

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
Graduação em Ciências Atuariais

Webster Alves Ferreira Lopes

IDENTIFICAÇÃO DE ALTOS RISCOS EM SAÚDE SUPLEMENTAR

Belo Horizonte
2022

Webster Alves Ferreira Lopes

IDENTIFICAÇÃO DE ALTOS RISCOS EM SAÚDE SUPLEMENTAR

Monografia apresentada ao departamento de Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Bacharel em Ciências Atuariais.

Orientador: Vinícius Diniz Mayrink

Coorientador(a): Jussiane Gonçalves da Silva

Belo Horizonte

2022

TERMO DE ACORDO DO ORIENTADOR
ENTREGA DE MONOGRAFIA – VERSÃO FINAL

Eu, Vinicius Diniz Mayrink, orientador de monografia do aluno: Webster Alves Ferreira Lopes, matrícula 2017016084, estou de acordo com a versão final da monografia de conclusão de curso intitulada "**Identificação de Altos Riscos em Saúde Suplementar**", entregue em 22 de fevereiro de 2022.



Assinatura

Belo Horizonte, 22 de fevereiro de 2022.

RESUMO

Geralmente, o setor de saúde suplementar é tendencioso a maiores frequências de utilização dos serviços prestados pelas operadoras desse segmento e conseqüentemente custos advindos desses usos. Além disso, os dados dessas demonstram pequena parte dos beneficiários responsáveis por maior parcela dos custos assistenciais. Sendo um ponto importante para gestão nesse setor: é a identificação desses beneficiários configurados como alto custo, podendo determiná-lo sendo um associado de alto risco. Com isso, há o interesse de estimar a probabilidade de um determinado beneficiário ser um risco em potencial para a operadora, de acordo com as características e utilização do plano de saúde.

Dado esse panorama, de impacto atuarial e sem contar a existência da dificuldade de se estabelecer qual é o custo final de um beneficiário. O objetivo desse trabalho é estimar a probabilidade de um beneficiário ser um alto risco para a operadora, através de um modelo regressão. Tomando a sinistralidade como base para criar a resposta deste método e os fatores de riscos disponíveis as covariáveis. Para este fim, será realizado um estudo de caso com os dados fornecidos por uma operadora de plano de saúde. A base: para os devidos fins acadêmicos – pedagógicos e seguindo as devidas providências acerca de Lei de Geral de Proteção de Dados (LGPD), sofreu modificações para não haver rastreabilidade.

Após a modelagem estatística, foram encontrados dois modelos capazes de: avaliarem os impactos dos fatores de riscos para a estimação da probabilidade de um beneficiário se tornar alto risco. Sendo um modelo para um plano de saúde ambulatorial e o outro para um ambulatorial+hospitalar. Ademais, a escolha do ponto de corte pela sinistralidade demonstrou ser uma alternativa para lidar com as implicações acarretadas pelo comportamento do custo assistencial. Conclui-se, para os devidos fins, que essa pesquisa alcançou seu objetivo, segundo o apresentado pela literatura, pormenor que esteja a predição realizada para esses beneficiários de alto risco, ela é válida por abordar um evento de difícil previsão.

SUMÁRIO

| | |
|--|----|
| 1. INTRODUÇÃO..... | 6 |
| 1.1. Justificativa | 6 |
| 1.2. Objetivo | 7 |
| 2. REVISÃO DA LITERATURA | 7 |
| 3. METODOLOGIA DE PESQUISA..... | 9 |
| 3.1. Metodologia..... | 9 |
| 3.1.1. <i>Regressão Logística</i> | 10 |
| 3.1.2. <i>Estimação dos Parâmetros</i> | 11 |
| 3.2. Dados..... | 13 |
| 3.2.1. <i>Descrição dos dados</i> | 14 |
| 4. RESULTADOS | 17 |
| 4.1. Ambulatorial..... | 17 |
| 4.2. Ambulatorial + Hospitalar..... | 24 |
| 5. CONCLUSÃO..... | 30 |
| REFERÊNCIAS..... | 35 |

1. INTRODUÇÃO

Efetivamente, são reconhecidas as melhorias significativas das condições de saúde ao longo do tempo, o declínio nos níveis de fecundidade e da mortalidade impactam sobre a estrutura etária da população. A redução da proporção de jovens, concomitante, a expansão dos idosos na população geram consequências de várias magnitudes: demográficas, socioeconômicas entre outras. Das mais importantes diz a respeito às necessidades de saúde.

A literatura apresenta a transição epidemiológica como transição de saúde (Schramm 2015), de forma que anexa elementos conceptivos e comportamentais da sociedade no que tange aos aspectos de saúde na população humana. Segundo Schramm (2015), a transição da saúde pode ser dividida em dois elementos principais: primeiramente, encontra-se a transição das condições de saúde (referindo-se às mudanças na frequência, magnitude e distribuição das condições de saúde, expressas através das mortes, doenças e incapacidades) e, em segundo, a resposta social organizada a estas condições que se instrumenta por meio dos sistemas de atenção à saúde (transição da atenção sanitária), determinada em grande medida pelo desenvolvimento social, econômico e tecnológico mais amplo.

Aquino 2017, descreve esse mercado como complexo, devido à conjuntura que está inserido, uma vez que fatores internos e externos obrigam os “players” analisarem os diferentes riscos envolvidos e propor critérios os quais conduzam para uma situação equilibrada. Dentre os principais aspectos que atraem a atenção dos gestores em saúde suplementar: a avaliação dos custos assistenciais, em especial o aumento dessas despesas e a identificação dos associados que produzem custos assistenciais mais elevados. Portanto, gerenciar os ônus assistenciais em plano de saúde é uma prática que pode restringir as despesas das operadoras, além de proporcionar melhorias de cuidados para com os pacientes e otimizar os recursos disponíveis.

1.1. Justificativa

Em um plano de saúde, geralmente, é tendencioso ocorrer volumosa frequência de assistência/utilização. Contudo, não necessariamente, um beneficiário de alto risco se resume a uma frequência demasiada, uma vez que, o custo do associado para operadora é composto pela utilização e do valor, em média, do procedimento prestado. Nessa sequência, vale apresentar uma situação corriqueira, os dados das operadoras demonstram sobre a distribuição dos custos assistenciais uma pequena parcela dos beneficiários (geralmente 5%) são responsáveis por uma grande fração das despesas.

Fato encontrado em diversos trabalhos como o de Aquino (2017), em seus resultados descritivos apontaram que menos de 1% da massa segurada acarretava cerca de 46,1% de todos os custos assistenciais. Outro estudo realizado em 2017 pelo Instituto de Estudos de Saúde Suplementar (IESS) para uma carteira de planos da modalidade de autogestão encontrou que grupo de alto custo (5% que mais gastam), a despesa assistencial por pessoa varia do mínimo de R\$ 20.815 no ano ao máximo de R\$ 5,7 milhões no ano. De acordo à conjuntura apresentada referente a realidade das operadoras de planos de saúde, torna-se necessário uma forma para identificar esses possíveis riscos.

1.2. Objetivo

Ressalto que o objetivo não é identificar o alto custo, a princípio, não é uma tarefa simples defini-lo, pois alguns tratamentos- procedimentos- exames entre outros podem ser de valores elevados de acordo com a sua complexidade e material necessário para realização, acarretando altas sinistralidades. Determinar esse dispêndio final gera muito trabalho e a literatura apresenta muitos critérios diferentes para esse fim, por exemplo, quem serão os 5% clientes mais caros, ou os 10% mais caros, ou aqueles que vão gerar alto custo, mas por enquanto não são. Esses cortes costumam ser pequenos, e se, esse grupo for muito ínfimo, atrapalha a análise.

Dados esses panoramas: mercadológico, de impacto atuarial e da dificuldade para lidar com o alto custo o objetivo desse trabalho: é estimar a probabilidade de ocorrência desses eventos de alto risco, via regressão logística e tomando como “*base-line*” a sinistralidade e nos fatores de risco disponíveis. Além de ser um, ponto importante para gestão um conhecimento prévio de beneficiários que representam alto risco e estimar a sua probabilidade de um ser um potencial risco para a operadora.

2. REVISÃO DA LITERATURA

O tema em questão é alvo de outras pesquisas, em especial os pesquisadores estrangeiros deferem muito desse linha de pesquisa. Muito deve-se debruçar sobre os escritos atuariais de países como Estados Unidos, Inglaterra, Canadá entre outros, pois esses possuem um pouco mais da ciência atuarial desenvolvida. Mas, na literatura brasileira encontram-se algumas linhas de pesquisa dentro dessa temática.

Em 1986, Lavange, Iannacchione e Garfinkel propuseram um modelo de regressão logística como método preditivo de usuários de alto custo. A pesquisa objetivou o desenvolvimento de um mecanismo que permita prévio conhecimento da massa segurada que incorrerá elevados custos assistenciais, e, consecutivamente, estabelecer políticas de controle desses custos. Os autores

aplicaram duas regressões logísticas - uma para os indivíduos entre 17 e 64 anos e outra para os com 65 anos ou mais. No que tange ao Brasil, a Agência Nacional de Saúde (ANS), através da Resolução Normativa nº 63 de 22 de dez de 2003, estabeleceu 10 faixas etárias para o reajuste de plano: 0-18, 19-23, 24-28, ..., 59 anos ou mais; usualmente alguns estudos de mercado adotaram essas faixas.

Bierman et al. (1999) em seus estudos utilizaram apenas de estatísticas descritivas. Com dados advindos de um questionário aplicado no ano de 1992, em uma amostra de 8.775 participantes de seguros privados de assistência à saúde nos EUA. Todos os indivíduos que possuíam mais de 64 anos foram avaliados da seguinte maneira: “No geral, a que com às outras pessoas da sua idade, você diria a sua saúde é: excelente, muito boa, boa, regular ou má?” (BIERMAN et al., 1999, p. 57).

Daquele trabalho 18% consideraram a sua saúde excelente, 56% responderam que possuíam uma boa ou uma muito boa condição de saúde, 17% afirmaram a sua saúde regular e 7% relataram possuir más condições de saúde. Dessa maneira, Bierman et al. (1999), relacionaram os custos assistenciais dos beneficiários no ano de 1993 e as respostas obtidas em 1992, expondo associação entre o questionário aplicado e o custo assistencial do ano subsequente, o qual os indivíduos responderam em 1992 que possuíam más condições de saúde apresentaram custo assistencial médio de US\$ 8.190 em 1993, contrastando com os segurados responderam que possuíam excelentes condições de saúde e obtiveram um custo assistencial médio de US\$ 1.627.

A predição realizada por Dove, Duncan e Robb (2003), utilizou uma base de dados com informações assistenciais, demográficas e de diagnóstico médico de 209 mil usuários pertencentes às organizações de gerenciamento em saúde nos Estados Unidos da América. O trabalho buscou à mensuração do nível de risco dos indivíduos para o período subsequente ao utilizado para a análise e tendo como “*base-line*” alto custo \geq \$2000,00 dólares. Na apuração do resultado, obteve-se uma área sob a curva ROC de 0,73, não sendo apresentado o coeficiente de determinação do modelo, mas concluem, no entanto, a predição realizada foi eficaz para o devido propósito.

Pelos achados de Aquino (2017), relata estatísticas descritivas que apontaram menos de 1% dos beneficiários da operadora de plano de saúde consome 46,1% de todos os custos assistenciais, permitindo uma primeira inferência de identificação prévia desses clientes que incorrerão em altos custos no ano seguinte. Ainda expôs o gerenciamento de clientes preditos como alto custo requer da operadora um investimento (gasto), os gestores do plano de saúde devem ter bastante cautela, pois, em seu trabalho, apesar do modelo ter um resultado de 96,6% de acurácia, as classificações não foram muito assertivas para os indivíduos de alto risco.

Diante desses relatos é perceptível a cautela considerada com a devida predição proposta. Devido à complexidade para lidar com o alto risco e além dos enclaves expostos ao mercado de saúde suplementar. Sendo mais frequentes, aqueles quanto à incerteza sobre o montante final dos custos assistenciais.

3. METODOLOGIA DE PESQUISA

3.1. Metodologia

Para a metodologia pretendida, um modelo linear generalizado (MLG), no caso a regressão logística. Esse trabalho busca avaliar a relação entre beneficiário e sua ocorrência de alto risco. Indo ao encontro do uso de um MLG, geralmente utiliza-se para avaliar e quantificar a relação entre variáveis: sendo uma resposta, denominada simplesmente de variável, e as explicativas (covariáveis ou regressores) e compostos por três elementos.

O primeiro componente refere a uma distribuição de probabilidade pertencente à família exponencial (**componente aleatório**) a qual engloba as distribuições como – normal, gama, normal inversa, entre outras para dados contínuos e quanto para dados de contagem – Poisson e binomial negativa e para dados binários a Bernoulli/Binomial. Uma distribuição de probabilidade é dita pertencer à família exponencial canônica se pode ser escrita da seguinte maneira:

$$f(\mathbf{y}_1, \dots, \mathbf{y}_n; \boldsymbol{\theta}) = \prod_{i=1}^n \exp\{\mathbf{y}_i \mathbf{b}(\boldsymbol{\theta}_i) + c(\boldsymbol{\theta}_i) + d(\mathbf{y}_i)\} = \exp\left\{ \sum_{i=1}^n \mathbf{y}_i \mathbf{b}(\boldsymbol{\theta}_i) + \sum_{i=1}^n c(\boldsymbol{\theta}_i) + \sum_{i=1}^n d(\mathbf{y}_i) \right\}$$

Sendo:

- $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)^T$: Diz a respeito do parâmetro de interesse da densidade de probabilidade de cada $f(\mathbf{y}_i, \dots, \mathbf{y}_n)$, atente-se ao índice, pois indica que o valor de $\boldsymbol{\theta}$ pode ser diferente para cada $\mathbf{y}_i, \dots, \mathbf{y}_n$;
- $f(\mathbf{y}_1, \dots, \mathbf{y}_n; \boldsymbol{\theta})$: Refere-se à função de densidade de probabilidade escrita na formulação da família exponencial canônica;
- $b(\boldsymbol{\theta}_i)$, $c(\boldsymbol{\theta}_i)$ e $d(\mathbf{y}_i)$ são funções que assumem valores no R^+ ;
- Se houver outros parâmetros serão denominados de ruídos (parâmetro de ruído).

O segundo, preditor linear (**componente sistemático**), é uma combinação linear dos regressores usando os coeficientes, cujo papel é engajar no modelo a informação das covariáveis. Além

disso, esse elemento é relacionado com a média ou parâmetro da distribuição de probabilidade da variável de interesse, por meio de uma função de ligação.

Por fim, o terceiro elemento, a **função de ligação**, responsável pela conexão entre os componentes: aleatório e sistemático. Tal ligação deve ser monótona e diferenciável no universo de análise. Condicionamente, é pertinente que a imagem dessa função esteja restrita ao conjunto de valores em que a média ou parâmetro estejam definidos.

3.1.1. Regressão Logística

Em diversas modelagens, há o interesse de análise do comportamento de uma variável resposta binária Y_i com relação a um conjunto de covariáveis X_i . Isto é, a variável resposta assume o valor 1 quando ocorre o evento de interesse, denominado como “sucesso”, para este trabalho “alto risco”, e assume o valor 0 quando não ocorre o evento de interesse, designado “fracasso” – “baixo risco”. Nesta conjuntura, podemos aplicar o modelo regressão logística para estimar as probabilidades de ocorrência do evento de interesse com base nas covariáveis.

Sendo assim a variável resposta $Y_i \sim \text{Bernoulli}(\theta)$, logo:

$$Y_i = \begin{cases} 0, & \text{se ocorrer o fracasso} \\ 1, & \text{se ocorrer o sucesso} \end{cases}$$

Sua função de densidade é dada por:

$$f(y; \theta) = \theta^y (1 - \theta)^{(1-y)}, y = (0, 1)$$

Se θ_i é a probabilidade de sucesso $P(Y_i = 1) = \theta_i$, concomitantemente:

$$E(Y_i) = \theta_i \text{ e } \text{Var}(Y_i) = \theta_i(1 - \theta_i)$$

Assumindo que cada Y_i são variáveis independentes, a função de densidade conjunta e a sua função de log-verossimilhança são respectivamente:

$$f(y_1, \dots, y_n; \theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{(1-y_i)}$$

$$l(\theta) = \sum_{i=1}^n y_i \log\left(\frac{\theta_i}{1 - \theta_i}\right) + \sum_{i=1}^n \log(1 - \theta_i)$$

A $E(Y_i) = \theta_i$ será modelada por meio de uma função de ligação - “logit”, a qual descreve a relação entre o preditor linear ($g(\theta_i)$) e θ_i , conseqüentemente, a função logística descreve a relação entre o preditor linear e θ_i . Sendo assim segue:

$g(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$ e $\mathbf{x}_i^\top = (\mathbf{1}, x_1, \dots, x_p)$ é o vetor de p covariáveis das i -ésima observação. Dessa forma:

$$g(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \log\left(\frac{\theta_i}{1-\theta_i}\right) \Rightarrow \theta_i = \frac{e^{(\mathbf{x}_i^\top \boldsymbol{\beta})}}{1+e^{(\mathbf{x}_i^\top \boldsymbol{\beta})}} \Rightarrow \theta_i = \frac{1}{1+e^{(-\mathbf{x}_i^\top \boldsymbol{\beta})}}$$

Conhecida como razão de “chance” ou “odds”: $\frac{\theta_i}{1-\theta_i}$, indica a ocorrência do evento de interesse e é mais provável do que a sua não ocorrência. Atente-se ao fato: a função “logit” é o log da “odds”: log-odds. Além disso para todo:

$$g(\theta_i) \in \mathbb{R} \Rightarrow \frac{1}{1+e^{(-\mathbf{x}_i^\top \boldsymbol{\beta})}} \in (0,1).$$

3.1.2. Estimação dos Parâmetros

Para estimar o vetor $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ utilizaremos o método da máxima verossimilhança. Seja \mathbf{x}_i as observações referentes ao i -ésimo indivíduo e $\mathbf{y} = (y_1, \dots, y_n)^\top$. Seja uma função de log-verossimilhança:

$$l(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) = \sum_{i=1}^n y_i g(\mathbf{x}_i) - \log[1 + e^{g(\mathbf{x}_i)}]$$

Derivando a expressão acima em relação aos parâmetros $\beta_j, j = 1, \dots, p$, as estimativas são dadas pela solução simultânea das equações abaixo, em que x_{ij} representa a j -ésima variável do i -ésimo indivíduo.

$$\sum_{i=1}^n [y_i - \theta_i(\mathbf{x}_i)] = 0 \text{ e } \sum_{i=1}^n x_{ij} [y_i - \theta_i(\mathbf{x}_i)] = 0, \quad j = 1, \dots, p.$$

As estimativas dos parâmetros podem ser encontradas a partir de métodos numéricos e denotamos por $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$. Essas estimativas serão obtidas pela função *glm* do software *R*, o qual emprega o método IWLS (*Iterative Weighted Least Squares*) e tem como base o Método Escore de Fisher.

Para interpretação dessas estimativas (coeficientes) considere um modelo com uma covariável:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x_{i1}$$

A razão de chances é dada por:

$$\frac{\theta_i}{1-\theta_i} = \exp(\beta_0 + \beta_1 x_{i1})$$

Se x_{i1} aumenta em d unidades, sendo $d \geq 0$, qual o efeito na razão de chances?

$$\begin{aligned} &= \exp(\beta_0 + \beta_1(x_{i1} + d)) \\ &= \exp(\beta_0 + \beta_1 x_{i1} + d\beta_1) \\ &= \exp(\beta_0 + \beta_1 x_{i1}) * \exp(d\beta_1) \\ &= \frac{\theta_i}{1-\theta_i} * \exp(d\beta_1) \end{aligned}$$

Portanto, se x_{i1} aumenta em d unidades, a razão de chances é multiplicada por $\exp(d\beta_1)$.

O fator $\exp(d\beta_1)$ indica o aumento/redução na razão de chances.

Para um modelo com k regressores:

$$\frac{\theta_i}{1-\theta_i} = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})$$

Se x_{i1} aumenta em d unidades, sendo $d \geq 0$, qual o efeito na razão de chances, mantendo as demais covariáveis constantes?

$$\begin{aligned} &= \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) \\ &= \exp(\beta_0 + \beta_1(x_{i1} + d) + \dots + \beta_k x_{ik}) \\ &= \frac{\theta_i}{1-\theta_i} * \exp(d\beta_1) \end{aligned}$$

O fator $\exp(d\beta_1)$ indica o aumento/redução na razão de chances, associada ao acréscimo de d unidades em x_{i1} , mantendo as demais covariáveis constantes.

Suponha agora que X_{1i} é uma covariável categórica com J categorias; isto é, $X_{1i} = j$ com $j \in \{1, \dots, J\}$. Para introduzir no modelo a informação desta covariável, são necessárias $J-1$ variáveis indicadoras, pois uma categoria é usada como base.

A relação entre θ_i e as demais covariáveis/regressores indicadoras será:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 I_{(X_1=1)} + \dots + \beta_{J-1} I_{(X_{j-1}=J-1)}$$

Se $X_{1i} = J$, a razão chance será:

$$\frac{\theta_i}{1 - \theta_i} = \exp(\beta_0 + \beta_1 0 + \dots + \beta_{J-1} 0) = \exp(\beta_0)$$

Se $X_{1i} \neq J$, a razão chance será:

$$\frac{\theta_i}{1 - \theta_i} = \exp(\beta_0 + \beta_{J-1}) = \exp(\beta_0) * \exp(\beta_{J-1})$$

Verbalmente: o efeito do nível j sobre a razão de chances é o efeito do nível base multiplicado por $\exp(\beta_j)$.

- $\beta_j < 0$, ocorre a diminuição na chance (em relação ao nível base);
- $\beta_j > 0$, ocorre um aumento na chance (em relação ao nível base).

3.2. Dados

Os dados utilizados provêm de uma operadora de saúde suplementar e para os devidos fins acadêmicos – pedagógicos e seguindo as devidas providências acerca de LGPD, a base sofreu modificações para não haver rastreabilidade. Nesta base estão presentes cerca de 57.351 vidas para o ano de 2019 e é elaborada de acordo com o uso do beneficiário. Dessa maneira, esse material é extenso em memória e observações. Além desse número de vidas, a base contém as variáveis: *Id, Data Nascimento, Data Exclusão, Sexo, Abrangência, Segmentação, Acomodação, Fator Moderador, Competência, Dependência, Data Ocorrência, Data Aviso, Data Pagamento, Intercâmbio, Regime, Tipo Serviço Geral, Tipo Receita, Ativos, Expostos, Receita, Custo Total, Quantidade, Quantidade Final*.

Posteriormente, foram agregados os dados de acordo com *Id* ativo-exposto durante o ano, *Tipo de Receita*: mensalidade. As variáveis selecionadas para trabalho foram: *Id, Sexo, Data de Nascimento, Abrangência, Segmentação, Acomodação, Dependência, Receita e Custo*. Outras variáveis não foram utilizadas por não haver modos de realizar validações como o tipo de receita coparticipação, uma vez que não possuía acesso à política de coparticipação praticada pela operadora, anexo a isso removeu o fator moderador. As restantes não foram utilizadas por demanda de tempo disponibilizado para esse tipo de trabalho, pois essas contribuíram significativamente para dimensão da base.

Com a *Data de Nascimento*, calculou-se a idade de cada beneficiário e dividiu por 100, para que não seja penalizada no método de estimação. Por outro lado, não foram criadas as faixas etárias da Resolução Normativa nº 63 de 12/2003. A justificativa é que a inclusão dessas faixas aumentaria o número de coeficientes, e pelo tempo proposto para esse tipo de trabalho, não seria ideal. Ressalto que é conveniente realizar esse estudo, pois alguma idade, seja mais avançada ou não, poderá influenciar a estimativa.

Por conseguinte, obteve-se o percentual de sinistralidade por beneficiário. Ademais, a sinistralidade reflete o quanto de receita é comprometido para custear as despesas assistenciais. Preferivelmente, é esperado que a sinistralidade esteja sempre abaixo de 100%, pois é necessário parte da receita para custear as outras despesas.

A variável resposta, *Alto Risco*, foi criada, adotando o critério de ponto de corte fundamentado na sinistralidade do beneficiário:

$$Alto\ Risco = \begin{cases} 0, & \text{sinistralidade} < 75\% \\ 1, & \text{sinistralidade} \geq 75\% \end{cases}$$

O corte escolhido está de acordo com a sinistralidade de 75% que é habitualmente utilizada no mercado para as carteiras de beneficiários. Em 2019, segundo os dados do painel da ANS, *Prisma Econômico-Financeiro da Saúde Suplementar*, a sinistralidade do setor para as operadoras médico-hospitalares atingiu a média de 82%. Entretanto, quando analisadas trimestralmente, 1º Trim. 81%; 2º Trim. 79%; 3º Trim. 77%; 4º Trim. 86%. Esses valores referem-se a um conjunto de “players”, dessa maneira para apenas um, é plausível a escolha de 75% como ponto de corte.

3.2.1. Descrição dos dados

A seguir serão apresentadas as descritivas desses dados.

| Variável | Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|----------------|----------|------------|--------------|---------------|--------------|-------------------|
| Idade | 0 | 20 | 33 | 34,5 | 48 | 103 |
| Receita | R\$ 0,98 | R\$ 635,50 | R\$ 1.688,30 | R\$ 2.782,17 | R\$ 3.421,39 | R\$ 42.889,20 |
| Custo | R\$ 0,00 | R\$ 50,00 | R\$ 474,00 | R\$ 31.345,00 | R\$ 1.442,00 | R\$ 28.189.062,00 |
| Sinistralidade | 0 | 0,02 | 0,26 | 19,97 | 0,76 | 122476,51 |

Tabela 1: Estatísticas Descritivas Dados

Analisando a Tabela 1 e a Figura 1 nota-se: para idade há poucos valores destoantes e sua média e mediana estão relativamente próximas. Quanto às outras, acontece o oposto, médias e medianas distantes e muitas observações destoantes. Sobre o custo, é esperado esse tipo de comportamento e a presença de gaps, consequentemente a sinistralidade também é afetada. Para a receita, utilizou-se o *Painel de Precificação* da ANS como fonte de consulta para o valor comercial da mensalidade para que seja possível a seleção de valores compatíveis.

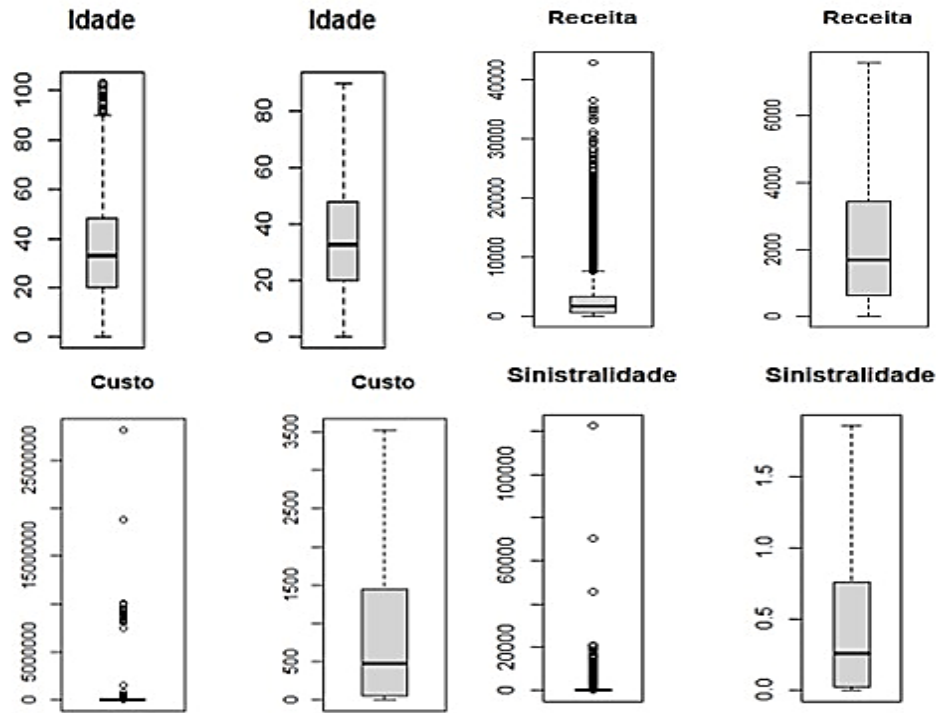


Figura 1: Boxplots Dados para cada variável duas representações

Para as variáveis categóricas: Abrangência, Dependência, Acomodação e Sexo, a Figura 2 apresenta as devidas proporções, respectivamente. A Abrangência refere ao local onde existe cobertura de procedimentos e eventos sendo: grupos de municípios, municipal, estadual e nacional. A Dependência se o beneficiário é o titular ou dependente. Acomodação sobre o local onde o paciente ficará acomodado em caso de internação – ambulatório, enfermaria e apartamento.

A Segmentação é a característica do plano correspondente à composição das coberturas, demonstrada as proporções no Gráfico 1. No banco há: ambulatorial (14.317), hospitalar com obstetrícia (1.014), ambulatorial+hospitalar com obstetrícia (41.552), ambulatorial+hospitalar sem obstetrícia (52) e referência (416).

Proporções Dados 2019

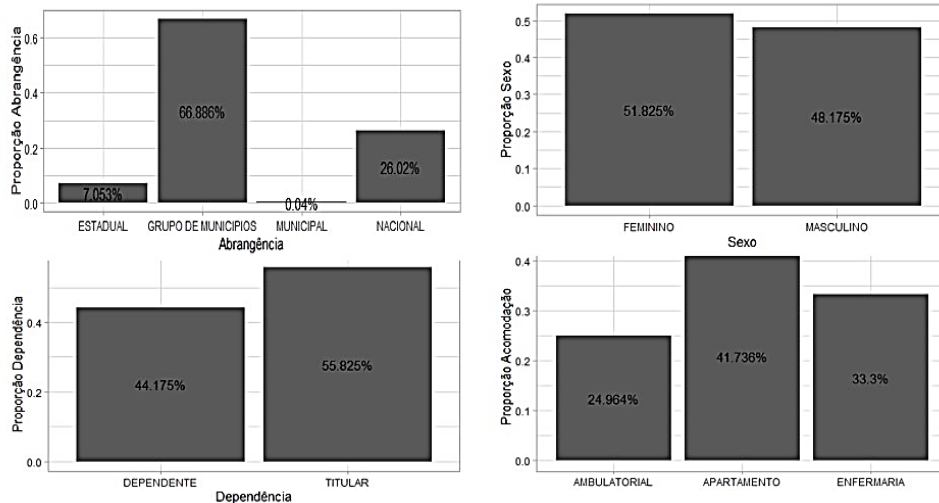


Figura 2: Proporções Dados 2019

A cobertura ambulatorial garante ao beneficiário a prestação de serviços de saúde que compreende consultas médicas em clínicas ou consultórios, exames, tratamentos e demais procedimentos ambulatoriais. A hospitalar com obstetria presta serviços em regime de internação hospitalar e está incluída a atenção ao parto. A composição ambulatorial+hospitalar é a junção de ambulatorial e hospitalar, podendo ser com ou sem obstetria. Quanto ao segmento referência, foi instituído pela Lei nº 9.656/98 englobando assistência médico-ambulatorial e hospitalar com obstetria e acomodação em enfermaria. Além disso sua cobertura mínima também foi estabelecida por essa lei, devendo o atendimento de urgência e emergência ser integral após as 24 horas da sua contratação.

Proporção Segmentação

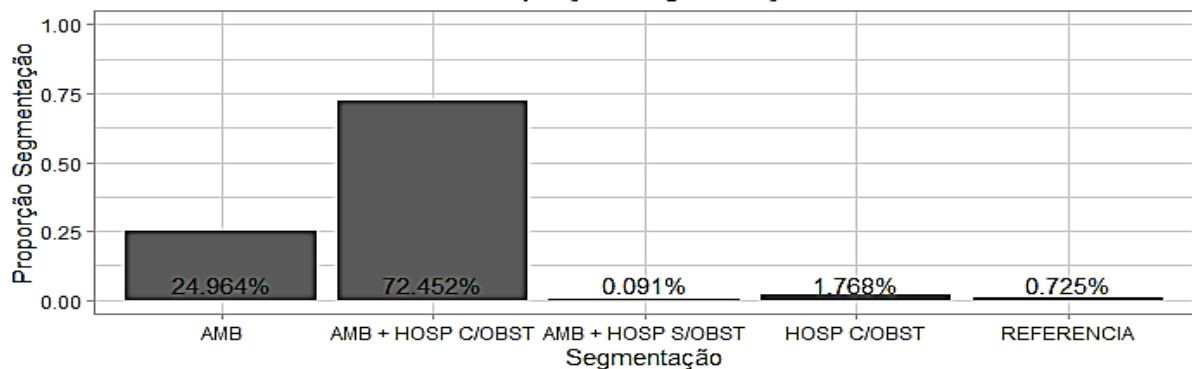


Gráfico 1: Proporções Segmentação

Devido ao tipo de cobertura ser diferente em cada desenho de plano, os dados foram seccionados de acordo com a segmentação. Alguns níveis, por demais, não são representativos quanto ao tamanho amostral, isso sem inspecionar as discrepâncias em cada segmento. Dado essa perspectiva,

é encontrada uma limitação, por isso foi decidido trabalhar com as segmentações ambulatorial, hospitalar com obstetria e ambulatorial+hospitalar com e sem obstetria.

Em todas as partes, optou-se por alterar a variável abrangência, conforme abaixo. Isso pelo pressuposto de se o beneficiário tem área de cobertura em todo país, possivelmente, ele pode acarretar maior risco, e, diminuir o número de covariáveis no modelo.

$$Abrangência = \begin{cases} 0, & \neq nacional \\ 1, & = nacional \end{cases}$$

Sobre a receita e o custo, procedeu-se destas formas - a receita a partir da consulta do painel de precificação, selecionou-se a segmentação e contratação averiguou o valor mínimo e máximo que um beneficiário pagou durante o ano e considerando o reajuste de mensalidade. Enquanto ao custo, selecionou-se os valores maiores que 0, se porventura, o custo é igual a 0, a sinistralidade também é, e assim o beneficiário não impactaria para a pretensão dessa análise.

4. RESULTADOS

Esta seção será apresentada de acordo com a divisão realizada e mostrará algumas estatísticas descritivas e os modelos obtidos para cada segmento, mas para o nível hospitalar será exposto no apêndice, por causa de problemas como verossimilhança monótona, muitas observações destoantes e ter resultado em um banco com poucas observações após a tratativa.

4.1. Ambulatorial

Nesta partição, a covariável *Acomodação* não é exibida nas descritivas e no modelo, pois no plano ambulatorial não são inclusas internações, fazendo com que essa variável categórica tenha somente um nível. Caso fosse incluída no modelo o algoritmo utilizado não converge. Sendo assim, as outras covariáveis serão devidamente demonstradas.

| Variável | Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|----------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Idade | 0 | 32 | 45 | 43 | 55 | 96 |
| Receita | R\$ 1.171,00 | R\$ 1.378,00 | R\$ 1.657,00 | R\$ 2.003,00 | R\$ 2.129,00 | R\$ 7.783,00 |
| Custo | R\$ 2,83 | R\$ 380,45 | R\$ 884,76 | R\$ 1.189,61 | R\$ 1.680,70 | R\$ 9.984,67 |
| Sinistralidade | 0,0015 | 0,2116 | 0,4867 | 0,6368 | 0,8982 | 2,8391 |

Tabela 2: Estatísticas Descritivas- Ambulatorial

Pela Tabela 2, depara-se com dados bem mais comportados para a aplicação do modelo. Quanto à Figura 3, a recodificação da abrangência penalizou quantitativamente o nível nacional, já era

conhecido que havia menos planos nacionais na base. O intuito dessa codificação é captar o pressuposto abordado anteriormente.

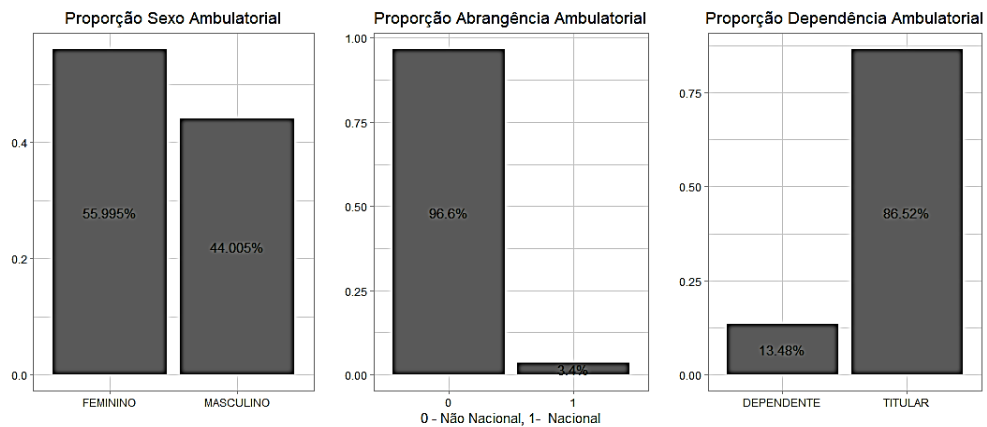


Figura 3: Proporções Variáveis Categóricas -Ambulatorial

As relações entre a variável resposta, *Alto Risco*, com as covariáveis será realizada de acordo com a *Sinistralidade*, uma vez que esta foi utilizada para a criação da resposta.

A correlação entre a sinistralidade e idade resultou em um valor negativo de $-0,0758$ fato não esperado, pois o comportamento previsto é conforme a idade aumente o risco também aumentasse. Ainda sobre o gráfico de dispersão (Gráfico 2), a relação entre a sinistralidade e a idade, aparentemente, não é linear. Uma maneira de contornar isso seria de ajustar o modelo com as faixas etárias, já que o efeito da sinistralidade parece diferente para crianças, jovens e idosos. Possivelmente, o coeficiente para a idade deverá ter um valor negativo. Aquino (2017), em seus resultados (Tabela 13) também encontrou valores negativos para as estimativas dos coeficientes da idade no modelo de classificação de custos assistências, sendo $-1,293$ e causando impacto na odds de $0,275$ para o classificador baixo custo, para o médio $-0,513$ e impacto de $0,599$.

Consoante à Figura 4, os níveis da dependência estão com comportamento semelhante e expressos ao mesmo patamar de sinistralidade e com medianas relativamente próximas. No tocante ao sexo, o masculino está com mediana mais baixa que o feminino, geralmente, é esperado essa ocorrência. Por fim, a abrangência seguiu o mesmo comportamento semelhante à covariável sexo, pois o não nacional está com a mediana mais deslocada que o nacional, ou seja, está mais susceptível ao risco, mas deve-se levar em consideração que aquele nível corresponde a cerca de 97% da base ambulatorial e, possivelmente, isso tenha colaborado para essa ocorrência.

Ambulatorial: Sinistralidade x Idade

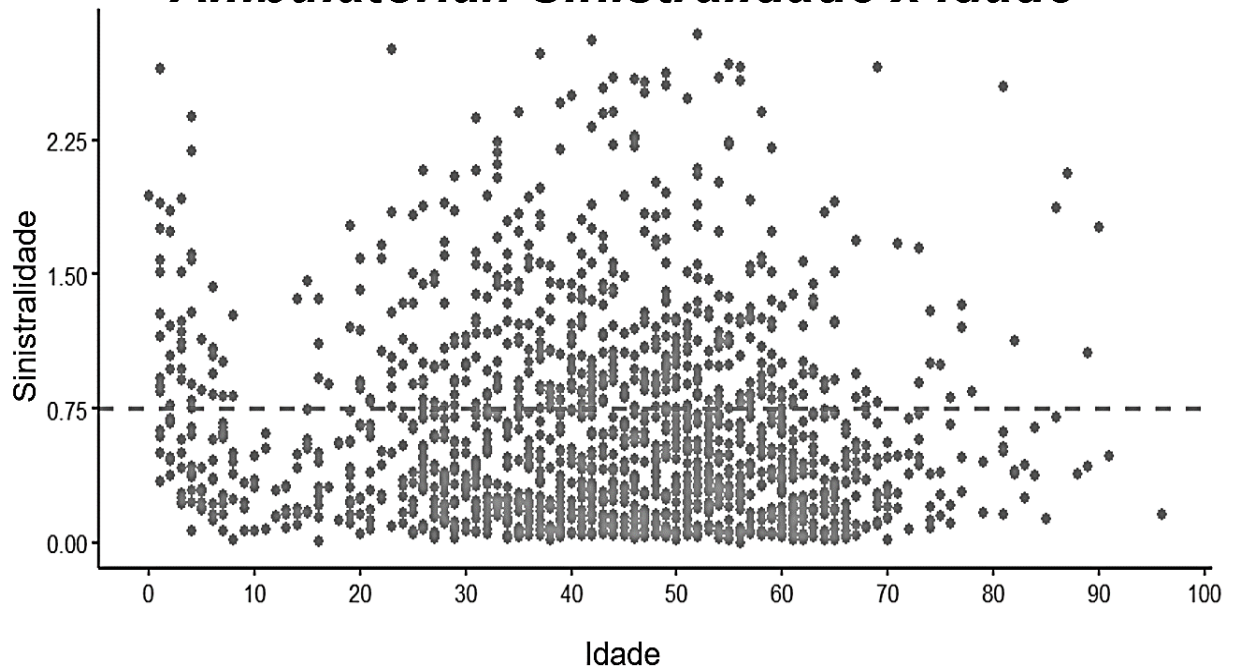


Gráfico 2: Dispersão Sinistralidade x Idade – Ambulatorial

Ambulatorial: Boxplots Sinistralidade

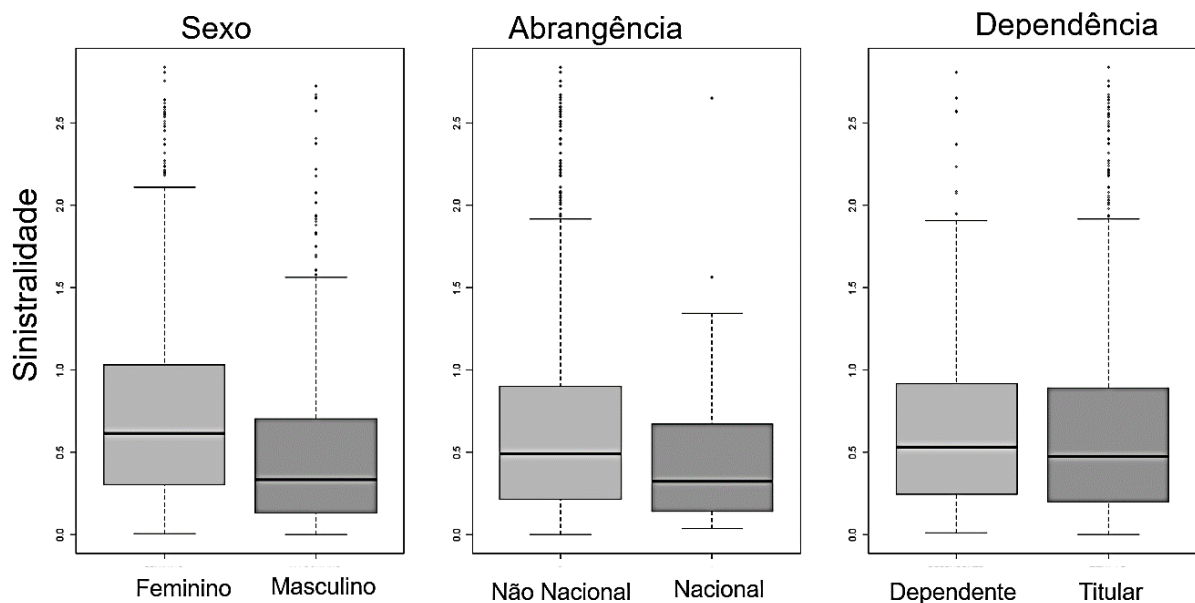


Figura 4: Relação Sinistralidade – Sexo, Abrangência e Dependência - Ambulatorial

Após apresentadas essas relações segue o primeiro modelo testado:

| Modelo Completo Ambulatorial | $\hat{\beta}_i$ | Erro | Teste Z | Valor p |
|------------------------------|-----------------|--------|---------|-------------------|
| Intercepto | 0,0414 | 0,2106 | 0,197 | 0,84414 |
| Sexo-Masculino | -0,8291 | 0,1154 | -7,1860 | $6,66 * 10^{-13}$ |
| Idade | -1,0073 | 0,3124 | -3,2250 | 0,0013 |
| Abrangência | -0,7907 | 0,3764 | -2,1010 | 0,0357 |
| Dependência-Titular | 0,0332 | 0,1633 | 0,2030 | 0,8039 |
| Deviance Residual | 1883,9 | gl | 1538 | |
| Deviance Nula | 1954 | gl | 1542 | |
| AIC | 1893,9 | | | |

Tabela 3: 1º Modelo – Ambulatorial

Os regressores: Sexo, Idade e Abrangência foram significativos para o modelo enquanto o intercepto e a Dependência não, para o nível de significância (α) de 5%, com valores p menores que o α . Quanto aos seus coeficientes ($\hat{\beta}_i$), os sinais desses regressores foram negativos, caso esperado conforme as análises feitas anteriormente devido ao “base-line” (Feminino – 0 anos – Não nacional - Dependente) de cada covariável.

Deste modelo, calculou-se as Distâncias de Cook para identificar pontos influentes. A literatura aborda o critério de a i -ésima observação será influente se a distância calculada for $> 4/n$, e n é o tamanho da amostra. Como o banco ambulatorial após as tratativas, tem o $n = 1543$, realizar esse critério penalizaria muitos pontos. Por isso, analisou-se o boxplot dessas distâncias (Gráfico 3) e por ele decidiu-se que os valores acima de 0,005 serão os influentes já que ocorre um gap em torno desse valor.

Distâncias de Cook - Ambulatorial

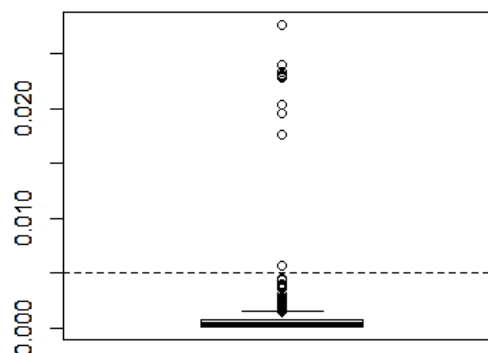


Gráfico 3: Boxplot de Cook – Modelo Ambulatorial

Posteriormente, ajustou-se um modelo sem a variável dependência, logo depois outro sem os pontos influentes. Mas, antes de removê-los é analisado a ocorrência de verossimilhança monótona

(VM). Verifica-se que a abrangência (Tabela 4) para o nível nacional está associado somente a classe baixo risco. A VM configura-se quando uma categoria da variável resposta está associado a uma única classe de uma variável explicativa. Para mais detalhes desse evento, consulte HEINZE 2002 e HEINZE 2001.

| Verossimilhança Monótona | | |
|--------------------------|-------------|------------|
| Abrangência | Baixo Risco | Alto Risco |
| Não Nacional | 993 | 497 |
| Nacional | 43 | 0 |

Tabela 4: VM – Ambulatorial

Como a variável abrangência é significativa, escolheu-se o ajuste com a presença dos pontos influentes, apresentado a seguir:

| Modelo Escolhido Ambulatorial | $\hat{\beta}_i$ | Erro | Teste Z | Valor p |
|-------------------------------|-----------------|--------|---------|-------------------|
| Intercepto | 0,0715 | 0,1497 | 0,4770 | 0,6331 |
| Sexo-Masculino | -0,8275 | 0,1151 | -7,1890 | $6,54 * 10^{-13}$ |
| Idade | -1,0119 | 0,3116 | -3,2480 | 0,0012 |
| Abrangência | -0,7939 | 0,3761 | -2,1110 | 0,0348 |
| Deviance Residual | 1884 | gl | 1539 | |
| Deviance Nula | 1954 | gl | 1542 | |
| AIC | 1892 | | | |

Tabela 5: Modelo Escolhido – Ambulatorial

Quanto aos coeficientes percebe-se o aumento nos valores do intercepto, idade e abrangência, e uma pequena redução na categoria sexo. Analisando a deviance residual do modelo escolhido, a fim de validar a sua adequação, primeiramente, nota-se que ela está muito maior que os graus de liberdade, valor o qual é a média da distribuição assintótica, sendo uma Qui-Quadrado. Posteriormente, calculou-se os valores extremos dessa distribuição que deixam 5% e 1% acima de área, respectivamente:

$$\chi^2_{(1539;0,05)} = 1631,38, \quad Dev. Res = 1884, \quad \chi^2_{(1539;0,01)} = 1671$$

Demonstrou estar fora do intervalo do quantil da $\chi^2_{(n-p,\alpha)}$ ($n = n^\circ$ de observações e $p = n^\circ$ parâmetros estimados) que deixam 5% e 1% acima de área, evidenciando que o modelo ainda precisa de melhorias. A Estatística de Pearson Generalizada retornou um valor de 1547,09 e seu valor-p de 0,43, o qual é maior que uma significância de 0,05. Sendo assim, há indícios estatísticos de que o modelo está bem ajustado. Além disso, os AIC's, praticamente, estão nos mesmos patamares.

A odds (razão de chance) e a interpretação desse modelo seguem:

$$\frac{\theta_i}{1 - \theta_i} = \exp(0,0715 - 0,8275S_{\text{masculino}} - 1,0119\text{Idade} - 0,7939\text{Abg}_{\text{nacional}})$$

- O aumento de 1 unidade na Idade, mantendo as demais covariáveis fixas, causará na odds o impacto de $e^{(-1,0119)} = 0,3635$, ocorrendo uma redução. Quanto a probabilidade $\frac{e^{(-1,0119)}}{1+e^{(-1,0119)}} = 0,2666$. Ou seja, estima-se que, com o aumento de 1 ano na idade a probabilidade de um beneficiário se tornar um alto risco é cerca de 27%.
- O impacto na odds de um beneficiário do sexo masculino, com as demais covariáveis fixas: $e^{(-0,8275)} = 0,4371$, demonstrando uma redução na razão de chances. Referente a probabilidade, estima-se que $\frac{e^{(-0,8275)}}{1+e^{(-0,8275)}} = 0,3042$. Logo, há evidências de um beneficiário do sexo masculino tem cerca de 30% de torna-se um alto risco.
- Em relação ao câmbio de não nacional para nacional sobre a odds, fixando as demais covariáveis: $e^{(-0,7939)} = 0,4521$, mensurando-se uma redução. Probabilisticamente, temos que, $\frac{e^{(-0,7939)}}{1+e^{(-0,7939)}} = 0,3113$, estima-se que um associado com a abrangência nacional tem cerca de 31% de incorrer em um alto risco.

Para os resíduos, investigou-se o comportamento da curva Lowess (Gráfico 4). Observa-se a curva estável ao redor do patamar de resíduo 0 e quase o cruza, mas predominantemente, está na área de resíduo negativo, evidenciando que o modelo pode ser melhorado. Utilizou-se o pacote, "HNP", simulou-se, os envelopes de confiança (Gráfico 5). Esses, são interpretados semelhantes ao qqplot a diferença é a presença das bandas de confiança. Adotou-se, um α de 5% de confiança para validar se os dados, de fato, são de uma distribuição de Bernoulli. Nesse gráfico, apenas 11 pontos de 1543 ficaram dentro do envelope, evidenciando que os dados são de uma Bernoulli.

Para avaliação da qualidade preditiva, plotou-se a curva roc e a área abaixo (auc) desta (Gráfico 5), com auxílio do pacote *pRoc*. Esse gráfico é feito utilizando a sensibilidade e especificidade. A sensibilidade, refere-se à probabilidade de acerto dado que, de fato, ocorreu o acerto. Enquanto a especificidade, a probabilidade de erro dado que, de fato, foi um erro. Em outras palavras, essas duas variáveis medem, respectivamente, a proporcionalidade de acertos e erros de um modelo ajustado. Primeiramente, a curva encontra-se acima da reta diagonal e tem desvio para o canto superior esquerdo do gráfico, o que é bom para o modelo, ou seja, ele se afasta da situação de as taxas de acerto e erro

serem iguais, mesmo que necessite de melhorias. Segundo esse gráfico, o limiar ideal para classificar um possível alto risco ocorreu em 0,417 ponto o qual possui sensibilidade e especificidade, respectivamente, 0,671 – 0,592.

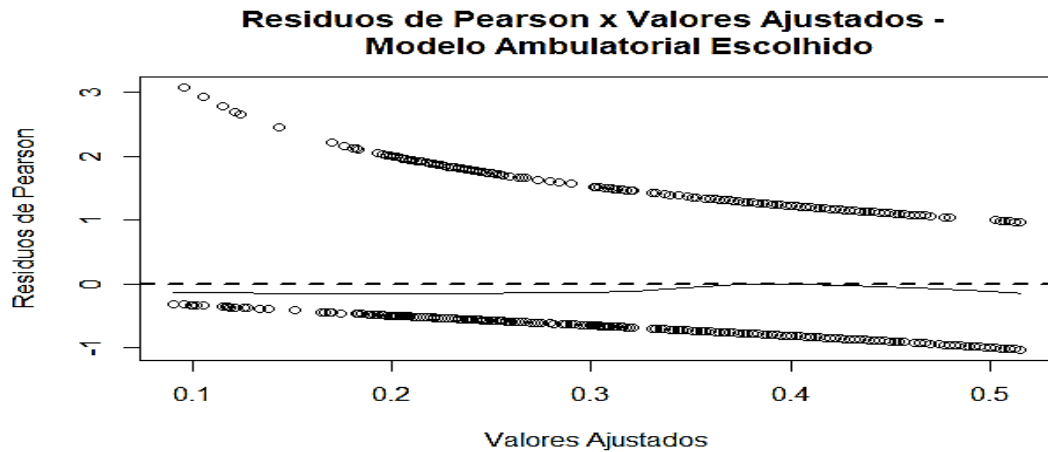


Gráfico 4: Resíduos de Pearson x Valores Ajustados – Modelo Ambulatorial

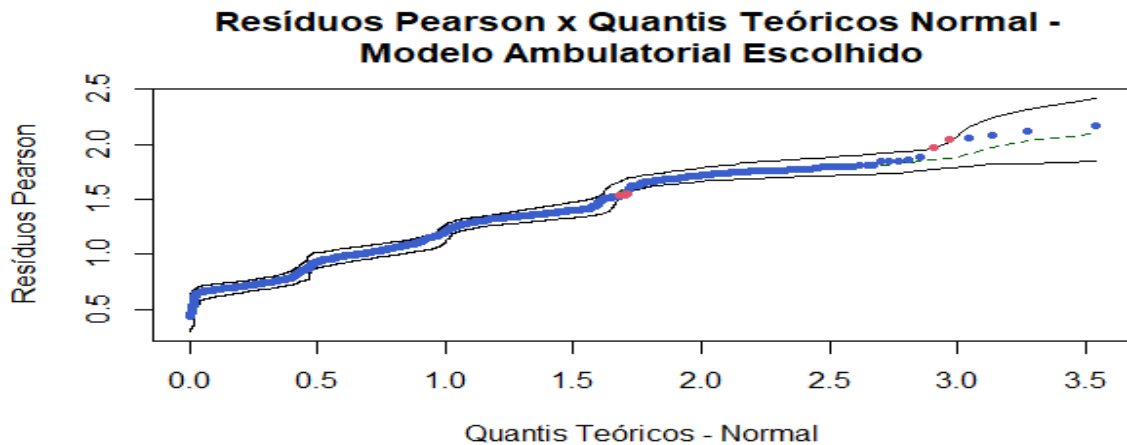


Gráfico 5: Resíduos de Pearson x Quantis Teóricos – Modelo Ambulatorial

Quanto à auc, obteve-se o valor de 0,643: sendo relativamente moderado, em razão de estar distante de 0,5; pois, 0,5 indicaria a situação de que o classificador decide aleatoriamente. Contudo, as possíveis melhorias que o modelo necessita devem impactar positivamente nesses parâmetros.

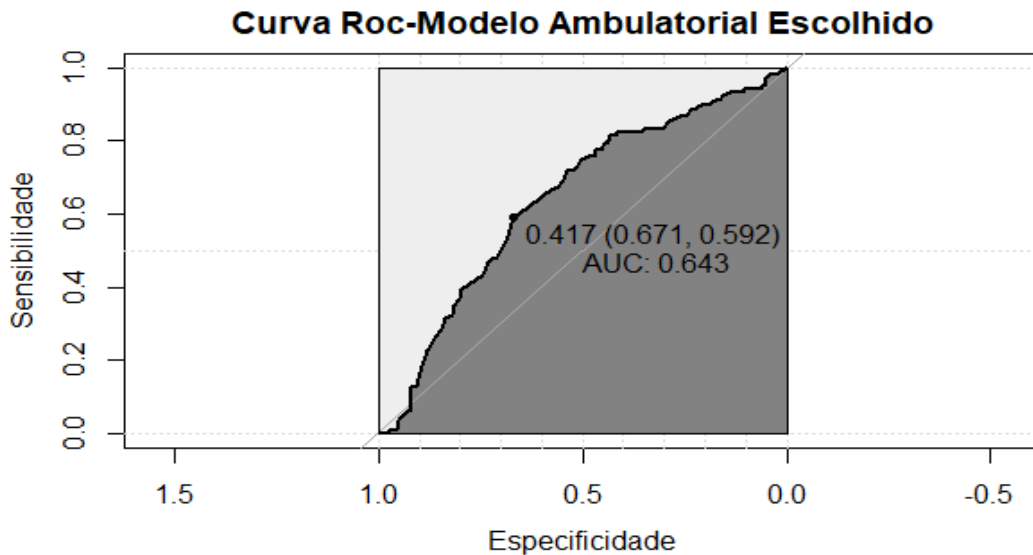


Gráfico 6: Curva Roc – Modelo Ambulatorial Escolhido

4.2. Ambulatorial + Hospitalar

De acordo com a Tabela 6, os dados estão bem mais comportados para a aplicação do modelo. Quanto à Figura 5, a recodificação da abrangência está razoavelmente aceitável, e referente à segmentação ambulatorial+hospitalar com e sem obstetrícia, espera-se captar alguma oscilação na odds em relação a ter ou não cobertura obstétrica, porém a proporção sem obstetrícia é muito pequena, possivelmente, o modelo não captará essa mudança.

| Variável | Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|----------------|--------------|--------------|--------------|--------------|---------------|----------------|
| Idade | 0 | 50 | 61 | 59 | 71 | 102 |
| Receita | R\$ 4.502,00 | R\$ 5.880,00 | R\$ 7.911,00 | R\$ 8.846,00 | R\$ 10.799,00 | R\$ 33.586,00 |
| Custo | R\$ 15,50 | R\$ 893,60 | R\$ 1.888,60 | R\$ 3.126,70 | R\$ 3.765,30 | R\$ 351.840,00 |
| Sinistralidade | 0,0013 | 0,1092 | 0,2317 | 0,3646 | 0,4651 | 2,0000 |

Tabela 6: Estatísticas Descritivas- Ambulatorial + Hospitalar

A correlação entre a sinistralidade e idade resultou em 0,0611 e é esperado, pois o comportamento previsto é conforme a idade aumente o risco. Quando analisado o gráfico de dispersão (Gráfico 6) o grupo de até 10 anos possui baixa sinistralidade e pouquíssimas observações atípicas. Deste panorama, espera-se um coeficiente que seja positivo, indo de encontro com o sucedido para o modelo ambulatorial.

Ambulatorial+Hospitalar - Sinistralidade x Idade

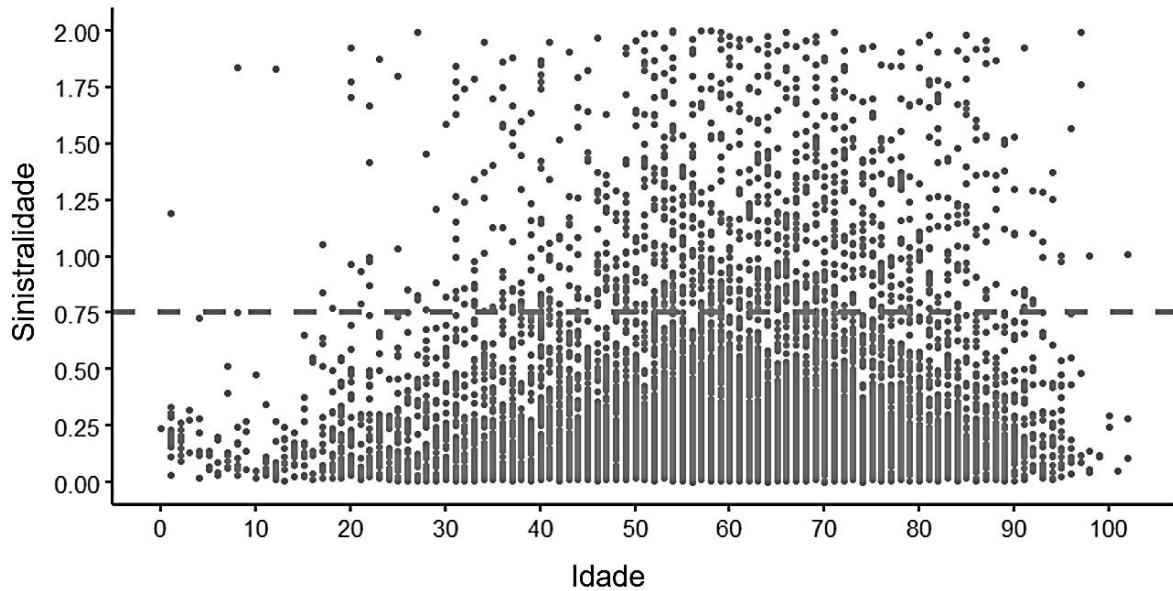


Gráfico 7: Dispersão Sinistralidade x Idade – Ambulatorial+Hospitalar

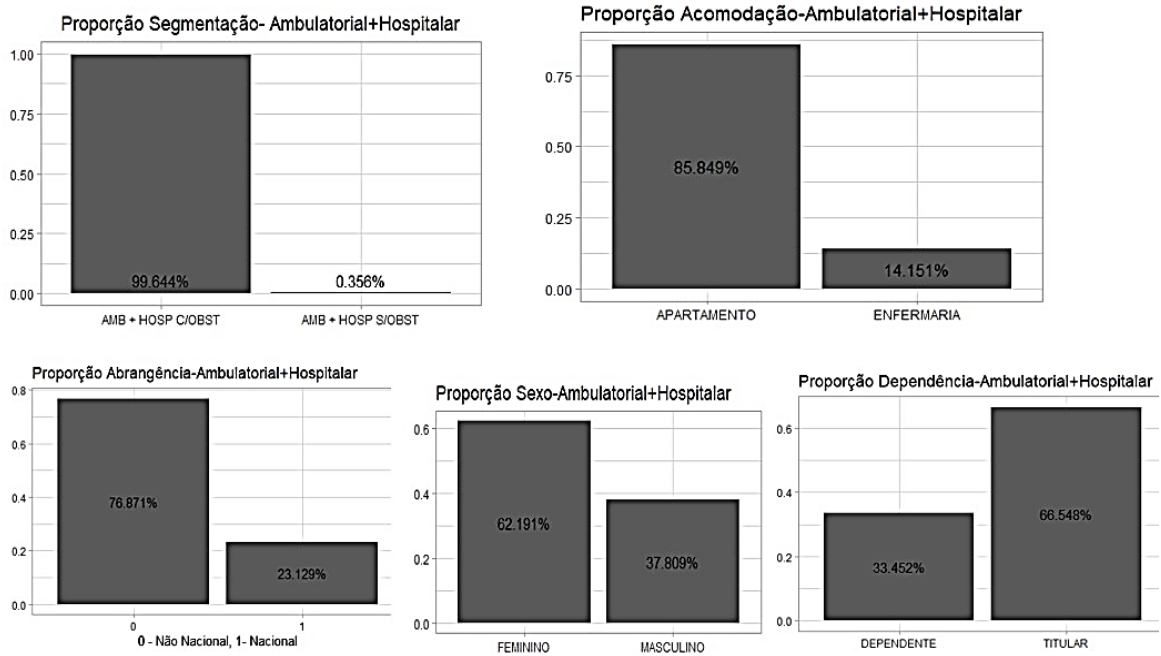


Figura 5: Proporções Variáveis Categóricas -Ambulatorial+Hospitalar

A distribuição da sinistralidade (Figura 6) de acordo com a abrangência e dependência estão, praticamente idênticas e susceptíveis aos níveis de sinistralidade. Sobre a segmentação, nota-se que a mediana do nível com obstetrícia está um pouco acima do que o sem obstetrícia, e assim há evidência desse nível refletir maiores riscos do que a classe sem. Para as outras categorias, acomodação

enfermaria e sexo feminino, seguem o mesmo comportamento de segmentação com obstetrícia. Desse cenário, supostamente, ocasionará redução na odds, pois esses compõem o “base-line” da análise.

Após essas análises, modelou-se o primeiro ajuste (Tabela 7). Desse, ao nível de significância de 0,05, o intercepto e as covariáveis sexo, idade e acomodação, foram significativas. As estimativas de seus coeficientes são coerentes e concordam com o parecer das descritivas desses dados. Os regressores dependência, abrangência e segmentação resultaram com valor-p elevados, consoante a isso, não são significativos para o modelo.

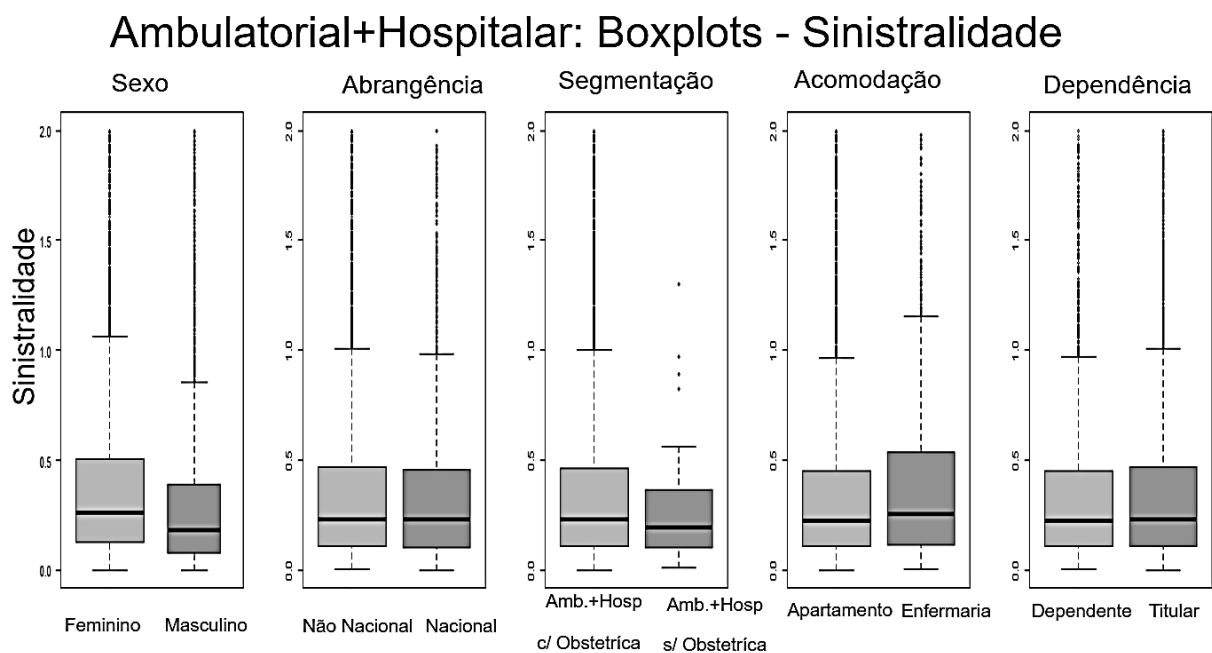


Figura 6: Relação Sinistralidade – Ambulatorial+Hospitalar

Posteriormente, identificou-se os pontos influentes para o ajuste com base no primeiro modelo. Seguindo o mesmo critério aplicado antes, do ponto de corte, para ser considerado influente foi de 0,003, conforme é demonstrado pelo Gráfico 8.

Em seguida, ajustou-se modelos removendo cada regressor não significativo até chegar em um, cujos regressores sejam expressivos. Quando encontrado, foi reajustado sem os pontos influentes. Salienta-se, antes da realização desse molde que, foi analisado a ocorrência de alguma VM e não sucedeu. Sequentemente, a Tabela 8 expõe esse ajuste.

| Modelo Completo Amb + Hosp | $\hat{\beta}_i$ | Erro | Teste Z | Valor p |
|--------------------------------|-----------------|--------|----------|-------------------|
| Intercepto | -2,3421 | 0,1319 | -17,7570 | $2,00 * 10^{-15}$ |
| Sexo-Masculino | -0,3459 | 0,0689 | -5,0210 | $5,13 * 10^{-7}$ |
| Idade | 0,8497 | 0,1946 | 4,3660 | $1,26 * 10^{-5}$ |
| Acomodação-Enfermaria | 0,2952 | 0,0854 | 3,4550 | 0,0005 |
| Dependência- Titular | 0,0220 | 0,0767 | 0,3330 | 0,7394 |
| Abrangência | -0,0220 | 0,0767 | -0,2870 | 0,7743 |
| Segmentação-Amb+Hosp S/Obst | 0,0254 | 0,5402 | 0,0470 | 0,9624 |
| Deviance Residual | 6753,2 | gl | 8698 | |
| Deviance Nula | 6692,2 | gl | 8692 | |
| AIC | 6706,2 | | | |

Tabela 7: 1º Modelo – Ambulatorial+Hospitalar

Distâncias de Cook - Ambulatorial + Hospitalar

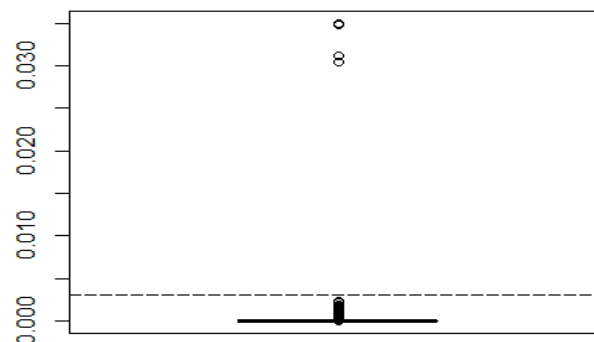


Gráfico 8: Boxplots Distâncias de Cook – Modelo Ambulatorial+Hospitalar

| Modelo Amb + Hosp Escolhido | $\hat{\beta}_i$ | Erro | Teste Z | Valor p |
|-----------------------------|-----------------|--------|----------|-------------------|
| Intercepto | -2,3302 | 0,1263 | -18,4450 | $2,00 * 10^{-15}$ |
| Sexo-Masculino | -0,3474 | 0,0688 | -5,0490 | $4,43 * 10^{-7}$ |
| Idade | 0,8423 | 0,1924 | 4,3780 | $1,19 * 10^{-5}$ |
| Acomodação-Enfermaria | 0,2963 | 0,0853 | 3,4720 | 0,0005 |
| Deviance Residual | 6736,9 | gl | 8694 | |
| Deviance Nula | 6676,8 | gl | 8691 | |
| AIC | 6684,8 | | | |

Tabela 8: Modelo – Ambulatorial+Hospitalar Escolhido

Quanto aos coeficientes, percebe-se aumento nos valores do intercepto e acomodação, e redução na idade e sexo. Para avaliar a adequação do ajuste procedeu-se da mesma forma que o modelo ambulatorial. Nota-se, que a deviance residual está abaixo dos graus de liberdade da Qui-Quadrado, mas está muito distante de 0 demonstrando que ela não é pequena. Ademais, pela análise dos valores críticos da $\chi^2_{(n-p,\alpha)}$ que deixam 5% e 1% acima de área, respectivamente, evidenciando a necessidade de melhorias.

$$\chi^2_{(8691;0,05)} = 8908,99; Dev.Res = 6677; \chi^2_{(8691;0,01)} = 9000,65$$

A Estatística de Pearson Generalizada, ela retornou um valor de 8662,94 e seu valor p de 0,58, o qual é maior que uma significância de 0,05. Sendo assim, há indícios estatísticos de o modelo está bem ajustado. Além disso, o AIC para o modelo escolhido é o menor entre os outros testados.

A odds desse modelo é dada:

$$\frac{\theta_i}{1 - \theta_i} = \exp(-2,3302 - 3,3474S_{masculino} + 0,8423Idade + 0,2963Aco_{enfermaria})$$

- O aumento de 1 ano na Idade, mantendo as demais covariáveis fixas, causará na odds o impacto de $e^{(0,8423)} = 2,3217$, demonstrando que, o aumento de 1 ano na idade, ocasiona-se ao beneficiário está mais susceptível a torna um alto risco. Probabilisticamente, $\frac{e^{(0,8423)}}{1 + e^{(0,8423)}} = 0,6989$. Ou seja, estima-se que, com o aumento de 1 ano na idade a probabilidade de um beneficiário se tornar um alto risco é cerca de 70%;
- O impacto na odds de um beneficiário do sexo masculino, com as demais covariáveis fixas, é: $e^{(-3,3474)} = 0,0352$, estima-se que, um associado do sexo masculino é menos propenso a ser um alto risco. Quanto a probabilidade de ele configura-se como alto risco, é mensurada $\frac{e^{(-3,3474)}}{1 + e^{(-3,3474)}} = 0,0339$;
- Em relação à acomodação, a enfermaria impacta a odds, fixando as demais covariáveis: $e^{(0,2963)} = 1,3449$, expõe que, um beneficiário relacionado à acomodação enfermaria é mais susceptível a configurar-se como um alto risco. A probabilidade deste, torna-se um risco em potencial é estimada em $\frac{e^{(0,2963)}}{1 + e^{(0,2963)}} = 0,5735$

Para análise dos resíduos, novamente, o comportamento da curva lowess (Gráfico 9). Observa-se uma curva, predominantemente, abaixo do resíduo de patamar 0 e segue a tendência dos resíduos negativos. Expondo, desta maneira que o ajuste necessita de melhorias. Com o auxílio do pacote, "HNP", simulou-se, os envelopes de confiança (Gráfico 10). Esses, são interpretados semelhantes ao qqplot a diferença é a presença das bandas de confiança. Adotou-se, um α de 5% de confiança para validar se os dados, de fato, são de uma distribuição de Bernoulli. Nesse gráfico, todos os pontos ficaram dentro do envelope, evidenciando que os dados advêm da distribuição mencionada.

Avaliando a qualidade preditiva, a curva ROC (Gráfico 11) encontra-se próxima da reta diagonal, situação de as taxas de acerto e erro serem, relativamente, próximas. Segundo esse gráfico, o limiar ideal para classificar ocorreu em 0,117 ponto, o qual possui sensibilidade e especificidade, respectivamente, 0,623 – 0,471. Quanto à auc, obteve-se o valor de 0,535, sendo próximo de 0,5 demonstrando a situação de que o classificador decide como se jogasse uma moeda. Portanto, esse modelo não é apropriado para classificação, mas pode ser utilizado para avaliar os impactos das covariáveis

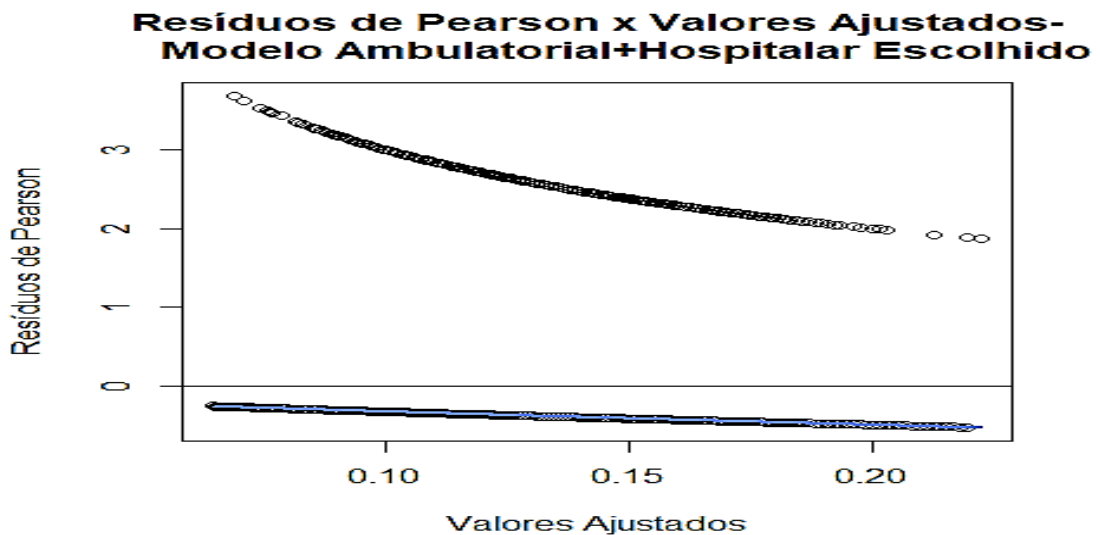


Gráfico 9: Resíduos de Pearson x Valores Ajustados – Modelo Ambulatorial+Hospitalar

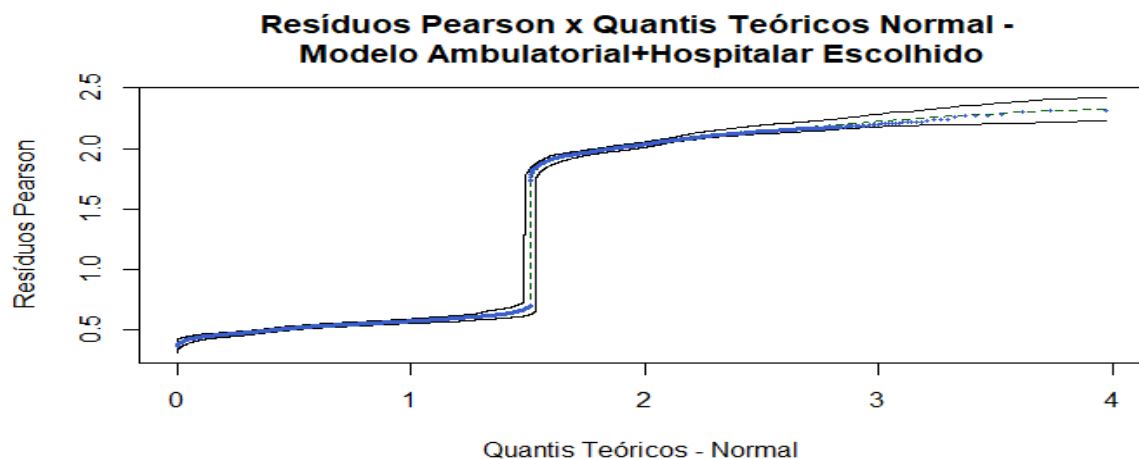


Gráfico 10: Resíduos de Pearson x Quantis Teóricos – Modelo Ambulatorial+Hospitalar

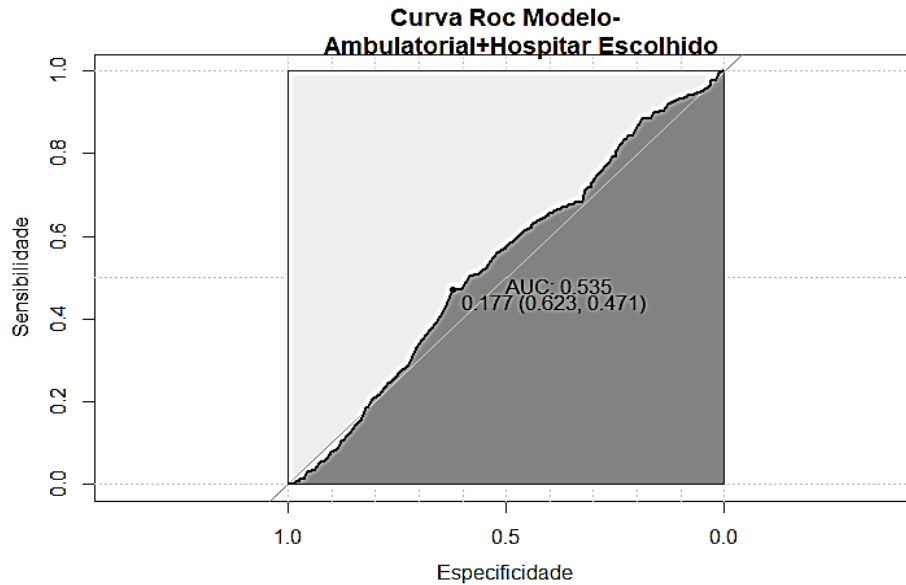


Gráfico 11: Curva Roc – Modelo Ambulatorial+Hospitalar Escolhido

5. CONCLUSÃO

Os modelos apresentados, embora tenham suas limitações, atenderam ao objetivo deste trabalho identificar e estimar os impactos dos beneficiários considerados altos riscos. O ajuste proposto à segmentação ambulatorial pode ser aplicado para realizar previsões, dado que obteve um auc de 0,643; logo há evidências estatísticas para esse efeito. Por outro lado, para o ambulatorial+hospitalar não seja um bom classificador, apesar disso, pode ser empregado em avaliação dos fatores de risco propiciados pelas características de um plano de saúde. Apesar de não ter proposto um modelo de risco para o hospitalar, devido aos entraves surgidos durante a análise, foi detectado que as variáveis explicativas Idade, Acomodação e Dependência carregam informações relevantes.

A escolha do ponto de corte pela sinistralidade demonstrou ser uma alternativa para lidar com as implicações que o comportamento do custo assistencial traz, das definições relativas, acerca do alto custo apresentadas na literatura e além de poderem ser escolhidos outros pontos de corte, de acordo com a realidade de uma operadora.

A divisão do banco pela segmentação apresentou ser um bom particionamento, embora seja necessário ter representatividades consideráveis para cada nível. Retratadas as limitações ao longo do desenvolvimento deste estudo, como o não uso de algumas variáveis, a exemplo, o fator moderador e variáveis retrativas do uso dos serviços de saúde, sendo estas que poderiam colaborar para identificação dos altos riscos. Para os modelos, a inclusão de variáveis que carreguem informações sobre: o uso dos serviços prestados, fatores moderados (redutores de risco), tempo médio de uma utilização entre os

serviços prestados, condições de saúde, entre outras colaborariam para melhorias, desde que sejam realizadas as devidas análises para inclusão de um novo regressor.

Conclui-se, para os devidos fins, mesmo com as limitações dessa pesquisa, que o objetivo foi alcançado, foram encontrados dois modelos capazes de: avaliarem os impactos dos fatores de riscos para a estimação da probabilidade de um beneficiário se tornar alto risco. Sendo um modelo para um plano de saúde ambulatorial e o outro para um ambulatorial+hospitalar. Ademais, a escolha do ponto de corte pela sinistralidade demonstrou ser uma alternativa para lidar com as implicações acarretadas pelo comportamento do custo assistencial. Desfecha-se, para os devidos fins, que essa pesquisa alcançou seu objetivo, segundo o apresentado pela literatura, pormenor que esteja a predição realizada para esses beneficiários de alto risco, ela é válida por abordar um evento de difícil previsão.

APÊNDICE

HOSPITALAR

Essa repartição ficou muito reduzida, caindo de 1014 para 46 observações, por causa da presença de diversos valores destoantes. Tentou-se mudar o ponto de corte de receita com base no painel de precificação, e mesmo assim, persistiram, no que tange à sinistralidade. A tabela abaixo demonstra alguns desses valores.

| Variável | Mínimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|----------------|------------|--------------|--------------|--------------|--------------|---------------|
| Idade | 4 | 33 | 38 | 42 | 51 | 85 |
| Receita | R\$ 337,00 | R\$ 711,60 | R\$ 1.120,30 | R\$ 2.293,30 | R\$ 2.188,80 | R\$ 12.109,70 |
| Custo | R\$ 60,00 | R\$ 2.828,00 | R\$ 5.021,00 | R\$ 5.725,00 | R\$ 6.900,00 | R\$ 31.764,00 |
| Sinistralidade | 0,0141 | 1,3914 | 3,7575 | 5,2435 | 8,8622 | 16,9058 |

Tabela 9: Estatísticas Descritivas- Hospitalar

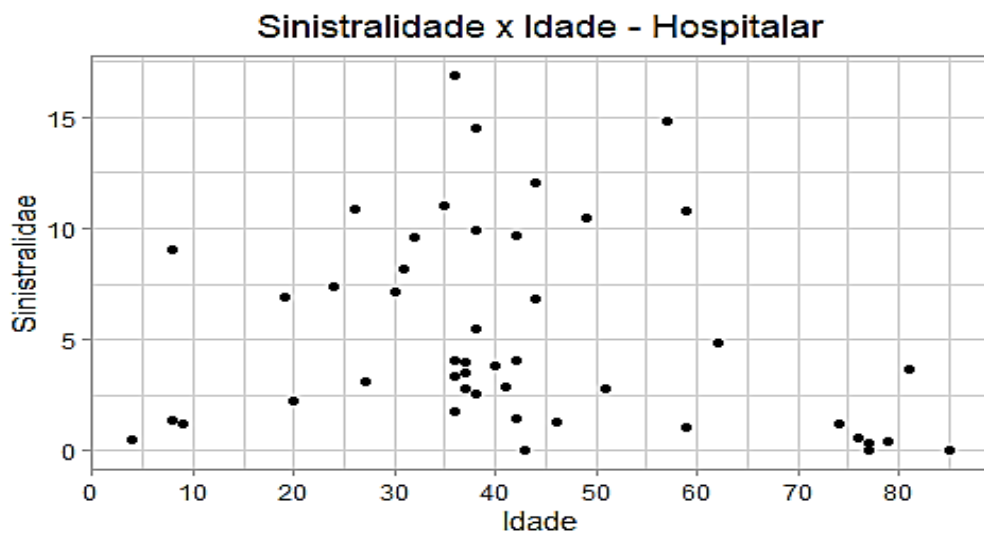


Gráfico 12: Dispersão Sinistralidade x Idade Hospitalar

Novamente, a correlação entre a sinistralidade foi de encontro ao esperado e resultou no valor de -0,2153, e concorrente a esse fato o gráfico de dispersão (Gráfico 12) demonstra um beneficiário com cerca de 10 anos com a sinistralidade elevada. Testou-se a remoção dele, e mesmo assim a correlação ficou negativa. Novamente, espera-se um coeficiente negativo para a idade. Ressalto que, ocorreram outras observações desse tipo e como essa partição já era pequena, decidiu-se manter essas verificações.

Pesquisando as relações de sinistralidade pelas covariáveis categóricas para o “base-line” da abrangência e sexo, aconteceu de estarem experimentando maiores níveis de sinistralidade, podendo

desconfiar de uma possível redução na odds, enquanto para acomodação e dependência ocorre o contrário.

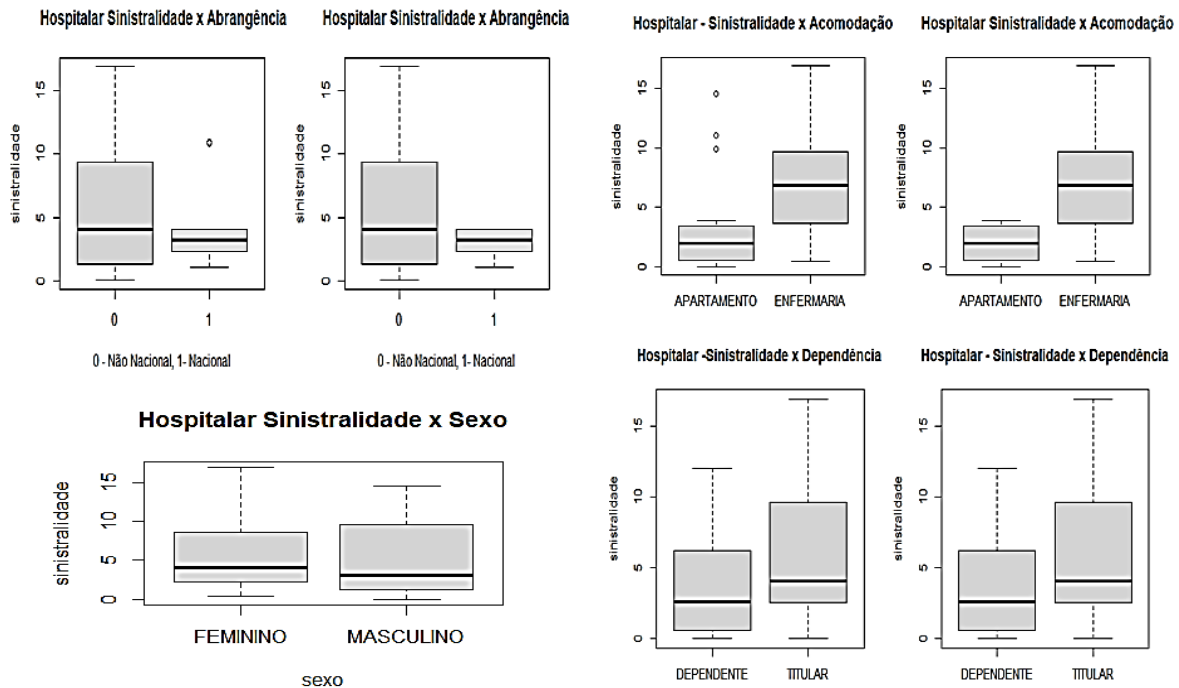


Figura 7: Relação Sinistralidade -Hospitalar

Ajustou-se um primeiro modelo para esse segmento (Tabela 10) e deparou-se com o problema de verossimilhança monótona. Neste caso, as estimativas não são precisas, devido ao algoritmo não convergir para que sejam devidamente calculadas. A Tabela 11 demonstra a ocorrência relatada.

| Modelo Completo Hospitalar | $\hat{\beta}_i$ | Erro | Teste Z | Valor p |
|----------------------------|-----------------|-----------|---------|---------|
| Intercepto | 12,0540 | 9,9660 | 1,2100 | 0,2260 |
| Sexo-Masculino | -10,7760 | 9,2200 | -1,1690 | 0,2420 |
| Idade | -21,3110 | 15,8890 | -1,3410 | 0,1800 |
| Acomodação-Enfermaria | 4,9970 | 5,7460 | 0,8700 | 0,3840 |
| Dependência-Titular | 14,6200 | 11,3880 | 1,2840 | 0,1990 |
| Abrangência | 14,8130 | 4365,5960 | 0,0030 | 0,9970 |
| Deviance Residual | 39,2345 | gl | 45 | |
| Deviance Nula | 9,4726 | gl | 40 | |
| AIC | 19,555 | | | |

Tabela 10: 1º Modelo- Hospitalar

| Verossimilhança Monótona | | |
|--------------------------|-------------|------------|
| Abrangência | Baixo Risco | Alto Risco |
| Não Nacional | 7 | 29 |
| Nacional | 0 | 10 |

Tabela 11: Verossimilhança Monótona _ Abrangência- Hospitalar

Tentou-se ajustar outros modelos sem a abrangência. No entanto, a VM novamente apareceu, como mostra a tabela 12. Aplicou-se, a anova com teste qui-quadrado no primeiro modelo (Tabela 13), para identificar quais regressores seriam significativos para esse modelo. Conforme seguem, a idade – acomodação e dependência foram significativas para um α de 0,05.

| Acomodação | Baixo Risco | Alto Risco |
|-------------|-------------|------------|
| Apartamento | 4 | 12 |
| Enfermaria | 0 | 24 |

Tabela 12: Verossimilhança Monótona _ Acomodação- Hospitalar

| | Gl | Dev.Res | Gl | Resi.Dev | Valor p |
|-------------|--------|---------|---------|----------|---------|
| Null | | | 45,0000 | 39,2340 | |
| Sexo | 1,0000 | 0,2874 | 44,0000 | 38,9470 | 0,5919 |
| Idade | 1,0000 | 8,9836 | 43,0000 | 29,9630 | 0,0027 |
| Acomodação | 1,0000 | 4,0648 | 42,0000 | 25,8990 | 0,0438 |
| Dependência | 1,0000 | 16,3439 | 41,0000 | 9,5550 | 0,0001 |
| Abrangência | 1,0000 | 0,0822 | 40,0000 | 9,4720 | 0,7743 |

Tabela 13: Anova- Análise Deviance-Modelo Hospitalar

REFERÊNCIAS

- AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (Brasil). **Painel de Precificação Planos de Saúde**. Rio de Janeiro, junho 2021. Disponível em: <https://app.powerbi.com/view?r=eyJrjoiNWQ0NzcyYzgtM2FhY00MGQ4LTlhMjktNTgzMWZmODY1ODEyYmE0ODBlLTRmYTctNDJmNC1iYmEzLTBmYjEzNzVmYmU1ZiJ9&pageName=ReportSection5c53e7c32090a5b7d403>.
- AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (Brasil). **Prisma Econômico-Financeiro da Saúde Suplementar**. Rio de Janeiro, junho 2021. Disponível em: <https://app.powerbi.com/view?r=eyJrjoiNjViNzQ2NmQtYWFhNy00MzQ0LWJjODUtOGI3OWNiY2ZjNDgwliwidCI6ljlkYmE0ODBlLTRmYTctNDJmNC1iYmEzLTBmYjEzNzVmYmU1ZiJ9>.
- AQUINO, Luana Ramos de. **Análise De Regressão Logística Para Predição De Custos Assistenciais No Setor De Saúde Suplementar**. Sergipe, 2017. Disponível em: <https://ri.ufs.br/bitstream/riufs/67272/Luana%20Ramos%20de%20Aquino.pdf>. Acesso em 27 mai. 2021.
- BIERMAN, A. S. et al. How Well Does a Single Question about Health Predict the Financial Health of Medicare Managed Care Plans? *Effective Clinical Practice*, v. 2, n. 2, p. 56–62, 1999.
- Da Silva, J. P. B. C. 2016. **“Modelos de Regressão Linear e Logística Usando o Software R”**. Master’s thesis, Universidade Aberta. Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. New Jersey: Wiley & Sons.
- DOVE, H. G.; DUNCAN, I.; ROBB, A. A prediction model for targeting low-cost, high-risk members of managed care organizations. *The American journal of managed care*, v. 9, n. 5, p. 381–9, maio 2003.
- IESS, Instituto de Estudos de Saúde Complementar. **Caracterização dos beneficiários de alto custo assistencial - Um estudo de caso**. Disponível em: <https://www.iess.org.br/?p=publicacoes&id=878&idtipo=15>. Acesso em: 30 de jul. de 2021.
- HEINZE, Georg; SCHEMPER, Michael. **A solution to the problem of monotone likelihood in Cox regression**. *Biometrics*, v. 57, n. 1, p. 114-119, 2001.
- HEINZE, Georg; SCHEMPER, Michael. **A solution to the problem of separation in logistic regression**. *Statistics in medicine*, v. 21, n. 16, p. 2409-2419, 2002.
- LAVANGE, Lisa M. et al. An application of logistic regression methods to survey data: Predicting high cost users of medical care. In: **Proc. Survey Research Methods Section**. 1986. Disponível em: <https://www.semanticscholar.org/paper/AN-APPLICATION-OF-LOGISTIC-REGRESSION-METHODS-TO-%3A-LaVange%20Iannacchione/b6360ea6775b9850c6488031caab68bd2dbd2965?p2df>. Acesso em 01 de ago. de 2021.
- Montgomery, D. C., Peck, E. A., Vinning G. G. (2012) **Introduction to Linear Regression Analysis, 5th ed.**, Haboken: John Wiley.
- Moral RA, Hinde J, Demétrio CGB (2017). **“Half-Normal Plots and Overdispersed Models in R: The hnp Package.”** *Journal of Statistical Software*, *81*(10), 1-23. doi: 10.18637/jss.v081.i10 (URL: <https://doi.org/10.18637/jss.v081.i10>).
- NUNES, André. **O Envelhecimento Populacional E As Despesas Do Sistema Único De Saúde**. Jan. 2004. Disponível em: <https://www.ipea.gov.br/portal/images/stories/PDFs/livros/Arq21Cap13.pdf>. Acesso em: 30 mai. 2021.
- PLANALTO. Lei nº 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais**. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 22 mar. 2021.
- SANTOS, Samara Lauer; TURRA, Cássio M.; NORONHA, Kenya. **Envelhecimento Populacional E Gastos Com Saúde: Uma Análise Das Transferências Intergeracionais E Intergeracionais Na Saúde Suplementar Brasileira**. 2018. Disponível em: <http://dx.doi.org/10.20947/S102-3098a0062>. Acesso em: 27 jun. 2021.
- SCHRAMM, Joyce Mendes de Andrade; OLIVEIRA, Andreia Ferreira; LEITE, Iúri da Costa; et al. **Transição Epidemiológica e o Estudo de Carga de Doença no Brasil**. Dez. 2004. Disponível em: <https://doi.org/10.1590/S1413-81232004000400011>. Acesso em: 27 jun. 2021.
- SOUZA, Helano Silva Eugênio; LEÃO, Luiz Carlos da Silva. **Tarifação De Um Plano De Saúde Autogestão Aplicando Os Modelos Lineares Generalizados**. Rio de Janeiro, 2012. Disponível em: dx.doi.org/10.12957/cadest. Acesso em: 27 mai. 2021.
- Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). **pROC: an open-source package for R and S+ to analyze and compare ROC curves**. *BMC Bioinformatics*, 12, p. 77. DOI: 10.1186/1471-2105-12-77 <http://www.biomedcentral.com/1471-2105/12/77>