LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

SINCE 1911

MESTRADO EM

CIÊNCIAS ACTUARIAIS

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTAGIO

TARIFAÇÃO EM NÃO-VIDA COM UM MLG, APLICAÇÃO PRÁTICA COM A FERRAMENTA R

FRANCISCO ALIER VALENTIM NOGUEIRA

OUTUBRO-2015

LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

SINCE 1911

MESTRADO EM
CIÊNCIAS ATUARIAIS

TRABALHO FINAL DE MESTRADO
RELATÓRIO DE ESTAGIO

TARIFAÇÃO EM NÃO-VIDA COM UM MLG, APLICAÇÃO PRÁTICA COM A FERRAMENTA R

FRANCISCO ALIER VALENTIM NOGUEIRA

ORIENTAÇÃO:

JOÃO ANDRADE E SILVA

SNEJINA MALINOVA

OUTUBRO-2015

Agradecimentos

Antes de mais gostaria de agradecer à Optimind Winter Portugal, por me terem dado a oportunidade de trabalhar na área de consultoria atuarial e me terem acolhido na empresa, de modo a poder concluir o meu mestrado.

Aos meus orientadores de estágio, João Andrade e Silva, por todo o apoio prestado, dúvidas, comentários e paciência e Snejina Malinova, por me ter acolhido na sua empresa e proporcionado uma experiência inicial dentro do mercado de trabalho. Deixo um especial agradecimento à minha colega Liliana Lemos por me ter ajudado no dia-a-dia durante o decurso do meu estágio.

Agradeço especialmente à minha família, nomeadamente à minha mãe, ao meu pai, à minha irmã e aos meus avós por terem sido fulcrais para a realização deste mestrado.

Por fim, agradeço aos meus colegas de mestrado e amigos pela sua dedicação, apoio e ajuda para a realização deste trabalho.

Resumo

No contexto atual de estagnação económica, de uma nova regulamentação prudencial e de um aumento geral da concorrência, as companhias de seguros visam reposicionar-se no mercado com o objetivo de aumentar a rentabilidade dos seus produtos, proporcionando ao mesmo tempo preços mais justos aos seus segurados. No entanto, o que acaba por acontecer a maioria das vezes é uma discrepância entre os sinistros reais e as previsões do modelo, originada pelo facto de os antigos coeficientes tarifários já não corresponderem à carteira em questão.

Este trabalho propõe métodos para fornecer apoio à decisão na seleção de variáveis e à escolha do modelo do prémio puro para uma companhia de seguros. Para este efeito será usada a teoria dos Modelos Lineares Generalizados (MLG) com recurso à ferramenta R e iremos aplicá-la a uma companhia de seguros Não-Vida. Os resultados finais foram analisados considerando esta metodologia, mas também tendo em conta uma perspetiva de rentabilidade da companhia.

Palavras-chave

Tarifação, Perfil de Risco, Modelos Lineares Generalizados, Família de Dispersão Exponencial, Família Tweedie, Variáveis Tarifárias, Frequência de Sinistros, Custo Médio dos Sinistros, Perda Agregada, Prémio Puro

Abstract

Under the current context of economic stagnation, of a new supervisory regime, and of a general increase in market competition, insurance companies aim to reposition themselves in the market with the goal of increasing their profitability, while providing fairer prices to their policyholders. However, what tends to happen most of the times is that there is a difference between the real claims and the predictions of the model used, originated by the fact that the old ratemaking variables are not adapted to the current portfolio.

The goal of this work is to propose methods to support business decision making of ratemaking variables and to the choice of a pure premium model of an insurance company. To this end, it will be used the theory behind Generalized Linear Models (GLM) by treating the data using the software tool R and applying it to a Non-Life insurance company. The final results will be commented using this methodology, but also by taking into account the profitability of the company.

Key Words

Ratemaking, Risk Profile, Generalized Linear Models, Exponential Family, Tweedie Family, Ratemaking coefficients, Claim Frequency, Claim Severity, Aggregated Losses, Pure Premium

Índice

Lista de tabelas

1. Introdução

Este Trabalho Final de Mestrado foi feito no âmbito da realização de um estágio curricular na empresa Optimind Winter Portugal, uma consultora de atuariado e de gestão de risco. Teve uma duração de três meses tendo o propósito de, para além da conclusão do mestrado em Ciências Atuariais no ISEG, dar início ao meu percurso profissional.

Durante o período de estágio, tive a oportunidade de tomar conhecimento de alguns dos outros projetos da empresa, tanto para o mercado português como para o mercado francês. Entre esses destaco a criação de uma ferramenta de tarifação para contratos de seguro de crédito à habitação para as garantias de morte, invalidez e incapacidade, que me permitiram obter um *know-how* maior sobre o ramo vida da atividade seguradora.

Para concluir o estágio, foi-me lançado o desafio de desenvolver um novo modelo tarifário de uma companhia de seguros Não-Vida para a sua carteira de contratos de seguro de motociclos. Para a realização desse mesmo objetivo, utilizei duas abordagens diferentes:

1) Adotando a abordagem clássica de tarifação *à priori* deste tipo de carteiras que passa pela construção de um Modelo Linear Generalizado (MLG) para duas variáveis diferentes: a frequência de sinistros e o custo médio dos sinistros;
2) Fazendo a modelização da perda agregada dos sinistros, recorrendo à família de distribuições tweedie.

Estas previsões serão feitas com base no perfil de risco dos segurados, a partir dos quais iremos criar um modelo de prémio puro a ser cobrado para cada grupo de risco homogéneo. Iremos finalizar com alguns exemplos de perfis de risco de segurados de modo a comparar o prémio puro a ser cobrado com base nas duas metodologias apresentadas.

2. Modelos : Metodologias

Neste capítulo serão apresentadas as bases teóricas e os passos necessários para a criação de um modelo linear generalizado que produza estimativas o mais próximo possíveis da realidade, ou seja, para definir o modelo que melhor se adequa à carteira em estudo.

2.1. Modelos e tarifação

A imposição de um tarifário que reflita bem os riscos associados a um determinado segurado constitui uma das bases mais importantes para as companhias de seguros. A abordagem clássica para a elaboração de um tarifário passa, essencialmente, pela segmentação *à priori* da carteira em grupos com um grau de risco similar e, a partir desse ponto, prever o comportamento de uma determinada variável de resposta com base nas características de risco dos segurados, previamente definidas. Dentro do seguro automóvel, as variáveis de resposta mais comuns são a frequência dos sinistros e o custo médio dos sinistros as quais, tradicionalmente, são modelizadas separadamente, dado que podem ter comportamentos diferentes.

É do nosso maior interesse construir um modelo que seja simples e que se enquadre, o mais próximo possível, na realidade. Após a construção do modelo, é possível estimar o prémio puro para cada grupo de risco homogéneo, combinando as estimativas da frequência de sinistros e do custo médio dos sinistros. Seguidamente é acrescido ao prémio puro uma margem de segurança para fazer face à aleatoriedade do risco e, por fim, são ainda adicionados os encargos administrativos e de gestão de modo a obter o prémio comercial. No entanto, para efeitos deste trabalho, iremos apenas focar-nos na estimação do prémio puro relativo a cada grupo de risco homogéneo.

Para as nossas previsões iremos construir, tal como foi referido, um Modelo Linear Generalizado (MLG) utilizando diversos fatores tarifários que reflitam a sinistralidade da nossa carteira.

2.2. Modelos Lineares Generalizados (MLG)

Esta metodologia encontra-se entre as mais populares para a criação de modelos tarifários, especialmente devido à sua simplicidade e à evolução das novas tecnologias, o que permitiu um aumento da eficiência e da facilidade de manuseamento com grandes quantidades de dados.

No decurso deste capítulo, iremos utilizar as abordagens definidas por MacCullagh & Nelder (1989) , assim como o livro desenvolvido por Kaas et al (2008).

2.2.1. Família de dispersão exponencial

A imposição de um MLG visa a alteração do pressuposto subjacente ao modelo de regressão linear clássico, i.e, a assunção de que as observações seguem uma distribuição normal, com uma função de ligação identidade[1].

Geralmente, um MLG dispõe de três características fundamentais:

1) As observações $Y_i, i = 1, \dots, n$, provêm de variáveis aleatórias independentes com uma função densidade pertencente à família de dispersão exponencial (ver 2.1);
2) As variáveis explicativas são conectadas através de um preditor linear $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \dots, n$ onde os β_j representam parâmetros a serem estimados, dependendo do número de parâmetros existentes p e x_{ij} o valor da variável j para a observação i;
3) Existe uma função de ligação que tem como propósito interligar o preditor linear com o valor esperado da variável de resposta através de $\eta_i = g(\mu_i)$ em que μ_i representa o valor esperado para a observação i.

[1] Informações sobre funções de ligação podem ser encontradas no capítulo 2.2.2.

Em relação à característica número 1, define-se que uma variável aleatória Y pertence à família de dispersão exponencial se a sua função densidade puder ser escrita sob a forma:

$$f(y;\theta;\phi) = exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y;\phi)\right\} \quad (2.1)$$

Onde:

- $a(\phi)$ é uma função não-negativa, diferenciável em $\mathbb{R}$;
- $b(\theta)$ é uma função diferenciável duas vezes em $\mathbb{R}$;
- $c(y;\phi)$ é uma função definida em $\mathbb{R}^2$;
- θ representa o parâmetro de escala;
- ϕ representa o parâmetro de dispersão;

Existem uma variedade de distribuições pertencentes à família de dispersão exponencial. Entre as distribuições mais conhecidas encontram-se as distribuições Poisson, Gama, Normal e a Binomial Negativa. A seleção da distribuição a usar vai, posteriormente, influenciar a estimação das previsões da variável de resposta assim como a estimação dos parâmetros de regressão. Por exemplo, se quisermos modelizar o número de sinistros, uma distribuição Poisson poderá provar ser a mais adequada dado que a variável tem uma natureza discreta, mas se quisermos modelizar a severidade dos sinistros, normalmente, uma distribuição Gama costuma mostrar-se adequada, dado que é uma variável contínua.

2.2.2. Função de ligação

Para o modelo tarifário, é necessária a escolha da função de ligação a ser usada. A função de ligação relaciona o preditor linear η com o valor esperado μ da variável de resposta Y e é escolhida tendo também presente a distribuição teórica adotada. Por exemplo, se a distribuição for Poisson, teremos que ter $\mu > 0$ pelo que a função de ligação escolhida terá que ter em conta esta restrição.

As distribuições pertencentes à família de dispersão exponencial têm uma função de ligação especial que ocorre quando $\theta = \eta$ onde θ é o parâmetro canónico definido

em (2.1). Estas funções de ligação são as chamadas funções de ligação canónicas. De seguida, encontram-se alguns exemplos de funções de ligação canónicas e as respetivas distribuições padrão associadas:

Nome	Função de Ligação	Distribuição associada
Identity Link	$g(\mu) = \mu$	Normal
Log Link	$g(\mu) = \ln(\mu)$	Poisson
Inverse Link	$g(\mu) = -\mu^{-1}$	Gama
Logit Link	$g(\mu) = \ln(\frac{\mu}{1-\mu})$	Binomial

Tabela 1 - Funções de Ligação.

O valor esperado de uma variável de resposta é calculado invertendo o respetivo preditor linear, ou seja, $\eta = g(\mu) <=> \mu = g^{-1}(\eta)$.

Para a criação dos modelos, recorreu-se a uma função de ligação "log" de modo a criar um modelo multiplicativo. Sendo assim, o valor esperado da nossa variável de resposta será calculado da seguinte forma:

$$\mu_i = \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\} \qquad (2.2)$$

Em que β_0 representa a estimativa do coeficiente de regressão base do modelo e x_{ij} o valor da variável *j* para a observação *i* ao qual é multiplicado pela estimativa do j-ésimo coeficiente β.

Dada uma determinada função de ligação, já é possível o cálculo do valor esperado da nossa variável de resposta mas, para esse efeito, é necessário estimar os parâmetros de regressão β_j. De modo a obter estimadores tão eficientes quanto possível, o método da máxima verosimilhança será usado. Este método consiste em maximizar o logaritmo da função densidade da família de densidade exponencial, que é definido por:

$$l(\beta, \phi) = \sum_{i=1}^{n} \ln f\,(y_i, \beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi) \right\} \qquad (2.3)$$

2.2.3. Métodos de seleção

O objetivo de qualquer estatístico é o de selecionar um modelo que consiga estimar valores o mais próximo possível da realidade e que produza uma boa aderência aos dados. Existem três métodos que podem ajudar na seleção de um modelo:

1) *Forward* – Parte do modelo nulo (em que consta apenas o coeficiente independente) e vai adicionando sequencialmente a variável explicativa que mais melhora o ajustamento do modelo aos dados.
2) *Backward* – Parte do modelo completo (em que constam todas as variáveis explicativas) e retira sequencialmente variáveis explicativas, uma a uma, de acordo com o ajustamento do novo modelo encontrado.
3) *Stepwise* – É uma mistura dos dois métodos descritos acima. Envolve a inclusão e a exclusão de variáveis e pode partir do modelo nulo ou do modelo completo. Em cada passo, avalia-se a adição/exclusão de uma nova variável ao modelo.

Para o decurso deste trabalho iremos utilizar o processo *backward,* de modo a selecionar as nossas variáveis tarifárias. Através deste método, iremos eliminar ou simplificar as variáveis de modo a produzir modelos encaixados, ou seja, modelos mais reduzidos (submodelos) que pertençam ao modelo completo. O nosso modelo ideal estará, portanto, entre o modelo nulo e o modelo completo.

Utilizando uma destas técnicas, comparamos os diferentes modelos através de uma análise dos desvios. Um desvio pode ser definido como uma medida de distância entre o modelo completo e o proposto modelo. Matematicamente é calculado usando a seguinte fórmula:

$$Dev = 2(l_{sat} - l) \quad , \tag{2.4}$$

onde l_{sat} representa o logaritmo da função de verosimilhança do modelo completo (modelo saturado) e l o logaritmo da função de verosimilhança do submodeloproposto. Calculando os desvios dos nossos submodelos, poderemos

encontrar o modelo mais adequado através de uma análise dos respetivos desvios residuais. A diferença dos desvios é simplesmente definida como:

$$\Delta Dev = Dev_{sub} - Dev \quad , \qquad (2.5)$$

em que Dev_{sub} representa o desvio do proposto submodelo e Dev o desvio do modelo atual.

Comparamos esta estatística com uma distribuição Chi-Quadrado com $m - q$ graus de liberdade, em que m e q representam os graus de liberdade do modelo atual e do submodelo proposto, respetivamente. Com base nesse resultado, iremos rejeitar ou não rejeitar a nossa hipótese nula. As hipóteses são as seguintes:

$H_0: \beta_j = 0, \forall\, j = 1, \dots, q$

$H_1: \beta_j \neq 0, \forall\, j = 1, \dots, q$

Em que q representa o número de variáveis que estão em análise e β_j os coeficientes de regressão da variável *j.*

Para perceber melhor este teste, considere o seguinte exemplo: suponha-se que temos um modelo inicial com p = 5 parâmetros e que o modelo seguinte tem apenas 3 variáveis, ou seja, foi eliminado um fator com 3 níveis (logo, 2 parâmetros) que eram β_4 e β_5. Neste caso, estamos a testar a hipótese nula $H_0: \beta_4 = \beta_5 = 0$ contra pelo menos um dos coeficientes ser diferente de zero. A rejeição da hipótese nula implica que a imposição do fator em análise é estatisticamente significativo para explicar o comportamento da nossa variável de resposta, pelo que o modelo inicial deve ser guardado. Caso não rejeitemos a hipótese nula, passamos para o novo modelo e eliminamos o fator.

Para realizar estes testes de hipótese, iremos produzir tabelas de desvios sob a seguinte forma:

Modelo	Df	Dev	ΔDf	Δ Dev	Pr(>Chi)
1					
...					
F					

Tabela 2 - Exemplo de uma tabela de desvios

Em que "Df" representa os graus de liberdade do modelo e "Δ Df" os graus de liberdade residuais do modelo. O objetivo será criar um modelo F cujas alterações façam rejeitar o teste de hipóteses descrito anteriormente, ou seja, o valor da última coluna da tabela 1.1, terá de ser superior a um nível de confiança α%.

2.2.4. Família Tweedie

Dentro da família das distribuições de dispersão exponencial, existe uma subclasse de famílias que tem ganho notoriedade dentro do ramo atuarial, especialmente para a modelização das perdas agregadas.

Uma variável pertencente à família de dispersão exponencial (definida na equação (2.1)) pertence à família tweedie se a sua função variância satisfizer:

$$V(\mu) = \mu^p \text{ com } p \in \mathbb{R} \qquad (2.6)$$

Em que o valor do parâmetro p é o que vai caracterizar a distribuição probabilística da variável em estudo. Por exemplo, se considerarmos p = 1 a distribuição representada é uma Poisson enquanto se considerarmos p = 2 a distribuição já será uma Gama.

A classe que nos vai ser de mais utilidade para este trabalho é a classe em que os valores p se encontram entre 1 e 2, que representam a família de distribuições Poisson composta. Este caso surge assumindo que o número de sinistros segue uma distribuição Poisson e que, por sua vez, o montante dos sinistros segue uma distribuição Gama, criando uma família de distribuições mista, ou seja, uma distribuição que não é puramente discreta nem puramente contínua.

As distribuições agregadas serão modelizadas através da seguinte fórmula:

$$S = \sum_{i=0}^{N} X_i \quad (2.7)$$

Em que N representa o número de sinistros e X_i o montante do i-ésimo sinistro.

Devido a esta característica, estes modelos são adequados para fazer uma modelização do prémio puro, sem ser necessária a análise separada da frequência e da severidade (Ohlsson & Johansson, 2010).

3. Tarifação à priori: Aplicação prática

Passaremos agora à vertente prática do trabalho, isto é, à análise de uma carteira de uma seguradora Não-Vida. Estes dados foram-me fornecidos pela empresa onde efetuei o meu estágio curricular, Optimind Winter Portugal, e provêm de uma companhia de seguros Não-Vida que se encontra sediada na Suécia. Apesar de a empresa dispor de diferentes linhas de negócio, iremos apenas analisar uma carteira de seguros para motociclos cujos sinistros ocorreram em diversas zonas na Suécia entre os períodos de 1994 e 1998. Esta análise será feita utilizando as metodologias que foram abordadas no capítulo anterior.

Primeiramente irão ser apresentadas e analisadas as variáveis disponíveis para a construção do tarifário. De seguida, vai-se proceder à seleção de três modelos diferentes: um modelo de previsão da frequência de sinistros, outro de previsão do custo médio dos sinistros e, por fim, um modelo de previsão da perda agregada dos sinistros. Para este efeito, a utilização de uma ferramenta é necessária. A escolhida foi a ferramenta R, um *open software* estatístico que irá ser fulcral para a realização do trabalho.

3.1. Descrição da carteira

A carteira em análise provém de um produto de seguros para motociclos de uma companhia de seguros, dentro do período de 1994 a 1998[2]. O número total de apólices é desconhecido mas dispomos de informação sobre o número de unidades de risco (que neste caso são o número de apólices por ano) para cada possível combinação dos fatores tarifários. É de notar que, em relação a montantes dos sinistros, a moeda usada é a moeda local da Suécia, representada em coroas suecas (SEK), mas, com o objetivo de facilitar a análise, os resultados finais serão convertidos para euros usando uma taxa de conversão de 1 EUR = 9,3 SEK[3]. Salienta-se que esta conversão não é a mais correta dado que os dados são de 1994-1998 e estamos a usar uma taxa de conversão do ano corrente.

3.1.1. Análise das variáveis tarifárias

As variáveis tarifárias existentes são as seguintes:

ID	Nome	Tipo de variável	Descrição
1	Gender	Categórica	Representa o género do segurado: - F (Feminino) - M (Masculino).
2	OwnersAge	Discreta	Representa a idade dos segurados, compreendida entre 16 e 99 anos.
3	Zone	Categórica	Representa a zona de ocorrência do sinistro. É subdividida em 7 categorias diferentes: 1 – Zonas centrais e semi-centrais das três maiores cidades na Suécia; 2 - Subúrbios e cidades de média dimensão;

[2] Estes dados são usados no livro "Non Life Insurance Pricing with Generalized Linear Models" de E. Ohlsson & B. Johansson, 2010
[3] A taxa de conversão usada foi à data de 12 de Outubro de 2015, disponível em: http://www.xe.com/currencyconverter/convert/?Amount=1&From=EUR&To=SEK#converter

			3 - Pequenas cidades, exceto as descritas nas classes 5 ou 7; 4 - Vilas e aldeias exceto as descritas nas classes 5 a 7; 5 - Vilas do norte; 6 - Aldeias do norte; 7 - Gotland (a maior ilha da Suécia).
4	Class	Categórica	Representa a categoria do veículo em relação ao denominado *EV ratio* definido como: (Potência do motor (kw)*100)/(Peso do veículo (kg) + 75). É subdividida em 7 classes diferentes: 1: EV≤5 2: 5 < EV ≤ 8 3: 8 < EV ≤ 12 4: 12 < EV ≤ 15 5: 15 < EV ≤ 19 6: 19 < EV ≤ 24 7: EV > 24
5	VehicleAge	Discreta	Representa a idade do motociclo segurado compreendida entre 1 e 99 anos.
6	BonusClass	Categórica	Representa o nível de bónus categorizado de 1 a 7. Um novo condutor começa com um bónus de classe 1; por cada ano sem sinistros a classe de bónus é aumentada em 1 até ao máximo de 7, representando o bónus mais elevado. Depois do primeiro sinistro o bónus decresce em 2; o condutor não pode voltar para a classe 7 com menos de 6 anos consecutivos sem sinistros.
7	Duration	Contínua	Representa a duração da apólice de seguro em anos, i.e., o número de anos em que o segurado se encontra exposto ao risco.
8	NumberClaims	Discreta	Número de sinistros reportados.
9	ClaimCost	Contínua	Representa o custo total dos sinistros registados.

Tabela 3 - Descritivo das variáveis explicativas

3.1.2. Tratamento dos dados

De modo a criar modelos adequados é importante realizar uma análise mais aprofundada dos dados da nossa carteira. No entanto, antes de proceder a qualquer tipo de análise, é importante referir que ao serem analisados os dados da carteira foram identificadas duas situações:

1) 4 casos em que existem sinistros registados mas o número de anos de exposição é igual a 0.
2) 2069 casos em que o número de anos de exposição é igual a 0 e não existem sinistros registados.

A situação 1 provém, provavelmente, de um erro de registo de sinistros pelo que irá originar conclusões falsas sobre a carteira caso não seja feito um tratamento adequado, que neste caso, será a exclusão destes para efeitos de análise. Para a situação 2, procedemos à eliminação dos 2069 casos em que a exposição é nula dado que, para efeitos de análise, não é lógico analisar unidades de risco que nunca estivaram expostas a qualquer tipo de risco. Durante este subcapítulo, apenas serão mostrados e comentados os resultados principais desta análise.

Começaremos por analisar as variáveis *Duration, NumberClaims* e *ClaimCost*. Com estas variáveis é possível analisar alguns rácios da indústria com os quais, no próximo capítulo, iremos criar os nossos modelos tarifários. Estes são calculados com as seguintes fórmulas:

$$FS = \frac{Número\ Sinistros}{Exposição}\ ; PP = \frac{Custo\ Sinistros}{Exposição}; CM = \frac{Custo\ Sinistros}{Número\ de\ Sinistros}$$

Em que *FS* representa a frequência dos sinistros, *PP* representa o Prémio Puro e *CM* representa o Custo Médio dos sinistros. Neste caso, a exposição é representada pela variável *Duration*.

Fazendo as somas e o rácio das variáveis com a *Duration*, obtêm-se os seguintes resultados globais:

Variáveis			
	Duration	NumberClaims	ClaimCost
Soma	65217,04	697	1.832.454€
Rácio	-	1,069%	28,10€

Tabela 4 - Somas globais

Podemos então verificar que na nossa carteira houve um total de 697 sinistros registados e que estes sinistros, na sua totalidade, tiveram um custo de 1.832.454 euros. A nossa carteira tem uma frequência de sinistros de 1,069% e um custo médio por ano de apólice de 28,10 euros, o que indica que ocorreram muito poucos sinistros durante o tempo em que as apólices se encontravam em vigor e que, quando efetivamente ocorreram, representaram um custo médio por ano de apólice de 28,10 euros.

Passaremos agora ao tratamento e análise de algumas variáveis que são essenciais para a modelização efetuada no capítulo seguinte. Informações relativamente aos rácios analisados das outras variáveis que não são referidas no corpo de texto poderão ser encontradas no Anexo 1.

Relativamente à variável *OwnersAge*, podemos verificar que esta está compreendida num intervalo e não se encontra dividida em classes. De modo a criar um modelo com esta variável, é de todo o nosso interesse em fazer uma divisão desta em classes de idades dos segurados. Para além do mais, ao fazer este agrupamento, torna-se mais fácil avaliar estas variáveis e efetuar previsões com menos volatilidade sobre a nossa variável de resposta.

Esta nova variável categórica será denominada de *GroupAge* e o agrupamento escolhido foi o seguinte:

Idade do segurado (*GroupAge*):

Classe 1: [16, 20]

Classe 2: [21, 29]

Classe 3: [30, 59]

Classe 4: [60, 92]

Este agrupamento é feito com base na divisão dos segurados em grupos de risco característicos e, neste caso, a divisão é feita de modo a separar os condutores de motociclos inexperientes (entre 16 e 20 anos); os condutores com alguma experiência e maturidade, sendo no entanto ainda considerados jovens condutores (entre 21 e 29 anos); os condutores experientes e já com uma idade adulta (entre 30 e 59 anos) e, por fim, os condutores mais velhos (com idades superiores ou iguais a 60 anos).

Verificando os rácios de indústria e unidades de risco por idade do segurado, obteve-se o seguinte resultado:

Por Idade do Segurado				
Classe	1	2	3	4
Unidades de Risco	1772	11808	43270	5586
Sinistros	42	286	336	29
FS	5,66%	5,23%	1,53%	0,84%
PP	72 €	159 €	39 €	12 €
CM	1.263 €	3.038 €	2.555 €	1.422 €

Tabela 5 - Rácios de indústria por idade do segurado

Como podemos ver na Tabela 5, a classe de idades que tem uma maior frequência de sinistralidade é a classe 1 com 5,66%, ou seja, a classe mais jovem e mais inexperiente é a que regista um maior número de acidentes dentro do período aos quais estiveram expostas ao risco, o que é algo lógico dado que ainda não têm muita experiência na condução de motociclos. No entanto, a frequência de sinistros da classe 2 regista apenas uma diferença de 0,43% comparativamente com a classe 1 o que significa que estas classes têm uma frequência de sinistralidade muito parecida,

mas se compararmos os custos médios, a classe que tem um custo médio mais elevado é a classe 2, ou seja, esta classe é a que tem acidentes com montantes mais elevados. Não obstante, parece que o número de unidades de risco divididos por estas classes é relevante para uma posterior modelização.

Outra variável que precisa de ser agrupada em classes é a variável *VehicleAge*. O agrupamento escolhido foi o seguinte:

Idade do veículo (VehicleAge):

Classe 1: [1, 5]

Classe 2: [6, 20]

Classe 3: [21, 99]

Como se pode verificar acima, o agrupamento é feito com base na vida útil dos motociclos em que temos, por ordem crescente de classes, um grupo constituído por motociclos até aos 5 anos de idade, outro constituído por motociclos até aos 20 anos de idade e, o último grupo, que é constituído por motociclos com idades superiores a 20 anos.

Fazendo a mesma análise feita à variável anterior, obteve-se a seguinte tabela:

Por Idade do Veiculo			
Classe	1	2	3
Unidades de Risco	16032	37958	8446
Sinistros	313	354	26
FS	4,24%	1,89%	0,47%
PP	164 €	30 €	10 €
CM	3.860 €	1.573 €	2.175 €

Tabela 6 - Rácios de indústria por idade do motociclo

Através da Tabela 6, verificamos que existe uma ordem decrescente de classe, em relação à frequência de sinistralidade, que é maior na classe dos motociclos mais recentes (entre 1 e 5 anos) e é menor na classe dos motociclos mais antigos (superior a 20 anos). Em relação ao custo médio este é mais elevado, tal como na frequência de sinistralidade, na classe 1.

Se considerarmos as nossas variáveis por zona de ocorrência do sinistro, obtemos os seguintes resultados:

Por Zona							
Classe	1	2	3	4	5	6	7
Unidades de Risco	8210	11390	12296	24183	2274	3716	367
Sinistros	182	166	122	195	9	18	1
FS	4,39%	2,88%	1,92%	1,69%	0,68%	0,82%	0,46%
PP	143 €	89 €	42 €	35 €	9 €	14 €	0,3 €
CM	3.257 €	3.096 €	2.212 €	2.065 €	1.251 €	1.721 €	70 €

Tabela 7 - Rácios da indústria por zona

Como se pode observar na Tabela 7, a zona 1 – zonas centrais e semi-centrais das três maiores cidades na Suécia, é a zona que tem uma maior frequência de sinistros, assim como o maior custo médio das 7 zonas em estudo. Relativamente à frequência de sinistros, este resultado é esperado dado que onde é mais provável a ocorrência de um sinistro de um motociclo é nas grandes áreas metropolitanas. Entre zonas parece existir um fator decrescente de risco dado que tanto a frequência como o custo médio vão decrescendo à medida que aumentamos a classe da zona. No entanto, esse padrão não é seguido na zona 6, pois verifica-se um aumento dos nossos rácios, comparativamente com a zona 5 (o prémio puro na zona 5 é de 9€ enquanto que na zona 6 é de 14€). Esta ocorrência pode ser explicada devido ao facto de a zona 5 ter um registo de 9 sinistros e a zona 6 ter um registo de 18 sinistros. A zona 7 é a zona em que existe uma menor frequência de sinistralidade, o custo médio é de apenas 70 euros, tem apenas 1 sinistro registado e o prémio puro é de apenas 30 cêntimos. De uma perspetiva atuarial, estes números para a classe 7 não fazem qualquer sentido já que são demasiado baixos para qualquer companhia implementar. Verifica-se que esta última classe tem uma representatividade estatística muito fraca.

Considerando os nossos rácios por classe de bónus, obtemos os seguintes resultados:

Por Bónus							
Classe	1	2	3	4	5	6	7
Unidades de Risco	13806	8534	6491	5813	4922	5203	17667
Sinistros	134	71	57	64	45	41	281
FS	1,98%	1,63%	1,72%	2,14%	1,70%	1,44%	3,24%
PP	46 €	40 €	57 €	67 €	53 €	49 €	77 €
CM	2.317 €	2.445 €	3.301 €	3.154 €	3.101 €	3.365 €	2.384 €

Tabela 8 - Rácios da indústria por classe de bónus

Podemos verificar que a classe onde existe uma maior frequência de sinistros é a classe 7, que representa a classe de bónus máximo deste sistema. Mas se considerarmos o custo médio, este é mais elevado na classe 6 e mais reduzido na classe 1. Entre todas as classes, a frequência de sinistros não regista uma grande variação encontrando-se entre 1,44% e 3,24% e, como podemos observar, este rácio regista valores bastante reduzidos dentro do sistema de *bónus-malus*. Por essa razão, face à carteira disponível, este sistema é inadequado, o que irá originar uma fraca aderência da variável explicativa às variáveis de resposta estudadas, como se vai poder verificar no capítulo seguinte.

3.2. Seleção do modelo

3.2.1. Modelização da frequência de sinistros

Começaremos então por definir o nosso primeiro modelo que irá prever a frequência de sinistros tendo em conta os grupos de risco dos segurados. É de reafirmar que o objetivo da criação dos modelos é de encontrar um modelo simples e que represente o melhor possível a realidade. Para o resto do trabalho, quando necessário, todos os testes de hipóteses serão feitos utilizando um nível de significância de 5% sendo que este é considerado como o mínimo aceite (Hair et al., 1998).

Para a criação dos modelos iremos utilizar a ferramenta R que, convenientemente, dispõe de diversas funções para a estimação e teste de modelos lineares generalizados. A função primária para a sua realização é a função descrita na seguinte forma:

```
>glm(variável_resposta ~ variáveis_explicativas +offset(log(Duration/365)),
family=quasipoisson (link = "log")))
```

Para além da função glm, iremos usar a seguinte função no R para produzir uma tabela de desvios:

```
>anova(Modelo1, Modelo2,…,ModeloN, test = "Chisq")
```

N representa o número de submodelos criados e o teste usado será o da Chi-Quadrado, como foi referenciado no Capítulo 2.

Utilizando as abordagens descritas em MacCullagh & Nelder (1989), de modo a prever a frequência de sinistros, iremos usar a distribuição de Poisson, mais especificamente, uma quasipoisson e uma função de ligação logarítmica, dado que estas, usualmente, descrevem bem a evolução desta variável de resposta. A imposição do elemento "quasi" serve para superar o problema de sobre dispersão, ou seja, quando a variância amostral é superior à média amostral, o que usualmente acontece. Outra particularidade para esta variável de resposta é a de ser necessário inserir uma variável denominada de *offset* que, sucintamente, representa o logaritmo da exposição ao risco e serve para tomar em consideração a existência de diferentes períodos temporais.

Como explicado no Capítulo 2, iremos partir do modelo completo e, a partir desse modelo, iremos retirar sequencialmente as nossas variáveis explicativas de modo a analisar a sua significância em termos do modelo. Para esse efeito a utilização da ferramenta R irá provar-se de bastante utilidade, já que realiza de uma forma simples e intuitiva os testes de hipóteses necessários para a avaliação da significância do modelo em questão.

Partindo então do modelo completo, obteve-se o *output* que poderá ser consultado no Anexo 2.

Como podemos observar no *output* do Anexo 2, o modelo completo, efetivamente, não é o mais adequado. Através de uma primeira análise, nota-se que os erros padrões – que representam o nível de incerteza associado às estimativas (Gelman & Hill, 2007) assumem, no geral, valores pequenos, mas olhando para o teste *t*, que o R convenientemente realiza, podemos ver que bastantes variáveis não são estatisticamente significativas a um nível de confiança de 5%. O elemento *intercept* que se observa nos *outputs* deste comando no R representa a estimativa da frequência de sinistros para um segurado que pertença aos níveis base dos fatores tarifários usados, ou seja, um segurado do género feminino que tenha menos de 20 anos, tenha um motociclo com menos de 5 anos que circule nas grandes cidades da Suécia, que tenha um rácio EV inferior a 5 e que pertença à classe de Bónus 1. Um segurado com estas características terá uma frequência de sinistros estimada de $e^{-2,24558}$= 10,58% (aproximadamente).

Analisando o *output* deste modelo, realizaremos sequencialmente as seguintes ações, testando-se em cada passo a significância estatística de cada uma delas:

- ✓ BonusClass: Eliminar a variável por esta não ser estatisticamente significativa;
- ✓ Gender: Eliminar a variável por esta não ser estatisticamente significativa. Devido à sua baixa frequência de sinistralidade entre classes (ver Anexo 1) este resultado não é surpreendente;
- ✓ Zone: Agregar as classes 6 e 7 com as classes 4 e 5, respetivamente, criando a variável *Zone2*. Esta agregação faz sentido dado que as estimativas dos coeficientes de regressão são bastante similares entre estes grupos de fatores. As classes 2 e 3 mantêm-se inalteradas devido à sua significância estatística;

- ✓ GroupAge: Agrupamento da classe 4 na classe 3 dado que as estimativas são similares e agrupamento da classe 2 na classe 1, criando a variável *GroupAge2*;
- ✓ Class: Agregar as classes 2, 5 e 7 na classe 1 e agregar a classe 4 na 3, criando a variável *Class2.1*. Estas agregações foram feitas com base nas estimativas observadas no modelo completo (as classes 3 e 4 têm uma influência negativa na frequência de sinistros enquanto as classes 2, 5 e 7 têm uma influência positiva). A classe 6 manteve-se inalterada dado que esta é estatisticamente significativa;
- ✓ VehicleAge: Manter a variável inalterada dado que todas as classes são estatisticamente significativas.

Realizando as ações descritas em cima, prossegue-se para a criação da tabela de desvios, de modo a confirmar se estas alterações realmente são significativas para a previsão da frequência de sinistros. Todas estas alterações foram feitas sequencialmente, até chegar ao nosso último modelo.

Através da função *anova*, obteve-se a seguinte tabela de desvios:

Modelo	Df	Dev	ΔDf	Δ Dev	Pr(>Chi)
1	62.411	5894,4			
2	62.417	5902,4	-6	-7,9591	0,6845
3	62.418	5907,2	-1	-4,8097	0,1227
4	62.420	5907,2	-2	-0,068800	0,9831
5	62.422	5908,7	-2	-1,4720	0,6945
6	62.426	5911,3	-4	-2,5596	0,8668

Tabela 9 - Análise dos desvios para a frequência de sinistros

O *output*, assim como a descrição dos modelos, podem ser encontrados no Anexo 3. Através da análise da Tabela 9, verificamos que todas as passagens são estatisticamente aceitáveis dado que o teste à nulidade dos parâmetros referente ás variáveis a eliminar ou modificar, aponta claramente para a não rejeição da hipótese nula, dado que os valores se encontram todos acima do limiar dos 5% de significância. O nosso modelo final, o modelo 6, a ser usado para a estimação da frequência de sinistros é o seguinte:

Coeficientes	Estimativa	Erro Padrão	t value	Pr(> |t|)
(Intercept)	-1,9728	0,1478	-13,346	< 2e-16
GroupAge2-3	-1,4009	0,1087	-12,890	< 2e-16
Zone2-2	-0,5698	0,1526	-3,735	0,000188
Zone2-3	-1,0623	0,1669	-6,366	0,000000
Zone2-4	-1,4536	0,1447	-10,048	< 2e-16
Zone2-5	-1,7196	0,4603	-3,736	0,000187
Class2.1.-3	-0,4755	0,1259	-3,776	0,000159
Class2.1.-6	0,5614	0,1408	3,987	0,000067
VehicleAge 2	-0,8379	0,1121	-7,475	0,000000
VehicleAge 3	-1,3303	0,2912	-4,568	0,000005

Tabela 10 - Modelo final da frequência de sinistros

Ora, através de uma primeira inspeção, este modelo parece bastante mais aceitável. Os erros padrão continuam a tomar valores reduzidos e todos os fatores são considerados estatisticamente significativos.

Através da análise das estimativas dos coeficientes de regressão, podemos verificar que o grupo de segurados que representam um maior risco para a frequência de sinistros da carteira em análise é o de indivíduos com menos de 30 anos, conduzindo um motociclo com menos de 5 anos de idade que circule nas zonas centrais e semi-centrais das três maiores cidades da Suécia e que tenha um motociclo com um rácio EV entre 19 e 24. Um segurado com estas características terá uma frequência de sinistros estimada de $e^{-1,9728+0,5614} = 24,38\%$ (aproximadamente). No extremo oposto, o grupo de segurados que representam um menor risco para a frequência de sinistros da carteira em análise é o de indivíduos que tenham mais de 30 anos, conduzindo um motociclo com mais de 20 anos de idade que circule nas vilas do norte da Suécia ou em Gotland e que tenha um rácio EV entre 8 e 15. Um segurado com estas características terá uma frequência de sinistros estimada de $e^{-1,9728-1,4009-1,7196-0,4755-1,3303} = 0,10\%$ (aproximadamente).

Considerando estas variáveis, pode verificar-se a existência de uma ordem de grandeza de risco, numa análise fator a fator. Se considerarmos a idade do segurado, à medida que esta aumenta a frequência de sinistros diminui, o que é algo lógico

dado que os segurados com idades superiores a 30 anos são mais maturos e conscientes para uma condução passiva. Em relação à zona, esta tem uma contribuição para a frequência de sinistros decrescente, que pode ser justificada pela diminuição do aglomerado populacional à medida que a classe aumenta. Considerando a idade do motociclo, dado que a maioria dos motociclos sinistrados se encontra na classe 2, é normal que seja esta a que mais contribui para a frequência de sinistros na carteira. Finalmente, em relação ao rácio EV, dado que a classe 6 é a que apresenta a maior frequência de sinistros de entre todas as classes da variável[4], é normal que esta tenha um contributo positivo para a nossa previsão.

Finalizando este segmento, o modelo que iremos utilizar para a modelização da frequência de sinistros inclui as variáveis *GroupAge2, Zone2, Class2.1 e VehicleAge*.

3.2.2. Modelização do custo médio dos sinistros

Passaremos agora para a criação do modelo de previsão do custo médio dos sinistros. Esta variável é de uma natureza diferente da variável anterior dado que tem uma natureza contínua, o que implica um processo diferente de análise e modelização. A distribuição que iremos adotar será a distribuição Gama, dado que usualmente apresenta uma boa adequabilidade a dados desta natureza, como é referido em Brockman et al (1992).

Primeiramente será necessário definir a nossa variável de resposta, que será o rácio entre o custo total dos sinistros (representado pela variável *ClaimCost*) e o número de sinistros (representado pela variável *NumberClaims*), ou seja:

$$Custo\ médio = \frac{Custo\ Total\ dos\ sinistros}{Número\ de\ sinistros}.$$

Das 64506 células 63835 não apresentam registo de qualquer sinistro. Como o custo médio dos sinistros implica que haja a ocorrência de pelo menos um sinistro, estes registos foram eliminados para efeitos de modelização desta variável de resposta. A

[4] Este facto pode ser verificado no Anexo 1.

função de ligação escolhida é a mesma que foi utilizada anteriormente, a função "log", que apesar de não ser uma função de ligação canónica, dá-nos a possibilidade de criar um modelo multiplicativo o que torna o modelo mais simples e intuitivo. Partiremos, mais uma vez, do modelo completo utilizando os fatores tarifários iniciais, os quais vão ser selecionados através do processo *backward.* O comando usado no R tem a seguinte forma:

```
glm(formula = variável_resposta ~ variáveis_explicativas, family = Gamma (link = "log"), weights = NumberClaims)
```

O *output* do modelo completo pode ser visto no Anexo 5. Através da análise desse mesmo *output,* é possível verificar que existem muitos níveis dos fatores tarifários que não são estatisticamente significativos a um nível de confiança de 5%. De modo a melhorar o nosso modelo, iremos descartar algumas variáveis e agrupar outras, sequencialmente, tal como se fez no modelo anterior:

- ✓ *Gender:* Eliminar a variável dado esta não ser significativa para a nossa carteira;
- ✓ *BonusClass:* Eliminar a variável dado esta não ser significativa para a nossa carteira;
- ✓ *VehicleAge:* Dado a classe 3 ter um *p-value* bastante superior a 5% (cerca de 22,9%), irá-se proceder ao agrupamento da classe 3 na classe 1 e deixamos a classe 2 inalterada, criando a variável *VehicleAge3*.
- ✓ *Zone:* Agrupamento das classes 4, 5 e 6 na classe 3 dado que têm estimativas similares e agrupamento da classe 2 na classe 1, criando a variável *Zone3*. Para além destas alterações, iremos agrupar a classe 7 na classe 1 pois, apesar de esta ser considerada estatisticamente significante para o custo médio, não se justifica, de uma perspetiva atuarial, usá-la dado que apenas foi registado um sinistro nessa zona. Havendo só um registo os nossos resultados seriam enviesados caso não fosse feito nenhum tratamento a esta classe;

- ✓ *Class:* Como esta variável foi usada no modelo anterior, iremos usá-la para explicar o custo médio, apesar de nenhuma das classes aparentar ser estatisticamente significativa através da análise do modelo completo. Agrupamos as classes 2, 4 e 5 na classe 1 assim como agrupamos as classes 3 e 7 na classe 6 devido a terem estimativas semelhantes, criando a variável *Class3.2*. Este agrupamento foi o único que fez sentido tento em conta as respetivas estimativas das classes;
- ✓ *GroupAge*: Procedemos ao agrupamento da classe 4 na classe 1, de modo a integrá-la no *intercept* e deixamos as classes 2 e 3 inalteradas, criando a variável *GroupAge3.2*.

De modo a confirmar a adequabilidade destas alterações, procedemos à criação de uma tabela de desvios, utilizando a mesma metodologia que foi usada no modelo anterior. A tabela produzida é a seguinte:

Modelo	Df	Dev	ΔDf	Δ Dev	Pr(>Chi)
1	641	1154.5			
2	647	1170.1	-6	-15,53390	0.1328
3	648	1171.5	-1	-1,46320	0.3363
4	649	1173.6	-1	-2,03290	0.2571
5	654	1182.2	-5	-8,65030	0.3619
6	659	1183.7	-5	-1,49350	0.9670
7	660	1183.9	-1	-0,16720	0.7452

Tabela 11 - Análise dos desvios dos modelos do custo médio dos sinistros

O *output*, assim como a descrição dos modelos, podem ser encontrados no Anexo 6. Através da análise da Tabela 11, verificamos que todas as passagens são estatisticamente aceitáveis dado que o teste à nulidade dos parâmetros referente ás variáveis a eliminar implica, mais uma vez, a não rejeição da hipótese nula, dado que os valores se encontram todos acima do limiar dos 5% de significância.

Após estas alterações, obteve-se o seguinte resultado para as estimativas de cada fator, correspondentes ao modelo 7:

Coeficientes	Estimativas	Desvio Padrão	t value	Pr(> |t|)
(Intercept)	9,8866	0,1776	55,664	< 2e-16
GroupAge3.2 – 2	0,7140	0,1696	4,209	0,000029
GroupAge3.2 - 3	0,4395	0,1667	2,637	0,008565
Zone3 – 3	-0,3684	0,0971	-3,796	0,000161
Class3.2 – 6	0,3502	0,0967	3,620	0,000317
VehicleAge3 - 2	-0,8261	0,0973	-8,491	< 2e-16

Tabela 12 - Modelo final do custo médio dos sinistros

Um *output* mais completo pode ser consultado no Anexo 7. Após a análise do *output*, podemos verificar que este modelo já se encontra bastante mais aceitável, podendo ainda verificar-se que as estimativas dos fatores tiveram uma diminuição do erro padrão, e já superam os testes de significância.

Observando as estimativas dos coeficientes de regressão, podemos notar que o perfil de risco que origina uma maior severidade de sinistros é o de um indivíduo com idade entre os 21 e os 29 anos, conduzindo um motociclo com idade inferior a 5 anos ou superior a 20 anos, que circule nas zonas centrais e semi-centrais das três maiores cidades da Suécia ou que circule nos subúrbios e cidades de média dimensão da Suécia ou que circule na ilha de Gotland e que tenha um rácio EV entre 8 e 12 ou superior a 19. Um segurado com estas características terá um custo médio estimado de $e^{9,8866+0,7140} / 9,3 = 4.318$ euros (aproximadamente). Por outro lado, o perfil de risco que origina uma menor severidade dos sinistros é o de um indivíduo que tenha menos de 21 anos ou mais de 60 anos, conduzindo um motociclo com idade entre os 5 e os 20 anos, que circule nas pequenas cidades ou nas vilas e aldeias da Suécia e que tenha um rácio EV menor que 8 ou entre 12 e 19. Um segurado com estas características terá um custo médio estimado de $e^{9,8866-0,3684-0,8261} / 9,3 =$ 640 euros (aproximadamente).

Tal como no modelo anterior, a classe de segurados que representam um maior risco para o custo médio dos sinistros são os condutores pertencentes à classe base, com exceção da classe referente à idade do segurado que indica que quem tem os sinistros mais gravosos são os condutores jovens com alguma experiência de condução. Este evento pode dever-se ao facto de, por usarem mais vezes o motociclo

e por se sentirem mais relaxados na sua condução, poderem originar acidentes com maiores custos por excesso de velocidade e uma condução menos passiva, por exemplo. A idade de veículo que representa um maior risco é a classe de veículos mais recentes, muito provavelmente pelo simples facto de que os veículos mais novos serem os que têm custos mais elevados de reparação em caso de sinistro, devido a serem relativamente recentes. O mesmo se pode dizer em relação aos veículos com uma vida útil superior a 20 anos dado que poderão originar custos mais elevados devido à antiguidade das peças do motociclo. Finalmente, em relação ao rácio EV, este apresenta resultados que não são muito normais, devido às agregações feitas, no entanto, é de salientar que esta agregação é a que faz sentido de um ponto de vista estatístico e, se verificarmos o custo médio dos motociclos pertencentes à classe 3 desta variável, podemos ver que é a que apresenta um maior custo médio de entre todas as classes do rácio EV[5].

Concluindo este subcapítulo, para a modelização do custo médio dos sinistros, as variáveis que são mais significantes para este modelo são a zona de ocorrência, a idade do segurado, a idade do veículo sinistrado e o rácio EV do motociclo, com os respetivos ajustamentos. Estas quatro variáveis, para além de representarem uma melhor aderência à nossa carteira, foram as usadas no modelo anterior (ver secção 3.2.1.) circunstância que irá tornar o nosso modelo de prémio puro mais fácil de obter e analisar.

3.2.3. Modelização da perda agregada dos sinistros

Após a criação dos modelos para a frequência e o custo médio dos sinistros, é possível proceder ao cálculo do prémio puro para cada grupo de risco homogéneo. Mas antes de dar seguimento a essa abordagem, iremos criar um modelo alternativo, que consiga prever a perda agregada dos sinistros, de modo a que não seja necessário modelizar a frequência e a severidade dos sinistros separadamente. Esta abordagem não é das mais usadas, dado que usualmente a frequência de sinistros é

[5] Este facto pode ser constatado no Anexo 1

bastante mais estável comparativamente com a severidade dos sinistros, o que implica que os estimadores dessa variável são mais corretos e precisos (B. Johansson & E. Ohlsson, 2010). De modo a fazer uma comparação entre prémios, iremos proceder à criação deste modelo.

Tal como já foi referido, uma distribuição pertence à família tweedie se a sua função variância puder ser apresentada sob a forma: $V(\mu) = \mu^p$ em que $p \epsilon \mathbb{R}$. Para o modelo apenas nos interessam os casos em que $p \epsilon [1,2]$ dado que esses são os representativos de uma distribuição Poisson-Gama composta (como foi referenciado na seção 2.2.4 deste trabalho).

Para este modelo a nossa variável de resposta é a perda agregada dos sinistros, ou seja, iremos modelizar a variável *TotalLoss* que é definida através da equação (2.7).

Para desenvolver o MLG, utilizamos o pacote *Tweedie,* criado por Peter K Dunn, assim como o pacote *Statmod*, criado por Gordon Smyth, que permitem usar diversas funções para cálculo e modelização de distribuições que pertençam à família tweedie. Acrescentando a família tweedie à, já conhecida, função *glm* dentro do R, para um determinado valor de p o comando a usar na ferramenta terá a seguinte forma:

```
glm(formula = variável_resposta ~ variáveis_explicativas + offset (log(Duration/365)), family = tweedie (var.power = p, link.power = 0))
```

O valor do parâmetro p a ser usado será 1.5, dado que o modelo com este valor é o que apresenta um menor desvio residual em termos do modelo completo (ver Anexo 8). Para esta verificação foram criados modelos diferentes em que cada modelo tinha um valor diferente para o parâmetro p ($p = 1,1; 1,2…$). Quando $p > 1,5$ notamos que o algoritmo do R não converge pelo que não foram criados modelos considerando $p > 1,5$.

Utilizando o já conhecido processo *backward*, começamos com o modelo completo e, a partir daí, selecionamos os fatores tarifários que são mais representativos para

a nossa variável de resposta. O *output* do modelo completo pode ser observado no Anexo 9. Como seria de esperar, os fatores tarifários escolhidos nos modelos anteriores são aqueles que apresentam uma melhor aderência à nossa carteira. Procedemos, tal como anteriormente, à agregação de alguns fatores e à exclusão de outros. Foram efetuadas sequencialmente as seguintes ações, testando-se em cada passo a significância estatística de cada uma delas:

- ✓ *BonusClass:* Eliminar a variável, dado que esta não é significativa;
- ✓ *Gender:* Eliminar a variável, dado que esta não é significativa;
- ✓ *GroupAge*: Procedemos ao agrupamento da classe 2 na classe 1, de modo a agrupar a classe no *intercept* e mantemos as classes 3 e 4 inalteradas, criando a variável *GroupAge4*.
- ✓ *Zone:* Agrupamento das classes 2 e 7 na classe 1, de modo a agrupar a classe no *intercept* e agrupamento da classe 4 na 3 e da classe 6 na 5, por estas terem estimativas similares, criando a variável *Zone4*;
- ✓ *VehicleAge:* Apesar de todas as classes serem significativas, iremos agregar a classe 3 na 2 por estas terem estimativas muito similares, criando a variável *VehicleAge4*.
- ✓ *Class:* Apenas a classe 6 se encontra mais próxima dos 5% de significância pelo que iremos criar uma variável binária denominada *Class4*, agregando as classes 2, 3, 4 e 5 na classe 1 e agregando a classe 7 na classe 6;

De modo a confirmar a adequabilidade destas alterações, procedemos à criação de uma tabela de desvios, utilizando a mesma metodologia praticada nos modelos anteriores. A tabela é a seguinte:

Modelo	Df	Dev	ΔDf	Δ Dev	Pr(>Chi)
1	62.411	6.779.808			
2	62.417	6.860.579	-6	-80771	0,310100
3	62.418	6.878.006	-1	-17427	0,215300
4	62.419	6.878.966	-1	-959	0,771200
5	62.423	6.948.583	-4	-69617	0,189300
6	62.424	6.948.703	-1	-120	0,917900
7	62.429	6.967.262	-5	-18.558	0,896900

Tabela 13 - Análise dos desvios dos modelos da perda agregada

O *output*, assim como a descrição dos modelos, podem ser encontrados no Anexo 10. Através da análise da Tabela 13, verificamos que todas as passagens são estatisticamente aceitáveis, dado que o teste à nulidade dos parâmetros referente ás variáveis a eliminar implica a não rejeição da hipótese nula, uma vez que os valores se encontram todos acima do limiar dos 5% de significância.

Após estas alterações, foram obtidas as estimativas dos coeficientes de regressão do nosso modelo final, o denominado modelo 7 na tabela anterior. Os resultados foram os seguintes:

Coeficientes	Estimativa	Desvio Padrão	t value	Pr(>|t|)
(Intercept)	8,5455	0,2871	29,765	< 2e-16
GroupAge4 – 3	-1,4846	0,2686	-5,527	3,27E-08
GroupAge4 – 4	-2,3639	0,6435	-3,673	0,0002400
Zone4 – 3	-1,0734	0,2607	-4,117	0,0000384
Zone4 – 5	-1,8937	0,6169	-3,070	0,0021450
Class4 – 6	1,1203	0,3136	3,573	0,0003530
VehicleAge4 - 2	-2,0414	0,2539	-8,039	9,18E-16

Tabela 14 – Modelo final para a perda agregada dos sinistros

Um *output* mais completo pode ser consultado no Anexo 11. Através de uma primeira inspeção o modelo apresenta uma boa adequabilidade aos nossos dados, sendo que as novas variáveis superam os testes de significância e os erros padrões se encontram todos próximos de zero.

Para este modelo, o perfil de risco de um segurado que resulta numa maior perda agregada para a companhia de seguros é o de um segurado que tenha entre 16 e 29 anos, com um motociclo com idade inferior a 5 anos, que circule nas zonas centrais e semi-centrais das três maiores cidades na Suécia ou que circule na ilha de Gotland e que tenha um motociclo com um rácio EV superior a 19. Um segurado com estas características tem uma perda agregada estimada de $e^{8,5455+1,1203}$ / 9,3 = 1.696 euros (aproximadamente). No outro extremo, o perfil de risco de um segurado que resulta numa menor perda agregada para a companhia de seguros é o de um segurado que tenha uma idade superior a 60 anos, com um motociclo com idade superior ou igual a 5 anos que circule em vilas e aldeias do norte da Suécia e com um

rácio EV inferior ou igual a 19. Um segurado com estas características tem uma perda agregada estimada de $e^{8,5455-2,3639-2,0414-1,8937}$/ 9,3 = 1,02 euros (aproximadamente).

É possível verificar que, tal como no modelo da frequência de sinistros, existe uma ordem decrescente em termos de risco dentro das classes nos nossos fatores, dado que o valor dos coeficientes de regressão vai decrescendo (originando valores mais negativos) à medida que a classe da variável tarifária aumenta, exceto os condutores que tenham um motociclo com um rácio EV superior a 19, cujo coeficiente estimado tem um sinal positivo, o que indica que os motociclos com um motor mais potente têm mais probabilidades de originar uma maior perda agregada.

Finalizando, iremos usar um modelo com quatro variáveis explicativas: *Class4, VehicleAge, Zone4 e GroupAge4*. No próximo capítulo, iremos discutir estes modelos para o cálculo do prémio puro, dependendo do perfil de risco homogéneo dos segurados na nossa carteira.

4. Prémio Puro

Após a criação dos modelos tarifários é possível modelizarmos o prémio puro a ser pago por um segurado, dependendo do grupo de risco a que ele/ela pertença. Tendo em conta que foram criados modelos através de abordagens diferentes, teremos que os usar para obter o prémio puro, ou seja:

1) Recorrer aos nossos modelos de frequência/custo médio e multiplicar as estimativas;
2) Recorrer ao modelo tweedie e proceder a uma estimativa direta.

De modo a calcular o prémio puro, em ambos os modelos, iremos usar a equação (2.2).

Para efeitos de análise da tarifa, foi elaborada uma tabela que resume os coeficientes de agravamento/desconto sob o prémio padrão, por fatores das nossas variáveis

tarifárias divididos pelas classes originais e por modelo estimado (a tabela pode ser consultada no Anexo 12). Esta tabela, sucintamente, indica as classes tarifárias que originam um agravamento ou desconto no prémio base para os dois modelos. Se o valor for superior a 1 a classe correspondente origina um agravamento sobre o prémio padrão, se o valor for inferior a 1 a classe correspondente origina um desconto sobre o prémio padrão e se o valor for igual a 1 a classe não tem qualquer efeito sob o prémio padrão. Por exemplo, considerando o modelo frequência/custo médio, um segurado que tenha 28 anos terá um agravamento sob o prémio padrão de 2,04 pelo irá pagar um prémio de 294 * 2,04 = 600 euros (aproximadamente).

Imediatamente se consegue reparar na discrepância entre os dois modelos e para justamente confirmar esse facto, é importante analisar a amplitude tarifária dos nossos modelos, i.e., o quociente entre o prémio puro mais elevado e o prémio puro mais reduzido, de modo a confirmar se os nossos modelos se podem aplicar num contexto económico real.

Para os modelos frequência/custo médio e tweedie, as amplitudes tarifárias são, respetivamente, de 1438 e 1638 (os cálculos podem ser consultados no Anexo 12). Imediatamente se verifica que estes valores não são, de todo, aplicáveis a qualquer tipo de mercado segurador. Isto significa que, no caso do modelo frequência/custo médio, um segurado com o maior perfil de risco da carteira irá pagar 1438 vezes mais, comparativamente a um segurado que esteja nas classes de risco mais reduzidas. O mesmo raciocínio se aplica ao prémio tweedie em que, neste caso, o segurado com o maior perfil de risco paga um prémio 1638 vezes superior. Se uma seguradora aplicasse estes modelos, teria de encontrar uma maneira de corrigir estas disparidades entre prémios verificadas, por exemplo, fazendo um ajuste manual aos prémios de modo a aproximar mais as amplitudes tarifárias verificadas.

Para ilustrar a situação, vamos assumir a existência de cinco tipos de segurados com características diferentes:

1) O primeiro segurado é um indivíduo com 17 anos de idade, com um motociclo com 2 anos de antiguidade com um rácio EV de 20 e que circula nas três maiores cidades da Suécia.
2) O segundo segurado é um indivíduo com 64 anos de idade, com um motociclo com 21 anos de antiguidade com um rácio EV de 4 e que circula na ilha de Gotland.
3) O terceiro segurado é um indivíduo com 32 anos de idade, com um motociclo com 3 anos de antiguidade com um rácio EV de 10 e que circula nas pequenas cidades da Suécia.
4) O quarto segurado é um indivíduo com 23 anos de idade, com um motociclo com 10 anos de antiguidade com um rácio EV de 23 e que circula nos subúrbios e cidades de média dimensão da Suécia.
5) O quinto segurado é um indivíduo com 41 anos de idade, com um motociclo com 17 anos de antiguidade com um rácio EV de 7 e que circula nas aldeias do norte da Suécia.

Este grupo de segurados foi escolhido de modo a tentar capturar um perfil de risco que tente abranger as classes que originam mais sinistros com elevados custos médios, assim como os segurados que têm menos probabilidades de vir a originar sinistros.

4.1. Prémio de frequência/custo médio

Relativamente ao nosso grupo de segurados, fazendo primeiro a previsão da frequência de sinistros, seguindo-se a previsão da severidade e, posteriormente, multiplicando os resultados, obtivemos os prémios puros correspondentes:

Segurado	Frequência	Custo Médio	Prémio Freq./Custo Médio
1	24,38%	3.001,25 €	731,71 €
2	0,16%	4.657,74 €	7,56 €
3	0,74%	3.222,41 €	23,72 €
4	5,97%	2.682,99 €	160,07 €
5	0,35%	993,85 €	3,44 €

Tabela 15 - Prémios de frequência/custo médio por tipo de segurado

Um cálculo mais detalhado pode ser consultado no Anexo 13.

4.2. Prémio Tweedie

Para o modelo de prémio puro tweedie, obtivemos os seguintes resultados, aos quais iremos fazer uma comparação percentual com os resultados anteriores:

Segurado	Prémio Freq./Custo Médio	Prémio Tweedie	Variação (%)	Variação Bruta
1	731,71 €	1.695,59 €	131,73%	963,88 €
2	7,56 €	6,75 €	-10,64%	-0,80 €
3	23,72 €	42,84 €	80,60%	19,12 €
4	160,07 €	220,17 €	37,55%	60,10 €
5	3,44 €	0,23 €	-93,31%	-3,21 €

Tabela 16 - Prémios de frequência/custo médio e prémios tweedie por tipo de segurado

Os passos intermédios dos cálculos podem ser consultados no Anexo 14.

No caso dos nossos grupos de segurados, é possível fazer-se alguns comentários. Desde já, existe uma diferença bastante elevada entre o prémio puro com a modelização separada e com a modelização conjunta, em particular, no caso do segurado 1 (que representa o segurado com o maior perfil de risco), para o qual o prémio tweedie é 131,73% mais elevado do que o prémio frequência/custo médio, o que representa uma variação de cerca de 964 euros.

Se considerarmos os segurados 3 e 4, estes apresentam variações entre prémios mais reduzidas do que o segurado 1 e, nestes casos, o prémio tweedie é superior ao prémio frequência/custo médio em 80,60% e 37,55%, respetivamente.

O resultado do prémio puro do segurado 5 é relativamente baixo, devido à sua baixa frequência de sinistralidade assim como um baixo custo médio. Dado que este

segurado tem todos os fatores que proporcionam um menor risco, não é surpreendente que tenha o prémio mais baixo do grupo. Apesar de a variação percentual ser cerca de -93,31%, em relação ao prémio frequência/custo médio, este valor é enganador dado que se se considerar apenas valores brutos, o prémio frequência/custo médio é superior ao prémio tweedie em apenas 3 euros, aproximadamente.

Finalmente, em relação ao segurado 2, este tem uma frequência de sinistros estimada muito similar à do segurado 5. No entanto, apresenta um custo médio bastante mais elevado e a variação entre prémios é de apenas 0,80 euros, o que significa que este segurado é o que apresenta uma menor variação dentro do grupo.

Estas discrepâncias devem-se, sobretudo, à baixa frequência de sinistros verificada na carteira em análise. Possivelmente, com uma base de dados com mais registos e um maior número de unidades de risco, seria possível obter resultados mais consistentes.

Na perspetiva de uma seguradora, se optassem apenas por usar o modelo de prémio frequência/custo médio, iriam cobrar prémios mais aceitáveis e consistentes do que os do modelo tweedie. Efetivamente, utilizando este grupo de segurados, os prémios tweedie parecem não ser muito consistentes, pois registamos casos mais "extremos", comparativamente aos prémios frequência/custo médio: para o caso do segurado 1 existe uma variação de quase 964 euros para o prémio frequência/custo médio e para o caso do segurado 5, o prémio a ser cobrado é tão reduzido que não se justifica impô-lo em qualquer tipo de garantia de seguros de motociclos.

5. Conclusão e desenvolvimentos futuros

Neste trabalho foi feita uma análise a uma carteira de seguros de motociclos, um ramo que não é muito abordado pela comunidade cientifica e, por isso, onde não existem publicações sobre o assunto. Para o efeito utilizamos duas abordagens amplamente usadas: a modelização separada da frequência e do custo médio dos sinistros e a modelização da perda agregada dos sinistros.

As variáveis que melhor conseguiram explicar o comportamento das nossas variáveis de resposta foram: *(i)* a idade dos segurados, *(ii)* a zona de ocorrência do sinistro, *(iii)* a idade do motociclo segurado e *(iv)* o rácio EV do motociclo (fazendo as simplificações estatisticamente necessárias). De uma perspetiva atuarial são as variáveis mais lógicas a usar para o cálculo do prémio puro. As restantes variáveis mostraram pouca aderência à nossa carteira, pelo que foram excluídas da análise.

Em relação aos dois modelos diferentes que foram apresentados, concluímos que o modelo de frequência/custo médio é aquele que apresenta resultados mais consistentes, quando comparado com o modelo de prémio puro tweedie. No entanto, pudemos verificar que a amplitude tarifária de ambos os modelos apresenta valores que, simplesmente, não são aplicáveis em qualquer tipo de mercado num contexto real.

Relativamente ao modelo de prémio tweedie, de uma perspetiva atuarial, a imposição de prémios puros com este modelo iria, muito provavelmente, originar grandes perdas para a companhia, mesmo acrescendo uma margem de lucro e os respetivos gastos administrativos e de gestão. Por um lado poderão vir a cobrar prémios demasiado elevados, o que irá possivelmente originar uma diminuição do número de segurados que poderão passar para outras seguradoras do mercado que pratiquem preços mais acessíveis e, logicamente, levará a uma diminuição no montante de prémios recebidos; por outro lado, cobrar prémios demasiado baixos iria obrigar a seguradora a pagar por indemnizações superiores aos prémios recebidos. Ambos os casos iriam deixar a seguradora com uma perda financeira,

pelo que seria necessário realizar ajustes manuais na tarifa de modo a corrigir os modelos e amplitudes tarifárias, criando prémios mais justos para os segurados e mais rentáveis para a companhia. Estas correções poderão dar origem a resultados mais próximos da realidade e mais consistentes.

Finalmente, para o caso da nossa carteira, apenas criámos modelos baseados numa tarifação *à priori* das variáveis explicativas tentando criar modelos o mais simples possível. Nitidamente pode ser verificado que ambos os modelos precisam de ser corrigidos de modo a poderem ser aplicados num contexto económico real, pelo que se sugere, para um trabalho futuro, a imposição de modelos de credibilidade ou até mesmo implementar modelos baseados em análises multivariadas sobre as variáveis explicativas. Mas, dado a frequência de sinistros observada ser tão reduzida, seria muito difícil utilizar alguns destes métodos sem ter uma base de dados mais consistente e com um maior número de registos.

Como no mundo da estatística se costuma afirmar, um modelo é tão bom quanto os seus dados mas, usando as abordagens descritas neste trabalho e uma base de dados com um maior número de registos, é possível impor um bom modelo de prémio puro simples e fácil de interpretar e, juntamente com os objetivos económicos da companhia, criar uma tarifa que reflita bem os riscos associados dos seus segurados cobrando prémios justos e que resultem em ganhos financeiros.

6. Anexos

Anexo 1 – Tabelas de análise de variáveis explicativas

Por Género		
Classe	Feminino	Masculino
Unidades de Risco	9482	52954
Sinistros	61	632
FS	1,12%	2,42%
PP	22 €	65 €
CM	1.961 €	2.693 €

Por Rácio EV							
Classe	1	2	3	4	5	6	7
Unidades de Risco	6748	5007	18460	11992	11502	8091	636
Sinistros	46	56	165	97	149	174	6
FS	1,14%	1,90%	1,79%	1,58%	2,79%	4,84%	1,98%
PP	26 €	31 €	63 €	38 €	66 €	123 €	51 €
CM	2.321 €	1.634 €	3.483 €	2.415 €	2.379 €	2.553 €	2.591 €

Anexo 2 - Modelo Completo para Frequência de Sinistros

```
Call:
glm(formula = NumberClaims ~ factor(Gender) + factor(GroupAge) +
    factor(VehicleAge) + factor(Class) + factor(Zone) + factor(BonusClas
    offset(log(Duration/365)), family = quasipoisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2502  -0.1509  -0.0974  -0.0639   4.8405

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.24558    0.36095  -6.221 4.96e-10 ***
factor(Gender)M       0.28077    0.19195   1.463 0.143550
factor(GroupAge)2    -0.24790    0.25786  -0.961 0.336376
factor(GroupAge)3    -1.67145    0.25793  -6.480 9.22e-11 ***
factor(GroupAge)4    -1.71446    0.35811  -4.788 1.69e-06 ***
factor(VehicleAge)2  -0.81566    0.11434  -7.134 9.88e-13 ***
factor(VehicleAge)3  -1.25079    0.30141  -4.150 3.33e-05 ***
factor(Class)2        0.24405    0.28740   0.849 0.395803
factor(Class)3       -0.33384    0.24897  -1.341 0.179965
factor(Class)4       -0.22822    0.26745  -0.853 0.393492
factor(Class)5        0.15596    0.25756   0.606 0.544814
factor(Class)6        0.68838    0.25833   2.665 0.007708 **
factor(Class)7        0.40929    0.62561   0.654 0.512966
factor(Zone)2        -0.56930    0.15368  -3.705 0.000212 ***
factor(Zone)3        -1.06898    0.16806  -6.361 2.02e-10 ***
factor(Zone)4        -1.46434    0.14951  -9.794  < 2e-16 ***
factor(Zone)5        -1.68365    0.48619  -3.463 0.000535 ***
factor(Zone)6        -1.40223    0.35305  -3.972 7.14e-05 ***
factor(Zone)7        -1.84649    1.42525  -1.296 0.195131
factor(BonusClass)2  -0.04154    0.20912  -0.199 0.842537
factor(BonusClass)3   0.02841    0.22709   0.125 0.900425
factor(BonusClass)4   0.25890    0.22071   1.173 0.240791
factor(BonusClass)5   0.03712    0.25035   0.148 0.882133
factor(BonusClass)6  -0.04800    0.25943  -0.185 0.853218
factor(BonusClass)7   0.20970    0.16459   1.274 0.202640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.019017)

    Null deviance: 6696.3  on 62435  degrees of freedom
Residual deviance: 5894.4  on 62411  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 8
```

Anexo 3 - Tabela de desvios dos modelos para a Frequência de Sinistros

```
> anova(g1, g2, g3, g4, g5, g6, test="Chisq")
Analysis of Deviance Table

Model 1: NumberClaims ~ factor(Gender) + factor(GroupAge) + factor(Zone) +
    factor(Class) + factor(VehicleAge) + factor(BonusClass) +
    offset(log(Duration/365))
Model 2: NumberClaims ~ factor(Gender) + factor(GroupAge) + factor(Zone) +
    factor(Class) + factor(VehicleAge) + offset(log(Duration/365))
Model 3: NumberClaims ~ factor(GroupAge) + factor(Zone) + factor(Class) +
    factor(VehicleAge) + offset(log(Duration/365))
Model 4: NumberClaims ~ factor(GroupAge) + factor(Zone2) + factor(Class) +
    factor(VehicleAge) + offset(log(Duration/365))
Model 5: NumberClaims ~ factor(GroupAge2) + factor(Zone2) + factor(Class) +
    factor(VehicleAge) + offset(log(Duration/365))
Model 6: NumberClaims ~ factor(GroupAge2) + factor(Zone2) + factor(Class2.1.) +
    factor(VehicleAge) + offset(log(Duration/365))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     62411     5894.4
2     62417     5902.4 -6  -7.9591   0.6845
3     62418     5907.2 -1  -4.8097   0.1227
4     62420     5907.2 -2  -0.0688   0.9831
5     62422     5908.7 -2  -1.4720   0.6945
6     62426     5911.3 -4  -2.5596   0.8668
```

Anexo 4 – Modelo final para a Frequência de Sinistros

```
Call:
glm(formula = NumberClaims ~ factor(GroupAge2) + factor(Zone2) +
    factor(Class2.1.) + factor(VehicleAge) + offset(log(Duration/365)),
    family = quasipoisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1912  -0.1537  -0.1001  -0.0653   4.7300

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.9728     0.1478 -13.346  < 2e-16 ***
factor(GroupAge2)3   -1.4009     0.1087 -12.890  < 2e-16 ***
factor(Zone2)2       -0.5698     0.1526  -3.735 0.000188 ***
factor(Zone2)3       -1.0623     0.1669  -6.366 1.96e-10 ***
factor(Zone2)4       -1.4536     0.1447 -10.048  < 2e-16 ***
factor(Zone2)5       -1.7196     0.4603  -3.736 0.000187 ***
factor(Class2.1.)3   -0.4755     0.1259  -3.776 0.000159 ***
factor(Class2.1.)6    0.5614     0.1408   3.987 6.69e-05 ***
factor(VehicleAge)2  -0.8379     0.1121  -7.475 7.82e-14 ***
factor(VehicleAge)3  -1.3303     0.2912  -4.568 4.94e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.002277)

    Null deviance: 6696.3  on 62435  degrees of freedom
Residual deviance: 5911.3  on 62426  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 8
```

Anexo 5– Modelo completo para o Custo Médio dos Sinistros

```
Call:
glm(formula = CM ~ factor(Gender.sin) + factor(GroupAge.sin) +
    factor(VehicleAge.sin) + factor(Zone.sin) + factor(BonusClass.sin) +
    factor(Class.sin), family = Gamma(link = "log"), weights = NumberClaims.sin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2400  -1.4831  -0.5879   0.3971   3.5075

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.78952    0.33830  28.937  < 2e-16 ***
factor(Gender.sin)M        0.14379    0.17379   0.827  0.40834
factor(GroupAge.sin)2      0.52580    0.23015   2.285  0.02266 *
factor(GroupAge.sin)3      0.27357    0.23152   1.182  0.23779
factor(GroupAge.sin)4     -0.35435    0.32671  -1.085  0.27851
factor(VehicleAge.sin)2   -0.89743    0.10419  -8.613  < 2e-16 ***
factor(VehicleAge.sin)3   -0.39751    0.26336  -1.509  0.13169
factor(Zone.sin)2          0.12368    0.13901   0.890  0.37395
factor(Zone.sin)3         -0.34419    0.15056  -2.286  0.02258 *
factor(Zone.sin)4         -0.33681    0.13609  -2.475  0.01359 *
factor(Zone.sin)5         -0.33976    0.43712  -0.777  0.43728
factor(Zone.sin)6         -0.49618    0.32003  -1.550  0.12154
factor(Zone.sin)7         -4.17961    1.26704  -3.299  0.00103 **
factor(BonusClass.sin)2    0.03228    0.18787   0.172  0.86364
factor(BonusClass.sin)3    0.35166    0.20561   1.710  0.08769 .
factor(BonusClass.sin)4   -0.10121    0.20044  -0.505  0.61377
factor(BonusClass.sin)5    0.24287    0.22711   1.069  0.28528
factor(BonusClass.sin)6    0.53645    0.23528   2.280  0.02293 *
factor(BonusClass.sin)7    0.10788    0.14838   0.727  0.46749
factor(Class.sin)2        -0.01390    0.25301  -0.055  0.95621
factor(Class.sin)3         0.34182    0.21383   1.599  0.11041
factor(Class.sin)4         0.01351    0.23029   0.059  0.95323
factor(Class.sin)5         0.04022    0.21890   0.184  0.85427
factor(Class.sin)6         0.39934    0.21711   1.839  0.06632 .
factor(Class.sin)7         0.50738    0.56701   0.895  0.37121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.583013)

    Null deviance: 1378.5  on 665  degrees of freedom
Residual deviance: 1154.5  on 641  degrees of freedom
AIC: 15179

Number of Fisher Scoring iterations: 10
```

Anexo 6 – Tabela de desvios dos modelos para o Custo Médio de Sinistros

```
> anova(gamma_1, gamma_2, gamma_3, gamma_4, gamma_5, gamma_6, gamma_7, test="Chisq")
Analysis of Deviance Table

Model 1: CM ~ factor(Gender.sin) + factor(GroupAge.sin) + factor(Zone.sin) +
    factor(Class.sin) + factor(VehicleAge.sin) + factor(BonusClass.sin)
Model 2: CM ~ factor(Gender.sin) + factor(GroupAge.sin) + factor(Zone.sin) +
    factor(Class.sin) + factor(VehicleAge.sin)
Model 3: CM ~ factor(GroupAge.sin) + factor(Zone.sin) + factor(Class.sin) +
    factor(VehicleAge.sin)
Model 4: CM ~ factor(GroupAge.sin) + factor(Zone.sin) + factor(Class.sin) +
    factor(VehicleAge3.sin)
Model 5: CM ~ factor(GroupAge.sin) + factor(Zone3.sin) + factor(Class.sin) +
    factor(VehicleAge3.sin)
Model 6: CM ~ factor(GroupAge.sin) + factor(Zone3.sin) + factor(Class3.2.sin) +
    factor(VehicleAge3.sin)
Model 7: CM ~ factor(GroupAge3.2.sin) + factor(Zone3.sin) + factor(Class3.2.sin) +
    factor(VehicleAge3.sin)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       641     1154.5
2       647     1170.1 -6 -15.5339   0.1328
3       648     1171.5 -1  -1.4632   0.3363
4       649     1173.6 -1  -2.0329   0.2571
5       654     1182.2 -5  -8.6503   0.3619
6       659     1183.7 -5  -1.4935   0.9670
7       660     1183.9 -1  -0.1672   0.7452
```

Anexo 7 – Modelo final para o Custo Médio de Sinistros

```
Call:
glm(formula = CM ~ factor(GroupAge3.2.sin) + factor(Zone3.sin) +
    factor(Class3.2.sin) + factor(VehicleAge3.sin), family = Gamma(link = "log"),
    weights = NumberClaims.sin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2911  -1.5187  -0.6422   0.3904   3.6266

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.88662    0.17761  55.664  < 2e-16 ***
factor(GroupAge3.2.sin)2   0.71395    0.16962   4.209 2.92e-05 ***
factor(GroupAge3.2.sin)3   0.43945    0.16666   2.637 0.008565 **
factor(Zone3.sin)3        -0.36837    0.09705  -3.796 0.000161 ***
factor(Class3.2.sin)6      0.35017    0.09672   3.620 0.000317 ***
factor(VehicleAge3.sin)2  -0.82610    0.09730  -8.491  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.612696)

    Null deviance: 1378.5  on 665  degrees of freedom
Residual deviance: 1183.9  on 660  degrees of freedom
AIC: 15163

Number of Fisher Scoring iterations: 7
```

Anexo 8– Desvio residual dos modelos Tweedie

P	Desvio Residual
1,1	83.797.682
1,2	42.388.006
1,3	22.350.722
1,4	12.321.292
1,5	6.779.808
1,6	Algorítmo não convergeu
1,7	Algorítmo não convergeu
1,8	Algorítmo não convergeu
1,9	Algorítmo não convergeu

Anexo 9– Modelo completo para a Perda Agregada dos Sinistros

```
Call:
glm(formula = TotalLoss ~ factor(Gender) + factor(GroupAge) +
    factor(Zone) + factor(Class) + factor(VehicleAge) + factor(BonusClass) +
    offset(log(Duration/365)), family = tweedie(var.power = 1.5,
    link.power = 0))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-33.10   -8.44   -6.32   -4.83  406.69

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.54792    0.79066  10.811  < 2e-16 ***
factor(Gender)M       0.39531    0.34925   1.132  0.25769
factor(GroupAge)2     0.02774    0.63269   0.044  0.96502
factor(GroupAge)3    -1.52761    0.62013  -2.463  0.01377 *
factor(GroupAge)4    -2.40125    0.79805  -3.009  0.00262 **
factor(Zone)2        -0.47112    0.33339  -1.413  0.15762
factor(Zone)3        -1.08564    0.35022  -3.100  0.00194 **
factor(Zone)4        -1.56097    0.31577  -4.943 7.70e-07 ***
factor(Zone)5        -2.25642    0.85605  -2.636  0.00839 **
factor(Zone)6        -2.08040    0.68768  -3.025  0.00249 **
factor(Zone)7        -6.18848    5.70021  -1.086  0.27763
factor(Class)2       -0.21164    0.57946  -0.365  0.71493
factor(Class)3       -0.30177    0.44931  -0.672  0.50182
factor(Class)4       -0.56988    0.49093  -1.161  0.24572
factor(Class)5       -0.12679    0.48766  -0.260  0.79486
factor(Class)6        0.87382    0.48842   1.789  0.07361 .
factor(Class)7        0.39562    1.12474   0.352  0.72503
factor(VehicleAge)2 -2.01454    0.23509  -8.569  < 2e-16 ***
factor(VehicleAge)3 -1.93358    0.45975  -4.206 2.61e-05 ***
factor(BonusClass)2 -0.34640    0.43657  -0.793  0.42752
factor(BonusClass)3  0.62270    0.40839   1.525  0.12732
factor(BonusClass)4 -0.27829    0.49302  -0.564  0.57244
factor(BonusClass)5  0.33341    0.46897   0.711  0.47712
factor(BonusClass)6  0.55191    0.44773   1.233  0.21770
factor(BonusClass)7  0.31297    0.33296   0.940  0.34724
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 11348.38)

    Null deviance: 9219077  on 62435  degrees of freedom
Residual deviance: 6779808  on 62411  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 11
```

Anexo 10 – Tabela de desvios dos modelos para a Perda Agregada dos Sinistros

```
> anova(tweedie_1, tweedie_2, tweedie_3, tweedie_4, tweedie_5, tweedie_6, tweedie_7, test="Chisq")
Analysis of Deviance Table

Model 1: TotalLoss ~ factor(Gender) + factor(GroupAge) + factor(Zone) +
    factor(Class) + factor(VehicleAge) + factor(BonusClass) +
    offset(log(Duration/365))
Model 2: TotalLoss ~ factor(Gender) + factor(GroupAge) + factor(Zone) +
    factor(Class) + factor(VehicleAge) + offset(log(Duration/365))
Model 3: TotalLoss ~ factor(GroupAge) + factor(Zone) + factor(Class) +
    factor(VehicleAge) + offset(log(Duration/365))
Model 4: TotalLoss ~ factor(GroupAge4) + factor(Zone) + factor(Class) +
    factor(VehicleAge) + offset(log(Duration/365))
Model 5: TotalLoss ~ factor(GroupAge4) + factor(Zone4) + factor(Class) +
    factor(VehicleAge) + offset(log(Duration/365))
Model 6: TotalLoss ~ factor(GroupAge4) + factor(Zone4) + factor(Class) +
    factor(VehicleAge4) + offset(log(Duration/365))
Model 7: TotalLoss ~ factor(GroupAge4) + factor(Zone4) + factor(Class4) +
    factor(VehicleAge4) + offset(log(Duration/365))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     62411    6779808
2     62417    6860579 -6   -80771   0.3101
3     62418    6878006 -1   -17427   0.2153
4     62419    6878966 -1     -959   0.7712
5     62423    6948583 -4   -69617   0.1893
6     62424    6948703 -1     -120   0.9179
7     62429    6967262 -5   -18558   0.8969
```

Anexo 11 – Modelo final para a Perda Agregada dos Sinistros

```
Call:
glm(formula = TotalLoss ~ factor(GroupAge4) + factor(Zone4) +
    factor(Class4) + factor(VehicleAge4) + offset(log(Duration/365)),
    family = tweedie(var.power = 1.5, link.power = 0))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-29.27   -8.67   -6.56   -5.14  482.80

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.5455     0.2871  29.765  < 2e-16 ***
factor(GroupAge4)3   -1.4846     0.2686  -5.527 3.27e-08 ***
factor(GroupAge4)4   -2.3639     0.6435  -3.673 0.000240 ***
factor(Zone4)3       -1.0734     0.2607  -4.117 3.84e-05 ***
factor(Zone4)5       -1.8937     0.6169  -3.070 0.002145 **
factor(Class4)6       1.1203     0.3136   3.573 0.000353 ***
factor(VehicleAge4)2 -2.0414     0.2539  -8.039 9.18e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 15286.93)

    Null deviance: 9219077  on 62435  degrees of freedom
Residual deviance: 6967262  on 62429  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 9
```

Anexo 12- Coeficientes de agravamento/desconto dos prémios

Fator	Classe	Frequência	Custo Médio	Freq/Custo Médio	Tweedie
Padrão	-	13,91%	2.115 €	294€	553€
Idade do Segurado	16-20	1,00	1,00	1,00	1,00
	21-29	1,00	2,04	2,04	1,00
	30-59	0,25	1,55	0,38	0,23
	>60	0,25	1,00	0,25	0,09
Idade Motociclo	0-5	1,00	1,00	1,00	1,00
	6-20	0,43	0,44	0,19	0,13
	21-99	0,26	1,00	0,26	0,13
Rácio EV	0-5	1,00	1,00	1,00	1,00
	6-8	1,00	1,00	1,00	1,00
	9-12	0,62	1,42	0,88	1,00
	13-15	0,62	1,00	0,62	1,00
	16-19	1,00	1,00	1,00	1,00
	20-24	1,75	1,42	2,49	3,07
	> 24	1,00	1,42	1,42	3,07
Zona	Centro cidades	1,00	1,00	1,00	1,00
	Subúrbios	0,57	1,00	0,57	1,00
	Peq. Cidades	0,35	0,69	0,24	0,34
	Vilas	0,23	0,69	0,16	0,34
	Vilas do Norte	0,18	0,69	0,12	0,15
	Aldeias do Norte	0,23	0,69	0,16	0,15
	Gotland	0,18	1,00	0,18	1,00

Estes resultados foram calculados com recurso às estimativas obtidas nos nossos modelos finais e através do uso da fórmula (2.2).

Tomando como exemplo o fator "Idade do segurado", classe "30-59" e denominando C_M como o coeficiente de desconto/agravamento do modelo M, os valores da tabela foram obtidos da seguinte maneira:

- Modelo frequência/custo médio

$$C_{frequência/custo\ médio} = C_{frequência} * C_{custo\ médio} = \exp(-1{,}4009) * \exp(0{,}4395) \approx$$

$$\approx 0{,}2463 * 1{,}552 \approx \mathbf{0,38}$$

- Modelo tweedie

$C_{tweedie} = \exp(-1{,}4846) \approx \mathbf{0,23}$

Anexo 13- Amplitude tarifária dos modelos estimados

Denominando $AT, P_{máx} e\ P_{mín}$ como Amplitude Tarifária, Prémio máximo e Prémio mínimo, respetivamente, obtivemos os seguintes resultados para os nossos modelos:

- Modelo frequência/custo médio

$$AT = \frac{P_{máx}}{P_{mín}} = \frac{\exp(-1{,}9728 + 0{,}5614) * \exp(9{,}8866 + 0{,}714 + 0{,}3502)}{\exp(-1{,}9728 - 1{,}4009 - 1{,}3303 - 1{,}7196) * \exp(9{,}8866 - 0{,}8261 - 0{,}3684)} \approx$$

$$\approx \frac{13896}{9{,}66} \approx \mathbf{1438}$$

- Modelo tweedie

$$AT = \frac{P_{máx}}{P_{mín}} = \frac{\exp(8{,}5455 + 1{,}1203)}{\exp(8{,}5455 - 2{,}3639 - 2{,}0414 - 1{,}8937)} \approx \frac{15769}{9{,}45} \approx \mathbf{1668}$$

Anexo 14- Cálculos intermédios para o prémio puro

Denominando PP_i, FS_i, CM_i, PT_i como, respetivamente, o prémio frequência/custo médio, a frequência de sinistros, o custo médio dos sinistros, o prémio tweedie do segurado *i*, (relembrando que a taxa de conversão usada é a de 1 EUR = 9,3 SEK), obtivemos os seguintes resultados para os nossos modelos de prémio puro:

- Prémio de frequência/custo médio

Segurado 1: $PP_1 = FS_1 * CM_1/9{,}3 = exp\{-1{,}9728 + 0{,}5614\} * exp\{9{,}8866 + 0{,}3502\}/$
$9{,}3 \approx 0{,}2438 * 27911{,}67/9{,}3 \approx \mathbf{732}€$

Segurado 2: $PP_2 = FS_2 * CM_2/9{,}3 = exp\{-1{,}9728 - 1{,}4009 - 1{,}3303 - 1{,}7196\} * exp\{9{,}8866 + 0{,}4395 + 0.3502\}/9{,}3 \approx 0{,}0016 * 43316{,}98/9{,}3 \approx$ **8€**

Segurado 3: $PP_3 = FS_3 * CM_3/9{,}3 = exp\{-1{,}9728 - 1{,}4009 - 0{,}4755 - 1{,}0623\} * exp\{9{,}8866 + 0{,}4395 + 0{,}3502 - 0{,}3684\}/9{,}3 = 0{,}0074 * 29968{,}44/9{,}3 =$ **24€**

Segurado 4: $PP_4 = FS_4 * CM_4/9{,}3 = exp\{-1{,}9728 - 0{,}8379 + 0{,}5614 - 0{,}5698\} * exp\{9{,}8866 + 0{,}714 - 0{,}8261 + 0{,}3502\}/9{,}3 \approx 0{,}0597 * 24951{,}77/9{,}3 \approx$ **160€**

Segurado 5: $PP_5 = FS_5 * CM_5/9{,}3 = exp\{-1{,}9728 - 1{,}4009 - 0{,}8379 - 1{,}4536\} * exp\{9{,}8866 + 0{,}4395 - 0{,}8261 - 0{,}3684\}/9{,}3 \approx 0{,}0035 * 9242{,}80/9{,}3 \approx$ **3€**

- Prémio Tweedie

Segurado 1: $PT_1 = \exp\{8{,}5455 + 1{,}1203\}/9{,}3 \approx$ **1.696€**

Segurado 2: $PT_2 = \exp\{8{,}5455 - 2{,}3639 - 2{,}0414\}/9{,}3 \approx$ **7€**

Segurado 3: $PT_3 = \exp\{8{,}5455 - 1{,}4846 - 1{,}0734\}/9{,}3 \approx$ **43€**

Segurado 4: $PT_4 = \exp\{8{,}5455 - 2{,}0414 + 1{,}1203\}/9{,}3 \approx$ **220€**

Segurado 5: $PT_5 = \exp\{8{,}5455 - 1{,}4846 - 2{,}3639 - 1{,}8937 - 2{,}0414\}/9{,}3 \approx$

$\approx$ **0,23€**

7. Bibliografia

1. Brockman, M.J. & Wright, T.S. (1992). Statistical Motor Rating: Making Effective Use of Your Data, Journal of the Institute of Actuaries 119, 457-543
2. Dunn, P. (2013). Tweedie: An R Package for Tweedie exponential family models [Online]. Disponível em: https://cran.r-project.org/web/packages/tweedie/index.html
3. Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models,* 1st edition, Cambridge University Press.
4. Kaas, R.,Goovaerts, M., Dhaene, J. & Denuit, M. (2008). *Modern Actuarial Risk Theory: Using R*, 2nd edition, Springer.
5. Kafková, S., Krivánková, L. (2014). *Generalized Linear Models in Vehicle Insurance*, Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, volume 62. Disponível em: http://dx.doi.org/10.11118/actaun201462020383
6. Nelder, J.A. & McCullagh P. (1989).*Generalized Linear Models*, 2nd edition, Chapman & Hall,London.
7. Ohlsson, E. & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*, EAA series/EAA Lecture Notes, Springer.
8. Smyth, G., Hu, Y., Dunn, P., Phipson, B. & Chen, Y. (2015). *Statmod: An R package for statistical modeling [Online].* Disponível em: https://cran.r-project.org/web/packages/statmod/index.html
9. Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, 4th edition, Springer.