MODELANDO A TAXA DE NEOPLASIA PULMONAR NO BRASIL VIA MODELOS LINEARES GENERALIZADOS

Thiago Rezende dos Santos[1,2], **Roger William Câmara Silva**[1]

Resumo: *O câncer de pulmão é uma das principais causas de morte na atualidade em todo mundo, tendo como um de seus principais fatores de risco o tabagismo, e assim, constitui um dos mais importantes problemas de saude pública no Brasil. Neste trabalho, utilizando dados correspondentes a taxa de incidência da doença nas cinco regiões brasileiras, discriminadas por sexo e ano, ajustamos um Modelo Linear Generalizado para modelar a variável resposta Taxa de Incidência de Neoplasia Pulmonar através das covariaveis mencionadas, a saber, Região, Sexo e Ano. Mostramos que o modelo é adequado e que todas as covariáveis são significativas ao nível de 5% de significância com exceção da variável Ano. As regiões sul e sudeste têm as maiores taxas e os homens apresentam uma propensão maior que as mulheres.*

Palavras-chave: Câncer, Modelo Gama, Regiões do Brasil.

Introdução

No organismo verificamos formas de crescimento celular, que podem ser controladas e não controladas. A hiperplasia, a metaplasia e a displasia são exemplos de crescimento controlado, enquanto que as neoplasias correspondem às formas de crescimento não controladas e são denominadas, na prática, de "tumores". A primeira dificuldade que enfrentamos no estudo das neoplasias é a sua definição, pois ela se baseia na morfologia e na biologia do processo tumoral, sofrendo modificações com a evolução do conhecimento. A definição mais aceita atualmente é: "Neoplasia é uma proliferação anormal do tecido, que foge parcial ou totalmente ao controle do organismo e tende à autonomia e à perpetuação com efeitos agressivos sobre o hospedeiro" [2].

Segundo o Instituto Nacional de Câncer (INCA), o câncer é a segunda causa de óbitos no país, com tendência de crescimento nos próximos anos, sendo assim uma questão de saúde pública, principalmente ao se levar em consideração que aproximadamente um terço dos novos casos de câncer no mundo poderiam ser evitados. A neoplasia maligna mais frequente no Brasil é a de pele não melanoma, com maiores taxas nas regiões Sudeste, Sul e Centro-Oeste. No sexo masculino, seguem-se as de próstata, de estômago e de pulmão (inclusive traqueia e brônquios). No sexo feminino, a neoplasia maligna de mama é a mais incidente, seguindo-se a de pele não melanoma e a de colo de útero. De maneira geral, as regiões Sudeste e Sul apresentam as taxas mais elevadas, em ambos os sexos.

Alguns exemplos de fatores de risco associados às localizações de neoplasias malignas são: tabagismo (90% dos casos de neoplasia do pulmão, traqueia e brônquios); consumo de álcool e dieta pobre em fibras (esôfago); consumo de sal e alimentos defumados (estômago); dietas ricas em gordura e colesterol (cólon e reto); radiação solar (pele); fatores genéticos (melanoma);

[1]Departamento de Estatística, ICEX, UFMG, thiagords@ufmg.br,rogerwcs@yahoo.com.br
[2]Autor de correspondência.

comportamento hormonal e reprodutivo (mama feminina); higiene precária e exposição ao vírus do papiloma humano (colo de útero); irritação mecânica crônica (boca).

Neste trabalho, consideramos apenas as neoplasias de pulmão (inclusive traqueia e brônquios), o mais comum de todos os tumores malignos, apresentando aumento de 2% ao ano na incidência mundial, segundo o INCA. Ainda segundo o INCA, a sobrevida média de pacientes com este tipo de câncer varia entre 7% e 10% em países desenvolvidos e entre 13% e 21% em países em desenvolvimento, sendo assim altamente letal.

Tendo em vista o exposto acima, torna-se necessária cada vez mais a compreensão dos fatores que interferem na incidência de neoplasia pulmonar. O objetivo deste trabalho é estabelecer a relação, caso exista, entre a incidência de neoplasia pulmonar no Brasil e as covaráveis, que temos disponíveis no banco de dados do [2], as taxas de incidência de neoplasia pulmonar para as cinco regiões do país, nos anos de 2005-2006, 2007-2008, 2009-2010, segundo o sexo. Por uma questão de simplicidade, o efeito das regiões são captadas via covariáveis.

Os dados, que consistem de 30 observações, estão disponíveis no site do DATASUS e representam as taxas estimadas de novos casos de neoplasias malignas, por 100 mil habitantes, discriminados segundo os critérios acima, e podem ser utilizadas para estimar o risco de ocorrência de casos novos de neoplasias malignas, além de dimensionar sua magnitude como problema de saúde pública. Essas taxas são construídas de acordo a uma metodologia internacionalmente reconhecida descrita por [1]. Maiores detalhes sobre o método de cálculo podem ser obtidas no site do DATASUS.

Neste estudo, encontramos algumas limitações, como por exemplo, as estimativas para as regiões do país baseiam-se em dados provenientes de alguns municípios, que são cobertos por RCBP. Essas estimativas estão sujeitas a variações, tanto na metodologia de cálculo quanto na cobertura do RCBP, o que recomenda cautela em análises temporais. As tendências tendem a aumentar em função da melhoria das condições de diagnóstico.

Os dados utilizados foram fornecidos pelo Ministério da Saúde/Instituto Nacional do Câncer (INCA): Registro de Câncer de Base Populacional (RCBP) e Cenepi/Sistema de Informações sobre Mortalidade (SIM).

Este artigo está organizado da seguinte forma: Na segunda Seção é apresentado a metodologia. Na terceira Seção é mostrada a análise dos dados de neoplasia pulmonar via MLG. Finalmente, é feita a conclusão na última Seção.

Metodologia

O modelo de análise de regressão linear é uma das técnicas mais usadas em análise de dados e há aplicações deste modelo em diferentes áreas do conhecimento. Neste artigo, consideramos a classe de modelos lineares generalizados (MLGs) desenvolvidos por [4] que desempenham hoje um papel muito importante na Estatística, uma vez que generalizam o modelo clássico de regressão linear, abrindo um leque de opções para a distribuição da variável resposta, bem como permitindo maior flexibilidade para ligação entre a média e a parte sistemática do modelo. Assim, a hipótese básica de normalidade não é mais exigida para a análise dos dados.

O modelo linear generalizado é definido por uma distribuição de probabilidade, membro da família exponencial de distribuições, para a variável resposta, um conjunto de variáveis independentes descrevendo a estrutura linear do modelo e uma função de ligação entre a média da variável resposta e o preditor linear.

Para a melhor escolha da referida distribuição de probabilidade para a variável resposta, é aconselhável examinar os dados buscando observar alguns aspectos, tais como: assimetria, natureza discreta ou contínua, intervalo de variação, etc. É importante salientar que os termos que compõem a matriz modelo podem ser de natureza contínua, qualitativa ou mista, e que devem ter uma contribuição significativa na explicação da variável resposta.

Observe que se $Y \sim Gama(\mu, \phi)$ (veja [3]), então, para $0 < y < \infty$, a função de densidade

de Y é

$$f_Y(y) = \exp\{\phi[-y/\mu - \log(\mu)] + \phi\log(\phi y) - \log(y) - \log\Gamma(\phi)\}, \quad (1)$$

onde $y > 0$, $\phi > 0$ e $\mu > 0$. Escrevendo esse modelo na forma da família exponencial, concluímos que $\theta = -1/\mu$, $b(\theta) = -\log(-\theta)$ e ϕ (parâmetro de dispersão). Assim, utilizando resultados dos MLG's, temos

$$E(Y) = \mu = -\frac{1}{\theta} \quad Var(Y) = \frac{\phi}{\theta^2} \quad V(\mu) = \phi\mu^2. \quad (2)$$

O MLG que utilizaremos na Seção Ajuste do MLG assume que a variável resposta Y_{ijk} a incidência de neoplasia pulmonar da região i no período j e para o sexo k tem distribuição Gama com função de ligação logaritmica para $i = 1, \ldots, 5$, $j = 1, \ldots, 3$ e $k = 1, 2$. Neste caso, o modelo Gama terá a seguinte fórmula funcional:

$$log(\mu_{ijk}) = \beta_0 + \beta_{1(i)} + \beta_{2(j)} + \beta_{3(k)}, \quad (3)$$

onde $\mu_{ijk} = E(Y_{ijk})$.

Análise dos dados de neoplasia pulmonar

Nesta seção é feita a análise dos dados com o auxílio do *software* R. Primeiramente, fazemos uma análise descritiva e, posteriormente, verificamos o ajuste de um MLG Gama. Para maior clareza do texto, apresentamos na Tabela 1 as variáveis envolvidas em nosso estudo.

Tabela 1: Variáveis envolvidas no estudo.

Variável Resposta			Notação	Codificação *Dummies*
		Taxa de incidência de neoplasia pulmonar traquéia e brônquios	Y	-
Covariáveis				
	Regiões	Região Nordeste	RNordeste	1 0 0 0
		Região Norte	RNorte	0 1 0 0
		Região Sudeste	RSudeste	0 0 1 0
		Região Sul	RSul	0 0 0 1
		Região Centro-oeste	RCentro-oeste	0 0 0 0
	Sexo	Masculino	Masc	1
		Feminino	Fem	0
	Ano	2004-2005	Ano	0 0
		2006-2007		1 0
		2008-2009		0 1

Análise descritiva

Apresentamos agora uma análise exploratória dos dados da taxa de incidência de neoplasia maligna de pulmão, traquéia e brônquios por 100 mil habitantes, buscando entender a relação entre incidência, sexo, região e ano. Temos 30 observações divididas nos periodos de 2004-2005, 2006-2007 e 2008-2009, regiões e sexo. A Figura 1 sugere que as regiões Sudeste e Sul possuem as maiores taxas de incidência média e também os maiores desvios-padrão, enquanto que as regiões nordeste e norte apresentam as menores taxas de incidência média e desvios-padrão. Podemos observar também que o número de casos de neoplasia pulmonar parece ser maior entre os homens. Nota-se que a distribuição da variável resposta é assimétrica à direita e os seus valores possíveis estão na semi-reta positiva, indicando que um modelo Gama seja adequado a esses dados. Ainda de acordo com a Figura 1, à medida que o tempo passa, o número de casos de neoplasia pulmonar por 100 mil habitantes não se altera.

Ajuste do MLG

A Tabela 2 apresenta as estimativas dos parâmetros e seus respectivos p-valores do modelo gama com interações entre as covariaveis. A covariável Ano não foi significativa (p-valor $> 0,50$), por isso não a apresentamos na tabela abaixo. Os níveis 0 denotam as categorias-base de

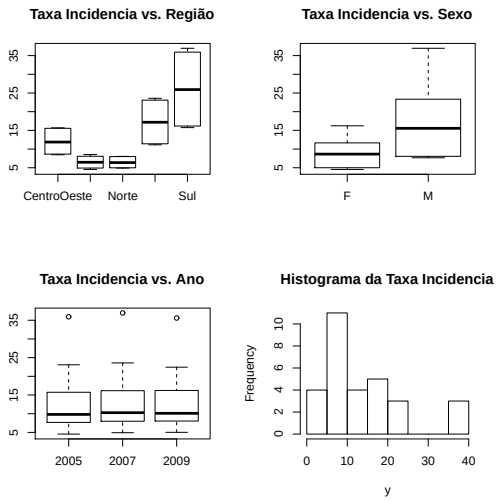

Figura 1: Análise exploratória dos dados do número de casos novos de neoplasia maligna de pulmão.

comparação. Note que todas as covariáveis são significativas ao nível de 5% de significância com exceção da interação repesctiva à região nordeste ($valor - p = 0,07$), mas resolvemos mantê-la no modelo pelo o valor-p ser limítrofe e por conviniência. Observe também que o p-valor do teste Deviance é maior que 0,10, isto é, não rejeitamos a hipótese nula de que o modelo é adequado ao nível 5% de significância. A presença de interação entre os fatores sexo e região significa que a diferença entre as incidências de neoplasia pulmonar de homens e mulheres não é a mesma à medida que variamos as regiões.

Tabela 2: Estimativas dos parâmetros do modelo Gama ajustado aos dados.

Covariáveis	Estimativas	p-valor
RCentro-oeste	0	-
RNordeste	-0,57	0,00
RNorte	-0,55	0,00
RSudeste	0,28	0,00
RSul	0,61	0,00
Fem	0	-
Masc	0,58	0,00
RCentro-oeste*Fem	0	-
RNordeste*Masc	-0,08	0,07
RNorte*Masc	-0,11	0,01
RSudeste*Masc	0,12	0,01
RSul*Masc	0,24	0,00
Intercepto	2,16	0,00
ϕ	0,001	-

Na Figura 2 há alguns pontos fora dos limites correspondentes à região nordeste, a qual possui valores baixos de incidência, ver Figura 1. Contudo, eles não impactam tanto as estimativas do modelo.

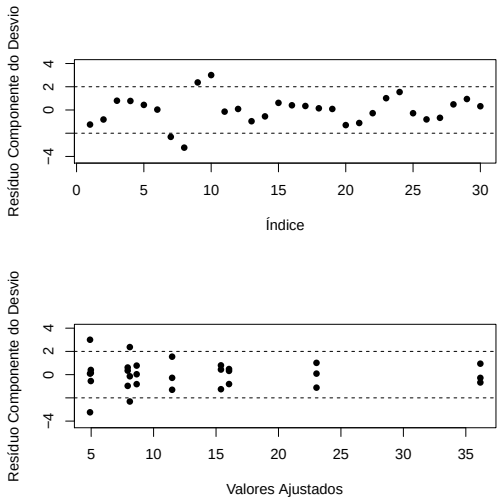

Figura 2: Análise diagnóstico do modelo Gama para os dados de neoplasia. Resíduos Deviance.

Isso pode ser comprovado através da análise do gráfico Q-Q Plot na Figura 3. Desta forma, não rejeitamos a hipótese de que o modelo é adequado.

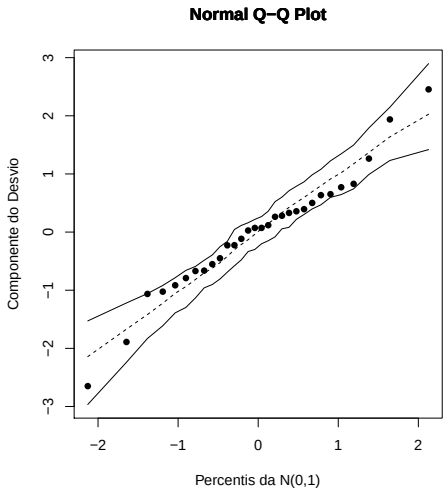

Figura 3: Gráfico de envelope de probabilidade para o modelo Gama ajustado aos dados.

Interpretação

O modelo final é aquele dado na Tabela 2. Mantendo as outras variáveis constantes, se a região for nordeste, norte, sudeste e sul o logarítmo da taxa média diminui 0,57 unidades, 0,55 unidades e aumenta 0,28 unidades e 0,61 unidades, respectivamente. A taxa média fica multiplicada por 0,57 quando é considerado a região nordeste se comparado com a região centro-oeste. A região sul e sudeste são as que apresentam maior taxa de neoplasia pulmonar se comparado com a região centro-oeste e as regiões nordeste e norte as menores.

Se o sexo for masculino, aumenta 0,58 unidades no logarítmo da taxa média. A taxa média de neoplasia fica multiplicada por 1,79 quando o indivíduo é do sexo masculino, isto é, os homens estão mais propensos à doença do que as mulheres. Há indícios de que existe um aumento maior da taxa de neoplasia pulmonar para o sexo masculino nas regiões Sul e Sudeste. Não efeito de

interação para a região nordeste, isto é, a diferença da incidência entre os homens e mulheres para essa região não altera se comparado com a região centro-oeste.

Conclusões

Neste trabalho, o nosso objetivo foi ajustar um modelo linear generalizado para explicarmos a variável taxa de incidência de neoplasia pulmonar, traquéia e brônquios através das covariáveis que temos disponíveis via o banco de dados do DATASUS. É notório que as regiões sul e sudeste têm taxas de incidência maior que as outras regiões. Isso pode ser explicado por diferentes hábitos e costumes entre as regiões. É interessante o fato de que a taxa de incidência de neoplasia pulmonar aumenta quando o sexo é masculino. Uma possível explicação para este fato é que os homens são mais resistentes em ir ao médico e fazer exames, além de que os hábitos de homens e mulheres são muito diferentes. Segundo o mesmo estudo da SBC, a taxa de fumantes entre os homens é de 28%, enquanto que nas mulheres esta taxa é de 17%, o que contribui para uma maior incidência entre os homens. A existência de interação entre sexo e região é um fato interessante, pois isso significa a incidência entre homens e mulheres não é a mesma quando mudamos as regiões.

Referências

[1] Black R. J., Bray, F., Ferlay, J. and Parkin, D. M., Cancer Incidence and Mortality in the European Union: Cancer Registry Data Estimates of National Incidence for 1990, *European Journal of Cancer*, **37(7)**, 1075-1107, 1997.

[2] Datasus, Dados sobre a morbidade por neoplasia maligna de pulmão, brônquios e traqueia. *Disponível em http://w3.datasus.gov.br/datasus/datasus.php.*

[3] McCullach, P., Nelder, J. A., *Generalized linear models*. 2.ed. London: Chapman and Hall, 1989. 511p.

[4] Nelder, J. A., Wedderburn, R. W. M., Generalized linear models, *Journal of the Royal Statistical Society, Series A (Royal Statistical Society)*, **135(3)**, 370-384, 1972.