UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

CIÊNCIAS ATUARIAIS

MODELOS LINEARES GENERALIZADOS APLICADOS À PRECIFICAÇÃO EM SEGURO SAÚDE

Pedro Ivo Pires Machado
Rafael Santos Calzavara

Rio de Janeiro
2013

Pedro Ivo Pires Machado
Rafael Santos Calzavara

MODELOS LINEARES GENERALIZADOS APLICADOS À PRECIFICAÇÃO EM SEGURO SAÚDE

Projeto de Graduação apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências Atuariais.

Orientadora:

Profª. Thaís Cristina Oliveira da Fonseca

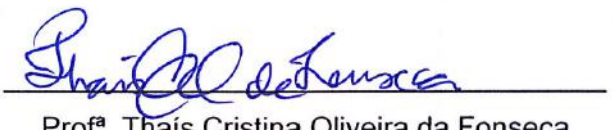
Rio de Janeiro
Abril 2013

Pedro Ivo Pires Machado
Rafael Santos Calzavara

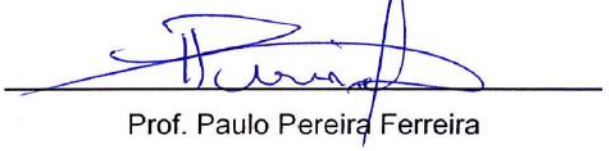
MODELOS LINEARES GENERALIZADOS APLICADOS À PRECIFICAÇÃO EM SEGURO SAÚDE

Projeto de Graduação apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências Atuariais.

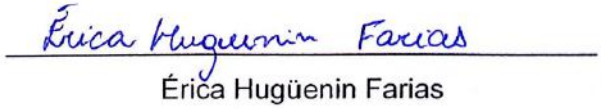
Examinado por:



Profª. Thaís Cristina Oliveira da Fonseca



Prof. Paulo Pereira Ferreira



Érica Hugüenin Farias

Rio de Janeiro
Abril 2013

Machado, Pedro Ivo e Calzavara, Rafael

Modelos Lineares Generalizados aplicados à Precificação em Seguro Saúde/ Pedro Ivo Machado e Rafael Calzavara – Rio de Janeiro; UFRJ/ Instituto de Matemática, 2013.

IX, 69 p.: il.; 29,7cm

Orientador: Thaís Cristina Oliveira da Fonseca

Projeto de Graduação – UFRJ/ IM/ Ciências Atuariais, 2013

Referências Bibliográficas: p. 64.

1. Modelos de Regressão. 2. Seguro Saúde 3. Modelos Lineares Generalizados. I Fonseca, Thaís Cristina Oliveira da. II. Universidade Federal do Rio de Janeiro, UFRJ, Ciências Atuariais. III. Título.

DEDICATÓRIA

Dedicamos este projeto às nossas famílias e amigos, que de formas diferentes e igualmente importantes nos incentivaram e acreditaram no nosso potencial.

AGRADECIMENTOS

Agradecemos a colaboração do corpo docente do Departamento de Métodos Estatísticos pelos ensinamentos de todos esses anos, em especial à Profª Thaís Fonseca pela paciência e compreensão com relação à todos os obstáculos encontrados. Um agradecimento especial também para Érica Hugüenin, que nos ajudou na obtenção da base de dados e contribuiu de diversas formas para o sucesso do trabalho.

“A modelagem em Ciência ainda permanece, pelo menos em parte, uma arte.”

(MC CULLAGH e NELDER, 1989)

RESUMO

Com o crescimento do acesso aos planos de assistência à saúde no Brasil, aliado à alta competitividade existente entre as empresas atuantes no mercado de Saúde Suplementar, o processo de formação do preço do Seguro Saúde exige uma alta sofisticação técnica com o objetivo de melhorar a acurácia na previsão dos gastos médicos decorrentes de eventos associados à condição de bem-estar do ser humano. Apesar da quase impossível missão de se estabelecer um preço à vida, este trabalho propõe uma discussão sobre técnicas de previsão de sinistros no segmento Saúde onde, a partir de uma base de dados provenientes de uma seguradora atuante neste ramo de negócio, fez-se uso de Modelos de Regressão como forma de associar as principais características do segurado que influenciam no valor total do sinistro. Obedecendo aos critérios básicos usualmente adotados no processo de modelagem, foram realizados tratamentos na base de dados, bem como uma pormenorização de todas as suas componentes e suas possíveis influências no modelo, através da técnica estatística de Análise Exploratória de Dados. Após o trabalho de entendimento da base, foram ajustados os chamados Modelos Clássicos de Regressão, utilizando critérios de seleção e bondade do ajuste, baseando-se em suposições de Normalidade e Independência do fenômeno estudado. Por fim, como extensão dos modelos clássicos, utilizaram-se Modelos Lineares Generalizados, que necessitam de premissas mais flexíveis para a sua aplicabilidade, de modo a melhorar a acertividade na previsão de novas observações. Foram eliciados também outros procedimentos de forma a melhorar este processo, como a utilização de metodologias alternativas e enriquecimento da base de dados.

Palavras-chave: Modelos de Regressão, Seguro Saúde, Modelos Lineares Generalizados.

ABSTRACT

The constant growth of the health care plans in Brazil, associated with the high competitiviness between the most active companies in the Private Health Insurance's market, made the pricing process of Health plans demands a high technical sophistication in order to improve the accuracy in predicting medical expenses coming from events associated with the welfare of the individual. Despite the almost impossible task of establishing a price on life, this paper proposes a discussion of prediction techniques for claims in the Healthcare Insurance Business where, from a database from an Insurer Company, Regression Models were used as a way of associate the main characteristics of the insured that influences the total claim amount . Following the basic criteria usually adopted in the modeling process, treatments were performed on the data base, as well as a detailing of all its components and their possible influences on the model through an Exploratory Data Analysis. After understanding the data base, were adjusted the so-called Classic Regression Models, using selection and goodness of fit criteria, based on assumptions of normality and independence of the studied phenomenon. Finally, as an extension of the classical models, were used generalized linear models, which require more flexible assumptions for its application, looking for an improve of accuracy in predicting new observations. Also other procedures were elucidated to improve this process, as the use of alternative methodologies and enrichment of the database.

Keywords: Regression Models, Healthcare Insurance, Generalized Linear Models.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Planos Privados de Assistência à Saúde no Brasil

Nos últimos anos o mercado de Saúde Suplementar brasileiro apresentou forte tendência de crescimento, decorrente do aumento da renda média salarial, o que possibilitou a inserção de classes mais baixas ao acesso de serviços dos mais variados tipos, incluindo-se os planos de assistência à saúde. De acordo com a Lei nº 9656/98, entende-se por plano de assistência à saúde como:

> (...) prestação continuada de serviços ou cobertura de custos assistenciais a preço pré ou pós estabelecido, por prazo indeterminado, com a finalidade de garantir, sem limite financeiro, a assistência à saúde, pela faculdade de acesso e atendimento por profissionais ou serviços de saúde, livremente escolhidos, integrantes ou não de rede credenciada, contratada ou referenciada, visando a assistência médica, hospitalar e odontológica, a ser paga integral ou parcialmente às expensas da operadora contratada mediante reembolso ou pagamento direto ao prestador, por conta e ordem do consumidor. (Lei nº 9656/98)

Podemos dizer que um plano de assistência à saúde é um contrato de cobertura financeira que garante ressarcimento em caso de oneração decorrente de procedimentos médicos. Este contrato pode ser enquadrado como um tipo de seguro, que é definido como um contrato onde um Segurador se submete à indenização de uma pessoa (segurado), mediante ao pagamento de um prêmio, de prejuízos decorrentes de eventos futuros.

Como os eventos médicos (doenças, acidentes, etc) e seu custo (exames médicos, cirurgias e outros procedimentos) são de natureza incerta, utiliza-se a teoria das probabilidades para determinar o quanto um determinado indivíduo precisa pagar por mês para ter a garantia de cobertura de sua "saúde", ou seja, o valor do prêmio.

O Mercado de Saúde Suplementar brasileiro é formado por diferentes tipos de estruturas empresariais que compartilham o objetivo de cobertura de assistência à saúde, dentre eles: cooperativas médicas, entidades de auto-gestão e seguradoras. Como a estrutura deste mercado é predominantemente uma concorrência imperfeita, ou seja, uma estrutura onde não existe parcimônia na distribuição equitativa de recursos, surge a necessidade de intervenção governamental para corrigir as chamadas 'falhas de mercado'. Essa intervenção surge com a criação da Agência

Nacional de Saúde (ANS), através da Lei nº 9961/2000, que define a agência como sendo:

> (...) autarquia sob regime especial, vinculada ao Ministério da Saúde,(...), como órgão de regulação, normatização, controle e fiscalização das atividades que garantam a assistência suplementar à saúde.". (Lei nº 9961/00)

O impacto financeiro dos referidos planos na área da Saúde Pública apresenta aspectos muito peculiares, devido à sua função social de garantia do bem-estar do segurado. Assim, a formação do preço é de fundamental importância para o equilíbrio entre assistência ao mercado consumidor de maneira sólida, além da solvência das pessoas jurídicas ofertantes deste serviço, viabilizando a dinâmica e a sobrevivência do mercado.

1.2 Técnicas de precificação em Saúde: práticas do mercado

Com a ideia de formação de preço em mente, existem inúmeros métodos aplicados à este problema. Algumas empresas fazem uso da sinistralidade da carteira na definição do preço, baseada em experiências anteriores de contratos semelhantes. Outra metodologia utilizada utiliza as características do segurado, que influenciam diretamente no custo total de sinistro para um determinado período, condizente do estabelecido na vigência do contrato. Essas características são então dispostas em uma tabela comparativa, em que a cada mudança de característica teremos um acréscimo (se positivo) ou um decréscimo (se negativo) no valor do sinistro médio. Assim, estabelecemos uma comparação de custo de sinistro para diferentes características, de modo que consigamos cobrar preços justos para cada classe, obedecendo ao principio de homogeneidade do seguro. O principal problema desta metodologia é a suposição de independência entre essas características, o que nem sempre (ou quase nunca), é verdadeira.

Dada a estrutura de correlação entres as características do segurado, é natural a busca de uma modelagem para o montante de sinistro em que sejam abarcadas não só essas características em si, mas também uma estrutura de correlação entre elas.

A utilização de Técnicas de Regressão, onde a variação de uma variável resposta pode ser explicada por um conjunto de covariáveis, se mostra bastante atrativa neste tipo de problema.

1.3 Objetivo

O presente trabalho tem por objetivo oferecer uma discussão sobre alternativas na modelagem em seguros saúde. Os modelos de precificação amplamente utilizados no mercado se baseiam no método chamado Método de Relatividades, onde através de algumas variáveis caracterizadoras do segurado (características de interesse) avaliadas individualmente estabelecem o prêmio médio a ser cobrado, não levando em consideração um efeito conjunto das mesmas.

Dada essa limitação, buscaremos o ajuste de modelos que explicam a variação do sinistro através de Modelos de Regressão ou seja, através de variáveis que subdividem a população segurada nas chamadas “classes de riscos”, podendo determinar o valor de sinistro esperado para cada classe, garantindo a homogeneização dos riscos e uma estrutura correlacionada de variáveis para a determinação do prêmio. Assim, podem ser encontradas as características do segurado que mais influenciam o preço do seguro, assim como o valor a ser cobrado por ele.

Este trabalho foi organizado didaticamente em capítulos que serão apresentados a seguir.

- O Capítulo 2 apresenta a base de dados utilizada no estudo, caracterizando as variáveis utilizadas nos modelos ajustados, assim como a sua importância no processo de precificação. É realizada também uma Análise Exploratória de Dados de modo a obter uma percepção da influência dos fatores de risco no modelo, além da natureza da variável resposta estudada.
- O Capítulo 3 apresenta a modelagem utilizando Modelos Clássicos de Regressão, explicando a metodologia utilizada e o seu desempenho com a base de dados. Como extensão aos modelos clássicos, foram aplicados os chamados Modelos Lineares Generalizados, seguindo a mesma estrutura de

exposição da metodologia e análise de desempenho.

- Sem a pretensão de esgotar o tema escolhido, o Capítulo 4 apresenta as considerações finais deste estudo.

1.4 Motivação

Em nossa breve experiência no mercado atuarial, o funcionamento de certos processos de precificação despertou nosso interesse, sendo o principal deles o de carteiras de saúde, que ainda carecem de sofisticação técnica apropriada.

Outros ramos de negócios em seguros utilizam os Modelos Lineares Genelarizados (MLG) para a modelagem de sinistro, tomando como exemplo o ramo de seguro de automóveis. Assim, nos questionamos sobre a adequabilidade da utilização de um ajuste usando os MLG para precificação de planos de assistência à saúde, tornando-se, então, o enfoque deste trabalho.

2 BASE DE DADOS

2.1 Composição da Base

O primeiro passo na construção do modelo é o entendimento da base de dados disponível para o estudo. A estrutura dos dados varia de empresa para empresa, de acordo com suas políticas internas de gestão de informação e planejamento estratégico. Assim, cria-se uma particularidade sobre qual tipo de modelo pode ser aplicado em cada instituição e qual a interpretação será resultante dessas análises. A dinâmica com que os dados são aproveitados é hoje uma grande vantagem competitiva, e se bem aproveitados fornecem modelos representativos da realidade com alto grau de precisão, elevando o patrimônio intelectual da companhia e sua experiência no segmento em que atua.

No mercado de saúde os dados são obtidos através da documentação que acompanha um determinado evento, seja internação, consultas, etc. Como existem uma quantidade considerável de variáveis possíveis a serem aproveitadas, as informações passam por um pré-tratamento, onde são resumidas e dispostas em formatos palatáveis para usos específicos. Para o presente estudo foi utilizada uma fração da base de dados de uma companhia seguradora especializada no segmento de seguro saúde, totalizando três anos de experiência no mercado. Os dados provém de planos do segmento empresarial, ou seja, planos coletivos vendidos à pessoas jurídicas para cobertura de seus funcionários e/ou dependentes. É importante frisar que os dados foram alterados de maneira a manter a confidencialidade das estratégias da companhia, mas sem comprometer a análise dos dados. Foram também excluídas as observações com valores vazios.

Como o objetivo principal está na avaliação do sinistro, a variável resposta foi o custo total de sinistro avaliada no período de estudo para uma determinada empresa, ou seja, o gasto da seguradora com eventos de funcionários e/ou dependentes da pessoa jurídica contratante do plano. Á essa variável daremos o nome de severidade.

Estabelecida a variável resposta, determinaremos o conjunto de variáveis que

podem ser utilizadas como variáveis explicativas nos modelos (covariáveis). Foram selecionados sete fatores que teoricamente influenciam para o custo de sinistro, sendo a validade desta afirmativa a ser testada posteriormente no modelo. A seleção dessas variáveis foi feita de modo intuitivo utilizando conhecimento prévio das práticas correntes do mercado, paralelizando os conceitos principais avaliados à ramos de seguros em que a utilização dos Modelos Lineares Generalizados é mais utilizada.

A seguir será caracterizada cada variável a ser testada pelo modelo:

2.1.1 Período

Variável categórica que mostra em qual período estamos analisando a experiência de sinistros da carteira (em qual dos 3 anos). O efeito de variação do custo médio de sinistro devido à reajuste de preço e atualização do rol de procedimentos (a relação de eventos cobertos pelo plano) é medido através da inflação médica, o que poderia gerar interpretações errôneas já que períodos reajustados fornecem sinistros maiores. Dito isto, os valores de sinistros foram descontados de reajustes, anulando o efeito da inflação médica.

Para efeito de modelagem a empresa avaliada em períodos distintos são considerados como observações distintas.

Variável: Período		
Classe	Frequência	Frequência Relativa
Período 1	3330	33%
Período 2	3331	33%
Período 3	3368	34%
Total	**10029**	**100%**

Tabela 1 - Tabela de Frequência: Período

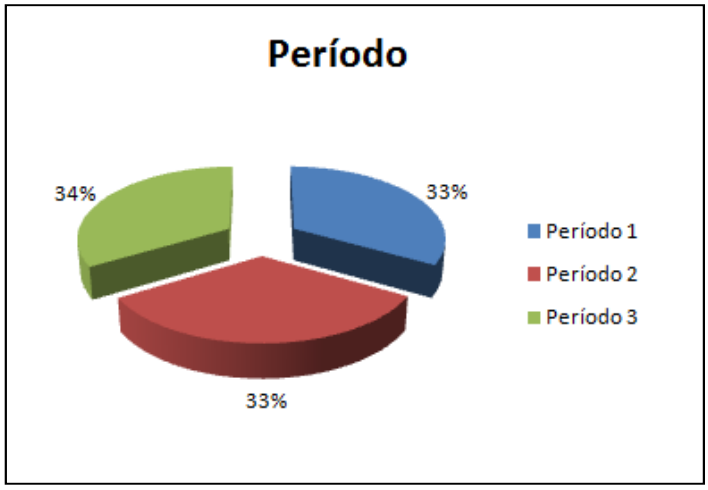


Figura 1 - Período

2.1.2 Região

Como os planos desta empresa possuem abrangência nacional, esta variável categórica torna-se imprescindível para a caracterização da variabilidade do sinistro por localidade, onde são empregados diferentes estratégias de venda e mercados consumidores diferenciados, respeitando a proporcionalidade de segurados em cada região e a característica das atividades exercidas, bem como a diversidade cultural em que está inserida. A base foi dividida em 13 regiões, como segue:

Variável: Região		
Classe	Frequência	Frequência Relativa
Região 1	843	8%
Região 2	955	10%
Região 3	900	9%
Região 4	941	9%
Região 5	586	6%
Região 6	601	6%
Região 7	503	5%
Região 8	791	8%
Região 9	796	8%
Região 10	544	5%
Região 11	877	9%
Região 12	890	9%
Região 13	802	8%
Total	**10029**	**100%**

Tabela 2 - Tabela de Frequência: Região

É natural também o interesse da companhia em ajustar o preço nas localidades de maior interesse estratégico, como por exemplo a região onde existe maior massa de segurados.

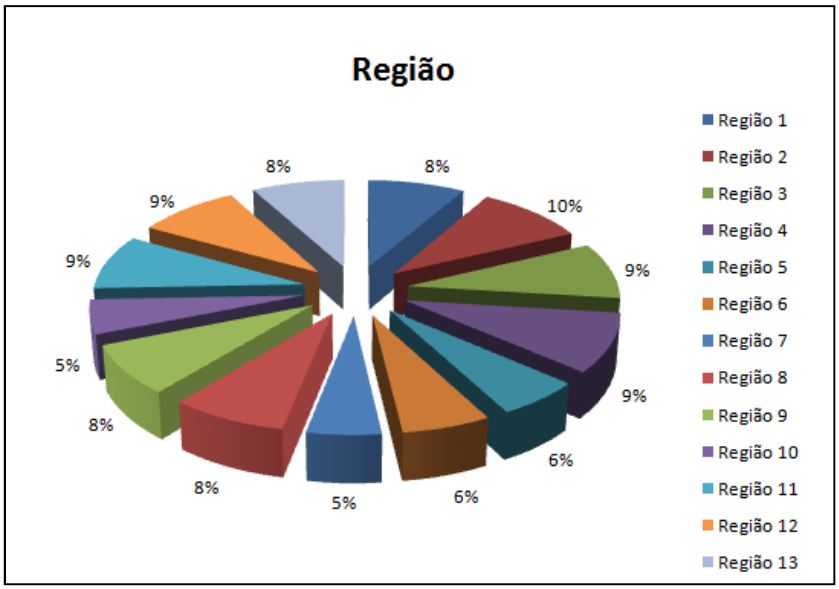


Figura 2 - Região

2.1.3 Plano

Na base constam dois tipos de planos: A e B. Esses dois planos diferem em tipos de cobertura e características contratuais. Esta variável possivelmente fornecerá respostas para questões do tipo: planos com mais coberturas fornecem maior custo para a seguradora ou, apesar de um plano ser mais caro que outro, existe maior custo proveniente de planos mais simples.

Variável: Plano		
Classe	Frequência	Frequência Relativa
Plano A	4892	49%
Plano B	5137	51%
Total	**10029**	**100%**

Tabela 3 - Tabela de frequência: Plano

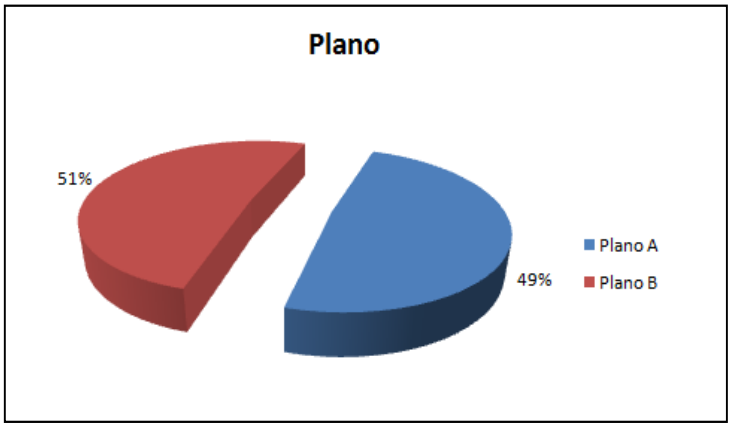


Figura 3 - Plano

2.1.4 Sexo

Essa variável nos mostra como a distinção dentre os procedimentos médicos entre homens e mulheres afetam o custo do sinistro. A especificidade dos atendimentos, ou seja, procedimentos utilizados exclusivamente por determinado gênero, pode ser a principal causa da diferença do sinistro. Citamos como exemplo o fato da gestação, cujos gastos estão alocados exclusivamente no sexo feminino. É interessante observar o impacto social associado à esta variável, pois caso uma precificação fosse feita com base exclusivamente na mesma, acarretaria um processo de anti-seleção, onde um gênero pagaria mais por um mesmo plano de assistência à saúde. A alternativa está em padronizar os procedimentos para os dois gêneros, gerando coberturas que não serão aproveitadas (planos com obstetrícia para homens, por exemplo).

Variável: Sexo		
Classe	Frequência	Frequência Relativa
Masculino	5034	50%
Feminino	4995	50%
Total	**10029**	**100%**

Tabela 4 - Tabela de frequência: Sexo

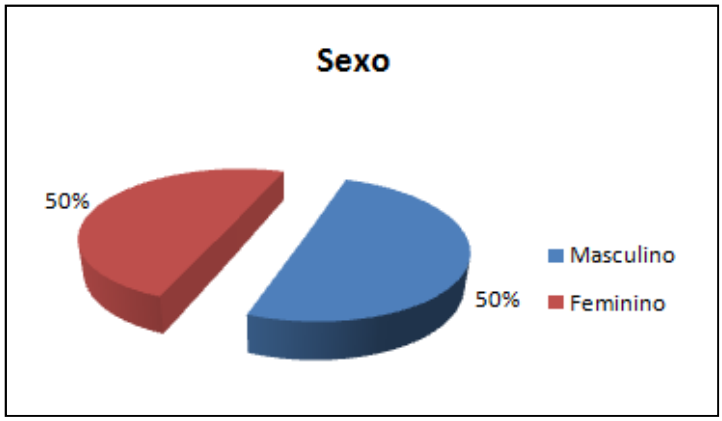


Figura 4 - Sexo

2.1.5 Tamanho do Grupo

Como estamos lidando com planos empresariais, o tamanho da empresa também será uma variável a ser analisada, o que pode trazer a percepção da relação entre o tamanho da empresa com o custo de sinistro.

Os agrupamentos foram realizados em categorias que variam de 1 a 5, onde 1 remete as empresas de menor tamanho e 5 as de maior contingente. Vale ressaltar que cada categoria possuem um número máximo e um número mínimo de funcionários.

Variável: Tamanho de Grupo		
Classe	Frequência	Frequência Relativa
Grupo 1	1715	17%
Grupo 2	2109	21%
Grupo 3	2161	22%
Grupo 4	1901	19%
Grupo 5	2143	21%
Total	**10029**	**100%**

Tabela 5 - Tabela de frequência: Tamanho do Grupo

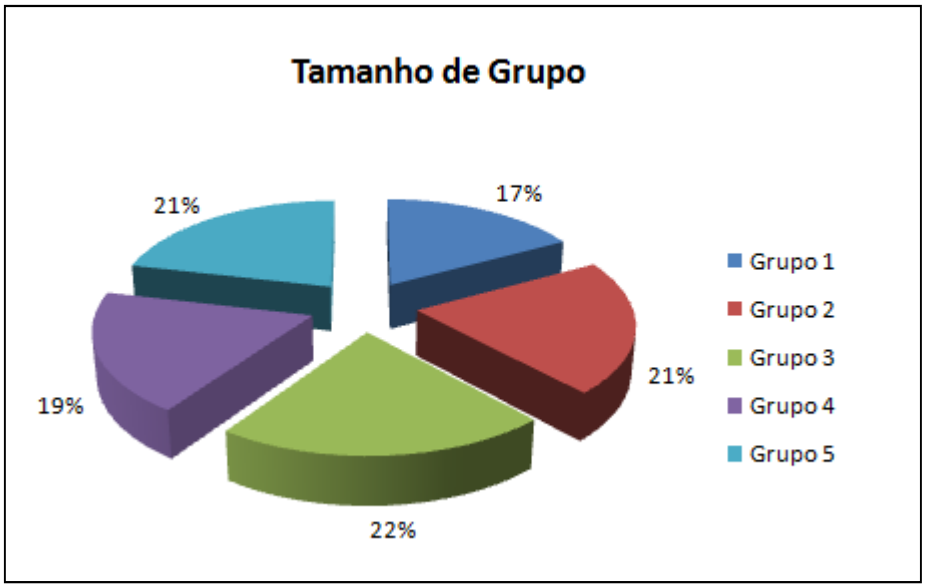


Figura 5 - Tamanho de Grupo

2.1.6 Faixa Etária

Talvez a mais importante variável a ser estudada, pois está diretamente ligada com a frequência de utilização do plano. Com o envelhecimento, é necessária uma maior atenção com os cuidados da saúde, o que consequentemente leva o indivíduo a utilizar mais o plano, gerando mais custo para o segurador. Assim, é intuitivo pensar que os mais velhos devem pagar mais caro pelo mesmo plano de assistência, já que teoricamente utilizariam mais o serviço. Esta afirmação gera um paradoxo, pois o crescimento do preço do seguro não seria viável para uma população em que a capacidade produtiva (e consequentemente a renda) está em declínio. O reajuste por faixa etária já é limitado pelo órgão regulador, prevenindo o consumidor de reajustes abusivos e incompatíveis com a sua capacidade de arcar com tal despesa. Mesmo com essa limitação regulatória, o modelo ideal deve ser capaz de mensurar essa variação por idade, o que possibilita a criação de mecanismos para contornar este problema. As idades foram agrupadas em 16 faixas etárias, como segue:

Variável: Faixa Etária		
Classe	Frequência	Frequência Relativa
00-01	695	7%
02-18	761	8%
19-23	730	7%
24-28	745	7%
29-33	752	7%
34-38	742	7%
39-43	734	7%
44-48	731	7%
49-53	723	7%
54-58	687	7%
59-63	650	6%
64-68	569	6%
69-73	482	5%
74-78	402	4%
79-83	325	3%
84->>	301	3%
Total	**10029**	**100%**

Tabela 6 - Tabela de frequência: Faixa Etária

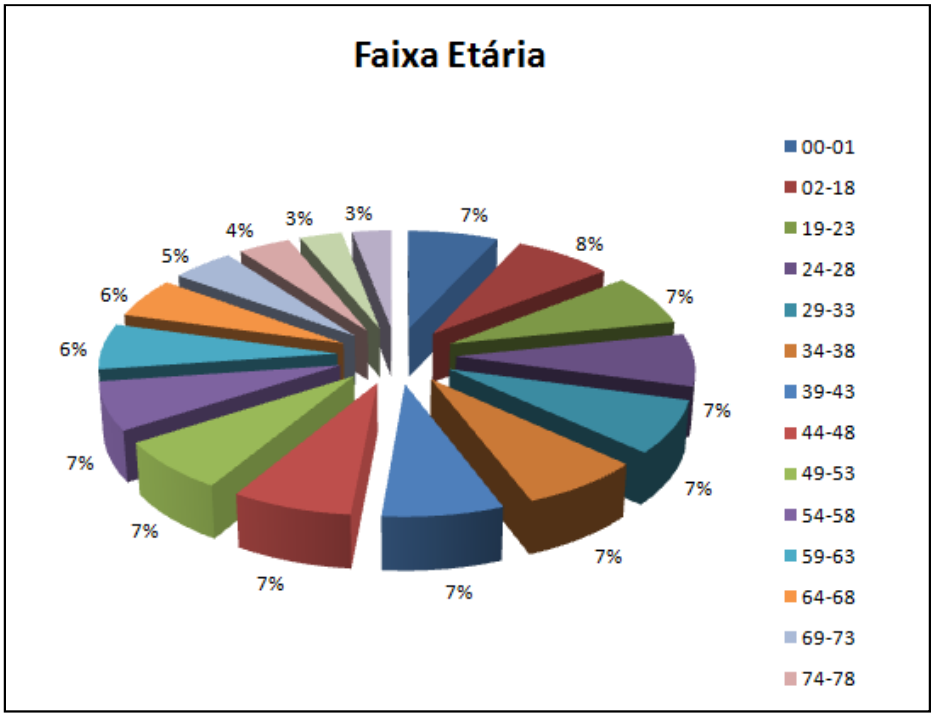


Figura 6 - Faixa Etária

2.1.7 Exposição

A variável Exposição expressa o número de vidas das instituições contratantes do seguro no período avaliado na base de dados suscetíveis à ocorrência de sinistros. Como as celebrações dos contratos de planos de assistência à saúde ocorrem em datas distintas uma das outras, o cálculo da exposição leva em consideração a proporção da quantidade de segurados em relação ao tempo de vigência do contrato. Para ilustrar o cálculo, supõe-se uma empresa com 10 segurados para um determinado tipo de plano com vigência total de um ano, sendo 5 segurados possuindo contratos vigentes a partir do dia 1º de junho de um ano de referência (contratos do do tipo I) e os outros 5 funcionários com início de vigência em 1º de agosto neste mesmo ano (contratos do tipo II). Ao final do ano analisado, o número considerado de expostos será:

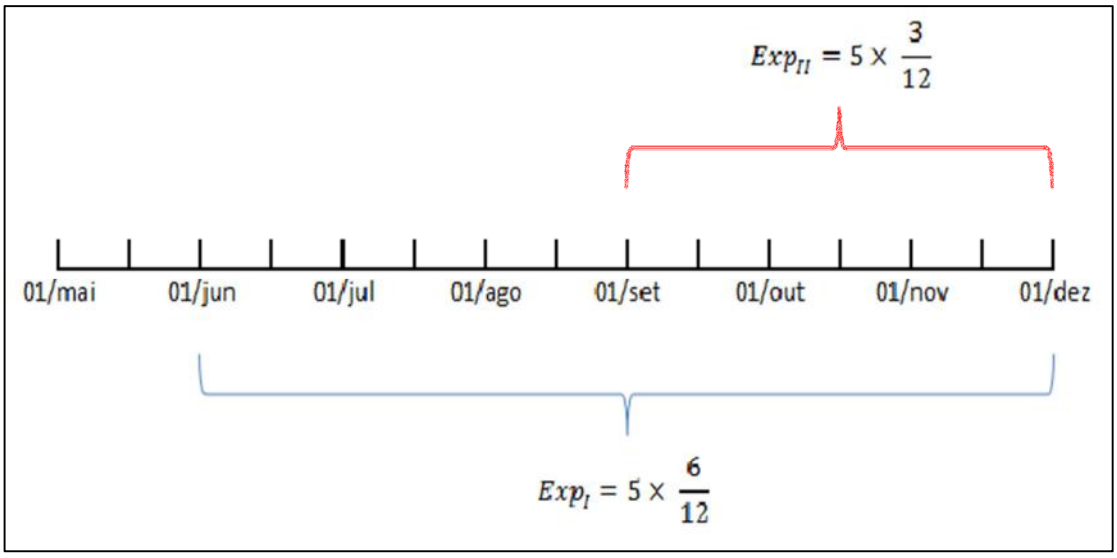


Figura 7 - Expostos

2.1.8 Frequência

A variável frequência é definida como o número de sinistros ocorrido para uma dada observação. Como cada unidade sob estudo é uma empresa contratante do plano de assistência à saúde, produzindo um determinado montante de sinistro, não só é importante o entendimento de quanto “custa” cada evento médico, mas sim quando e com que frequência eles ocorrem. Alguns modelos de precificação utilizam duas modelagens trabalhadas em paralelo: a primeira, um modelo para dados de contagem utilizando distribuições adequadas, como a Bernoulli, Binomial, Poisson, etc ; enquanto a segunda traz a noção quantitativa do valor do sinistro, atribuindo distribuições contínuas positivas para o seu ajuste, como a Gamma, Normal, Normal Inversa, etc. Estes modelos são usualmente chamados de Modelos de Frequência e Severidade, atribuindo o valor do sinistro associado à sua probabilidade de ocorrência.

Como a base de dados utilizada no estudo não possui tal informação, concentraremos a modelagem e as interpretações dos modelos resultantes somente na parte de Severidade, apesar de reconhecer a importância da quantidade de sinistros para o desenvolvimento de modelos mais acertivos.

2.2 Análise Exploratória

Como primeiro passo da modelagem, é necessário conhecermos os comportamentos das variáveis sob análise. Assim, procederemos uma análise exploratória de dados para inferir sobre quais modelos podem ser mais aplicáveis, assim como a escolha da distribuição da variável resposta. Como forma de ilustrar as propriedades mais importantes para o ajuste dos modelos propostos, utilizaremos o gráfico de box-plot. Segundo Montgomery:

> (...) o diagrama de caixa (box-plot) é uma apresentação gráfica que descreve simultaneamente várias características importantes de um conjunto de dados, tais como centro, dispersão, desvio da simetria e identificação das observações que estão surpreendentemente longe do seio dos dados (os chamados "outliers"). (MYERS, MONTGOMERY, *et al.*, 2010)

Para a análise da distribuição da variável resposta, construiremos um histograma, que, segundo a definição fornecida por Montgomery:

> (...) o histograma, assim como o diagrama de ramo-e-folhas, fornece uma impressão visual da forma da distribuição da medidas, assim como informação sobre a dispersão dos dados." (MYERS, MONTGOMERY, *et al.*, 2010)

2.2.1 Variável Resposta

Primeiramente analisaremos a variável resposta: a severidade. Como devemos considerar somente os riscos vigentes por segurado, devido à possibilidade de extensão do modelo para contratos individuais, dividimos o sinistro pela quantidade de expostos de cada observação, obtendo assim uma variável resposta como sendo a severidade por unidade exposta. A seguir apresentamos o box-plot para esta variável:

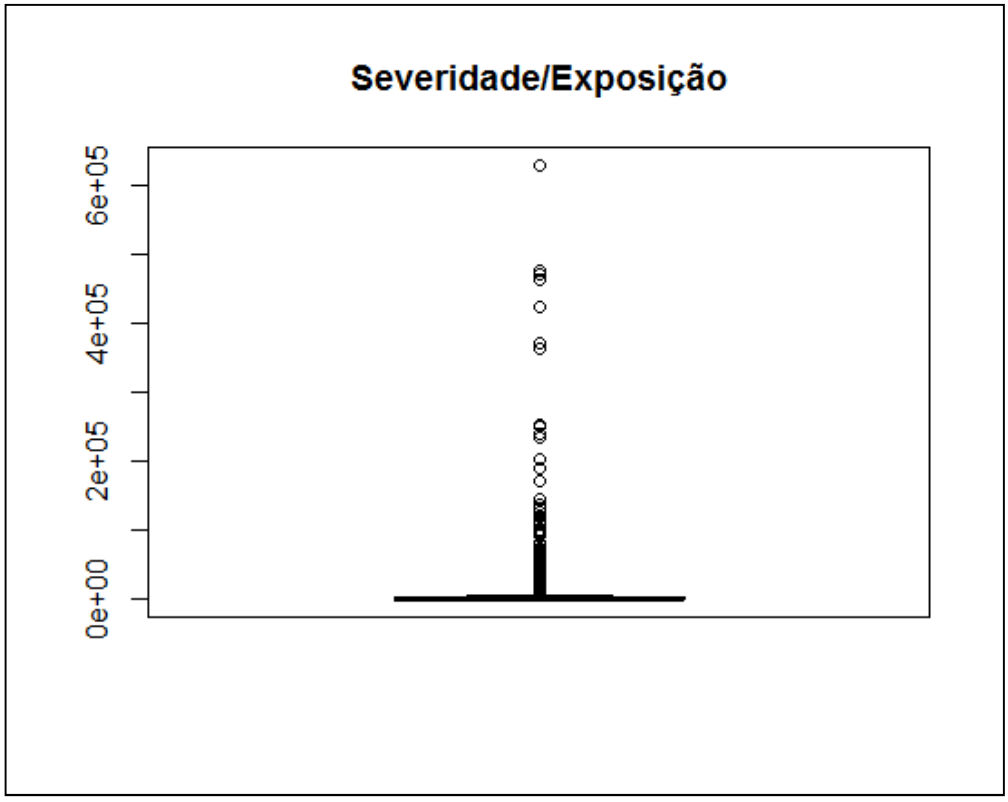


Figura 8 - Box-Plot Severidade/Exposição

O boxplot da variável resposta fornece algumas interpretações importantes. Podemos observar uma grande quantidade de sinistros acima do limite de 3 desvios, o que pode ser um indício da presença de muitos outliers. Este comportamento é plausível para a linha de negócio estudada, dado que o seguro saúde é baseado em um rol de procedimentos, ou seja, uma lista de procedimentos médicos mínimos a serem cobertos pelo plano de assistência à saúde, aumentando a variabilidade da perda financeira associada ao sinistro. Dado este fato, podem haver tanto eventos onde os custos não são muito variáveis como consultas médicas, e eventos onde o gasto é considerável, como internações em centros de tratamento intensivo e cirurgias de alta complexidade. Como a base não é segregada por procedimentos, este comportamento era esperado.

A partir dessa primeira análise faremos um box plot ocultando os outliers de maneira a facilitar a visualização:

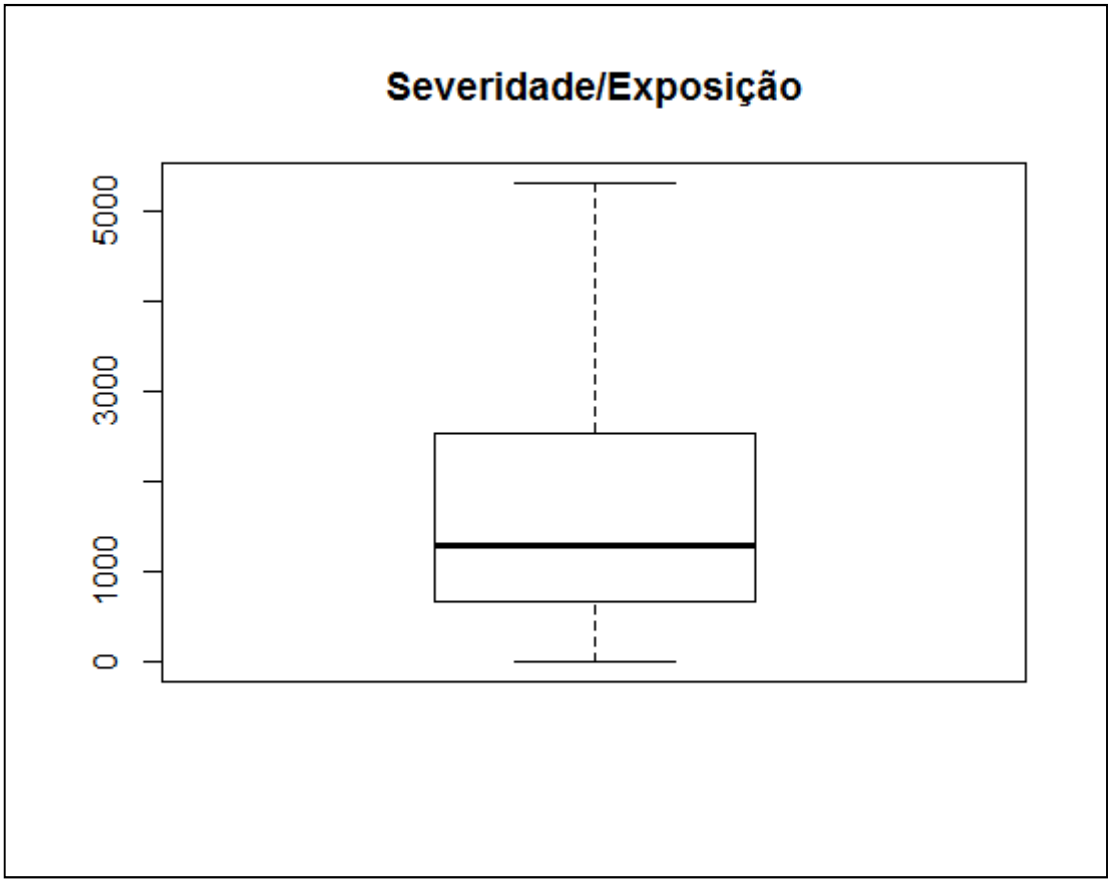


Figura 9 - Box-Plot Severidade/Exposição sem outliers

Elaborando um histograma da variável resposta, percebe-se uma assimetria positiva

acentuada, interpretado como muitos sinistros de baixo valor e poucos sinistro de alto valor. Essa disposição dos dados nos auxilia na escolha de uma distribuição assimétrica para a variável resposta.

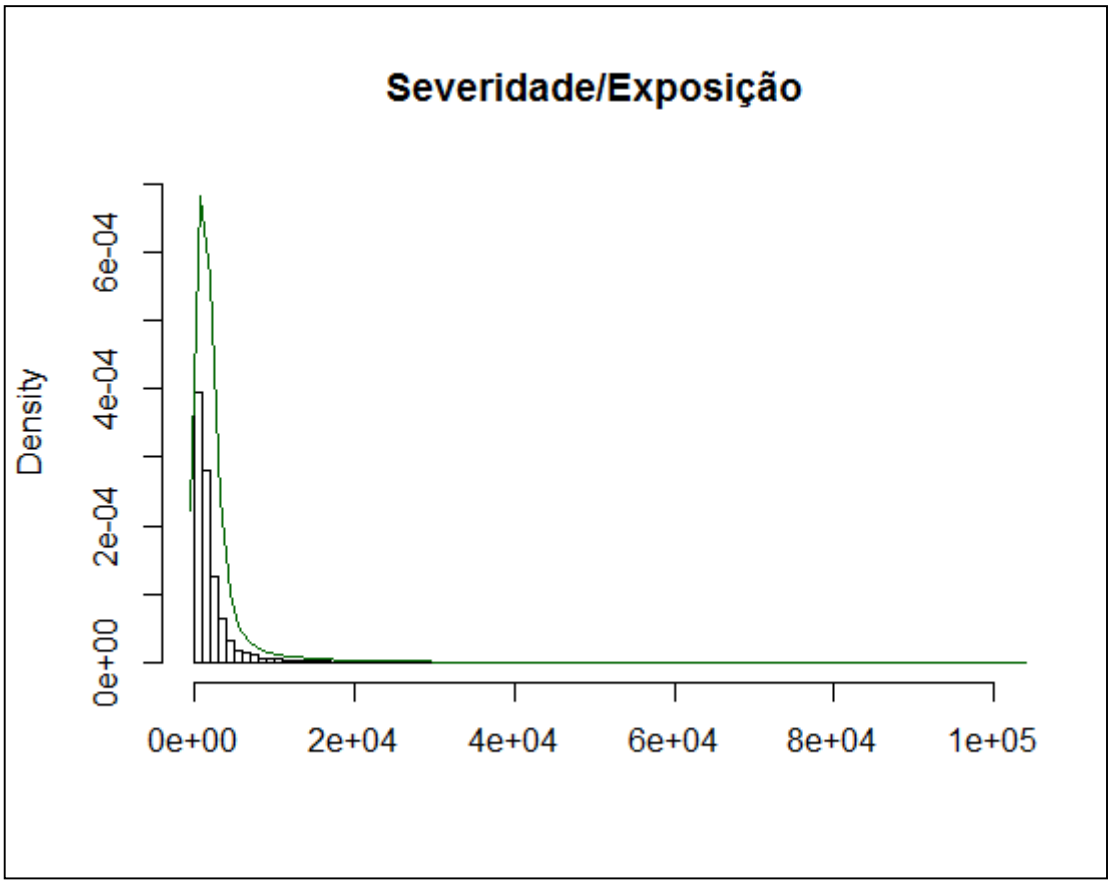


Figura 10 - Histograma - Severidade/Exposição

Para as análises das possíveis covariáveis do modelo, dadas as suas características categóricas, serão construídos gráficos de box-plot de cada fator em relação à variável resposta, ocultando os possíveis outliers para melhor visualização.

2.2.2 Período

Observando o box-plot da variável resposta segregado por período, observamos uma distribuição semelhante nas três categorias. Como os dados foram tratados de forma a anular o efeito da inflação médica, ou seja, do aumento do custo dos

procedimentos médicos ao longo do tempo, é esperado que as observações apresentem comportamentos semelhantes para os diferentes instantes analisados. Podemos notar uma leve distorção no terceiro período, mas não de maneira a modificar substancialmente o modelo.

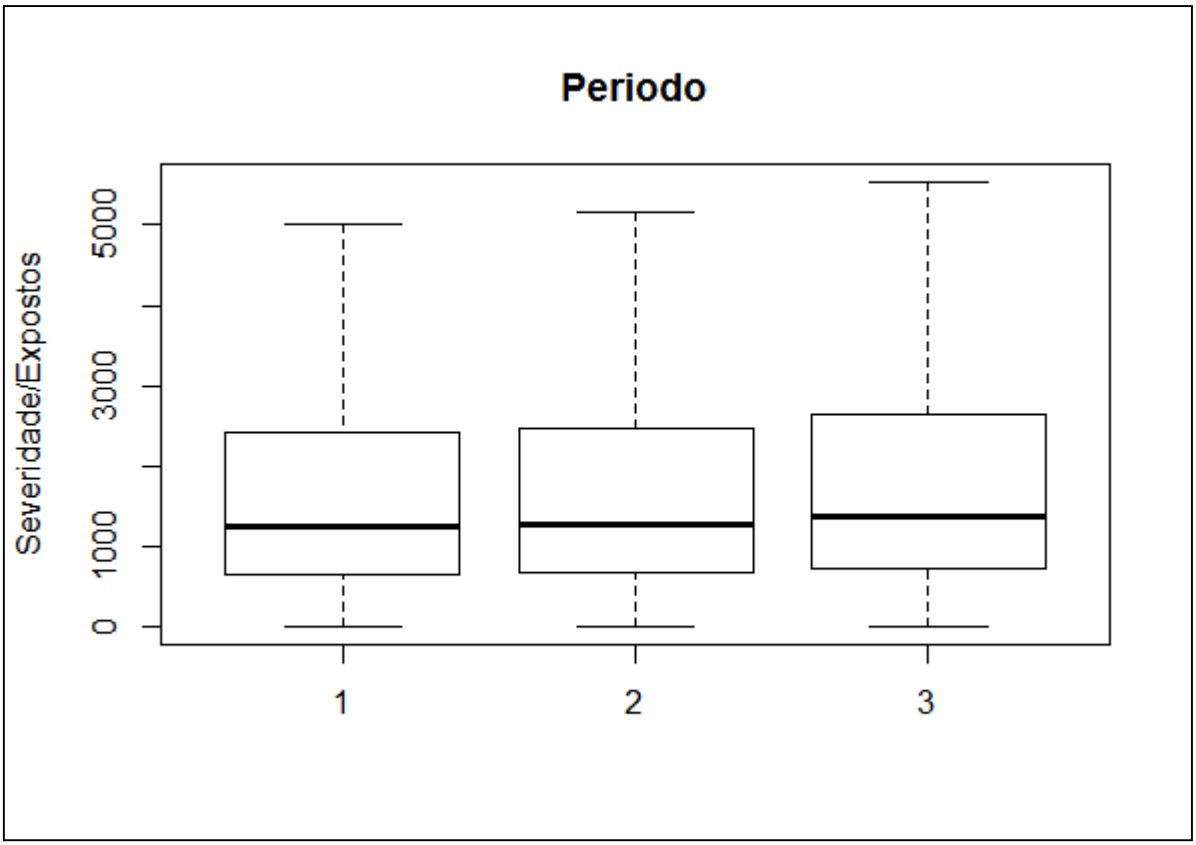


Figura 11 - Box-Plot Período

Assim, podemos supor que a evolução temporal disponível na base de dados não afeta o modelo de maneira significativa, mas sim ilustra um mesmo padrão de comportamento dos sinistros ao longo do tempo.

2.2.3 Região

A variável região apresenta bastantes diferenças em relação à severidade, o que é um indicativo de um fator importante na modelagem. Podemos atribuir essa variação ao fato da dinâmica de vendas ser diferenciada em cada localidade, respeitando as especificidades de cada mercado. Podemos inferir que as regiões de maior significância estratégica para a companhia carecem de ajustes mais sofisticados,

pois em uma área de alta competitividade a vantagem não está somente no menor preço, mas sim a capacidade da companhia de alinhar uma sólida solvência com uma precificação bem distribuída entre os fatores de risco.

Podemos observar que em algumas região a dispersão parece bem maior que outras (por exemplo, comparando-se as regiões 2 e 10). Isso pode ser explicado pela diferença no número e tipo de procedimento utilizado em cada região, onde uma determinada cobertura pode ser oferecida somente em determinadas regiões.

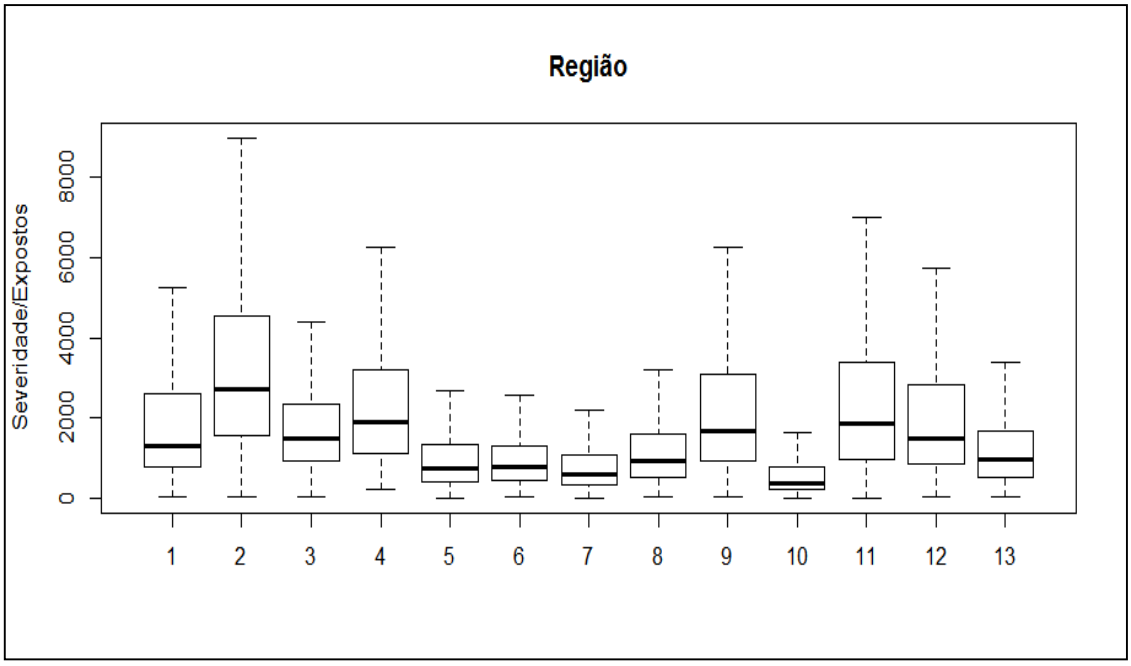


Figura 12 - Box-Plot Região

2.2.4 Plano

O tipo do plano oferece interpretações pertinentes, e seus efeitos a serem captados pelo modelo. Pela diferença contratual entre os dois produtos é esperado que o que ofereça uma cobertura mais ampla teoricamente possua maior volume sinistro. Assim, podemos observar uma mudança de comportamento entre os dois produtos, sendo que o produto B possui uma amplitude maior que o A.

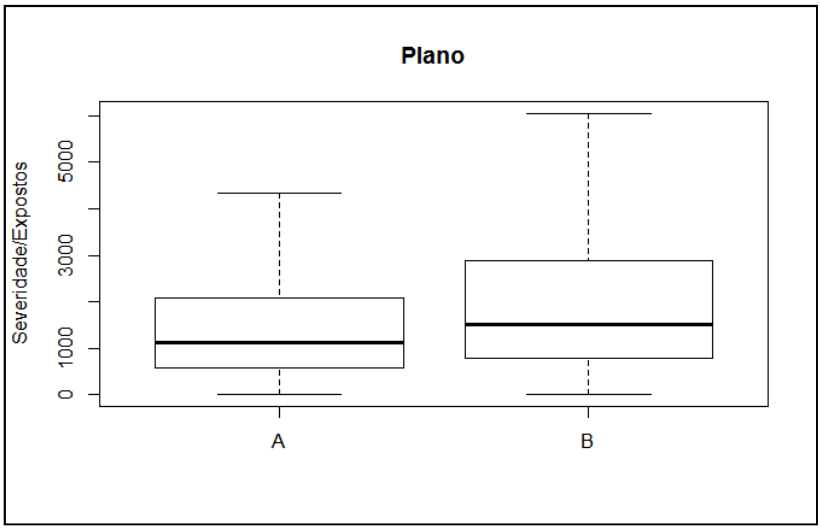


Figura 13 - Box-Plot Plano

A estrutura contratual do produto não é levada em consideração nesta modelagem, dado que a base de dados não possui segregação por coberturas.

2.2.5 Sexo

Pela segregação da base por gênero, podemos ver uma clara diferença entre as categorias, onde o sexo feminino possui um comportamento diferenciado em relação ao masculino, como segue:

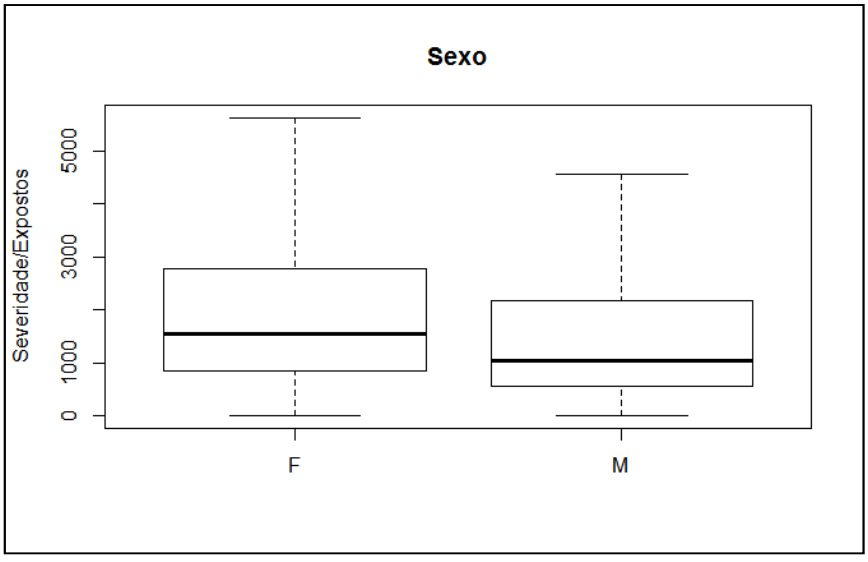


Figura 14 - Box-Plot Sexo

A diferença entre sexos pode existir devido à vários motivos, onde a gestação é um deles. Durante a gravidez é necessário um tratamento que envolve vários procedimentos médicos, o que acarreta um aumento no custo médico nesta fase da vida.

Assim como os seguros de automóveis, é natural uma diferenciação de risco por gênero. O modelo deve ser capaz de quantificar o quanto a variação da categoria influencia no custo do sinistro, sendo esta variável possivelmente relevante para o modelo.

2.2.6 Tamanho do grupo

O tamanho de grupo não ilustra diferenças relevantes com relação ao tamanho da empresa. O número de funcionários das empresas contratante do seguro aparentemente influencia somente o volume de sinistro, ou seja, as empresas com mais funcionários produzem mais sinistros.

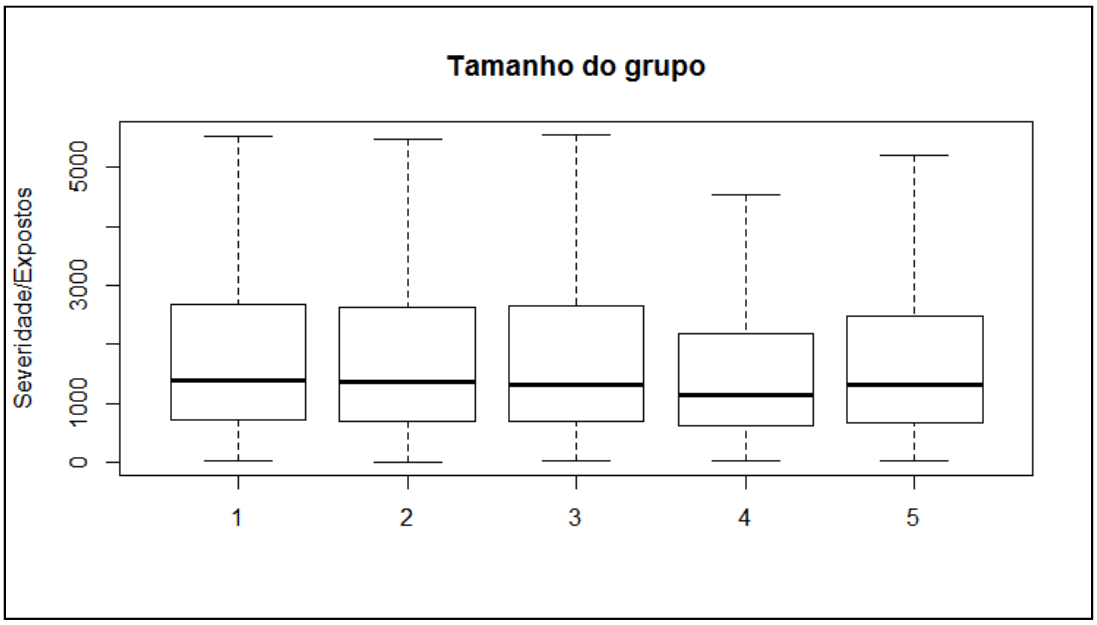


Figura 15 - Box-Plot Tamanho do Grupo

Como a variável resposta é a severidade por exposição, o volume de sinistro dado o tamanho da empresa é contrabalanceado com seu número de expostos, fornecendo valores de severidade semelhantes entre companhias de diferentes tamanhos.

2.2.7 Faixa Etária

A faixa etária é a variável categórica que fornece a maior relevância no modelo de precificação. Como a saúde de maneira geral está diretamente relacionada com a idade do indivíduo e o seu envelhecimento, observamos um aumento no sinistro a medida que observamos idades mais avançadas, como segue:

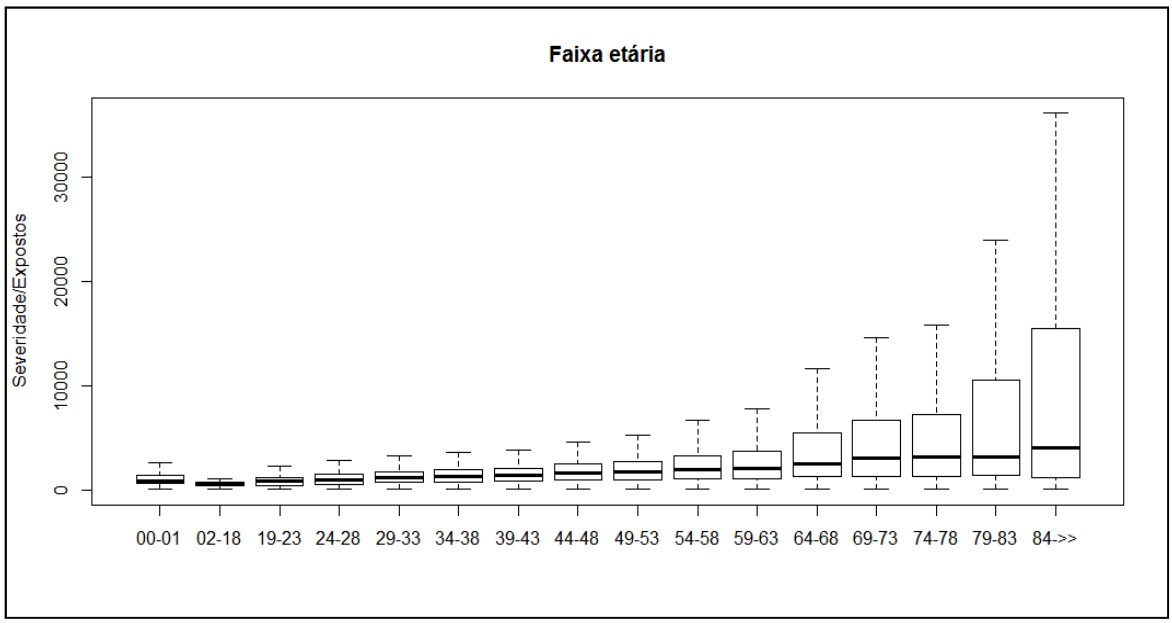


Figura 16 - Box-Plot Faixa Etária

Assim, podemos afirmar que a precificação se baseia fortemente na variável idade. É válido ressaltar também, além da relevância, a simplicidade de obtenção da variável, que não envolve análises complexas de quadros clínicos, mas um simples campo de data de nascimento.

3 MODELAGEM

3.1 Introdução

Segundo Duncan:

> (...)um modelo é uma abstração do mundo real que tenta capturar as alterações do complexo comportamento da natureza em uma simples formulação matemática. Um modelo pode ser expresso através de palavras, equações, diagramas e outras estruturas lógicas. (DUNCAN, 2011)

Sendo uma simplificação da realidade, o modelo considerado ideal não é capaz de capturar todos os efeitos inerentes do fenômeno estudado, mas sim os principais aspectos de sua natureza. O objetivo é obter um conhecimento de fácil manipulação que nos ajude a entender os aspectos macroinfluentes, além auxiliar as tomadas de decisões.

Para a escolha do melhor modelo são considerados todos os aspectos desenvolvidos no processo, desde a construção e tratamento da base de dados às análises provenientes dos ajustes estatísticos. O objetivo da pesquisa deve ser o mais claro possível, de modo a guiar o pesquisador na obtenção dos melhores resultados. Formulado o problema e construída a base de dados, algumas características são essenciais para um modelo ser considerado adequado, entre eles:

- Parcimônia. Como representação da realidade, o modelo precisa ser o mais conciso possível com um máximo de ganho, aproveitando-se apenas as variáveis mais influentes na variação da variável de interesse. A sofisticação excessiva, como a inclusão de variáveis pouco significativas, pode gerar um custo operacional alto para um ganho relativamente pequeno.
- Bondade do ajuste. Além da parcimônia, o modelo precisa se ajustar bem aos dados. Assim, utilizam-se métricas para essa qualidade de ajuste, como por exemplo o R^2 ajustado para os modelos clássicos de regressão, que avalia o quanto da variação das covariáveis influencia a variação da variável resposta.
- Capacidade Preditiva. Um dos objetivos principais do modelo, além da

compreensão dos fatores influentes sob estudo, é o de "simular" a realidade, ou seja, a partir de novas observações sermos capazes de obter valores o mais realistas possível. Assim, testamos essa capacidade a partir da comparação dos valores ajustados (preditos) com os valores efetivamente observados, estabelecendo critérios estatísticos para a avaliação dessa característica do modelo.

Para o presente trabalho, buscando obter as características citadas anteriormente, utilizaremos os critérios de seleção de modelos descritos por (MC CULLAGH e NELDER, 1989), baseados na metodologia desenvolvida por (BOX e JENKINS, 1976), separando a modelagem em 3 macroprocessos:

- Seleção do modelo: A partir das variáveis presentes na base de dados e de conhecimento prévio do ramo de negócio, baseando-se em estudos anteriores e percepções do mercado, faremos a construção do modelo de acordo com as técnicas estatísticas necessárias, utilizando critérios para a inclusão ou exclusão de variáveis de modo a obter o ganho ótimo possível com a informação disponível.
- Estimação: Calcular os parâmetros presentes no modelo, utilizando medidas de precisão dessas estimativas sob critérios de bondade de ajuste, através de algoritmos de estimação escolhidos convenientemente.
- Teste de adequação e Capacidade de Predição: Realizar análise de resíduos e testes de bondade de ajuste para verificar a qualidade do modelo, assim como inputar novas observações e comparar com os valores efetivamente observados, obtendo assim uma percepção da capacidade de previsão do modelo.

3.2 Modelos de Regressão Múltipla

3.2.1 Metodologia

Os Modelos Clássicos de Regressão surgiram através dos trabalhos de Gauss e Legendre no início do século XIX. Os conceitos de mínimos quadrados e a

distribuição Normal dos erros, aplicados ao estudo de corpos celestes, foram expandidos para outras ciências, e esta técnica foi amplamente utilizada desde então. A aplicabilidade desta tecnologia está atrelada à premissas essenciais de independência e homocedasticidade, o que por vezes não é possível em muitos fenômenos. Assim, para contornar o problema, foram criados artifícios que "forçam" essa estrutura, como por exemplo a transformação da variável resposta para obtenção da normalidade. Os dados de seguros são um exemplo da dificuldade de aplicação dessa técnica, devido à sua natureza assimétrica (muitos sinistros de baixo valor e poucos sinistros de alto valor).

Para a formulação do modelo, consideramos uma coleção de n observações independentes e identicamente distribuídas. Denotaremos como $Y_i$sendo o i-ésimo valor correspondente à variável resposta e X_{ij} como sendo o i-ésimo valor da j-ésima variável explicativa. Assim, montamos uma equação de forma que a variável resposta seja uma combinação linear das covariáveis ponderados por um vetor de parâmetros, somados a um termo correspondente ao erro. Supondo k covariáveis, obtemos a equação abaixo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \sum_{j=1}^{k} \beta_j x_{ik} + \epsilon_i$$

Considerando as n observações, obtemos um sistema de n equações. Para facilitar a visualização, utilizaremos a notação matricial:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \epsilon$$

Onde:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Consideraremos o modelo de regressão como sendo influenciado por duas componentes: a componente aleatória e a componente sistemática.

3.1.1.1 Componente Aleatória

Seguindo a estrutura de normalidade dos erros supostas pelo modelo e pelas propriedades da distribuição Normal, temos que o vetor de observações $\boldsymbol{y}$ são realizações da variável $\boldsymbol{Y}$, que possui distribuição Normal com os momentos:

$$E[Y] = \mu \qquad \text{e} \qquad cov[Y] = \sigma^2 \mathbf{I}$$

Sendo I a matriz identidade.

3.1.1.2 Componente Sistemática

Também chamado de preditor linear, esta componente é o input dos dados do modelo, ou seja, é a parcela da equação onde são consideradas as covariáveis selecionadas. O conceito de preditor linear será importante à medida que formularemos os Modelos Lineares Generalizados. Supondo um modelo com k fatores:

$$\boldsymbol{\eta} = \sum_{j=1}^{k} \beta_j x_{ik}$$

Os fatores podem ser dos tipos qualitativo ou quantitativo. Para as variáveis quantitativas basta substituir o valor correspondente do modelo. Para as variáveis qualitativas de vários níveis, como por exemplo a variável região (Região 1, Região 2, etc) utilizam-se variáveis denominadas dummy, que assumem o valor 1 caso pertença a um nível e 0 caso não pertença. Para uma variável com j níveis, teremos j variáveis dummy e consequentemente j covariáveis.

3.1.1.3 Suposições do Modelo

De acordo com a estrutura das suas componentes, pode-se elaborar um conjunto de premissas que regem a aplicabilidade desta técnica, ditando os procedimentos de

estimação e inferência. São eles:

- A variável resposta Y é função linear nos parâmetros das variáveis explicativas X;
- Os valores das variáveis explicativas são fixos;
- Os erros são independentes e normalmente distribuídos com média 0 e variância finita e constante, ou seja, $\epsilon \sim N(0, \sigma^2 \mathbf{I})$.
- Os erros são não-correlacionados.

3.1.1.4 Método de Estimação

Como o vetor paramétrico $\boldsymbol{\beta}$ é desconhecido, precisa-se estimá-lo. Assim, utilizaremos o método de Mínimos Quadrados, ou seja, o vetor $\boldsymbol{\beta}$ que minimiza a soma dos desvios ao quadrado, denotado por L.

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta)$$

3.1.1.5 Inferência

Para testar a significância das covariáveis, utilizaremos os testes de hipóteses comumente utilizados na regressão linear múltipla, nos atendo aos valores correspondentes das estatísticas e análise dos p-valores dos testes de hipótese correspondentes.

3.1.1.6 Bondade do Ajuste

Sobre a qualidade de ajuste do modelo, faremos uso do coeficiente de determinação múltiplo. Segundo (MONTGOMERY e PECK, 1982), o R² ajustado é uma medida de redução da variabilidade de Y obtida através do uso das covariáveis.

$$0 \leq R^2 \leq 1$$

Vale ressaltar que mesmo para valores de R² próximos de 1, essa propriedade não garante uma boa qualidade do modelo, como por exemplo a sua capacidade de predição.

3.1.1.7 Análise dos Resíduos

De modo a verificar as suposições citadas anteriormente, procederemos uma análise de resíduos e, através de gráficos, interpretar os efeitos do ajuste e a adequação do modelo. Também serão realizados testes para verificar a presença de outliers, observações influentes e leverages.

3.1.2 Aplicação

Após a Análise Exploratória de Dados e a descrição da metodologia a ser utilizada na modelagem, elaboramos modelos na tentativa de obter uma interpretação das significâncias das variáveis utilizadas no problema, bem como de obter predições para novos valores. Como previamente definido, a variável resposta é a razão entre a severidade e a exposição, de modo que a interpretação fornecida foi o custo que cada indivíduo gera em um plano de saúde contratado.

Como primeiro modelo, utilizamos a distribuição normal para a variável resposta, ou seja, aplicamos um modelo clássico de regressão. Em consonância com a Análise Exploratória de Dados, não incluímos a variável período, pois os Box-plots dessa variável indicaram pouca variação no modelo. Assim, o Modelo 1 foi construído da seguinte forma:

$$\frac{severidade}{exposição} \sim região + plano + sexo + fxetária + tamgrupo$$

Aplicando este modelo no software estatístico R, obtivemos os seguintes resultados:

```
Call:
lm(formula = (resp) ~ (regiao + plano + sexo + fxetaria + tamgrupo))

Residuals:
   Min     1Q Median     3Q    Max
-18808  -1673   -417    556 611965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2717.57     847.20   3.208  0.00134 **
regiao2         591.23     707.50   0.836  0.40336
regiao3       -1262.21     716.36  -1.762  0.07810 .
regiao4        -598.61     709.50  -0.844  0.39886
regiao5       -1174.54     807.45  -1.455  0.14581
regiao6       -1215.90     801.28  -1.517  0.12919
regiao7        -970.87     845.68  -1.148  0.25098
regiao8       -1289.27     739.83  -1.743  0.08142 .
regiao9        1897.23     738.53   2.569  0.01022 *
regiao10      -1572.81     825.36  -1.906  0.05673 .
regiao11        398.64     720.63   0.553  0.58015
regiao12        581.85     718.29   0.810  0.41793
regiao13      -1726.31     737.06  -2.342  0.01919 *
planoB          -17.91     299.21  -0.060  0.95227
sexoM          -154.31     298.55  -0.517  0.60527
fxetaria02-18 -1235.50     784.29  -1.575  0.11521
fxetaria19-23  -861.51     791.82  -1.088  0.27662
fxetaria24-28  -729.83     788.12  -0.926  0.35445
fxetaria29-33  -518.37     786.37  -0.659  0.50979
fxetaria34-38  -416.97     788.79  -0.529  0.59708
fxetaria39-43  -188.70     790.79  -0.239  0.81140
fxetaria44-48   103.83     791.55   0.131  0.89564
fxetaria49-53   389.28     793.68   0.490  0.62381
fxetaria54-58   840.80     803.73   1.046  0.29553
fxetaria59-63  1322.63     815.32   1.622  0.10479
fxetaria64-68  3485.63     845.69   4.122 3.79e-05 ***
fxetaria69-73  5090.47     887.73   5.734 1.01e-08 ***
fxetaria74-78  5589.98     940.33   5.945 2.86e-09 ***
fxetaria79-83 14278.33    1009.42  14.145  < 2e-16 ***
fxetaria84->> 14746.65    1036.40  14.229  < 2e-16 ***
tamgrupo2      -360.43     488.46  -0.738  0.46059
tamgrupo3      -891.35     486.72  -1.831  0.06708 .
tamgrupo4       214.63     501.74   0.428  0.66883
tamgrupo5      -445.78     488.07  -0.913  0.36108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14940 on 9995 degrees of freedom
Multiple R-squared: 0.06862,     Adjusted R-squared: 0.06554
F-statistic: 22.31 on 33 and 9995 DF,  p-value: < 2.2e-16
```

Segundo os p-valores avaliados para testar a significância das variáveis, observa-se um ajuste insatisfatório, pois poucas delas são significantes. Conjuntamente com teste de bondade do ajuste R^2 ajustado de valor baixo concluímos que o modelo não é adequado. Também podemos perceber um valor elevado para a variância, o que sugere a mudança de escala da variável resposta. Para contornar este problema da variância, aplicamos uma transformação logarítmica na variável resposta. Assim, construímos o Modelo 2 da seguinte forma:

$$\log\left(\frac{severidade}{exposição}\right) \sim região + plano + sexo + fxetária + tamgrupo$$

Os resultados do ajuste estão apresentados a seguir:

```
Call:
lm(formula = log(resp) ~ (regiao + plano + sexo + fxetaria +
    tamgrupo))

Residuals:
    Min      1Q  Median      3Q     Max
-6.3501 -0.4142 -0.0464  0.3680  5.3550

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.870286   0.049071 140.007  < 2e-16 ***
regiao2        0.575934   0.040979  14.054  < 2e-16 ***
regiao3        0.033389   0.041493   0.805 0.421012
regiao4        0.255876   0.041095   6.226 4.96e-10 ***
regiao5       -0.541135   0.046769 -11.570  < 2e-16 ***
regiao6       -0.531081   0.046411 -11.443  < 2e-16 ***
regiao7       -0.690381   0.048983 -14.094  < 2e-16 ***
regiao8       -0.396398   0.042852  -9.250  < 2e-16 ***
regiao9        0.226333   0.042776   5.291 1.24e-07 ***
regiao10      -1.147728   0.047806 -24.008  < 2e-16 ***
regiao11       0.245215   0.041740   5.875 4.37e-09 ***
regiao12       0.113893   0.041604   2.738 0.006201 **
regiao13      -0.386343   0.042692  -9.050  < 2e-16 ***
planoB         0.264393   0.017331  15.256  < 2e-16 ***
sexoM         -0.299887   0.017292 -17.342  < 2e-16 ***
fxetaria02-18 -0.547211   0.045427 -12.046  < 2e-16 ***
fxetaria19-23 -0.226236   0.045863  -4.933 8.24e-07 ***
fxetaria24-28 -0.003879   0.045649  -0.085 0.932279
fxetaria29-33  0.192846   0.045547   4.234 2.32e-05 ***
fxetaria34-38  0.268426   0.045688   5.875 4.36e-09 ***
fxetaria39-43  0.332560   0.045804   7.261 4.14e-13 ***
fxetaria44-48  0.478597   0.045848  10.439  < 2e-16 ***
fxetaria49-53  0.570676   0.045971  12.414  < 2e-16 ***
fxetaria54-58  0.708260   0.046553  15.214  < 2e-16 ***
fxetaria59-63  0.715844   0.047224  15.158  < 2e-16 ***
fxetaria64-68  0.913154   0.048983  18.642  < 2e-16 ***
fxetaria69-73  1.024942   0.051419  19.933  < 2e-16 ***
fxetaria74-78  0.996763   0.054465  18.301  < 2e-16 ***
fxetaria79-83  1.203814   0.058467  20.590  < 2e-16 ***
fxetaria84->>  1.225848   0.060029  20.421  < 2e-16 ***
tamgrupo2      0.108318   0.028292   3.829 0.000130 ***
tamgrupo3      0.100182   0.028192   3.554 0.000382 ***
tamgrupo4     -0.003805   0.029061  -0.131 0.895825
tamgrupo5      0.081331   0.028270   2.877 0.004023 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8652 on 9995 degrees of freedom
Multiple R-squared: 0.4174,      Adjusted R-squared: 0.4155
F-statistic:   217 on 33 and 9995 DF,  p-value: < 2.2e-16
```

A utilização da transformação logarítmica surtiu um efeito positivo no modelo, já que obtivemos a maioria das variáveis avaliadas como significativas. Alguns fatores isolados apresentaram p-valores altos para o teste de hipótese: a região 3, a faixa etária de 24 a 28 anos e o tamanho de grupo 4. A exclusão destes fatores não é indicada pois a estratégia de precificação deve observar essas categorias, já que existem observações para elas. Assim, mesmo com significância aquém do desejado, as mantivemos no modelo.

O R^2 ajustado forneceu um valor considerado plausível para o teste de bondade do ajuste. A interpretação obtida é que 41,55% da variação dos dados é explicada pelo

modelo.

Com o objetivo da escolha de um modelo ótimo, ou seja, que tenha uma capacidade de ajuste adequada com o menor número de variáveis possível, utilizamos um critério de seleção de modelos chamado AIC, sigla para Akaike Information Criterion, que estabelece uma relação entre acurácia e complexidade O modelo escolhido será o de menor valor do AIC. As variáveis adotadas em cada modelo será um subconjunto do Modelo 2, sendo uma perspectiva do ganho/perda se variáveis forem retiradas da modelagem.

Modelo	Variáveis	GL	AIC
Modelo 2	regiao+plano+sexo+fxetaria+tamgrupo	35	25.593,29
Modelo 3	regiao+plano+sexo+fxetaria	31	25.616,12
Modelo 4	regiao+plano+sexo+tamgrupo	20	28.177,44
Modelo 5	regiao+plano+fxetaria+tamgrupo	34	25.888,61
Modelo 6	regiao+sexo+fxetaria+tamgrupo	34	25.822,14
Modelo 7	plano+sexo+fxetaria+tamgrupo	23	27.790,46

Tabela 7- AIC dos modelos ajustados

A partir da tabela acima, consideramos que o modelo considerando todas as variáveis é o mais adequado de acordo com critério adotado.

Dada a significância das covariáveis, o teste de bondade de ajuste e da seleção do melhor modelo, realizamos uma análise de resíduos para a verificação das suposições de normalidade e independência.

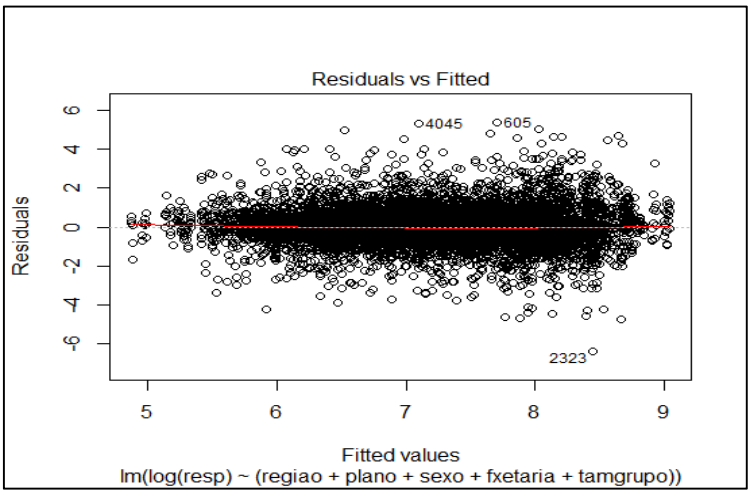


Figura 17- Gráfico "Valores ajustados vs. Resíduos" do modelo de regressão

Para a checagem de suposição da normalidade dos resíduos, fizemos uso do gráfico QQ-plot, que compara os quantis empíricos com os quantis da distribuição Normal. Idealmente os pontos devem estar próximo à linha reta indicativa do gráfico.

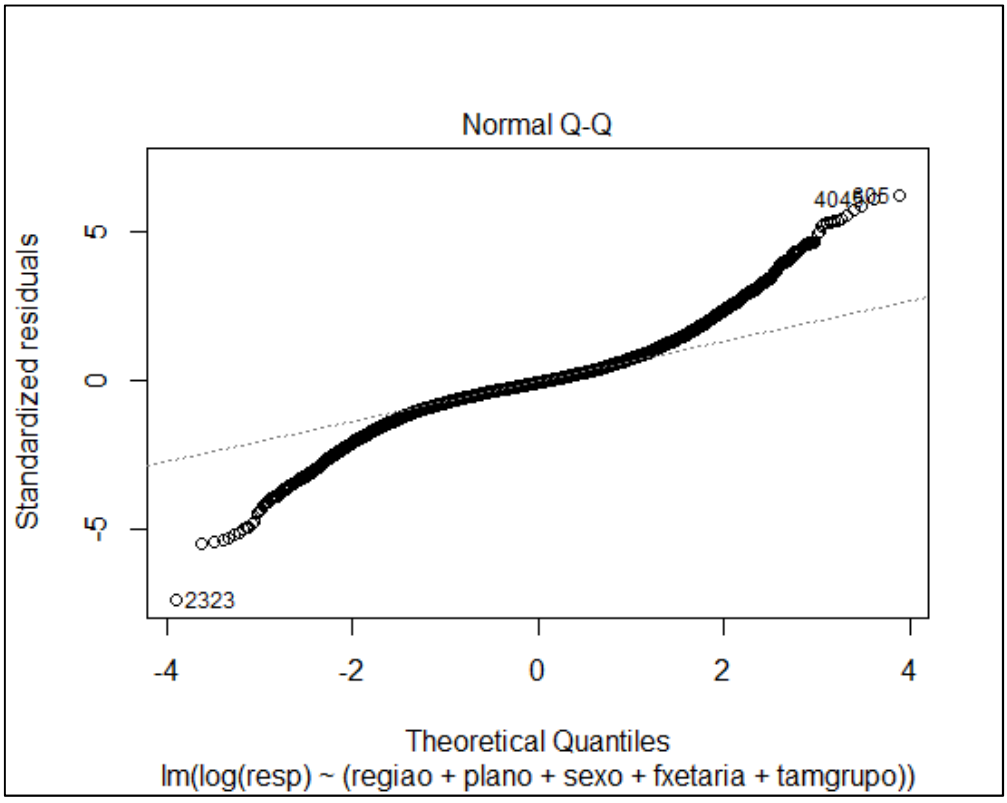


Figura 18 - Gráfico QQ-plot do modelo de regressão

Observamos que a suposição de Normalidade não foi satisfeita para os valores extremos da variável resposta. Dada a natureza dos dados em estudo, onde existem sinistros que consideram procedimentos médicos de alta complexidade (e consequentemente maior custo), o modelo não captura corretamente este efeito. Assim, há a possibilidade da existência de observações influente e leverages.

Para verificar a existência de observações influentes, ou seja, pontos que se exercem efeitos consideráveis nos valores ajustados, avaliaremos os elementos h_{ii} da matriz H, dada por $H = X(X^TX)^{-1}X^T$, que pode ser interpretado como a quantidade de leverage (influência) a i-ésima observação exerce no modelo. Será considerada uma observação influente se:

$$h_{ii} > 2\bar{h}$$

Onde $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n}$

Realizado o teste, não foram encontradas observações influentes na base de dados.

Como segundo critério, utilizaremos a medida da Distância de Cook que, segundo Montgomery:

> (...) é uma métrica da distância entre a estimativa de β pelo método dos mínimos quadrados baseada nas n observações, e a estimativa obtida quando o i-ésimo ponto for removido. (MONTGOMERY e PECK, 1982)

A distância de Cook é calculada da seguinte forma:

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})}$$

Caso $D_i > 1$, a i-ésima observação é considerada influente. Dado este novo critério, nenhuma observação foi considerada influente.

Dados os valores extremos encontrados na cauda da distribuição empírica comparada com a distribuição Normal, há a suspeita de possíveis valores aberrantes, ou outliers. Conservadoramente, consideraremos como observações discrepantes os valores dos resíduos compreendidos fora do intervalo [-2,2]. Foram encontradas 592 observações fora do limite estabelecido. Essas observações foram excluídas e um novo modelo foi construído a partir dessa nova base de dados e as considerações feitas anteriormente sobre a significância das covariáveis.

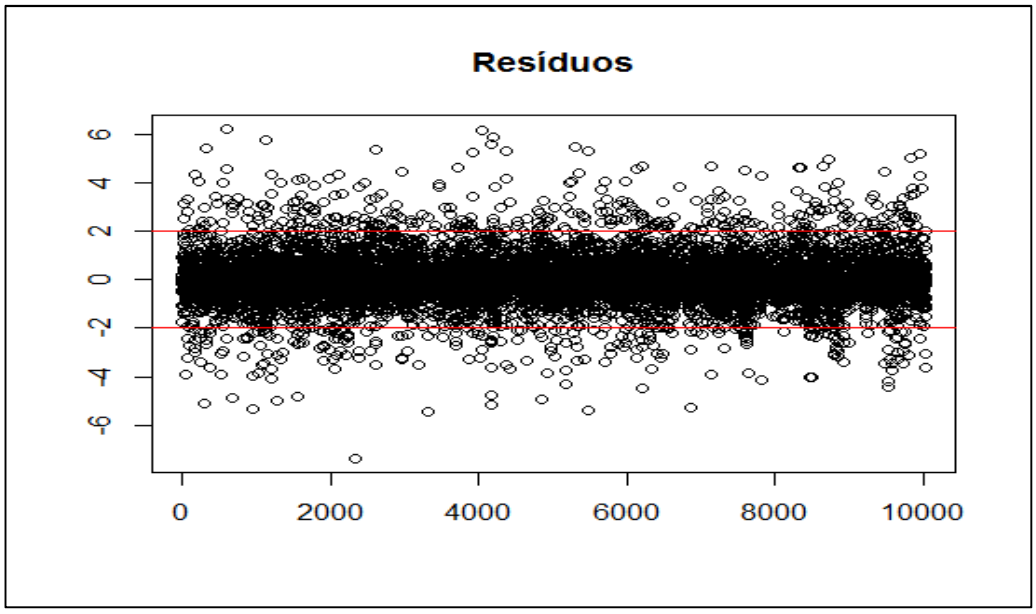


Figura 19 - Gráfico dos resíduos do modelo de regressão

Os resultados do Modelo 8 seguem abaixo:

```
Call:
lm(formula = log(resp2) ~ (regiao2 + plano2 + sexo2 + fxetaria2 +
    tamgrupo2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.73470 -0.37169 -0.03585  0.34863  1.90510

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.78358    0.03571 189.988  < 2e-16 ***
regiao22        0.60405    0.02964  20.381  < 2e-16 ***
regiao23        0.04492    0.03013   1.491   0.1360
regiao24        0.24458    0.02976   8.218 2.34e-16 ***
regiao25       -0.56392    0.03447 -16.360  < 2e-16 ***
regiao26       -0.53869    0.03384 -15.920  < 2e-16 ***
regiao27       -0.71939    0.03634 -19.796  < 2e-16 ***
regiao28       -0.39761    0.03123 -12.730  < 2e-16 ***
regiao29        0.19175    0.03115   6.155 7.80e-10 ***
regiao210      -1.09110    0.03519 -31.004  < 2e-16 ***
regiao211       0.24562    0.03045   8.067 8.07e-16 ***
regiao212       0.06189    0.03036   2.038   0.0415 *
regiao213      -0.34980    0.03115 -11.230  < 2e-16 ***
plano2B         0.25981    0.01265  20.532  < 2e-16 ***
sexo2M         -0.32535    0.01262 -25.779  < 2e-16 ***
fxetaria202-18 -0.48494    0.03270 -14.831  < 2e-16 ***
fxetaria219-23 -0.16070    0.03315  -4.847 1.27e-06 ***
fxetaria224-28  0.06077    0.03296   1.844   0.0652 .
fxetaria229-33  0.26721    0.03285   8.134 4.68e-16 ***
fxetaria234-38  0.36531    0.03293  11.093  < 2e-16 ***
fxetaria239-43  0.41148    0.03307  12.442  < 2e-16 ***
fxetaria244-48  0.55222    0.03317  16.648  < 2e-16 ***
fxetaria249-53  0.64381    0.03322  19.380  < 2e-16 ***
fxetaria254-58  0.75422    0.03365  22.416  < 2e-16 ***
fxetaria259-63  0.79260    0.03454  22.949  < 2e-16 ***
fxetaria264-68  0.96302    0.03620  26.600  < 2e-16 ***
fxetaria269-73  1.07538    0.03830  28.078  < 2e-16 ***
fxetaria274-78  1.04594    0.04120  25.387  < 2e-16 ***
fxetaria279-83  1.08778    0.04474  24.313  < 2e-16 ***
fxetaria284->>  1.20378    0.04844  24.849  < 2e-16 ***
tamgrupo22      0.11788    0.02076   5.679 1.40e-08 ***
tamgrupo23      0.13912    0.02061   6.749 1.58e-11 ***
```

```
tamgrupo24      0.02405    0.02132   1.128   0.2594
tamgrupo25      0.10403    0.02065   5.039 4.77e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6123 on 9403 degrees of freedom
Multiple R-squared: 0.5762,      Adjusted R-squared: 0.5747
F-statistic: 387.4 on 33 and 9403 DF,  p-value: < 2.2e-16
```

Nota-se uma melhora significativa no ajuste do modelo, dado que o R^2 ajustado aumentou para 57,47%, mantendo a significância das covariáveis. Novamente não foram encontradas observações influentes. Analisando o gráfico QQ-plot, observa-se um melhor ajuste nas caudas, mas ainda aquém do objetivo da obtenção de um modelo ótimo.

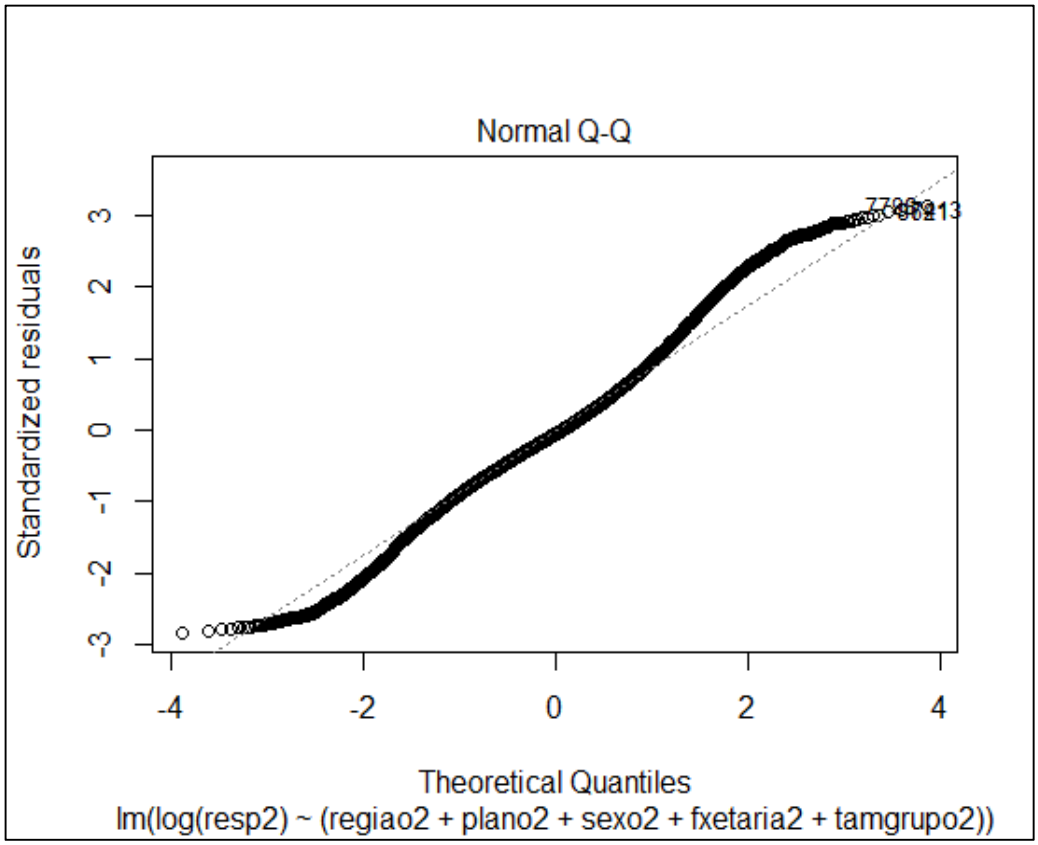


Figura 20 - QQ-plot do novo modelo de regressão ajustado

Dadas as conclusões elucidadas anteriormente, e sob a justificativa da não normalidade dos resíduos, consideramos o modelo clássico de regressão como não adequado à base de dados avaliada. Assim, ajustaremos um modelo considerando outra distribuição da variável resposta, baseando-se na metodologia dos Modelos Lineares Generalizados.

3.2 Modelos Lineares Generalizados

3.2.1 Metodologia

Em meados do século XX, foram formulados por Nelder e Wedderburn os chamados Modelos Lineares Generalizados (MLG). Esses modelos são uma extensão dos modelos clássicos, englobando não somente as variáveis com distribuição normal de erros, mas outras distribuições como: Gama, Poisson, Binomial, Normal Inversa, entre outras. Assim, a distribuição da variável resposta não necessita ser Normal, mas sim pertencente à uma família de distribuições, chamada Família Exponencial.

3.2.1.1 Família Exponencial

Segundo (TURKMAN e SILVA, 2000), uma variável aleatória Y tem distribuição pertencente á família exponencial se sua função de probabilidade (no caso discreto) ou sua função densidade de probabilidade (no caso contínuo) puder ser escrita na forma:

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

onde a(.), b(.) e c(.) são funções conhecidas. $\phi > 0$ é denominado parâmetro de dispersão e θ é denominado parâmetro canônico, que caracteriza a distribuição. Esses dois parâmetros também são conhecidos como parâmetros de escala.

Pode-se mostrar que, sob condições de regularidade, que:

$$E[Y] = \mu = b'(\theta)$$

$$var[Y] = a(\phi)b''(\theta)$$

Podemos reescrever a $var[Y]$ como $var[Y] = \phi^{-1}V(\mu)$, onde $V(\mu) = d\mu/d\theta$. $V(\mu)$ é a chamada função de variância, explicitando a relação entre a variância e a média. Vale ressaltar que a função de variância caracteriza a distribuição, ou seja,

dada essa função é possível estabelecer a distribuição da variável resposta.

Para exemplificar a formulação das distribuições pertencentes à família exponencial, serão demonstradas aqui duas distribuições pertencentes à esta família: Normal e Gamma.

3.2.1.1.1 Distribuição Normal

Seja Y a variável aleatória de distribuição Normal com média μ e variância σ^2. A sua função densidade de probabilidade é dada por:

$$f(y;\mu;\sigma) = \frac{-1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$$

Reescrevendo a equação acima, obtemos:

$$f(y;\mu;\sigma) = \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right\}$$

Assim, $\theta = \mu$; $b(\theta) = \frac{\mu^2}{2}$; $a(\phi) = \phi$; $\phi = \sigma^2$ e

$$c(y_i, \phi) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$$

Vale ressaltar que no caso da distribuição Normal $V(\mu) = 1$, ou seja, a variância é constante e igual ao parâmetro de dispersão, justificando a premissa de variância constante neste modelo.

3.2.1.1.2 Distribuição Gamma

Seja Y a variável aleatória de distribuição Gamma de parâmetro de forma μ e parâmetro de escala α/μ . A sua função densidade de probabilidade é dada por:

$$f(y;\mu;\alpha) = \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right)^{\alpha} y^{\alpha-1} e^{-\alpha\left(\frac{y}{\mu}\right)}$$

Reescrevendo a equação acima, obtemos:

$$f(y;\mu;\alpha) = \exp\left\{\frac{-y\frac{1}{\mu} - \ln(\mu)}{\frac{1}{\alpha}} + \alpha\ln(\alpha) + (\alpha-1)\ln(y) - \frac{\alpha y}{\mu} - \ln\Gamma(\alpha)\right\}$$

Assim, $\theta = -\frac{1}{\mu}$; $b(\theta) = \ln(\mu)$; $a(\phi) = \phi$; $\phi = \frac{1}{\alpha}$;

$$c(y_i, \phi) = \alpha\ln(\alpha) + (\alpha-1)\ln(y) - \frac{\alpha y}{\mu} - \ln\Gamma(\alpha)$$

$$E[Y] = \mu$$

$$var[Y] = \frac{\mu^2}{\alpha}$$

Observe que a função de variância é dada por $V(\mu) = \mu^2$, ou seja, a variância da distribuição de Y é uma função quadrática da média.

3.2.1.2 Estrutura

Como mencionado anteriormente, os MLG são uma abordagem alternativa para transformação de dados quando as suposições usuais de normalidade e variância constante não são satisfeitas. Além da distribuição da variável resposta, esses modelos apresentam a característica da relação entre as partes sistemática e aleatória ocorrer através da média da distribuição. Segundo Montgomery, podemos reescrever os MLG com a seguinte estrutura:

- O vetor resposta $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ independentes e com vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)$.
- Cada observação y_i possui distribuição de probabilidade (ou densidade de probabilidade) pertencente à família exponencial.

- A parte sistemática do modelo envolve as covariáveis $x_1, x_2, \cdots, x_k$.
- O modelo é construído através dos dados via preditor linear η, onde:

$$\eta = \sum_{j=1}^{k} \beta_j x_{ik}$$

- A relação entre a média da distribuição da variável resposta e o preditor linear é feita através de uma função g(.), denominada Função de Ligação.

  $$\eta_i = g(\mu_i) \quad i = 1,2, \dots, n$$

  Assim, o primeiro momento da distribuição pode ser reescrito como:

  $$E[y_i] = g^{-1}(\eta_i)$$

- A função de ligação é monótona e diferenciável.
- A variância da distribuição é uma função da média μ_i.

3.2.1.3 Função de Ligação

A função de ligação relaciona o componente aleatório ao componente sistemático, ou seja, relaciona o valor esperado ao seu preditor linear. Se a função de ligação é escolhida de fazendo que $g(\mu_i) = \theta_i = \eta_i$, então o preditor linear modelará diretamente o parâmetro canônico μ_i, caracterizando a chamada função de ligação canônica. Os modelos que consideram esse tipo de função de ligação são denominados modelos canônicos.

Podemos ver abaixo as funções de ligação canônicas das distribuições usualmente utilizadas:

Distribuição	*Função de ligação canônica*
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = log\mu$
Binomial	Logística: $\eta = \log\left(\frac{\pi}{1-\pi}\right)$
Gamma	Recíproca: $\eta = \frac{1}{\mu}$
Normal Inversa	Recíproca do quadrado: $\eta = \frac{1}{\mu^2}$

Tabela 8 - Funções de ligação

A escolha da função de ligação depende dos dados do fenômeno sob estudo. Apesar da utilização das ligações canônicas fornecerem propriedades estatísticas mais interessantes, essa escolha não garante uma melhoria na qualidade do ajuste nem na capacidade de predição do modelo.

3.2.1.4 Método de Estimação

Segundo (MC CULLAGH e NELDER, 1989), para a estimação do vetor paramétrico desconhecido β, é utilizado o método de Máxima Verossimilhança obedecendo um critério de bondade de ajuste entre os valores observados e os valores ajustados pelo modelo. Os parâmetros estimados são os valores que minimizam o critério de bondade de ajuste. Primeiramente, devemos nos preocupar com as estimativas encontradas pela máxima verossimilhança ou log da verossimilhança dos parâmetros para os dados observados. Segundo (DUNCAN, 2011), "exceto para os modelos de distribuição Normal, não existe solução explícita para o vetor estimado $\hat{\beta}$". Os estimadores de máxima verossimilhança são obtidos através de métodos de aproximação, como os métodos de Newton-Raphson e Score de Fisher.

Como o parâmetro de dispersão ϕ também é desconhecido, utilizamos a estatística de Pearson que, para grandes valores de n, possui uma distribuição aproximada de uma χ^2 com $n-p$ graus de liberdade. Salientamos que no modelo clássico de regressão, como $V(\hat{\mu}_i) = 1$, $\hat{\phi}$ coincide com o estimador de σ^2.

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}}{n - p}$$

3.2.1.5 Inferência

Para testar a significância das covariáveis, utilizamos os testes de hipóteses utilizados no software R, que são baseados nas estatísticas de Wald que, segundo (TURKMAN e SILVA, 2000), são construídos em torno do conceito de normalidade assintótica do estimador de máxima verossimilhança.

3.2.1.6 Seleção do Modelo

Além da estimação dos parâmetros e a análise de sua significância, é preciso estabelecer critérios para a seleção do(s) melhor(es) modelo(s). Assim, verificamos desde a escala da variável resposta e a escolha da função de ligação até a combinação de covariáveis mais parcimoniosa, considerando o menor número de fatores possível sem perda significativa da informação.

Os modelos selecionados para o ajuste dos dados geralmente são escolhidos de uma classe em particular que será muito relevante nos dados sob estudo. A escolha da escala para análise é um importante aspecto na seleção do modelo. Uma escolha usual é entre a análise de $\boldsymbol{Y}$,a escala original, ou $log\boldsymbol{Y}$. Com a introdução dos MLG os problemas com escala foram reduzidos já que normalidade e variância constante não são necessárias, apesar de que a maneira como a variância depende da média deve ser conhecida. Entretanto, a aditividade dos efeitos continua sendo um importante componente nos MLG e pode ser usada uma transformação de escala se necessário. Nos MLG aditividade é postulada como propriedade dos valores esperados.

Na seleção do modelo ainda existe a questão da escolha das variáveis explicativas a serem incluídas na parte sistemática do modelo. Na sua forma mais simples nós

teremos um certo número de candidatos $x_1, \ldots, x_p$, e precisamos então achar um subconjunto desses dados que serão melhores para a construção dos valores ajustados $\widehat{\boldsymbol{\mu}} = \sum \boldsymbol{x_j \beta_j}$. Como a base de dados possui um número pequeno de covariáveis, trabalhamos com o modelo considerando todos os fatores sob análise.

3.2.1.7 Bondade do Ajuste

A verificação da qualidade do ajuste é feita através de uma medida que avalia a distância entre os valores ajustados dos valores efetivamente observados. Considerando o vetor de observações independentes $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ e o vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)$, o log da função de verossimilhança será dado por:

$$l(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_i \log f_i(y_i; \theta_i)$$

Considerando um modelo $l(\boldsymbol{y}; \boldsymbol{y})$ onde os valores ajustados são iguais aos valores observados e um modelo $l(\boldsymbol{\mu}; \boldsymbol{y})$ com um número menor de parâmetros. Assim, estabelecemos uma função de distância entre os dois modelos, denominada função desvio ou *deviance*, como segue:

$$D^*(\boldsymbol{y}; \boldsymbol{\mu}) = 2l(\boldsymbol{y}; \boldsymbol{y}) - 2l(\boldsymbol{\mu}; \boldsymbol{y})$$

Valores pequenos da função desvio indica que o ajuste com um número menor de parâmetros fornece um ajuste similar ao modelo saturado. Para o modelo clássico de regressão, a função desvio coincide com a soma dos quadrados dos resíduos.

Segundo (PAULA, 2012), quando temos uma dispersão relativamente pequena, é possível compararmos os valores observados $D^*(\boldsymbol{y}; \boldsymbol{\mu})$ com os quantis de uma distribuição de qui-quadrado com $n - p$ graus de liberdade. Sendo assim consideraremos a hipótese nula de que o modelo é adequado a avaliaremos a função desvio dividida pela estimativa do parâmetro de dispersão, rejeitando a hipótese nula caso:

$$\frac{D^*(\boldsymbol{y}; \boldsymbol{\mu})}{\hat{\phi}} > \chi^2_{n-p}$$

3.2.1.8 Análise de Resíduos

Assim como no modelo clássico de regressão, realizamos uma análise de resíduos, e através de gráfico fizemos considerações sobre a qualidade do modelo, assim como a detecção de possíveis outliers, observações influentes e leverages.

A metodologia principal na avaliação dos resíduos foi o uso do desvio residual. Reescrevendo a equação da função desvio, encontramos a contribuição da i-ésima observação para este desvio, como segue:

$$r_i^D = sinal(y_i - \hat{\mu}_i)\frac{2}{\phi}\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) + b\left(\tilde{\theta}_i\right) - b\left(\hat{\theta}_i\right)\right]$$

A partir da equação anterior, obtemos o desvio residual padronizado, dividindo r_i^D por $\sqrt{\hat{\phi}(1-h_{ii})}$, onde h_{ii} é o i-ésimo componente da matriz **H**. Assim, obtemos os resíduos:

$$r_i^{D\prime} = \frac{r_i^D}{\sqrt{\hat{\phi}(1-h_{ii})}}$$

Onde aproximadamente $r_i^{D\prime} \sim N(0,1)$.

3.2.2 Aplicação

Dado que os modelos ajustados adotando-se os modelos clássicos de regressão mostraram-se insatisfatórios, principalmente devido à não veracidade da premissa normalidade dos resíduos, adotamos a utilização de outra distribuição contínua para a variável resposta e, baseando-se na metodologia dos MLG, foram construídos modelos com um objetivo de melhor ajuste, significância dos fatores analisados, satisfação das premissas decorrentes da modelagem e alta capacidade de predição.

Como os dados em estudo tratam sobre sinistros decorrentes de procedimentos médicos e que a variável resposta é o logaritmo desta severidade por unidade em

exposição, utilizamos a distribuição Gamma, já que a mesma é de natureza contínua e estritamente positiva. O primeiro modelo ser testado contiveram as mesmas componentes utilizadas anteriormente, considerando as variáveis região, tipo de plano, sexo, faixa etária e tamanho do grupo, novamente não considerando o período na análise. Assim, o Modelo 9 foi estruturado como segue:

$$\log\left(\frac{severidade}{exposição}\right) \sim região + plano + sexo + fxetária + tamgrupo$$

De acordo com a metodologia utilizada no MLG, a relação entre a média da distribuição da variável resposta e o preditor linear é feita através da função de ligação. Assim, devido à facilidade de ajuste, escolhemos a função logarítmica como função de ligação, ou seja:$\eta = \log(\mu) = \boldsymbol{X\beta}$

Os resultados, a partir do software R, estão dispostos abaixo:

```
Call:
glm(formula = (log(resp)) ~ (regiao + plano + sexo + fxetaria + 
    tamgrupo), family = Gamma(link = "log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max  
-1.13358  -0.05826  -0.00665   0.05034   0.64488  

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.928e+00  6.772e-03 284.680  < 2e-16 ***
regiao2        7.451e-02  5.655e-03  13.175  < 2e-16 ***
regiao3        5.502e-03  5.726e-03   0.961 0.336670    
regiao4        3.343e-02  5.671e-03   5.895 3.87e-09 ***
regiao5       -7.685e-02  6.454e-03 -11.907  < 2e-16 ***
regiao6       -7.620e-02  6.405e-03 -11.897  < 2e-16 ***
regiao7       -1.007e-01  6.760e-03 -14.902  < 2e-16 ***
regiao8       -5.618e-02  5.914e-03  -9.500  < 2e-16 ***
regiao9        3.049e-02  5.903e-03   5.164 2.46e-07 ***
regiao10      -1.758e-01  6.597e-03 -26.641  < 2e-16 ***
regiao11       3.268e-02  5.760e-03   5.674 1.44e-08 ***
regiao12       1.512e-02  5.742e-03   2.633 0.008467 ** 
regiao13      -5.339e-02  5.892e-03  -9.062  < 2e-16 ***
planoB         3.638e-02  2.392e-03  15.211  < 2e-16 ***
sexoM         -4.324e-02  2.386e-03 -18.120  < 2e-16 ***
fxetaria02-18 -8.432e-02  6.269e-03 -13.451  < 2e-16 ***
fxetaria19-23 -3.518e-02  6.329e-03  -5.559 2.78e-08 ***
fxetaria24-28 -1.773e-03  6.300e-03  -0.281 0.778357    
fxetaria29-33  2.746e-02  6.286e-03   4.368 1.26e-05 ***
fxetaria34-38  3.832e-02  6.305e-03   6.078 1.26e-09 ***
fxetaria39-43  4.713e-02  6.321e-03   7.457 9.61e-14 ***
fxetaria44-48  6.750e-02  6.327e-03  10.669  < 2e-16 ***
fxetaria49-53  7.956e-02  6.344e-03  12.541  < 2e-16 ***
fxetaria54-58  9.898e-02  6.424e-03  15.406  < 2e-16 ***
fxetaria59-63  9.874e-02  6.517e-03  15.151  < 2e-16 ***
fxetaria64-68  1.239e-01  6.760e-03  18.323  < 2e-16 ***
fxetaria69-73  1.380e-01  7.096e-03  19.450  < 2e-16 ***
fxetaria74-78  1.339e-01  7.516e-03  17.812  < 2e-16 ***
fxetaria79-83  1.586e-01  8.069e-03  19.653  < 2e-16 ***
fxetaria84->>  1.604e-01  8.284e-03  19.366  < 2e-16 ***
tamgrupo2      1.537e-02  3.904e-03   3.936 8.34e-05 ***
tamgrupo3      1.439e-02  3.890e-03   3.698 0.000219 ***
tamgrupo4      3.583e-05  4.011e-03   0.009 0.992872    
tamgrupo5      1.177e-02  3.901e-03   3.016 0.002566 ** 
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.01425711)

    Null deviance: 251.36  on 10028  degrees of freedom
Residual deviance: 146.56  on  9995  degrees of freedom
AIC: 25560

Number of Fisher Scoring iterations: 4
```

Observando os p-valores dos testes de hipótese propostos para a análise de significância de cada fator, nota-se que quase todos são relevantes para a variação da severidade. Os fatores com baixa significância foram mantidos levando-se em conta o mesmo critério estabelecido no modelo clássico de regressão.

Para a escolha do modelo ótimo, utilizamos novamento o Akaike Information Criterion e escolhemos o de menor valor. Consideramos inicialmente o modelo contendo as variáveis que demonstraram ser significantes e, como anteriormente, adotamos um subconjunto do mesmo, obtendo uma percepção sobre o ganho/perda da inclusão de cada variável no modelo.

Modelo	Variáveis	GL	AIC
Modelo 9	regiao+plano+sexo+fxetaria+tamgrupo	35	25.559,70
Modelo 10	regiao+plano+sexo+fxetaria	31	25.582,55
Modelo 11	regiao+plano+sexo+tamgrupo	20	28.098,28
Modelo 12	regiao+plano+fxetaria+tamgrupo	34	25.873,71
Modelo 13	regiao+sexo+fxetaria+tamgrupo	34	25.781,19
Modelo 14	plano+sexo+fxetaria+tamgrupo	23	27.811,84

Tabela 9 – AIC – Seleção de Modelos MLG

A partir deste critério, como o modelo considerando todos os fatores apresenta o menor AIC, optamos por utilizar todas as variáveis.

Como o parâmetro de dispersão é desconhecido, usaremos o estimador descrito no capítulo da metodologia. Assim, obteremos indícios da ordem de grandeza da variabilidade dos dados estudados, além de usar o resultado para a realização do teste de bondade do ajuste. A estimativa encontrada foi $\widehat{\phi} = 0{,}01425996$.

Para o teste de bondade de ajuste, comparando a razão entre $D^*(\boldsymbol{y};\boldsymbol{\mu})$ e a estimativa do parâmetro de dispersão com o quantil da distribuição qui-quadrado com $n-p$ graus de liberdade, rejeitamos a hipótese nula de ajuste do modelo aos

dados. Assim, sob a justificativa da presença de valores de sinistro não condizentes com o comportamento usual (outliers), foi realizado um tratamento de outliers e a partir dos critérios estabelecidos as observações foram deletados e o modelo julgado mais adequados retestado, contendo as variáveis região, tipo de plano, sexo, faixa etária e tamanho do grupo.

O critério adotado para a exclusão de outliers levou em consideração o desvio residual standardizado, sendo excluídas as observações contidas fora do intervalo [-2,2]. Foram encontradas 582 observações consideradas outliers, e o modelo resultante dos dados não considerando essas observações está disposto a seguir:

```
Call:
glm(formula = (log(resp3)) ~ (regiao3 + plano3 + sexo3 + fxetaria3 +
    tamgrupo3), family = Gamma(link = "log"))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.241800  -0.052908  -0.005749   0.046986   0.251230

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.918600   0.004985 384.905  < 2e-16 ***
regiao32        0.076262   0.004111  18.553  < 2e-16 ***
regiao33        0.006051   0.004182   1.447   0.1480
regiao34        0.031230   0.004129   7.564 4.27e-14 ***
regiao35       -0.079846   0.004825 -16.547  < 2e-16 ***
regiao36       -0.076200   0.004727 -16.119  < 2e-16 ***
regiao37       -0.102286   0.005089 -20.099  < 2e-16 ***
regiao38       -0.056554   0.004350 -13.000  < 2e-16 ***
regiao39        0.027310   0.004324   6.316 2.80e-10 ***
regiao310      -0.163033   0.004980 -32.739  < 2e-16 ***
regiao311       0.033768   0.004216   8.009 1.29e-15 ***
regiao312       0.009262   0.004210   2.200   0.0278 *
regiao313      -0.049511   0.004331 -11.432  < 2e-16 ***
plano3B         0.035814   0.001764  20.300  < 2e-16 ***
sexo3M         -0.044990   0.001759 -25.577  < 2e-16 ***
fxetaria302-18 -0.077895   0.004582 -17.001  < 2e-16 ***
fxetaria319-23 -0.028093   0.004649  -6.043 1.57e-09 ***
fxetaria324-28  0.006594   0.004618   1.428   0.1534
fxetaria329-33  0.036408   0.004603   7.909 2.89e-15 ***
fxetaria334-38  0.049887   0.004611  10.820  < 2e-16 ***
fxetaria339-43  0.057495   0.004630  12.418  < 2e-16 ***
fxetaria344-48  0.077405   0.004645  16.663  < 2e-16 ***
fxetaria349-53  0.090233   0.004658  19.370  < 2e-16 ***
fxetaria354-58  0.103889   0.004714  22.040  < 2e-16 ***
fxetaria359-63  0.111329   0.004839  23.008  < 2e-16 ***
fxetaria364-68  0.134732   0.005042  26.720  < 2e-16 ***
fxetaria369-73  0.149218   0.005320  28.048  < 2e-16 ***
fxetaria374-78  0.145582   0.005685  25.606  < 2e-16 ***
fxetaria379-83  0.153548   0.006139  25.013  < 2e-16 ***
fxetaria384->>  0.171338   0.006554  26.143  < 2e-16 ***
tamgrupo32      0.014466   0.002897   4.994 6.02e-07 ***
tamgrupo33      0.016809   0.002876   5.844 5.26e-09 ***
tamgrupo34      0.002543   0.002975   0.855   0.3927
tamgrupo35      0.012097   0.002884   4.194 2.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.00729134)

    Null deviance: 164.300  on 9446  degrees of freedom
Residual deviance:  68.279  on 9413  degrees of freedom
AIC: 17471

Number of Fisher Scoring iterations: 4
```

Podemos observar que a exclusão das observações segundo o critério definido melhorou o ajuste, diminuindo o AIC e mantendo a significância dos fatores. Realizando novamente o teste de bondade do ajuste, aceitamos a hipótese nula de que o modelo se ajusta aos dados, dado que a razão entre $D^*(\boldsymbol{y};\boldsymbol{\mu})$ e a estimativa do parâmetro de dispersão (de valor 9364,397) é menor que o quantil da distribuição qui-quadrado com $n-p$ graus de liberdade (de valor 10228,69).

Para a verificação de observações influentes e leverages, foram realizados os mesmos testes aplicados nos modelos clássicos de regressão, levando em consideração os elementos h_{ii} da matriz H e a distância de Cook. Não foram encontradas observações influentes.

A análise de resíduos demonstrou o mesmo comportamento resultante dos modelos lineares clássicos. Segundo a metodologia aplicada, os desvios residuais (deviance residuals) possuem distribuição Normal. Fazendo o gráfico QQ-plot, podemos observar a falta de ajuste para valores extremos, dificultando a predição.

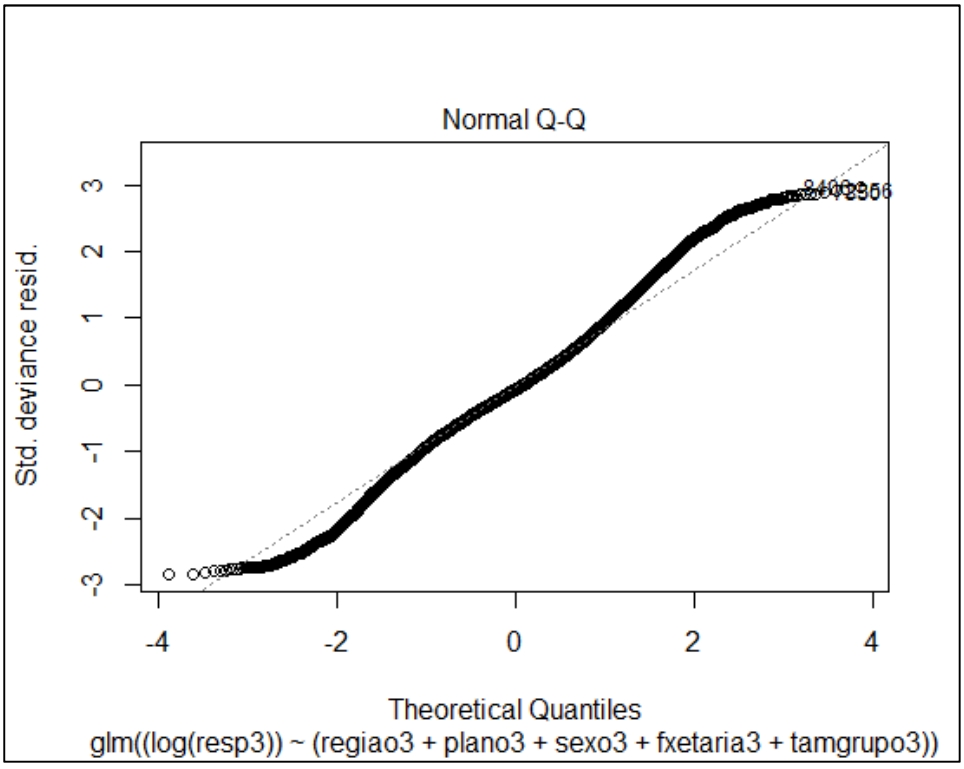


Figura 21 - QQ-plot dos resíduos deviance do MLG ajustado

Para uma análise da capacidade de previsão do modelo, foram selecionados três conjuntos de observações, denominados grupos 1, 2 e 3. Os grupos 1 e 2 representam as observações que se encontram nas caudas da distribuição formada pelos resíduos deviance enquanto o grupo 3 é composto por observações que se encontram próximas ao sinistro médio encontrado na base de dados. A partir dessa seleção, calculamos o erro quadrático médio para cada grupo, como formulação abaixo:

$$EMQ = \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)}{n}$$

Onde $\hat{Y}_i$é o valor ajustado do modelo, Y_i o valor efetivamente observado e n o número de observações no grupo.

Grupo	Valor Observado	Valor Estimado	Diferença
Grupo 1	4.591	5.626	1.035
	5.077	6.089	1.012
	4.975	6.252	1.278
	5.824	7.239	1.415
	6.838	8.514	1.677
Grupo 2	7.405	6.001	-1.403
	7.378	6.186	-1.192
	6.720	5.613	-1.107
	8.283	6.767	-1.516
	9.702	8.009	-1.694
Grupo 3	5.319	5.367	0.047
	6.210	6.233	0.023
	5.171	5.180	0.009
	5.957	5.962	0.004
	7.229	7.195	-0.033

Com esses resultados seremos capazes de calcular o erro quadrático médio que segue:

	EQM
Grupo 1	1.708
Grupo 2	1.957
Grupo3	0.001

A partir desses valores podemos confirmar o que observamos anteriormente, a falta de ajuste para valores mais extremos, dificultando a predição para esses valores de severidade.

Como a suposição de normalidade dos desvios residuais não é satisfeita para as caudas da distribuição, consideramos o modelo de distribuição Gamma para a variável resposta como não adequado.

4 CONCLUSÃO

Este trabalho constituiu em uma aplicação de Modelos Lineares Generalizados à uma base de dados de seguro saúde, utilizando dois modelos que, apesar de estruturas e premissas diferentes, ofereceram interpretações semelhantes.

Além do ajuste em si, foram expostas as etapas mais importantes no processo de modelagem, desde a construção da base de dados até as interpretações do modelo seguindo critérios estatísticos consistentes. O entendimento do conceito de modelagem vai além de um simples conjunto de procedimentos, mas sim um esforço do pesquisador em, dos recursos disponíveis e conhecimento do fenômeno analisado, obter os melhores resultados a partir de modelos o mais simplificado possível.

A análise da base disponível para estudo forneceu insumos importantes sobre o tipo e a qualidade dos dados imputados no modelo. A partir dessa avaliação, foi possível reconhecer quais variáveis necessitam de um tratamento mais apurado. Para bancos de dados com muitas informações, muitas vezes não utilizadas pelo seu difícil manuseio, é necessário um pré-tratamento ponto de criar a informação necessária para a modelagem. Assim, uma das justificativas para a falta de acurácia do modelo está diretamente ligada a base, pois foi possível a construção de modelos apenas para a severidade, dado não constavam dados de frequência, variável importante para o entendimento da probabilidade de ocorrência de um determinado evento.

Pôde-se avaliar que, apesar de um ajuste satisfatório, houve perda na precisão na previsão de valores contidos nas caudas, ou seja, sinistros de valores muito elevados não foram captados no modelo. Assim, pode-se sugerir que se utilize metodologias alternativas para estes casos, como as fundamentadas na Teoria dos Valores Extremos.

Outras estruturas também podem ser utlizadas para a modelagem em Saúde. Um exemplo é uma análise segregada por doença, onde é realizado um estudo sobre o sinistro esperado relacionado ao evento causador da despesa médica, agrupando esta despesa em grupos de doenças (doenças pulmonares, doenças coronarianas,

etc). Outro exemplo é o input de variáveis mais específicas, como indicadores médicos de nível de glicose ou pressão arterial.

A construção de modelos para a construção de uma precificação mais acertiva é um objetivo que exige um conhecimento técnico apurado, somado ao planejamento estratégico da companhia e o conhecimento na área de negócio. Estudos característicos da área de Ciências Atuariais focadas em modelagem e métodos de pricing sem dúvida auxiliarão na formação de uma visão mais holística, essencial para uma melhor performace do profissional que atue no ramo de Saúde Suplementar.

REFERÊNCIAS

BRASIL. **Lei nº 9656, de 3 de junho de 1998**. Dispõe sobre os planos e seguros privados de assistência à saúde. Diário Oficial da União. 3 de junho de 1998. Seção 1. Disponível em: < http://www.planalto.gov.br/ccivil_03/Leis/L9656.htm>.

BRASIL. **Lei nº 9961, de 28 de janeiro de 2000**. Cria a Agência Nacional de Saúde – ANS e dá outras providências. Diário Oficial da União, 28 de janeiro de 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L9961.htm>.

ANDERSON, D. et al. **A Practiotioner's Guide to Generalized Linear Models**. Society of Actuaries. [S.l.]. 2007.

BOX, G. E. P.; JENKINS, G. M. **Time Series Analysis:** Forecasting and Control. San Francisco: Holden-Day, 1976.

DE JONG, P.; HELLER, G. Z. **Generalized Linear Models for Insurance Data**. Cambridge: Cambridge University Press, 2008.

DUNCAN, I. **Healthcare Risk Adjustmentes and Predictive Modeling**. Winsted: ACTEX Publications, 2011.

MC CULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2nd. ed. London: Chapman & Hall, 1989.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to Linear Regression Analysis**. New York: John Wiley & Sons, 1982.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engeneers**. 2nd. ed. New York : John Wiley & Sons , 1999.

MORETTIN, P. A.; BUSSAB, W. D. O. **Estatística Básica**. 5ª. ed. São Paulo: Editora Saraiva, 2004.

MYERS, R. H. et al. **Generalized Linear Models with Applications in Engeneering and the Sciences**. 2nd ed. ed. New Jersey: John Wiley & Sons, 2010.

PAULA, G. A. **Modelos de Regressão com Apoio Computacional**. Universidade de São Paulo. São Paulo, p. 417. 2012.

TURKMAN, M. A. A.; SILVA, G. L. **Modelos Lineares Generalizados - da teoria à prática**. Universidade de Lisboa. Lisboa, p. 151. 2000.

APÊNDICE A – PROGRAMAÇÃO

```
###Variaveis
resp = base[,9]/base[,8]
periodo = base[,2]
regiao = as.factor(base[,3])
plano = base[,4]
sexo = base[,5]
fxetaria = base[,6]
tamgrupo = as.factor(base[,7])

### Análise Exploratória de Dados
par(mfrow=c(1,1))
boxplot((resp),prob=T,outline=TRUE, main="Severidade/Exposição",ylab="", xlab="")
boxplot((resp),prob=T,outline=FALSE, main="Severidade/Exposição",ylab="", xlab="")
boxplot((resp)~periodo,prob=T,outline=FALSE, main="Periodo",ylab="Severidade/Expostos",
xlab="")
boxplot((resp)~regiao,prob=T,outline=FALSE, main="Região",ylab="Severidade/Expostos",
xlab="")
boxplot((resp)~plano,prob=T,outline=FALSE, main="Plano",ylab="Severidade/Expostos",
xlab="")
boxplot((resp)~sexo,prob=T,outline=FALSE, main="Sexo",ylab="Severidade/Expostos",
xlab="")
boxplot((resp)~fxetaria,prob=T,outline=FALSE, main="Faixa
etária",ylab="Severidade/Expostos", xlab="")
boxplot((resp)~tamgrupo,prob=T,outline=FALSE, main="Tamanho do
grupo",ylab="Severidade/Expostos", xlab="")
hist(resp,main="histograma razão
sev/exp",prob=T,breaks=500,xlim=c(0,10e+04),ylim=c(0,7e-04),xlab="")
lines(density(resp),col="dark green")

###Modelo Linear Clássico de Regressão
##Primeiro modelo
model.fit1 <- lm((resp)~(regiao+plano+sexo+fxetaria+tamgrupo))
summary(model.fit1)
model.fit2 <- lm(log(resp)~(regiao+plano+sexo+fxetaria+tamgrupo))
summary(model.fit2)
plot(model.fit2)

##AIC para a seleção do melhor modelo, considerando a exclusão de variaveis
model.fit2 <- lm(log(resp)~(regiao+plano+sexo+fxetaria+tamgrupo))
model.fit3 <- lm(log(resp)~(regiao+plano+sexo+fxetaria))
model.fit4 <- lm(log(resp)~(regiao+plano+sexo+tamgrupo))
model.fit5 <- lm(log(resp)~(regiao+plano+fxetaria+tamgrupo))
model.fit6 <- lm(log(resp)~(regiao+sexo+fxetaria+tamgrupo))
model.fit7 <- lm(log(resp)~(plano+sexo+fxetaria+tamgrupo))
```

```
AIC(model.fit2,model.fit3,model.fit4,model.fit5,model.fit6,model.fit7)

##Teste de Observações Influentes
#Primeiro critério
hvalues = hatvalues(model.fit2)
aux=2*mean(hvalues)
sum(hvalues>aux)
#Segundo critério
cookvalues=cooks.distance(model.fit2)
sum(cookvalues>1)

##Tratamento de Outliers
residuoslm=rstudent(model.fit2)
plot(residuoslm)
abline(h=c(-2,2), col="red")
sum(residuoslm>=2,residuoslm<=-2)

#Exclusão das observações da base
base.out2=cbind(base[,1:9],residuoslm)
base.out2=base.out2[(base.out2[,10]<=2),]
base.out2=base.out2[(base.out2[,10]>=-2),]

##Novo modelo (com tratamento de outliers)
#Variáveis da nova base
resp2 = base.out2[,9]/base.out2[,8]
periodo2 = base.out2[,2]
regiao2 = as.factor(base.out2[,3])
plano2 = base.out2[,4]
sexo2 = base.out2[,5]
fxetaria2 = base.out2[,6]
tamgrupo2 = as.factor(base.out2[,7])

##Modelo
model.fit8 <- lm(log(resp2)~(regiao2+plano2+sexo2+fxetaria2+tamgrupo2))
summary(model.fit8)
plot(model.fit8, which=2)

##############################################################################
#################################################################

###Modelos Lineares Generalizados - Distribuição Gamma

model.fit9 <- glm((log(resp))~(regiao+plano+sexo+fxetaria+tamgrupo), family = Gamma(link
= "log"))
summary(model.fit9)
plot(model.fit9)
```

```
##AIC para a seleção do melhor modelo, considerando a exclusão de variaveis
model.fit10 <- glm(log(resp)~(regiao+plano+sexo+fxetaria), family = Gamma(link = "log"))
model.fit11 <- glm(log(resp)~(regiao+plano+sexo+tamgrupo), family = Gamma(link = "log"))
model.fit12 <- glm(log(resp)~(regiao+plano+fxetaria+tamgrupo), family = Gamma(link =
"log"))
model.fit13 <- glm(log(resp)~(regiao+sexo+fxetaria+tamgrupo), family = Gamma(link =
"log"))
model.fit14 <- glm(log(resp)~(plano+sexo+fxetaria+tamgrupo), family = Gamma(link = "log"))
AIC(model.fit9,model.fit10,model.fit11,model.fit12,model.fit13,model.fit14)

##Estimativa para o parâmetro de dispersão
fichapeu=sum(((log(resp)-model.fit9$fit)/model.fit9$fit)^2)/9993

##Teste bondade do ajuste
aux=146.56/fichapeu
1-pchisq(aux,9995)

#Quantil da distribuição qui-quadrado
qchisq(0.95,9995)

##Tratamento de Outliers
residuosglm=rstandard(model.fit9, type='deviance')
plot(residuosglm)
abline(h=c(-2,2), col="red")
sum(residuosglm>=2,residuosglm<=-2)

#Exclusão das observações da base
base.out3=cbind(base[,1:9],residuosglm)
base.out3=base.out3[(base.out3[,10]<=2),]
base.out3=base.out3[(base.out3[,10]>=-2),]

##Novo modelo (com tratamento de outliers)
#Variáveis da nova base
resp3 = base.out3[,9]/base.out3[,8]
periodo3 = base.out3[,2]
regiao3 = as.factor(base.out3[,3])
plano3 = base.out3[,4]
sexo3 = base.out3[,5]
fxetaria3 = base.out3[,6]
tamgrupo3 = as.factor(base.out3[,7])

##Modelo
model.fit15 <- glm((log(resp3))~(regiao3+plano3+sexo3+fxetaria3+tamgrupo3), family =
Gamma(link = "log"))
summary(model.fit15)
plot(model.fit15)
```

```
##Estimativa para o parâmetro de dispersão
fichapeu=sum(((log(resp3)-model.fit15$fit)/model.fit15$fit)^2)/9413

##Teste bondade do ajuste
aux=68.279/fichapeu
1-pchisq(aux,9413)

#Quantil da distribuição qui-quadrado
qchisq(0.95,9413)

##Teste de Observações Influentes
#Primeiro critério
hvalues = hatvalues(model.fit15)
aux=2*mean(hvalues)
sum(hvalues>aux)
#Segundo critério
cookvalues=cooks.distance(model.fit15)
sum(cookvalues>1)

##Análise de resíduos
residuosglm=rstandard(model.fit15, type='deviance')
```