por um montante necessário para fazer face às comissões, Bónus Médio a que se destina a tarifa em construção, custos de exploração da seguradora e margem de lucro.

Todos estes encargos convertem o Prémio de Risco em Prémio Comercial. Contudo, a entidade seguradora há de satisfazer mais alguns agravamentos ao Prémio Comercial que dão origem ao Prémio Total (valor pago pelo tomador de seguro à seguradora). O Prémio Total será o Prémio Comercial acrescido das taxas e impostos legais.

No caso do seguro automóvel é acrescido 2,5% do Prémio Comercial para o Fundo de Garantia Automóvel[1], 2% do Prémio Comercial para o Instituto Nacional de Emergência Médica (INEM)[2], 9% do Prémio Comercial para o Imposto de Selo[3] e 0,21% do Prémio Comercial para a Prevenção Rodoviária[4]. Também é cobrado um valor adicional de 2.69€ pela carta verde, em

[1] Fundo que se destina a pagar indemnizações devidas em caso de acidente automóvel ocorrido com veículo sem seguro válido, assim como danos corporais maus em que não seja possível identificar o autor do sinistro.

[2] O INEM tem a seu cargo as ações de socorro pré-hospitalar, assim como o transporte e receção de doentes.

[3] O valor apurado com esta taxa vai diretamente para a Autoridade Tributária.

[4] A Prevenção Rodoviária é uma associação, sem fins lucrativos, que tem o objetivo de prevenir os acidentes rodoviários e as suas consequências.

que 2.50€ são lucro e 0.19€ são para Imposto de Selo e INEM.

Em resumo, para obter o prémio pago pelo tomador de seguro à seguradora, é necessário determinar o Prémio de Risco, de seguida o Prémio Comercial e por fim, obter o Prémio Total (prémio pretendido).

Pelo exposto acima, a modelação do Prémio de Risco encaixa-se muito naturalmente no paradigma dos Modelos Lineares Generalizados. Neste caso, irão ser utilizados modelos desta classe para estimar a frequência esperada de sinistralidade e o valor esperado do custo por sinistro. As variáveis resposta são a frequência ou o custo por sinistro, com os fatores de classificação disponíveis utilizados como variáveis explicativas.

Os modelos para a frequência, por norma, por serem modelos de contagem, assumem como função de ligação a função logaritmo (log), com uma estrutura de erro Poisson. Desta forma, o número de sinistros, normalmente, é considerado um processo de Poisson (Murphy, Brockman and Lee (2000)). De esperar, no entanto, será a possibilidade de ocorrência de um número excessivo de zeros e ainda de falta de sobredispersão ou subdispersão, desviando-nos do modelo de Poisson mais simples.

Os modelos para o custo, por norma, por serem modelos com variável resposta contínua e apresentarem acentuada assimetria, assumem também a função logaritmo (log) como função de ligação, com uma estrutura de erro Gama (Murphy, Brockman and Lee (2000)). Como alternativa, pode considerar-se o modelo Normal para a variável resposta logaritmizada, de forma a permitir uma relação não-linear entre a variável de interesse e as variáveis explicativas.

1.1 Enquadramento e Objetivos

A Companhia de Seguros que me disponibilizou toda a base de dados com a toda a informação real necessária para a realização deste projeto, associando-se a profissionais em ascensão e a eventos desportivos com notoriedade em várias modalidades, apoia também diversas atividades culturais e desportivas, tais como o Automobilismo, Atletismo, Ténis, Vela e Hipismo.
No âmbito pedagógico mantém estreitas relações e parcerias a diversas associações e instituições de cariz sociopedagógico e cultural, privilegiando o apoio a ações direcionadas a crianças.

Assim, projeta-se como uma empresa do futuro e assume-se como o parceiro de confiança em todas as situações, criando valor económico e social e contribuindo, decisivamente, para o progresso e bem-estar da nossa comunidade.

O objetivo deste trabalho é construir uma fórmula de cálculo de um prémio de risco, o mais adequado possível às características do cliente, isto é, encontrar um modelo adequado para a frequência e um modelo adequado para o custo por sinistro de tal modo que se consiga prever o número esperado de sinistros por unidade de tempo para um indivíduo com determinado perfil, assim como o custo esperado por sinistro que esse indivíduo possa vir a custar à Companhia.

1.2 Estrutura do Trabalho

A Base de Dados consiste numa carteira de Responsabilidade Civil do ramo Automóvel com registos desde 1 de janeiro de 2011 a 30 de junho de 2013 (dois anos e meio). Estão

integradas na base de dados as apólices que tenham estado em vigor num determinado momento, independentemente da duração do período de tempo em que estiveram em vigor. As variáveis presentes na base de dados são relativas às características do indivíduo detentor da apólice, incluindo a sua localização geográfica (distrito, concelho e freguesia de morada), bem como características do veículo.

Este relatório está organizado com a seguinte estrutura:

No capítulo 2, é apresentado o Suporte Teórico. São apresentadas brevemente as ferramentas de análise exploratória dos dados, do seu tratamento, e da construção dos modelos e sua comparação. Uma vez que muitas das características são de natureza categórica com um elevado número de categorias, como por exemplo a região de morada do detentor da apólice, serão apresentadas as metodologias de *clustering* utilizadas para o agrupamento num número razoável de *clusters* homogéneos. Segue-se a apresentação dos Modelos Lineares Generalizados, onde se começa por definir o modelo mais simples, o modelo linear, sendo depois apresentados os modelos para variáveis resposta seguindo outras distribuições tais como: Bernoulli na regressão logística, regressão de Poisson, regressão Gama e dois modelos para excesso de zeros: *Hurdle* Poisson e *Hurdle* Binomial Negativo.

Na atividade seguradora, não é esperado que cada segurado pague por aquilo que virá a receber em caso de sinistro, mas que, sendo o prémio de seguro considerado aceitável por parte do segurado para o risco que cobre, a receita total da seguradora seja suficiente para cobrir todas as despesas. Assim, para comparar o desempenho dos diferentes modelos estimados, são calculadas diversas medidas de erro, que serão descritas também no capítulo 2. Apresentam-se neste capítulo, também, algumas ferramentas de diagnóstico dos modelos.

No capítulo 3, na Análise Exploratória dos Dados, serão apresentadas as variáveis em estudo, assim como o seu tratamento de dados, onde é realizada uma breve análise exploratória de cada variável.

No capítulo 4, é apresentada a descrição relativa à construção dos modelos, com exemplos práticos e discussão de resultados.

Por fim, no capítulo 5, serão apresentadas as principais conclusões retiradas deste projeto.

Capítulo 2

Suporte Teórico

Este capítulo serve como uma base teórica consistente que permite dar apoio na elaboração prática do projeto, isto é, um auxílio em todos os procedimentos até à obtenção do modelo pretendido e, posteriormente, na obtenção de conclusões.

2.1 Identificação de grupos homogéneos e redução da dimensionalidade - *Clustering*

O *Clustering*, também denominado por Análise de *Clusters*, é um procedimento da Estatística Multivariada que tem como finalidade agrupar um conjunto de dados em subgrupos homogéneos, denominados por *clusters*.

O *clustering* enquadra-se no conjunto de técnicas ditas de classificação não supervisionada, em que o objetivo é a alocação dos objetos em classes distintas que não estão definidas *a priori*. O objetivo é proceder ao agrupamento de elementos com base nas suas características, de tal modo que elementos pertencentes a um mesmo *cluster* sejam o mais semelhantes possível, e os elementos pertencentes a *clusters* diferentes sejam o mais dissemelhantes possível. Enquanto técnica exploratória, espera-se que permita identificar padrões, que os *clusters* tenham significado e sejam interpretáveis.

Em estudos envolvendo diversas variáveis, cada objeto é caracterizado por um vetor em que cada componente corresponde à medição de uma quantidade, a uma contagem, a uma pontuação numa escala ou à observação de uma característica categórica. Convencionou-se em Estatística que estes vetores devem ser reunidos numa matriz de dados, geralmente designada por $\mathbf{X}$, de dimensão $n \times p$, em que a cada linha corresponde o vetor que caracteriza um objeto ou indivíduo. A notação habitual é

$$\mathbf{X} = [x_{ij}], \qquad i = 1, \ldots, n; \quad j = 1, \ldots, p, \tag{2.1}$$

sendo x_{ij} o valor da variável j para o objeto i (Antunes (2010)).

Identificar *clusters* compostos por objetos semelhantes implica, pois, tendo em atenção a complexidade do vetor que caracteriza os objetos, definir semelhança ou proximidade. A distância euclidiana corresponde à distância geométrica entre dois objetos no plano multidimensional. Esta distância, entre o i-ésimo e o j-ésimo objetos caracterizados por p variáveis, é traduzida pela seguinte expressão:

$$d_E(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{2.2}$$

seguindo a notação apresentada em (2.1).

Esta medida de distância tem interpretação física e só deve ser considerada nas medições efetuadas na mesma unidade de medida. Se tal não acontecer, calcular esta distância faz pouco sentido e, inclusivamente, não existe uma unidade em que se possa expressar.

Como na prática é frequente que os objetos sejam caracterizados por variáveis medidas em unidades diferentes, a eliminação da unidade de medição por via da padronização é o recurso habitual. A padronização dos dados consiste na divisão dos valores de cada variável pelo desvio padrão amostral da variável. Desta forma, todas as variáveis irão ter igual "importância" no cálculo da distância, removendo-se o efeito de escala e também as unidades de medição. O desvio padrão para a k-ésima variável é estimado pelo desvio padrão amostral,

$$\widehat{s}_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}, \tag{2.3}$$

onde

$$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \tag{2.4}$$

é a média das observações realizadas sobre a variável X_k.

A matriz de distâncias entre objetos, de dimensão $n \times n$, constitui a base de trabalho para a aplicação de diversos métodos de análise classificatória para a constituição de *clusters*. Estes métodos podem ser hierárquicos ou não-hierárquicos.

No presente trabalho, o *clustering* será utilizado para reduzir a dimensionalidade (mais concretamente a diversidade) de categorias de algumas variáveis que descrevem as apólices, nomeadamente *distrito*, *concelho* e *marca* do automóvel de forma a simplificar, posteriormente, o processo de modelação.

Atendendo aos objetivos principais do trabalho - a estimação do prémio de risco - interessa-nos agrupar as variáveis categóricas num menor número de classes que sejam homogéneas entre si no que respeita às características de maior interesse - o número médio de sinistros e o custo médio por sinistro. Assim, para cada uma das variáveis categóricas a reagrupar é construída uma matriz com tantas linhas quanto o número de categorias (são estes os objetos a agrupar) e com duas colunas, uma referente ao número médio de sinistros por apólice nessa categoria e outra referente ao custo médio por sinistro. Por se tratarem de variáveis com escalas muito diferentes, as variáveis são padronizadas.

A matriz $\mathbf{X}$ a que nos referimos agora, tem dimensão $k \times 2$ onde k representa o número de categorias da variável categórica a reagrupar. Uma vez que cada categoria é caracterizada por um vetor bidimensional de variáveis quantitativas, a medida de semelhança a utilizar será a distância euclidiana.

- **Métodos Não Hierárquicos, baseados em particionamento**

Nos algoritmos não hierárquicos, baseados em particionamento, dado um conjunto de n objetos organizados numa matriz de dados $\mathbf{X}$, a tarefa consiste em identificar um número K de *clusters*, $C_1, ..., C_K$, tais que cada objeto pertence a um e um só *cluster* C_k, $k = 1, \ldots, K$.

Existem diversas variantes deste algoritmo. Em todas elas, fixa-se à partida o número, K, de classes que se pretende constituir e (regra geral) faz-se uma classificação inicial dos n indivíduos nas K classes (Cadima (2010)). Num processo automático, os K centros da primeira iteração são determinados aleatoriamente e cada elemento passa a pertencer à classe cujo centro está mais próximo, considerando a distância euclidiana. Nas iterações seguintes, os centros dos *clusters* são recalculados tendo em conta a sua constituição e os pontos novamente alocados, de forma a pertencerem ao *cluster* de cujo centro estão mais próximos. O processo repete-se até que se obtenha convergência, a qual poderá consistir em não haver mais mudanças ou outro critério que se considere satisfatório.

- **Métodos Hierárquicos**

Existem dois tipos de métodos hierárquicos: os aglomerativos (que vão agrupando) e os divisivos (que vão desagregando), menos utilizados. O dendrograma é a representação gráfica, com aspeto de árvore, que ilustra toda a sequência de aglomeração ou de divisão, exibindo os grupos formados por agrupamento de observações em cada passo e os seus níveis de similaridade ou de distância. O nível de similaridade ou de distância é medido ao longo do eixo vertical e as diferentes observações são organizadas convenientemente ao longo do eixo horizontal.

Referir-nos-emos nesta secção apenas aos métodos aglomerativos pois são os mais intuitivos e mais frequentemente utilizados. Nestes métodos, em geral, são determinadas a partir de n *clusters* iniciais (em que cada *cluster* é constituído por apenas um elemento), sucessivas fusões de *clusters* (o par considerado mais semelhante), reduzindo-se em cada fusão, em uma unidade, o número de *clusters*, até que todos os elementos passam a pertencer a um só *cluster*.

É, portanto, necessário definir o critério de fusão. A escolha do método depende não só do tipo de *clusters* que se pretende obter como também do tipo de vetor que caracteriza os objetos a agrupar. Posto de outra forma, a escolha do critério de fusão está condicionada pela medida de distância ou de dissemelhança entre objetos utilizada.

Entre os critérios de fusão mais populares encontram-se os critérios *single linkage* (vizinho mais próximo) e *complete linkage* (vizinho mais longe) e o método do centróide. Os dois primeiros operam sobre a matriz de distância ou de dissemelhança dos dados, não requerendo o acesso aos dados. O método do centróide exige o acesso à matriz de dados para que nas diferentes etapas os centróides possam ser calculados bem como as distâncias a estes.

- Método do vizinho mais próximo (*single linkage*)

Este método, que se deve a Sneath (1957), assenta no uso da matriz de distâncias ou de dissemelhanças e define a distância entre *clusters* como sendo a menor das distâncias entre objetos, pertencendo um a cada *cluster*:

$$d(C_k, C_l) = \min_{\mathbf{x}_i \in C_k} \min_{\mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j). \tag{2.5}$$

Este critério tende a produzir *clusters* desequilibrados e dispersos, com aspeto alongado, de "encadeamento", especialmente em grandes conjuntos de dados. Não é dada relevância à estrutura do cluster (Everitt (2011)).

- Método do vizinho mais longe (*complete linkage*)

Este método, que se deve a Sorensen (1948), assenta também no uso da matriz de distâncias ou de dissemelhanças e define a distância entre *clusters* como sendo a maior das distâncias entre objetos, pertencendo um a cada *cluster*:

$$d(C_k, C_l) = \max_{\mathbf{x}_i \in C_k} \max_{\mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j). \tag{2.6}$$

Este critério tende a produzir *clusters* compactos, ditos "esféricos", com aproximadamente o mesmo diâmetro (distância máxima entre objetos). Neste método também não é dada relevância à estrutura do *cluster* (Everitt (2011)).

- **Avaliação da qualidade do agrupamento**

A homogeneidade dos *clusters* é avaliada por uma função conveniente, que pode envolver, por exemplo, a distância entre cada objeto (ponto de $\mathbb{R}^p$) e o centróide do *cluster* a que o ponto pertence. O centróide do *cluster* é considerado o ponto representativo do *cluster* e, dependendo

da forma de avaliar a homogeneidade do *cluster*, a melhor solução será aquela para a qual a referida função é máxima ou mínima. Estas funções designam-se genericamente por funções *score*.

A escolha da função *score* depende da definição de distância considerada. Seja $d(\mathbf{x}, \mathbf{y})$ a distância entre dois pontos $\mathbf{x}, \mathbf{y}$ (na aplicação prática do algoritmo, serão duas linhas da matriz de dados $\mathbf{X}$).

A maioria das funções *score* realçam dois aspetos: (1) os clusters devem ser compactos; e (2) devem estar o mais afastados possível uns dos outros. Assim, é bastante intuitivo que o particionamento $\mathcal{C}$(*clustering*) seja avaliado segundo

- a variação intra-*clusters* (*within clusters variation*): $wc(\mathcal{C})$; e
- a variação entre *clusters* (*between clusters variation*): $bc(\mathcal{C})$,

onde $wc(\mathcal{C})$ mede o quão compactos os *clusters* são e $bc(\mathcal{C})$ mede o quão afastados se encontram.

O centro do *cluster* C_k, $\mathbf{r}_k$, define-se como o seu centróide, considerando-se que é o ponto que representa o *cluster*:

$$\mathbf{r}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}, \tag{2.7}$$

onde n_k é o número de pontos no k-ésimo *cluster*.

Uma medida simples para $wc(\mathcal{C})$ é a soma dos quadrados das distâncias dos pontos aos centróides dos *clusters* a que os pontos pertencem,

$$wc(\mathcal{C}) = \sum_{k=1}^{K} wc(C_k) = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k)^2. \tag{2.8}$$

No caso em que $d(\mathbf{x}, \mathbf{r}_k)$ é a distância euclideana, $wc(\mathcal{C})$ designa-se por soma dos quadrados dentro dos *clusters* (*within-cluster-sum-of-squares*).

A variação entre os *clusters* pode ser avaliada através da distância entre os centros dos *clusters*:

$$bc(\mathcal{C}) = \sum_{1 \leq j < k \leq K} d(\mathbf{r}_j, \mathbf{r}_k)^2. \tag{2.9}$$

A qualidade de um particionamento $\mathcal{C}$ é avaliada através de uma combinação monótona de $wc(\mathcal{C})$ e $bc(\mathcal{C})$, como por exemplo o rácio $\frac{bc(\mathcal{C})}{wc(\mathcal{C})}$.

A variação intra-*clusters*, $wc(\mathcal{C})$, é também uma medida global pois para que C_k tenha uma contribuição pequena para a medida, é necessário que todos os pontos do *cluster* estejam relativamente próximos do centro do *cluster*. Portanto, procurar um particionamento que possua um valor pequeno desta quantidade conduz à obtenção de *clusters* esféricos. O algoritmo *K-means* é um exemplo da utilização deste princípio - utiliza a média como ponto central dos *clusters* e para d a distância euclideana de forma a encontrar o particionamento $\mathcal{C}$ que minimiza a variação intra-*clusters*, para medidas num espaço euclideano $\mathbb{R}^p$.

2.2 Modelos Lineares Generalizados

Começa-se por apresentar o modelo mais simples, o modelo de Regressão Linear, que consiste na verificação da existência de uma relação funcional de tipo linear entre uma variável dependente e uma ou mais variáveis independentes.

Em muitos estudos estatísticos somos confrontados com problemas, em que o objetivo principal é o de estudar a relação entre variáveis estatísticas, isto é, analisar a influência que uma ou mais variáveis (*explicativas*), medidas em indivíduos ou objetos, têm sobre uma variável de

interesse a que damos o nome de *variável resposta* (Turkman e Silva (2000)).

O modelo linear normal, "criado" no início do século XIX por Legendre e Gauss, dominou a modelação estatística até meados do século XX, embora vários modelos não lineares ou não normais tenham entretanto sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas pelo modelo linear normal (Turkman e Silva (2000)).

Devido ao grande número de modelos que englobam e à facilidade de análise associada ao rápido desenvolvimento computacional que se tem verificado nas últimas décadas, os Modelos Lineares Generalizados têm vindo a desempenhar um papel cada vez mais importante na análise estatística, apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo facto das distribuições se restringirem à família exponencial e por exigirem a independência das respostas (Turkman e Silva (2000)).

2.2.1 Modelo Linear

Estes modelos são utilizados quando a variável resposta apresenta uma distribuição Normal. No entanto, o modelo escolhido deve ser coerente com o que realmente acontece na prática. Deve-se ter em conta algumas considerações:

- O modelo selecionado deve ser condizente tanto no grau como no aspeto da curva (no caso da regressão linear simples – tem apenas uma variável explicativa), para representar em termos práticos o problema em estudo;
- O modelo deve conter apenas as variáveis que são relevantes para explicar o problema.

Como é natural, para este projeto, irão ser consideradas mais do que uma variável explicativa e portanto a regressão linear a aplicar é a regressão linear múltipla. Esta análise tem por objetivo estabelecer uma equação que possa ser usada para predizer valores de Y, variável resposta, para os valores dados das diversas variáveis independentes. O modelo é traduzido pela seguinte expressão:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon, \tag{2.10}$$

onde $x_1, x_2, ..., x_p$ são variáveis explicativas (também denominadas por regressoras, preditoras ou covariáveis), $\beta_0, \beta_1, \beta_2, ..., \beta_p$ são os parâmetros desconhecidos do modelo e ε é o erro aleatório associado a Y. Logo $E[\varepsilon] = 0$ e $V[\varepsilon] = \sigma^2$.

Como $E[\varepsilon] = 0$, implica que:

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{2.11}$$

O modelo normal (modelo linear clássico) pressupõe que a variância da resposta seja constante. Contudo, surgem por vezes, situações de variáveis resposta de natureza contínua, em que a variância não é constante.

Se a variável resposta apresentar acentuada assimetria positiva, uma possibilidade consiste em recorrer a uma transformação logarítmica, para estabilizar a variância.

Seja Y a variável resposta. Se $log(Y) \sim N(\mu, \sigma)$, então $Y \sim lognormal(\mu, \sigma)$ e, nesse caso, a expressão que relaciona as variáveis é dada por:

$$\mu = E[log(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{2.12}$$

Como $\varepsilon \sim N(0, \sigma)$, tem-se que:

$$\mu = E[log(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{2.13}$$

e por conseguinte, o valor esperado de Y é traduzido pela seguinte expressão:

$$E[Y] = e^{\mu + \sigma^2/2} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p} \times e^{\sigma^2/2}. \tag{2.14}$$

A aplicação dos Modelos Lineares Generalizados tem-se verificado em diferentes áreas científicas, nomeadamente na construção de tarifas no mercado segurador. Estes modelos permitem que a variável resposta, Y, não apresente distribuição normal e apresentam também algumas características fundamentais:

i.) As variáveis resposta, Y_i, são variáveis aleatórias independentes e seguem uma distribuição da família exponencial, tais como: a distribuição normal (Gaussiana), Poisson, Gama, binomial ou a inversa da normal. No entanto, os Modelos Lineares Generalizados podem estender-se a distribuições de famílias exponenciais multivariadas ou não-exponenciais, como é o exemplo da distribuição binomial negativa, assim como a algumas situações em que a distribuição de Y não está completamente especificada.

ii.) A componente sistemática consiste numa combinação linear de variáveis preditoras, havendo p variáveis preditoras e n observações:

$$v_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{2.15}$$

$\forall_i \in \{1, ..., n\}$.

As variáveis preditoras não têm de ser todas da mesma natureza, ou seja, umas podem ser categóricas e outras quantitativas, por exemplo.

iii.) A função de ligação é uma função diferenciável e monótona, g, que associa as componentes aleatória e sistemática, através de uma relação da forma:

$$g(\mu_i) = v_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{2.16}$$

$\forall_i \in \{1, ..., n\}$.

Estes modelos permitem que tanto as variáveis explicativas como a variável resposta sejam de natureza contínua ou discreta.

A escolha da função de ligação vai depender do tipo de resposta e do estudo em particular.

As três etapas fundamentais para a estimação destes modelos são: formulação; ajustamento; seleção e validação dos modelos.

Na **formulação** do modelo há que ter em consideração a escolha da distribuição para a variável resposta. Assim, uma análise preliminar dos dados é fundamental para que se possa fazer uma escolha adequada da família de distribuições a considerar. É necessário efetuar uma escolha das variáveis preditoras, tendo em atenção a natureza destas e por fim escolher a função de ligação adequada.

O **ajustamento** do modelo passa pela estimação dos parâmetros, isto é, pela estimação dos coeficientes β associados e do parâmetro de dispersão, ϕ, caso exista.

Nos Modelos Lineares Generalizados, o parâmetro β (parâmetro de interesse) é estimado pelo Método da Máxima Verosimilhança e o parâmetro de dispersão ϕ, quando existe, é estimado pelo Método dos Momentos.

A **seleção e validação** do modelo tem por objetivo encontrar submodelos com um número moderado de parâmetros que ainda sejam adequados aos dados, detetar discrepâncias entre os dados e os valores preditos, averiguar a existência de *outliers* ou/e observações influentes, etc.

2.2.2 Família Exponencial

Como já foi referido anteriormente, os Modelos Lineares Generalizados pressupõem que a variável resposta tenha uma distribuição pertencente à Família Exponencial.

Diz-se que a variável aleatória Y tem distribuição pertencente à Família Exponencial se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever da seguinte forma:

$$f(y;\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{2.17}$$

onde:

- $f(y;\theta,\phi)$ é a função de probabilidade para variáveis aleatórias discretas, Y, ou a função densidade de probabilidade para variáveis contínuas, Y;
- $a(.), b(.), c(.)$ são funções que variam consoante a distribuição da família exponencial a que pertencem;
- $\theta = g_c(\mu)$, o parâmetro canónico da família exponencial em questão, é a função do valor esperado $\mu \equiv E[Y]$. $g_c(.)$ não depende de ϕ;
- ϕ é o parâmetro de dispersão, que em algumas famílias é um parâmetro fixo conhecido, enquanto que noutras famílias é um parâmetro desconhecido, calculado a partir dos dados juntamente com θ.

Se X é uma variável aleatória com distribuição pertencente à família exponencial, definida em 2.17 tem-se:

$$E[X] = \mu = b'(\theta); \tag{2.18}$$

$$V[X] = a(\phi)b''(\theta). \tag{2.19}$$

2.2.3 Regressão de Poisson

A distribuição de Poisson surge com muita frequência associada à contagem de acontecimentos aleatórios (quando se pode admitir que não há acontecimentos simultâneos), como é o caso da ocorrência de sinistros.

Uma variável aleatória discreta tem distribuição de Poisson (λ_i), se toma valores em $\mathbb{N}_0$ com função massa de probabilidade dada por:

$$P[Y_i = y_i] = e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}, \tag{2.20}$$

com $\lambda_i > 0$ e $y_i \in \{0, 1, 2, \dots\}$.

Neste modelo, a variância toma o mesmo valor que o valor médio:

$$E[Y_i] = V[Y_i] = \lambda_i. \tag{2.21}$$

i.) A distribuição de Poisson (λ) pertence à família exponencial, definida em 2.17.

No caso desta distribuição tem-se:

- $f(y;\lambda) = e^{-\lambda}\frac{\lambda^y}{y!}$
- $\theta = log(\lambda)$
- $\phi = 1$
- $a(\phi) = 1$
- $b(\theta) = e^{\theta} = \lambda$
- $c(y,\phi) = -log(y!)$

ii.) A função de ligação para uma componente aleatória com distribuição de Poisson é a função logarítmica, daí este modelo se designar por Modelo *Log-Linear*.

Usando uma função de ligação logarítmica (*log link*), pode calcular-se o número esperado de ocorrências (λ_i), de um indivíduo futuro, com perfil $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, por unidade de tempo (neste caso é o ano), para cada caso i, da seguinte forma:

$$\lambda_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}. \tag{2.22}$$

λ_i designa-se como taxa do processo e a sua estimação, em função do perfil, constitui o interesse na construção deste tipo de modelos.

Quando os dados disponíveis dizem respeito a contagens realizadas sobre períodos de tempo de duração diferente (diferentes tempos de exposição), então é necessário incluir no modelo tal informação (termo *offset*).

Considerando que cada caso i esteve exposto ao risco um tempo E_i, então Y_i (número de sinistros do indivíduo i em E_i, tempo de exposição) segue uma distribuição de *Poisson* de parâmetro μ_i onde,

$$\mu_i = E_i\lambda_i = E_i e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}. \tag{2.23}$$

De forma equivalente,

$$\frac{\mu_i}{E_i} = \lambda_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}. \tag{2.24}$$

Na estimação do modelo, porque cada contagem se refere a um tempo de exposição potencialmente distinto, E_i, considera-se o modelo na forma:

$$log(\mu_i) = log(E_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{2.25}$$

sendo $log(E_i)$ o termo *offset*.

Interpretação dos Parâmetros do Modelo

Os parâmetros do modelo são interpretados em termos das suas exponenciais, em que e^{β_0} representa o número de sinistros esperados para um indivíduo cujas características sejam as correspondentes à categoria de referência das variáveis nominais (categóricas) e apresenta valor zero para as variáveis quantitativas.

Para uma variável categórica, binária, seja x_j, e^{β_j} representa o termo multiplicativo a introduzir no cálculo do valor esperado de Y quando um indivíduo i possui a característica ($x_{ij} = 1$).

Comparando dois indivíduos semelhantes em todas as restantes variáveis, o risco relativo é dado por $RR = e^{\beta_j}$, representando o número de ocorrências mais que se esperam para o indivíduo com $x_j = 1$ relativamente ao que tem $x_j = 0$, em termos multiplicativos.

No caso das variáveis explicativas nominais com mais categorias $(k+1)$ são definidas variáveis *dummy*, estimando-se por isso, k parâmetros associados a essa variável, sejam $\beta_{j1}, ..., \beta_{jk}$.

A interpretação de cada parâmetro é semelhante ao caso binário, comparando-se o número esperado de ocorrências de um perfil com cada uma das categorias com o perfil que possui a categoria de referência.

As características principais do modelo de Poisson são:

- proporcionar, em geral, uma descrição satisfatória de dados experimentais cuja variância é proporcional à média;
- poder ser deduzido teoricamente de princípios elementares com um número mínimo de restrições;
- se se registarem ocorrências independentemente e aleatoriamente no tempo, com taxa média de ocorrência constante, o modelo determina o número de ocorrências, num intervalo de tempo especificado.

O número de ocorrências do processo durante intervalos de tempo não-sobrepostos corresponde a variáveis aleatórias independentes.

2.2.4 Regressão Gama

Admitindo que as respostas são variáveis aleatórias $Y_i \sim Ga(v, \frac{v}{\mu_i})$ independentes, com valor médio $\mu_i = exp(z_i^T \beta)$, este é um modelo da classe dos Modelos Lineares Generalizados adequado para variável resposta contínua no caso em que esta é estritamente positiva e apresenta enviesamento/assimetria à direita.

O valor esperado é dado por: $E[Y_i] = \mu_i$, onde $z_i = (1, x_{i1}, ..., x_{ip})^T$.

i.) A distribuição Gama pertence à família exponencial, definida em 2.17.

No caso desta distribuição tem-se:

- $f(y; \mu, \alpha) = \frac{1}{y\Gamma(\alpha)} \left(\frac{\alpha y}{\mu}\right)^{\alpha} exp\left(-\frac{\alpha y}{\mu}\right)$
- $\theta = -\frac{1}{\mu}$
- $\phi = \frac{1}{\alpha}$
- $a(\phi) = \phi = \frac{1}{\alpha}$
- $b(\theta) = -log(-\theta) = log(\mu)$
- $c(y, \phi) = \alpha log(\alpha y) - log(y\Gamma(\alpha))$

ii.) A função de ligação para uma componente aleatória com distribuição Gama é a função logarítmica.

Note-se que a função de ligação considerada não é a função de ligação canónica. A função de ligação canónica obtém-se quando $z_i^T \beta = \theta_i$, o que neste caso corresponde a ter $-\frac{1}{\mu_i} = z_i^T \beta$, portanto a função de ligação canónica é a função recíproca.

2.3 Modelos para Excesso de Zeros

Para modelar dados com grande número de zeros é possível usar os modelos com excesso de zeros (*Zero Inflated Models*) e os modelos de duas partes, também conhecidos por modelos de barreira (*Hurdle Models*). Estes modelos podem ser usados com várias distribuições.

Neste trabalho serão usados os *Hurdle Models* com as distribuições de Poisson e binomial negativa.

Modelos de Barreira ou Modelos *Hurdle*

Em muitas situações, os dados de contagem contêm um número elevado de zeros.

Esta frequência elevada de valores nulos pode levar a que os dados não sejam devidamente ajustados por uma distribuição de Poisson ou binomial negativa, distribuições normalmente utilizadas para dados desta natureza. Para ultrapassar esta limitação são necessários modelos com capacidade para, simultaneamente, acomodar o excesso de zeros e descrever adequadamente as contagens.

Os modelos *Hurdle*, conhecidos por modelos de barreira, servem a este propósito. Estes modelos são compostos por duas partes:

- Os dados são considerados como zeros *vs* não-zeros e um modelo binomial (regressão logística) é usado para modelar a probabilidade de um valor não nulo ser observado.
- As contagens positivas (diferentes de zero) são modeladas através de uma distribuição de Poisson truncada ou binomial negativa truncada.

As contagens positivas são interpretadas como tendo ultrapassado a barreira dos zeros. A barreira não tem de se encontrar necessariamente no valor zero, podendo estar em qualquer valor de acordo com o problema em análise.

2.3.1 *Hurdle* Poisson

A função de probabilidade do modelo de regressão de Poisson com barreira em zero é dada por:

$$P[Y_i = y_i] = \begin{cases} 1 - \pi_i, & y_i = 0 \\ \frac{\pi_i e^{-\lambda_i} \lambda_i^{y_i}}{(1-e^{-\lambda_i}) y_i!}, & y_i > 0 \end{cases} \tag{2.26}$$

onde π_i é a probabilidade de se observar uma contagem não nula e $\lambda_i = e^{\beta_0 + \beta_{i1} x_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}$.

O processo de contagem de Poisson aqui considerado, exclui a probabilidade de existirem zeros, portanto estamos perante uma distribuição de Poisson truncada. A probabilidade de se observar uma contagem diferente de zero é igual à probabilidade de não existir um zero, multiplicada pela função de probabilidade de uma distribuição de Poisson truncada.

Para este modelo, as expressões para o valor médio e variância são as seguintes, respetivamente:

$$E[Y_i] = \frac{\pi_i}{1 - e^{-\lambda_i}} \times \lambda_i; \tag{2.27}$$

$$V[Y_i] = \frac{\pi_i}{1 - e^{-\lambda_i}} \times \left(\lambda_i + \lambda_i^2\right) - \left(\frac{\pi_i}{1 - e^{-\lambda_i}} \times \lambda_i\right)^2. \tag{2.28}$$

2.3.2 *Hurdle* Binomial Negativo

A função de probabilidade do modelo de regressão binomial negativa com barreira é dada por:

$$P[Y_i = y_i] = \begin{cases} 1 - \pi_i, & y_i = 0 \\ \dfrac{\pi_i \left[\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i} \right]}{1 - \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}}, & y_i > 0 \end{cases} \tag{2.29}$$

onde π_i é a probabilidade de se observar uma contagem não nula e $\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}$.

Tal como no modelo anterior, o processo de contagem é independente do processo das contagens nulas, logo a probabilidade de se observar uma contagem diferente de zero é igual à probabilidade de não existir um zero, multiplicada pela função de probabilidade de uma distribuição Binomial Negativa truncada.

Para este modelo, as expressões para o valor esperado e para a variância são as seguintes, respetivamente:

$$E[Y_i] = \frac{\pi_i}{1 - \left(\frac{\frac{1}{\alpha}}{\mu_i + \frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}} \times \mu_i; \tag{2.30}$$

$$V[Y_i] = \frac{\pi_i}{1 - \left(\frac{\frac{1}{\alpha}}{\mu_i + \frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}} \times \left(\mu_i + \mu_i^2(1+\alpha)\right) - \left(\frac{\pi_i}{1 - \left(\frac{\frac{1}{\alpha}}{\mu_i + \frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}} \times \mu_i \right)^2. \tag{2.31}$$

2.4 Ajustamento do Modelo

2.4.1 Estimação de $\boldsymbol{\beta}$

Como já foi referido anteriormente, o vetor de parâmetros $\boldsymbol{\beta}$ é estimado através do Método da Máxima Verosimilhança, método que permite a aplicação de testes de hipóteses sobre os parâmetros do modelo e aferir da qualidade do ajustamento, uma vez que estes estimadores são assintoticamente normais.

Considerando o Modelo Linear Generalizado, tem-se:

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi, \omega_i) = \prod_{i=1}^{n} exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, \omega_i) \right\}. \tag{2.32}$$

A log-verosimilhança é dada por:

$$ln[L(\beta)] = l(\beta) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, \omega_i) \right] = \sum_{i=1}^{n} l_i(\beta), \tag{2.33}$$

onde $l_i(\beta)$ representa a contribuição de cada observação y_i para a verosimilhança.

A Estimativa de Máxima Verosimilhança para os elementos de $\boldsymbol{\beta}$ obtém-se através da equação de verosimilhança:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_j} = 0, \tag{2.34}$$

onde $j = 1, ..., p$ e $\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\partial l_i(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\beta)}{\partial \beta_j}$.

Relembrando 2.18[1] e 2.19[2], vem:

$$\begin{cases} \frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \\ \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{var(Y_i)}{a(\phi)} \\ \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} = [b''(\mu_i)]^{-1} = \frac{a(\phi)}{var(Y_i)} \\ \frac{\partial \eta_i}{\partial \beta_i} = x_{ij} \end{cases} . \tag{2.35}$$

Desta forma, tem-se:

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \tag{2.36}$$

As equações de verosimilhança para $\boldsymbol{\beta}$ são obtidas por:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \tag{2.37}$$

onde $j = 1, ..., p$.

Assumindo a existência e unicidade das Estimativas de Máxima Verosimilhança, as equações de verosimilhança não têm, em geral, uma solução analítica.

Nas situações mais usuais, que envolvem muitos dados, é necessário recorrer a técnicas numéricas iterativas que permitem obter estimativas para $\boldsymbol{\beta}$, tais como, por exemplo, o método dos *scores* de Fisher, método utilizado neste projeto.

A função *score* é o vetor p-dimensional:

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta), \tag{2.38}$$

onde $s_i(\beta)$ é o vetor de componentes $\frac{\partial l_i(\beta)}{\partial \beta_j}$ definido em 2.34.

O elemento genérico de ordem j da função *score* é:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}. \tag{2.39}$$

O método dos *scores* de Fisher parte de uma estimativa inicial $\widehat{\boldsymbol{\beta}}^{(0)}$, e através da relação definida de seguida, calcula as sucessivas iteradas:

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + [I(\widehat{\beta}^{(k)})]^{-1} s(\widehat{\beta}^{(k)}), \tag{2.40}$$

em que, $I(\beta) = E\left[-\frac{\partial s(\beta)}{\partial \beta}\right]$ é a matriz de covariância da função *score*, conhecida como a matriz de informação de Fisher.

[1] $b'(\theta_i) = \mu_i$.

[2] $var(Y_i) = a(\phi) b''(\theta_i)$.

2.4.2 Estimação do parâmetro de dispersão ϕ

Em algumas distribuições, tal como a distribuição de Poisson, o parâmetro ϕ é conhecido, no entanto, em geral, este parâmetro não é conhecido *a priori*, e tem de ser estimado.

A estimação deste parâmetro não é necessária para a obtenção do vetor $\boldsymbol{\beta}$, mas sim para determinar algumas estatísticas.

Tal como para o vetor de parâmetros $\boldsymbol{\beta}$, o parâmetro de dispersão pode ser estimado pelo Método da Máxima Verosimilhança, que no entanto não permite a obtenção de uma fórmula explícita para ϕ, e pode ser uma alternativa mais lenta.
Em geral, utilizam-se os seguintes estimadores para ϕ:

- O estimador de Momentos ou a Estatística χ^2 de Pearson, definido como:

$$\widehat{\phi} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{V(\mu_i)} \tag{2.41}$$

- O estimador baseado na Estatística de Pearson Generalizada, um estimador consistente e assintoticamente centrado:

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \tag{2.42}$$

- O estimador "*Deviance*", definido como:

$$\widehat{\phi} = \frac{D}{n-p}, \tag{2.43}$$

  sendo D a "*Deviance*", que em termos gerais, é uma medida de quanto os valores ajustados diferem das observações[3].

2.5 Seleção e Validação do Modelo

Ao estudar um problema onde existem algumas variáveis explicativas a considerar, uma das análises a efetuar é decidir qual o modelo mais adequado, ou seja, o modelo que permite uma boa interpretação do problema em estudo e que se ajusta bem aos dados.

Diagnóstico em Modelos Lineares Generalizados

As técnicas utilizadas para o diagnóstico nos modelos lineares generalizados são semelhantes às técnicas utilizadas no modelo linear, com algumas adaptações, devido à estrutura destes modelos. No entanto, dado que os erros não têm estrutura normal, a utilização destas técnicas está menos facilitada.

2.5.1 Resíduos

A análise de resíduos é bastante útil na avaliação da qualidade de ajustamento do modelo no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear, como também serve para ajudar a identificar observações mal ajustadas, ou seja, que não são bem explicadas pelo modelo.

[3] A *Deviance* não será estudada neste projeto.

Os resíduos ordinários medem discrepâncias entre os valores observados da variável resposta e os valores ajustados:

$$e_i = y_i - \widehat{y_i}, \tag{2.44}$$

relativamente à i-ésima observação.

Se o modelo estiver corretamente especificado, os resíduos ordinários são variáveis aleatórias com média nula e variância $\sigma^2(1-h_i)$, onde h_i representa o i-ésimo termo da diagonal da matriz de projeção generalizada $\mathbf{H}$ (também denominada por *hat-matrix*), que pode ser calculado através da expressão: $h_i = \mathbf{x}(i)^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}(i)$.

Para uma análise adequada dos resíduos, é conveniente que estes sejam padronizados e reduzidos, ou seja, que tenham variância constante unitária e, preferencialmente, que sejam aproximadamente normalmente distribuídos.
Podem também ser utilizados para testar a presença de *outliers*.

Pode definir-se o resíduo de Pearson por:

$$R_i{}^P = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{var(Y_i)}}}. \tag{2.45}$$

Os resíduos de Pearson padronizados são dados por:

$$\frac{R_i{}^P}{1-h_{ii}}. \tag{2.46}$$

A desvantagem destes resíduos é que a sua distribuição é, por norma, bastante assimétrica para modelos não normais.

2.5.2 Medida de Alavancagem (*Leverage Measure*)

Observações que se encontram relativamente longe do centro dos valores amostrais têm potencialmente uma maior influência sobre os valores dos coeficientes de regressão, sendo conhecidos como pontos de alavanca.
As medidas mais comuns de alavancagem são os h_i (*hat-values*). Quando se observam vários h_i com valor elevado, pode ser problemático, uma vez que observações com alto poder de alavancagem influenciam muito os resultados.

A definição geral de *leverage* da j-ésima observação no valor predito da i-ésima resposta é a amplitude da derivada do i-ésimo valor predito $\widehat{\mu_i}$ em relação ao valor observado da j-ésima resposta, y_j. No caso dos Modelos Lineares Generalizados, esta medida é dada pelo ij-ésimo elemento da matriz $\mathbf{H}$ generalizada[4].

Espera-se que as observações distantes do espaço formado pelas variáveis explicativas apresentem valores apreciáveis de h_i. Como $\mathbf{H}$ é matriz de projeção e h_i encontra-se no intervalo $[0,1]$, sugere-se que h seja superior a $\frac{2p}{n}$ para indicar os pontos de alavancagem.

2.5.3 Medidas de Influência (*Influence Measures*)

Uma observação que é simultaneamente um *outlier* e tem poder de alavancagem (*high leverage*) tem um efeito acentuado sobre os coeficientes de regressão no sentido que se a observação

[4] $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$, onde $\mathbf{W}$ é a matriz diagonal de ordem n.

for removida, os coeficientes se alteram significativamente.

A medida de influência mais comum é a distância de Cook. É a forma de encontrar observações com *leverage* elevado e mal ajustadas. Esta distância mede o efeito de retirar uma observação do conjunto de observações usado para fazer o ajustamento do modelo, uma vez que observações com resíduos e/ou *leverage* elevados podem distorcer o resultado e a precisão da regressão. Às observações com uma distância de Cook elevada requere-se uma análise mais aprofundada.

A distância de Cook é traduzida pela seguinte expressão:

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \times SQE}, \tag{2.47}$$

onde SQE é a soma dos quadrados dos resíduos e p é o número de parâmetros ajustados no modelo.

Neste trabalho serão considerados os modelos Poisson, *Hurdle* Poisson e *Hurdle* binomial negativo para a frequência, e os modelos de regressão linear múltipla, regressão linear múltipla com resposta logaritmizada e Gama para o custo por sinistro.

Na estimação dos modelos para cada variável resposta (a frequência e o custo por sinistro) foram considerados os mesmos conjuntos de covariáveis sobre uma parte da base de dados (conjunto de treino), tendo os modelos sido comparados entre si com base nos erros de previsão resultantes da sua aplicação ao remanescente da base de dados, o conjunto de teste.

Capítulo 3

Análise Exploratória dos Dados

Inicialmente, toda a base de dados é extraída do *Microsoft Access* para o *Microsoft Excel* e os dados são filtrados por registos pertencentes a particulares e pequenas empresas, contendo apenas colunas em que as variáveis caracterizam o tomador de seguro e o respetivo veículo. Para fazer toda a análise, o *software* estatístico utilizado foi o *R*.

3.1 Variáveis em Estudo

Foram selecionadas as variáveis explicativas a usar no modelo, com base nas características do veículo e do tomador de seguro. Neste projeto foi considerado que toda essa informação seria relevante, uma vez que todas elas poderiam ter impacto no cálculo da tarifa por serem, por norma, as variáveis habitualmente consideradas na construção das mesmas.

No modelo para a frequência, a variável resposta é o número de sinistros (variável quantitativa discreta) e no modelo para o custo por sinistro, a variável resposta é o custo médio por sinistro associado a cada apólice (variável quantitativa contínua). Há também uma variável que vai funcionar como um termo *offset*, que será a exposição ao risco (variável numérica). Este termo é imprescindível para estudar o perfil de risco do tomador de seguro, pois dá a indicação de tempos de risco diferentes para cada perfil. O que se pretende é a obtenção de um modelo para o número esperado de sinistros por unidade de tempo (ano), e só assim o modelo consegue atribuir os fatores mais precisos a cada classe de cada variável explicativa.

As variáveis explicativas são as seguintes:

- Distrito (variável categórica);
- Concelho (variável categórica);
- Subscritor (variável categórica);
- Idade do Condutor (variável numérica);
- Idade de Carta (variável numérica);
- Idade do Veículo (variável numérica);
- Marca (variável categórica);
- Escalão de Cilindrada (variável categórica);
- Categoria Agregada (variável categórica);
- Tipo de Uso (variável categórica).

Nesta fase, numa primeira análise, a base de dados que vai permitir efetuar um estudo mais consistente é a base de dados com todos os registos acerca do veículo e do tomador de seguro devidamente preenchidos. No período em estudo (dois anos e meio), foram registados acidentes em, aproximadamente, 9% das apólices.

Nas variáveis categóricas, devido à pouca representatividade de algumas das classes é necessário proceder ao agrupamento destas numa só.

Neste projeto, o *clustering* será utilizado para reduzir o número de classes de algumas variáveis.

Nas variáveis em que será necessário agrupar por classes, seja por método hierárquico ou por método não-hierárquico, fará mais sentido formar *clusters* por frequência média de sinistralidade e custo médio por sinistro em cada variável a estudar, uma vez que são essas duas variáveis resposta que no modelo encontrado vão permitir o cálculo do prémio de risco.

Sempre que seja necessário recorrer ao *clustering*, a distância utilizada para o agrupamento em *clusters* será a distância euclidiana[1].

Apresenta-se de seguida pequenas análises estatísticas a cada variável, para que se torne possível conhecer melhor a base de dados.

3.2 Tratamento de Dados

3.2.1 Distrito

Uma vez que existem 21 distritos, distribuídos por Portugal Continental, arquipélago da Madeira, Porto Santo (arquipélago da Madeira) e arquipélago dos Açores, há interesse em agregá-los de forma a que fiquem no mesmo grupo, distritos que possam ser considerados semelhantes entre si, no que respeita a custo médio por sinistro e frequência média de sinistralidade em cada distrito. Este procedimento visa simplificar o posterior trabalho de construção do modelo, diminuindo o número de parâmetros a estimar.

Recorrendo ao *clustering*, obtém-se um dendrograma com os distritos agregados em *clusters*. Os distritos pertencentes ao mesmo *cluster* são os mais parecidos entre si, em termos de custo médio por sinistro e frequência média de sinistralidade em cada distrito. Existe a necessidade de padronizar essas variáveis, para que as mesmas sejam consideradas com o mesmo peso.

O número de *clusters* escolhido foi 6, como é visível no dendrograma da figura 3.1. A escolha do número de classes foi feita após observar o dendrograma e perceber de que forma os distritos estavam agregados.

Como também se percebe, Porto Santo ficou numa classe diferente do resto do arquipélago da Madeira. Desta forma, Porto Santo junta-se à Madeira, uma vez que o número de apólices em Porto Santo é muito pouco significativo, como é visível no gráfico da figura 3.4. Assim, o número de *clusters* é reduzido a cinco, uma vez que Porto Santo se encontra sozinho numa classe.

É de salientar que a frequência média de sinistralidade em cada distrito é traduzida pelo quociente entre o número de sinistros nesse distrito e o número de apólices existentes nesse mesmo distrito e que o custo médio por sinistro é traduzido pelo quociente entre o custo total com sinistros nesse distrito e o número de sinistros nesse mesmo distrito.

[1] Apresentada no suporte teórico em 2.2.

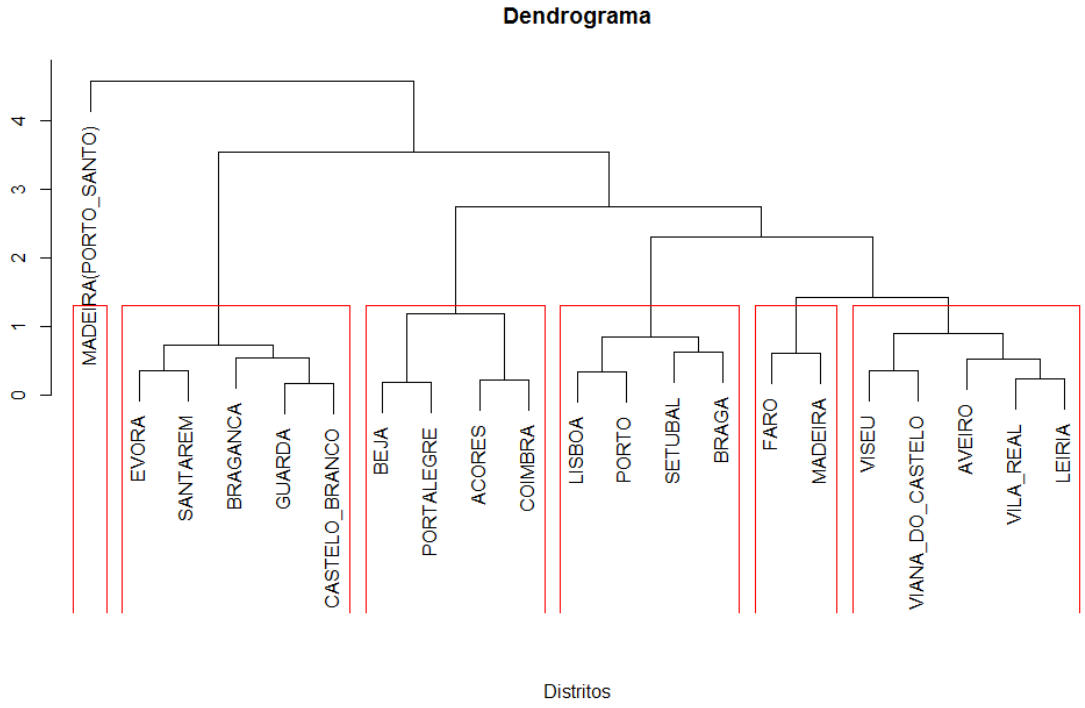

Figura 3.1: Agregação de distritos por custo médio por sinistro e frequência média de sinistralidade em cada distrito

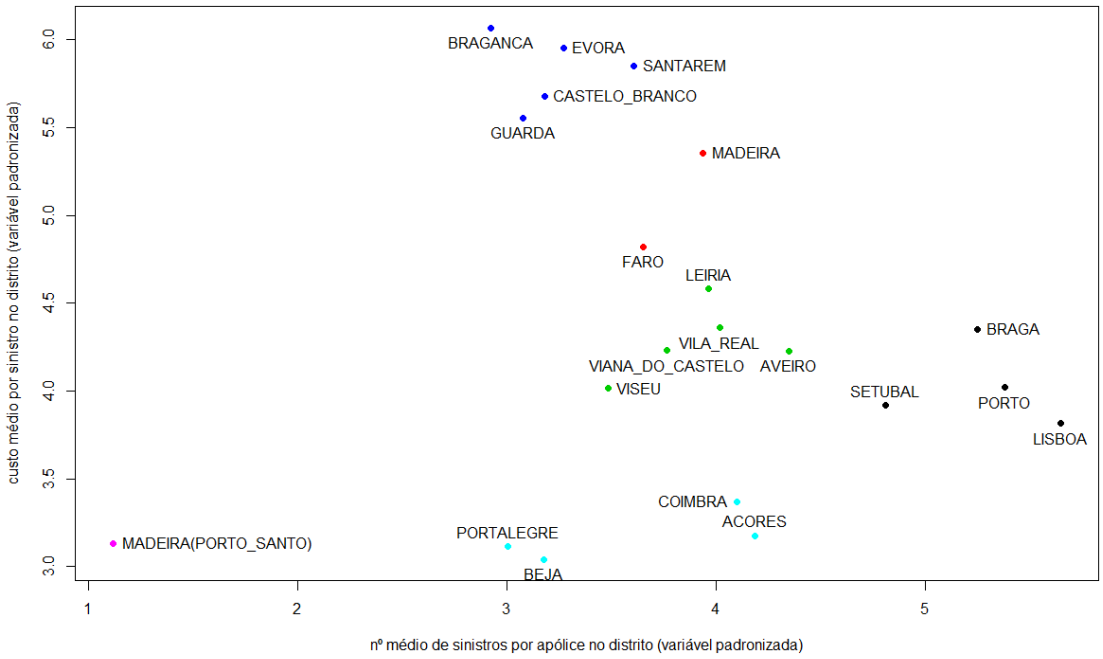

Figura 3.2: *Clustering* - Agregação de distritos por custo médio por sinistro e frequência média de sinistralidade em cada distrito

Classe 1	Braga Lisboa Porto Setúbal
Classe 2	Faro Madeira Porto Santo
Classe 3	Aveiro Leiria Viana do Castelo Vila Real Viseu
Classe 4	Bragança Castelo Branco Évora Guarda Santarém
Classe 5	Açores Beja Coimbra Portalegre

Tabela 3.1: Agregação final de distritos

Após terminado o *clustering*, foi efetuada uma análise às diferenças entre os *clusters*. Foi utilizada a distribuição binomial para a comparação da variável resposta "teve sinistro" entre eles e a regressão linear para a comparação da variável resposta "custo com sinistros". Na tabela da figura 3.3 é possível concluir que os grupos 3 e 5 não apresentam diferenças significativas entre si em relação à frequência de sinistralidade, mas que na maioria das comparações, os grupos apresentam diferenças significativas ($\approx 0\%$) entre si. Em relação aos custos com sinistros, pode-se concluir que os grupos 1, 2 e 3 não apresentam diferenças muito significativas entre si.

Distrito

Cluster	1	2	3	4	5
1		***/*	***/	***/***	***/*
2	***/*		***/	***/.	**/**
3	***/	***/		***/***	/*
4	***/***	***/.	***/***		***/***
5	***/*	**/**	/*	***/***	

Níveis de significância
0% < *** ≤ 0,1%
0,1% < ** ≤ 1%
1% < * ≤ 5%
5% < . ≤ 10%
10% < ≤ 100%

Figura 3.3: Análise às diferenças de frequência/custos de sinistralidade entre os *clusters* de distritos

É possível verificar, pelos gráficos das figura 3.2, 3.4 e 3.5 que os distritos com custos médios mais baixos (Açores, Beja, Coimbra e Portalegre) pertencem todos ao mesmo *cluster*. A figura 3.5 por ter os *boxplots* desenhados à escala logarítmica, dá uma melhor perceção dos intervalos de custos do que no gráfico da figura 3.4. É também possível verificar, através destes gráficos, a razão pela qual a agregação aparece distribuída por aqueles *clusters*, mesmo tendo sido consideradas duas variáveis: custo médio por sinistro e frequência média de sinistralidade em cada distrito.

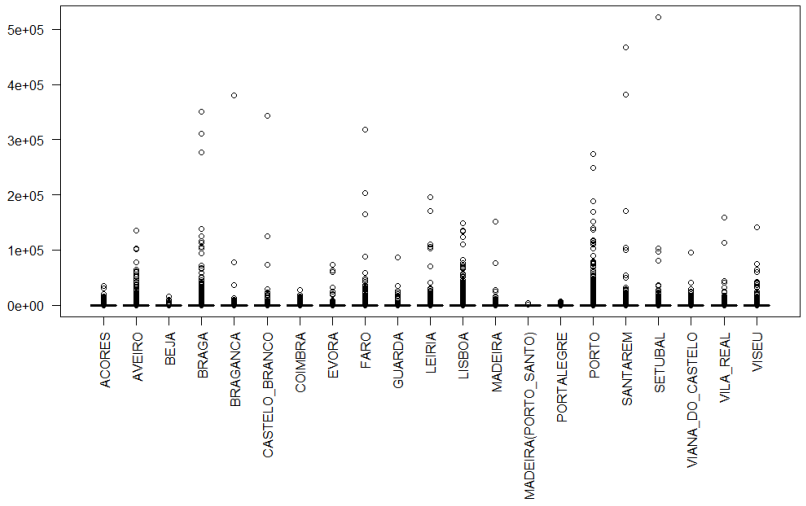

Figura 3.4: Representação do custo com sinistros por distrito

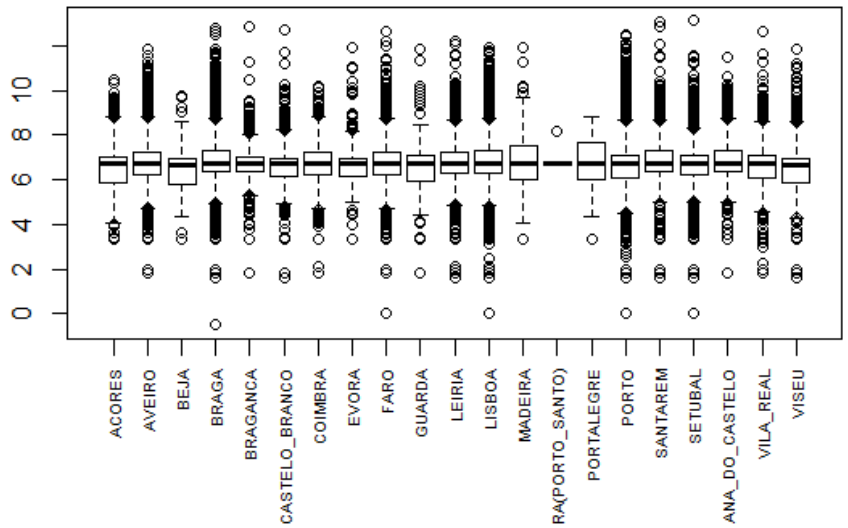

Figura 3.5: Representação do custo com sinistros por distrito na escala logarítmica

Mais de 55% das apólices em estudo são apólices registadas no Porto, Lisboa e Braga. Estes são os distritos com mais apólices registadas nesta carteira. É natural que nestes distritos os custos totais com sinistros sejam bastante elevados, porque são realmente os distritos com mais apólices subscritas, mas ainda assim, os custos médios por sinistro nestas cidades, não são de todo os mais elevados, como é possível observar nos mapas das figuras 3.6 e 3.7. Estes mapas foram construídos como curiosidade, de forma a que nos fosse possível verificar quais os distritos que apresentam custo médio por sinistro mais e menos elevado, assim como a frequência média de sinistralidade mais e menos elevada[2].

As zonas a vermelho são aquelas cujo custo médio por sinistro é mais elevado ou aquelas em que há mais sinistros. De seguida segue-se o cor-de-laranja, depois o amarelo, o verde e por fim o azul. Tal informação está visível nas legendas dos mapas das figuras 3.6 e 3.7.

Como é visto através do *clustering*, o arquipélago da Madeira[3] pertence ao grupo de Faro e o arquipélago dos Açores[4] pertence ao grupo de Beja, Coimbra e Portalegre.

Analisando os mapas, facilmente se percebe que os distritos com custo médio por sinistro mais elevado não são aqueles onde existe maior frequência de sinistralidade.

O grupo de distritos que apresenta maior frequência de sinistralidade é o grupo de Braga, Lisboa, Porto e Setúbal, o que é natural, pois é onde se encontram mais de metade das apólices subscritas na Companhia. Estes distritos não são, claramente, os que apresentam maior custo por sinistro, o que também é natural, pois os sinistros que ocorrem, normalmente, acabam por ser sinistros com pouca gravidade, devido ao enorme tráfego nestas cidades.

Com base nesta carteira, 69% dos sinistros ocorridos no período em estudo, pertence ao grupo 1[5] (Braga, Lisboa, Porto e Setúbal) e o custo médio por sinistro nessa classe é o segundo mais baixo.

[2]Poderão ser consultados, em anexo, os mapas relativos aos arquipélagos da Madeira e Açores nas figuras 6.2 e 6.3.

[3]Os mapas podem ser consultados em anexo (figuras 6.2 e 6.3).

[4]Os mapas podem ser consultados em anexo (figuras 6.6 e 6.7).

[5]A tabela 3.1 corresponde às agregações por distrito.

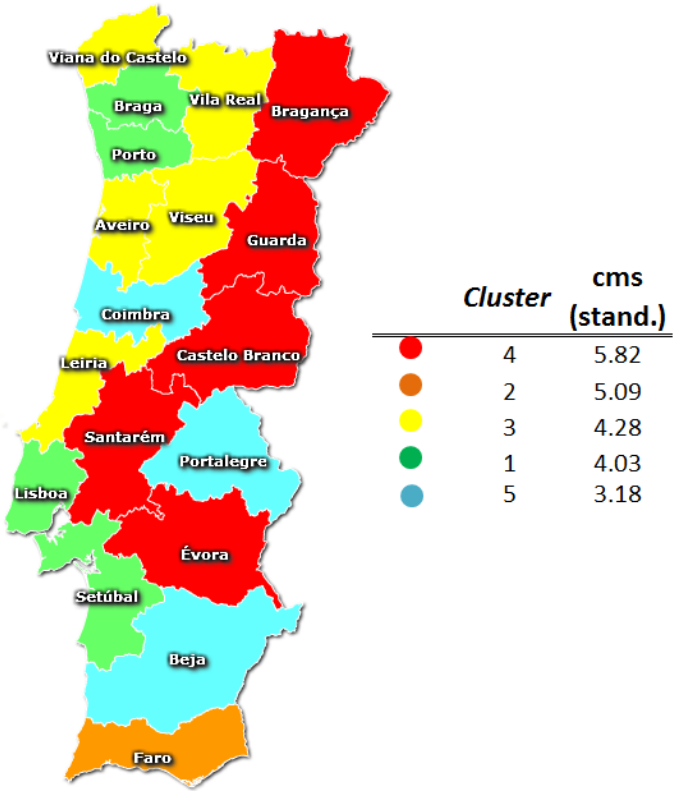

Figura 3.6: Custo médio por sinistro (variável standardizada)

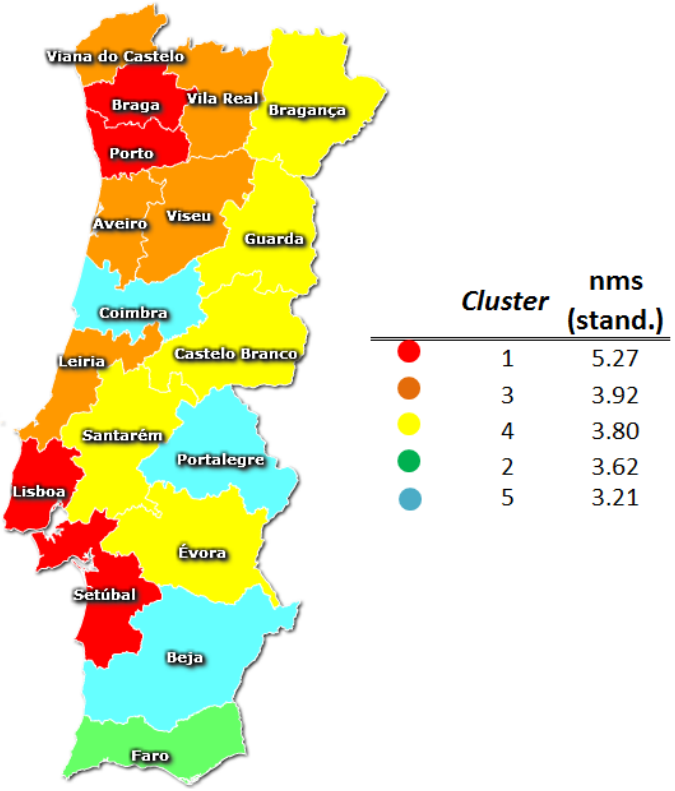

Figura 3.7: Frequência média de sinistralidade (variável standardizada)

3.2.2 Concelho

O número de concelhos em carteira é de 307, daí ser estritamente necessário agregá-los.

Os concelhos são agregados pelas suas semelhanças, sendo o custo médio por sinistro e a frequência média de sinistralidade em cada concelho os critérios a considerar. Foram agregados pelo método das k-médias (método não-hierárquico). Neste caso o k escolhido foi 6 de forma a obter 6 classes. Os elementos cujos critérios em que a distância aos elementos de cada classe é menor do que a distância ao ponto médio de qualquer outra classe, ficam agregados aos elementos dessa mesma classe, isto é, cada elemento com um certo custo médio por sinistro e frequência média de sinistralidade no concelho, fica na classe onde está mais próximo do centro (média da classe).

Na figura 6.1, em anexo, é possível verificar o dendrograma, dividido em 6 classes, pelo método hierárquico. Verifica-se que existe um grupo com 233 concelhos, outro com 5, outro com 18, outro com 44, outro com 6 e outro com 1.

Chegou-se à conclusão que por mais que se dividam os concelhos em mais de 6 *clusters*, Vila Nova de Foz Côa acaba por continuar sempre sozinha num só *cluster*. O custo médio por sinistro nesse concelho é bastante elevado, porque em 159 subscrições, ocorreram 2 sinistros acima de 11000€. Claramente, Vila Nova de Foz Côa é um *outlier*.

Também pelo facto de se ter 233 concelhos agregados numa classe e ter, comparativamente a esta classe, muito poucos concelhos nas outras, optou-se por considerar daqui em diante, o resultado obtido pelo método não-hierárquico, k-médias, uma vez que os concelhos se encontram mais uniformemente distribuídos entre *clusters*[6].

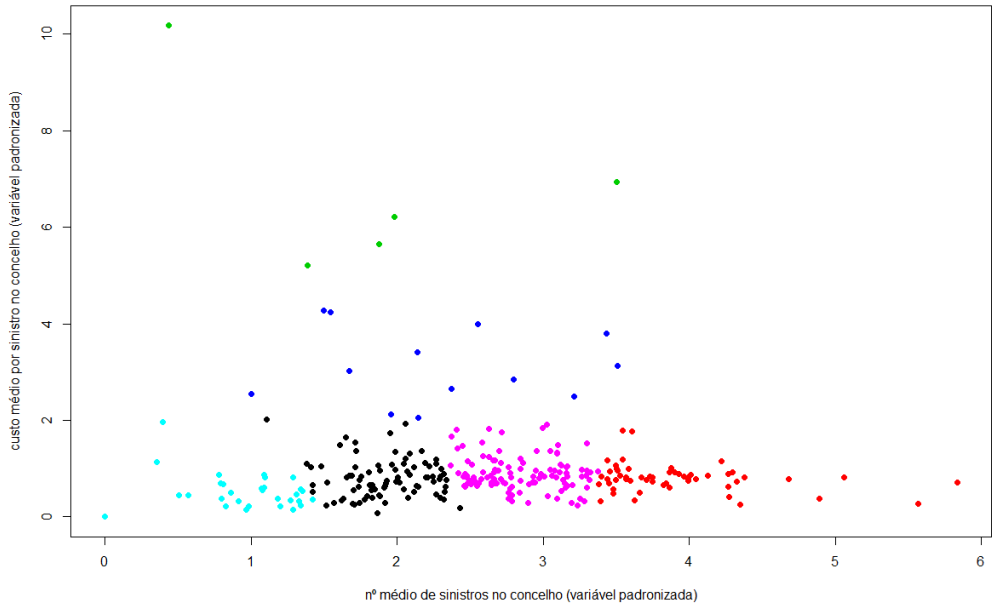

Figura 3.8: *Clustering* - Agregação de concelhos por custo médio por sinistro e frequência média de sinistralidade em cada concelho (método k-médias)

Após terminado o *clustering*, foi efetuada uma análise às diferenças entre os *clusters*. Foi utilizada a distribuição binomial para a comparação da variável resposta "teve sinistro" entre

[6] A tabela final de agregação por concelho pode ser consultada em anexo (tabela 6.1).

eles e a regressão linear para a comparação da variável resposta "custo com sinistros". Na tabela da figura 3.9 é possível concluir que os grupos 1, 4 e 5 não apresentam diferenças muito significativas entre si em relação à frequência de sinistralidade, mas que na maioria das comparações, os grupos apresentam diferenças significativas ($\approx 0\%$) entre si. Em relação aos custos com sinistros, pode-se concluir que os grupos 2, 3, 4 e 6 não apresentam diferenças muito significativas entre si.

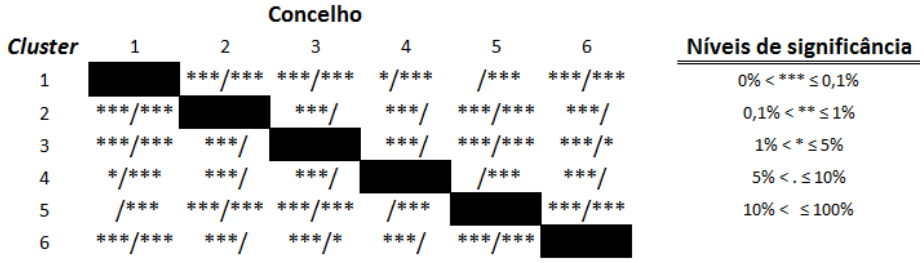

Figura 3.9: Análise às diferenças de frequência/custos de sinistralidade entre os *clusters* de concelhos

Os mapas de Portugal para os concelhos são construídos, seguindo a mesma metodologia dos distritos.

Observando os mapas, facilmente se percebe que os concelhos com custos médios mais elevados com sinistros são precisamente os concelhos onde ocorrem menos sinistros. Estes concelhos pertencem ao grupo 5 (Mesão Frio, Portel, São Brás de Alportel, Sátão e Vila Nova de Foz Côa). Mais uma vez, isto deve-se ao facto destes concelhos terem muito poucas apólices subscritas na Companhia e os poucos sinistros que ocorreram terem apresentado custos bastante elevados.

De salientar que, mais de metade da sinistralidade desta carteira pertence ao grupo a azul da figura 3.10 que é o grupo com menos custos associados a sinistros. O grupo a vermelho, com mais custos associados a sinistros, representa 12 vezes mais de custos em relação ao *cluster* preenchido a azul.

Assim, justifica-se ponderar a construção dos modelos separando os casos mais extremos.

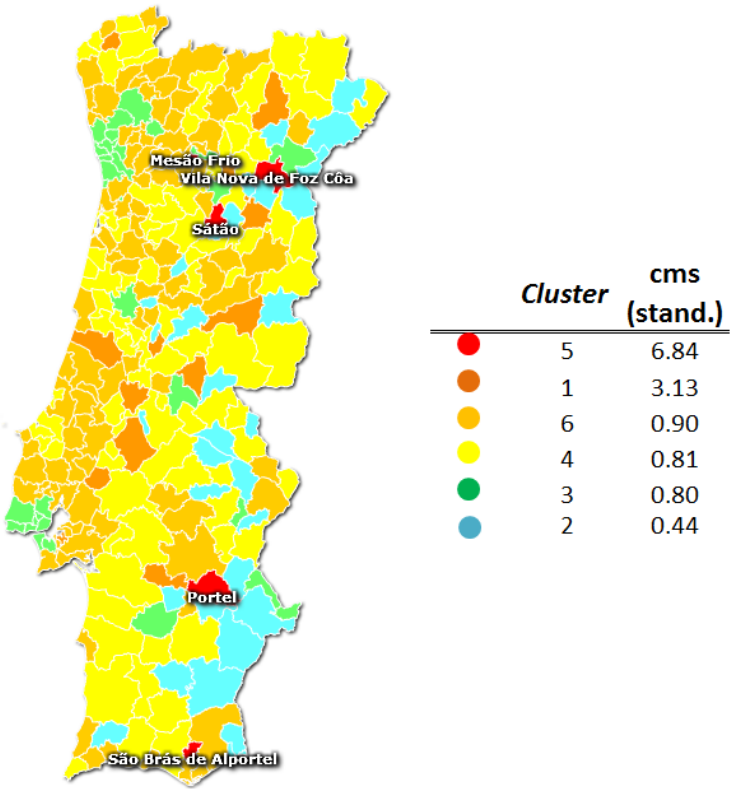

Figura 3.10: Custo médio por sinistro (variável standardizada)

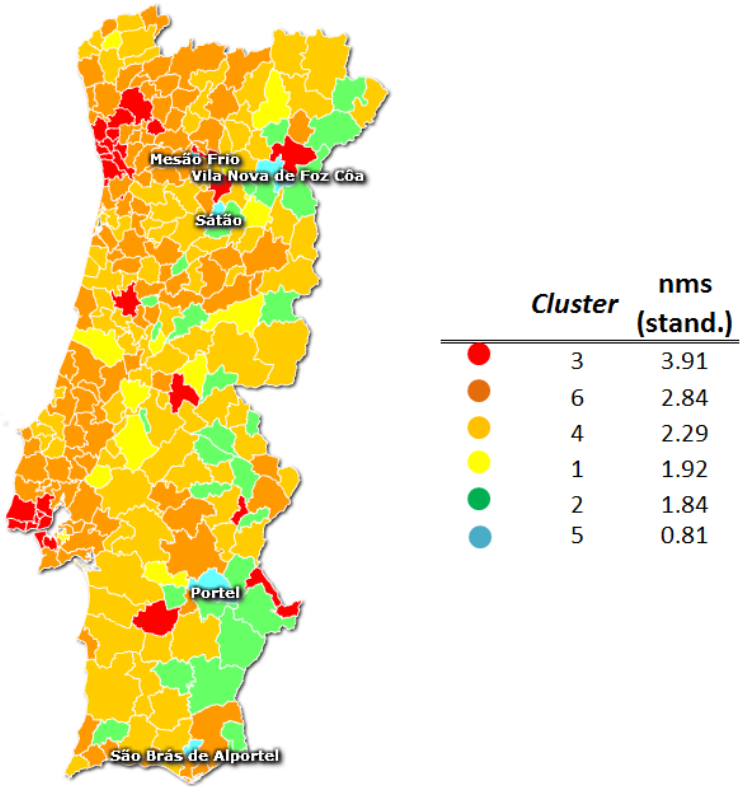

Figura 3.11: Frequência média de sinistralidade (variável standardizada)

3.2.3 Subscritor

Trata-se de uma variável que indica o género do tomador de seguro ou se se trata de uma pequena empresa.

Atribuiu-se o nome de subscritor, pois era a forma de manter a informação sobre género e empresa numa mesma variável, uma vez que era esta a organização inicial dos dados.

Subscritor	Freq. Rel.
Masculino	61,55%
Feminino	23,04%
Empresa	15,41%

Tabela 3.2: Frequência relativa de registos nos subscritores

É possível verificar que a maioria dos segurados são do sexo masculino, quase o triplo do número de subscritores do sexo feminino e o quádruplo do número de subscritores "empresas".

Subscritor	Prop. de apólices	Prop. de sinistros	Prop.(sinistro|subscritor)
Masculino	61,55%	58,16%	8,41%
Feminino	23,04%	24,18%	9,34%
Empresa	15,41%	17,66%	10,19%

Tabela 3.3: Tabela da distribuição de apólices e sinistros por tipo de subscritor e da sinistralidade condicionada no subscritor, ordenada por proporção de apólices

Apesar da maioria de apólices pertencer ao sexo masculino, é interessante verificar que não são estes que, em termos relativos, têm mais sinistros, mas sim os que têm menos.

3.2.4 Idade do Condutor

Esta variável indica a idade que o tomador de seguro tem. A média da idade do condutor é 47.47, ou seja, aproximadamente 47 anos. Existem condutores (nesta carteira) com 18 anos de idade no mínimo e com 95 anos, no máximo. A idade do condutor que ocorre mais vezes é 38 anos.

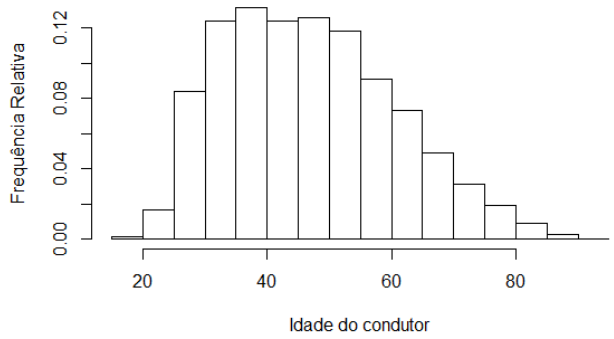

Figura 3.12: Frequência relativa das idades do condutor

Observando o gráfico da figura 3.12, verifica-se que a maioria dos segurados desta carteira está entre os 30 e os 55 anos, inclusive, isto é, 64% das apólices subscritas pertencem a este intervalo de idades do condutor.

3.2.5 Idade da Carta

Esta variável dá a indicação do número de anos de carta do tomador de seguro. A média da idade de carta é 22.16, ou seja, aproximadamente 22 anos. Existem tomadores de seguro com 0 anos de idade de carta, ou seja, com carta há menos de 1 ano e existem tomadores de seguro com 77 anos de idade de carta, no máximo. A idade de carta que ocorre mais vezes é 20 anos.

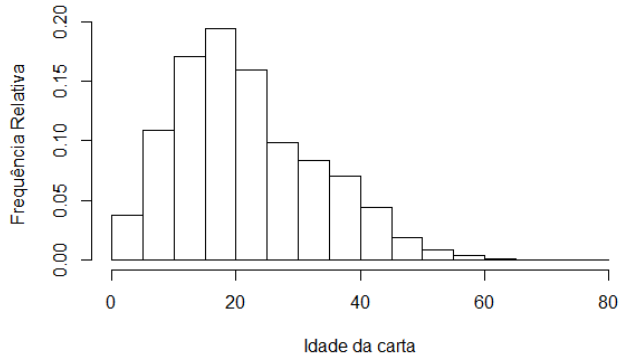

Figura 3.13: Frequência relativa da idade da carta

Observando o gráfico da figura 3.13, verifica-se que a maioria dos segurados desta carteira tem entre 10 e 25 anos de carta, inclusive, isto é, 55% das apólices subscritas pertencem a este intervalo de idades da carta de condução.

3.2.6 Idade do Veículo

Esta variável indica o número de anos que o veículo tem. A média da idade do veículo é 12.95, ou seja, aproximadamente 13 anos. Nesta carteira, existem veículos com 0 anos de idade no mínimo e com 105 anos, no máximo. A idade do veículo que ocorre mais vezes é 13 anos.

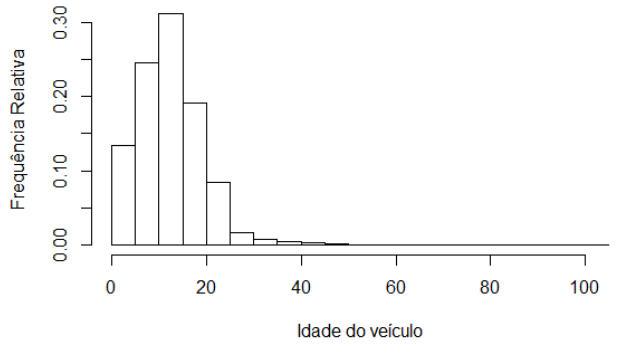

Figura 3.14: Frequência relativa das idades do veículo

Observando o gráfico da figura 3.14, verifica-se que o maior número de veículos desta carteira está entre os 5 e os 20 anos, inclusive, isto é, 79% das apólices subscritas pertencem a este intervalo de idades do veículo seguro.

3.2.7 Marca

Esta variável indica a marca do veículo e é composta por 29 marcas diferentes, com mais 4 grupos que consideramos marca para este estudo: pequena quantidade (representa os veículos de marca menos frequente); luxo (representa os veículos mais luxuosos); pesados (representa todos os veículos considerados "pesado" na Companhia; diversos (representa todos os veículos que por erro na Base de Dados não se percebe qual a sua marca). No entanto, existem outras marcas pouco representativas e portanto existe necessidade de as agrupar. A tabela seguinte 3.4 mostra a frequência relativa de registos de cada marca nesta carteira.

Marca	Freq. Rel.
Renault	11,28%
Opel	9,49%
Volkswagen	7,94%
Diversos	7,13%
Peugeot	7,03%
Ford	6,52%
Citroen	5,98%
Fiat	5,81%
Mercedes-Benz	5,30%
Toyota	4,59%
Seat	3,87%
BMW	3,24%
Audi	3,08%
Mitsubishi	2,41%
Nissan	2,38%
Honda	2,31%
Peq. Quant.	2,09%
Hyundai	1,20%
Volvo	1,09%
Smart	0,91%
Suzuki	0,91%
Pesados	0,80%
Mazda	0,78%
Skoda	0,71%
Land Rover	0,55%
Alfa Romeo	0,46%
Lancia	0,44%
Kia	0,43%
Luxo	0,35%
Chevrolet	0,33%
Mini	0,32%
Jeep	0,22%
Dacia	0,07%

Tabela 3.4: Frequência relativa de registos nas marcas

Foi feita novamente uma agregação por *clusters*, pelo método hierárquico, tal como para os distritos, e dada a quantidade de marcas, optou-se por se fazer a agregação em 6 *clusters* por custo médio por sinistro e frequência média de sinistralidade em cada marca. Existe a necessidade de padronizar essas variáveis, para que as mesmas sejam consideradas com o mesmo peso.

O dendrograma é mostrado na figura 3.15 e a tabela referente à agregação pode ser consultada em anexo (tabela 6.2).

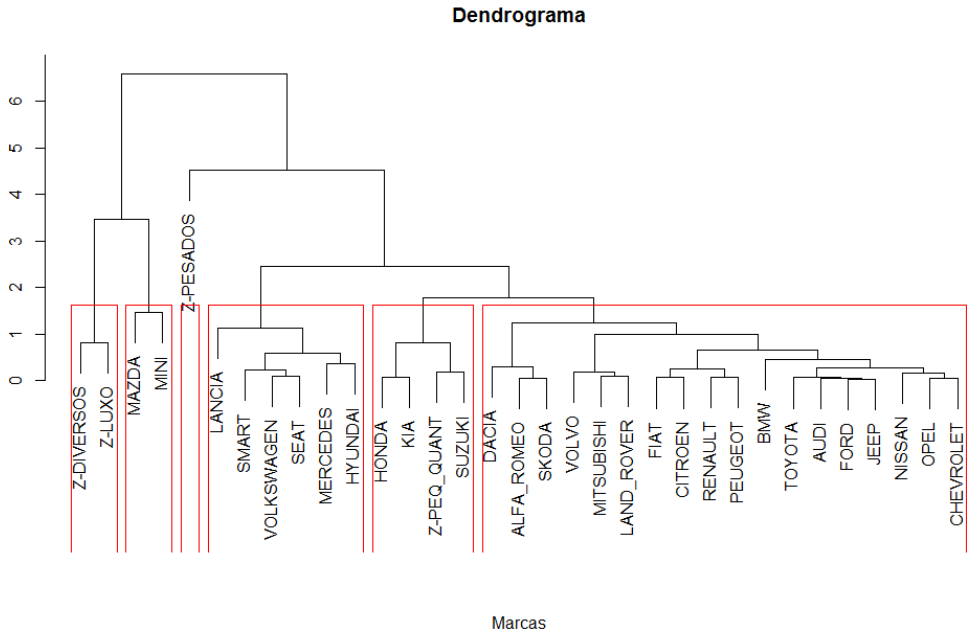

Figura 3.15: Agregação de marcas por custo médio por sinistro e frequência média de sinistralidade em cada marca

Após terminado o *clustering*, foi efetuada uma análise às diferenças entre os *clusters*. Foi utilizada a distribuição binomial para a comparação da variável resposta "teve sinistro" entre eles e a regressão linear para a comparação da variável resposta "custo com sinistros". Na tabela da figura 3.16 é possível concluir que os grupos 1 e 3 não apresentam diferenças significativas entre si em relação à frequência de sinistralidade, mas que na maioria das comparações, os grupos apresentam diferenças significativas ($\approx 0\%$) entre si. Em relação aos custos com sinistros, pode-se concluir que o grupo 6 ("Pesados") é parecido com qualquer grupo à exceção do grupo 5 (Mazda; Mini). Os conjuntos 1 e 2 e 3 e 4 também não apresentam diferenças significativas entre si[7].

[7] A tabela de agregação de *clusters* encontra-se em anexo (tabela 6.2).

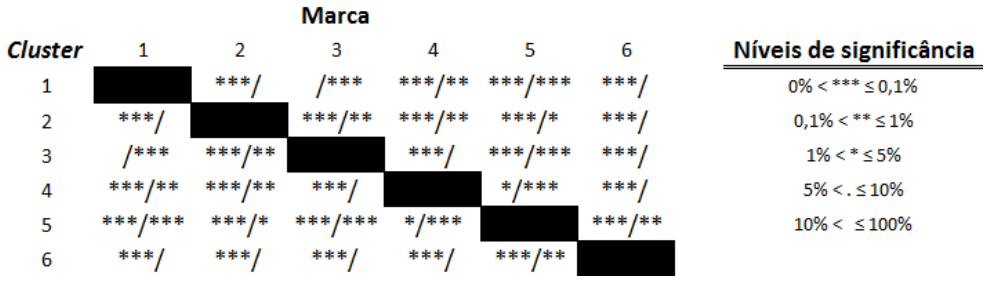

Figura 3.16: Análise às diferenças de frequência/custos de sinistralidade entre os *clusters* das marcas

3.2.8 Escalão de Cilindrada

A cilindrada é definida como o volume varrido pelo deslocamento de uma peça móvel numa câmara hermeticamente fechada durante um movimento unitário. Vem expressa em centímetros cúbicos (cm^3) e é uma variável bastante importante e a ter em conta no cálculo do prémio de risco, pois é um fator diferenciador nas características dos veículos. Os escalões de classificação utilizados pela Companhia são:

$$\text{Escalão } 1 \in [0; 1500];$$

$$\text{Escalão } 2 \in [1501; 2500];$$

$$\text{Escalão } 3 \in [2501; 5000].$$

De seguida, é apresentada a tabela de frequências relativas sobre os escalões de cilindrada (tabela 3.5).

Escalão de cilindrada	Freq. Rel.
Escalão 1	51,86%
Escalão 2	43,22%
Escalão 3	4,92%

Tabela 3.5: Frequência relativa de registos por escalão de cilindrada

Verifica-se que a maioria dos veículos pertencem ao escalão 1, ou seja, apresentam até 1500 cm^3 de cilindrada, inclusive.

Escalão de cilindrada	Prop. de apólices	Prop. de sinistros	Custo médio por sinistro	Prop. (sinistro|escalão)
Escalão 1	51,86%	90,06%	1.850,62€	8,15%
Escalão 2	43,22%	8,94%	1.980,21€	9,71%
Escalão 3	4,92%	1,00%	1.933,59€	9,57%

Tabela 3.6: Tabela da distribuição de apólices e sinistros por tipo de escalão de cilindrada, custo médio por sinistro e distribuição da sinistralidade condicionada no escalão de cilindrada, ordenada por proporção de apólices

A tabela 3.6 dá a informação que tanto o custo médio por sinistro, como a frequência de sinistralidade são mais elevados no escalão 2.

3.2.9 Categoria Agregada

Para mais fácil leitura, a base de dados disponibilizada já continha as categorias agregadas pela Companhia, como é mostrado na tabela 3.7.

Categoria	Prop. de apólices
Ligeiro de Passageiros	72,53%
Ligeiro de Mercadorias	17,78%
Diversos	4,44%
Motociclo	2,81%
Pesado de Mercadorias	1,92%
Reboque	0,30%
Garagista	0,17%
Pesado de Passageiros	0,05%

Tabela 3.7: Proporção de apólices por categoria agregada pela Companhia

Uma vez que se continua a ter classes pouco representativas, optou-se por agregar algumas delas.

Como existe pouca representatividade nos garagistas, reboques e pesados de passageiros, os garagistas e reboques passam a fazer parte dos diversos e os pesados ficam num só grupo (pesados de passageiros e pesados de mercadorias).

Categoria	Prop. de apólices
Ligeiro de Passageiros	72,53%
Ligeiro de Mercadorias	17,78%
Diversos	4,91%
Motociclo	2,81%
Pesado	1,96%

Tabela 3.8: Proporção de apólices pela nova agregação de categorias

3.2.10 Tipo de Uso

O número de apólices no tipo de uso está distribuído como mostra a tabela 3.9.

Entre "Vida Privada" e "Uso Particular" não há distinção, trata-se de haver apólices com tarifas antigas em que o tipo de uso era gravado como "Uso Particular" e hoje em dia é gravado como "Vida Privada". Assim, estes dois tipos de uso são agregados num só. Como é visível na tabela 3.9, o "Uso Particular" e a "Vida Privada" representam 84% dos dados, o que significa que mais uma vez poderá fazer sentido agregar os restantes tipos de uso.

O procedimento a ser seguido, nesta variável, é o mesmo que na categoria agregada. Tendo em conta a representatividade dos dados da tabela 3.9, faz todo o sentido deixar o "Uso Particular"/"Vida Privada" e o "Uso Profissional" em grupos distintos, pois são os tipos de uso mais representativos, assim como os "Táxis" noutro grupo distinto dos outros tipos de uso, pois é usada uma tarifa diferente para estes. Assim, são formadas então, 4 classes e os grupos são agregados como mostra a tabela 3.10.

Tipo de uso	Prop. de apólices
Vida Privada	78,532%
Uso Profissional	15,309%
Uso Particular	5,494%
Profissional - mercadorias (nacional/internacional)	0,310%
Profissional - táxi	0,29%
Profissional - mercadorias (nacional)	0,143%
Profissional - praça ou letra T	0,101%
Profissional - aluguer com condutor	0,040%
Profissional - transporte de matérias perigosas (tipo I)	0,030%
Profissional - instrução e exame	0,004%
Profissional - passageiros (nacional/internacional)	0,003%
Profissional - passageiros (nacional)	0,002%
Profissional - aluguer sem condutor	0,001%

Tabela 3.9: Proporção de apólices por tipo de uso

Grupo 1	Grupo 2	Grupo 3	Grupo 4
Uso Particular	Uso Profissional	Praça ou letra T	Aluguer com condutor
Vida Privada		Táxi	Aluguer sem condutor
			Instrução e exame
			Mercadorias (nacional)
			Mercadorias (nacional/internacional)
			Passageiros (nacional)
			Passageiros (nacional/internacional)
			Transporte de matérias perigosas (tipo I)

Tabela 3.10: Agregação em classes por tipos de uso

3.2.11 Número de Sinistros

O número de sinistros é variável resposta no modelo para a frequência de sinistralidade, e é sobre esta variável que se pretende analisar o efeito das variáveis explicativas descritas anteriormente.

O primeiro gráfico da figura 3.17 mostra claramente o excesso de zeros. 91,36% dos segurados não tiveram sinistro em dois anos e meio, enquanto que o segundo gráfico da mesma figura permite visualizar a frequência da variável resposta quando os zeros são excluídos. Tal proporção é mostrada na tabela 3.11.

Dos sinistrados, a maioria teve apenas 1 sinistro em dois anos e meio. Com as percentagens apresentadas na tabela 3.11, é natural que não seja possível visualizar algumas barras nos gráficos da figura 3.17.

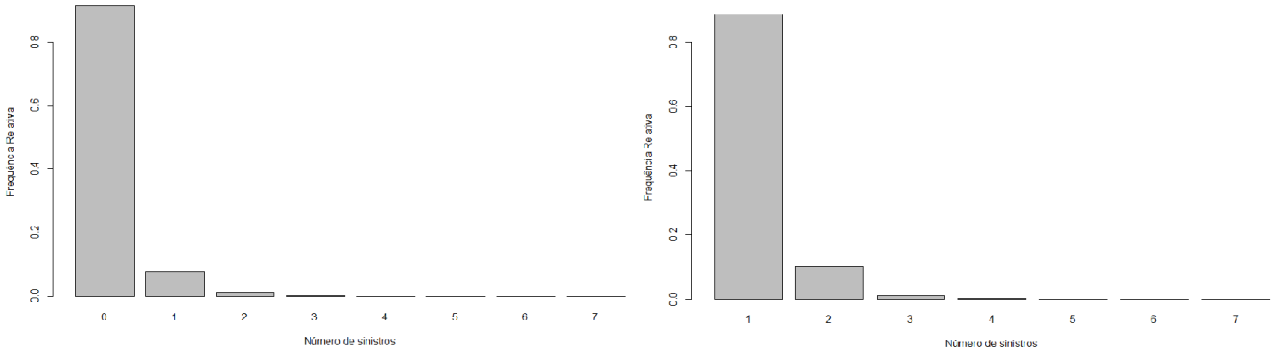

Figura 3.17: Ocorrência de sinistro incluindo e excluindo os zeros

Número de sinistros	Freq. Rel.
1	$88,551\%$
2	$10,118\%$
3	$1,117\%$
4	$0,182\%$
5	$0,019\%$
6	$0,008\%$
7	$0,006\%$

Tabela 3.11: Frequência relativa de sinistros, excluindo os zeros

É importante dizer que a variância desta variável é aproximadamente igual ao seu valor médio ($s^2 \approx 0,11; \overline{x} \approx 0,10$). Conclui-se que esta variável não tem sobredispersão nem subdispersão ($s^2 \approx \overline{x}$) e portanto poderá ser utilizada a distribuição de Poisson para esta variável resposta.

3.2.12 Custo com Sinistros

O custo com sinistros é variável resposta no modelo para o custo médio por sinistro, e é sobre esta variável que se pretende analisar o efeito das variáveis explicativas descritas anteriormente.

Foram construídos *boxplots* ("caixa de bigodes") para se avaliar a evolução do grau de sinistralidade de ano para ano (figura 3.18).

Para uma visão mais clara dos dados, foram construídos *boxplots* com os custos logaritmizados (figura 3.19).

A tabela 3.12 apresenta os extremos e quartis com a respetiva média de custo com sinistros por apólice em cada ano e a tabela 3.13 apresenta os extremos e quartis com a respetiva média do logaritmo do custo com sinistros por apólice em cada ano.

Analisando os gráficos das figuras 3.18 e 3.19 e tabelas 3.12 e 3.13, verifica-se que o número de sinistros tem vindo a diminuir de ano para ano e que os custos também têm vindo a diminuir muito significativamente.

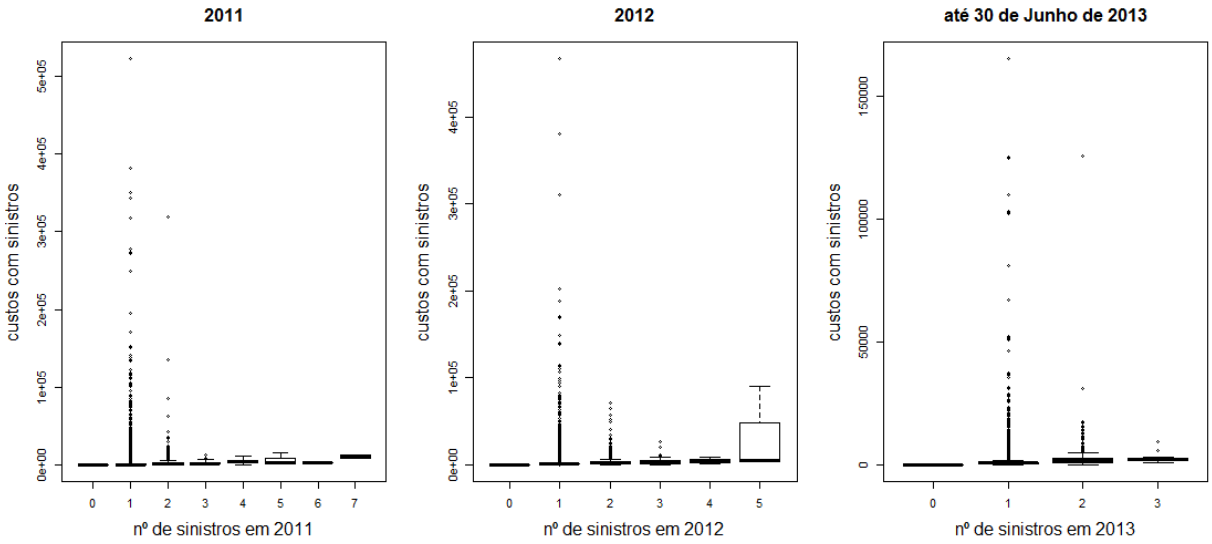

Figura 3.18: Custo com sinistros (por número de sinistros por apólice nos anos 2011, 2012 e 1º semestre de 2013)

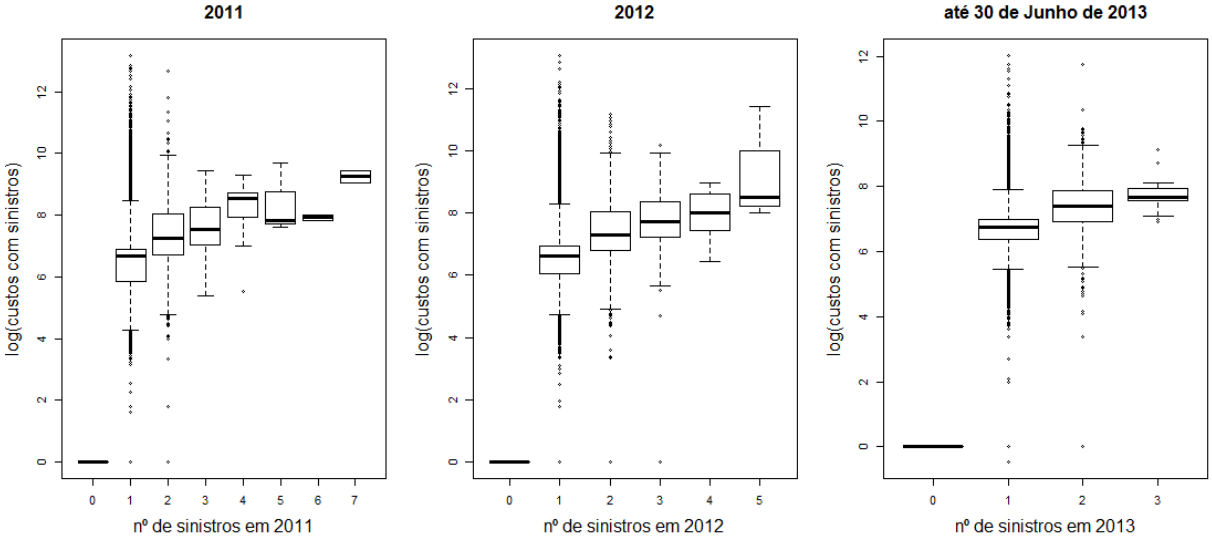

Figura 3.19: Log(Custo com sinistros) (por número de sinistros por apólice nos anos 2011, 2012 e 1º semestre de 2013)

	Mínimo	1º Quartil	2º Quartil	3º Quartil	Máximo	Média
2011	0,00 €	0,00 €	0,00 €	0,00 €	522.296,50 €	75,80 €
2012	0,00 €	0,00 €	0,00 €	0,00 €	467.448,90 €	66,20 €
1º Semestre de 2013	0,00 €	0,00 €	0,00 €	0,00 €	165.363,13 €	23,85 €

Tabela 3.12: Tabela de extremos e quartis e média de custos com sinistros

	Mínimo	1º Quartil	2º Quartil	3º Quartil	Máximo	Média
2011	0,0000	0,0000	0,0000	0,0000	13,1660	0,2648
2012	0,0000	0,0000	0,0000	0,0000	13,0550	0,2367
1º Semestre de 2013	0,0000	0,0000	0,0000	0,0000	12,0159	0,0978

Tabela 3.13: Tabela de extremos e quartis e média de custos com sinistros logaritmizados (Às apólices sem sinistros corresponde um custo nulo. Nas restantes apólices logaritmizou-se o custo.)

3.2.13 Exposição ao Risco

A cada registo está associado um número de dias no qual a apólice esteve em vigor, podendo tomar um mínimo de um dia e um máximo de 912 dias. Este é o tempo que o objeto ou serviço seguro na Companhia está sujeito a sofrer um dano futuro e incerto, ou de data incerta.

Nesta carteira, a exposição ao risco vem expressa em anos, com um mínimo de 0,0027 anos e um máximo de 2,5 anos, que significa precisamente o que foi referido acima: 0,0027 anos é equivalente a 1 dia e 2,5 anos é equivalente a 912 dias.

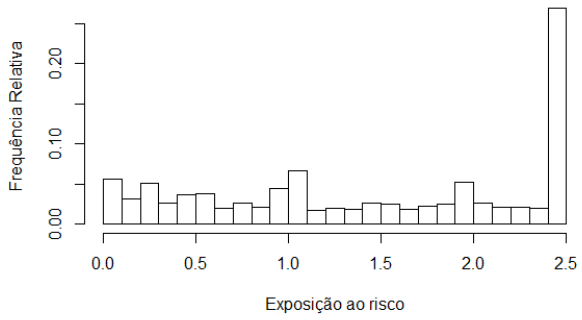

Figura 3.20: Tempo, em anos, que as apólices estiveram em vigor

Esta variável é considerada como um termo *offset* no ajustamento do modelo, não se tratando de uma variável explicativa ou resposta, como as restantes.

Capítulo 4

Aplicação

Nesta fase do trabalho, seguindo o exposto no suporte teórico, é feita a estimação de variados modelos para a frequência de sinistros e também para o custo médio por sinistro por apólice, com base num conjunto de variáveis explicativas escolhidas à partida e que constituem o conjunto de informação habitualmente considerada pelas companhias na definição do prémio.

Um aspeto importante a ter aqui em conta é que quase certamente nenhum segurado pagará pelo seu seguro aquilo que virá a receber dele. A maioria dos segurados não tem sinistros e, aqueles que têm, representam custos em regra bastante superiores ao que pagaram pelo seu seguro. O objetivo na utilização de modelos para estabelecimento do prémio é, portanto, apenas conseguir diferenciar de forma razoável os segurados que apresentam mais risco daqueles que apresentam menor risco. Haverá sempre uma diferença grande entre os valores observados dos custos com os sinistros e o prémio pago por cada segurado. O que importa nesta atividade é que, havendo alguma diferenciação no prémio de acordo com o risco, no global, a receita gerada pelo conjunto dos prémios puros seja suficiente para cobrir os custos dos segurados.

Este foi o princípio que regeu a análise que se fez neste trabalho, comparando-se os diferentes modelos estimados com recurso a medidas de erro. A convencional análise de resíduos, é feita apenas no sentido de validação dos pressupostos do modelo de regressão linear. É avaliada graficamente a distribuição dos resíduos padronizados segundo as variáveis explicativas do modelo. A normalidade dos resíduos é avaliada graficamente com recurso a um gráfico quantil-quantil. A expectativa de que os resíduos individualmente sejam pequenos, faz neste contexto pouco sentido pelo acima exposto. Ainda assim, apresenta-se de forma sucinta uma tal análise.

4.1 O Modelo

Como já foi apresentado antes, no suporte teórico, o número esperado de sinistros para um cliente i é dado por:

$$\lambda_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}, \tag{4.1}$$

e o valor esperado para o custo por sinistro é dado por:

$$E\left[custo_{sinistro}|x_{i1}, x_{i2}, ..., x_{ip}\right] = exp\left(E\left(log(custo)\right) + \frac{\widehat{\sigma^2}}{2}\right) = e^{\gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+...+\gamma_p x_{ip}} \times e^{\frac{\widehat{\sigma^2}}{2}}. \tag{4.2}$$

Como o prémio de risco é dado pelo produto do número esperado de sinistros e o valor esperado do custo por sinistro, então o prémio de risco é dado pela expressão:

$$Prémio\ de\ Risco = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}} \times e^{\gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+...+\gamma_p x_{ip}} \times e^{\frac{\widehat{\sigma^2}}{2}}. \tag{4.3}$$

Os modelos têm 29 parâmetros cada um. No modelo de regressão escolhido para a frequência, tem-se $\beta_0, \beta_1, \beta_2, ..., \beta_{28}$, onde β_0 é o intercepto e $\beta_1, \beta_2, ..., \beta_{28}$ são os coeficientes das variáveis,

como será apresentado em tabela mais adiante na discussão de resultados. No caso das variáveis categóricas, a primeira categoria apresentada é a categoria referência, dizendo os coeficientes respeito às restantes categorias.

Tal como no modelo para a frequência, no modelo de regressão para o custo por sinistro, tem-se $\gamma_0, \gamma_1, \gamma_2, ..., \gamma_{28}$, onde γ_0 é o intercepto e $\gamma_1, \gamma_2, ..., \gamma_{28}$ são os coeficientes das variáveis, como será apresentado também em tabela mais adiante na discussão de resultados. No caso das variáveis categóricas, a primeira categoria apresentada é a categoria referência, dizendo os coeficientes respeito às restantes categorias.

4.2 Construção dos Modelos

Com base nesta carteira, é traçado um perfil para cada indivíduo, com o objetivo de posteriormente se avaliar qual o risco que um novo segurado pode vir a trazer à Companhia.

Relativamente à codificação das variáveis categóricas no *R*, por defeito, o *software* toma como categoria referência a primeira categoria de cada variável, sendo possível alterar. Assim, quando os fatores são carateres alfabéticos, a categoria referência é a correspondente à primeira palavra por ordem alfabética ou, se por outro lado as variáveis forem dadas por carateres numéricos, a categoria referência é a classe com valor mais baixo.

Para a estimação de um Modelo Linear Generalizado é necessário trabalhar com casos completos, deixa-se de considerar todas de apólices e passa-se a considerar apenas todos os casos com informação completa de todas as variáveis em cada apólice que equivalem a, aproximadamente, 86% das apólices que foram consideradas na análise exploratória dos dados.

Uma vez que se dispõe de um número muito elevado de registos, a Base de Dados foi dividida em dois blocos, utilizando-se o primeiro para construir os modelos e o segundo para validar os mesmos. A estes grupos deu-se a designação de grupo de treino e grupo de teste.

O grupo de treino tem uma dimensão 1.5 vezes superior ao do grupo de teste, tendo ambos mais de uma centena de milhar de casos.

Foram construídos modelos para a frequência e modelos para o custo por sinistro, onde se utiliza os dados de treino.

Nos modelos para a frequência, foram considerados o modelo de regressão Poisson, o modelo *Hurdle* Poisson e o modelo *Hurdle* binomial negativo, como já foram apresentados no suporte teórico. De forma a tornar possível a utilização dos modelos *Hurdle* no *R*, é necessário obter o *package pscl*.

Para o custo por sinistro, foram considerados o modelo de regressão linear múltipla, o modelo de regressão linear múltipla com variável resposta logaritmizada e o modelo de regressão gama, já que esta é uma distribuição utilizada quando a variável resposta é de natureza contínua e apresenta assimetria positiva. Estes modelos foram também apresentados anteriormente no suporte teórico.

São calculados os valores ajustados em cada modelo para cada registo. Os valores estimados para a frequência são o produto entre a frequência esperada por unidade de tempo e o tempo que essa apólice esteve exposta ao risco na Companhia. Para o custo, os valores estimados para os vários modelos, para cada indivíduo, são o custo esperado por sinistro.

4.3 Resultados e Discussão

A avaliação do erro de ajustamento é fundamental para a escolha do modelo. Este conhecimento adicional fornece uma melhor perceção sobre o quão precisa pode vir a ser a previsão em termos globais. Os desvios negativos ocorrem quando o valor ajustado tem um valor mais elevado que o valor observado.

As decisões podem ser influenciadas de duas formas distintas pelos erros de previsão: uma forma consiste na escolha entre alternativas de previsão e a outra consiste na avaliação do sucesso ou fracasso da técnica utilizada.

4.3.1 Medidas de Erro

No estudo das técnicas de previsão, as medidas de precisão são de extrema importância. É fundamental incluir informação acerca da medida em que a previsão pode desviar-se do valor real da variável. Este conhecimento adicional fornece uma melhor perceção sobre o quão precisa pode ser a previsão.

Existem várias medidas de erro: Erro Médio (EM), Desvio Médio Absoluto (DMA), Erro Quadrático Médio (EQM), entre outras.

A diferença entre o valor real (custo) e a previsão do valor (custo esperado) dá origem ao erro de previsão:

$$e_i = custo_i - custo\ esperado_i, \tag{4.4}$$

para cada caso i.

O EM é dado pelo quociente entre a soma dos erros de previsão e o número de casos considerados:

$$EM = \frac{\sum_{i=1}^{n} e_i}{n}. \tag{4.5}$$

O DMA é dado pelo quociente entre a soma dos valores absolutos dos erros de previsão e o número de casos considerados:

$$DMA = \frac{\sum_{i=1}^{n} |e_i|}{n}. \tag{4.6}$$

O EQM é dado pelo quociente entre a soma dos quadrados dos erros de previsão e o número de casos considerados:

$$EQM = \frac{\sum_{i=1}^{n} e_i^2}{n}. \tag{4.7}$$

Na tabela 4.1 são apresentadas as medidas de erro apresentadas anteriormente para cada modelo, calculadas com os dados de teste:

		Modelos para o custo por sinistro					
		Regressão Linear		Regressão Linear (log)		Regressão Gama	
Modelos para a frequência	Regressão Poisson	EM	$-166,33$	EM	$-126,82$	EM	$-165,18$
		DMA	$429,16$	DMA	$393,60$	DMA	$427,70$
		EQM	$2.351,23$	EQM	$2.343,04$	EQM	$2.350,61$
	Hurdle Poisson	EM	$-170,79$	EM	$-130,75$	EM	$-169,60$
		DMA	$432,99$	DMA	$396,92$	DMA	$431,50$
		EQM	$2.352,23$	EQM	$2.343,67$	EQM	$2.351,57$
	Hurdle Binomial Negativo	EM	$-171,58$	EM	$-131,53$	EM	$-170,40$
		DMA	$433,65$	DMA	$397,57$	DMA	$432,17$
		EQM	$2.352,31$	EQM	$2.343,75$	EQM	$2.351,65$

Tabela 4.1: Erros de previsão dos modelos

Entre os modelos, aquele que apresentou todos os erros mais próximos de zero foi o modelo considerado para o prémio de risco, o modelo que combina o modelo de regressão de Poisson (para a frequência) com o modelo de regressão linear múltipla com variável resposta logaritmizada (para o custo por sinistro). Ainda assim, é possível observar que os vários modelos multiplicativos apresentam as medidas de erro bastante semelhantes entre si.

Nas tabelas apresentadas de seguida, tabelas 4.2 e 4.3, são apresentados os fatores a ser utilizados no modelo multiplicativo, que correspondem à exponencial de cada valor estimado da variável, dependendo das características do condutor e do veículo.

Para as variáveis categóricas, o valor a entrar no modelo multiplicativo será o fator que se insere no escalão associado à característica. Quando a categoria referência da variável é a característica escolhida, então esse fator toma o valor 1, valor a entrar no modelo multiplicativo. Para as variáveis quantitativas (idade do veículo, idade da carta, idade do condutor), o valor a ser utilizado no cálculo do prémio de risco é o fator relativo a essas variáveis explicativas elevado à respetiva idade.

De seguida, são apresentadas as tabelas de regressão dos modelos escolhidos para a frequência e para o custo médio por sinistro com as exponenciais dos parâmetros estimados para cada modelo (tabelas 4.2 e 4.3).[1]

[1]Em anexo, poderão ser consultadas as tabelas relativas aos outros modelos em estudo (tabelas 6.3, 6.4, 6.5 e 6.6).

		Regressão Poisson		e^{β_j}
	β_0		(Intercept)	0.0199 ***
Categoria		Diversos	categoria_agregDIVERSOS	1.0000
	β_1	Lig. de Mercad.	categoria_agregLIGEIRO_DE_MERCADORIAS	3.1670 ***
	β_2	Lig. de Passag.	categoria_agregLIGEIRO_DE_PASSAGEIROS	2.5065 ***
	β_3	Motociclo	categoria_agregMOTOCICLO	0.8545 •
	β_4	Pesado	categoria_agregPESADO	1.5389 *
Distrito		Classe 1	distrito_agreg11	1.0000
	β_5	Classe 2	distrito_agreg12	0.8952 ***
	β_6	Classe 3	distrito_agreg13	0.9220 ***
	β_7	Classe 4	distrito_agreg14	0.9136 **
	β_8	Classe 5	distrito_agreg15	1.0086
Concelho		Classe 1	concelho_agreg11	1.0000
	β_9	Classe 2	concelho_agreg12	0.4404 ***
	β_{10}	Classe 3	concelho_agreg13	1.3186 ***
	β_{11}	Classe 4	concelho_agreg14	0.8414 **
	β_{12}	Classe 5	concelho_agreg15	0.7788 •
	β_{13}	Classe 6	concelho_agreg16	1.0997 •
Escalão de Cilindrada		Escalão 1	escalao_cilindrada[0,1500]	1.0000
	β_{14}	Escalão 2	escalao_cilindrada[1501,2500]	1.0896 ***
	β_{15}	Escalão 3	escalao_cilindrada[2501,5000]	1.2923 ***
Idades	β_{16}	Carta	idade_carta	0.9848 ***
	β_{17}	Condutor	idade_condutor	1.0067 ***
	β_{18}	Veículo	idade_veículo	1.0051 ***
Marca		Classe 1	marca_agreg1	1.0000
	β_{19}	Classe 2	marca_agreg2	0.6491 ***
	β_{20}	Classe 3	marca_agreg3	1.0427 **
	β_{21}	Classe 4	marca_agreg4	1.0389
	β_{22}	Classe 5	marca_agreg5	0.9184
	β_{23}	Classe 6	marca_agreg6	1.3157 **
Subscritor		Empresa	subscritorEMPRESA	1.0000
	β_{24}	Feminino	subscritorFEMININO	1.0539
	β_{25}	Masculino	subscritorMASCULINO	1.0466
Tipo de Uso		Classe 1	tipo_uso_agreg1	1.0000
	β_{26}	Classe 2	tipo_uso_agreg2	1.2717 ***
	β_{27}	Classe 3	tipo_uso_agreg3	2.7065 ***
	β_{28}	Classe 4	tipo_uso_agreg4	2.1468 •

Tabela 4.2: Coeficientes multiplicativos para o cálculo do prémio de risco - Exponenciais dos β's estimados para a frequência de sinistralidade

		Regressão Linear (log)		e^{γ_j}
	γ_0		(Intercept)	1014.3230 ***
Categoria		Diversos	categoria_agregDIVERSOS	1.0000
	γ_1	Lig. de Mercad.	categoria_agregLIGEIRO_DE_MERCADORIAS	1.0228
	γ_2	Lig. de Passag.	categoria_agregLIGEIRO_DE_PASSAGEIROS	0.9690
	γ_3	Motociclo	categoria_agregMOTOCICLO	0.8371
	γ_4	Pesado	categoria_agregPESADO	0.6815
Distrito		Classe 1	distrito_agreg11	1.0000
	γ_5	Classe 2	distrito_agreg12	0.9958
	γ_6	Classe 3	distrito_agreg13	1.0014
	γ_7	Classe 4	distrito_agreg14	0.9854
	γ_8	Classe 5	distrito_agreg15	0.9553
Concelho		Classe 1	concelho_agreg11	1.0000
	γ_9	Classe 2	concelho_agreg12	0.5965 **
	γ_{10}	Classe 3	concelho_agreg13	0.7967 ***
	γ_{11}	Classe 4	concelho_agreg14	0.8193 **
	γ_{12}	Classe 5	concelho_agreg15	1.4452 .
	γ_{13}	Classe 6	concelho_agreg16	0.8076 **
Escalão de Cilindrada		Escalão 1	escalao_cilindrada[0,1500]	1.0000
	γ_{14}	Escalão 2	escalao_cilindrada[1501,2500]	1.0060
	γ_{15}	Escalão 3	escalao_cilindrada[2501,5000]	1.0521
Idades	γ_{16}	Carta	idade_carta	0.9988
	γ_{17}	Condutor	idade_condutor	0.9994
	γ_{18}	Veículo	idade_veículo	1.0058 ***
Marca		Classe 1	marca_agreg1	1.0000
	γ_{19}	Classe 2	marca_agreg2	0.9919
	γ_{20}	Classe 3	marca_agreg3	0.9717
	γ_{21}	Classe 4	marca_agreg4	0.9458
	γ_{22}	Classe 5	marca_agreg5	1.1448 .
	γ_{23}	Classe 6	marca_agreg6	1.0889
Subscritor		Empresa	subscritorEMPRESA	1.0000
	γ_{24}	Feminino	subscritorFEMININO	0.8754
	γ_{25}	Masculino	subscritorMASCULINO	0.9020
Tipo de Uso		Classe 1	tipo_uso_agreg1	1.0000
	γ_{26}	Classe 2	tipo_uso_agreg2	0.9215
	γ_{27}	Classe 3	tipo_uso_agreg3	0.9679
	γ_{28}	Classe 4	tipo_uso_agreg4	1.6777

Tabela 4.3: Coeficientes multiplicativos para o cálculo do prémio de risco - Exponenciais dos γ's estimados para o custo por sinistro

Analisando as tabelas 4.2 e 4.3, consegue-se perceber que os coeficientes do modelo para a frequência são os que mais influenciam no aumento do prémio de risco, uma vez que as exponenciais dos parâmetros estimados são significativamente maiores, em geral, do que as exponenciais dos parâmetros estimados para o modelo para o custo por sinistro.

No modelo Poisson, tabela 4.2, conclui-se que o facto de pertencer ao grupo dos ligeiros, principalmente ligeiro de mercadorias, contribui significativamente para o aumento do valor do prémio de risco, uma vez que estes são os que mais contribuem para o aumento da frequência de sinistralidade. Para um ligeiro de mercadorias, o coeficiente multiplicativo é de 3.1670, o que significa que em termos de número de sinistros por unidade de tempo (ano), se espera para um ligeiro de mercadorias, 3.1670 vezes mais sinistros do que para um "diversos", que é a classe de referência.

Por sua vez, pertencer ao grupo 2 nos concelhos, baixa significativamente o valor do prémio de risco, uma vez que o valor esperado de sinistros é 56% inferior relativamente à classe de referência.

Conclui-se também que o facto de ser taxista contribui muito significativamente no modelo e apresenta também um aumento significativo no valor do prémio de risco.

Tal como era de esperar, é possível verificar que quanto maior a idade da carta, menos é a probabilidade de ocorrência de sinistro. E, quanto maior a idade do condutor, maior é a probabilidade de ocorrência de sinistro.

No modelo de regressão linear com variável resposta logaritmizada, 4.3, consegue-se perceber que os veículos com maior cilindrada são aqueles cujo valor com sinistros é mais elevado, tal como seria de esperar. O custo com sinistros sobe também quando o tipo de uso pertence ao grupo 4, grupo este que engloba veículos de aluguer e veículos com características mais específicas, tais como transporte de mercadorias, transporte de matérias perigosas ou até mesmo transporte de passageiros. Entende-se que são transportes de maior cilindrada, pesados, daí a justificação dos custos.

Quanto ao concelho, o grupo 5 apresenta de facto um incremento no valor do prémio de risco, uma vez que neste conjunto de dados houveram realmente sinistros com valores acima de uma dezena de milhar de euros e portanto era de esperar este comportamento no modelo.

No modelo, em geral, conclui-se que as categorias ligeiro de mercadorias, ligeiros de passageiros e pesados, contribuem para um aumento da frequência esperada, comparativamente à categoria referência "diversos". Os veículos ligeiros são os que mais têm um impacto negativo no cálculo do prémio de risco, principalmente se for ligeiro de mercadorias. Se o tipo de uso pertencer ao grupo 3 (táxis) ou 4 (diversos tipos de uso profissional), também irá ter uma grande influência no valor final do cálculo do prémio de risco, comparativamente à categoria referência "uso particular". Quanto à cilindrada do veículo, o terceiro escalão é o que vai ter mais impacto no aumento da frequência e custos esperados. Na marca do veículo, os pesados contribuem para o aumento da frequência esperada, mas o grupo 5 (Mazda; Mini) é o grupo com um maior aumento a nível de custos esperados por sinistro. Quanto ao subscritor, são sem dúvida as mulheres que têm mais sinistros, mas também as que têm menos custos com sinistros. Como já seria de esperar, em geral, as mulheres são mais cuidadosas e conduzem mais a medo em relação aos homens, daí terem os chamados "toques de cidade"e nada de sinistros graves. Também a idade da carta e do condutor contribuem de alguma forma, como já referi anteriormente, sendo que quanto maior a idade da carta conjugada com a idade do condutor, menos a probabilidade de ocorrência de sinistro. Mais idade do condutor significa, em geral, maior responsabilidade deste, assim como, mais experiência. Estes são dois dos principais fatores na descida do valor do prémio de risco.

É calculado o MSE (média da soma dos quadrados entre grupos) que é a estimativa da

variância ($\widehat{\sigma^2}$), no modelo escolhido para o custo por sinistro, e é dado pela seguinte expressão:

$$\widehat{\sigma^2} = MSE = \frac{soma\ de\ quadrados\ dos\ resíduos}{graus\ de\ liberdade\ dos\ resíduos} = \frac{39426.62}{25506} \approx 1.5458. \tag{4.8}$$

Com base nos resultados, procede-se ao cálculo do prémio base. O prémio base corresponde ao valor estimado para o prémio de risco considerando que todas as variáveis categóricas são iguais à categoria referência e que as variáveis quantitativas são nulas. É dado pela seguinte expressão:

$$Prémio\ Base = e^{\beta_{0\,frequência}} \times e^{\gamma_{0\,custo\ médio}} \times e^{\frac{MSE}{2}} = 0.0199 \times 1014.3230 \times e^{\frac{1.5458}{2}} = 43.72€. \tag{4.9}$$

O prémio de risco base a que se chegou foi de 43.72 €, o que parece ser um montante bastante razoável para a cobertura de Responsabilidade Civil do ramo Automóvel.

O prémio de risco para qualquer cenário é dado por:

$$Prémio\ de\ Risco = Prémio\ Base \times \prod_{j=1}^{p} e^{\beta_j x_{ij}} \times \prod_{j=1}^{p} e^{\gamma_j x_{ij}}. \tag{4.10}$$

Neste trabalho, a medida de erro a ser utilizada é o Erro Quadrático Médio (EQM), uma vez que em todos os modelos multiplicativos é aquele que está mais próximo de zero.

Serão agora enunciados dois exemplos para o cálculo do prémio de risco.

Exemplo 1

		Coeficientes	
		Frequência	**Custo por Sinistro**
Categoria	Ligeiro de Passageiros	2.5065	0.9690
Cilindrada	escalão 1	1.0000	1.0000
Concelho	Sesimbra (*cluster* 6)	1.0997	0.8076
Distrito	Setúbal (*cluster* 1)	1.0000	1.0000
Idade de Carta	7	0.8983	0.9916
Idade do Condutor	26	1.1896	0.9845
Idade do Veículo	5	1.0258	1.0293
Marca	Audi (*cluster* 3)	1.0427	0.9717
Subscritor	Feminino	1.0539	0.8754
Tipo de Uso	Uso Particular (*cluster* 1)	1.0000	1.0000

Tabela 4.4: Exemplo prático 1

Multiplicando todos estes fatores pelo prémio base, 43.72 €, obtém-se um Prémio de Risco por ano de 97.11 €.

Exemplo 2

		Coeficientes	
		Frequência	**Custo por Sinistro**
Categoria	Ligeiro de Passageiros	3.1670	1.0228
Cilindrada	escalão 2	1.0896	1.0060
Concelho	Lisboa (*cluster* 3)	1.3186	1.7967
Distrito	Lisboa (*cluster* 1)	1.0000	1.0000
Idade de Carta	20	0.9698	0.9976
Idade do Condutor	40	1.1505	0.9875
Idade do Veículo	8	1.0051	1.0058
Marca	Mercedes-Benz (*cluster* 1)	1.0427	0.9717
Subscritor	Empresa	1.0000	1.0000
Tipo de Uso	Profissional - táxi (*cluster* 3)	1.2717	0.9215

Tabela 4.5: Exemplo prático 2

Multiplicando todos estes fatores pelo prémio base, 43.47€, obtém-se um Prémio de Risco por ano de 213.93€.

Os exemplos 4.4 e 4.5 mostram o resultado do impacto das exponenciais dos parâmetros observados nas tabelas 4.2 e 4.3 observadas anteriormente, dependendo das características do veículo e do tomador de seguro.

4.4 Diagnóstico dos Modelos

Dados invulgares podem por vezes, danificar as estimativas das regressões, mas podem por si só, revelar também informações importantes. Estes incluem *outliers*, dados com elevada alavancagem e observações com grande influência.

Outliers são valores da variável resposta (Y) que são invulgares, dados os valores das variáveis preditoras.

4.4.1 Regressão de Poisson

Nesta secção apresenta-se uma análise exploratória dos resíduos do modelo de Poisson estimado para a frequência de sinistralidade. Como a ocorrência de sinistros é um acontecimento raro e as condições de ocorrência não dependem exclusivamente das características do condutor do veículo e da região, mas de outras como, por exemplo, condições meteorológicas e outras, tais como, consumo de álcool, cansaço e também aspetos sasonais como dia da semana e época do ano, é de esperar que o modelo para a frequência não seja capaz de prever um número médio de ocorrências sequer superior a um por unidade de tempo.

Por consequência, os resíduos são essencialmente de dois tipos: negativos e de pequeno valor absoluto, correspondendo às situações em que não houve sinistro, ou positivos com valores até 7, correspondendo aos casos em que houveram sinistros. Note-se num entanto que a proporção de resíduos mais afastados de zero é de apenas 8.5%. É um valor muito pouco significativo tendo em conta a dimensão dos dados.

A leitura dos gráficos que a seguir se apresentam devem ter em conta estes aspetos.

Nas figuras 4.1 e 4.2 apresentam-se os gráficos dos resíduos *versus* as variáveis preditoras categóricas que se traduzem num conjunto de *boxplots* para as várias classes dos preditores.

Após a correta estimação do modelo, espera-se que em todas as classes das variáveis preditoras se observe o mesmo centro e a mesma amplitude interquartis, refletindo a ausência de correlação entre resíduos e preditores.

Na figura 4.3 apresenta-se os gráficos dos resíduos *versus* as variáveis preditoras quantitativas e *Fitted values* (frequência estimada).

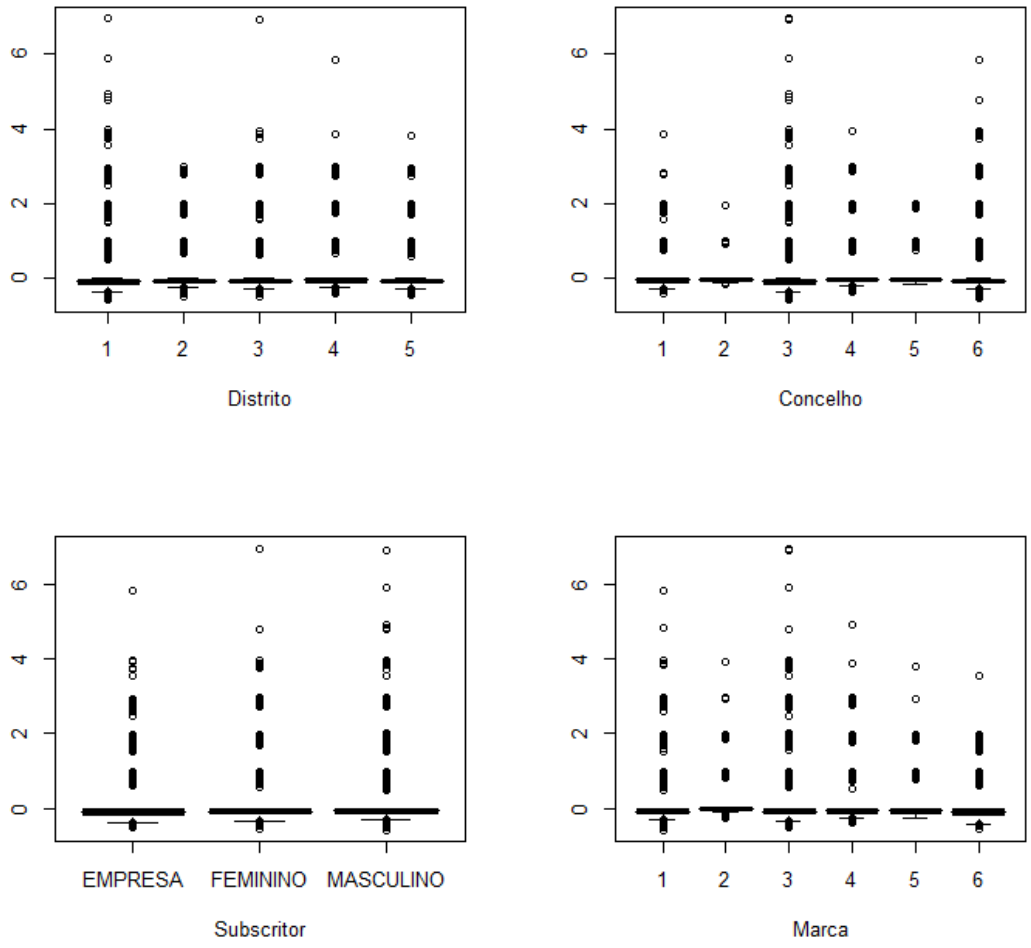

Figura 4.1: Regressão de Poisson - Resíduos

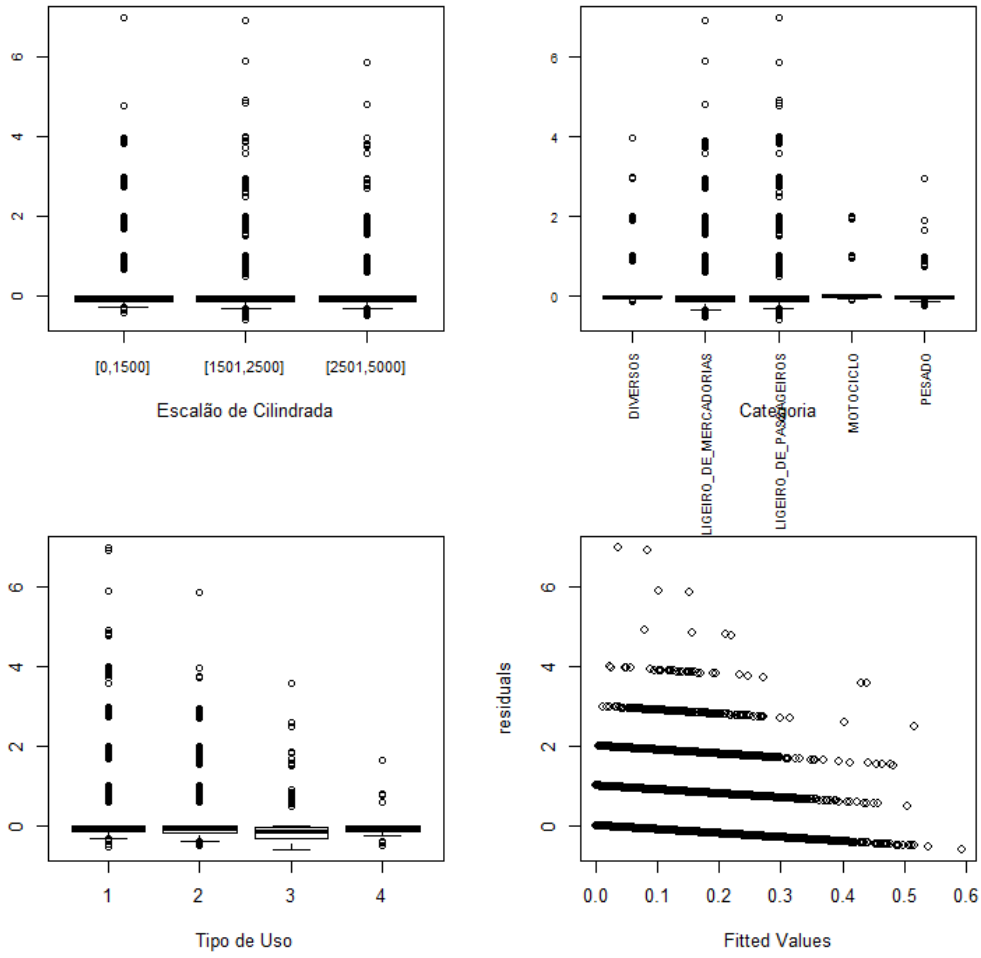

Figura 4.2: Regressão de Poisson - Resíduos

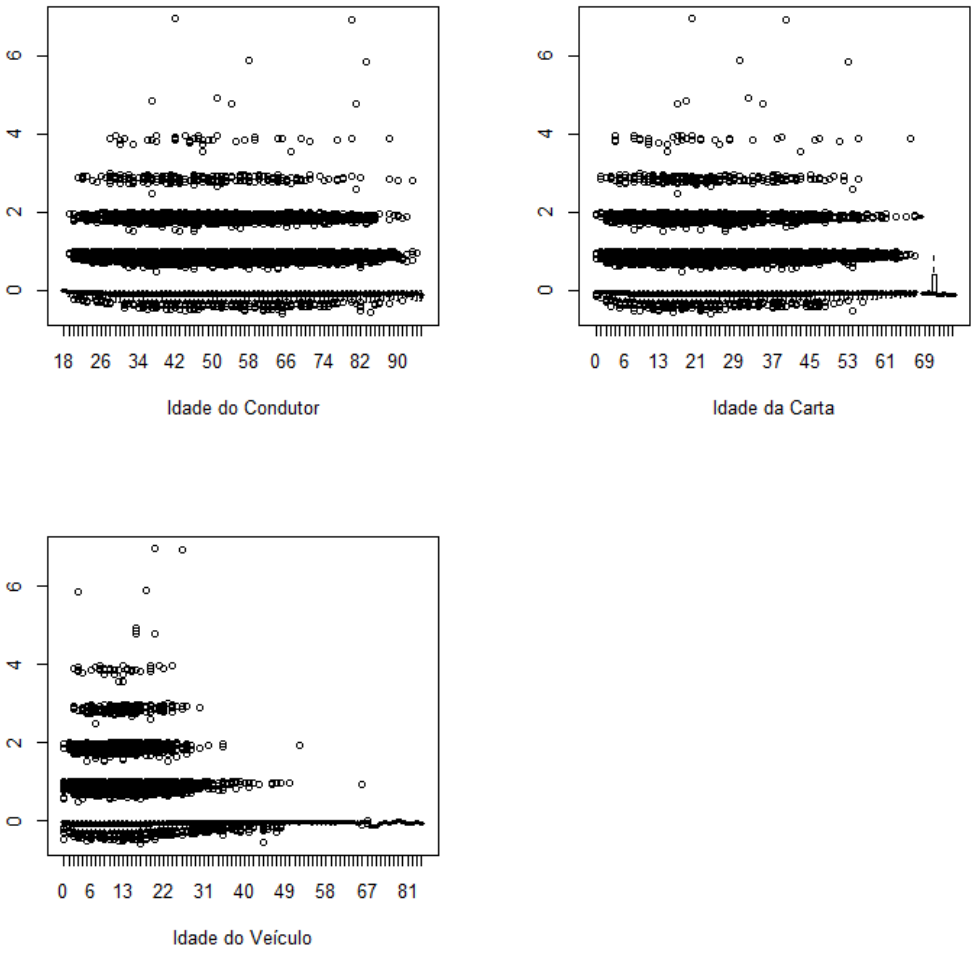

Figura 4.3: Regressão de Poisson - Resíduos

4.4.2 Regressão Linear Múltipla com variável resposta logaritmizada

Nesta secção apresenta-se uma análise exploratória dos resíduos padronizados (*standardized residuals*) do modelo de regressão linear múltipla com variável resposta logaritmizada estimado para o custo por sinistro. É de esperar que o custo por sinistro seja um valor a pagar acima do valor que o tomador de seguro paga à Companhia por ano. Estes resíduos são calculados em torno das apólices com custo por sinistro acima de zero.

Estes resíduos apresentam simetria entre si, isto é, estão bem comportados em torno do zero. Apresentam-se algumas estatísticas sumárias dos resíduos padronizados: mínimo -5.88864; 1º quartil -0.40865; mediana 0.09689; média 0.0; 3º quartil 0.34758; máximo 5.08056. Apesar dos valores mais extremos dos resíduos padronizados serem considerados elevados para uma população normal padrão, no conjunto dos resíduos padronizados, apenas uma proporção 0.0087 excede 3 em valor absoluto. Comparando com a proporção esperada sob o pressuposto de normalidade, a diferença encontrada é muito pequena ($P(|Z| > 3) = 0.0027$).

Na figura 4.4 encontra-se um gráfico quantil-quantil construído com os resíduos padronizados. Espera-se que estes apresentem um comportamento aproximadamente normal. Podemos considerar que o pressuposto de normalidade é razoavelmente cumprido, registando-se apenas algum afastamento nas caudas.

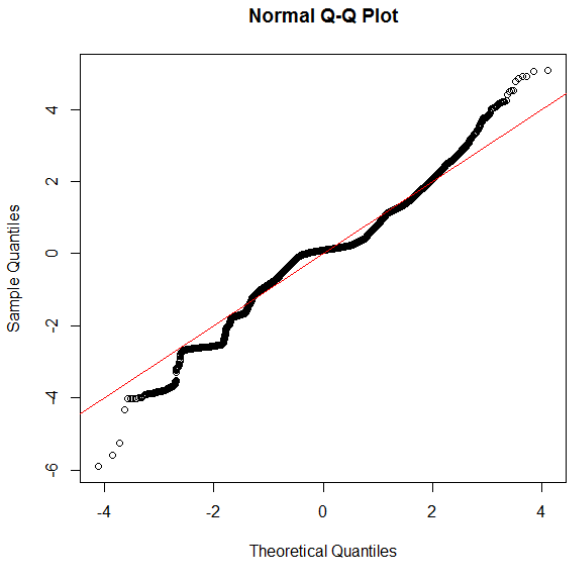

Figura 4.4: Regressão Linear com variável resposta logaritmizada - gráfico quantil-quantil para os resíduos padronizados

Nas figuras 4.5 e 4.6 apresentam-se os gráficos dos resíduos padronizados *versus* as variáveis preditoras categóricas que se traduzem num conjunto de *boxplots* para as várias classes dos preditores.

Após a correta estimação do modelo, espera-se que em todas as classes das variáveis preditoras se observe o mesmo centro e a mesma amplitude interquartis, refletindo a ausência de correlação entre resíduos e preditores.

Na figura 4.7 apresenta-se os gráficos dos resíduos padronizados *versus* as variáveis preditoras quantitativas e *Fitted values* (custos médios estimados).

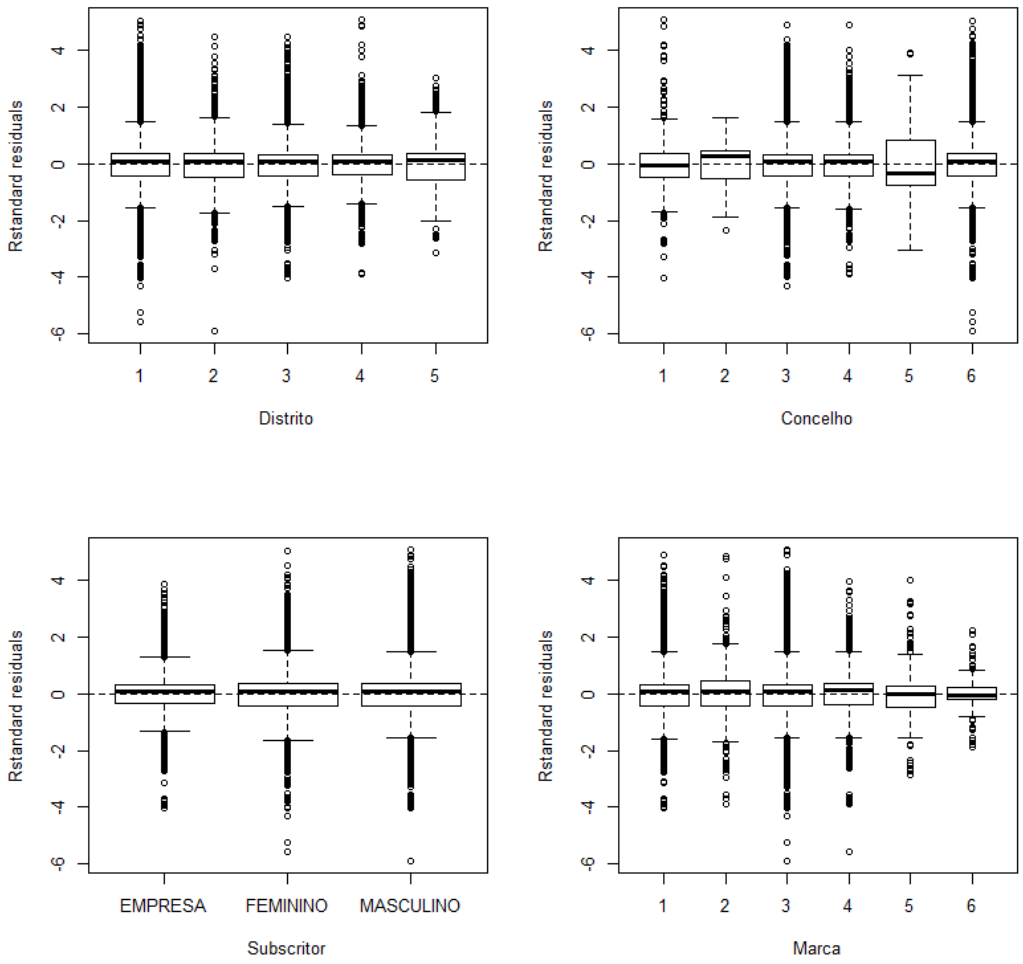

Figura 4.5: Regressão Linear com variável resposta logaritmizada - Resíduos padronizados

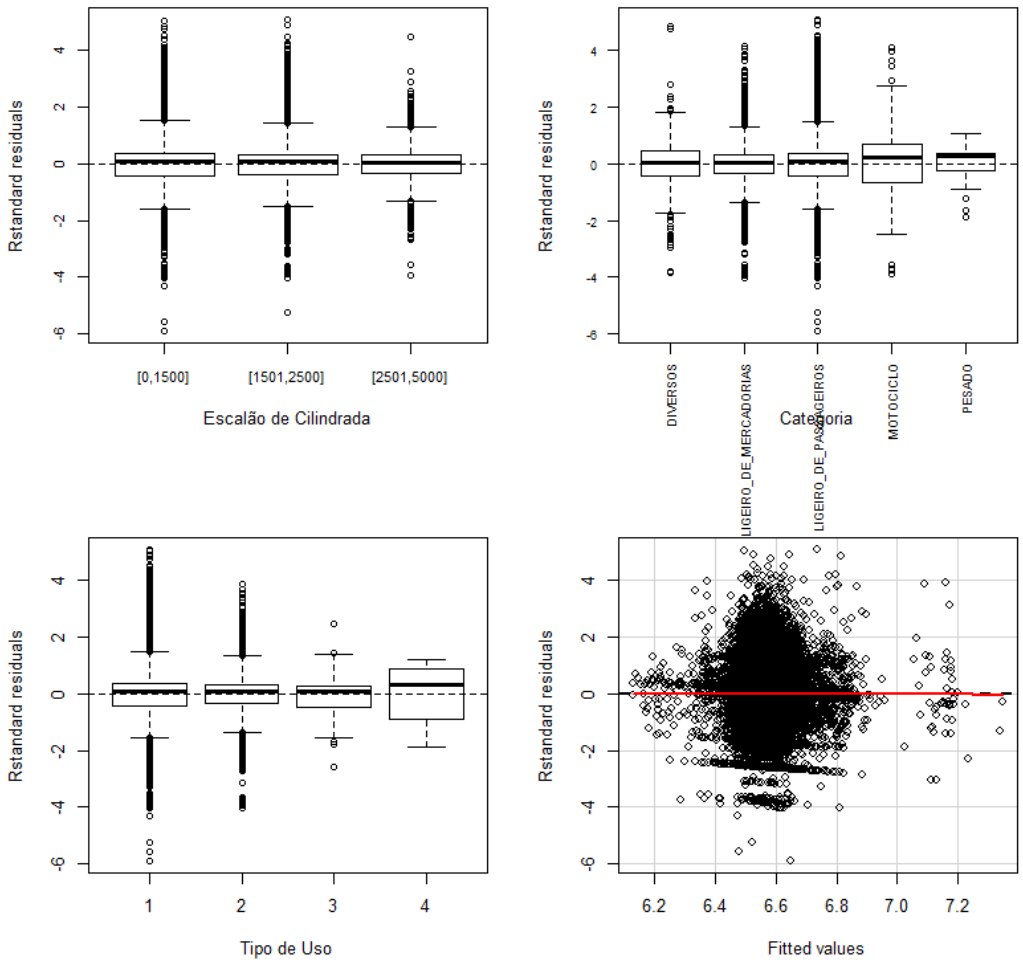

Figura 4.6: Regressão Linear com variável resposta logaritmizada - Resíduos padronizados

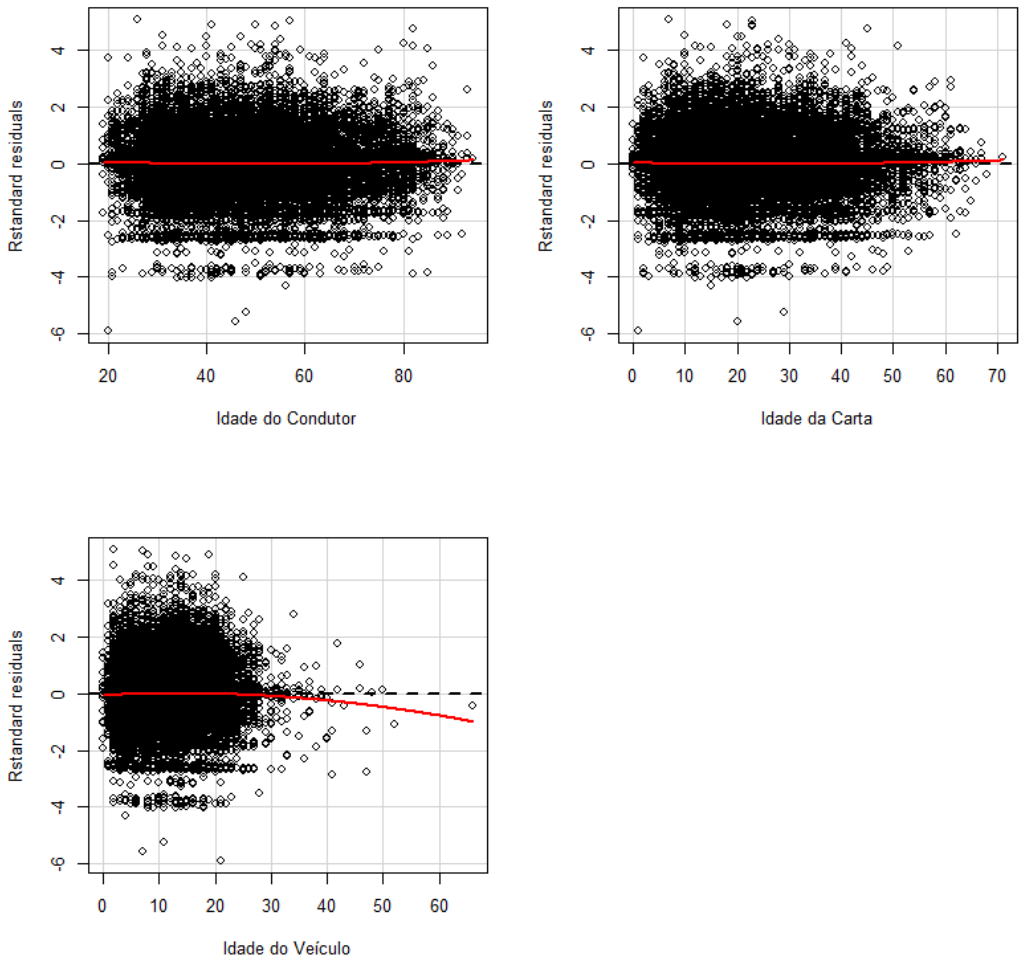

Figura 4.7: Regressão Linear com variável resposta logaritmizada - Resíduos padronizados

Capítulo 5

Conclusões

Como já foi referido anteriormente, o cálculo do prémio de risco será um valor inicialmente calculado, antes da apresentação do valor final ao tomador de seguro, com base nas características do veículo e tomador de seguro.

À análise descritiva destas variáveis seguiu-se a construção de um modelo para o prémio de risco considerando-se que este deve resultar da conjunção de duas componentes: a frequência de sinistralidade e o custo médio por sinistro. Para cada uma das componentes do modelo foram construídos diversos modelos da classe dos Modelos Lineares Generalizados, tendo-se concluído que a melhor solução resulta da combinação de um modelo de regressão de Poisson para a frequência e de um modelo de regressão linear com variável resposta logaritmizada para o custo por sinistro.

Os modelos estimados permitiram a construção de uma fórmula de cálculo do prémio de risco do tipo multiplicativo de fácil implementação e utilização. Para o seu uso, em termos operacionais, bastará que sejam recolhidos sobre o tomador de seguro e o veículo as informações seguintes: distrito, concelho, subscritor (se é particular ou se se trata de uma pequena empresa), idade do condutor, idade de carta, idade do veículo, marca do veículo, escalão de cilindrada do veículo, categoria do veículo (se é ligeiro de passageiros, de mercadorias, etc.) e o seu tipo de uso (para que efeito será utilizado o veículo).

O facto de existirem sinistros de valores muito elevados em localidades com muito poucas apólices subscritas fez com que houvesse uma inflação nos custos muito grande e assim existissem *clusters* que não eram tão esperados. De qualquer das formas, foi possível concluir que nas grandes cidades, tais como, Lisboa, Porto e Braga, os sinistros, em geral, correspondem a custos bastante baixos, ainda que sejam as cidades com mais frequência de sinistralidade, como era esperado pelo grande número de apólices subscritas. Estes custos mais baixos que em outras localidades com menos apólices subscritas podem ser explicados por haver um maior número de veículos em circulação e estes serem obrigados a circular mais lentamente.

Outra conclusão interessante é perceber que os condutores do sexo feminino contribuem para a frequência de sinistralidade negativamente, mas que contribuem positivamente para os custos com sinistros. São as mulheres, em maior número que dão os chamados "toques de cidade".

Este estudo teve como finalidade, entender o impacto que as várias características do veículo e do tomador de seguro podem tomar no cálculo de uma tarifa e de como cada uma delas é importante para determinar o prémio de risco.

Bibliografia

Antunes, Marília (2010). *Texto de apoio à disciplina de CRM e Prospecção de Dados.* Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa.

Cadima, Jorge (2010). *Apontamentos de Estatística Multivariada.* Instituto Superior de Agronomia, Universidade Técnica de Lisboa, 117-140.

Denuit, Michel (2011-2012). *Introduction: Overview of GLMs and of their Actuarial Applications.* Louvain School of Statistics, Biostatistics and Actuarial Science (LSBA), UCL, Belgium.

Everitt, Brian S., Landau, Sabine, Leese, Morven e Stahl, Daniel (2011). *Custer Analysis.* 5th Edition. John Wiley & Sons, Ltd., UK.

Faraway, Julian J. (2006). *Extending the Linear Model with R - Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Chapman and Hall/CRC (Taylor and Francis Group), Boca Raton.

Fox, John (2002). *An R and S-PLUS Companion to Applied Regression.* 1st edition, SAGE Publications, UK.

Fox, John (2008). *Applied Regression Analysis and Generalized Linear Models.* 2nd edition, SAGE Publications, UK.

Jespersen, Nicolai Schipper (2010). *Non-life insurance risk models under inflation. Master Thesis.* Copenhagen Business School.

Jong, Piet and Heller, Gillian Z. (2008). *Generalized Linear Models for Insurance Data.* Cambridge University Press, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd edition, Chapman & Hall, UK.

Murphy, Karl P., Brockman, Michael J. and Lee, Peter K. W. (2000). Using Generalized Linear Models to Build Dynamic Pricing Systems. *Article*, 107-140.

Santos, Susete Tomás (2008). *Construção de uma Tarifa de Responsabilidade Civil Automóvel. Tese de Mestrado.* Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa.

Turkman, Maria Antónia A. e Silva, Giovani L. (2000). *Modelos Lineares Generalizados - da teoria à prática.* Lisboa.

Werner, Geoff and Guven, Serhat (2007). GLM Basic Modeling: Avoiding Common Pitfalls. *Article of the Casualty Actuarial Society Forum, Winter 2007*, 257-272.

Wikilivros (2013). *Logística/Técnicas de previsão/Medidas de precisão da previsão.*

Yoo, Jong H. Introducing the Generalized Linear Models. Korea. *Article*, 408-424.

(http://www.actuaries.org/EVENTS/Seminars/EAAC_Bali/24%20(408-424)%20JongHwan_Yoo.pdf)

Capítulo 6

Anexos

6.0.1 Tabelas e Figuras

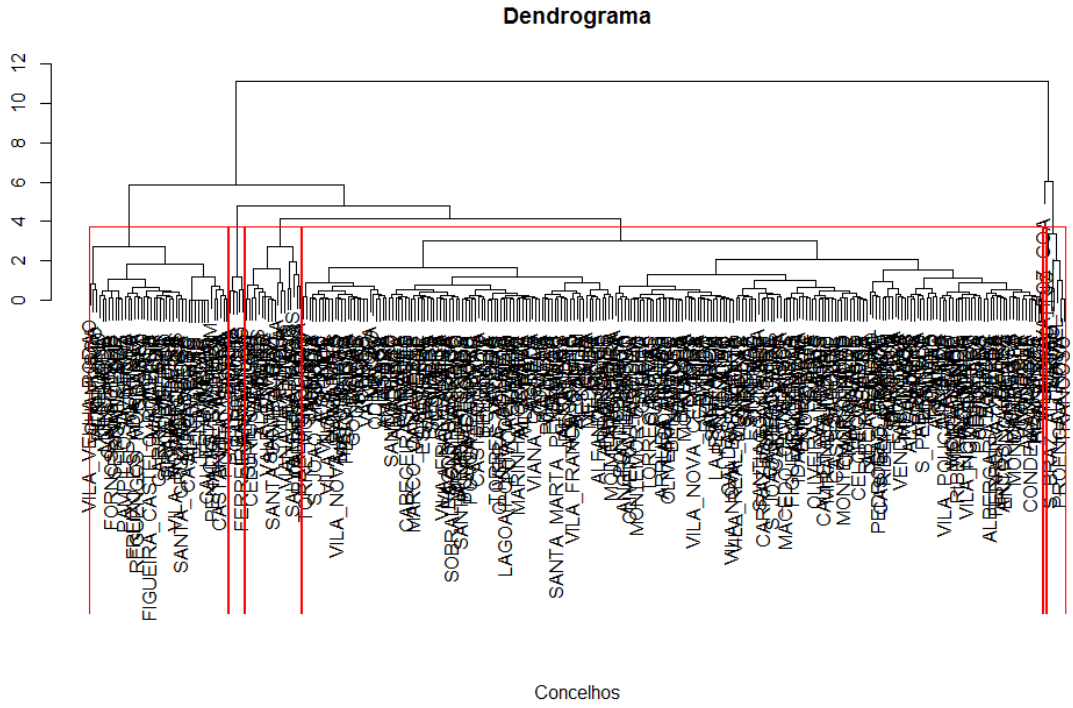

Figura 6.1: Agregação de concelhos por custo médio por sinistro e frequência média de sinistralidade em cada concelho

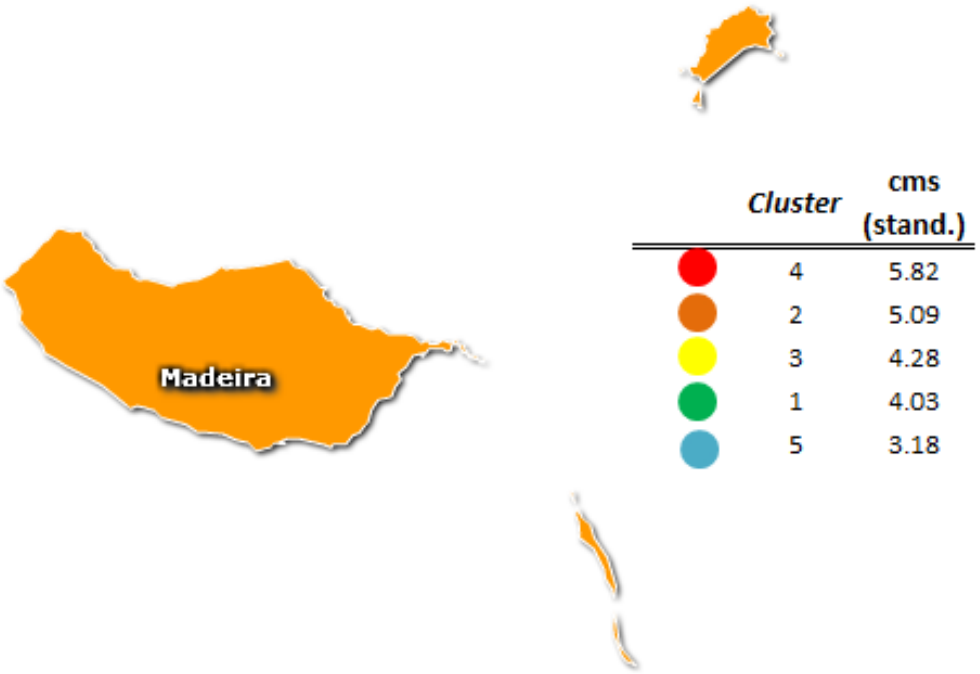

Figura 6.2: Distrito (arquipélago da Madeira) – Custo médio por sinistro (variável standardizada)

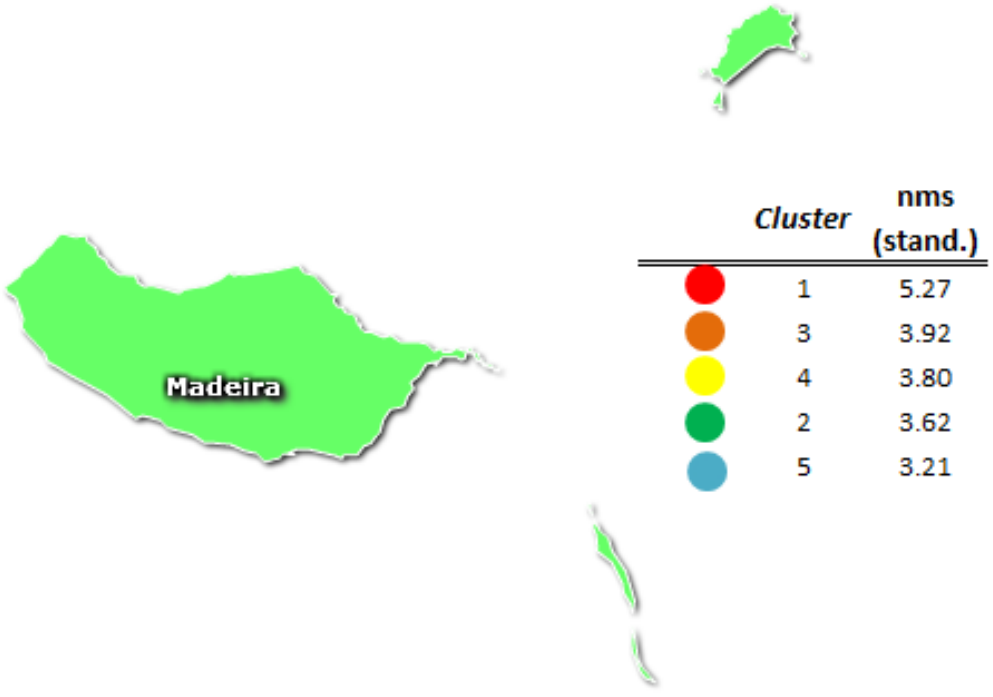

Figura 6.3: Distrito (arquipélago da Madeira) - Frequência média de sinistralidade (variável standardizada)

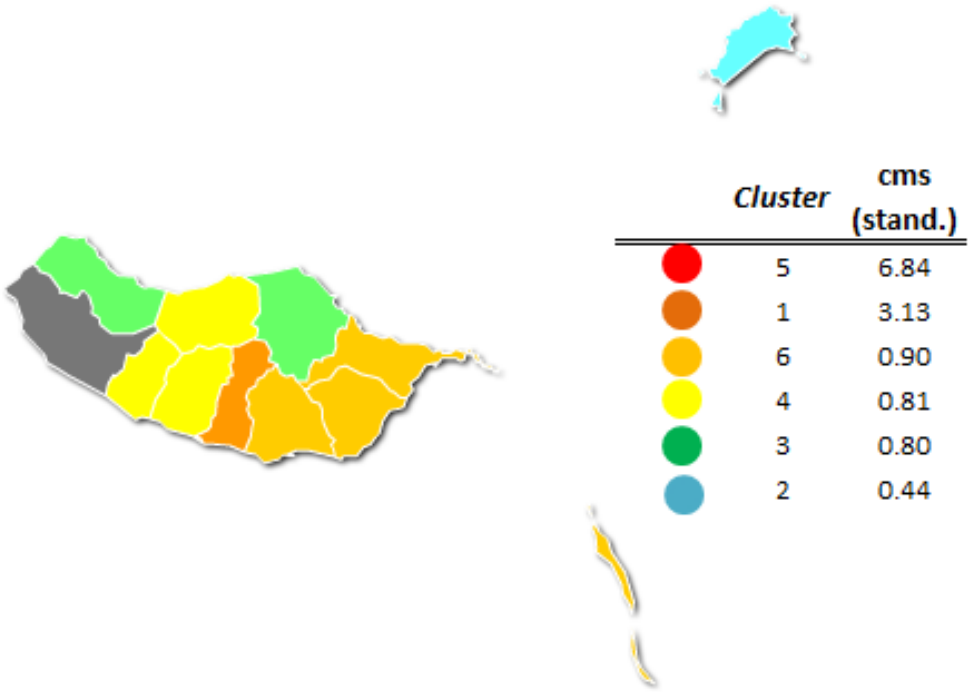

Figura 6.4: Concelhos (arquipélago da Madeira) – Custo médio por sinistro (variável standardizada)

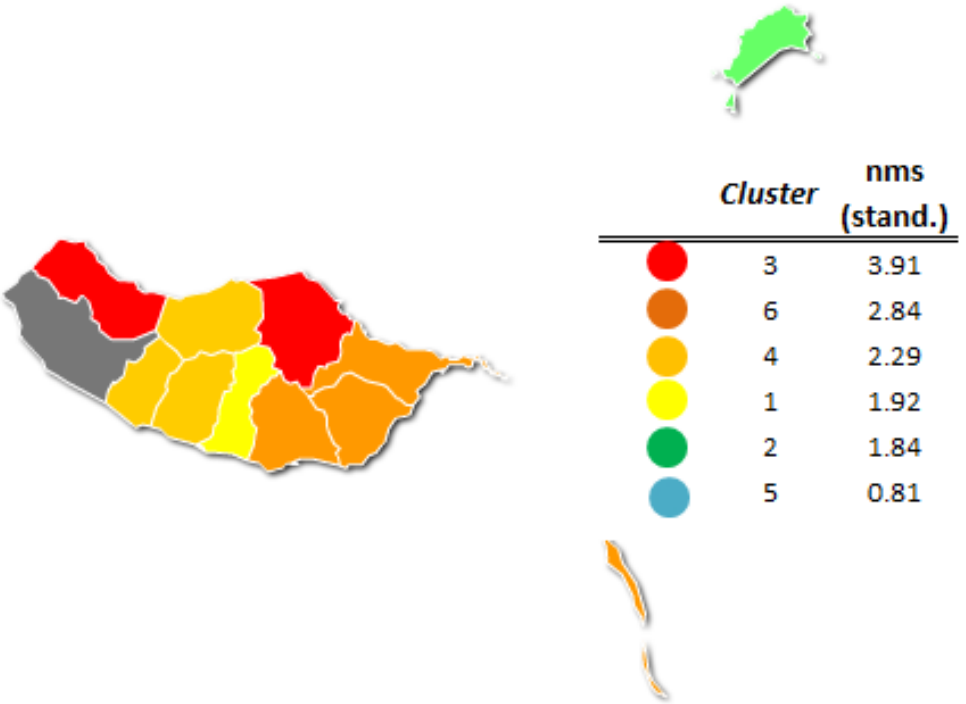

Figura 6.5: Concelhos (arquipélago da Madeira) - Frequência média de sinistralidade (variável standardizada)

Figura 6.6: Distrito (arquipélago dos Açores) - Custo médio por sinistro (variável standardizada)

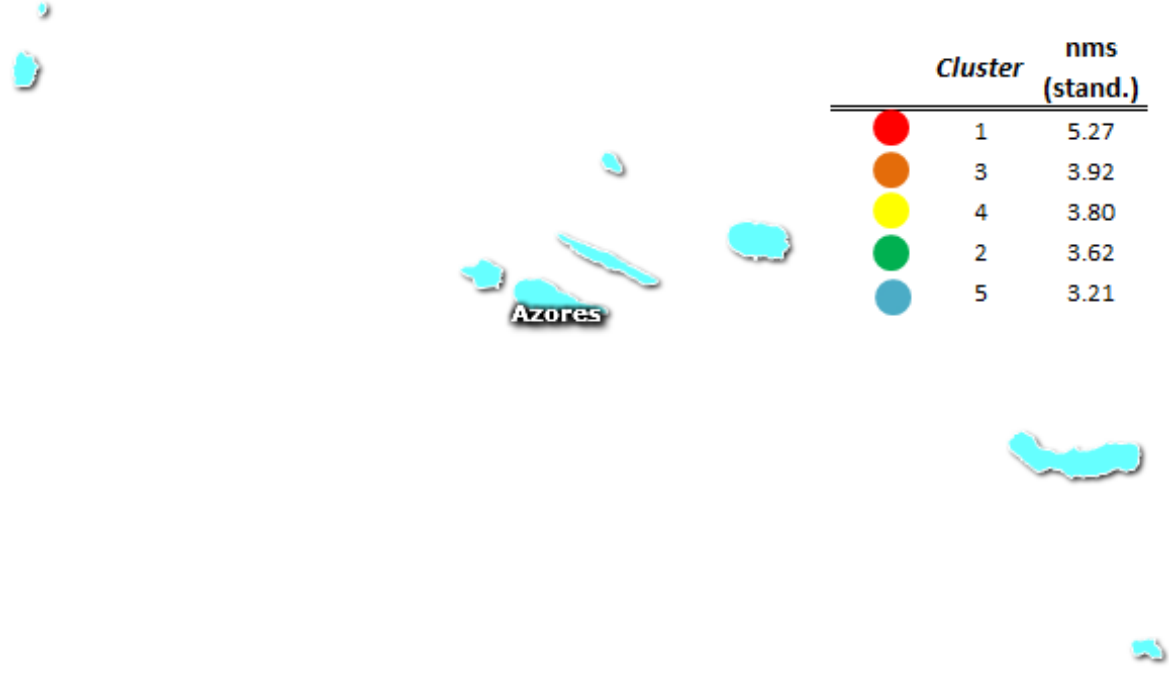

Figura 6.7: Distrito (arquipélago dos Açores) - Frequência média de sinistralidade (variável standardizada)

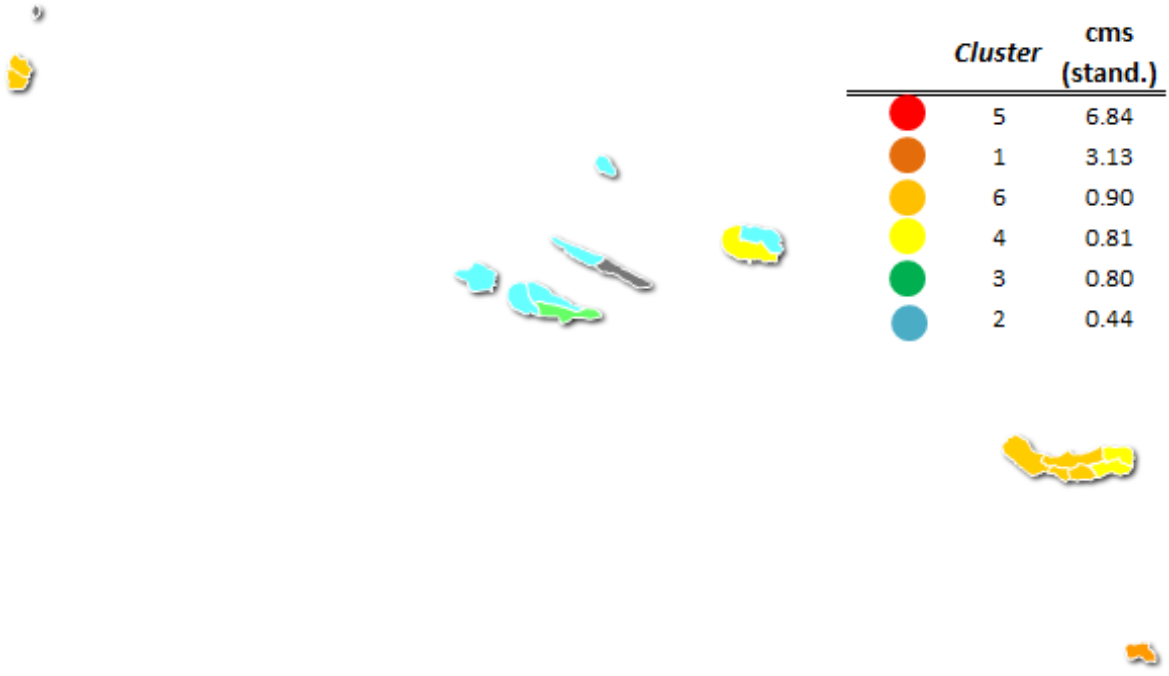

Figura 6.8: Concelhos (arquipélago dos Açores) – Custo médio por sinistro (variável standardizada)

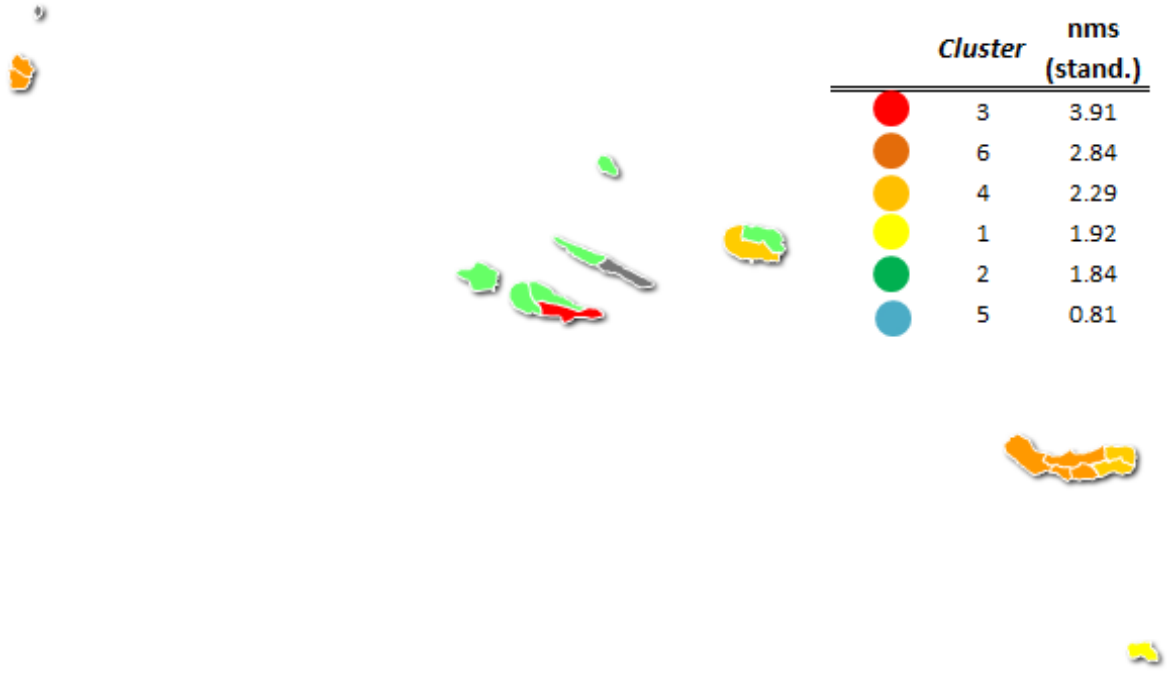

Figura 6.9: Concelhos (arquipélago dos Açores) - Frequência média de sinistralidade (variável standardizada)

Classe 1	Câmara de Lobos, Chamusca, Fundão, Mirandela, Moita, Paredes de Coura, Pedrógão Grande, Pombal, Proença-a-Nova, Salvaterra de Magos, Tabuaço, Tomar, Trancoso, Viana do Alentejo, Vila do Porto.
Classe 2	Aguiar da Beira, Alter do Chão, Alvito, Calheta (R.A.A.), Carregal do Sal, Castanheira de Pera, Castro Marim, Castro Verde, Constância, Crato, Figueira de Castelo Rodrigo, Freixo de Espada à Cinta, Horta, Madalena, Marvão, Meda, Mértola, Mogadouro, Monchique, Monforte, Moura, Pampilhosa da Serra, Penalva do Castelo, Penamacor, Penedono, Porto Santo, Praia da Vitória, Reguengos de Monsaraz, Santa Cruz da Graciosa, São Roque do Pico, Serpa, Sousel, Velas, Vidigueira, Vila Flor, Vila Nova de Poiares, Vila Real de Santo António, Vila Velha de Ródão, Vila Viçosa, Vimioso.
Classe 3	Almada, Amadora, Barrancos, Borba, Braga, Cascais, Coimbra, Espinho, Felgueiras, Ferreira do Alentejo, Gondomar, Guimarães, Lajes do Pico, Lisboa, Loures, Mação, Maia, Matosinhos, Moimenta da Beira, Mourão, Odivelas, Oeiras, Peso da Régua, Porto, Porto Moniz, São João da Madeira, Santana, Seixal, Sintra, Torre de Moncorvo, Trofa, Valongo, Vila do Conde, Vila Nova de Famalicão, Vila Nova de Gaia.
Classe 4	Abrantes, Alandroal, Albergaria-a-Velha, Alcácer do Sal, Alijó, Aljustrel, Almeida, Almeirim, Almodôvar, Alpiarça, Angra do Heroísmo, Ansião, Arcos de Valdevez, Armamar, Arouca, Avis, Beja, Bombarral, Boticas, Bragança, Cadaval, Caldas da Rainha, Calheta (R.A.M.), Caminha, Campo Maior, Cantanhede, Carrazeda de Ansiães, Castelo Branco, Castelo de Vide, Castro Daire, Celorico da Beira, Condeixa-a-Nova, Coruche, Entroncamento, Estarreja, Estremoz, Ferreira do Zêzere, Figueiró dos Vinhos, Fornos de Algodres, Fronteira, Gavião, Góis, Golegã, Grândola, Idanha-a-Nova, Loulé, Macedo de Cavaleiros, Mangualde, Manteigas, Mealhada, Mira, Miranda do Douro, Monção, Mondim de Basto, Montemor-o-Novo, Montemor-o-Velho, Mortágua, Murça, Murtosa, Nazaré, Nelas, Nisa, Nordeste, Óbidos, Odemira, Oleiros, Oliveira de Frades, Oliveira do Hospital, Ourique, Penela, Pinhel, Ponta do Sol, Ponte da Barca, Ponte de Lima, Ponte de Sor, Portalegre, Povoação, Redondo, Ribeira Brava, Ribeira de Pena, São João da Pesqueira, São Vicente, Sabrosa, Sabugal, São Pedro do Sul, Santiago do Cacém, Sardoal, Sernancelhe, Sertã, Silves, Soure, Tarouca, Tavira, Torres Novas, Vagos, Valpaços, Vendas Novas, Vila do Bispo, Vila Nova da Barquinha, Vila Nova de Cerveira, Vinhais, Viseu, Vouzela.
Classe 5	Mesão Frio, Portel, São Brás de Alportel, Sátão, Vila Nova de Foz Côa.
Classe 6	Águeda, Albufeira, Alcanena, Alcobaça, Alcochete, Alcoutim, Alenquer, Alfândega da Fé, Aljezur, Alvaiázere, Amarante, Amares, Anadia, Arganil, Arraiolos, Arronches, Arruda dos Vinhos, Aveiro, Azambuja, Baião, Barcelos, Barreiro, Batalha, Belmonte, Benavente, Cabeceiras de Basto, Cartaxo, Castelo de Paiva, Celorico de Basto, Chaves, Cinfães, Covilhã, Cuba, Elvas, Esposende, Évora, Fafe, Faro, Figueira da Foz, Funchal, Gouveia, Guarda, Ílhavo, Lagoa (Faro), Lagoa, Lagos, Lajes das Flores, Lamego, Leiria, Lourinhã, Lousã, Lousada, Machico, Mafra, Marco de Canaveses, Marinha Grande, Melgaço, Miranda do Corvo, Montalegre, Montijo, Mora, Olhão, Oliveira de Azeméis, Oliveira do Bairro, Ourém, Ovar, Paços de Ferreira, Palmela, Paredes, Penacova, Penafiel, Peniche, Ponta Delgada, Portimão, Porto de Mós, Póvoa de Lanhoso, Póvoa de Varzim, Resende, Ribeira Grande, Rio Maior, Santa Comba Dão, Santa Cruz, Santa Cruz das Flores, Santa Maria da Feira, Santa Marta de Penaguião, Santarém, Santo Tirso, Seia, Sesimbra, Setúbal, Sever do Vouga, Sines, Sobral de Monte Agraço, Tábua, Terras de Bouro, Tondela, Torres Vedras, Vale de Cambra, Valença, Viana do Castelo, Vieira do Minho, Vila Franca do Campo, Vila Franca de Xira, Vila Nova de Paiva, Vila Pouca de Aguiar, Vila Real, Vila de Rei, Vila Verde, Vizela.

Tabela 6.1: Agregação final de concelhos

Classe 1	Hyundai
	Lancia
	Mercedes-Benz
	Seat
	Smart
	Volkswagen
Classe 2	Diversos
	Luxo
Classe 3	Alfa Romeo
	Audi
	BMW
	Chevrolet
	Citroen
	Dacia
	Fiat
	Ford
	Jeep
	Land Rover
	Mitsubishi
	Nissan
	Opel
	Peugeot
	Renault
	Skoda
	Toyota
	Volvo
Classe 4	Honda
	Kia
	Peq. Quant.
	Suzuki
Classe 5	Mazda
	Mini
Classe 6	Pesados

Tabela 6.2: Agregação final das marcas

		Regressão *Hurdle* Poisson		e^{β_j}
	β_0		(Intercept)	0.1221 ***
Categoria		Diversos	categoria_agregDIVERSOS	1.0000
	β_1	Lig. de Mercad.	categoria_agregLIGEIRO_DE_MERCADORIAS	1.4769 .
	β_2	Lig. de Passag.	categoria_agregLIGEIRO_DE_PASSAGEIROS	1.1664
	β_3	Motociclo	categoria_agregMOTOCICLO	0.6134
	β_4	Pesado	categoria_agregPESADO	1.0568
Distrito		Classe 1	distrito_agreg11	1.0000
	β_5	Classe 2	distrito_agreg12	1.0110
	β_6	Classe 3	distrito_agreg13	0.8723 *
	β_7	Classe 4	distrito_agreg14	0.9467
	β_8	Classe 5	distrito_agreg15	1.0259
Concelho		Classe 1	concelho_agreg11	1.0000
	β_9	Classe 2	concelho_agreg12	0.2564 .
	β_{10}	Classe 3	concelho_agreg13	0.0614
	β_{11}	Classe 4	concelho_agreg14	0.7962
	β_{12}	Classe 5	concelho_agreg15	0.8407
	β_{13}	Classe 6	concelho_agreg16	0.9451
Escalão de Cilindrada		Escalão 1	escalao_cilindrada[0,1500]	1.0000
	β_{14}	Escalão 2	escalao_cilindrada[1501,2500]	1.0345
	β_{15}	Escalão 3	escalao_cilindrada[2501,5000]	1.4530 ***
Idades	β_{16}	Carta	idade_carta	0.9927 **
	β_{17}	Condutor	idade_condutor	1.0029
	β_{18}	Veículo	idade_veículo	1.0097 **
Marca		Classe 1	marca_agreg1	1.0000
	β_{19}	Classe 2	marca_agreg2	0.8171
	β_{20}	Classe 3	marca_agreg3	1.0631
	β_{21}	Classe 4	marca_agreg4	1.0607
	β_{22}	Classe 5	marca_agreg5	0.9998
	β_{23}	Classe 6	marca_agreg6	1.0872
Subscritor		Empresa	subscritorEMPRESA	1.0000
	β_{24}	Feminino	subscritorFEMININO	0.7474
	β_{25}	Masculino	subscritorMASCULINO	0.8074
Tipo de Uso		Classe 1	tipo_uso_agreg1	1.0000
	β_{26}	Classe 2	tipo_uso_agreg2	0.9852
	β_{27}	Classe 3	tipo_uso_agreg3	2.6571 ***
	β_{28}	Classe 4	tipo_uso_agreg4	1.0714

Tabela 6.3: Coeficientes multiplicativos para o cálculo do prémio de risco - Exponenciais dos β's estimados para a frequência de sinistralidade

		Regressão *Hurdle* Binomial Negativa		e^{β_j}
	β_0		(Intercept)	0.0116 ***
Categoria		Diversos	categoria_agregDIVERSOS	1.0000
	β_1	Lig. de Mercad.	categoria_agregLIGEIRO_DE_MERCADORIAS	1.5542 .
	β_2	Lig. de Passag.	categoria_agregLIGEIRO_DE_PASSAGEIROS	1.2022
	β_3	Motociclo	categoria_agregMOTOCICLO	0.6069
	β_4	Pesado	categoria_agregPESADO	1.0514
Distrito		Classe 1	distrito_agreg11	1.0000
	β_5	Classe 2	distrito_agreg12	0.9883
	β_6	Classe 3	distrito_agreg13	0.8581 *
	β_7	Classe 4	distrito_agreg14	0.9285
	β_8	Classe 5	distrito_agreg15	1.0220
Concelho		Classe 1	concelho_agreg11	1.0000
	β_9	Classe 2	concelho_agreg12	0.2402 .
	β_{10}	Classe 3	concelho_agreg13	1.0956
	β_{11}	Classe 4	concelho_agreg14	0.8212
	β_{12}	Classe 5	concelho_agreg15	0.8802
	β_{13}	Classe 6	concelho_agreg16	0.9629
Escalão de Cilindrada		Escalão 1	escalao_cilindrada[0,1500]	1.0000
	β_{14}	Escalão 2	escalao_cilindrada[1501,2500]	1.0341
	β_{15}	Escalão 3	escalao_cilindrada[2501,5000]	1.5061 ***
Idades	β_{16}	Carta	idade_carta	0.9918 **
	β_{17}	Condutor	idade_condutor	1.0032
	β_{18}	Veículo	idade_veículo	1.0124 ***
Marca		Classe 1	marca_agreg1	1.0000
	β_{19}	Classe 2	marca_agreg2	0.7984
	β_{20}	Classe 3	marca_agreg3	1.0684
	β_{21}	Classe 4	marca_agreg4	1.0543
	β_{22}	Classe 5	marca_agreg5	1.0134
	β_{23}	Classe 6	marca_agreg6	1.1254
Subscritor		Empresa	subscritorEMPRESA	1.0000
	β_{24}	Feminino	subscritorFEMININO	0.7277
	β_{25}	Masculino	subscritorMASCULINO	0.7927
Tipo de Uso		Classe 1	tipo_uso_agreg1	1.0000
	β_{26}	Classe 2	tipo_uso_agreg2	0.9946
	β_{27}	Classe 3	tipo_uso_agreg3	2.7522 **
	β_{28}	Classe 4	tipo_uso_agreg4	0.9604

Tabela 6.4: Coeficientes multiplicativos para o cálculo do prémio de risco - Exponenciais dos β's estimados para a frequência de sinistralidade

		Regressão Linear		γ_j
	γ_0		(Intercept)	$7811.9005 ***$
Categoria		Diversos	categoria_agregDIVERSOS	0.0000
	γ_1	Lig. de Mercad.	categoria_agregLIGEIRO_DE_MERCADORIAS	$-1590.8741 **$
	γ_2	Lig. de Passag.	categoria_agregLIGEIRO_DE_PASSAGEIROS	$-1370,5535 *$
	γ_3	Motociclo	categoria_agregMOTOCICLO	-393.6239
	γ_4	Pesado	categoria_agregPESADO	-2215.3646
Distrito		Classe 1	distrito_agreg11	0.0000
	γ_5	Classe 2	distrito_agreg12	29.9608
	γ_6	Classe 3	distrito_agreg13	-90.8861
	γ_7	Classe 4	distrito_agreg14	172.7046
	γ_8	Classe 5	distrito_agreg15	-431.1798
Concelho		Classe 1	concelho_agreg11	0.0000
	γ_9	Classe 2	concelho_agreg12	$-4700.3634 ***$
	γ_{10}	Classe 3	concelho_agreg13	$-3939.0325 ***$
	γ_{11}	Classe 4	concelho_agreg14	$-3788.8360 ***$
	γ_{12}	Classe 5	concelho_agreg15	$4257.4605 ***$
	γ_{13}	Classe 6	concelho_agreg16	$-3684.3409 ***$
Escalão de Cilindrada		Escalão 1	escalao_cilindrada[0,1500]	0.0000
	γ_{14}	Escalão 2	escalao_cilindrada[1501,2500]	114.7594
	γ_{15}	Escalão 3	escalao_cilindrada[2501,5000]	-168.3825
Idades	γ_{16}	Carta	idade_carta	-1.6151
	γ_{17}	Condutor	idade_condutor	-0.6290
	γ_{18}	Veículo	idade_veículo	-13.7966
Marca		Classe 1	marca_agreg1	0.0000
	γ_{19}	Classe 2	marca_agreg2	496.2416
	γ_{20}	Classe 3	marca_agreg3	$-229.9759 \cdot$
	γ_{21}	Classe 4	marca_agreg4	$-452.6418 \cdot$
	γ_{22}	Classe 5	marca_agreg5	$1198.2530 *$
	γ_{23}	Classe 6	marca_agreg6	-249.1228
Subscritor		Empresa	subscritorEMPRESA	0.0000
	γ_{24}	Feminino	subscritorFEMININO	-664.5154
	γ_{25}	Masculino	subscritorMASCULINO	-443.9313
Tipo de Uso		Classe 1	tipo_uso_agreg1	0.0000
	γ_{26}	Classe 2	tipo_uso_agreg2	-609.7323
	γ_{27}	Classe 3	tipo_uso_agreg3	-591.0860
	γ_{28}	Classe 4	tipo_uso_agreg4	556.3261

Tabela 6.5: Coeficientes aditivos para o cálculo do prémio de risco - γ's estimados para o custo por sinistro

		Regressão Gama		e^{γ_j}
	γ_0		(Intercept)	1.0501×10^4 ***
Categoria		Diversos	categoria_agregDIVERSOS	1.0000
	γ_1	Lig. de Mercad.	categoria_agregLIGEIRO_DE_MERCADORIAS	0.6431
	γ_2	Lig. de Passag.	categoria_agregLIGEIRO_DE_PASSAGEIROS	0.7322
	γ_3	Motociclo	categoria_agregMOTOCICLO	1.3183
	γ_4	Pesado	categoria_agregPESADO	0.3386
Distrito		Classe 1	distrito_agreg11	1.0000
	γ_5	Classe 2	distrito_agreg12	1.0356
	γ_6	Classe 3	distrito_agreg13	0.9443
	γ_7	Classe 4	distrito_agreg14	0.9635
	γ_8	Classe 5	distrito_agreg15	0.7774 *
Concelho		Classe 1	concelho_agreg11	1.0000
	γ_9	Classe 2	concelho_agreg12	0.1656 ***
	γ_{10}	Classe 3	concelho_agreg13	0.2948 ***
	γ_{11}	Classe 4	concelho_agreg14	0.3363 ***
	γ_{12}	Classe 5	concelho_agreg15	1.8269
	γ_{13}	Classe 6	concelho_agreg16	0.3446 ***
Escalão de Cilindrada		Escalão 1	escalao_cilindrada[0,1500]	1.0000
	γ_{14}	Escalão 2	escalao_cilindrada[1501,2500]	1.0613
	γ_{15}	Escalão 3	escalao_cilindrada[2501,5000]	0.9546
Idades	γ_{16}	Carta	idade_carta	0.9968
	γ_{17}	Condutor	idade_condutor	1.0016
	γ_{18}	Veículo	idade_veículo	0.9943
Marca		Classe 1	marca_agreg1	1.0000
	γ_{19}	Classe 2	marca_agreg2	1.0749
	γ_{20}	Classe 3	marca_agreg3	0.8894 *
	γ_{21}	Classe 4	marca_agreg4	0.7781 *
	γ_{22}	Classe 5	marca_agreg5	1.6204 *
	γ_{23}	Classe 6	marca_agreg6	0.8571
Subscritor		Empresa	subscritorEMPRESA	1.0000
	γ_{24}	Feminino	subscritorFEMININO	0.7552
	γ_{25}	Masculino	subscritorMASCULINO	0.8352
Tipo de Uso		Classe 1	tipo_uso_agreg1	1.0000
	γ_{26}	Classe 2	tipo_uso_agreg2	0.8086
	γ_{27}	Classe 3	tipo_uso_agreg3	0.7514
	γ_{28}	Classe 4	tipo_uso_agreg4	1.3779

Tabela 6.6: Coeficientes multiplicativos para o cálculo do prémio de risco - Exponenciais dos γ's estimados para o custo por sinistro