**PROGNÓSTICO DO CÂNCER DE PULMÃO: UMA ABORDAGEM VIA MODELOS DA ANÁLISE DE SOBREVIVÊNCIA**

¹Dionisio Alves da Silva Neto; ²Paulo Cerqueira dos Santos Júnior

1Acadêmico de Estatística, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil; 2Professor Doutor em Estatística, Universidade Federal do Pará (UFPA), Belém, Pará, Brasil.

**Eixo Temático:** Eixo Transversal.

**E-mail do Autor Principal:** dionisioneto899@gmail.com

**Resumo**

**Introdução**: A análise de Sobrevivência é dada pelo conjunto de técnicas estatísticas elaboradas para avaliar o tempo até a ocorrência de um evento, denominado de tempo de falha. Por este motivo, é bastante comum utilizar dos conceitos desenvolvidos nesta área para estudar o tempo até a morte de um paciente ou a duração total até a recuperação após uma determinada debilidade. **Objetivo**: Neste estudo, o objetivo volta-se à investigação do tempo de sobrevivência para o óbito ocasionado pelo câncer de pulmão e verificar se a presença de covariáveis corrobora para este malefício. **Metodologia**: Este trabalho utilizou-se do banco de dados aberto providenciado por um importante centro de pesquisa norte-americano. A implementação do estimador de Kaplan-Meier e diversas propostas probabilísticas para os dados nos permitiu observar em quais momentos a doença atinge o estágio final, gerando a morte dos pacientes. O modelo de regressão de Cox, ajustado pelo método de seleção *stepwise*, selecionou as covariáveis significativas para compreender quais fatores diminuem tempo de sobrevida. **Resultados**: Através do estimador não-paramétrico de Kaplan-Meier, conseguimos inferir que metade dos indivíduos não ultrapassam os 200 primeiros dias após o diagnóstico. Em adição, dentre os principais modelos probabilísticos abordados na literatura, concluiu-se que os dados do tempo de falha são genuinamente explicados pela distribuição de probabilidade Weibull. Ao analisar a influência das covariáveis coletadas ao longo do experimento, resultou-se na significância do sexo masculino e da qualidade do cuidado próprio como fatores de proteção da doença. **Conclusão**: Este tipo de informação é extrema necessidade dentro das Ciências da Saúde. Investigar o tempo até a morte por meios de modelos probabilísticos, associados aos fatores de risco e proteção, pode guiar o diagnóstico médico baseado em fatos e, por consequência, liderar atitudes que aumentem a sobrevida de um paciente ou população.

**Palavras-chave:** DistribuiçãoWeibull; Modelo de Cox; Curva de Sobrevivência; Taxa de Falha.

**1 INTRODUÇÃO**

A Análise de sobrevivência, também denominada de Teoria da Confiabilidade, é o ramo da Bioestatística que se volta para a investigação do tempo de ocorrência até um determinado evento. Também chamado de tempo de falha, esta quantificação é criada quando atividades de monitoramento, considerando uma unidade de tempo, são cronometradas exaustivamente em relação a um evento de interesse e, com isso, pode-se determinar o tempo percorrido desde um início temporal (BOGAERTS; KOMÁREK; LESAFFRE, 2017).

As pesquisas médicas e o interesse na análise da sobrevida de pacientes sujeitos a determinadas doenças, debilidades e tratamentos em estudos clínicos e experimentais foram o que deram origem a esta área na atuação da ciência de dados. De fato, esta ciência tem manuscritos datados desde do século 17, no qual a primeira tábua da vida foi criada genuinamente por John Graunt, no ano de 1662 (MORABIA, 2013). Na era atual, devido ao avanço computacional e desenvolvimento de métodos estatísticos de ponta, a Análise de Sobrevivência vem se estabelecendo como um dos pilares da Estatística moderna, o qual vem ganhando um grande destaque nos últimos 20 anos (COLOSSIMO; GIOLO, 2006).

A análise do tempo de falha usualmente engloba a designação de um dos três principais encargos da Teoria da Confiabilidade. De modo profundo, é também de interesse investigar as funções de sobrevivência e suas interpretações, comparar dois ou mais tratamentos em termos de eficácia e avaliar o efeito de covariáveis na sobrevivência de grupos/indivíduos (SUN, 2006).

Apesar de serem verdadeiramente classificadas como informações incompletas no banco de dados, é de suma importância incluí-las dentro da análise estatística. Colossimo e Giolo (2006) apontam que as razões que justificam esta medida são: (*i*) os dados censurados são capazes de fornecer informações sobre o tempo de sobrevida dos elementos; (*ii*) a exclusão das informações censuradas pode gerar estimadores e estimativas viciadas nos cálculos inferenciais.

**2 METODOLOGIA**

A pesquisa é de natureza quantitativa do tipo descritivo e experimental, a qual está interligada com o estudo divulgado pelo grupo de estudos clínicos *North Central Cancer Treatment Group*, entidade líder em pesquisa sobre o câncer nos Estados Unidos. Ademais, a pesquisa é caracterizada como longitudinal, a qual é caracterizada pelo acompanhamento de variáveis, sobre um indivíduo ou grupo, ao longo do tempo sob o controle do pesquisador. Desse modo, o objetivo do trabalho era verificar o tempo de sobrevivência dos pacientes associado à sua performance no cotidiano e características genéticas. Como resultado, a amostra coletada totalizou em um grupo de 228 pacientes diagnosticados com câncer de pulmão.

O processo de coleta de dados ocorreu entre julho de 1987 a junho de 1990, por meio de um questionário estruturado com 30 itens sobre informações demográficas, de saúde e do cotidiano, o qual foi direcionado para todos os pacientes atendidos pelos médicos colaboradores do instituto. As principais covariáveis coletadas agregavam características sobre: idade, sexo, qualidade de cuidado próprio, qualidade de executar tarefas comuns, quantidade de calorias consumidas nas refeições e quantidade de peso perdido nos últimos seis meses; além do momento de diagnóstico e óbito pela doença.

Como critérios de inclusão considerou-se os indivíduos adultos e com moderado grau de instrução escolar, cujos apresentassem o estágio avançado, incurável e maligno do câncer de pulmão. Como critérios de exclusão considerou-se apenas os pacientes que completaram todas as informações requeridas do questionário, assim como a inexistência de qualquer outra enfermidade mais maligna do que a doença em investigação.

**2.1 MÉTODOS ESTATÍSTICOS**

Declara-se que o tempo de sobrevivência *T* é uma variável aleatória contínua e não negativa, isto é, *T* pode assumir qualquer valor dentro do conjunto dos reais a partir de zero. Por alguns embasamentos das teorias probabilísticas associadas a esta variável aleatória, obrigatoriamente tem-se uma função densidade de probabilidade *f(t)* e função de distribuição acumulada F(t) para explicar o seu comportamento. A partir da função de distribuição, temos a definição da função de sobrevivência como a probabilidade de um indivíduo sobreviver acima ou igual a um determinado instante no tempo *t,* dado que este sobreviveu até este determinado momento. Em outras palavras, a função de sobrevivência representa a área complementar da função de distribuição do tempo de falha:

Em contraste com a função de sobrevivência, outra função importante para a interpretação do tempo de falha é a função taxa de falha, a qual explica o potencial instantâneo por unidade de tempo para a ocorrência do evento, quando se sabe que o indivíduo sobreviveu até um tempo *t*. A definição para a função taxa de falha é dada por:

No qual denota um acréscimo infinitamente pequeno.

 O estimador não-paramétrico de Kaplan-Meier serve para descrevermos a função de sobrevivência na sua forma empírica, mas com a consideração de censura. A forma matemática desta metodologia é dada por:

 O modelo de regressão de Cox, ou modelo de Riscos Proporcionais, proporciona não apenas a probabilidade de que um de ocorrência de um determinado evento, mas também a estimativa até o tempo de ocorrência por meio da função de risco. A popularidade desta metodologia está presente na interpretação das covariáveis que influenciam o tempo de sobrevivência pelo **Razão de Taxas de Falha** ou **Risco Relativo**. A estrutura do modelo de Cox é dada por:

em que é chamada de função de risco basal, é um vetor de dimensão de covariáveis e é um vetor dos coeficientes.

**3 RESULTADOS E DISCUSSÕES**

Ao longo do estudo, desenvolvemos 6 famosos modelos de probabilidade para explicar o tempo de sobrevida relacionado ao câncer de pulmão. As distribuições testadas foram: Exponencial, Weibull, Log-Normal, Log-Logístico, Gama e Gama generalizado.

Percebemos que o modelo exponencial, por ser simples em sua forma, destoou bastante da forma empírica da função de sobrevivência e, por isso, este modelo não seria adequado para explicar os dados do tempo de sobrevivência do câncer de pulmão. Por outro lado, o modelo de sobrevivência de Weibull abordou um ajuste esplendidamente adequado, ao se aproximar bastante do formato empírico da função de sobrevivência estimada.

O ajuste do modelo Log-Normal não conseguiu explicar corretamente o comportamento da função de sobrevivência empírica, englobando vários erros em diversos períodos temporais abordados pelo estimador de Kaplan-Meier. Por outro lado, embora o ajuste do modelo Log-Logístico tenha sido satisfatório além do tempo mediano de sobrevivência, no final da curva de Kaplan-Meier, ocorreu um sobreajuste da proposta, tal falha gera uma expectativa de sobrevivência maior do que o esperado.

Para os modelos gama e gama generalizado, constatou-se a qualidade de ambos modelos pela proximidade de suas curvas suaves próximas ao formato escada do estimador de Kaplan-Meier. Com destaque para a forma generalizada, a qual fornece menos defeitos ao longo do ajuste e, principalmente, no final da curva estimada da função de sobrevivência.

Além da forma visual, por meio de gráficos, uma outra forma de se avaliar e selecionar os modelos de sobrevivência é através dos critérios e seleção AIC (*Akaike information criterion*), BIC (*Bayesian information criterion*) e HC (*Hannan-Quinn Information Criterion*). Essas propostas se baseiam essencialmente no valor de maximização da função de log-verossimilhança, no tamanho amostral e na quantidade de parâmetros envolvidos no processo (AKAIKE, 1984; SCHWARZ, 1978; AKAIKE & QUINN, 1958). A regra de decisão para cada medida consiste em selecionar o modelo que contém a menor valoração.

Na **Tabela 1**, observa os valores de AIC, BIC e HC para os modelos de sobrevivência exponencial, Weibull, Log-Normal, Log-Logístico, Gama e Gama generalizado. Deste modo, nota-se em destaque o desempenho do modelo Weibull, o qual poderia não ser descoberto apenas na visualização gráfica. De fato, ao analisarmos dentre os diversos modelos propostos para os dados de câncer, realizou-se que os mais apropriados seriam o Weibull e o gama generalizado, porém, os detalhes sobre qual era a melhor forma paramétrica não eram possíveis na inspeção gráfica.

**Tabela 1:** Valores das medidas de seleção AIC (*Akaike information criterion*), BIC (*Bayesian information criterion*) e HC **(***Hannan-Quinn Information Criterion*), por modelo de sobrevivência.

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo | AIC | BIC | HC |
| Exponencial | 2326,676 | 2330,106 | 2328,060 |
| Weibull | **2311,702** | **2318,561** | **2314,470** |
| Log-Normal | 2342,538 | 2349,397 | 2345,305 |
| Log-Logístico | 2325,861 | 2332,720 | 2328,629 |
| Gama | 2313,469 | 2320,328 | 2316,237 |
| Gama generalizado | 2313,380 | 2323,668 | 2317,531 |

**Fonte**: Construído pelo autor, 2023.

Com a descoberta do modelo paramétrico Weibull como melhor forma de explicar a curva de sobrevivência, devermos estudar a influência de covariáveis que possam influenciar no potencial de risco instantâneo para morte do paciente. Desse modo, a **Tabela 2** aborda os resultados ajuste do modelo de Cox com a utilização das covariáveis idade (age), sexo (sex), qualidade de cuidado próprio (ph.ecog), qualidade em executar tarefas comuns (ph.karno), quantidade de calorias consumidas em refeições (meal.cal) e quantidade de peso perdido nos últimos seis messes (wt.loss), acompanhadas do intercepto estimado. Quando avaliamos o nível descritivo (p-valor) para cada covariável, considerando o nível de 5% de significância, percebemos que apenas as variáveis sexo, qualidade no cuidado próprio e qualidade em executar tarefas do cotidiano comprovaram-se estatisticamente significativas. Em outras palavras, a significâncias destas covariáveis nos permite a inferir que existe uma influência dos pesos estimada para uma generalização populacional.

**Tabela 2:** Resultado da estimação global para o coeficiente, coeficiente na escala exponencial, estatística de teste (Z) e p-valor, por variável explicativa.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variável | Coeficiente |  | Exp (Coeficiente) | Z | P-valor |
| Intercepto | 7,6740  |  | 2151,7566  | 7,60 | 3,1e-14 |
| Age | -0,0085  |  | 0,9916  | -1,04  | 0,2997 |
| Sex | 0,4034  |  | 1,4970  | 2,82  | 0,0048 |
| Ph.ecog | -0,5777  |  | 0,5612  | -3,90 | 9,8e-05 |
| Ph.karno | -0,0149  |  | 0,9852 | -1,96  | 0,0495 |
| Meal.cal | 0,000048  |  | 1,0000  | 0,26  | 0,7915 |
| Wt.loss | 0,00752  |  | 1,0075 | 1,41  | 0,1596 |

**Fonte**: Construído pelo autor, 2023.

O ajuste do modelo permite a interpretação das covariáveis em nível amostral, porém, para termos interpretações populacionais devemos manter os coeficientes aprovados no teste de significância. Por consequência, utilizamos o método *stepwise*, o qual é amplamente abordado na modelagem de regressão múltipla, para combinarmos todos os modelos possíveis sob a conceitualização de significância dos parâmetros estimados. Em síntese, o método retornou como resultado a formação de um modelo utilizando apenas as variáveis sexo e qualidade em cuidado próprio, evidenciando que a covariável qualidade em realizar tarefas no cotidiano é apenas significativa quando inserida em uma estimação conjunta com todas as outras.

A partir da **Tabela 3**, podemos interpretar os resultados e influências das covariáveis no tempo de contração do câncer de pulmão. Pelo valor exponencial do coeficiente estimado e, assumindo a propriedade de riscos proporcionais, podemos inferir que o grupo feminino manifesta 49% mais chance de morrer pelo câncer de pulmão em relação ao grupo masculino. Ademais, a cada um acréscimo na pontuação da qualidade de cuidado próprio, reduzimos em 30% a chance de óbito em relação à doença. Portanto, a regressão de Cox mostra que o pertencimento ao sexo feminino e o descaso em ter hábitos saudáveis e boa alimentação são fatores de risco para a contração do câncer de pulmão.

**Tabela 3:** Resultado da estimação *stepwise* para o coeficiente, coeficiente na escala exponencial, estatística de teste (Z) e p-valor, por variável explicativa.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variável | Coeficiente |  | Exp (Coeficiente) | Z | P (Z < z) |
| Intercepto | 5,8196 |  | 336,8342 | 30,60 | < 2e-16 |
| Sex | 0,4014  |  | 1,4939 | 3,24  | 0,0012 |
| Ph.ecog | -0,3557  |  | 0,7007 | -4,31 | 1,7e-05 |

**Fonte**: Construído pelo autor, 2023.

**4 CONSIDERAÇÕES FINAIS**

Neste presente trabalho foram abordados modelos de sobrevivência não-paramétricos, paramétricos e semiparamétricos para o prognóstico do tempo de sobrevivência após a contração câncer de pulmão. Pelo estimador de Kaplan e Meier, verificamos que dentro da coorte do estudo, o tempo mediano para a contração da doença foi de aproximadamente 280 dias, como também os percentis 25 e 75 foram perto 200 e 600 dias, nesta respectiva ordem. Dentre as seis propostas potencias para explicar a curva elaborada pelo método descritivo de Kaplan e Meier, o modelo Weibull foi o que melhor adequou-se aos dados segundo os critérios AIC, BIC e HC e, além disso, pelo princípio da parcimônia, devemos adotar o mesmo pela metodologia de um processo menos oneroso. Para avaliar o efeito de covariáveis sobre o potencial de risco instantâneo da contração do câncer de pulmão, a metodologia *stepwise* retornou como resultado a importância e significância das covariáveis sexo e qualidade no cuidado próprio como bases para explicar o risco de desenvolvimento deste malefício.

**REFERÊNCIAS**

AKAIKE, H. Likelihood of a model and information criteria. **Journal of Econometrics**, v. 16, n. 1, p. 3–14, 1984.

APPOLINÁRIO, F. **Dicionário de metodologia científica**: um guia para a produção do conhecimento científico. São Paulo: Atlas, 2007.

BOGAERTS, K.; KOMÁREK, A.; LESAFFRE, E. **Survival analysis with interval-censored data: A practical approach with examples in R, SAS, and BUGS**. Boca Raton: Chapman and Hall/CRC, 2017.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Editora Blücher, 2006.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society:
Series B (Methodological),** v. 34, n. 2, p. 187–202, 1972

GIL, A. C. **Metodologia da pesquisa**. São Paulo: Atlas, 2002.

GONSALVES, E. P. **Iniciação à pesquisa científica**. 3. ed. Campinas: Alínea, 2003.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 41, n. 2, p. 190-195, 1979.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, v. 53, n. 282, p. 457-481, 1958.

LOPRINZI, C. L.; LAURIE, H. S.; WIEAND, H. S.; KROOK, J. E.; NOVOTNY, P. J.; BARTEL, J. W. K.; LAW, M.; BATEMAN, M.; KLATT, N. E. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. **Journal of Clinical Oncology**, v. 12, n. 3, p. 601-607, 1994.

MORABIA, A. Epidemiology’s 350th Anniversary: 1662–2012. **Epidemiology (Cambridge, Mass.)**, v. 24, n. 2, p. 179, 2013.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, v. 6, n. 2, p. 461-464, 1978.

SUN, J. **The Statistical Analysis of Interval-censored Failure Time Data**. New York: Springer, 2006.