Universidade Federal do
Paraná Setor de Ciências
Exatas Departamento de
Estatística
Programa de Especialização em *Data Science e Big Data*

João Marcos Melo Santos

Modelos Probabilísticos em Análise de Sobrevivência de portadores da doença de Chagas

Curitiba
2022

João Marcos Melo Santos

Modelos Probabilísticos em Análise de Sobrevivência de portadores da doença de Chaga

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. José Luiz Padilha

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Modelos Probabilísticos em Análise de Sobrevivência em portadores da doença de Chagas

João Marcos Melo Santos[1]
Prof. José Luiz Padilha[2]

Resumo

Em estatística a técnica que melhor analisa dados com informações censuradas é chamada de análise de sobrevivência. Existem várias técnicas paramétricas e não paramétricas que são amplamente usadas neste tipo de análise. Este estudo considera os modelos paramétricos e selecionou as distribuições exponencial, Weibull e log-normal para modelar o tempo de sobrevida de indivíduos portadores da doença de chagas dos pacientes do hospital das clínicas de Minas Gerais entre os anos de 1999 e 2019, no intuito de identificar quais fatores influenciam no tempo mediano de sobrevivência.

Palavras-chave: Análise de Sobrevivência, modelos paramétricos, exponencial, Weibull, log-normal.

Abstrair

In statistics the technique that best analyzes data with censored information is called survival analysis. There are several parametric and nonparametric techniques that are widely used in this type of analysis. This study considers the parametric models and selected exponential, Weibull and log-normal distributions to model the survival time of individuals with chagas disease of patients in the hospital of the clinics of Minas Gerais between 1999 and 2019, in order to identify which factors influence the median survival time.

Keywords: *Survival Analysis, parametric models, exponential, Weibull, log-normal.*

1 Introdução

A doença de Chagas é causada pelo protozoário parasita Trypanosoma cruzi, que por sua vez causa miocardite aguda e, posteriormente, uma miocardite crônica fibrosante, de baixa intensidade e incessante, que produz dano miocárdico progressivo e resulta tardiamente na cardiomiopatia crônica da doença de chagas. Em função das ações de controle de vetores realizadas a partir da década de 1970, o Brasil recebeu em 2006 a certificação Internacional da interrupção da transmissão vetorial por Triatoma infestans, espécie exótica e responsável pela maior parte da transmissão vetorial no passado. Porém, estima-se que existe aproximadamente 12 milhões de portadores da doença crônica nas Américas, e que haja no Brasil, atualmente, pelo menos um milhão de pessoas infectadas por Trypanosoma cruzi.

Entre 2007 e 2019, foram registrados 3.118 casos confirmados agudos da doença de chagas, com média anual de 239 casos apresentando maior incidência nos últimos anos de 2018 e 2019 (0.182 e 0.183 casos por 100.000 habitantes, respectivamente) para o Brasil (BRASIL, 2022).

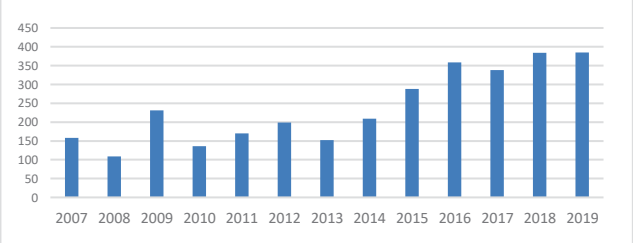

Figura 1: Casos da doença de chagas registrados no Brasil de 2007 a 2019.

2 Conceituação e Definição

2.1 Público alvo

Os dados utilizados no estudo tem como origem a base de informações do acompanhamento de doentes de chagas coletadas no período entre os anos de 1999 e 2019 no Hospital das Clínicas da Universidade Federal de Minas Gerais.

2.2 Objetivo

Objetivo deste estudo é analisar a influência das covariáveis no risco de morte de pacientes com a doença de chagas, através de modelos de regressão paramétricas em análise de sobrevivência.

[1] Aluno do programa Especialização em Data Science & Big Data, joao.marcos_123_@hotmail.com
[2]Professor do Departamento de Estatística - DEST/UFPR.

3 Metodologia

3.1 Tratamento dos dados

Os dados analisados constam com 28 variáveis e 615 observações, sendo que, 357 são do sexo feminino e 258 são do sexo masculino e cujos, possuem idade média de 50,31 e 51,91 anos, respectivamente.

3.2 Imputação de Dados Faltantes

Verificou-se que o conjunto de dados trazia inicialmente um grande número de dados faltantes. Apenas a resposta (tempo e indicador do evento) e 2 covariáveis não possuíam valores ausentes. Com isso, apenas 8% dos indivíduos possuíam informação completa em todas as variáveis.
Daí em diante, o ajuste de modelos contendo todas as variáveis torna-se inviável caso não exista um tratamento nos dados ausentes em razão de que na maioria das técnicas de modelagem não são ajustados caso uma linha tenha pelo menos um dado ausente.
Sabendo disto, foi optado por preencher os valores faltantes através do método de imputação implementado pelo pacote MICE do software R. A função mice utiliza o procedimento chamado 'full conditional specification', no qual para cada variável com valores ausentes é construído um modelo de regressão condicional em todas as demais variáveis no banco de dados. Foi adotado o procedimento recomendado por White & Royston (2009), em que o tempo de sobrevivência é substituído pela estimativa da função taxa de falha acumulada obtida pelo estimador de Nelson-Aalen.
Foi aplicada a imputação simples utilizando as configurações padrões da função mice. A justificativa da imputação simples deve ao fato de que estamos interessados em explorar aspectos dos modelos paramétricos de sobrevivência e não buscamos realizar uma inferência definitiva das covariáveis no risco no evento.

4 Modelagem do Problema

4.1 Análise Exploratória dos Dados

Primeiramente, foi avaliado cada covariável individualmente, através das curvas de sobrevivência obtidas pelo estimador de Kaplan-Meier, juntamente com o valor-p do teste logrank. Como o teste só pode ser realizado em variáveis de grupo, as variáveis contínuas foram categorizadas em intervalos de mesma amplitude. Ao testar as curvas de sobrevivência para cada covariável, encontramos 17 covariáveis significativas, como exposto na Figura 1, (p-valor < 0,05), conforme o teste de logrank, entre 26 variáveis testadas.

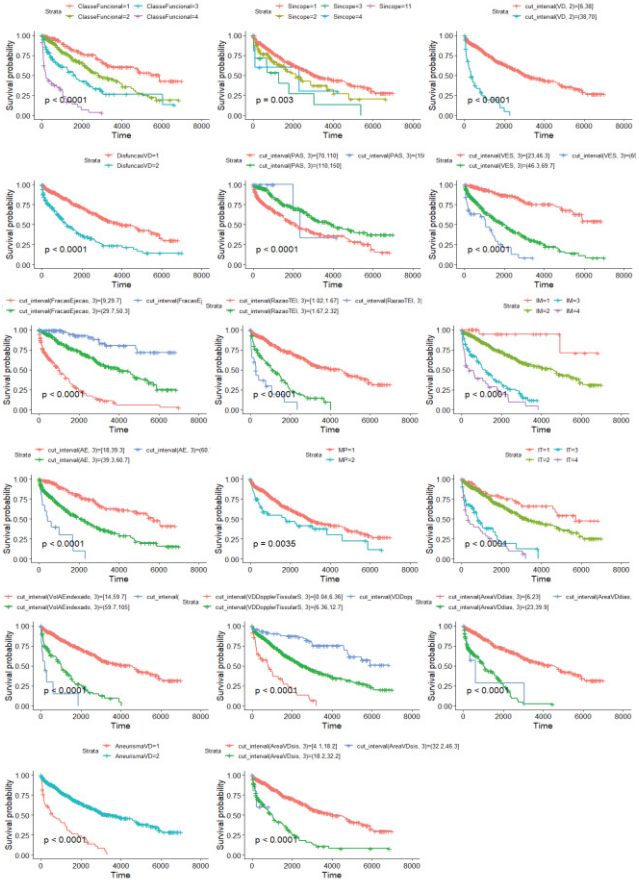

Figura 2.: Curvas de sobrevivência estimadas por Kaplan-Meier e p-valor do teste Logrank.

As 17 covariáveis encontradas foram consideradas candidatas nos modelos de regressão descritos a seguir.

4.2 Modelos de Regressão

Os modelos probabilísticos de análise de sobrevivência são métodos amplamente usados em estudos biomédicos e industriais, em que, utilizam o tempo até a ocorrência de um evento, como variável resposta, e variáveis explicativas de interesse, para estimar a sobrevida.
Uma característica importante que levou a escolha de um método específico como análise de sobrevivência é a presença de censura, que torna-se a observação incompleta. Isso ocorre quando os indivíduos permanecem sem mudança de estado ao término do estudo, ou por causas não relacionadas com a doença de interesse, ou abandonam o estudo.
Os modelos paramétricos são eficientes na obtenção dos estimadores de parâmetros de regressão por sua ampla adequação. Os modelos exponencial, Weibull e log-normal são os mais usados por responder bem a várias situações práticas, pois, em diversos casos conseguem descrever com eficiência a distribuição do tempo de vida. Assim, os três modelos foram testados inicialmente como candidatos para descrever o tempo de vida de pacientes com doença de Chagas.
De modo geral, no Quadro 01, tem-se a função de

densidade de probabilidade $f(t)$, função de sobrevivência $S(t)$, função taxa de falha λ(t) e função percentil t_p, dos modelos exponencial, Weibull e log-normal, da forma seguinte:

$f(t)$	$S(t)$	$\lambda(t)$	t_p
Exponencial			
$\frac{1}{\alpha}\ exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$	$exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$	$\frac{1}{\alpha}$	$t_p = -\alpha\ log(1-p)$
Weibull			
$\frac{\gamma}{\alpha^\gamma}\ t^{\gamma-1}\ exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}$	$exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}$	$\frac{\gamma}{\alpha^\gamma}\ t^{\gamma-1}$	$t_p = \alpha[log(1-p)]^{1/\gamma}$
Log-normal			
$\frac{1}{\sqrt{2\pi}t\sigma} exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\}$	$\Phi\left(\frac{-log(t)-\mu}{\sigma}\right)$	$\frac{f(t)}{S(t)}$	$t_p = \exp\{z_p\sigma + \mu\}$

Quadro 01. Relação das funções densidade de probabilidade, sobrevivência, taxa de falha e percentil das distribuições exponencial, Weibull e log-normal. Todos as quantidades γ, α, t, no quadro acima, são positivos.

Para os modelos de regressão de sobrevivência citados, os seguintes parâmetros são escritos como função de covariáveis: α = exp(xβ) para a exponencial, α = exp(xβ) e para a Weibull μ = exp(xβ), em que β é o coeficiente da regressão.

4.3 Ajuste do modelo

Para o ajuste do modelo, foi desenvolvido as estimativas das funções de sobrevivência variável tempo usando os modelos exponencial, de Weibull, log-normal e o Kaplan-Meier. Em seguida, comparou-se, graficamente, as funções de sobrevivência estimadas de cada modelo, individualmente, com Kaplan-Meier, conforme descrito em Colosimo e Giolo (2006).
Assim, o "melhor" modelo é aquele cujos pontos da função de sobrevivência estimada estão mais próximos dos valores obtidos pelo estimador de Kaplan-Meier, como demonstrado na Figura 2 abaixo.

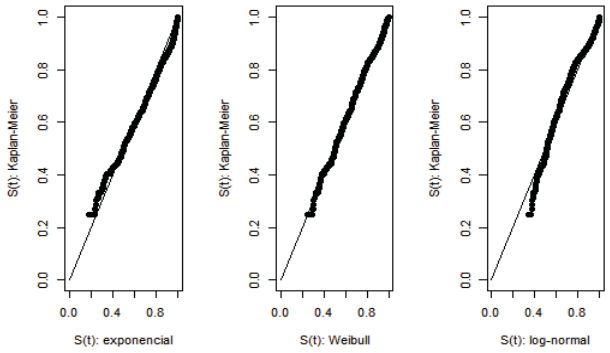

Figura 2.: Gráficos das sobrevivências estimadas por Kaplan-Meier versus as sobrevivências estimadas pelos modelos exponencial, Weibull e log-normal.

Na comparação gráfica da Figura 2, o modelo da distribuição Weibull apresentou o melhor ajuste em comparação com os modelos exponencial e log-normal.
Neste sentido, foi comparado também, as curvas de sobrevivência estimadas pelos modelos exponencial, Weibull e log-normal versus a curva de sobrevivência estimada por Kaplan-Meier. E como previsto, de acordo com a comparação gráfica da Figura 2, de que o modelo Weibull apresenta o melhor ajuste, a Figura 3, também apontou o mesmo indício de que o modelo Weibull é o melhor ajuste.

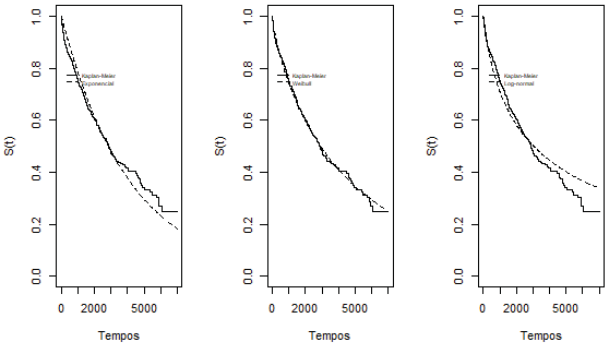

Figura 3.: Curvas de sobrevivência estimadas pelos modelos de exponencial, Weibull e log-normal versus a curva de sobrevivência estimada por Kaplan-Meier.

Como demonstrado no passo anterior, o modelo Weibull indica a melhor adequação aos dados comparados aos outros modelos exponencial e log-normal. Contudo, o presente estudo adotou o modelo de Weibull como ajuste na análise de sobrevivência.
Com as covariáveis significativas, encontradas na seção de Análise Exploratória dos Dados, foi ajustado o modelo de regressão assumindo a distribuição Weibull, como já identificado anteriormente, o melhor ajuste. No modelo em questão, como variáveis significativas foram ajustadas conjuntamente excluindo-se a cada iteração uma única variável, chegando então ao ajuste final, como exposto na Tabela 01.
Adentrando à análise dos resultados, foram escolhidos duas variáveis relevantes para avaliação.
A covariável Classe funcional, importante no ponto de vista clínico, é um instrumento usado para medir (classificar) a extensão da insuficiência cardíaca (NYHA - New York Heart Association) esta classificação identifica o nível de severidade da insuficiência cardíaca do paciente, podendo variar do nível 1 (sem limitações para atividade física) ao nível 4 (limitação acentuada para atividade física).
Em observação ao resultado de regressão, a razão dos tempos medianos de sobrevivência, foi significativo apensas para o nível 4, com 0.475 em comparação ao nível 1, isto implica em que o tempo mediano de sobrevida dos indivíduos que possuem limitação acentuada para atividade física é menor em relação aos indivíduos que não possuem nenhum tipo de limitação.
A seguinte variável analisada, a Fração ejeção, responsável por medir a quantidade do sangue no ventrículo esquerdo que é ejetada a cada batimento cardíaco. A insuficiência cardíaca ocorre quando o fornecimento de sangue (Fração ejeção) é insuficiente

para satisfazer as necessidades do corpo. Com o resultado da Tabela 01, o tempo mediano de sobrevivência dos enfermos com doença de chagas considerando a Fração ejeção é 1.030.

Tabela 01.: Tempo mediano e respectivo intervalo de confiança.

Covariável	Razão	2,5%	97,5%
Constante	1203,5	296,0	4892,2
ClasseFuncional2	0,837	0,607	1,154
ClasseFuncional3	0,726	0,492	1,071
ClasseFuncional4	0,476	0,306	0,741
PAS	1,010	1,002	1,017
FracaoEjecao	1,030	1,016	1,044
RazaoTEI	0,508	0,341	0,759
MP2	0,687	0,488	0,968
VolAEindexado	0,990	0,983	0,997
AneurismaVD2	1,832	1,291	2,599
VDDopplerTissularS	1,066	1,013	1,122
AreaVDsis	1,072	1,031	1,115
AreaVDdias	0,931	0,902	0,960

Tabela 02.: Estimativas dos parâmetros do modelo de regressão Weibull.

Covariável	Estimativa	Erro-Padrão	P-valor
Constante	7,09295	0,71554	< 0.0001
ClasseFuncional2	-0,17787	0,16369	0,277
ClasseFuncional3	-0,32054	0,19841	0,106
ClasseFuncional4	-0,74264	0,22604	0,001
PAS	0,00963	0,00381	0,011
FracaoEjecao	0,02941	0,00705	< 0.0001
RazaoTEI	-0,67637	0,2044	0,001
MP2	-0,37519	0,17505	0,032
VolaEindexado	-0,0103	0,0036	0,004
AneurismaVD2	0,60534	0,17855	0,001
VDDopplerTissularS	0,06406	0,02593	0,014
ÁreaVDsis	0,0696	0,02009	0,001
AreaVDdias	-0,07173	0,01578	< 0.0001
Log(escala)	-0,09522	0,05221	0,068

Paralelamente ao estudo de Filho, (2020), que faz esta mesma análise de sobrevivência utilizando modelo de regressão de cox. Observamos que, apesar do autor ter usado uma técnica não paramétrica, as variáveis significativas para o modelo são praticamente as mesmas encontradas neste estudo, com exceção de MP, PAS e EA.

4.4 Validação do modelo

Para a validação do modelo paramétrico de regressão de sobrevivência é importante a observação do ajuste do modelo com reação à sua adequação aos dados de interesse, para que os resultados sejam julgados como adequados.

Conforme os estudos de Colosimo e Giolo (2006), as técnicas gráficas utilizadas aos resíduos da regressão são muitas utilizadas para validar a distribuição dos erros. Como bem observados em Klein e Moeschberger (1997), deve ser usado principalmente para rejeitar modelos inadequados e não para "provar" que um determinado modelo é correto.

Um resíduo bastante útil para verificação do ajuste de um modelo paramétrico é o resíduo deviance. Este resíduo é preferível aos resíduos Cox-Snell ou martingal por serem mais simétricos e podem também ser usados para identificar a melhor forma funcional para uma dada covariável. Se o modelo for adequado, esperamos que os resíduos se distribuam aleatoriamente em torno de zero. Os resíduos deviance versus cada covariável no modelo são apresentados na Figura 4.

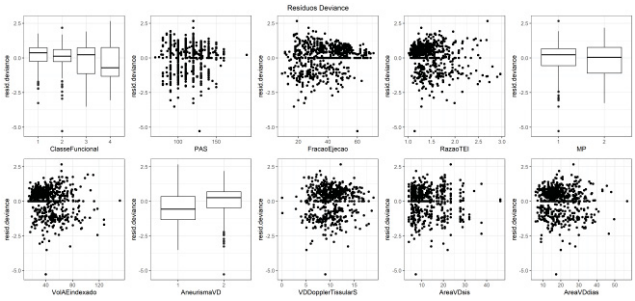

Figura 4.: Resíduos deviance versus cada covariável no modelo.

Exceto por alguns outliers, não observamos desvios consideráveis, indicando que o modelo Weibull apresenta bom ajuste aos dados. Para as variáveis categóricas, a distribuição dos resíduos é comparável entre os níveis. Para as variáveis contínuas, não há, em geral, um claro padrão que aponte problemas com a análise. Concluímos que as variáveis foram modeladas satisfatoriamente.

5 Conclusões

O estudo aqui desenvolvido foi analisar o melhor modelo de regressão paramétrico para análise de sobrevivência com intenção de identificar os efeitos das variáveis de controle na ocorrência do evento de interesse.

Agradecimentos

Agradeço a meus pais que deram suporte ao meu desenvolvimento contínuo e que sempre me apoiaram, à coordenação e corpo docente do Curso de Data Science Big Data da UFPR que se comprometeram com o desenvolvimento e conhecimento de todos, e finalmente agradeço ao meu professor e orientador José Luiz

Padilha Da Silva que teve paciência em me orientar nessa última jornada do curso.

Referências

ALENCAR, M.M. F et al. Epidemiologia da doença de chagas aguda no Brasil de 2007 a 2018. Pesquisa, Sociedade e Desenvolvimento, v. 9, n.10, e8449109120, 2020 (CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/ 10,33448/rsd-v9i10.9120.

COLOSIMO, E. A. ; GIOLO, S. R. Análise de sobrevivência aplicada. Editora Blucher, 2006.

FILHO, D. L. S. Modelo de Regressão de Cox - Uma Aplicação a Dados de Pacientes com a Doença de Chagas. Trabalho de Conclusão de Curso, Especialização em Data Science & Big Data, Universidade Federal do Paraná, Curitiba, nº 01, 01 2020.

KLEIN, J.P.; MOESCHBERGER, M. L. Estatística para biologia e saúde. Stat. Biol. Health, Nova Iorque, v. 27238, 1997.

WHITE, I.R.; ROYSTON, P. Imputing faltando valores covariados para o modelo Cox. Stat Med. 2009 Jul 10;28(15):1982-98. doi: 10.1002/sim.3618. PMID: 19452569; PMCID: PMC2998703.

Brasil, Ministério da Saúde. Banco de dados do Sistema Único de Saúde-DATASUS. Disponível em https://datasus.saude.gov.br/. Acessado em 10 de março de 2022.