UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Exatas

Programa de Pós-graduação em Estatística

LARISSA BRANDÃO ROCHA MARTINEZ FERNANDEZ

Aplicação do Pacote *survstan* em R para Modelagem de Dados de Sobrevivência: Uma Comparação com o Pacote *survival*

Belo Horizonte

2023

LARISSA BRANDÃO ROCHA MARTINEZ FERNANDEZ

Aplicação do Pacote *survstan* em R para Modelagem de Dados de Sobrevivência: Uma Comparação com o Pacote *survival*

Monografia de Especialização apresentado ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística

Orientador: Prof. Dr. Fabio Nogueira Demarqui

Belo Horizonte

2023

Fernandez, Larissa Brandão Rocha Martinez.

F363a Aplicação do pacote survstan em R para modelagem de dados de sobrevivência [recurso eletrônico]: uma comparação com o pacote survival / Larissa Brandão Rocha Martinez Fernandez — 2023.
1 recurso online (27 f. il, color).

Orientador: Fabio Nogueira Demarqui
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística
Referências: 26-27.

1. Estatística. 2. Confiabilidade (Probabilidades). I. Demarqui, Fabio Nogueira. II. Universidade Federal de Minas Gerais. I. Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 - Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais **E-mail: pgest@ufmg.br**
Instituto de Ciências Exatas Tel: 3409-5923 – FAX:

Departamento de Estatística
P Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

ATA DO 322ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE LARISSA BRANDÃO ROCHA MARTINEZ FERNANDEZ.

Aos quatorze dias do mês de dezembro de 2023, às 16:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Larissa Brandão Rocha Martinez Fernandez**, intitulado: "Aplicação do Pacote survstan em R para Modelagem de Dados de Sobrevivência: Uma Comparação com o Pacote survival", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Fábio Nogueira Demarqui – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 14 de dezembro de 2023.

gov.br Documento assinado digitalmente
FABIO NOGUEIRA DEMARQUI
Data: 15/12/2023 08:28:58-0300
Verifique em https://validar.iti.gov.br

Prof. Fábio Nogueira Demarqui (Orientador)
DEST/ICEX/UFMG

gov.br Documento assinado digitalmente
MARISTELA DIAS DE OLIVEIRA
Data: 18/12/2023 10:09:16-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Maristela Dias de Oliveira
Departamento de Estatística - UFBA

RESUMO

As análises de confiabilidade são de grande relevância na engenharia, sendo uma forma de avaliar e predizer a ocorrência de eventos indesejados e, então, embasar a melhoria de processos e produtos visando a maior eficiência nos resultados. Para modelar o comportamento das falhas de um sistema, se faz necessário selecionar uma distribuição de probabilidade para os tempos de falha, além do modelo de regressão que melhor explica a influência de covariáveis. O *software* R é uma das ferramentas que podem auxiliar o processo de análise, que de outra forma envolveria cálculos complexos. Em vista disso, este trabalho busca demonstrar a aplicação do *survstan*, um pacote de funções em R que visa fornecer meios para o ajuste de modelos de sobrevivência. Para esse fim, um conjunto de dados retirado da literatura serviu de entrada para as funções do pacote *survstan*, que foi utilizado para a seleção da distribuição de probabilidade mais adequada, além do ajuste do modelo de regressão de Tempos de Falha Acelerados e obtenção dos parâmetros de interesse dos modelos. A funcionalidade do pacote para executar estas funções foi comparada com o pacote *survival*, um dos pacotes mais utilizados em R para modelagem de dados de sobrevivência. Na comparação, o pacote *survstan* apresentou resultados consistentes com os obtidos através do pacote *survival*, com menos esforço empregado na interpretação dos resultados.

Palavras-chave: Confiabilidade, Survival, Survstan

ABSTRACT

Reliability analyses are greatly relevant in engineering, being a way of evaluating and predicting the occurrence of unwanted events and then, support processes and products improvement, aiming for greater efficiency in results. To model the behaviour of a system's failures, it is needed to select a probability distribution for failure times, in addition to the best regression model that will explain the influence of covariates. R software is one of the tools that can support the analysis process, which would otherwise involve complex calculations. In view of this, this work seeks to demonstrate the application of survstan, a R package of functions that aims to provide a toolkit for fitting survival models. For this purpose, a dataset extracted from literature was fed as input to survstan functions, which were used to select the appropriate probability distribution, as well as fitting the Accelerated Failure Times regression model and obtaining models' parameters of interest. The functionality of the package to perform these functions was compared with the survival package, one of the most used in R for survival analysis tasks. In comparison, survstan presented consistent results with those obtained through survival, with less effort applied in writing code and interpreting results.

Key words: Reliability, Survival Analysis, Survstan

SUMÁRIO

1 INTRODUÇÃO

Os estudos de confiabilidade são essenciais para avaliar e predizer o desempenho de produtos e/ou processos, e podem ser utilizados para embasar alterações de projeto, seleção de fornecedores, períodos e custos de garantia, bem como mudanças em estratégias de manutenção.

Uma forma de definir confiabilidade é como a probabilidade de que um sistema, equipamento ou objeto desempenhe sua função, sob condições predeterminadas, por um período determinado de tempo. (Meeker e Escobar,1998). Segundo Lafraia, 2001, a análise de confiabilidade é caracterizada pela avaliação de risco de falha de um sistema ou produto, sendo a falha uma “situação indesejável”.

Uma definição mais ampla pode ser aplicada. Os métodos estatísticos utilizados nas análises de engenharia de confiabilidade são análogos a estudos de outros campos, como a medicina. Neste caso, é possível generalizar o conceito de “falha” para qualquer evento de interesse definido, como por exemplo, a morte de um paciente, e os dados de tempo até a ocorrência deste evento, desde uma origem definida, são considerados os “dados de vida”. Nas ciências médicas, o termo mais comum para estes estudos é “Análise de sobrevivência” (Legrand, 2021). A definição de tempo também pode ser ampliada, restrita apenas às condições de ser uma variável aleatória, contínua e que assuma valores a partir de 0 (Legrand, 2021). Podem ser utilizadas unidades como quilômetros ou ciclos de operação, por exemplo, a depender da situação que está sendo definida.

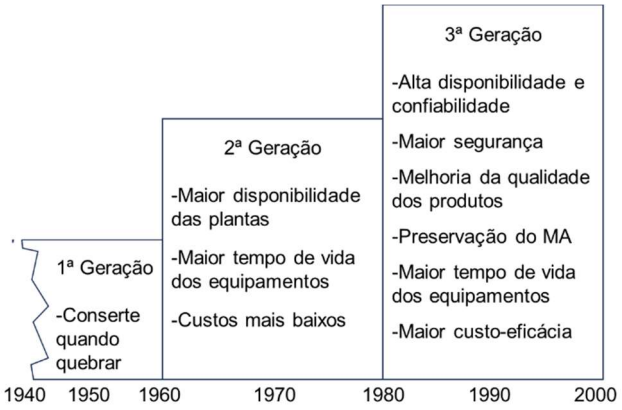
Uma das principais características dos dados de tempo de falha é a presença de censura, ou seja, situações em que a observação não foi feita por tempo suficiente para que o tempo de falha seja conhecido com exatidão (Kleinbaum, David G., e Mitchel Klein, 2012). Os métodos de análise de confiabilidade foram desenvolvidos especificamente para lidar com a presença de dados censurados, já que a utilização das técnicas estatísticas clássicas não é adequada nestas situações (Colosimo e Giolo, 2006).

Rocha, 2019, cita como marco histórico para os estudos de confiabilidade a publicação em 1939 do estudo de vida de rolamentos feito pelo engenheiro e matemático Waloddi Weibull, que mais tarde se tornaria um dos principais nomes do campo de confiabilidade, com sua distribuição de probabilidades conhecida como

“distribuição de Weibull”. Denson, 1998, define a década de 1950 como o período de inauguração da disciplina de Engenharia de confiabilidade, como um dos desdobramentos da 2ª Guerra Mundial.

Entretanto, segundo Lafraia, 2001, apenas a partir de 1980, com a 3ª geração da manutenção, foi possível observar a implantação das técnicas de análise de confiabilidade nos setores de engenharia em países de primeiro mundo, e no Brasil, nos setores de telecomunicações, elétrico, armamentista e nuclear.

Figura 1 – Gerações da Manutenção



Fonte: Adaptado de Lafraia (2001).

Com o crescimento do campo de aplicação e o desenvolvimento tecnológico e informacional, hoje é possível encontrar diversos *softwares* específicos para a elaboração de estudos de confiabilidade. Enquanto a utilização destes *softwares* é um facilitador do ponto de vista do engenheiro, que terá o apoio de algoritmos prontos para a execução de cálculos complexos, a maioria dos *softwares* especializados exigem a aquisição de licenças específicas, o que os torna menos acessíveis. Um outro aspecto a ser considerado é que as premissas e métodos de análise utilizados na construção destes sistemas não são personalizáveis pelo usuário. Isto posto, o *software* R (R Core Team, 2023) tem sido largamente utilizado como uma alternativa viável, por ser um ambiente de programação de uso gratuito, linguagem simples e de aplicação mais flexível, visto que dispõe de diversos pacotes de funções que podem ser utilizadas para as análises de sobrevivência, além de oferecer a possibilidade de criação de funções específicas.

Neste contexto, um dos pacotes de funções mais tradicionais presentes no R é o pacote *survival* (Therneau, 2023), cujo desenvolvimento foi iniciado em 1985,

antes mesmo de estar disponível no R, para o qual migrou anos depois (Therneau e Lumley, 2023). Suas funções são amplamente utilizadas, sendo referenciado em diversos estudos como ferramenta de análise e indicado na página do CRAN (The Comprehensive R Archive Network) como um dos pacotes relevantes no tema de análises de sobrevivência.

1.1 Objetivo

Este trabalho tem como objetivo demonstrar a aplicação do pacote *survstan* (Demarqui, 2023) como ferramenta para análises de confiabilidade em R, no contexto de modelos de tempos de falha acelerados.

1.1.1 Objetivos específicos

É possível apontar como objetivos específicos:

- Apresentar algumas das distribuições de probabilidade de utilização mais frequente em estudos de confiabilidade;
- Descrever o modelo de regressão de tempos de falha acelerados disponível no pacote *survstan*;
- Apresentar algumas vantagens do pacote *survstan* em relação ao pacote *survival* no que tange a usabilidade.

2 MODELOS PARAMÉTRICOS

Seja T uma variável aleatória que descreve o tempo até a falha de um equipamento. É possível definir:

- A função de distribuição acumulada, ou seja, a probabilidade de haver um evento (falha) até o tempo t, como:

$$F(t) = P(T \leq t) \tag{2.1}$$

- A função de sobrevivência, ou a probabilidade de não haver falha na observação até um tempo t, pode ser descrita como:

$$S(t) = P(T > t) = 1 - F(t) \tag{2.2}$$

- A função taxa de falha, como sendo a probabilidade de o elemento falhar até o tempo $t + \Delta t$, dado que não sofreu falha até o tempo t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$
$$h(t) = \frac{f(t)}{S(t)} \tag{2.3}$$

Ao assumir diferentes distribuições para a variável T, é possível obter as funções taxa de falha e de sobrevivência através das relações descritas nas equações de 2.1 a 2.4

Assumir uma distribuição de probabilidade específica para T permite aumentar a precisão sobre as inferências das funções taxa de falha e sobrevivência (Collett, 2023). Os modelos de regressão em que uma distribuição de probabilidade é especificada são chamados de modelos paramétricos. Algumas das distribuições mais utilizadas para modelagem de dados de sobrevivência são descritas neste capítulo.

2.1 Distribuição Exponencial

Nesta distribuição em que a taxa de falha é considerada constante e igual ao parâmetro λ, a função densidade de probabilidade f(t) e a função de sobrevivência S(t) são dadas pelas equações (2.4) e (2.5):

$$f(t|\lambda) = \lambda * e^{-\lambda t} \tag{2.4}$$

$$S(t|\lambda) = e^{-\lambda t} \quad (2.5)$$

A distribuição exponencial apresenta a função h(t) = λ, ou seja, a taxa de falha é constante a qualquer momento da vida do objeto estudado (Colosimo e Giolo, 2006).

2.2 Distribuição de Weibull

A distribuição proposta por Weibull é, na prática, a mais aplicada para modelagem de dados em análises de confiabilidade e sua forma mais utilizada baseia-se em dois parâmetros α>0 (parâmetro de forma) e γ>0 (parâmetro de escala). O equacionamento da probabilidade de falha, função de sobrevivência e taxa de falha estão descritas em (2.6),(2.7) e (2.8).

$$f(t|\alpha,\gamma) = \frac{\alpha}{\gamma^{\alpha}} t^{\alpha-1} e^{(-(\frac{t}{\gamma})^{\alpha})} \quad (2.6)$$

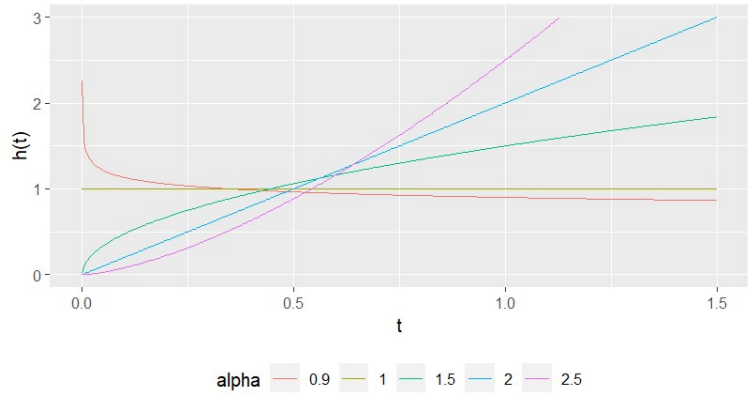
$$S(t|\alpha,\gamma) = e^{(-(\frac{t}{\gamma})^{\alpha})} \quad (2.7)$$

$$h(t|\alpha,\gamma) = \frac{\alpha}{\gamma^{\alpha}} t^{\alpha-1} \quad (2.8)$$

A função taxa de falha na distribuição de Weibull é caracterizada por ser uma função monótona. Esta função é decrescente quando o parâmetro de forma α é menor que 1 e crescente quando α>1, podendo assumir uma forma côncava quando 1<α<2, linear para α=2 e convexa quando α>2. Para α=1, h(t) é constante e igual a 1/γ, tal qual apresentado na distribuição exponencial. Por isto, a exponencial é um caso

particular da distribuição de Weibull (Colosimo e Giolo, 2006). É possível ver as formas que a função h(t) apresenta para diferentes valores de α na Figura 2.

Figura 2 – Função taxa de falha da distribuição Weibull.



Fonte: Autora (2023)

2.3 Distribuição Lognormal

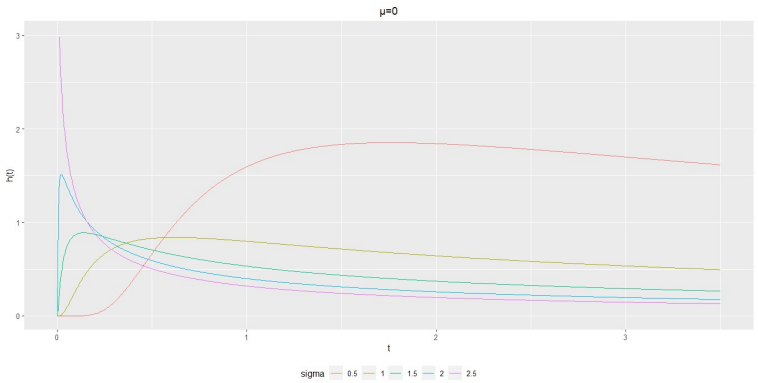
A distribuição lognormal, assim como a de Weibull, também considera uma variável aleatória t que assume apenas valores positivos, sendo também utilizada em análises de confiabilidade. Esta distribuição tem o parâmetro de localização μ e parâmetro de escala σ. Seja ϕ a função de distribuição acumulada de uma variável aleatória que segue uma distribuição normal padrão, as funções de distribuição da probabilidade de falha e função de sobrevivência da distribuição Lognormal estão descritas em (2.9) e (2.10).

$$f(t|\mu,\sigma) = \frac{1}{\sqrt{2\pi}t\sigma} e^{\left(-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right)} \tag{2.9}$$

$$S(t|\mu,\sigma) = \phi\left(\frac{-\log(t)+\mu}{\sigma}\right) \tag{2.10}$$

A função taxa de falha h(t) não possui uma forma analítica explícita para a distribuição Lognormal, porém pode ser encontrada pela relação mostrada na equação (2.3). Diferentemente do que ocorre com a distribuição Weibull, as curvas de h(t) para a Lognormal não são sempre monótonas, e sua forma varia a depender do parâmetro de escala σ. Para os casos de distribuição Lognormal com parâmetro σ≤1, a função taxa de falha é monótona decrescente. Quando σ>1, a função apresenta uma forma unimodal.

Figura 3 – Função taxa de falha da distribuição Lognormal.



Fonte: Autora (2023)

2.4 Distribuição Loglogística

A variável com distribuição log-logística possui uma função de distribuição de probabilidade definida em (2.11), parametrizada por α>0 e γ>0 sendo parâmetros de forma e escala, respectivamente.

$$f(t|\alpha,\gamma) = \frac{\frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}}{\left(1+\left(\frac{t}{\gamma}\right)^{\alpha}\right)^{2}} \tag{2.11}$$

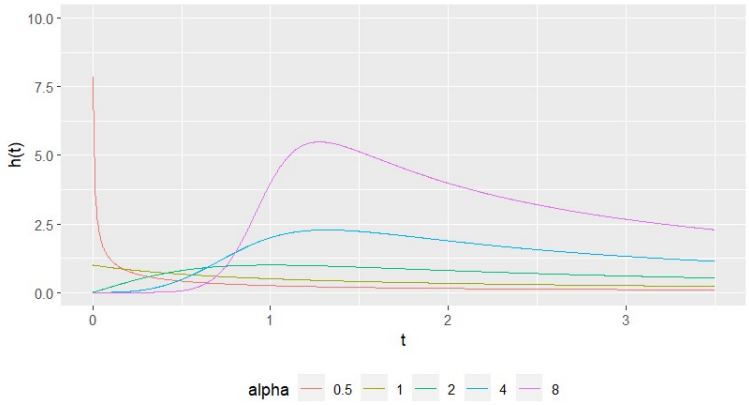
$$S(t|\alpha,\gamma) = \frac{1}{1+\left(\frac{t}{\gamma}\right)^{\alpha}} \tag{2.12}$$

$$h(t|\alpha,\gamma) = \frac{\frac{\alpha}{\gamma}\left(\frac{t}{\gamma}\right)^{\alpha-1}}{1+\left(\frac{t}{\gamma}\right)^{\alpha}} \tag{2.13}$$

É possível ver na Figura 4 que a forma da função taxa de falha do modelo loglogístico apresenta um comportamento estritamente decrescente quando o

parâmetro de forma α assume valores menores ou iguais a 1. Quando α>1, a curva apresenta-se unimodal.

Figura 4 – Função taxa de falha da distribuição Log-logística.



Fonte: Autora (2023)

2.5 Distribuição de Birnbaum-Saunders

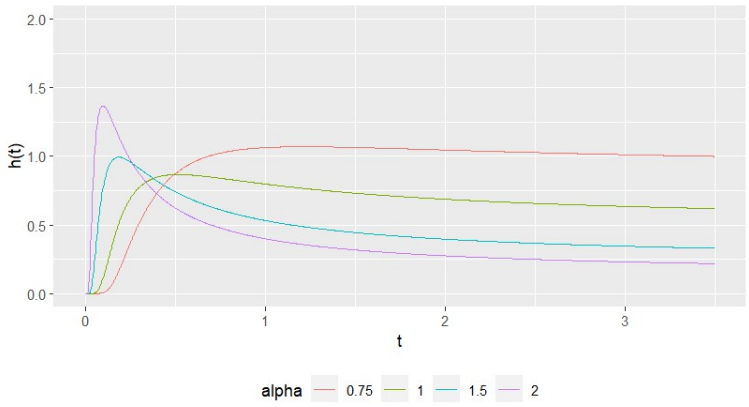
A distribuição de Birnbaum-Saunders originou-se através do estudo de vida sob fadiga componentes mecânicos, sendo esta sua aplicação mais conhecida. Entretanto, seu campo de aplicações hoje se estende para além das análises de engenharia, englobando outras situações em que se considera um dano cumulativo excedendo um certo patamar crítico, como o impacto de fatores de risco para doença cardíaca crônica, acumulação de toxinas no corpo humano e até mesmo ocorrência de desastres naturais (Leiva, 2016). Esta distribuição é assimétrica, unimodal e também possui parâmetros de forma α e escala γ.

$$f(t|\alpha,\gamma) = \frac{\sqrt{\frac{t}{\gamma}} + \sqrt{\frac{\gamma}{t}}}{2\alpha t}\phi\left(\sqrt{\frac{t}{\gamma}} + \sqrt{\frac{\gamma}{t}}\right)(t) \qquad (2.14)$$

$$S(t|\alpha,\gamma) = \phi\left(\sqrt{\frac{t}{\gamma}} + \sqrt{\frac{\gamma}{t}}\right)(t) \qquad (2.15)$$

Assim como na distribuição Lognormal, não possui uma expressão fixa para a função taxa de falha, devendo ser utilizada a relação expressa em (2.3). Na Figura 5, é possível notar que a função taxa de falha assume uma forma unimodal para qualquer valor do parâmetro de forma α (Leiva, 2016). Contudo, quanto maior o valor de α, observa-se que o valor máximo de h(t) ocorre em menos tempo.

Figura 5 – Função taxa de falha da distribuição de Birnbaum-Saunders.



Fonte: Autora (2023)

3 MODELOS DE REGRESSÃO

Modelos de regressão são métodos que permitem analisar covariáveis independentes de forma simultânea, sendo especialmente úteis para estudos em que se faz necessário analisar a influência de diferentes fatores nas características da distribuição de probabilidade de falha. (Legrand, 2021).

A escolha do modelo de regressão define como as covariáveis irão influenciar a distribuição “linha de base”. Neste trabalho, serão analisados os modelos de tempo de falha acelerado, dada sua grande aplicabilidade em problemas de engenharia. Outros modelos de regressão populares na literatura de análises de sobrevivência e confiabilidade, mas que estão fora do escopo deste trabalho, são o modelo de Cox (riscos proporcionais) e o modelo de chances proporcionais (Klein, John P., et al., 2014).

3.1 Modelos de tempos de falha acelerados

A premissa assumida nos modelos de tempo de falha acelerados é de que o efeito das covariáveis guarda proporcionalidade em relação aos tempos de sobrevivência. Desta forma, o modelo descreve como as covariáveis interferem de forma a alongar ou contrair o tempo até a ocorrência da falha (Kleinbaum, David G., e Mitchel Klein, 2012). Em testes de vida acelerados, parâmetros como tensão aplicada, temperatura, umidade ou frequência de operação são extrapolados em relação aos parâmetros normais de operação, visando obter de forma rápida resultados sobre a confiabilidade de um produto manufaturado (Meeker e Escobar,1998).

3.1.1 Equacionamento usado no pacote *survival*

O pacote *survival* permite o ajuste de modelos de tempos de falha acelerados através da função “survreg”. Nesta função, o modelo utilizado é baseado na escala logarítmica do tempo, conforme (3.1), em que ε é um erro aleatório que pode assumir diferentes distribuições padrão e σ>0 é um parâmetro de escala.

$$Y = \log T = x\beta + \sigma\epsilon \tag{3.1}$$

A depender da distribuição escolhida para ε, a distribuição para T será alterada. Alguns exemplos estão dispostos na tabela

Tabela 1: Relação entre distribuição para ε e distribuição para T, com forma de cálculo dos parâmetros de T.

Distribuição para ε	Distribuição para T	Parâmetros
Valor extremo padrão com σ≠1	Weibull	$\alpha = 1/\sigma$ $\gamma = e^{x\beta}$
Logística padrão	Loglogística	$\alpha = 1/\sigma$ $\gamma = e^{x\beta}$
Normal padrão	Lognormal	$\sigma = \sigma$ $\mu = x\beta$
Valor extremo padrão com σ=1	Exponencial	$\lambda = \frac{1}{e^{x\beta}}$

Fonte: Autora (2023)

3.1.2 Equacionamento usado no pacote *survstan*

Os modelos de tempo de falha acelerados foram implementados no pacote *survstan* conforme a equação (3.2), sendo ν uma variável aleatória com distribuição basal escolhida para representar o tempo até a falha e com função de sobrevivência S_0 (ver Seção 2).

$$T = e^{x\beta}\nu \tag{3.2}$$

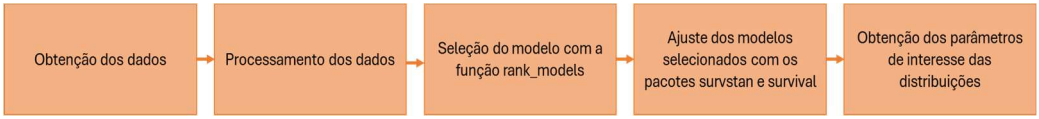
$$S(t) = S_0(te^{-x\beta}) \tag{3.3}$$

A principal diferença entre os dois modelos é a escala do tempo. No pacote *survstan*, o modelo ajustado considera a escala original do tempo, em vez de utilizar a escala logarítmica. Esta modificação torna o modelo resultante mais simples de analisar.

4 METODOLOGIA

Este capítulo trata da metodologia empregada neste estudo, a qual está esquematizada de forma resumida no fluxograma apresentado na Figura 6. Serão descritos os dados utilizados e a principal ferramenta de análise, o pacote *survstan*.

Figura 6 – Etapas do trabalho



Fonte: Autora (2023)

4.1 Os dados

Neste trabalho, as funções disponibilizadas no pacote *survstan* serão demonstradas utilizando uma base de dados apresentada por Freitas e Colosimo (1997), que contém resultados de um experimento feito com o objetivo de estudar o tempo de vida de um componente mecânico, sob a consideração de que a variável "temperatura" atua como fator de aceleração para a ocorrência de falha. O experimento foi realizado em 40 amostras, e cada 10 amostras foram expostas a diferentes temperaturas medidas em Kelvin: 300, 350, 400 e 500, sendo 300K a temperatura de operação do componente.

Meeker e Escobar (1998) apresentam a relação tempo x aceleração da reação química proposta por Arrhenius como sendo um modelo amplamente utilizado para descrever os efeitos da temperatura numa determinada reação química. A equação está definida na equação (4.1, em que $R(x^*)$ é a taxa de reação química, x^* é a temperatura em Kelvin, γ_0 e E_a são constantes específicas de cada material e k_b a constante de Boltzmann.

$$R(x^*) = \gamma_0 e^{\left(\frac{-E_a}{k_b * x^*}\right)} \tag{4.1}$$

Sob a assunção de que a falha do componente acontecerá quando a reação química alcançar um nível crítico N_c, a situação analisada pode ser representada pela equação (4.2) em que T é o tempo até a falha.

$$N_c = T * R(x^*) \tag{4.2}$$

Substituindo-se a relação de Arrhenius (4.1) na expressão (4.2), é possível definir o tempo T pela equação (4.3), que também pode ser reescrita na escala logarítmica, como em (4.4). A partir deste equacionamento, é possível ver que será considerada a influência da temperatura de forma inversamente proporcional ($^{1}/_{x^*}$), por isto é necessário fazer a transformação da variável para o seu inverso, ou seja, a covariável presente no modelo é 1/Temperatura.

$$T = \frac{N_c}{\gamma_0} e^{\left(\frac{-E_a}{k_b * x^*}\right)} \tag{4.3}$$

$$\log T = \log \frac{N_c}{\gamma_0} + \frac{-E_a}{k_b} * \frac{1}{x^*} \tag{4.4}$$

4.2 O pacote em R

O pacote *survstan*, em linguagem R, fornece ferramentas para modelagem de dados de sobrevivência, podendo ser utilizado para ajustar dados de vida com censura à direita. Neste pacote, foram implementadas as funções necessárias para ajustar dados às principais distribuições citadas na Seção 2, combinadas com cinco modelos de regressão diferentes. Além do modelo de tempos de falha acelerados citado em 0, o pacote disponibiliza também:

- Modelos de riscos proporcionais;
- Modelos de razão de chances (*odds*) proporcionais;
- Modelos de taxas de falha aceleradas;
- Modelos de Yang e Prentice.

Diferentemente da função “survreg*”* do pacote *survival*, uma das principais bibliotecas em R para análises de sobrevivência, que opera na escala logarítmica, o pacote *survstan* foi implementado utilizando a escala de tempo. Essa abordagem uniformiza a entrada e saída de dados para todos os modelos de regressão implementados, além de simplificar a análise dos resultados, pois os modelos ajustados não possuem intercepto, resultando em saídas que correspondem diretamente aos parâmetros de interesse.

Para ajustar um modelo tempos de falha acelerados, a função utilizada está descrita em (4.5), em que os parâmetros podem ser definidos como:

$$\text{aftreg(formula,data,baseline=”weibull”,dist=NULL,init=0,...)} \quad (4.5)$$

- “formula”: uma descrição do modelo a ser ajustado, por exemplo: Surv(tempo, status) ~ covariável
- “data”: conjunto de dados, geralmente um *dataframe*, em que as variáveis descritas em “formula” estão contidas
- “baseline”: a distribuição basal escolhida para o ajuste dos dados. As distribuições disponíveis no pacote estão descritas em 2.
- “dist”: uma forma alternativa para definição da distribuição basal, disponibilizada para permitir compatibilidade com a função “survreg” do pacote *survival*
- “Init”: especificação do valor inicial para os parâmetros de otimização (parâmetro utilizado na função “optimizing” do pacote rstan.
- (...): outros argumentos que podem ser passados para as funções que estão embutidas na função “aftreg”.

O pacote *survstan* oferece também a função *rank_models*, que ajusta o conjunto de dados de entrada em todas as distribuições de linha de base disponíveis e ordena os modelos gerados segundo o Critério de Informação de Akaike (AIC), que é uma métrica que permite comparar modelos diferentes de forma objetiva.

A expressão que define o AIC para um dado modelo está descrita em (4.6), em que $\hat{L}$ é o valor máximo da função de verossimilhança, e k é a quantidade de parâmetros do modelo (Legrand, 2021).

Segundo Colosimo e Giolo (2006), no processo de estimação de parâmetros dos modelos, o método de máxima verossimilhança é apropriado para estudos de tempos de falha, visto que é capaz de incorporar informações de dados censurados.

$$AIC = -2\ln\hat{L} + 2k \quad (4.6)$$

A interpretação do resultado deve ser feita de forma comparativa, com o modelo de menor AIC sendo o preferível para representar os dados. O valor absoluto de cada AIC não tem significado relevante em relação à qualidade do modelo. (Portet, 2020).

Neste trabalho, a função “rank_models” será aplicada ao conjunto de dados, permitindo a seleção da distribuição que gera o melhor resultado. Então, os mesmos

dados serão fornecidos à função “aftreg” do pacote *survstan* e à função “survreg” do pacote *survival,* para que seus resultados sejam interpretados e comparados.

5 RESULTADOS E DISCUSSÕES

5.1 Função “rank_models” – seleção da linha de base

A função “rank_models” foi utilizada para o conjunto de dados com as variáveis de tempo, inverso da temperatura e a variável “status” que indica se houve ou não falha em cada observação.

```
rank_models(
  formula = Surv(tempo, status)~invTemp,
  data = dados_1,
  survreg = "aftreg",
  baseline = c("exponential", "weibull", "lognormal", "loglogistic", "fatigue"))
```

O resultado obtido está exposto na Tabela 2. O termo “fatigue” se refere à distribuição de Birnbaum-Saunders.

Tabela 2 – Resultado da função rank_models

Ordem	*Baseline*	Nº Parâmetros	AIC
1	*weibull*	3	228,341
2	*loglogistic*	3	228,411
3	*lognormal*	3	229,394
4	*fatigue*	3	230,090
5	*exponential*	2	241,443

Fonte: Autora (2023)

É possível perceber que a distribuição com melhor resultado no ajuste dos dados é a de Weibull. Entretanto, o AIC da log-logística, em 2º lugar, se aproxima bastante do 1º lugar. Portanto, esta distribuição será incluída também na etapa de análise detalhada dos modelos de tempo de falha acelerados, para efeito comparativo.

5.2 Função “survreg” – utilizando o pacote *survival*

A Figura 7 mostra a aplicação da função “survreg” no conjunto de dados de estudo.

Figura 7 – Entrada de dados na função “survreg”

```
survreg_w <- survreg(Surv(tempo, status) ~ invTemp, data = dados_1, dist = "weibull")
survreg_ll <- survreg(Surv(tempo, status) ~ invTemp, data = dados_1, dist = "loglogistic")
```

Fonte: Autora (2023)

O resultado gerado está retratado na Figura 8, para a distribuição de Weibull, e Figura 9, para a log-logística.

Figura 8 – Resultado da função survreg aplicando a distribuição de Weibull

```
Call:
survreg(formula = Surv(tempo, status) ~ invTemp, data = dados_1,
    dist = "weibull")
             Value Std. Error     z       p
(Intercept)   3.191      0.528  6.05 1.5e-09
invTemp     589.680    216.711  2.72  0.0065
Log(scale)   -0.920      0.199 -4.62 3.9e-06

Scale= 0.398
```

Fonte: Autora (2023)

Figura 9 – Resultado da função survreg aplicando a distribuição log-logística

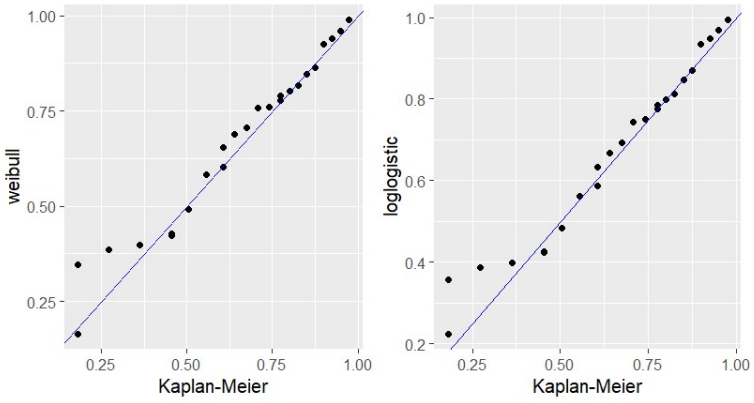
```
Call:
survreg(formula = Surv(tempo, status) ~ invTemp, data = dados_1,
    dist = "loglogistic")
             Value Std. Error     z       p
(Intercept)   3.003      0.557  5.39 6.9e-08
invTemp     600.794    218.410  2.75  0.0059
Log(scale)   -1.130      0.196 -5.77 7.7e-09

Scale= 0.323
```

Fonte: Autora (2023)

Após a obtenção do modelo de regressão, é necessária a criação de uma função que gere os gráficos para possibilitar a análise de resíduos. Foi implementada uma função que gera, para um determinado modelo, o gráfico dos resíduos de Cox-Snell.

Figura 10 – Gráficos para análise de adequação dos modelos ajustados com a função "survreg"



Fonte: Autora (2023)

Em ambas as distribuições, é possível ver na Figura 10 que os modelos tiveram um bom ajuste em boa parte do gráfico, apresentando, contudo, uma região de discrepância. Na análise gráfica, não é perceptível uma diferença na qualidade de ajuste entre os modelos da Weibull e Log-logística.

O resultado da função “survreg” não apresenta de forma direta os parâmetros de interesse da distribuição basal utilizada. Para obtê-los, é necessário fazer alguns cálculos que irão variar de acordo com cada distribuição (ver 3.1.1). Os parâmetros para cada curva estão definidos na seção 2. Os valores extraídos dos modelos ajustados estão descritos na Tabela 3.

Tabela 3 – Parâmetros das curvas de Weibull e Log-logística obtidas na função “survreg”

Distribuição	**α**	**γ**
Weibull	2,51	24,31
Log-logística	3,09	20,14

Fonte: Autora (2023)

5.3 - Função “aftreg” – utilizando o pacote *survstan*

A Figura 11 mostra como a função “aftreg” foi utilizada para modelar os mesmos dados e mesmas *baselines* do item 5.2.

Figura 11 - Entrada de dados na função “aftreg”

```
fit1_w <- aftreg(Surv(tempo, status) ~ invTemp, data = dados_1, dist = "weibull")
fit1_ll <-aftreg(Surv(tempo, status) ~ invTemp, data = dados_1, dist = "loglogistic")
```

Os resultados dos modelos gerados são vistos na Figura 12 e Figura 13.

Figura 12 - Resultado da função aftreg aplicando a distribuição de Weibull

```
Call:
aftreg(formula = Surv(tempo, status) ~ invTemp, data = dados_1,
    dist = "weibull")

Accelerated failure time model fit with weibull baseline distribution

Regression coefficients:
        Estimate StdErr z.value  p.value
invTemp   589.66 216.70  2.7211 0.006506 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Baseline parameters:
       estimate        se      2.5%    97.5%
alpha   2.50986   0.19925   2.11933   2.9004
gamma  24.30762  52.77662 -79.13264 127.7479
```

Fonte: Autora (2023)

Figura 13 – Resultado da função aftreg aplicando a distribuição Log-logística

```
Call:
aftreg(formula = Surv(tempo, status) ~ invTemp, data = dados_1,
    dist = "loglogistic")

Accelerated failure time model fit with loglogistic baseline distribution

Regression coefficients:
        Estimate StdErr z.value  p.value
invTemp   600.77 218.41  2.7507 0.005947 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Baseline parameters:
        estimate         se       2.5%   97.5%
alpha  3.0956750  0.1957126  2.7120853  3.4793
gamma 20.1450984  0.0055665 20.1341884 20.1560
---
```

Fonte: Autora (2023)

É possível observar que o resultado da função “aftreg” fornece diretamente os parâmetros de interesse das distribuições basais utilizadas no modelo de regressão, além dos intervalos de confiança. Os parâmetros obtidos estão apresentados na Tabela 4.

Tabela 4 - Parâmetros das curvas de Weibull e Log-logística obtidas na função “aftreg”

Distribuição	**α**	**γ**
Weibull	2,51	24,31
Log-logística	3,09	20,14

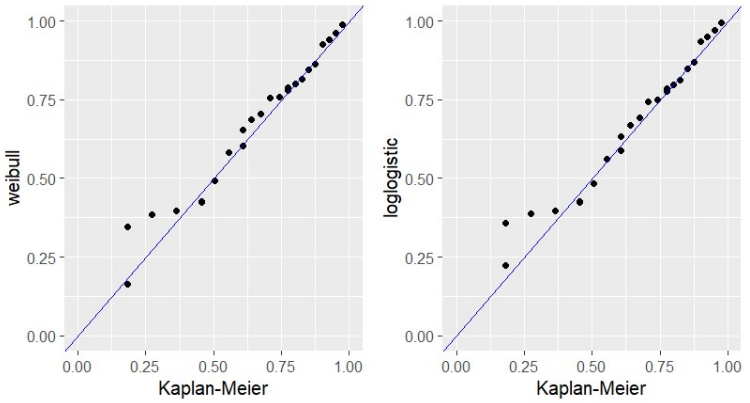
Fonte: Autora (2023)

Para a análise de adequação do modelo gerado, o pacote *survstan* fornece a função “ggresiduals”, que gera 3 tipos de gráficos:

- Resíduos de Cox-Snell – quanto mais próximos os pontos estiverem de uma reta, melhor o ajuste do modelo;
- Resíduos de Martingale, para avaliação da forma funcional da covariável – se a curva apresentada não estiver linear, pode ser necessária uma transformação da variável;
- Resíduos Deviance, para verificação de possíveis outliers – é esperado um comportamento aleatório dos resíduos em torno de zero (Colosimo e Giolo, 2006)

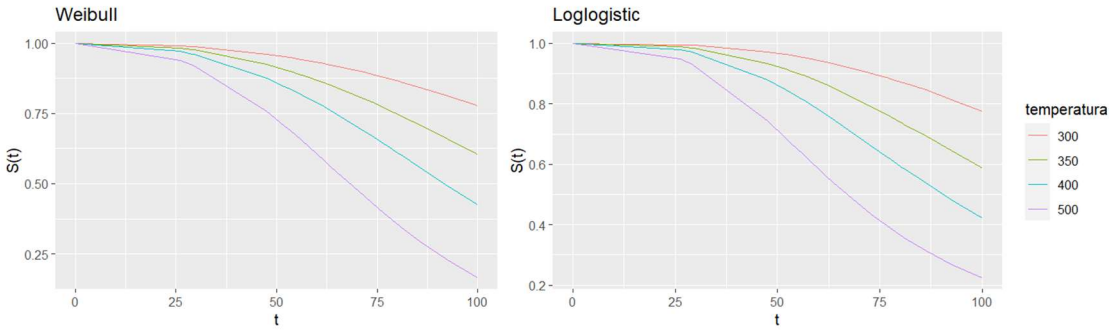
Os gráficos obtidos pela função “ggresiduals” estão apresentados na Figura 14 para os modelos de Weibull e Log-logístico, e se assemelham ao resultado visto na Figura 10.

Figura 14 – Gráficos de resíduos de Cox-Snell gerados pela função “ggresiduals” para o modelo de Weibull



Fonte: Autora (2023)

Após obter os parâmetros da distribuição, conforme Tabela 4, é possível plotar a função de sobrevivência S(t) para todos os níveis da variável “Temperatura”. Nos dois casos, as curvas apresentam comportamento similar, e é possível ver a relação inversa da temperatura com a probabilidade de sobrevivência. Usando como exemplo a curva de Weibull, S(t=100) é aproximadamente 0,8 na temperatura de 300K, enquanto para a temperatura de 500K esta probabilidade é menor que 0,2.



6 CONCLUSÕES

Este trabalho demonstra a viabilidade de utilização do pacote *survstan* para realização de análises de confiabilidade com modelos de tempos de falha acelerados, obtendo resultados consistentes em relação à referência utilizada, a função “survreg” do pacote *survival,* com a necessidade de utilização de menos linhas de código. Verificou-se também a maior facilidade de utilização e de interpretação das saídas

geradas pelas funções do modelo, posto que apresenta funções simplificadas para seleção de modelos, com um critério objetivo de comparação, e para análise de resíduos.

6.1 Sugestão para trabalhos futuros

Neste trabalho foi utilizada apenas um modelo de regressão dentre os cinco disponíveis no pacote *survstan*. Para avaliar a funcionalidade do pacote como um todo, sugere-se que sejam testadas também as funções dos outros modelos descritos na seção 3. Além disto, pode também ser feito um estudo utilizando simulação de bases de dados a partir de diversas distribuições de probabilidade, verificando assim a capacidade do pacote de ajustar os modelos de regressão a esses dados de acordo com os resultados esperados.

REFERÊNCIAS

Birnbaum, Z. W., e S. C. Saunders. ***Estimation for a Family of Life Distributions with Applications to Fatigue***. Journal of Applied Probability, vol. 6, nº 2, 1969, p. 328–47. *JSTOR*, https://doi.org/10.2307/3212004.

Colosimo, Enrico Antônio, e Sueli Ruiz Giolo. ***Análise de sobrevivência aplicada***. Editora Blucher, 2006.

Collett, David. ***Modelling Survival Data in Medical Research***. 4ª ed, Chapman and Hall/CRC, 2023. DOI.org (Crossref), https://doi.org/10.1201/9781003282525

Demarqui, Fabio. ***survstan: Fitting Survival Regression Models via "Stan"***. 2023, https://github.com/fndemarqui/survstan.

Demarqui F (2023). ***survstan: Fitting Survival Regression Models via 'Stan'***. R package version 0.0.3, <https://CRAN.R-project.org/package=survstan>.

Demarqui, Fabio N. ***Introdução à Análise de Sobrevivência e Confiabilidade***. 2023.

Denson, W. ***The history of reliability prediction***. *IEEE Transactions on Reliability*, vol. 47, nº 3, setembro de 1998, p. SP321–28. *IEEE Xplore*, https://doi.org/10.1109/24.740547.

Freitas, Marta Afonso, e Enrico Antônio Colosimo. ***Confiabilidade: Análise de Tempo de Falha e Testes de Vida Acelerados***. 1997.

Klein, John P., et al., organizadores. ***Handbook of Survival Analysis***. CRC Press, Taylor & Francis Group, 2014.

Kleinbaum, David G., e Mitchel Klein. ***Survival analysis: a self-learning text***. 3rd ed, Springer, 2012.

Kumar, Mukesh, et al. ***Parametric Survival Analysis Using R: Illustration with Lung Cancer Data***. *Cancer Reports*, vol. 3, nº 4, agosto de 2020, p. e1210. *DOI.org (Crossref)*, https://doi.org/10.1002/cnr2.1210.

LAFRAIA, João Ricardo Barusso. ***Manual de Confiabilidade, Mantenabilidade e Disponibilidade***. 3. ed., QualityMark, 2001.

Legrand, Catherine. ***Advanced Survival Models***. First edition, CRC Press, 2021.

Leiva, Víctor. ***The Birnbaum-Saunders Distribution***. Elsevier/AP, Academic Press is an imprint of Elsevier, 2016.

Meeker, William Q., e Luis A. Escobar. ***Statistical Methods for Reliability Data***. 1ª edição, John Wiley & Sons Inc, 1998.

Portet, Stéphanie. ***A Primer on Model Selection Using the Akaike Information Criterion***. *Infectious Disease Modelling*, vol. 5, 2020, p. 111–28. *DOI.org (Crossref)*, https://doi.org/10.1016/j.idm.2019.12.010.

R Core Team (2023). ***R: A Language and Environment for Statistical Computing***. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rocha, Henrique Martins. ***Confiabilidade***. Fundação CECIERJ, 2019.

The CRAN team. ***CRAN Task Views***. The Comprehensive R Archive Network, 2023, https://cran.r-project.org/index.html.

Therneau, Terry, e T. Lumley. ***R survival package***. R core team, 2023, https://rweb.webapps.cla.umn.edu/R/library/survival/doc/survival.pdf.

Therneau T (2023). ***A Package for Survival Analysis in R***. R package version 3.5-5, <https://CRAN.R-project.org/package=survival>.