

MARCOS VINICIUS DE ANDRADE
RODRIGO CAETANO DE SOUZA

UMA APLICAÇÃO DE MODELOS LINEARES GENERALIZADOS NA
MODELAGEM DE CUSTOS DOS USUÁRIOS DE UM PLANO DE SAÚDE

Trabalho de conclusão de curso apresentado à disciplina Laboratório de Estatística, Curso de Estatística, Setor de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial de obtenção do título de Estatístico.

Orientador: Prof.^o Dr.^o Cesar Augusto Taconeli

CURITIBA

2013

Aos nossos pais, que sempre nos deram
todo o apoio e suporte para que nossos
sonhos se tornassem realidade.

Ao meu irmão Renan (in memoriam).

Pela nossa amizade, que com certeza
continuará pelo resto de nossas vidas.

AGRADECIMENTOS

A Deus primeiramente, pela benção e proteção, por nos dar forças e nos guiar nos momentos de dificuldade.

Aos familiares, pelo apoio às nossas escolhas, que não mediram esforços para que chegássemos até esta etapa de nossas vidas.

Ao Professor Doutor Cesar Augusto Taconeli, pela disposição, paciência, competência ao compartilhar seus conhecimentos e, principalmente, por contribuir nesta conquista.

À Professora Mestra Fernanda Buhner Rizzato, pela disponibilidade em aceitar o convite para participar da banca deste trabalho.

À minha avó Otilia, que sempre me apoiou e incentivou constantemente.

À minha namorada Grazielle, pelo amor e paciência nos meus “maus” momentos e principalmente pela compreensão nos tempos de ausência.

Aos amigos da Casa do Estudante Universitário do Paraná, por me darem todas as condições de levar adiante meus estudos da melhor forma possível.

Aos amigos, pelo apoio e incentivo constantes.

*Há duas espécies de chatos: os chatos propriamente ditos e
[...] os amigos, que são nossos chatos prediletos.*

MÁRIO QUINTANA

RESUMO

Introdução: Com o aumento dos custos relacionados a procedimentos médicos, existe a necessidade de se utilizar métodos estatísticos para identificar fatores associados a maiores ou menores custos para as operadoras de planos de saúde, com o objetivo de redução de gastos. **Objetivo:** Propor modelos estatísticos que permitam analisar e estimar valores gastos com saúde, para a população de usuários, que demandaram um custo superior à franquia de resseguro de uma operadora de planos de saúde do Estado do Paraná, utilizando técnicas de Modelos Lineares Generalizados. **Métodos:** Foram propostos e ajustados modelos de regressão para a variável valor excedido à franquia de resseguro utilizando as distribuições Gama e Normal. Os parâmetros dos modelos foram estimados utilizando o método de máxima verossimilhança, e então, foram utilizados os testes de razão de verossimilhança e Wald para testar hipóteses de interesse, referentes aos parâmetros dos modelos. Para o ajuste do modelo, adotou-se, inicialmente, um nível de significância de $\alpha = 0,20$ como ponto de corte para inclusão das variáveis no modelo. Com base no conjunto de variáveis que atendessem tal critério, a seleção de variáveis baseou-se no método *backward*, utilizando como medida de qualidade de ajuste o Critério de Informação de Akaike (AIC). Dispondo-se de um modelo, adotou-se como medida de qualidade de ajuste a deviance escalonada. Para realizar o diagnóstico do modelo, foram usados gráficos de resíduos e cálculo de medidas de influência. **Resultados:** Num primeiro momento, o ajuste do modelo Normal mostrou-se inadequado, pois foram violados os pressupostos de normalidade e homocedasticidade dos erros. Obteve-se então, com o modelo de distribuição Gama e função de ligação logarítmica, um ajuste satisfatório aos dados em relação ao valor excedido. Verificou-se, por meio deste modelo, que variáveis relacionadas ao uso do plano juntamente com os procedimentos utilizados, que possuem a Classificação Internacional de Doenças (CID), estão mais correlacionadas ao valor excedido do que outras variáveis explicativas como, por exemplo, tipo de plano contratado e grau de dependência em relação ao titular do plano. A variável Idade ($p = 0,051$) chamou a atenção, pois apresentou um

coeficiente negativo (-0,006), indicando que quanto mais novo o usuário, menor será o seu gasto excedido.

Conclusão: Neste presente trabalho foi possível identificar, analisar e discutir os fatores associados ao valor excedido à franquia de resseguro, gerados por usuários de uma operadora de planos de saúde do estado do Paraná. Para estudos futuros, seria interessante analisar o tempo que o usuário levou até exceder o valor da franquia contratada e seus respectivos fatores associados, a fim de que sejam dados subsídios para a operadora de planos de saúde na tomada de ações com o intuito de minimizar os altos valores gastos com este usuário.

Palavras-chave: Planos de saúde. Modelos Lineares Generalizados. Teste da razão de verossimilhança. Distribuição Normal. Distribuição Gama. Estimação por Máxima Verossimilhança.

SUMÁRIO

1	INTRODUÇÃO	8
2	MATERIAL E MÉTODOS	10
2.1	MATERIAL	10
2.2	MÉTODOS	14
2.2.1	A Família Exponencial de Distribuições	14
2.2.2	Estrutura do Modelo Linear Generalizado	16
3	RESULTADOS E DISCUSSÃO	19
3.1	Análise Descritiva	19
3.2	Ajuste do Modelo Linear Normal	21
3.3	Processo de ajuste do modelo linear generalizado com resposta Gama	23
3.4	Predição com base no modelo ajustado com distribuição Gama	29
3.5	Processo de ajuste do modelo linear generalizado com resposta Normal Inversa	30
4	CONCLUSÃO	33
	REFERÊNCIAS	35

1 INTRODUÇÃO

Plano de saúde é um plano privado de assistência à saúde, com o objetivo de proteção ao beneficiário na possibilidade deste vir a incorrer em despesas médicas. A seguradora especializada em saúde é constituída em sociedade seguradora, com fins lucrativos, que comercializa seguros de saúde e oferece, obrigatoriamente, reembolso das despesas médicos-hospitalares, ou odontológicas, ou que comercializa ou opera seguro que preveja a garantia de assistência à saúde, estando sujeita ao disposto na Lei n.º 10.185, de 12 de fevereiro de 2001, sendo vedada a operação em outros ramos de seguro (ANS, 2009).

Com o aumento da renda dos brasileiros e a situação positiva do mercado de trabalho, aliados às menores taxas de desemprego já registradas e ao crescimento de vínculos formais com carteira assinada, a venda de planos de saúde aumentou consideravelmente, principalmente quanto aos planos coletivos e empresariais, que já somam 80% do mercado. Os planos coletivos cresceram 14% em 2010, mais que os individuais e familiares, que aumentaram em 5,4% o número de usuários (CFM, 2013).

A receita dos planos de saúde no primeiro trimestre de 2011 foi de R\$ 18,4 bilhões, 3,8% a mais que no mesmo período de 2010. O índice do crescimento do mercado de planos de saúde é confirmado pelo número de registros de operadoras. A quantidade registrada (12 novas empresas) foi maior que o número de cancelamento (4), no que se refere ao primeiro trimestre de 2011. Atualmente existem 1.042 planos ativos com beneficiários, dos quais 70% são de pequeno porte, com até 20.000 clientes. O valor médio cobrado pelos planos de saúde em 2010 foi de R\$ 127,00 per capita/mês, com aumento de 9,1% em relação a 2009, quando os planos custavam, em média, R\$ 116,37 (CFM, 2013).

Os gastos com a saúde podem ser resumidos, basicamente, por três segmentos diferentes: medicamentos, honorários de médicos, dentistas e auxiliares dos serviços médicos e, finalmente, gastos em hospitalização e tratamento. Atualmente, os gastos em hospitalização e tratamento encabeçam as despesas do consumo em saúde (ZUCCHI; NERO; MALI, 2000).

Duas causas distintas contribuem para o aumento do gasto com a saúde. A primeira diz respeito ao refinamento dos diagnósticos, onde o uso cada vez mais frequente de exames radiológicos e de análises laboratoriais é responsável pelo aumento dos gastos em saúde. Com isso, cada vez mais os preços dos diagnósticos e, conseqüentemente, o preço de uma diária hospitalar estão mais altos. Já a segunda causa é a difusão do progresso tecnológico que é o principal elemento do incremento das despesas em saúde, que aumentam à medida que novas técnicas são incorporadas (ZUCCHI; NERO; MALI, 2000).

Dados fornecidos pela Agência Nacional de Saúde Suplementar (ANS) mostram que o gasto médio por internação no país aumentou 15,5% entre 2007 e 2011. Os altos custos com gastos hospitalares foram ocasionados pelo aumento do número de internações, que teve um crescimento de 16,6% no período de 2011 a 2012. Em seguida, aparecem os tratamentos (15,1%), as consultas (13,3%) e os exames (9,8%) (D'ALAMA, 2013).

Visto o aumento dos custos com procedimentos médicos, o uso de metodologias estatísticas aplicadas à análise de custo de usuário de planos de saúde torna-se fundamental, viabilizando identificar fatores associados a maiores ou menores custos e fornecer previsões, subsidiando, do ponto de vista da operadora de planos de saúde, o planejamento e precificação de seus planos conforme o perfil e as demandas do cliente.

Com base no exposto, o presente trabalho tem como finalidade apresentar uma aplicação de modelos lineares generalizados na modelagem de custos dos usuários de um plano de saúde.

2 MATERIAL E MÉTODOS

2.1 MATERIAL

O banco de dados utilizado neste trabalho foi disponibilizado por uma operadora de planos de saúde que atua no Estado do Paraná, abrangendo 187 segurados que geraram, em um ano, um custo superior a R\$100.000,00, que é a franquia de resseguro contratada pela operadora, nos anos de 2010 a 2012. O resseguro é a operação pela qual a operadora de planos de saúde transfere a outra seguradora total ou parcialmente, o risco assumido do beneficiário. Sendo assim, quando a franquia de resseguro é atingida, o excedente a esse valor é pago pela empresa contratada pela operadora, evitando assim picos de sinistralidade que possam vir a ocorrer, mantendo a solvência da seguradora e a diluição de seus riscos.

Além de o gasto ser superior à franquia contratada, eles se enquadram em ao menos uma das seguintes situações: internação hospitalar, internação domiciliar (desde que caracterizada como prolongamento da internação hospitalar) ou ainda, no caso de atendimento ambulatorial, em especial os tratamentos de quimioterapia e radioterapia.

Para efeito de correção monetária dos valores, foi necessário que as quantias gastas fossem reajustadas de acordo com a inflação. O indicador de inflação utilizado para o reajuste foi o Índice de Variação dos Custos Médico-Hospitalares (IVCMH), divulgado pelo Instituto de Estudos de Saúde Suplementar (IESS), que no ano de 2011 foi de 12,9% e em 2012 foi de 15,4%.

O Quadro 1 apresenta uma visão simplificada das variáveis explicativas utilizadas no estudo.

Código	Descrição	Tipo de Variável	Categoria
Vlr	Valor excedente à franquia	Numérica	
Grau_dp	Grau de dependência em relação ao titular do plano	Categórica	Titular Dependente Agregado
Empr	Empresa em que o usuário mantém vínculo empregatício	Categórica	Empr 1 Empr 2 Empr 3 Empr 4 Empr 5 Empr Outras
Idd	Idade (em anos)	Numérica	
Sexo	Gênero	Categórica	Masculino Feminino
Cidade	Cidade	Categórica	Curitiba e Região Outras
Acom	Tipo de acomodação	Categórica	Apartamento Coletiva
Cap1	Capítulo 1 do CID	Categórica	0 - Ausência 1 - Presença
Cap2	Capítulo 2 do CID	Categórica	0 - Ausência 1 - Presença
Cap3	Capítulo 3 do CID	Categórica	0 - Ausência 1 - Presença
Cap4	Capítulo 4 do CID	Categórica	0 - Ausência 1 - Presença
Cap5	Capítulo 5 do CID	Categórica	0 - Ausência 1 - Presença
Cap6	Capítulo 6 do CID	Categórica	0 - Ausência 1 - Presença
Cap7	Capítulo 7 do CID	Categórica	0 - Ausência 1 - Presença
Cap8	Capítulo 8 do CID	Categórica	0 - Ausência 1 - Presença
Cap9	Capítulo 9 do CID	Categórica	0 - Ausência 1 - Presença
Cap10	Capítulo 10 do CID	Categórica	0 - Ausência 1 - Presença
Cap11	Capítulo 11 do CID	Categórica	0 - Ausência 1 - Presença
Cap12	Capítulo 12 do CID	Categórica	0 - Ausência 1 - Presença

QUADRO 1 - RELAÇÃO DAS VARIÁVEIS ANALISADAS

continua

conclusão

Código	Descrição	Tipo de Variável	Categoria
Cap13	Capítulo 13 do CID	Categórica	0 - Ausência
			1 - Presença
Cap14	Capítulo 14 do CID	Categórica	0 - Ausência
			1 - Presença
Cap15	Capítulo 15 do CID	Categórica	0 - Ausência
			1 - Presença
Cap16	Capítulo 16 do CID	Categórica	0 - Ausência
			1 - Presença
Cap17	Capítulo 17 do CID	Categórica	0 - Ausência
			1 - Presença
Cap18	Capítulo 18 do CID	Categórica	0 - Ausência
			1 - Presença
Cap19	Capítulo 19 do CID	Categórica	0 - Ausência
			1 - Presença
Cap20	Capítulo 20 do CID	Categórica	0 - Ausência
			1 - Presença
Cap21	Capítulo 21 do CID	Categórica	0 - Ausência
			1 - Presença
CapMiss	Quando não há informação referente ao CID	Categórica	0 - Ausência
			1 - Presença
ConsEI	Número de Consultas Eletivas	Numérica	
ConsEm	Número de Consultas de Emergência	Numérica	
Exa	Número de Exames	Numérica	
Int	Número de Internamentos	Numérica	
Amb	Número de Eventos Ambulatoriais	Numérica	
Hos	Número de Eventos Hospitalares	Numérica	
Mat	Número de Materiais Hospitalares	Numérica	
Med	Número de Medicamentos	Numérica	

QUADRO 1 - RELAÇÃO DAS VARIÁVEIS ANALISADAS
 FONTE: O autor (2013)

Algumas observações se fazem necessárias para a adequada compreensão das variáveis relacionadas no Quadro 1. Inicialmente foi denominada a variável empresa como o local onde os usuários do plano de saúde mantinham vínculo empregatício. Na sequência, foi codificado como “Empr 1” a empresa com maior número de funcionários na amostra, “Empr 2” para a segunda mais frequente e assim sucessivamente para as cinco primeiras empresas. As demais empresas foram agrupadas como “Empr Outras”.

A Classificação Internacional de Doenças (CID) é a ferramenta de diagnóstico padrão para a epidemiologia, gestão da saúde e para fins clínicos. O CID é usado para classificar doenças e outros problemas de saúde inscritos em muitos tipos de registros de saúde e vital, incluindo certidões de óbito (World Health Organization – WHO, 1985). Para facilitar a análise, os CIDs foram agrupados em 21 Capítulos. Na etapa de modelagem, os capítulos que apresentaram uma frequência abaixo de 5% foram agrupados em uma categoria chamada “Outros Capítulos”, composta por Capítulo 15, Capítulo 16, Capítulo 17 e Capítulo 20. Além disso, foi criada a categoria “Capítulo *Missing*”, que ocorreu quando, no prontuário médico, não houve a informação do CID por parte do médico. O Quadro 2 apresenta os 21 Capítulos com suas respectivas descrições.

Nome	Descrição
Capítulo 1	Algumas doenças infecciosas e parasitárias
Capítulo 2	Neoplasmas (tumores)
Capítulo 3	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários
Capítulo 4	Doenças endócrinas, nutricionais e metabólicas
Capítulo 5	Transtornos mentais e comportamentais
Capítulo 6	Doenças do sistema nervoso
Capítulo 7	Doenças do olho e anexos
Capítulo 8	Doenças do ouvido e da apófise mastoide
Capítulo 9	Doenças do aparelho circulatório
Capítulo 10	Doenças do aparelho respiratório
Capítulo 11	Doenças do aparelho digestivo
Capítulo 12	Doenças da pele e do tecido subcutâneo
Capítulo 13	Doenças do sistema osteomuscular e do tecido conjuntivo
Capítulo 14	Doenças do aparelho geniturinário
Capítulo 15	Gravidez, parto e puerpério
Capítulo 16	Algumas afecções originadas no período perinatal
Capítulo 17	Malformações congênicas, deformidades e anomalias cromossômicas
Capítulo 18	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte
Capítulo 19	Lesões, envenenamentos e algumas outras consequências de causas externas
Capítulo 20	Causas externas de morbidade e de mortalidade
Capítulo 21	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde

QUADRO 2 – DESCRIÇÃO DOS CAPÍTULOS DOS CIDs

FONTE: CID-10

2.2 MÉTODOS

Os Modelos Lineares Generalizados (Nelder e Wedderburn, 1972) configuram extensões dos modelos normais lineares. A ideia básica consiste em ampliar as opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições. Dentre tais distribuições destacam-se a Binomial, Poisson, Gama, Normal e Normal Inversa, configurando possibilidades de modelos de regressão para dados dos tipos discretos e contínuos, com distribuição simétrica ou assimétrica (PAULA, 2013).

2.2.1 A Família Exponencial de Distribuições

Considere uma variável aleatória com função densidade (ou função de probabilidade) pertencente à família exponencial, podendo ser expressa na forma:

$$f(Y; \theta, \phi) = \exp\{\phi^{-1}[y\theta - b(\theta)] + c(Y, \phi)\}, \quad (1)$$

tendo-se que $E(Y) = \mu = b'(\theta)$ e $Var(Y) = \phi^{-1}V(\mu)$ em que $V = V(\mu) = d\mu/d\theta$ é a função de variância e $\phi^{-1} > 0$ ($\phi > 0$) é o parâmetro de dispersão (precisão) do modelo (Paula, 2013).

Nos tópicos seguintes, iremos apresentar as distribuições pertencentes à família exponencial que serão consideradas no presente trabalho e suas propriedades.

- **Distribuição Normal**

Uma variável aleatória tem distribuição Normal de média μ e variância σ^2 , denotada por $Y \sim N(\mu, \sigma^2)$, se sua função densidade de probabilidade for dada por:

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

Verifica-se que:

$$\theta = \mu, b(\theta) = \frac{\theta^2}{2}, \phi = \sigma^{-2}, c(y, \phi) = \frac{1}{2} \frac{\log \phi}{2\pi} - \frac{\phi y^2}{2} \text{ e } V(\mu) = 1.$$

- **Distribuição Gama**

Uma variável aleatória tem distribuição Gama de média μ e coeficiente de variação $\phi^{-1/2}$, denotada por $Y \sim G(\mu, \phi)$, se sua função densidade de probabilidade for dada por:

$$f(y) = \frac{1}{\Gamma(\phi)} \left(\frac{\phi y}{\mu}\right)^\phi \exp\left(-\frac{\phi y}{\mu}\right) d(\log y),$$

em que $y > 0$, $\phi > 0$, $\mu > 0$ e $\Gamma(\phi) = \int_0^\infty t^{\phi-1} e^{-t} dt$ é a função Gama. Verifica-se que:

$$\theta = -\frac{1}{\mu}, b(\theta) = -\log(-\theta), c(y, \phi) = (\phi - 1) \log y + \phi \log \phi - \log \Gamma(\phi).$$

Temos que $\phi = E^2(Y)/(CV)^2$, em que $CV = \sqrt{\text{Var}(y)/E(y)}$ é o coeficiente de variação. A função de variância da distribuição Gama é dada por $V(\mu) = \mu^2$. É possível verificar que, à medida que ϕ aumenta, a distribuição Gama fica mais simétrica em torno da média. Mais especificamente à medida que ϕ aumenta a distribuição Gama se aproxima de uma distribuição Normal de média μ e variância $\mu^2 \phi^{-1}$.

- **Distribuição Normal Inversa**

Uma variável aleatória tem distribuição Normal Inversa de média μ e parâmetro de precisão ϕ , denotada por $Y \sim NI(\mu, \phi)$, se sua função densidade de probabilidade for dada por:

$$f(y) = \frac{\phi^{1/2}}{\sqrt{2\pi y^3}} \exp\left\{-\frac{\phi(y-\mu)^2}{2\mu^2 y}\right\},$$

em que $y > 0$, $\mu > 0$. Verifica-se que:

$$\theta = -\frac{1}{2\mu^2}, b(\theta) = -(-2\theta)^{1/2} \text{ e } c(y, \phi) = \frac{1}{2} \log\left\{\frac{\phi}{2\pi y^3}\right\} - \frac{\phi}{2y}.$$

A distribuição Normal Inversa se aproxima da distribuição Normal quando $\phi \rightarrow \infty$. Ou seja, para ϕ grande temos que $Y \sim N(\mu, \phi^{-1}V(\mu))$. A função de variância fica aqui dada por $V(\mu) = \mu^3$.

2.2.2 Estrutura do Modelo Linear Generalizado

Um modelo linear generalizado é constituído por três componentes:

- Componente aleatório - Variáveis aleatórias independentes Y_1, Y_2, \dots, Y_n com distribuição pertencente à família exponencial;
- Componente sistemático - Consiste numa combinação linear de variáveis preditoras $(x_1, x_2, x_3, \dots, x_p)$, conforme apresentado em (2);

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} \quad i = 1, \dots, n. \quad (2)$$

- Função de ligação - uma função diferenciável e monótona g que associa os componentes aleatório e sistemático, conforme apresentado em (3).

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ji} \quad i = 1, \dots, n. \quad (3)$$

O método utilizado para estimar os parâmetros neste trabalho foi o método de máxima verossimilhança, que produz estimadores consistentes com eficiência assintótica. (CORDEIRO; DEMÉTRIO, 2007). Supondo ϕ conhecido, o logaritmo da função de verossimilhança de um modelo linear generalizado com respostas independentes pode ser expresso na forma:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n \phi \{y_i \theta_i - b(\theta_i)\} + \sum_{i=1}^n c(y_i, \phi).$$

A qualidade de ajuste do modelo linear generalizado é avaliada através da função desvio:

$$D(\underline{y}; \hat{\mu}) = 2 \sum_{i=1}^n \{y_i (\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\},$$

que é resultante da diferença entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de máxima verossimilhança $\hat{\beta}$, onde $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\tilde{\theta}_i = \theta_i(\tilde{\mu}_i)$ são as estimativas de máxima verossimilhança de θ para os modelos com p parâmetros e saturado, respectivamente (PAULA, 2013).

Caso o parâmetro de dispersão seja desconhecido, deve-se estimá-lo, por exemplo, pelo método da máxima verossimilhança ou método dos momentos, e incorporá-lo aos resultados. Neste caso é definido o desvio escalonado por:

$$D^*(y, \hat{\mu}) = \frac{D(y, \hat{\mu})}{\hat{\phi}},$$

onde $D(y, \hat{\mu})$ é o desvio, conforme definido anteriormente. Em ambos os casos, a fim de avaliar a qualidade do ajuste, tem-se que assintoticamente, para um bom ajuste do modelo, a função desvio segue distribuição qui-quadrado com $n - p$ graus de liberdade, onde n é o tamanho da amostra e p é o número de parâmetros do modelo.

Para realizar a verificação da diferença entre os desvios de dois modelos é utilizado o teste de razão de verossimilhança (TRV), onde a hipótese a ser testada é do tipo $H_0: \beta_q = 0$ contra $H_1: \beta_q \neq 0$, sendo β_q um subconjunto do vetor de parâmetro $\underline{\beta}$. Para exemplificar o uso das diferenças de desvios para testar hipóteses em modelos encaixados, pode-se aplicá-lo a um modelo linear generalizado com dois fatores, A e B, onde é testado o efeito da inclusão do fator B dado que o fator A já está no modelo.

Uma das técnicas utilizada para diagnosticar a adequação dos modelos é a análise de resíduos, a qual tem por finalidade detectar a presença de pontos mal ajustados e avaliar o ajustamento da distribuição proposta para a variável resposta. Outro método importante na análise de diagnóstico é a detecção de observações influentes, isto é, pontos que exercem um peso desproporcional nas estimativas dos parâmetros do modelo. A distância de Cook é uma das técnicas utilizadas para

avaliar o impacto do afastamento do vetor de estimativas, provocado pela retirada de uma observação particular (CORDEIRO; DEMÉTRIO, 2007).

No desenvolvimento do presente trabalho, inicialmente, foi realizada uma análise descritiva dos dados, calculando medidas resumo (média, mediana, desvio padrão e coeficiente de variação) do valor que foi excedido à franquia contratada, segmentado pelas categorias das variáveis explicativas. Após a análise descritiva foi ajustado um modelo linear com distribuição Normal e posteriormente, foi ajustado um modelo considerando distribuição Gama para a resposta, a fim de encontrar um modelo que explique adequadamente a distribuição do valor excedido. Na triagem das variáveis para o ajuste do modelo de distribuição Gama foi testado o efeito da inclusão de cada uma das variáveis ao modelo nulo, verificando sua significância via teste de razão de verossimilhança. Então foram selecionadas apenas as variáveis que tiveram a probabilidade de significância inferior a 0,20. Considerando as variáveis incluídas nesse primeiro passo, procedeu-se com a execução do método de seleção de variáveis *backward*, baseado na minimização do Critério de Informação de Akaike (AIC) para a seleção do modelo com melhor ajuste. Os modelos foram ajustados usando o método de estimação da máxima verossimilhança e os parâmetros foram testados via teste de Wald e razão de verossimilhança (CASELLA; BERGER, 2010). Para a análise de diagnóstico foi realizado o estudo de pontos de alavanca, pontos de influência e análise dos resíduos, além do cálculo de medidas de qualidade de ajuste, como o desvio escalonado.

Todas as análises realizadas neste trabalho foram feitas com o auxílio dos softwares R (R CORE TEAM, 2012) e Microsoft Excel 2013 (MICROSOFT, 2013).

3 RESULTADOS E DISCUSSÃO

3.1 Análise Descritiva

Num primeiro momento, apresenta-se na Figura 1 a distribuição do valor excedido considerando todos os indivíduos. Nota-se que, em geral, os dados apresentam uma distribuição assimétrica à direita. Adicionalmente, tem-se a média do valor excedido igual a R\$ 104 mil, desvio padrão igual a R\$ 117 mil e um coeficiente de variação igual a 1,125, o que indica uma alta variabilidade nos dados.

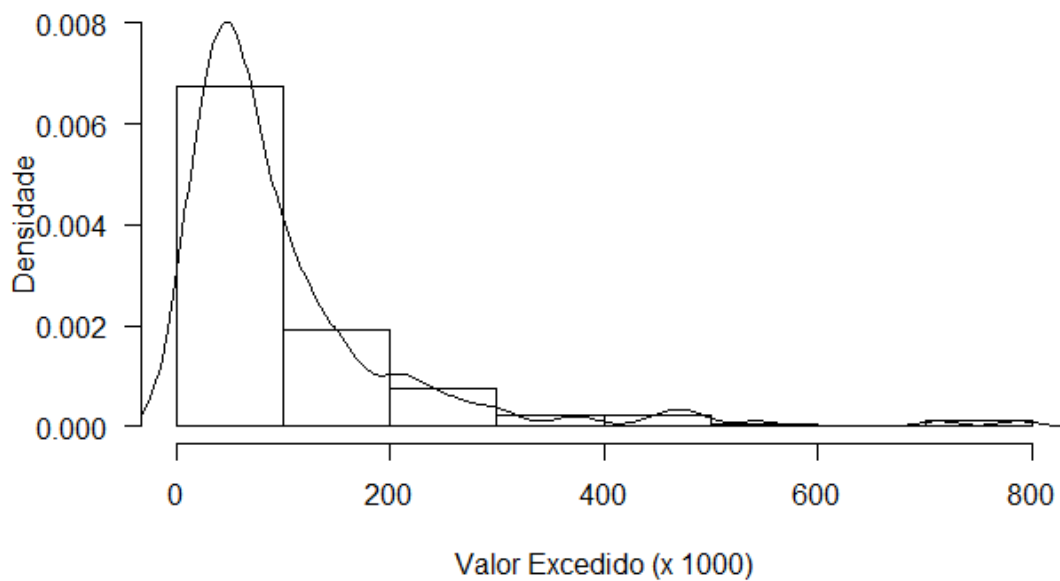


FIGURA 1 – Distribuição do Valor Excedido para os 187 indivíduos da população
FONTE: O autor (2013)

A Tabela 1 apresenta uma análise descritiva do valor excedido segundo as variáveis categorizadas.

TABELA 1 – ANÁLISE DESCRITIVA DO VALOR EXCEDIDO SEGUNDO AS CATEGORIAS DAS VARIÁVEIS EXPLICATIVAS.

					continua
Variável	Categoria	Média (R\$)	Mediana (R\$)	Desvio Padrão (R\$)	Coefficiente de Variação
Gênero	Masculino	117.127,74	69.107,67	137.120,77	1,17
	Feminino	92.364,03	64.044,87	92.720,94	1
Acomodação	Coletiva	103.867,83	63.973,41	113.116,82	1,09
	Individual	104.706,24	68.682,60	119.192,53	1,14
Abrangência	Curitiba e Região	104.982,94	74.199,44	111.589,28	1,06
	Outros	103.898,87	63.054,20	121.917,97	1,17
Empresa	Empresa 1	92.375,69	44.962,50	151.847,62	1,64
	Empresa 2	59.848,58	48.608,55	49.455,50	0,83
	Empresa 3	87.766,05	69.053,78	64.949,80	0,74
	Empresa 4	94.308,53	78.199,87	65.794,92	0,70
	Empresa 5	124.284,25	68.682,60	146.759,54	1,18
	Outras	111.758,69	75.035,58	114.230,04	1,02
CAPITULO1	Sim	129.312,80	73.368,83	160.198,93	1,64
	Não	97.422,92	63.015,09	100.689,52	1,03
CAPITULO2	Sim	94.030,74	62.486,48	136.519,03	1,21
	Não	113.256,70	69.107,67	87.962,56	0,78
CAPITULO3	Sim	173.059,79	79.398,35	212.665,81	2,29
	Não	92.830,98	64.044,87	86.427,33	0,93
CAPITULO4	Sim	108.477,91	70.825,44	97.570,73	0,95
	Não	103.164,63	63.897,68	122.478,39	1,19
CAPITULO5	Sim	104.842,43	60.003,29	108.013,02	1,03
	Não	104.369,27	65.901,51	118.045,28	1,13
CAPITULO6	Sim	142.921,57	101.883,98	158.867,28	1,75
	Não	90.742,13	61.053,93	94.401,74	1,04
CAPITULO7	Sim	116.587,60	93.594,16	108.824,91	1,08
	Não	100.779,89	63.015,09	119.240,98	1,18
CAPITULO8	Sim	87.708,66	42.349,61	109.549,35	1,03
	Não	106.194,17	68.257,54	117.754,91	1,11
CAPITULO9	Sim	103.376,27	68.257,54	119.118,57	1,13
	Não	105.688,27	63.936,79	114.602,33	1,08
CAPITULO10	Sim	126.757,43	67.504,59	151.144,91	1,71
	Não	88.426,53	62.132,49	81.063,76	0,92
CAPITULO11	Sim	109.648,58	58.884,85	134.296,39	1,31
	Não	102.504,68	73.368,83	110.120,15	1,07
CAPITULO12	Sim	96.592,77	61.583,57	96.779,97	0,91
	Não	105.909,48	65.569,62	120.555,81	1,14
CAPITULO13	Sim	92.276,02	61.098,08	103.293,93	0,94
	Não	109.603,92	72.505,83	122.184,01	1,11
CAPITULO14	Sim	87.661,01	61.032,60	112.444,89	1,01
	Não	111.760,72	70.825,44	118.362,37	1,06
CAPITULO18	Sim	108.190,40	65.901,51	123.203,01	1,3
	Não	95.115,70	62.525,59	99.962,84	1,05

TABELA 1 – ANÁLISE DESCRITIVA DO VALOR EXCEDIDO SEGUNDO AS CATEGORIAS DAS VARIÁVEIS EXPLICATIVAS.

Variável	Categoria	Média (R\$)	Mediana (R\$)	Desvio Padrão (R\$)	conclusão
					Coefficiente de Variação
CAPITULO19	Sim	111.191,23	82.092,88	100.683,43	0,98
	Não	102.686,60	63.973,41	120.890,62	1,18
CAPITULO21	Sim	106.688,87	31.487,52	42.558,73	0,4
	Não	92.513,96	65.735,56	118.144,13	1,12
CAPITULO MISSING	Sim	108.754,82	68.257,54	121.026,07	1,31
	Não	87.397,45	61.582,54	93.145,10	1,01
OUTROS CAPITULOS	Sim	86.589,85	73.368,83	114.085,14	1,05
	Não	106.549,54	52.082,46	126.892,01	1,45

FONTE: O autor (2013)

Na Tabela 1, nota-se, a título de exemplo, que os usuários do gênero masculino geraram um gasto médio superior aos usuários do gênero feminino (R\$ 117 mil contra R\$ 92 mil). Indivíduos que apresentaram procedimentos referentes ao Capítulo 3 também apresentaram um gasto médio muito superior àqueles que não os apresentaram (R\$ 173 mil contra R\$ 92 mil). Também se observa que os coeficientes de variação apresentam valores superiores a 0,7, indicando novamente que existe uma alta variabilidade dos dados.

3.2 Ajuste do Modelo Linear Normal

Inicialmente optou-se por ajustar um modelo de regressão com resposta Normal e função de ligação identidade, contendo todas as variáveis explicativas. Verificou-se que as variáveis Idade, Capítulo 1, Capítulo 3, Capítulo 6, Capítulo 7, Capítulo 10, Capítulo 14 e número de Internamentos foram marginalmente significativas ao nível de 10% confiança. O modelo ajustado gerou um coeficiente de determinação (R^2) de 0,148, indicando que o modelo explica 14,85% da variação total do valor excedido, evidenciando um ajuste insatisfatório. Para fins de avaliação da qualidade do ajuste, a Figura 2 apresenta a distribuição dos resíduos padronizados pelos valores ajustados.

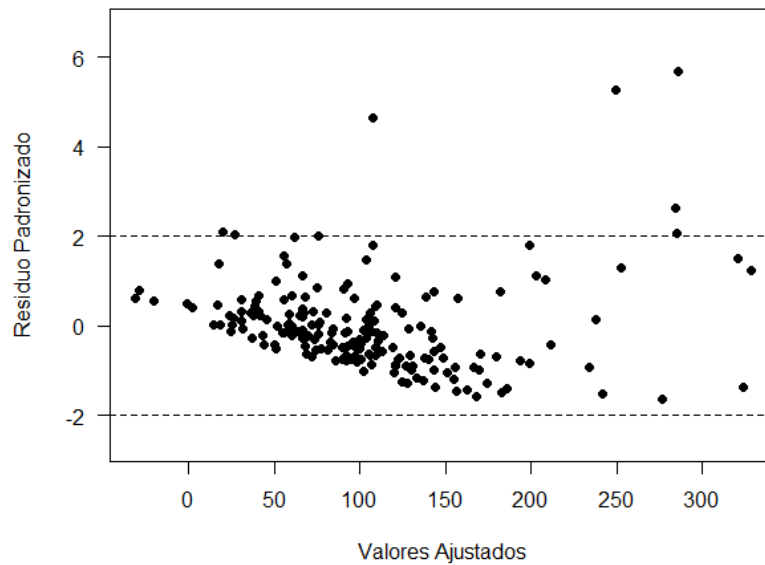


FIGURA 2 – Distribuição dos resíduos padronizados *versus* valores ajustados para o modelo com distribuição Normal e função de ligação identidade
 FONTE: O autor (2013)

Observando a Figura 2 notamos que os resíduos não atendem ao pressuposto de homocedasticidade, apresentando um aumento na variância conforme os valores ajustados aumentam. A seguir, a Figura 3 mostra os resíduos padronizados no gráfico probabilístico Normal com envelope simulado.

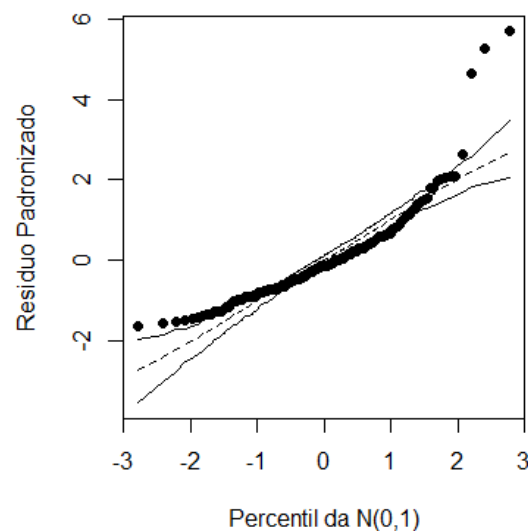


FIGURA 3 – Gráfico probabilístico Normal para os resíduos padronizados com envelope simulado para o modelo com distribuição Normal e função de ligação identidade
 FONTE: O autor (2013)

Observa-se que os resíduos não se ajustam bem à distribuição Normal, pois há diversos pontos fora do envelope simulado, além de uma clara tendência não linear. Portanto, temos também um indicativo de violação do pressuposto de normalidade dos erros. Com base nestes resultados, cogitou-se a possibilidade de se ajustar um modelo de regressão linear aplicando alguma transformação na variável resposta. Entretanto, como a interpretação dos resultados sob transformação dos dados é complicada, optou-se por encontrar outra distribuição à qual os dados melhor se ajustem.

3.3 Processo de ajuste do modelo linear generalizado com resposta Gama

Como os dados aparentemente tem distribuição assimétrica e com a verificação da violação dos pressupostos de normalidade e homocedasticidade dos erros para o modelo linear Normal, partiu-se para a modelagem com base em distribuições assimétricas, que permitam ainda acomodar uma relação não constante entre variância e média, como a Gama e Normal Inversa, iniciando pela distribuição Gama.

A fim de selecionar um conjunto inicial de covariáveis para o modelo, foi ajustado um modelo nulo, somente com o intercepto, analisando-se o efeito da inclusão de cada uma das covariáveis separadamente utilizando o teste de razão de verossimilhança. Ao nível de 20% de significância foram selecionadas, para a sequência do ajuste do modelo de regressão, as seguintes covariáveis: Idade, Sexo, Período em que o indivíduo esteve no plano, Capítulo 1, Capítulo 2, Capítulo 3, Capítulo 6, Capítulo 10, Capítulo 14, número de Exames realizados e número de Internamentos.

TABELA 2 – VARIÁVEIS SIGNIFICATIVAS PARA O MODELO GAMA A 20 % DE SIGNIFICÂNCIA PELO TRV

VARIÁVEL	Valor F	P-Valor
IDADE	4,82	0,03
SEXO	2,79	0,10
PERIODO	5,55	0,02
CAPITULO 1	2,86	0,09
CAPITULO 2	1,68	0,20
CAPITULO 3	11,46	< 0,01
CAPITULO 6	8,71	0,00
CAPITULO 10	6,46	0,01
CAPITULO 14	2,39	0,12
EXAMES	2,94	0,09
INTERNAMENTO	16,47	< 0,01

FONTE: O autor (2013)

Depois de feita a triagem inicial das variáveis, ajustou-se um modelo com distribuição Gama e função de ligação logarítmica com base nas variáveis retidas nesta primeira etapa. A partir deste modelo aplicou-se o processo de seleção de variáveis *backward*, baseado na minimização do AIC, com os resultados sendo apresentados na Tabela 3.

TABELA 3 – DESCRIÇÃO DO PROCESSO DE SELEÇÃO DE COVARIÁVEIS UTILIZANDO O MÉTODO *BACKWARD*

Passo 1		Passo 2		Passo 3		Passo 4	
Modelo	AIC	Modelo	AIC	Modelo	AIC	Modelo	AIC
- Sexo	2086,21	- Cap 2	2084,249	- Per	2082,75	Modelo 3	2082,75
- Cap 2	2086,25	- Per	2084,724	- Cap 6	2084,02	- Cap 6	2082,93
- Per	2086,72	- Cap 6	2085,829	Modelo 2	2084,25	- Cap 10	2083,44
- Cap 6	2087,75	Modelo 1	2086,206	- Cap 3	2084,81	- Cap 1	2083,46
Todas	2088,20	- Cap 3	2086,808	- Cap 1	2085,00	- Cap 3	2083,79
- Cap 1	2088,69	- Cap 1	2086,887	- Exa	2085,00	- Idade	2084,24
- Cap 3	2088,80	- Exa	2086,909	- Idade	2085,08	- Exa	2084,39
- Exa	2088,82	- Cap 10	2087,007	- Cap 10	2085,12	- Cap 14	2086,44
- Idade	2088,88	- Idade	2087,069	- Cap 14	2087,98	- Intern	2092,41
- Cap 10	2089,00	- Cap 14	2089,938	- Intern	2093,94		
- Cap 14	2091,94	- Intern	2095,858				
- Intern	2096,27						

FONTE: O autor (2013)

Com base na Tabela 3 verifica-se que no “Passo 1” o modelo composto por todas as variáveis, exceto a variável Sexo, apresentou o menor AIC, indicando ser o melhor ajuste. Posteriormente, ajustou-se um novo modelo sem a variável Sexo (Modelo 1) e comparou-se seu AIC com os AICs dos modelos sem cada uma das

variáveis restantes, tendo-se o menor AIC para o modelo sem as variáveis Sexo e Capítulo 2 (Neoplasmas), que foi adotado no passo seguinte.

Como no “Passo 4” o modelo com o melhor ajuste foi o modelo ajustado no “Passo 3”, sem as variáveis Sexo, Capítulo 2 e Período, conclui-se que o melhor modelo de regressão com distribuição Gama e com função de ligação logarítmica, é composto pelas variáveis Idade, Capítulo 1, Capítulo 3, Capítulo 6, Capítulo 10, Capítulo 14, número de Exames e número de Internamentos.

A seguir, na Tabela 4 são apresentadas as estimativas do parâmetro do modelo ajustado com este conjunto de covariáveis, seus respectivos erros padrões, estatísticas testes e p-valores, além de intervalos de 95% de confiança, que são calculados com base na exponencial das estimativas.

TABELA 4 – ESTIMATIVA DOS PARÂMETROS DO MODELO GAMA, ACOMPANHADOS DOS ERROS PADRÕES, ESTATÍSTICAS DE TESTES E P-VALORES

Parâmetros	Estimativa	Exp (Estimativa)	Erro Padrão	IC (95%)	P-Valor
INTERCEPTO	4,204	66,954	0,275	[39,032 ; 114,849]	< 0,001
IDADE	-0,006	0,994	0,003	[0,988 ; 1,000]	0,051
(CAPITULO 1)Sim	0,245	1,278	0,154	[0,945 ; 1,727]	0,112
(CAPITULO 3)Sim	0,314	1,369	0,185	[0,953 ; 1,967]	0,092
(CAPITULO 6)Sim	0,216	1,241	0,152	[0,922 ; 1,671]	0,156
(CAPITULO 10)Sim	0,216	1,241	0,134	[0,954 ; 1,615]	0,110
(CAPITULO 14)Sim	-0,331	0,718	0,139	[0,547 ; 0,944]	0,019
EXAMES	0,001	1,001	0,000	[1,000 ; 1,002]	0,080
INTERNAMENTO	0,495	1,640	0,172	[1,172 ; 2,296]	0,004

FONTE: O autor (2013)

A partir das estimativas apresentadas na Tabela 4, o modelo ajustado pode ser escrito da seguinte forma:

$$\hat{\mu}_i = E(\widehat{Y/x_i}) = \exp\{4,204 - 0,006\text{Idd}_i + 0,245\text{Cap1}_i + 0,314\text{Cap3}_i + 0,216\text{Cap6}_i + 0,216\text{Cap10}_i - 0,331\text{Cap14}_i + 0,001\text{Exa}_i + 0,495\text{Int}_i\},$$

onde Idd_i corresponde a variável numérica idade, Cap1_i é a variável indicadora ao Capítulo 1, assumindo valor 1 se o usuário apresenta procedimentos contemplados pelo Capítulo 1 e 0, caso contrário. Da mesma forma, Cap3_i , Cap6_i , Cap10_i , Cap14_i são as variáveis indicadoras dos Capítulos 3, 6, 10, 14 e Exa_i e Int_i correspondem, respectivamente, aos números de exames e internamentos, e $\hat{\mu}_i$ o valor excedido ajustado pelo modelo, para os indivíduos. A estimativa do parâmetro de dispersão,

obtida pelo método da máxima verossimilhança, foi $\hat{\phi} = 0,682$, com erro padrão de 0,138.

O desvio para o modelo ajustado foi $D = 141,39$, com $n - p = 179$ graus de liberdade. Logo, o desvio escalonado fica dado por $D^* = 207,3$. Verificou-se a qualidade do ajuste do modelo através da distribuição qui-quadrado assintótica, tendo-se obtido um p-valor de 0,17, o que não fornece um indicativo de mau ajuste do modelo. Na Figura 4 são apresentados gráficos de diagnóstico do modelo ajustado com distribuição Gama.

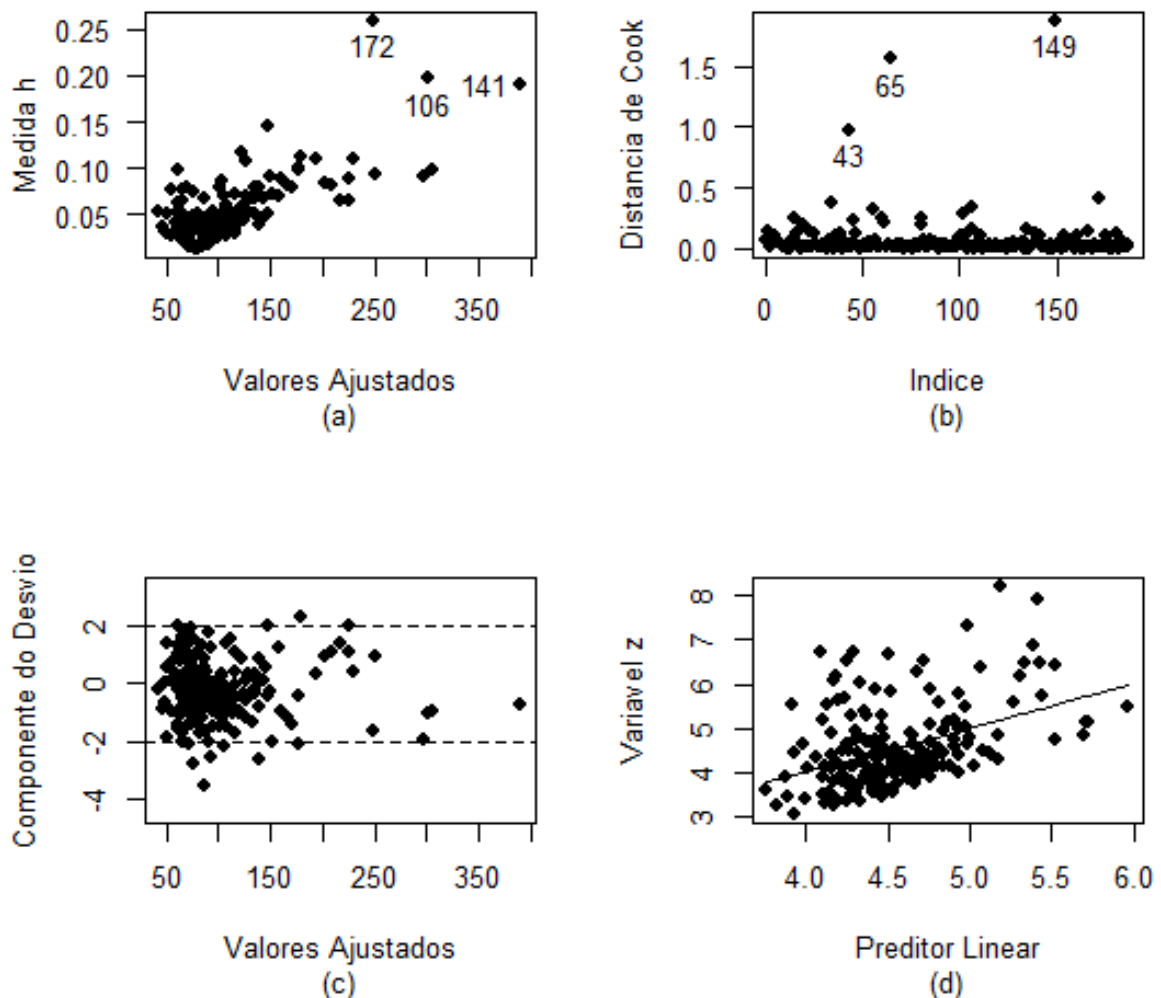


FIGURA 4: Gráficos de diagnóstico referentes ao modelo de distribuição Gama ajustado aos dados sobre o valor excedido

FONTE: O autor (2013)

Observa-se na Figura 4 em (a) três possíveis pontos de alavanca (#172, #106 e #141). Já em (b) outros três pontos são destacados como possivelmente influentes (#43, #65 e #149), não havendo indicativo de heterocedasticidade (c), e não havendo indícios de não linearidade em (d), o que avaliza a parte sistemática e a função de ligação do modelo.

A fim de avaliar o efeito dos pontos destacados nos resultados, foram ajustados novos modelos sem tais observações, com estimativas e p-valores, apresentados na Tabela 5. Retirando as observações #43, #65 e #149 conjuntamente, nota-se que as variáveis Capítulo 3 e número de Exames tornam-se não significativas para o modelo ao nível de 10% de significância, indicando que essas três observações, conjuntamente, induzem os efeitos destas duas variáveis. Retirando as observações #172, #106 e #141, não foram observadas mudanças inferenciais significativas.

TABELA 5 – ESTIMATIVAS E COEFICIENTES DO MODELO AJUSTADO E DO MODELO SEM POSSÍVEIS PONTOS INFLUENTES E DE ALAVANCA

Parâmetro	Modelo com todas as observações		Modelo sem as observações 172, 106 e 141		Modelo sem as observações 43, 65 e 149	
	Estimativa	Valor p	Estimativa	Valor p	Estimativa	Valor p
INTERCEPTO	4,204	< 0,001	4,099	< 2e-16	4,360	< 2e-16
IDADE	-0,006	0,051	-0,005	0,097	-0,006	0,034
(CAPITULO 1)Sim	0,245	0,112	0,243	0,114	0,120	0,424
(CAPITULO 3)Sim	0,314	0,092	0,371	0,047	0,077	0,678
(CAPITULO 6)Sim	0,216	0,156	0,220	0,148	0,178	0,229
(CAPITULO 10)Sim	0,216	0,11	0,221	0,099	0,117	0,366
(CAPITULO 14)Sim	-0,331	0,019	-0,364	0,010	-0,413	0,002
EXAMES	0,001	0,08	0,001	0,028	0,000	0,331
INTERNAMENTO	0,495	0,004	0,511	0,006	0,470	0,005

FONTE: O autor (2013)

O Quadro 3 apresenta as características dos usuários que apresentaram-se como possíveis pontos de alavanca e influência. A observação #65 teve um valor excedido de R\$ 487 mil, sendo o quarto indivíduo que mais gerou gastos, tendo realizado 1.019 exames, um internamento e com 59 anos. O indivíduo #43 é o que teve o maior gasto (R\$ 787 mil), tendo realizado 32 exames, um internamento e com 47 anos de idade. O indivíduo #149 foi o segundo que mais gerou gastos (R\$ 712 mil), tendo realizado 72 exames, dois internamentos e com 80 anos de idade.

Indivíduo	Idade	Cap1	Cap3	Cap6	Cap10	Cap14	Numero de Exames	Numero de Internamentos	Gasto Excedente
43	47	Sim	Sim	Sim	Sim	Não	32	1	R\$ 787.400
65	59	Não	Não	Sim	Sim	Sim	1019	1	R\$ 487.200
106	69	Não	Sim	Sim	Sim	Não	1264	1	R\$ 127.300
141	78	Sim	Não	Sim	Sim	Não	153	3	R\$ 204.900
149	80	Sim	Sim	Não	Sim	Sim	72	2	R\$ 712.200
172	86	Não	Sim	Não	Sim	Não	1501	1	R\$ 57.300

QUADRO 3 – CARACTERÍSTICAS DOS PONTOS DE ALAVANCA ENCONTRADOS NO MODELO COM DISTRIBUIÇÃO GAMA

FONTE: O autor (2013)

A Figura 5 mostra que os resíduos componentes do desvio padronizados apresentam um ajuste satisfatório à distribuição Normal, configurando uma tendência linear no gráfico quantil-quantil e não ultrapassando os limites do envelope simulado.

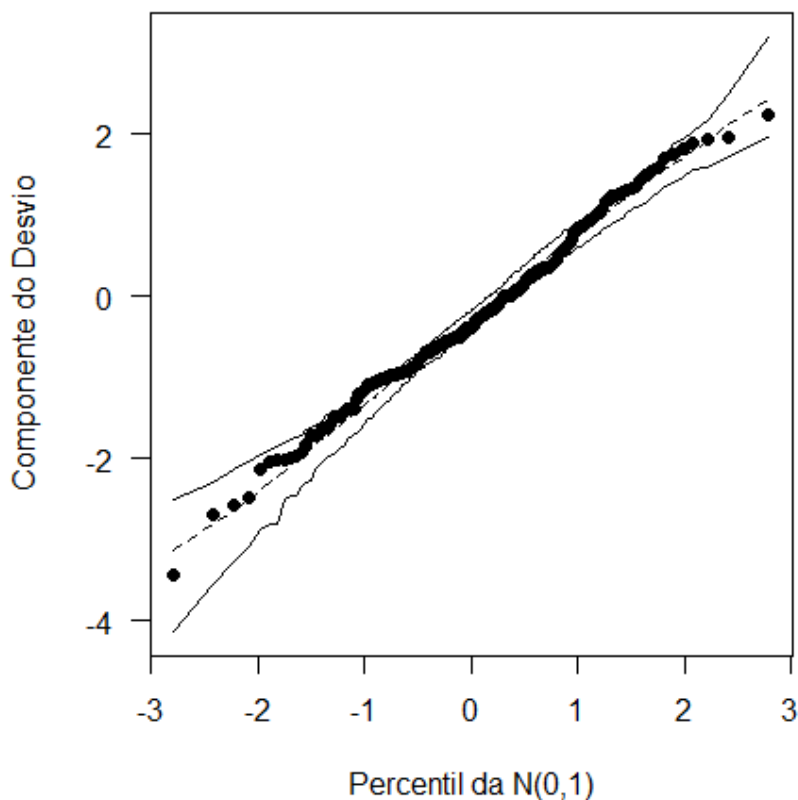


FIGURA 5: Gráfico Normal de probabilidades para os resíduos componentes da deviance padronizados referente ao modelo com distribuição Gama

FONTE: O autor (2013)

Tendo diagnosticado o modelo e verificado o bom ajuste proporcionado, com base no modelo ajustado e nas estimativas apresentadas na Tabela 4, podemos extrair as seguintes interpretações para as variáveis significativas ao nível de 10%:

- O valor excedido diminui conforme a idade do usuário. Aproximadamente 1% ($e^{-0,01} = 0,99$) é a redução estimada para o valor excedido ao considerar um ano a menos de idade;
- O valor excedido à franquia é maior para pacientes que utilizaram procedimentos referentes ao Capítulo 3, que trata de doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários. Estima-se um gasto excedente médio 36,89% maior ($e^{0,314} = 1,3689$) para tais pacientes em relação àqueles que não apresentam o referido capítulo ($p = 0,092$);
- Indivíduos que não apresentam o Capítulo 14, referente a doenças do aparelho geniturinário, têm um gasto excedente médio 28,17% maior do que aqueles que realizaram procedimentos relacionados ao Capítulo 14 ($p = 0,019$);
- Em relação à quantidade de exames realizados é estimado que, a cada 100 exames realizados, o valor excedente à franquia aumenta, em média, 10,52% ($p = 0,08$) o gasto excedente.
- Para a variável número de Internamentos, a cada unidade acrescida, o aumento esperado é de 64% ($e^{0,495} = 1,64$) no valor excedido à franquia.

Em todas as conclusões, os resultados são válidos fixando-se os valores das demais variáveis.

3.4 Predição com base no modelo ajustado com distribuição Gama.

A fim de prever situações que podem vir a ocorrer, a Tabela 6 apresenta uma simulação de casos de usuários que estão presentes na base de dados e de alguns casos extremos, que podem vir a ocorrer, a fim de estimar o gasto excedente pelo modelo de distribuição Gama e compará-lo com o valor observado, quando disponível.

TABELA 6 – SIMULAÇÕES DE GASTO EXCEDENTE A FRANQUIA COM BASE EM PERFIS EXISTENTES E CASOS EXTREMOS

Situação	Idd	Cap1	Cap3	Cap6	Cap10	Cap14	Exames	Int	Gasto Estimado (mil)	Gasto Real (mil)
1	78	Não	Não	Não	Não	Sim	30	1	R\$ 50,90	-
2	80	Sim	Não	Sim	Não	Sim	320	2	R\$ 174,86	R\$ 117,30
3	84	Não	Não	Não	Não	Sim	42	1	R\$ 49,70	R\$ 48,40
4	29	Não	Não	Não	Não	Sim	75	1	R\$ 71,40	R\$ 60,20
5	32	Não	Sim	Não	Sim	Não	312	2	R\$ 345,16	R\$ 36,40
6	90	Não	Não	Não	Sim	Sim	284	1	R\$ 75,79	R\$ 70,20
7	49	Sim	Sim	Sim	Sim	Não	144	1	R\$ 254,70	-

FONTE: O autor (2013)

Na Tabela 6 tem-se que na situação 1, foi simulado o perfil de um indivíduo que tem 78 anos de idade, utilizou procedimentos referentes ao Capítulo 14, realizou 30 exames e teve um internamento, o modelo nos dá a estimativa de valor excedente à franquia de R\$ 50,9 mil reais.

As situações 3 e 4 da Tabela 6 correspondem a indivíduos que foram utilizados nas análises deste trabalho, tendo-se, além do gasto estimado, o valor do gasto real. O indivíduo da situação 3 apresentou uma diferença de apenas R\$ 1,3 mil entre o gasto estimado e o gasto real, enquanto que o indivíduo da situação 4 apresentou uma diferença um pouco maior, de R\$11,2 mil.

Na situação 7 foi simulado um caso em que o indivíduo tem 49 anos, ausência do Capítulo 14, e utilização de procedimentos referentes aos demais capítulos, bem como uma quantidade alta de exames. Seu gasto estimado foi de R\$ 254,7 mil, bem maior se comparado à média do valor gasto na base em estudo (R\$ 104 mil).

3.5 Processo de ajuste do modelo linear generalizado com resposta Normal Inversa

A fim de tentar encontrar outra distribuição à qual os dados pudessem se adequar, e para efeito de comparação de modelos, tentou-se ajustar um modelo linear generalizado com distribuição Normal Inversa.

Logarítmica	Identidade	Inversa quadrática	Inversa
Grau de dependência	Idade	Internamento	
Idade	Período		
Período	Consulta eletiva		
Cidade	Internamento		
Capítulo 2			
Capítulo 9			
Consulta eletiva			
Consulta emergência			
Exames			
Internamento			

QUADRO 4 – FUNÇÕES DE LIGAÇÃO COM VARIÁVEIS QUE APRESENTARAM PROBLEMAS DE AJUSTE PARA A DISTRIBUIÇÃO NORMAL INVERSA

FONTE: O autor (2013)

Inicialmente ajustou-se um modelo para cada covariável explicativa, com quatro funções de ligação (logarítmica, identidade, inversa e inversa quadrática). O Quadro 4 mostra as covariáveis que produziram erro no ajuste do modelo, tendo a função de ligação logarítmica o maior número de variáveis que não possibilitaram o ajuste do modelo.

Posteriormente, analisou-se o efeito da inclusão de cada uma das covariáveis separadamente, exceto as covariáveis apresentadas no Quadro 4 para as quatro funções de ligação, utilizando o teste de razão de verossimilhança como critério para inclusão. Foram mantidas no processo de seleção somente as covariáveis significativas ao nível de 20%, conforme apresentado na Tabela 7.

Na sequência da modelagem dos dados ocorreram problemas referentes à convergência do algoritmo de estimação, tendo-se conseguido ajustar modelos apenas com número reduzidos de variáveis, não se verificando resultados satisfatórios. Decidiu-se então pela utilização do modelo com distribuição Gama, o qual mostrou um ajuste satisfatório.

TABELA 7 - VARIÁVEIS SIGNIFICATIVAS A 20% VIA TRV

Função de ligação	F	P-Valor
Logarítmica		
CAPÍTULO3	3,273	0,07206
CAPÍTULO6	2,676	0,10357
Inversa		
PERÍODO	1,936	0,16573
CAPÍTULO3	3,273	0,07206
CAPÍTULO6	2,676	0,10357
CAPÍTULO10	2,095	0,14952
INTERNAMENTO	4,846	0,02895
Identidade		
CAPÍTULO3	3,273	0,07206
CAPÍTULO6	2,676	0,10357
CAPÍTULO10	2,095	0,14952
Inversa quadrática		
PERÍODO	1,903	0,16938
CAPÍTULO3	3,273	0,07206
CAPÍTULO6	2,676	0,10357
CAPÍTULO10	2,095	0,14952

FONTE: O autor (2013)

4 CONCLUSÃO

Neste presente trabalho foi possível identificar, analisar e discutir os fatores associados ao valor excedido à franquia de resseguro, que foram gerados por usuários de uma operadora de planos de saúde do Estado do Paraná, com o apoio de técnicas de modelos lineares generalizados.

Num primeiro momento, ajustou-se um modelo de regressão com distribuição Normal. Porém, como este não se mostrou apropriado, e como se julgou não ser adequado tentar a transformação da variável resposta, partiu-se para o ajuste de outros modelos, como a distribuição Gama e Normal Inversa. Este último, porém, também não se mostrou adequado, visto que no processo de ajuste o modelo apresentou problemas de convergência, o que inviabilizou a obtenção de um ajuste satisfatório. Uma vez que os dados mostraram um ajuste adequado à distribuição Gama, a sequência do trabalho se deu com base na mesma.

Identificou-se que variáveis relacionadas ao uso do plano, juntamente com os CIDs, estão mais correlacionadas ao valor excedido do que outras variáveis explicativas como, por exemplo, tipo de plano contratado, grau de dependência e região onde o usuário reside.

Dentre as variáveis presentes no modelo ajustado, notou-se que o número de internamentos ($p = 0,004$) foi uma das variáveis mais importantes para o modelo, o que na prática se justifica, visto que os custos associados a ele são muito elevados. A variável Idade ($p = 0,051$) chama a atenção, pois apresentou um coeficiente negativo, indicando que quanto mais novo o usuário, menor será o seu gasto excedido.

Um dos problemas que pode ter prejudicado o ajuste do modelo é a considerável quantidade de ausência da informação do capítulo do CID, uma vez que não é mais obrigatório por parte do médico especificar o CID no prontuário do paciente. Um dos meios para se tentar identificar doenças e procedimentos e vinculá-los aos CIDs é utilizando a *expertise* de profissionais da saúde, analisando os procedimentos utilizados e, posteriormente, aplicando-se mineração de dados.

Para estudos futuros, seria interessante analisar o tempo que o usuário levou até exceder o valor da franquia contratada e seus respectivos fatores associados, a fim de que sejam dados subsídios para a operadora de planos de saúde na tomada de ações com o intuito de minimizar os altos valores gastos com este usuário.

REFERÊNCIAS

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR. **Glossário temático saúde suplementar**. Brasília, 2009. p. 70. Disponível em: <http://www.ans.gov.br/portal/upload/biblioteca/glossario_tematico_ss.pdf>. Acesso em 08/09/2013.

CASELLA, G.; BERGER, R. L. **Inferência estatística**. 1.ed. Stamford: Cengage, 2010.

CONSELHO FEDERAL DE MEDICINA (CFM). **Informações sobre a saúde suplementar no Brasil**. Disponível em: <http://portal.cfm.org.br/index.php?option=com_content&view=article&id=22200:informacoes-sobre-a-saude-suplementar-no-brasil&catid=3>. Acesso em 08/09/2013.

D'ALAMA, Luna. **Custos dos planos individuais de saúde têm maior alta desde 2007**, Fevereiro. 2013. Disponível em < <http://g1.globo.com/economia/noticia/2013/02/custos-dos-planos-individuais-de-saude-tem-maior-alta-desde-2007.html>>. Acesso em 21/09/2013.

DEMÉTRIO, C. G. B; CORDEIRO, G. M. **Minicurso para o 12º SEAGO e a 52ª Reunião Anual da RBRAS UFSC, Santa Maria, RS**. Disponível em: <http://www.lce.esalq.usp.br/clarice/livroSeagro.pdf>. Acesso em 23/11/2013.

INSTITUTO DE ESTUDOS DE SAÚDE SUPLEMENTAR (IESS). **Série Histórica Variação dos Custos Médico-Hospitalares (%)**. São Paulo, SP. Disponível em: <<http://www.iess.org.br/VCMHIESSdbdez12.pdf>>. Acesso em 22/09/2013.

MICROSOFT. **Microsoft Excel**. Redmond, Washington: Microsoft, 2013. Computer Software.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **CID-10 Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde**. São Paulo: Universidade de São Paulo; 1997. vol.1.

PAULA, G, A. **Modelos de Regressão com apoio computacional**. Disponível em: <http://www.ime.usp.br/~giapaula/texto_2013.pdf>. Acesso em: 21/09/2013.

R DEVELOPMENT CORE TEAM. 2013. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Disponível em: <<http://www.R-project.org/>>.

ZUCCHI, P.; NERO, C. D.; MALI, A. M. **Gastos em saúde: os fatores que agem na demanda e na oferta dos serviços de saúde**. *Saúde e Sociedade*, São Paulo, p 127- 150,2010. Disponível em: <<http://www.ppge.ufrgs.br/giacomo/arquivos/eco02072/zucchi-delnero-malik-2000.pdf>>. Acesso em 19/10/2013.