

**Vida Residual em pacientes
com Insuficiência Cardíaca:
Uma abordagem semiparamétrica**

Victor Gonçalves Duarte

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientador: Prof. Dr. Antonio Carlos Pedroso de Lima

São Paulo, fevereiro de 2017

Vida Residual em pacientes com Insuficiência Cardíaca: Uma abordagem semiparamétrica

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 12/06/2017. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Antonio Carlos Pedroso de Lima (orientador) - IME-USP
- Prof^a. Dr^a. Lucia Pereira Barroso - IME-USP
- Prof. Dr. Mário de Castro Andrade Filho - ICMC - USP

Agradecimentos

Gostaria de agradecer minha família pelo apoio durante toda a graduação e o mestrado. Eles sempre apoiaram e incentivaram o estudo mesmo quando não entendiam nada do que eu estava fazendo.

Ao meu orientador, o professor Antonio Carlos agradeço pela paciência. Nós sabemos o quanto a comunicação foi difícil nesses anos, em alguns meses quase não nos falamos, mas ele continuou apoiando, ensinando e dando bronca. Aos membros da banca, Prof^a Lucia e Prof Mário que leram com atenção e carinho minha dissertação e deram contribuições valiosas para o trabalho. Aproveito também para agradecer alguns professores que marcaram a minha estada no IME: Prof^a Elsabeti Kira por ensinar Probabilidade I de um jeito tão bom que nunca precisei revisar a matéria para lembrar o que tinha aprendido e também pelas longas conversas que tivemos nos corredores do IME, Prof Julio Singer por me fazer entender quatro matérias diferentes lecionando apenas uma, Prof^a Gisela Tunes com quem fiz aulas de análise de sobrevivência na graduação e na pós-graduação e Prof Luís Gustavo por ensinar muito bem a matéria obrigatória do mestrado.

JP pela amizade e pelo emprego que me indicou (rsrs), espero ajudá-lo no seu mestrado como ele ajudou no meu; Helder pelas conversas filosóficas, tenísticas, futebolísticas e, por que não dizer, as estatísticas também; Renan, 79 e Gisele pela diversão que temos jogando jogos de tabuleiro, vendo animes ou cantando; Yanfei, Louise, Manuella, Cissa e Guilherme pela ajuda nas matérias.

Por último, agradeço a pessoa mais importante para mim, alguém que conheço desde que estudamos juntos na 5^a série. Obrigado, Etienne, minha amiga, namorada, companheira e todos os adjetivos possíveis que posso dar a ela. Ela esteve comigo em todos os momentos difíceis, os ataques de pânico quando achei que não seria possível terminar o trabalho, as revisões dos textos e etc. Também compreendeu que a irritação que eu demonstrava não era culpa dela e ajudou em tudo que podia.

Resumo

Usualmente a análise de sobrevivência considera a modelagem da função da taxa de falha ou função de risco. Uma alternativa a essa visão é estudar a vida residual, que em alguns casos é mais intuitiva do que a função de risco. A vida residual é o tempo de sobrevida adicional de um indivíduo que sobreviveu até um dado instante t_0 . Este trabalho descreve técnicas semiparamétricas e não paramétricas para estimar a média e a mediana de vida residual em uma população, testes para igualdade dessas medidas em duas populações e também modelos de regressão. Tais técnicas já foram testadas anteriormente em dados com baixa presença de censura; aqui elas são aplicadas a um conjunto de dados de pacientes com insuficiência cardíaca que possui uma alta quantidade de observações censuradas.

Palavras-chave: vida residual, semiparamétrico, análise de sobrevivência.

Abstract

Usually, survival analysis is based on the modeling of the hazard function. One alternative approach is to consider the residual life, which would be more intuitive than the hazard function. Residual lifetime is the remaining survival time of a person given he or she survived a given time point t_0 . We describe semiparametric and non-parametric techniques for mean and median residual life estimation in a one-sample population, as well as tests for two-sample cases and regression models. Such techniques were previously tested for moderate censored data; here we apply them to heart-failure patients data with a high rate of censoring.

Keywords: residual life, semiparametric, survival analysis.

Sumário

Lista de Figuras	ix
Lista de Tabelas	xi
1 Análise de Sobrevivência	1
1.1 Dados de sobrevivência	1
1.2 Tempo de sobrevivência	2
1.3 Estimador e teste não paramétricos	2
1.4 Modelo de Cox	4
1.5 Motivação do trabalho	7
1.6 Apresentação do trabalho	8
2 Vida Média Residual	9
2.1 Definição	9
2.2 Caracterização	10
2.3 Teste em duas populações	14
2.4 Modelo de regressão para o tempo médio residual	16
3 Vida Mediana Residual	25
3.1 Abordagem via função de estimação	25
3.2 Abordagem via verossimilhança empírica	29
3.3 Modelo de regressão para o tempo mediano residual	32
4 Estudo de Simulação	37
4.1 Estimação da vida média e mediana residual em uma população	37
4.2 Avaliação da censura	38
4.3 Teste da vida média e mediana em duas populações	40
4.4 Síntese dos resultados	45
5 Aplicação	47
5.1 Análise descritiva	47
5.2 Teste dos efeitos associados com variáveis	50
5.3 Modelos de regressão	54
5.3.1 Vida média residual	54
5.3.2 Vida mediana residual	55
6 Discussão e Conclusão	59

Referências Bibliográficas

61

Lista de Figuras

4.1	Estimador de Kaplan-Meier segundo categorias de taxa de creatinina	40
4.2	Vida média residual segundo categorias de taxa de creatinina	40
4.3	Vida mediana residual segundo categorias de taxa de creatinina	41
4.4	Estimador de Kaplan-Meier por faixa de fração de ejeção	43
4.5	Vida média residual por faixa de fração de ejeção	43
4.6	Vida mediana residual por faixa de fração de ejeção	43
5.1	Estimador de Kaplan-Meier para o conjunto de dados	47
5.2	Vida média residual	48
5.3	Vida mediana residual	48
5.4	Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Taxa de Creatinina	48
5.5	Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Fração de Ejeção	49
5.6	Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Classe Funcional	49
5.7	Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Etiologia	49
5.8	Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Diâmetro Diastólico do Ventrículo Esquerdo	50
5.9	Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Idade	50

Lista de Tabelas

1.1	Tabela de contingência para o tempo t_j	4
4.1	Resultados da simulação da vida média residual: Proporção de cobertura (amplitude média dos intervalos)	38
4.2	Resultados da simulação da vida mediana residual: Proporção de cobertura (amplitude média dos intervalos)	38
4.3	Resultados da simulação da vida média residual: Proporção de cobertura (amplitude média dos intervalos)	39
4.4	Resultados da simulação da vida mediana residual: Proporção de cobertura (amplitude média dos intervalos)	39
4.5	Resultados do teste de razão de verossimilhanças empíricas para a vida média residual	41
4.6	Resultados do teste de razão de verossimilhanças empíricas para a vida mediana residual	42
4.7	Resultados do teste da função de estimação para a vida mediana residual	42
4.8	Porcentagem de vezes que $r_{t_0}^{md}$ pertencia ao intervalo de confiança (amplitude média do intervalo)	42
4.9	Resultados do teste de razão de verossimilhanças empíricas para a vida média residual	44
4.10	Resultados do teste de razão de verossimilhanças empíricas para a vida mediana residual	44
4.11	Resultados do teste da função de estimação para a vida mediana residual	44
4.12	Porcentagem de vezes que $r_{t_0}^{md}$ estava dentro do intervalo de confiança (amplitude média do intervalo)	45
4.13	Resultados do teste da razão de verossimilhanças empíricas para a vida média residual	45
4.14	Resultados do teste da razão de verossimilhanças empíricas para a vida mediana residual	45
4.15	Resultados do teste da função de estimação para a vida mediana residual	45
5.1	Taxa de Creatinina $\geq 2,6$ mg/dL versus $< 2,6$ mg/dL	51
5.2	Taxa de Creatinina $\geq 1,3$ mg/dL versus $< 1,3$ mg/dL	51
5.3	Fração de Ejeção $\geq 25\%$ versus $< 25\%$	51
5.4	Fração de Ejeção $\geq 30\%$ versus $< 30\%$	52
5.5	Fração de Ejeção $\geq 35\%$ versus $< 35\%$	52
5.6	Classe Funcional III ou IV versus I ou II	52
5.7	Classe Funcional IV versus I, II ou III	53
5.8	Chagas versus Outras doenças	53
5.9	Diâmetro Diastólico do ventrículo esquerdo ≥ 75 mm versus < 75 mm	54

5.10 Idade ≥ 65 anos versus < 65 anos	54
5.11 Seleção de variáveis. Modelo de regressão para o tempo médio residual	55
5.12 Seleção de variáveis para $t_0 = 0$. Modelo de regressão para o tempo mediano residual	55
5.13 Seleção de variáveis para $t_0 = 12$. Modelo de regressão para o tempo mediano residual	56
5.14 Seleção de variáveis para $t_0 = 24$. Modelo de regressão para o tempo mediano residual	56
5.15 Seleção de variáveis para $t_0 = 36$. Modelo de regressão para o tempo mediano residual	57
5.16 Seleção de variáveis para $t_0 = 48$. Modelo de regressão para o tempo mediano residual	57
5.17 Seleção de variáveis para $t_0 = 60$. Modelo de regressão para o tempo mediano residual	57

Capítulo 1

Análise de Sobrevivência

O objeto de estudo da análise de sobrevivência é o tempo até a ocorrência de um determinado evento. Este tempo pode ser chamado de tempo de falha (Colosimo e Giolo, 2006) ou, dependendo do caso, tempo de vida. Em uma mesma pesquisa pode haver mais de um evento de interesse; por exemplo, em um acompanhamento de pacientes com câncer, o evento pode ser a morte do paciente ou a remissão da doença.

1.1 Dados de sobrevivência

Em estudos médicos, muitas vezes não é possível se observar o tempo até o evento de interesse; dizemos, então, que a observação foi censurada. Quando o evento de interesse for a morte por determinada doença, por exemplo, a censura pode ocorrer pela falta de acompanhamento do paciente ou morte causada por um motivo diferente do estudado. A análise de sobrevivência é necessária para o estudo de dados censurados, pois ela incorpora as informações dessas observações.

Apesar do nome sobrevivência, ela também pode ser aplicada em áreas que não sejam médicas. Um exemplo pode ser visto em Asadi e Bayramoglu (2005) que estudam a *vida média residual* de um sistema de componentes paralelos. Já Benda (2003) usa o modelo de riscos proporcionais de Cox para analisar a reincidência de crimes.

Colosimo e Giolo (2006) citam dois motivos para que as informações censuradas sejam utilizadas na análise estatística. Primeiramente, mesmo incompletos, os dados censurados fornecem informação sobre o tempo de vida dos pacientes. Além disso, a não utilização dessas observações pode levar a conclusões enviesadas.

A origem do tempo de falha e o evento de interesse devem ser definidos de forma precisa para que não ocorram problemas no momento da análise. Em estudos de câncer, por exemplo, a origem do tempo pode ser a data do diagnóstico, o dia em que os sintomas começaram ou até o início de um tratamento, enquanto o evento de interesse pode ser a morte ou a remissão.

As censuras podem ser classificadas como sendo à direita, à esquerda e intervalar. Existem diferentes tipos de censura à direita. A censura do tipo I acontece quando o estudo termina antes que se observe o evento de interesse para todos os indivíduos. A censura do tipo II ocorre quando o estudo termina no momento em que um número pré-definido de indivíduos tenha o evento de interesse; por exemplo, ao testar o tempo de vida útil de lâmpadas termina-se o estudo quando 30 lâmpadas pararem de funcionar.

O tipo de censura que mais ocorre na prática médica é a censura aleatória, que ocorre quando um paciente morre por uma causa que não é a que está sendo estudada ou o indivíduo deixa de comparecer às consultas com o médico e não se tem o acompanhamento final dele. A representação deste caso de censura pode ser feita usando duas variáveis aleatórias. Seja T o tempo de falha de um indivíduo, C o tempo de censura do mesmo indivíduo e T independente de C . Então, serão observados X e δ definidos como

$$X = \min(T, C) \quad \text{e} \quad \delta = I(T \leq C).$$

Os outros tipos de censura não serão considerados neste trabalho, porém mais detalhes sobre eles podem ser vistos em [Colosimo e Giolo \(2006\)](#).

1.2 Tempo de sobrevivência

O tempo de falha pode ser especificado pela sua função de distribuição F . No contexto de análise de sobrevivência, é definida a função de sobrevivência, que é a probabilidade de um indivíduo não falhar até o tempo t , ou seja,

$$S(t) = P(T > t), \quad t \geq 0;$$

e conseqüentemente temos a relação $S(t) = 1 - F(t)$.

A taxa de falha é associada à probabilidade de que o evento ocorra em um intervalo $[t_1, t_1 + \Delta t)$, dado que ele não ocorreu antes de t_1 .

Quando tornamos esse intervalo bem pequeno, ou seja, $\Delta t \rightarrow 0$, temos a taxa de falha no momento t , definida como

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}, \quad t \geq 0.$$

A função de risco acumulada é muito útil em processos de estimação não paramétricos, apesar de não ter uma interpretação direta. Essa função é definida como

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

Essas funções possuem relações diretas entre elas, portanto conhecer uma delas implica no conhecimento das demais; algumas das relações são as seguintes:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} (\log S(t)),$$

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log S(t) \quad \text{e}$$

$$S(t) = e^{-\Lambda(t)} = \exp \left\{ - \int_0^t \lambda(u) du \right\}.$$

1.3 Estimador e teste não paramétricos

A primeira parte de uma análise de dados é a análise descritiva. No contexto de dados censurados, em que o tempo de sobrevivência é o objeto de interesse, o gráfico do estimador não paramétrico da função de sobrevivência é comumente usado para descrever os dados. O estimador de Kaplan-Meier ([Kaplan e Meier, 1958](#)) é o mais conhecido e será apresentado a seguir.

Considere um estudo com n pacientes e k falhas distintas em que $t_1 < t_2 < \dots < t_k$ são os tempos de falha distintos e ordenados, d_j é o número de falhas em t_j , $j = 1, \dots, k$, e n_j é o número de indivíduos em risco no tempo imediatamente anterior a t_j (t_j^-), ou seja, os indivíduos que não sofreram falha e nem foram censurados até o tempo t_j^- . Assim, o estimador de Kaplan-Meier é definido como

$$\hat{S}(t) = \prod_{j:t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j < t} \left(1 - \frac{d_j}{n_j} \right), \quad t \geq 0.$$

Kaplan e Meier (1958) mostraram que o estimador $\hat{S}(t)$ é o estimador de máxima verossimilhança de $S(t)$, enquanto Breslow e Crowley (1974) provaram que o estimador é fracamente consistente para $S(t)$ e assintoticamente normal, para t fixo. A variância assintótica é estimada pela fórmula de Greenwood (Greenwood, 1926):

$$\widehat{\text{Var}}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j < t} \frac{d_j}{n_j(n_j - d_j)}. \quad (1.1)$$

Sendo assim, um intervalo de confiança aproximado de $100(1 - \alpha)\%$ para $S(t)$ pode ser calculado por

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{S}(t))}, \quad t \text{ fixo},$$

em que $z_{\alpha/2}$ é o percentil $\alpha/2$ da distribuição Normal padrão.

O intervalo de confiança definido acima pode gerar valores não pertencentes ao intervalo $[0,1]$. Portanto, Kalbfleisch e Prentice (2002) propuseram uma transformação para $S(t)$ da forma $\hat{U}(t) = \log[-\log(\hat{S}(t))]$ e um estimador da variância de $\hat{U}(t)$ é

$$\widehat{\text{Var}}(\hat{U}(t)) = \frac{\sum_{j:t_j < t} \frac{d_j}{n_j(n_j - d_j)}}{[\log \hat{S}(t)]^2}.$$

Então, um intervalo de confiança de aproximadamente $100(1 - \alpha)\%$ para $S(t)$ é dado por

$$[\hat{S}(t)]^{\exp \left\{ \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{U}(t))} \right\}}.$$

Outro estimador de $S(t)$ importante é o estimador de Nelson-Aalen. Nelson (1972) propôs o estimador $\tilde{\Lambda}(t)$ definido por

$$\tilde{\Lambda}(t) = \sum_{j:t_j < t} \frac{d_j}{n_j},$$

em que d_j e n_j são definidos de forma semelhante ao caso do estimador de Kaplan-Meier. Aalen (1978) provou as propriedades assintóticas de $\tilde{\Lambda}(t)$ usando processos de contagem e propôs um estimador para a variância de $\tilde{\Lambda}(t)$:

$$\widehat{\text{Var}}(\tilde{\Lambda}(t)) = \sum_{j:t_j < t} \frac{d_j}{n_j^2}.$$

Sendo assim, o estimador da função de sobrevivência derivado do estimador de Nelson-Aalen é dado por

$$\tilde{S}(t) = \exp \{ \tilde{\Lambda}(t) \}.$$

A variância desse estimador pode ser estimada por

$$\widehat{\text{Var}}(\tilde{S}(t)) = [\tilde{S}(t)]^2 \sum_{j:t_j < t} \frac{d_j}{n_j^2},$$

ou substituindo $\hat{S}(t)$ por $\tilde{S}(t)$ na Equação (1.1).

Bohoris (1994) mostrou que $\tilde{S}(t) \geq \hat{S}(t)$ para todo t , apesar de apresentarem estimativas muito próximas na maioria das vezes. Uma vantagem no uso do estimador de Kaplan-Meier é a sua disponibilidade em diversos pacotes estatísticos, enquanto o estimador de Nelson-Aalen não é encontrado com tanta frequência. Entretanto, ambos estimadores coincidem assintoticamente.

Foram apresentados dois estimadores não paramétricos para a função de sobrevivência de uma população. Para comparar duas populações com relação às suas respectivas curvas de sobrevivência o teste mais conhecido é o teste *logrank* (Mantel, 1966).

Sejam $t_1 < t_2 < \dots < t_k$ os tempos de falha distintos na amostra combinada de duas populações. Então, para cada tempo t_j construímos a Tabela 1.1:

	Grupo		Total
	1	2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Total	n_{1j}	n_{2j}	n_j

Tabela 1.1: Tabela de contingência para o tempo t_j .

Harrington e Fleming (1982) propuseram a estatística de teste Z_{HF} , que tem distribuição assintótica qui-quadrado com 1 grau de liberdade, é dada por

$$Z_{HF} = \frac{[\sum_{j=1}^k u_j (d_{2j} - w_{2j})]^2}{\sum_{j=1}^k u_j^2 V_j},$$

em que $w_{2j} = \frac{n_{2j} d_j}{n_j}$, $V_j = \frac{n_{2j}(n_j - n_{2j}) d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$ e $u_j = [\hat{S}(t_{j-1})]^\rho$.

Outros valores podem ser usados para u_j . Um caso particular ocorre quando $\rho = 0$, em que Z_{HF} se reduz a estatística do teste *logrank*.

1.4 Modelo de Cox

O modelo de regressão de Cox (Cox, 1972) é o mais usado na análise de sobrevivência. Considere p covariáveis e $\mathbf{x} = (x_1, \dots, x_p)'$ o vetor de covariáveis. Então, o modelo de Cox é escrito na forma

$$\lambda(t|\mathbf{x}) = \lambda_0(t)g(\mathbf{x}'\boldsymbol{\beta}), \quad t \geq 0,$$

em que $\boldsymbol{\beta}$ é o vetor com os parâmetros do modelo, $g(\mathbf{x}'\boldsymbol{\beta})$ é uma função não negativa, de modo que $g(0) = 1$. Usualmente $g(\cdot)$ é a função exponencial e λ_0 é chamada de taxa de risco de base, ou basal. O modelo de Cox é formado por um componente paramétrico $g(\mathbf{x}'\boldsymbol{\beta})$ e um componente não paramétrico λ_0 , por esta razão o modelo é denominado semiparamétrico.

O modelo de Cox implica que para dois indivíduos i e j , a razão entre suas respectivas taxas de risco não varia ao longo do tempo, em outras palavras, o modelo supõe que as taxas de risco sejam proporcionais ao longo do tempo, como pode ser notado algebricamente abaixo:

$$\frac{\lambda(t|\mathbf{x}_i)}{\lambda(t|\mathbf{x}_j)} = \frac{\lambda_0(t)e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\lambda_0(t)e^{\mathbf{x}_j'\boldsymbol{\beta}}} = e^{\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_j'\boldsymbol{\beta}}.$$

Considere uma variável X_1 , que assume valor 1 ou 0, representando a exposição ou não exposição a um fator de interesse. Ajustando o modelo de Cox, a interpretação de β é que e^β é a razão da taxa de risco da população que foi exposta ao fator de interesse pela taxa de risco dos que não foram expostos ao fator de interesse.

Cox (1975) propôs o método de máxima verossimilhança parcial para evitar a necessidade de estimar $\lambda_0(t)$. Seja uma amostra de n indivíduos com $k \leq n$ falhas distintas nos tempos $t_1 < \dots < t_k$; o argumento utilizado é a probabilidade de a observação i falhar em t_i dado que há uma falha em t_i e sabendo quais observações estão sob risco em t_i é aproximadamente igual a

$$\frac{\lambda_i(t|\mathbf{x}_i)}{\sum_{j \in R(t_i)} \lambda_j(t|\mathbf{x}_j)} = \frac{\lambda_0(t)e^{\mathbf{x}'_i\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} \lambda_0(t)e^{\mathbf{x}'_j\boldsymbol{\beta}}} = \frac{e^{\mathbf{x}'_i\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{x}'_j\boldsymbol{\beta}}}, \quad (1.2)$$

em que $R(t_i)$ é o conjunto de índices dos indivíduos em risco no tempo t_i . A função de verossimilhança será o produto dos termos calculados por (1.2) para cada tempo distinto de falha:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^k \frac{e^{\mathbf{x}'_i\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{x}'_j\boldsymbol{\beta}}} = \prod_{i=1}^n \left(\frac{e^{\mathbf{x}'_i\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{x}'_j\boldsymbol{\beta}}} \right)^{\delta_i},$$

sendo δ_i o indicador de falha. O estimador de máxima verossimilhança parcial para $\boldsymbol{\beta}$ denotado $\hat{\boldsymbol{\beta}}$ é encontrado solucionando a equação abaixo:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left[\mathbf{x}_i - \frac{\sum_{j \in R(t_i)} \mathbf{x}_j e^{\mathbf{x}'_j\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{x}'_j\boldsymbol{\beta}}} \right] = \mathbf{0}. \quad (1.3)$$

$\mathbf{U}(\boldsymbol{\beta})$ é chamado de vetor escore e é formado pelas derivadas de primeira ordem do logaritmo da função de verossimilhança parcial.

Em dados reais, é comum encontrar empates entre tempos de falha ou entre falhas e censuras. A função de verossimilhança parcial supõe que os tempos de sobrevivência são contínuos e, portanto, não deveriam ocorrer empates nos valores observados. Para os empates entre falhas e censuras, usualmente considera-se que a censura ocorreu após a falha, então a observação censurada estaria em risco no momento da falha.

Quando o número de empates em qualquer tempo não é grande, uma aproximação adequada é (Colosimo e Giolo, 2006):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^k \frac{e^{\mathbf{s}'_i\boldsymbol{\beta}}}{\left[\sum_{j \in R(t_i)} e^{\mathbf{x}'_j\boldsymbol{\beta}} \right]^{d_i}},$$

em que \mathbf{s}_i é a soma dos vetores de covariáveis dos indivíduos empatados em t_i e d_i é o número de observações empatadas em t_i .

Outra extensão que o modelo de Cox permite é a inclusão de covariáveis dependentes do tempo. Um exemplo de covariável com essa característica é a classificação (soropositivo assintomático, ARC e AIDS) de pacientes infectados pelo vírus HIV (Colosimo e Giolo, 2006), que pode se alterar ao longo do tempo. O modelo de Cox com covariáveis dependentes do tempo é dado por

$$\lambda(t|\mathbf{x}(t)) = \lambda_0(t)e^{\mathbf{x}'(t)\boldsymbol{\beta}},$$

que não é um modelo de taxas de risco proporcionais, porque

$$\frac{\lambda(t|\mathbf{x}_i(t))}{\lambda(t|\mathbf{x}_j(t))} = e^{\mathbf{x}'_i(t)\boldsymbol{\beta} - \mathbf{x}'_j(t)\boldsymbol{\beta}}.$$

A estimativa $\hat{\beta}$ do coeficiente β é a solução da equação

$$\mathbf{U}(\beta) = \sum_{i=1}^n \delta_i \left[\mathbf{x}_i(t_i) - \frac{\sum_{j \in R(t_i)} \mathbf{x}_j(t_i) e^{\mathbf{x}'_j(t_i) \hat{\beta}}}{\sum_{j \in R(t_i)} e^{\mathbf{x}'_j(t_i) \hat{\beta}}} \right] = \mathbf{0}, \quad (1.4)$$

sendo a expressão acima bem similar à Equação (1.3).

Andersen e Gill (1982) mostraram que o estimador resultante da Equação (1.4) é consistente e assintoticamente normal sob certas condições de regularidade. Sendo assim, é possível utilizar a estatística de Wald e o teste da razão de verossimilhanças, que serão apresentados a seguir, para realizar testes de hipóteses sobre o parâmetro da regressão de Cox com covariáveis dependentes do tempo.

A função de verossimilhança parcial permite a realização de três testes de hipóteses sobre β utilizando a estimativa $\hat{\beta}$ resultante da Equação (1.3): razão de verossimilhanças, Wald e Escore. A hipótese testada será $H_0 : \beta = \beta_0$ (Klein e Moeschberger, 2003).

Seja $l(\beta) = \log L(\beta)$, a estatística do teste da razão de verossimilhanças é dada por

$$\chi_{RV}^2 = 2[l(\hat{\beta}) - l(\beta_0)],$$

que tem distribuição assintótica qui-quadrado com p graus de liberdade, sob H_0 .

A matriz de informação observada $\mathbf{I}(\beta)$ pode ser calculada como

$$\mathbf{I}(\beta) = -\frac{\partial^2}{\partial \beta \partial \beta'} l(\beta).$$

A estatística do teste de Wald é dada por

$$\chi_w^2 = (\hat{\beta} - \beta_0)' \mathbf{I}(\hat{\beta}) (\hat{\beta} - \beta_0),$$

que tem distribuição assintótica qui-quadrado com p graus de liberdade, sob H_0 .

O terceiro teste é o teste Escore, $\mathbf{U}(\beta)$ é assintoticamente normal p -variada com média $\mathbf{0}$ e matriz de covariância $\mathbf{I}(\beta)$. A estatística é definida como

$$\chi_{esc}^2 = \mathbf{U}(\beta_0)' \mathbf{I}(\beta_0)^{-1} \mathbf{U}(\beta_0).$$

A estatística de teste χ_{esc}^2 tem distribuição assintótica qui-quadrado com p graus de liberdade, sob H_0 .

Previamente foi dito que a suposição básica do modelo de Cox é a proporcionalidade da função de risco entre os níveis das variáveis estudadas, porém nem sempre isso ocorre na prática. Uma adaptação que pode ser feita é dividir os dados em m estratos de acordo com a variável cuja suposição de proporcionalidade não seja válida.

O modelo de Cox estratificado fica definido, então, como

$$\lambda(t|\mathbf{x}_{ij}) = \lambda_{0j}(t) e^{\mathbf{x}'_{ij} \beta}, \quad t \geq 0,$$

com $j = 1, \dots, m$ e $i = 1, \dots, n_j$, n_j é o número de observações no estrato j .

A função de log-verossimilhança parcial é construída de maneira similar ao mostrado anteriormente e resulta em

$$l(\boldsymbol{\beta}) = l_1(\boldsymbol{\beta}) + \cdots + l_m(\boldsymbol{\beta}),$$

em que l_j é o logaritmo da função de verossimilhança obtida utilizando apenas os dados dos indivíduos do estrato j (Kalbfleisch e Prentice, 2002).

O modelo estratificado supõe que as covariáveis têm o mesmo efeito na função de risco de cada estrato. A diferença entre os estratos é captada pelo componente não paramétrico, que é estimado separadamente para cada um deles.

Foram apresentadas algumas técnicas clássicas de análise de sobrevivência, em que o foco da análise é a função de sobrevivência ou a função de risco. O estimador de Kaplan-Meier e o modelo de regressão de Cox são amplamente usados em dados censurados e são estimativas não paramétricas e semiparamétricas, que serão características das análises feitas no decorrer do trabalho para a *vida residual*.

1.5 Motivação do trabalho

Há o interesse de médicos em avaliar os fatores que levam parte dos pacientes com insuficiência cardíaca a viver mais de cinco anos. A partir desse problema, perguntamos não somente se o paciente irá viver mais de cinco anos, mas por quanto tempo e com que qualidade. Por isso, vamos aplicar técnicas de estudo de vida residual para avaliar o tempo de sobrevivência dos pacientes, dado que eles viveram cinco anos.

A análise da vida residual pode ser mais intuitiva do que analisar a função de sobrevivência ou a função de risco (Jeong, 2014). Por exemplo, afirmar a um paciente que ao tomar um medicamento ele ganhará em média mais cinco anos de vida parece mais natural do que dizer que a redução na taxa de risco do grupo que tomou o medicamento foi 25% menor.

O conjunto de dados que motivou o estudo refere-se a 2128 pacientes com insuficiência cardíaca acompanhados de julho de 2003 até abril de 2015 pela clínica médica do Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo (InCor - FMUSP). Setenta e três variáveis estão no arquivo original e uma análise feita anteriormente, por meio do Centro de Estatística Aplicada (CEA), considerou algumas delas relevantes para um estudo de sobrevivência. Neste trabalho iremos nos referir a esses dados como InCor.

O tempo de acompanhamento considerado para o estudo foi o tempo decorrido, em meses, da entrada do paciente no protocolo até um dos casos: óbito (por qualquer causa) ou último contato com o paciente vivo. A primeira situação representa a “falha” e a segunda é considerada a censura.

Utilizando o gráfico de Kaplan-Meier como análise descritiva, chega-se a uma lista menor de variáveis que têm alguma relação com o tempo de sobrevivência: Idade, Etiologia (doença de Chagas, doença cardíaca hipertensiva, cardiomiopatia isquêmica, cardiomiopatia dilatada, cardiomiopatia alcoólica), Classe Funcional (I - sem sintomas, II - limitado a grandes esforços, III - limitado para esforços moderados e IV - limitado mesmo em repouso), Diâmetro diastólico do ventrículo esquerdo (mm), Taxa de creatinina no sangue (mg/dL) e a Fração de Ejeção do ventrículo esquerdo.

A Fração de Ejeção é a variável de maior interesse dos pesquisadores e representa a porcentagem de sangue que deixa o ventrículo esquerdo a cada contração do coração. No Capítulo 5 será feita uma análise mais detalhada do conjunto de dados do ponto de vista da vida residual, analisando a vida média e mediana residual.

1.6 Apresentação do trabalho

O Capítulo 2 discute a definição de *vida média residual* e suas propriedades teóricas. Um estimador não paramétrico é apresentado juntamente com sua variância. Também é considerado um teste para a igualdade entre a *vida média residual* de duas populações. Um modelo de regressão para a *vida média residual* semelhante ao modelo de Cox de riscos proporcionais é apresentado.

No Capítulo 3 é definida a *vida mediana residual* e apresentado um estimador não paramétrico para ela. Sob a abordagem via função de estimação é considerado um teste para a igualdade entre as medianas de vida residual de duas populações e um modelo de regressão semiparamétrico. Usando uma função de verossimilhança empírica, é considerado um teste em duas populações para a igualdade da mediana e média de vida residual.

O Capítulo 4 apresenta estudos de simulação para os estimadores não paramétricos da vida média e mediana residual feitos em situações de alta presença de censura. Também são feitas simulações para os testes de hipóteses mostrados nos Capítulos 2 e 3.

No Capítulo 5 é realizada uma análise descritiva com relação à vida média e mediana residual no banco de dados apresentado na Seção 1.5. Testes de hipóteses são realizados para identificar variáveis que expliquem as medidas de vida residual e os modelos de regressão são aplicados aos dados.

Capítulo 2

Vida Média Residual

2.1 Definição

O tempo de vida residual é o tempo de sobrevivida (T) de um indivíduo que excede a um instante t , quando ele permanece vivo. Uma medida resumo natural para se avaliar a vida residual é a média, definida por

$$m(t) = E(T - t | T > t), \text{ para } t \geq 0,$$

denominada *vida média residual*. Note que

$$\begin{aligned} m(t) &= E(T - t | T > t) = \int_{-\infty}^{\infty} (u - t) dF_{T|T>t}(u | T > t) \\ &= \int_{-\infty}^{\infty} (u - t) f_{T|T>t}(u | T > t) du = \frac{\int_t^{\infty} (u - t) f(u) du}{P(T > t)}. \end{aligned}$$

Utilizando integração por partes no numerador e usando o fato que

$$f(u) = -\frac{d}{du} S(u),$$

obtemos

$$\int_t^{\infty} (u - t) f(u) du = \lim_{u \rightarrow \infty} u S(u) + \int_t^{\infty} S(u) du.$$

O primeiro termo no lado direito da expressão anterior é igual a 0, pois

$$0 \leq \lim_{u \rightarrow \infty} u S(u) = \lim_{u \rightarrow \infty} u \int_u^{\infty} f(v) dv,$$

e para todo $u \leq v$, $u f(v) \leq v f(v)$. Então,

$$\lim_{u \rightarrow \infty} u \int_u^{\infty} f(v) dv \leq \lim_{u \rightarrow \infty} \int_u^{\infty} v f(v) dv = 0.$$

Sendo assim, temos que

$$m(t) = \frac{\int_t^{\infty} S(u) du}{S(t)}.$$

Portanto, temos uma relação de equivalência entre o tempo médio residual e a função de sobrevivência. Alternativamente, a partir da expressão anterior temos a relação

$$\frac{1}{m(t)} = \frac{S(t)}{\int_t^{\infty} S(u) du} = -\frac{d}{dt} \log \left(\int_t^{\infty} S(u) du \right) = -\frac{d}{dt} \log \left(S(t) m(t) \right).$$

Utilizando o resultado acima, obtemos a igualdade

$$\begin{aligned} \int_0^t \frac{1}{m(u)} du &= - \int_0^t \frac{d}{du} \left(\log(S(u)m(u)) \right) du = - \int_0^t d \left(\log(S(u)m(u)) \right) \\ &= - \left(\log(S(t)m(t)) - \log(S(0)m(0)) \right) = - \log \left(\frac{S(t)m(t)}{m(0)} \right). \end{aligned}$$

Multiplicando por -1 e aplicando a função exponencial temos

$$\exp \left(- \int_0^t \frac{1}{m(u)} du \right) = \frac{S(t)m(t)}{m(0)}.$$

Podemos, então, escrever a função de sobrevivência em função da *vida média residual* por meio da expressão

$$S(t) = \frac{m(0)}{m(t)} \exp \left(- \int_0^t \frac{du}{m(u)} \right),$$

chamada de *fórmula de inversão*.

Utilizando a seguinte relação entre a função de risco e a função de sobrevivência

$$\lambda(t) = - \frac{d}{dt} (\log S(t)),$$

podemos escrever a função de risco em função da *vida média residual*

$$\log S(t) = \log m(0) - \log m(t) - \int_0^t \frac{du}{m(u)}.$$

Então,

$$\lambda(t) = - \frac{d}{dt} \left(\log m(0) - \log m(t) - \int_0^t \frac{du}{m(u)} \right) = 0 + \frac{\frac{dm(t)}{dt}}{m(t)} + \frac{1}{m(t)} = \frac{\frac{dm(t)}{dt} + 1}{m(t)},$$

que relaciona a função de taxa de falha com a média do tempo de vida residual.

2.2 Caracterização

Hall e Wellner (1981) caracterizaram a função $m(t)$ e demonstraram diversas propriedades. A seguinte proposição estabelece limites superiores e inferiores para $m(t)$ e também sua continuidade.

Proposição: Seja T uma variável aleatória não negativa e absolutamente contínua, com $F(t)$ sua função de distribuição, $S(t)$ a função de sobrevivência e $\tau_F = \inf\{x : F(x) = 1\} \leq \infty$. Considere $(f(t))^+ = f(t)I(f(t) > 0)$. Então,

- a. $m(t) \leq (\tau_F - t)^+$, para todo t ;
- b. $m(t) \leq \frac{E(T)}{S(t)} - t$, para todo t ;
- c. $m(t) < \left[\frac{E(T^r)}{S(t)} \right]^{\frac{1}{r}} - t$, para todo t e qualquer $r > 1$;
- d. $m(t) \geq \frac{(E(T)-t)^+}{S(t)}$, para $t < \tau_F$;
- e. $m(t) \geq (E(T) - t)^+$, para todo t ;

- f. $m(t) > \frac{E(T) - F(t) \left[\frac{E(T^r)}{F(t)} \right]^{\frac{1}{r}}}{S(t)} - t$, para $t < \tau_F$ e qualquer $r > 1$;
- g. $m(t)$ é não negativa e contínua à direita;
- h. $v(t) \equiv m(t) + t$ é não decrescente;
- i. $m(t)$ tem limites à esquerda em $(0, \infty)$ e incrementos positivos nas descontinuidades, se houver alguma;
- j. $m(t^-) > 0$ para $t \in (0, \tau_F)$. Se $\tau_F < \infty$, $m(\tau_F^-) = 0$ e $m(t)$ é contínua em τ_F ;
- k. $\int_0^t \frac{1}{m(u)} du \rightarrow \infty$ para $t \rightarrow \tau_F$.

Para demonstrar as propriedades acima são estabelecidas cinco desigualdades resultantes da igualdade

$$m(t) + t = E(T|T > t) = \frac{E(I(T > t)T)}{S(t)},$$

sendo que $I(\cdot)$ é a função indicadora, de tal forma que

$$E(I(T > t)T) = E(T) - E(I(T \leq t)T).$$

As desigualdades são:

- I. $E(I(T > t)T) \leq E(T)$;
- II. $E(I(T > t)T) \leq \tau_F S(t)$;
- III. $E(I(T \leq t)T) \leq tF(t)$;
- IV. $E(I(T > t)T) \leq E(T^r)^{1/r} S(t)^{1-\frac{1}{r}}$, para $r > 1$;
- V. $E(I(T \leq t)T) \leq E(T^r)^{1/r} F(t)^{1-\frac{1}{r}}$, para $r > 1$.

As desigualdades (IV) e (V) são encontradas a partir da desigualdade de Hölder.

A propriedade (a) é provada desenvolvendo a desigualdade (II):

$$\begin{aligned} E(I(T > t)T) &\leq \tau_F S(t) \\ m(t) + t &\leq \tau_F, \end{aligned}$$

e para $t < \tau_F$, $m(t) \leq \tau_F - t$.

A partir da desigualdade (I), (b) pode ser provado que

$$\begin{aligned} \frac{E(I(T > t)T)}{S(t)} &\leq \frac{E(T)}{S(t)} \\ m(t) + t &\leq \frac{E(T)}{S(t)} \\ m(t) &\leq \frac{E(T)}{S(t)} - t. \end{aligned}$$

A propriedade (c) segue da desigualdade (III):

$$\begin{aligned}
E(I(T > t)T) &\leq E(T^r)^{1/r} S(t)^{1-\frac{1}{r}} \\
m(t) + t &\leq E(T^r)^{1/r} S(t)^{-\frac{1}{r}} \\
m(t) &< \left[\frac{E(T^r)}{S(t)} \right]^{\frac{1}{r}} - t.
\end{aligned}$$

Usando a relação entre a função de distribuição e a função de sobrevivência na desigualdade (IV), obtemos

$$\begin{aligned}
E(T) - E(I(T > t)T) &\leq t(1 - S(t)) \\
-E(I(T > t)T) &\leq -E(T) + t - tS(t) \\
E(I(T > t)T) &\geq E(T) - t + tS(t) \\
m(t) + t &\geq \frac{E(T)}{S(t)} - \frac{t}{S(t)} + t,
\end{aligned}$$

resultando na propriedade (d).

Como $0 \leq S(t) \leq 1$ para todo t , a propriedade (e) segue naturalmente de (d). A partir da desigualdade (V) temos

$$\begin{aligned}
E(T) - E(I(T > t)T) &\leq \left(\frac{E(T^r)}{F(t)} \right)^{1/r} F(t) \\
-E(I(T > t)T) &\leq \left(\frac{E(T^r)}{F(t)} \right)^{1/r} F(t) - E(T) \\
E(I(T > t)T) &\geq E(T) - F(t) \left(\frac{E(T^r)}{F(t)} \right)^{1/r} \\
m(t) + t &\geq \frac{E(T) - F(t) \left[\frac{E(T^r)}{F(t)} \right]^{\frac{1}{r}}}{S(t)},
\end{aligned}$$

que resulta na propriedade (f).

A propriedade (g) é verdadeira, pois $\int_t^\infty S(u)du$ é contínua e $S(t)$ é não negativa e contínua à direita. Para mostrar a propriedade (h), consideramos $x > 0$, $t + x < \tau$ e $v(t)$ como definido anteriormente. Temos então

$$v(t+x) - v(t) = \frac{\int_{t+x}^\infty S(u)du}{S(t+x)} - (t+x) - \left(\frac{\int_t^\infty S(u)du}{S(t+x)} - (t) \right) = \frac{\int_{t+x}^\infty S(u)du}{S(t+x)} - \frac{\int_t^\infty S(u)du}{S(t)} - x.$$

Como $S(t+x) \leq S(t)$,

$$\frac{\int_{t+x}^\infty S(u)du}{S(t+x)} - \frac{\int_t^\infty S(u)du}{S(t)} - x \geq \frac{\int_{t+x}^\infty S(u)du}{S(t)} - \frac{\int_t^\infty S(u)du}{S(t)} - x = -\frac{\int_t^{t+x} S(u)du}{S(t)} + x.$$

Uma vez que, $0 \leq S(t) \leq 1$,

$$-\frac{\int_t^{t+x} S(u)du}{S(t)} + x \geq -\int_t^{t+x} S(u)du + x = \int_t^{t+x} 1du - \int_t^{t+x} S(u)du = \int_t^{t+x} F(u)du.$$

Como $0 \leq F(u) \leq 1$,

$$v(t+x) - v(t) \geq \int_t^{t+x} F(u)du \geq 0.$$

Logo, $v(t)$ é não decrescente. A propriedade (i) é uma consequência direta da propriedade (h).

Seja $t < \tau$, então

$$m(t^-) = \frac{\int_{t^-}^{\infty} S(u)du}{S(t^-)} \geq \int_{t^-}^{\infty} S(u)du \geq \int_{t^-}^{\tau} S(u)du > 0.$$

Se $t < \tau < \infty$, $m(t) + t \leq m(\tau) + \tau = \tau$, então $m(t) \leq \tau - t$ e $m(t) \leq \tau - \tau^-$. Portanto, a propriedade (j) está provada.

O limite

$$\lim_{t \rightarrow \tau} S(t)m(t) = 0$$

e a igualdade

$$\int_0^t \frac{1}{m(u)} du = -\log \left(\frac{S(t)m(t)}{m(0)} \right)$$

mostram que a propriedade (k) é verdadeira:

$$\lim_{t \rightarrow \tau} \int_0^t \frac{1}{m(u)} du = \lim_{t \rightarrow \tau} -\log \left(\frac{S(t)m(t)}{m(0)} \right) = \infty.$$

Hall e Wellner (1981) também caracterizaram a classe de distribuições com a *vida média residual* linear na forma $m(t) = at + b$ e mostraram que as únicas distribuições com essa propriedade são Exponencial, Pareto e uma determinada classe adaptada da distribuição Beta.

Sejam a e b constantes, usando a *fórmula de inversão*, temos

$$S(t) = \frac{b}{at+b} \exp \left(-\int_0^t \frac{1}{au+b} du \right).$$

Como $\frac{d}{dt} \frac{\log(at+b)}{a} = \frac{1}{at+b}$, a expressão acima é igual a

$$\begin{aligned} & \frac{b}{at+b} \exp \left(-\left(\frac{\log(at+b)}{a} - \frac{\log(a \cdot 0 + b)}{a} \right) \right) = \frac{b}{at+b} \exp \left(-\frac{\log(at+b) - \log(b)}{a} \right) \\ & = \frac{b}{at+b} \exp \left(\frac{\log\left(\frac{b}{at+b}\right)}{a} \right) = \frac{b}{at+b} \exp \left(\log \left(\left(\frac{b}{at+b} \right)^{1/a} \right) \right) = \frac{b}{at+b} \left(\frac{b}{at+b} \right)^{1/a}. \end{aligned}$$

Portanto, a função de sobrevivência de uma distribuição que tenha *vida média residual* na forma linear é dada por

$$S(t) = \left(\frac{b}{at + b} \right)_+^{\frac{1}{a} + 1}.$$

2.3 Teste em duas populações

Berger *et al.* (1988) apresentam um teste não paramétrico para testar a igualdade da *vida média residual* de duas populações independentes. Seja X o tempo de vida de uma população, com função de distribuição F e a *vida média residual* $m_F(t)$ associada a X . G e $m_G(t)$ são definidos de forma análoga para uma população com tempo de vida Y . As hipóteses do teste são

$$\begin{aligned} H_0 : m_F(t) &\leq m_G(t), \quad \text{para algum } t \in [T_1, T_2] \\ H_a : m_F(t) &> m_G(t), \quad \text{para todo } t \in [T_1, T_2]. \end{aligned} \quad (2.1)$$

Sejam duas amostras independentes X_1, X_2, \dots, X_{n_1} e Y_1, Y_2, \dots, Y_{n_2} com função de distribuição F e *vida média residual* $m_F(t)$ associadas a X e G e $m_G(t)$ as correspondentes funções para Y . Definem-se as seguintes sub-hipóteses para um valor de t fixado:

$$\begin{aligned} H_{0t} : m_F(t) &\leq m_G(t) \\ H_{at} : m_F(t) &> m_G(t). \end{aligned} \quad (2.2)$$

H_0 será rejeitado, se e somente se, H_{0t} for rejeitada a um nível de significância α para algum $t \in [T_1, T_2]$.

A estatística de teste para as hipóteses (2.2) é dada por

$$Z_{n_1 n_2}(t) = \frac{\hat{m}_F(t) - \hat{m}_G(t)}{\sqrt{\frac{S_{n_1}^2(t)}{n_1(t)} + \frac{S_{n_2}^2(t)}{n_2(t)}}},$$

em que $\hat{m}_F(t) = \hat{\mu}_F - t$ com $\hat{\mu}_F$ sendo a média dos X_i 's maiores do que t , $S_{n_1}^2(t)$ é a variância amostral dos X_i 's maiores do que t e $n_1(t)$ é o número de X_i 's maiores do que t . As quantidades $\hat{m}_G(t)$, $S_{n_2}^2(t)$ e $n_2(t)$ são definidas de forma análoga para Y_i .

A estatística $Z_{n_1 n_2}(t)$ pode ser reescrita como

$$\begin{aligned} Z_{n_1 n_2}(t) &= \frac{\sqrt{n_1} [(\hat{m}_F(t) - m_F(t)) - (\hat{m}_G(t) - m_G(t)) + (m_F(t) - m_G(t))]}{\sqrt{n_1} \sqrt{\frac{S_{n_1}^2(t)}{n_1(t)} + \frac{S_{n_2}^2(t)}{n_2(t)}}} \\ &= \frac{\sqrt{n_1} [(\hat{m}_F(t) - m_F(t)) - (\hat{m}_G(t) - m_G(t))]}{D_{n_1 n_2}} + \frac{\sqrt{n_1} (m_F(t) - m_G(t))}{D_{n_1 n_2}} \\ &= A_{n_1 n_2}(t) + B_{n_1 n_2}(t), \end{aligned}$$

com

$$D_{n_1 n_2} = \sqrt{n_1} \sqrt{\frac{S_{n_1}^2(t)}{n_1(t)} + \frac{S_{n_2}^2(t)}{n_2(t)}}.$$

Berger *et al.* (1988) formularam o seguinte teorema que calcula o nível de significância do teste das hipóteses dadas em (2.2).

Teorema 1 *Sejam X_1, X_2, \dots, X_{n_1} e Y_1, Y_2, \dots, Y_{n_2} como definidos anteriormente e z_α igual ao quantil $(1 - \alpha)$ da distribuição normal padrão. Se $E(X_1^2) < \infty$ e $E(Y_1^2) < \infty$ e $\min(n_1, n_2) \rightarrow \infty$, então*

1. $S_{n_1}^2(t) \rightarrow \text{Var}(X_1 | X_1 > t)$ e $S_{n_2}^2(t) \rightarrow \text{Var}(Y_1 | Y_1 > t)$, com probabilidade 1;

2. $A_{n_1 n_2}(t) \rightarrow N(0, 1)$ em distribuição;

$$3. P(Z_{n_1 n_2}(t) > z_\alpha) \rightarrow \begin{cases} \alpha, & \text{se } m_F(t) = m_G(t) \\ 0, & \text{se } m_F(t) < m_G(t) \\ 1, & \text{se } m_F(t) > m_G(t) \end{cases}.$$

Para provar o resultado 1, definimos $X^* = X.I(X > t)$ e $X_1^*, \dots, X_{n_1(t)}^*$ as observações que tiveram valor maior que t . Temos então

$$\begin{aligned} S_{n_1}^2(t) &= \frac{1}{n_1(t)} \sum_{i=1}^{n_1(t)} \left(X_i^* - \frac{\sum_{j=1}^{n_1(t)} X_j^*}{n_1(t)} \right)^2 = \frac{1}{n_1(t)} \sum_{i=1}^{n_1(t)} \left((X_i^*)^2 - 2X_i^* \frac{\sum_{j=1}^{n_1(t)} X_j^*}{n_1(t)} + \left(\frac{\sum_{j=1}^{n_1(t)} X_j^*}{n_1(t)} \right)^2 \right) \\ &= \frac{\sum_{i=1}^{n_1(t)} (X_i^*)^2}{n_1(t)} - 2 \frac{\sum_{i=1}^{n_1(t)} \sum_{j=1}^{n_1(t)} X_i^* X_j^*}{n_1^2(t)} + \left(\frac{\sum_{j=1}^{n_1(t)} X_j^*}{n_1(t)} \right)^2 = \frac{\sum_{i=1}^{n_1(t)} (X_i^*)^2}{n_1(t)} - \left(\frac{\sum_{j=1}^{n_1(t)} X_j^*}{n_1(t)} \right)^2. \end{aligned}$$

Pela Lei Fraca dos Grandes Números, a primeira parcela da soma acima converge em probabilidade para $E((X^*)^2)$ e a segunda parcela converge para $E(X^*)^2$. Sendo assim,

$$S_{n_1}^2(t) \rightarrow \text{Var}(X^*) = \text{Var}(X|X > t), \quad \text{com probabilidade 1.}$$

Usando o resultado 1, o teorema de Slutsky e o Teorema Central do Limite para as parcelas $\sum_{i=1}^{n_1} I(X_i > t)(X_i - t - m_F(t))$ e $\sum_{i=1}^{n_2} I(Y_i > t)(Y_i - t - m_G(t))$, o resultado 2 é provado.

Se $m_F(t) < m_G(t)$, então $m_F(t) - m_G(t) < 0$. Logo, $B_{n_1 n_2}(t)$ converge para $-\infty$ quando $n_1 \rightarrow \infty$. Sendo assim, $P(Z_{n_1 n_2}(t) > z_\alpha) \rightarrow 0$. Analogamente, quando $m_F(t) > m_G(t)$, $B_{n_1 n_2}(t)$ converge para ∞ . De modo que, $P(Z_{n_1 n_2}(t) > z_\alpha) \rightarrow 1$.

Se $m_F(t) = m_G(t)$, então $B_{n_1 n_2}(t) = 0$ e $Z_{n_1 n_2}(t) = A_{n_1 n_2}(t)$, que converge em distribuição para a distribuição Normal padrão. Portanto,

$$P(Z_{n_1 n_2}(t) > z_\alpha) \rightarrow \alpha.$$

Para testar as hipóteses em (2.1), Berger *et al.* (1988) propuseram as hipóteses (2.2) por meio do método de união-interseção que, como desenvolvido em Berger (1982), será descrito a seguir.

Seja θ o parâmetro de interesse do teste de hipóteses, Θ o espaço paramétrico e $\Theta_i, i = 1, \dots, k$, subconjuntos de Θ . As hipóteses são definidas da seguinte forma:

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1, \end{aligned} \tag{2.3}$$

em que $\Theta_0 = \bigcup_{i=1}^k \Theta_i$ e $\Theta_1 = \Theta_0^c \cap \bigcap_{i=1}^k \Theta_i^c$.

Hipóteses individuais também são definidos como

$$\begin{aligned} H_{0i} &: \theta \in \Theta_i \\ H_{1i} &: \theta \in \Theta_i^c, \end{aligned} \tag{2.4}$$

para $i = 1, \dots, k$.

Seja R_i a região de rejeição do teste com significância α para testar (2.4). Berger (1982) mostrou que o teste para (2.3) tem significância no máximo α , pois seja R a região de rejeição de H_0 , sendo $R = \bigcap_{i=1}^k R_i$. Então sob H_0 , temos

$$P(R) = P\left(\bigcap_{i=1}^k R_i\right) \leq P(R_i) \leq \alpha.$$

A região de rejeição da hipótese (2.2) é dada por $\{Z_{n_1 n_2}(t) > z_\alpha\}$. Usando o método de união-interseção, obtemos a região de rejeição do teste para as hipóteses em (2.1):

$$\bigcap_{t \in [T_1, T_2]} \{Z_{n_1 n_2}(t) > z_\alpha\} \text{ ou } \inf_{t \in [T_1, T_2]} Z_{n_1 n_2}(t) > z_\alpha.$$

Pelo Teorema 1, temos que para um t fixado, o teste para (2.2) é assintoticamente de nível α . Seja $t^* \in [T_1, T_2]$. Então,

$$P\left(\bigcap_{t \in [T_1, T_2]} Z_{n_1 n_2}(t) > z_\alpha\right) \leq P(Z_{n_1 n_2}(t^*) > z_\alpha) \leq P(A_{n_1 n_2}(t^*) > z_\alpha) \rightarrow \alpha;$$

portanto, o teste para (2.1) também é assintoticamente de nível α .

Uma limitação do teste apresentado em Berger *et al.* (1988) é que ele não considera observações censuradas em sua estatística de teste. A abordagem via verossimilhança empírica que será vista no Capítulo 3 leva em consideração as observações incompletas.

2.4 Modelo de regressão para o tempo médio residual

Oakes e Dasu (1990) propuseram um modelo de regressão baseado na proporcionalidade da *vida média residual*:

$$m(t) = \alpha m_0(t), \quad (2.5)$$

em que $\alpha > 0$ é uma constante no tempo e $m_0(t)$ é a *vida média residual* quando $\alpha = 1$, sendo chamada de *vida média residual* basal.

Usando a *fórmula de inversão* pode-se escrever $S(t)$ em função de $S_0(t)$ e $m_0(t)$:

$$\begin{aligned} S(t) &= S(t; \alpha) = \frac{m(t)}{m_0(t)} \exp\left(-\int_0^t \frac{du}{m(u)}\right) = \frac{m_0(t)}{m_0(t)} \exp\left(-\frac{1}{\alpha} \int_0^t \frac{du}{m_0(u)}\right) \\ &= \frac{m_0(t)}{m_0(t)} \exp\left(-\int_0^t \frac{du}{m_0(u)}\right) \exp\left(\int_0^t \frac{du}{m_0(u)}\right) \exp\left(-\frac{1}{\alpha} \int_0^t \frac{du}{m_0(u)}\right) \\ &= S_0(t) \exp\left((1 - 1/\alpha) \int_0^t \frac{du}{m_0(u)}\right) = S_0(t) \exp\left(\left(\frac{1}{\alpha} - 1\right) \log\left(\frac{S_0(t)m_0(t)}{m_0(0)}\right)\right) \\ &= S_0(t) \left(\frac{S_0(t)m_0(t)}{m_0(0)}\right)^{\frac{1}{\alpha}-1} = S_0(t) \left(\int_t^\infty \frac{S_0(u)du}{m_0(0)}\right)^{\frac{1}{\alpha}-1}. \end{aligned}$$

Note que como S_0 é uma função de sobrevivência e $0 < \alpha \leq 1$, $S(t)$ é sempre uma função de sobrevivência.

A função densidade $f(t; \alpha)$ também pode ser escrita em função de $S_0(t)$ e $m_0(t)$. Seja $m'_0(t) = \frac{dm_0(t)}{dt}$:

$$\begin{aligned}
\frac{dS(t; \alpha)}{dt} &= \frac{d m_0(t)}{dt} \exp \left(- \int_0^t \frac{du}{m_0(u)} \right) \\
&= m_0(t) \left[- \frac{m_0'(t)}{m_0^2(t)} \exp \left(- \frac{1}{\alpha} \int_0^t \frac{du}{m_0(u)} \right) + \frac{1}{m_0(t)} \exp \left(- \frac{1}{\alpha} \int_0^t \frac{du}{m_0(u)} \right) \frac{-1}{\alpha} \frac{1}{m_0(t)} \right] \\
&= - \frac{m_0'(t)}{m_0^2(t)} \exp \left(- \frac{1}{\alpha} \int_0^t \frac{du}{m_0(u)} \right) \left(\frac{dm_0(t)}{dt} + \frac{1}{\alpha} m_0(t) \right),
\end{aligned}$$

resultando na equação

$$f(t; \alpha) = \frac{m_0'(t)}{m_0^2(t)} e^{\int_0^t \frac{du}{\alpha m_0(u)}} \left(m_0'(t) + \frac{1}{\alpha} m_0(t) \right).$$

Oakes e Dasu (2003) discutem algumas famílias de distribuições em que (2.5) ocorre. Uma delas é a distribuição exponencial com $S_0(t) = e^{-\rho t}$, $\rho > 0$ e $S(t; \alpha) = e^{-\frac{\rho t}{\alpha}}$, $\alpha > 0$. Outra distribuição com essa característica é a Pareto com função de sobrevivência igual a $S_0(t) = (1 + \rho t)^{-\gamma}$, ($\gamma > 0, \rho > 0$), o que resulta em

$$m_0(t) = \frac{\int_t^\infty (1 + \rho u)^{-\gamma} du}{(1 + \rho t)^{-\gamma}} = \frac{\frac{(1 + \rho t)^{1-\gamma}}{\rho(1-\gamma)}}{(1 + \rho t)^{-\gamma}} = \frac{1 + \rho t}{\rho(1 - \gamma)}.$$

A função de sobrevivência fica igual a

$$\begin{aligned}
S(t; \alpha) &= (1 + \rho t)^{-\gamma} \left(\int_t^\infty \frac{(1 + \rho u)^{-\gamma}}{\rho(1-\gamma)} du \right)^{\frac{1}{\alpha}-1} = (1 + \rho t)^{-\gamma} \left(\frac{(1 + \rho t)^{1-\gamma}}{\rho(1-\gamma)} \right)^{\frac{1}{\alpha}-1} \\
&= (1 + \rho t)^{-\gamma} (1 + \rho t)^{(1-\gamma)(\frac{1}{\alpha}-1)} = \left(\frac{1}{1 + \rho t} \right)^{1 + \frac{\gamma-1}{\alpha}}.
\end{aligned}$$

A família de distribuições definida por Oakes e Dasu (2003) segue a suposição que $m_0(t) = e^{-\lambda t}$, com $\alpha < 1$ e $0 < \lambda < 1$. Assim, pela fórmula de inversão temos

$$S_0(t) = \frac{1}{e^{-\lambda t}} \exp \left(- \int_0^t e^{\lambda u} du \right) = e^{\lambda t} \exp \left(- \frac{1}{\lambda} (e^{\lambda t} - 1) \right) = e^{\lambda t} \exp \left(\frac{1 - e^{\lambda t}}{\lambda} \right),$$

de modo que

$$S(t; \alpha) = e^{\lambda t} \exp \left(\frac{1 - e^{\lambda t}}{\lambda} \right) \left(\int_t^\infty e^{\lambda u} \exp \left(\frac{1 - e^{\lambda u}}{\lambda} \right) du \right)^{\frac{1}{\alpha}-1}.$$

Fazendo uma transformação de variável para resolver a integral acima, por exemplo $x = \frac{1 - e^{\lambda u}}{\lambda}$, resulta em

$$S(t; \alpha) = e^{\lambda t} \exp \left(\frac{1 - e^{\lambda t}}{\lambda} \right) \left(e^{\frac{1 - e^{\lambda t}}{\lambda}} \right)^{\frac{1}{\alpha}-1} = \exp \left\{ \lambda t - \frac{(e^{\lambda t} - 1)}{\lambda \alpha} \right\}.$$

Para se fazer inferências sobre α pelo método de máxima verossimilhança, calcula-se o logaritmo da função de verossimilhança

$$\log f(t; \alpha) = \log m_0(0) - 2 \log m_0(t) - \frac{1}{\alpha} \int_0^t \frac{du}{m_0(u)} + \log \left\{ \frac{1}{\alpha} + \frac{dm_0(t)}{dt} \right\}.$$

A função escore, que é a derivada da expressão acima com relação a α , resulta em

$$\frac{\partial \log f(t; \alpha)}{\partial \alpha} = \frac{1}{\alpha^2} \int_0^t \frac{du}{m_0(u)} - \frac{1}{\alpha + \alpha^2 m'_0(t)},$$

com $m'_0(t) = \frac{dm_0(t)}{dt}$. Para o cálculo da informação de Fisher a segunda derivada de $\log f(t; \alpha)$ é dada por

$$\frac{\partial^2 \log f(t; \alpha)}{\partial \alpha^2} = -\frac{2}{\alpha^2} \int_0^t \frac{du}{m(u; \alpha)} + \frac{1}{\alpha^2} \left[1 - \frac{m'(t; \alpha)^2}{(1 + m'(t; \alpha))^2} \right],$$

em que $m(t; \alpha) = \alpha m_0(t)$ e $m'(t; \alpha) = \frac{dm(t; \alpha)}{dt}$.

Usando a relação entre a função de risco e a *vida média residual* mostrada no final da Seção 2.1, temos

$$m'(t) = -1 + m(t) \frac{f(t)}{S(t)},$$

para simplificar a notação, $f(t; \alpha) = f(t)$, $S(t; \alpha) = S(t)$ e $m(t; \alpha) = m(t)$. A informação de Fisher $I(\alpha)$ para cada observação é calculada por

$$\begin{aligned} I(\alpha) &= E \left\{ -\frac{\partial^2 \log f(t; \alpha)}{\partial \alpha^2} \right\} = \frac{1}{\alpha^2} \left[2 \int_0^\infty \int_0^t \frac{du}{m(u)} f(t) dt - 1 + \int_0^\infty \frac{m'(t)^2}{(1 + m'(t))^2} f(t) dt \right] \\ &= \frac{1}{\alpha^2} \left[\int_0^\infty \int_u^\infty \frac{f(t)}{m(u)} dt du - 1 + \int_0^\infty \frac{\frac{m(t)^2 f(t)^2}{S(t)^2} - 2 \frac{m(t) f(t)}{S(t)} + 1}{\frac{m(t)^2 f(t)^2}{S(t)^2}} f(t) dt \right] \\ &= \frac{1}{\alpha^2} \left[2 \int_0^\infty \frac{1}{m(u)} \int_u^\infty f(t) dt du - 1 + \int_0^\infty f(t) dt - 2 \int_0^\infty \frac{S(t)}{m(t)} dt + \int_0^\infty \frac{S(t)^2}{m(t)^2 f(t)} dt \right] \\ &= \frac{1}{\alpha^2} \left[2 \int_0^\infty \frac{S(u)}{m(u)} du - 1 + 1 - 2 \int_0^\infty \frac{S(t)}{m(t)} dt + \int_0^\infty \frac{S(t)^2}{m(t)^2 f(t)} dt \right] \\ &= \frac{1}{\alpha^2} \int_0^\infty \frac{S(t; \alpha)^2}{m(t; \alpha)^2 f(t; \alpha)} dt. \end{aligned}$$

Portanto, quando $S_0(t)$ é conhecida e com $\hat{\alpha}$ sendo o estimador de máxima verossimilhança de α , $\sqrt{n}(\hat{\alpha} - \alpha)$ tem distribuição assintótica normal com média 0 e variância $I(\alpha)^{-1}$.

Maguluri e Zhang (1994) estenderam o modelo (2.5) com a inclusão de covariáveis, fazendo

$$m(t|\mathbf{z}) = e^{\mathbf{z}'\boldsymbol{\beta}} m_0(t). \quad (2.6)$$

Karlin e Taylor (1975) obtêm um processo de renovação estacionário formado por variáveis aleatórias $\{X_k\}$ não negativas, com X_2, X_3, \dots identicamente distribuídas com função de distribuição F e X_1 com função de distribuição G dada por

$$G(t) = \frac{\int_0^t (1 - F(u)) du}{E(X_2)}.$$

A função de sobrevivência \bar{G} associada a $G(t)$ é dada por

$$\bar{G}(t) = 1 - \frac{\int_0^t S(u)du}{E(X_2)} = 1 - \frac{\int_0^t S(u)du}{\int_0^\infty S(u)du} = \frac{\int_t^\infty S(u)du}{\mu}.$$

Como $\bar{G}'(t) = -\frac{S(t)}{\mu}$, $m_{X_2}(t)$, a função de *vida média residual* associada a X_2 , é dada por

$$m(t) = \frac{\int_t^\infty S(u)du}{S(t)} = -\frac{\bar{G}(t)}{\bar{G}'(t)} = \frac{1}{\lambda_{\bar{G}}(t)},$$

em que $\lambda_{\bar{G}}(t)$ é a função de risco associada a \bar{G} .

Para $F(t|\mathbf{z}) = P(T \leq t|\mathbf{z})$, $S(t|\mathbf{z}) = 1 - F(t|\mathbf{z})$, considere um processo de renovação em que as renovações têm função de distribuição igual a $F(t|\mathbf{z})$. O tempo de recorrência, V , é definido como o tempo até a próxima renovação partindo de um tempo fixado, de tal forma que a função densidade de V é dada por $f_V(\cdot|\mathbf{z}) = \frac{\hat{S}(t|\mathbf{z})}{m(0|\mathbf{z})}$. Sendo assim, a função de risco associada a V é igual a

$$\lambda_V(t|\mathbf{z}) = \frac{f_V(t|\mathbf{z})}{S_V(t|\mathbf{z})} = \frac{\frac{S(t|\mathbf{z})}{m(0|\mathbf{z})}}{\frac{\int_t^\infty S(u|\mathbf{z})du}{m(0|\mathbf{z})}} = \frac{1}{m(t|\mathbf{z})}.$$

Portanto, o modelo (2.6) induz a um modelo de riscos proporcionais de Cox na forma $\lambda_V(t|\mathbf{z}) = e^{-\mathbf{z}'\beta} \lambda_{V,0}(t)$. O estimador de β pode ser encontrado por meio da solução da equação

$$\hat{E}(\mathbf{z}) - \int_0^\tau \frac{\hat{E}[\mathbf{z}e^{-\mathbf{z}'\beta}I(V \geq t)]}{\hat{E}[e^{-\mathbf{z}'\beta}I(V \geq t)]} d\hat{F}_V(t) = \mathbf{0}, \quad (2.7)$$

com τ sendo o limite superior dos tempos. Maguluri e Zhang (1994) utilizaram o argumento de que o uso de V é apenas teórico e mostraram que

$$\begin{aligned} E(I(V \geq t)|\mathbf{z}) &= P(V \leq t|\mathbf{z}) = S_V(t|\mathbf{z}) = \frac{\int_t^\infty S(u|\mathbf{z})du}{m(0|\mathbf{z})} = \frac{\int_t^\infty (u-t)f(u|\mathbf{z})du}{m_0(0|\mathbf{z})e^{\mathbf{z}'\beta}} \\ &= \frac{E((T-t)I(T \geq t))}{m_0(0|\mathbf{z})e^{\mathbf{z}'\beta}} = \frac{E[e^{-\mathbf{z}'\beta}(T-t)^+|\mathbf{z}]}{m_0(0)} e \\ dF_V(t) &= f_V(t)dt = \frac{S(t|\mathbf{z})}{m(0|\mathbf{z})} = \frac{E(I(T \geq t))}{E(T)} = \frac{E[e^{-\mathbf{z}'\beta}I(T > t)]}{E[e^{-\mathbf{z}'\beta}T]} dt \end{aligned}$$

em que $(T-t)^+ = (T-t)I(T > t)$.

Com essas igualdades, a expressão (2.7) pode ser escrita como

$$\hat{E}(\mathbf{z}) - \int_0^\tau \frac{\hat{E}[\mathbf{z}e^{-2\mathbf{z}'\beta}(T-t)^+]}{\hat{E}[e^{-2\mathbf{z}'\beta}(T-t)^+]} \frac{\hat{E}[e^{-\mathbf{z}'\beta}I(T \geq t)]}{\hat{E}[e^{-\mathbf{z}'\beta}T]} dt = \mathbf{0}.$$

Por não considerar censura, um estimador empírico para as esperanças pode ser usado na expressão acima, resultando em

$$\hat{\rho}(\beta) = \frac{\sum_{i=1}^n \mathbf{z}_i}{n} - \int_0^\tau \frac{\sum \mathbf{z}_i e^{-2\mathbf{z}'_i\beta}(T_i-t)^+}{\sum e^{-2\mathbf{z}'_i\beta}(T_i-t)^+} \frac{\sum e^{-\mathbf{z}'_i\beta}I(T_i \geq t)}{\sum e^{-\mathbf{z}'_i\beta}T_i} dt = \mathbf{0}, \quad (2.8)$$

com $\hat{\beta}$ sendo a solução de (2.8).

A expressão limite de $\hat{\rho}$ é ρ (Maguluri e Zhang, 1994), dada por

$$\rho(\beta) = E(\mathbf{z}) - \int_0^\infty E_{t,\beta}(\mathbf{z}) E_\beta \left(\frac{S(t|\mathbf{z})}{E(T|\mathbf{z})} \right) dt$$

em que

$$E_{t,\beta}(g(\mathbf{z})) = \frac{E(g(\mathbf{z})e^{2\mathbf{z}'\beta(T-t)^+})}{E(e^{2\mathbf{z}'\beta(T-t)^+})} \text{ e } E_{\beta}(g(\mathbf{z})) = \frac{E(g(\mathbf{z})e^{\mathbf{z}'\beta T})}{E(e^{\mathbf{z}'\beta T})}.$$

A proposição e o teorema a seguir mostram a convergência assintótica de $\hat{\beta}$.

Proposição 1: Seja ρ como definido acima. Então,

$$\rho'(\beta) = - \int_0^{\infty} 2\text{var}_{t,\beta}(\mathbf{z})E_{\beta} \left\{ \frac{S(t|\mathbf{z})}{E(T|\mathbf{z})} \right\} dt + \int_0^{\infty} \text{cov}_{\beta} \left\{ \mathbf{z}, \frac{S(t|\mathbf{z})}{E(T|\mathbf{z}=0)} \right\} \text{cov}_{t,\beta}(\mathbf{z}, e^{\mathbf{z}'\beta}) dt < 0.$$

A prova da Proposição 1 encontra-se no Apêndice A.2 de Maguluri e Zhang (1994) e usa as seguintes relações:

$$\frac{\partial}{\partial \beta} E_{t,\beta}(\mathbf{z}) = 2\text{var}_{t,\beta}(\mathbf{z}) \text{ e } \frac{\partial}{\partial \beta} E_{\beta}(g(\mathbf{z})) = \text{cov}_{\beta}(\mathbf{z}, g(\mathbf{z})).$$

Antes de apresentar o teorema, vamos definir algumas quantidades:

$$S^{(k)}(t) = E((T-t)^k I(T \geq t)),$$

$$\begin{aligned} \xi = \mathbf{z} - \int_0^{\infty} w(t)f(t) & \left\{ \frac{\mathbf{z}e^{-2\mathbf{z}'\beta_0(T-t)^+}}{E(\mathbf{z}e^{-2\mathbf{z}'\beta_0(T-t)^+})} - \frac{e^{-2\mathbf{z}'\beta_0(T-t)^+}}{E(e^{-2\mathbf{z}'\beta_0(T-t)^+})} \right. \\ & \left. + \frac{e^{-\mathbf{z}'\beta_0 I(T>t)}}{E(e^{-\mathbf{z}'\beta_0})S(t|\mathbf{z})} - \frac{e^{-\mathbf{z}'\beta_0 T}}{E(T|\mathbf{z}=0)} \right\} dt, \end{aligned}$$

em que

$$w(t) = \frac{E(\mathbf{z}e^{-2\mathbf{z}'\beta(T-t)^+})}{E(e^{-2\mathbf{z}'\beta(T-t)^+})} \quad \text{e} \quad f(t) = \frac{E(e^{\mathbf{z}'\beta} I(T > t))}{E(e^{\mathbf{z}'\beta T})}.$$

Teorema 2 Suponha que \mathbf{z} seja limitado; então, $\hat{\beta} \rightarrow \beta_0$, quase certamente, em que β_0 é o verdadeiro valor de β . Se $\text{Var}(\xi) < \infty$ e

$$\int_0^{\infty} S(t) \frac{\sqrt{S^{(2)}(t)}}{S^{(1)}(t)} dt < \infty,$$

então

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2), \text{ com } \sigma^2 = \frac{\text{Var}(\xi)}{\rho'(\beta_0)^2}.$$

O Teorema 2 é provado no Apêndice A.1 de Maguluri e Zhang (1994).

Chen *et al.* (2005) estudaram a estimação de β para o modelo (2.6) com a presença de censura utilizando a abordagem da Probabilidade Ponderada Inversa de Censura (*inverse probability of censoring weighted, IPCW*). Chen *et al.* (2005) estipulam que funções bem definidas são funções identidade ou funções escada, por exemplo. Sejam $X_i = \min(T_i, C_i)$ e $\delta_i = I(T_i \leq C_i)$. Supondo que C_i tem função de distribuição $F_C(\cdot)$ e função de sobrevivência $S_C(\cdot)$, para qualquer função bem definida H :

$$\begin{aligned} E\left(\frac{H(X_i, \mathbf{z}_i, t)\delta_i}{S_C(X_i)}\right) &= E\left(\frac{H(T_i, \mathbf{z}_i, t)\delta_i}{S_C(T_i)}\right) = E\left\{E\left(\frac{H(T_i, \mathbf{z}_i, t)\delta_i}{S_C(T_i)}\middle|\mathbf{z}_i\right)\right\} \\ &= E\left\{E\left(\frac{H(T_i, \mathbf{z}_i, t)}{S_C(T_i)}E(I(C_i \geq T_i))\middle|\mathbf{z}_i, T_i\right)\right\} = E[H(T_i, \mathbf{z}_i, t)], \end{aligned}$$

para $H(X_i, \mathbf{z}_i, t) = \mathbf{z}_i$, $H(X_i, t) = (X_i - t)^+$ e $H(X_i, t) = I(X_i \geq t)$. Então,

$$\begin{aligned} E\left\{\frac{\mathbf{z}_i\delta_i}{S_C(X_i)}\right\} &= E(\mathbf{z}_i), \quad E\left\{\frac{(X_i - t)^+\delta_i}{S_C(X_i)}\right\} = E((T_i - t)^+) \quad e \\ E\left\{\frac{I(X_i \geq t)\delta_i}{S_C(X_i)}\right\} &= E[I(T_i \geq t)]. \end{aligned}$$

Chen *et al.* (2005) substituíram \mathbf{z}_i , $(T_i - t)^+$ e $I(T_i \geq t)$ por $\frac{\mathbf{z}_i\delta_i}{\hat{S}_C(X_i)}$, $\frac{(X_i - t)^+\delta_i}{\hat{S}_C(X_i)}$ e $\frac{I(X_i \geq t)\delta_i}{\hat{S}_C(X_i)}$, respectivamente, na equação (2.8). Com $\hat{S}_C(\cdot)$ sendo um estimador para S_C , por exemplo o estimador de Kaplan-Meier, um conjunto de equações de estimação de $\boldsymbol{\beta}$ foi proposto:

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\hat{S}_C(X_i)} \{\mathbf{z}_i - \tilde{\mathbf{z}}_i(\boldsymbol{\beta}, \hat{S}_C)\},$$

em que

$$\tilde{\mathbf{z}}_i(\boldsymbol{\beta}, \hat{S}_C) = \int_0^\tau \frac{n\mathbf{z}_i e^{-2\mathbf{z}'_i\boldsymbol{\beta}}(X_i - t)^+ I(X_i > t)}{\sum_{j=1}^n \frac{e^{-2\mathbf{z}'_j\boldsymbol{\beta}}(X_j - t)^+\delta_j}{\hat{S}_C(X_j)}} \frac{\sum_{j=1}^n \frac{e^{-\mathbf{z}'_j\boldsymbol{\beta}}\delta_j}{\hat{S}_C(X_j)}}{\sum_{j=1}^n \frac{e^{-\mathbf{z}'_j\boldsymbol{\beta}}\delta_j X_j}{\hat{S}_C(X_j)}} dt.$$

O estimador $\hat{\boldsymbol{\beta}}$ será a solução da equação $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$. Para o cálculo da distribuição assintótica de $\hat{\boldsymbol{\beta}}$, considera-se o estimador de Kaplan-Meier na forma martingal:

$$\frac{\hat{S}_C(t) - S(t)}{S(t)} = - \int_0^t \frac{\hat{S}_C(s-)}{S_C(s)} \frac{dM_C(s)}{Y(s)},$$

com $Y(s) = \sum_{j=1}^n I(X_j \geq s)$, $M_C(t) = \sum_{j=1}^n M_{C,j}(t) = \sum_{j=1}^n I(X_j \leq t, \delta_j = 0) - \int_0^t I(X_j \geq s) d\Lambda_C(s)$ e $\Lambda_C(s) = -\log(1 - S_C(s))$. Como $\{M_{C,j}(t), j = 1, \dots, n\}$ são martingais com respeito a uma filtragem \mathcal{F}_C gerada por um conjunto de σ -álgebras dado por $\sigma\{I(C_i \leq s), s \leq t; I(T_i \leq u), \mathbf{z}_i(u), \leq u \leq \tau, i = 1, \dots, n\}$, $\mathbf{U}(\boldsymbol{\beta}) \xrightarrow{P} \mathbf{u}(\boldsymbol{\beta})$, em que

$$\mathbf{u}(\boldsymbol{\beta}) = E\left[\frac{\delta}{S_C(X)} \mathbf{z} - \tilde{\mathbf{z}}(\boldsymbol{\beta}, S_C)\right] = E[\mathbf{z}] - \int_0^\tau \frac{E[\mathbf{z}e^{-2\mathbf{z}'\boldsymbol{\beta}(T-t)^+}]}{E[e^{-2\mathbf{z}'\boldsymbol{\beta}(T-t)^+}]} \frac{E[e^{-\mathbf{z}'\boldsymbol{\beta}}I(T > t)]}{E[e^{-\mathbf{z}'\boldsymbol{\beta}T}]} dt$$

e, similar ao caso sem censura, se $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\beta}_0$ verdadeiro valor de $\boldsymbol{\beta}$, então $\mathbf{u}(\boldsymbol{\beta}_0) = 0$. E assim, como no caso sem censura,

$$\mathbf{u}'(\boldsymbol{\beta}) = -2 \int_0^\tau \text{Var}(\mathbf{z}) E\left[\frac{S(t|\mathbf{z})}{m(0|\mathbf{z})}\right] dt + \int_0^\tau \text{cov}[\mathbf{z}, S(t|\mathbf{z})] \text{cov}[\mathbf{z}, e^{\mathbf{z}'\boldsymbol{\beta}}] dt,$$

e $\mathbf{u}'(\boldsymbol{\beta})$ é definida negativa (Chen *et al.*, 2005).

Chen *et al.* (2005) mostraram que $\hat{\boldsymbol{\beta}}$ é consistente e único na vizinhança de $\boldsymbol{\beta}_0$ (verdadeiro valor de $\boldsymbol{\beta}$). Por meio da expansão de Taylor, é mostrado que $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ é assintoticamente equivalente

a $\frac{\partial \mathbf{U}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \sqrt{n} \mathbf{U}(\boldsymbol{\beta}_0)$. Sendo assim, $\mathbf{U}(\boldsymbol{\beta}) \xrightarrow{\mathcal{P}} \mathbf{u}(\boldsymbol{\beta})$ uniformemente e $\frac{\partial \mathbf{U}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \xrightarrow{\mathcal{P}} \mathbf{u}'(\boldsymbol{\beta}_0)$. Também foi mostrado no Apêndice A de [Chen et al. \(2005\)](#) que $\sqrt{n} \mathbf{U}(\boldsymbol{\beta}_0)$ tem distribuição assintótica normal. Logo,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N\left(0, [\mathbf{u}'(\boldsymbol{\beta}_0)]^{-1} \Sigma(\boldsymbol{\beta}_0, S_C) \{[\mathbf{u}'(\boldsymbol{\beta}_0)]^{-1}\}'\right),$$

a expressão de $\Sigma(\boldsymbol{\beta}_0, S_C)$ é encontrada no Apêndice de [Chen et al. \(2005\)](#).

A prova desse resultado é feita a partir da aplicação do método Delta em $\mathbf{U}(\boldsymbol{\beta}_0)$ e na razão $\frac{\delta_i}{\hat{S}_C(X_i)}$, quando aplicado nesse último resulta em uma expressão similar ao estimador de Kaplan-Meier escrito na forma martingal. Com alguma manipulação algébrica e usando o Teorema Central do Limite multivariado, [Chen et al. \(2005\)](#) mostraram a convergência de $\hat{\boldsymbol{\beta}}$.

Portanto, a distribuição de $\hat{\boldsymbol{\beta}}$ pode ser aproximada por uma distribuição normal com média $\boldsymbol{\beta}_0$ e matriz de variâncias-covariâncias estimada por

$$\left[\frac{\partial \mathbf{U}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right]^{-1} \hat{\Sigma}(\hat{\boldsymbol{\beta}}, S_C) \left\{ \left[\frac{\partial \mathbf{U}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \right]^{-1} \right\}',$$

em que $\frac{\partial \mathbf{U}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}$ é $\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ aplicada em $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

Em aplicações práticas, no entanto, o cálculo da estimativa da variância é complicada e [Chen et al. \(2005\)](#) sugerem fazer a estimação via *bootstrap*.

[Sun e Zhang \(2009\)](#) propuseram uma generalização do modelo (2.6), usando uma função g :

$$m(t|\mathbf{z}) = g(m_0(t) + \mathbf{z}'\boldsymbol{\beta}). \quad (2.9)$$

Quando $g(x) = e^x$, (2.9) é o modelo (2.6) e se $g(x) = x$, (2.9) se reduz ao chamado modelo aditivo. Uma escolha da função g que deixa o modelo mais flexível é a transformação de Box-Cox dada por $g(x) = \frac{(x+1)^\rho - 1}{\rho}$, com $\rho = 0$ implicando em $g(x) = \log(x+1)$.

[Sun e Zhang \(2009\)](#) estabeleceram três condições para a função g :

- g é derivável até segunda ordem;
- g é estritamente crescente e
- $g(m_0(t) + \mathbf{z}'\boldsymbol{\beta})$ é uma função de *vida média residual* para todos os valores possíveis de \mathbf{z} .

Para o cenário com censura considerado anteriormente e com $\boldsymbol{\beta}_0$ o verdadeiro valor de $\boldsymbol{\beta}$, define-se

$$M_i(t) = \frac{\delta_i I(X_i > t)}{S_C(X_i)} [(X_i - t) - g\{m_0(t) + \mathbf{z}'_i \boldsymbol{\beta}_0\}], \quad i = 1, \dots, n.$$

Sob o modelo (2.9), $M_i(t)$ é um processo estocástico de média 0, portanto para um dado $\boldsymbol{\beta}$, um estimador $\hat{m}_0(t; \boldsymbol{\beta})$ razoável para $m_0(t)$ é a solução de

$$\sum_{i=1}^n \frac{\delta_i I(X_i > t)}{\hat{S}_C(X_i)} [(X_i - t) - g(m_0(t) + \mathbf{z}'_i \boldsymbol{\beta})] = 0, \quad 0 \leq t \leq \tau. \quad (2.10)$$

Para estimar $\boldsymbol{\beta}_0$, [Sun e Zhang \(2009\)](#), usando a abordagem *IPCW* propuseram a seguinte equa-

ção de estimação:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \frac{\delta_i I(X_i > t) \mathbf{z}_i}{\hat{S}_C(X_i)} [(X_i - t) - g(\hat{m}_0(t; \boldsymbol{\beta}) + \mathbf{z}'_i \boldsymbol{\beta})] dH(t) = \mathbf{0}, \quad (2.11)$$

em que $H(t)$ é uma função não decrescente em $[0, \tau]$. Sun e Zhang (2009) usaram três diferentes funções $H(t)$ em seu artigo: $H_1(t) = \sum_{i=1}^n \delta_i I(X_i \leq t)$, H_2 uma função escada com saltos de tamanho 1 em sete pontos fixados de tempo e $H_3 = 0$, se $t < 0$ e $H_3 = 1$, caso contrário.

Antes de apresentar o teorema que mostra a convergência de $\hat{\boldsymbol{\beta}}$ e $\hat{m}_0(t)$, vamos definir algumas quantidades:

$$N_i^c(t) = I(X_i \leq t, \delta_i = 0), \hat{\pi}(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \geq t),$$

$$\hat{\Lambda}^c(t) = \frac{1}{n} \sum_{i=1}^n \int_0^t \frac{dN_i^c(u)}{\hat{\pi}},$$

$$\hat{M}_i^c(t) = N_i^c(t) - \int_0^t I(X_i \geq u) d\hat{\Lambda}^c(u),$$

$$\hat{M}_i(t) = \frac{\delta_i I(X_i > t)}{\hat{S}_C(X_i)} \left((X_i - t) - g(\hat{m}_0(t) + \mathbf{z}'_i \boldsymbol{\beta}) \right) \quad e$$

$$\hat{Q}(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \geq t) \int_0^\tau \hat{m}_i(u) (\mathbf{z}_i - \bar{\mathbf{Z}}_a(u; \hat{\boldsymbol{\beta}})) dH(u), \text{ com}$$

$$\bar{\mathbf{Z}}_a(t; \hat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^n \frac{\delta_i}{\hat{S}_C(X_i)} I(X_i > t) g'(\hat{m}_0(t) + \mathbf{z}'_i \boldsymbol{\beta}) \mathbf{z}_i}{\sum_{i=1}^n \frac{\delta_i}{\hat{S}_C(X_i)} I(X_i > t) g'(\hat{m}_0(t) + \mathbf{z}'_i \boldsymbol{\beta})},$$

$g'(x)$ é a derivada de $g(x)$ e $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'$ para \mathbf{v} um vetor coluna.

Teorema 3 *Sob as condições de regularidade*

- $H(t)$ converge quase certamente e uniformemente em $t \in [0, \tau]$ para $\tilde{H}(t)$, que é uma função não aleatória e limitada;
- $\boldsymbol{\beta}$ é limitado e $S_C(t)$ é contínua;
- $m_0(t)$ é diferenciável em $[0, \tau]$;
- \mathbf{A} : $E \left[\int_0^\tau I(T_i > t) (\mathbf{z}_i - \bar{\mathbf{z}}_a(t))^{\otimes 2} g(m_0(t) + \mathbf{z}'_i \boldsymbol{\beta}_0) dH(t) \right]$, em que $\bar{\mathbf{z}}_a(t)$ é o limite de $\bar{\mathbf{Z}}_a(u; \boldsymbol{\beta}_0)$,

os seguintes resultados são verdadeiros:

- a. $\hat{\boldsymbol{\beta}}$ e $\hat{m}_0(t)$ sempre existem e são únicos e consistentes;
- b. $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ tem distribuição assintótica normal com média 0 e matriz de variâncias-covariâncias que pode ser consistentemente estimada por $\hat{\mathbf{A}}^{-1} \Sigma_a \hat{\mathbf{A}}^{-1}$, em que $\Sigma_a = \frac{\sum_{i=1}^n \hat{\xi}_i^{\otimes 2}}{n}$,

$$\hat{\xi}_i = \int_0^\tau \hat{M}_i(t) (\mathbf{z}_i - \bar{\mathbf{Z}}_a(u; \hat{\boldsymbol{\beta}})) dH(t) + \int_0^\tau \frac{\hat{Q}(t)}{\hat{\pi}(t)} d\hat{M}_i^c(t) \quad e$$

$$\hat{\mathbf{A}} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau (\mathbf{z}_i - \bar{\mathbf{Z}}(t; \hat{\boldsymbol{\beta}}))^{\otimes 2} g(\hat{m}_0(t) + \mathbf{z}'_i \hat{\boldsymbol{\beta}}) dH(t);$$

- c. $\sqrt{n}(\hat{m}_0(t) - m_0(t))$, ($0 \leq t \leq \tau$) converge fracamente para um processo gaussiano de média 0 e função de covariância que pode ser consistentemente estimada por $\hat{\Gamma}_a(s, t) = \frac{1}{n} \sum_{i=1}^n \hat{\phi}_i(s) \hat{\phi}_i(t)$, em que

$$\hat{\phi}(t) = \frac{1}{\hat{\Phi}_a(t)} \left(\hat{M}_i(t) + \int_0^\tau \frac{\hat{R}_a(t, u)}{\hat{\pi}(u)} d\hat{M}_i^c(u) \right) - \bar{\mathbf{Z}}_a(t; \hat{\boldsymbol{\beta}})' \hat{\mathbf{A}}^{-1} \hat{\xi}_i,$$

$$\hat{\Phi}_a(t) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i I(X_i > t)}{\hat{S}_C(X_i)} g(\hat{m}_0(t) + \mathbf{z}'_i \hat{\boldsymbol{\beta}}) \text{ e } \hat{R}_a(t, u) = \frac{1}{n} \sum_{i=1}^n \hat{M}_i(t) I(X_i \geq u).$$

A prova do Teorema 3 é feita no relatório de suplemento técnico de Sun e Zhang (2009). A normalidade assintótica de $\hat{m}_0(t)$ e a consistência do estimador $\hat{\Gamma}_a$ permitem a construção de intervalos de confiança pontuais para $m_0(t)$, para t fixado (Sun e Zhang, 2009).

Capítulo 3

Vida Mediana Residual

3.1 Abordagem via função de estimação

A função de *vida mediana residual*, $\theta(t)$ é definida por

$$P(T - t > \theta(t) | T > t) = \frac{1}{2} \Rightarrow \frac{P(T > \theta(t) + t)}{P(T > t)} = \frac{1}{2} \Rightarrow S(\theta(t) + t) = \frac{1}{2}S(t) \Rightarrow$$
$$\theta(t) = S^{-1}\left(\frac{1}{2}S(t)\right) - t, \text{ supondo } S(\cdot) \text{ inversível.}$$

Uma notação que será usada ao longo deste texto é $\theta(t) = \text{mediana}(T - t | T > t)$.

Seja θ_{t^*} a *vida mediana residual* para um tempo t^* , um estimador $\hat{\theta}_{t^*}$ para θ_{t^*} pode ser escrito como

$$\hat{\theta}(t^*) = \hat{S}^{-1}\left(\frac{1}{2}\hat{S}(t^*)\right) - t^*,$$

em que $\hat{S}(t)$ é o estimador de Kaplan-Meier. Utilizando as expressões acima, Jeong (2014) propôs a *função de estimação* $u(\theta_{t^*})$ como

$$u(\theta_{t^*}) = \hat{S}(t^* + \theta_{t^*}) - \frac{1}{2}\hat{S}(t^*), \quad (3.1)$$

e $\hat{\theta}_{t^*}$ é a solução de $u(\theta_{t^*}) = 0$.

Como $\hat{S}(t)$ é uniformemente consistente para $S(t)$ (Fleming e Harrington, 1991), $u(\theta_{t^*})$ converge para $\dot{u}(\theta_{t^*})$, em que

$$\dot{u}(\theta_{t^*}) = S(t^* + \theta_{t^*}) - \frac{1}{2}S(t^*).$$

De acordo com Jeong (2014), $\dot{u}(\theta_{t^*}) = 0$ e esse resultado, conjuntamente com a convergência de $u(\theta_{t^*})$, prova que $\hat{\theta}_{t^*}$ é um estimador consistente para θ_{t^*} .

Sejam $N_i(t) = I(X_i \leq t, \delta_i = 1)$, $N(t) = \sum_{i=1}^n N_i(t)$, $Y_i(t) = I(X_i \geq t)$, $\sum_{i=1}^n Y_i(t)$ e a função de risco acumulada $\Lambda(t)$, então $M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda(s)$ é um processo martingal (Fleming e Harrington, 1991) assim como $M(t) = \sum_{i=1}^n M_i(t)$. O estimador de Kaplan-Meier pode ser escrito na forma de processo de contagem:

$$\hat{S}(t) = \prod_{i=1}^n \left(1 - \int_0^t \frac{I(Y(s) > 0)}{Y(s)} dN(s) \right).$$

Fleming e Harrington (1991) usaram os seguintes resultados para U, V e W funções contínuas à direita e limitadas em um intervalo finito $[0, t]$:

$$U(t)V(t) = U(0)V(0) + \int_0^t U(s^-)dV(s) + \int_0^t V(s)dU(s) \quad \text{e} \quad d\frac{1}{W(s)} = -\frac{dW(s)}{W(s)W(s^-)}$$

e mostraram que, se $S(t) > 0$, então

$$\frac{\hat{S}(t)}{S(t)} = 1 - \int_0^t \frac{\hat{S}(s^-)}{S(s)} \left(\frac{dN(s)}{Y(s)} - d\Lambda(s) \right).$$

Utilizando a expressão acima, temos

$$\begin{aligned} \hat{S}(t) - S(t) &= -S(t) \int_0^t \frac{\hat{S}(s^-)}{S(s)} \left(\frac{dN(s)}{Y(s)} - [I(Y(s) > 0)d\Lambda(s) + I(Y(s) = 0)d\Lambda(s)] \right) \\ &= -S(t) \int_0^t \frac{\hat{S}(s^-)}{S(s)} \left(\frac{I(Y(s) > 0)dM(s)}{Y(s)} - I(Y(s) = 0)d\Lambda(s) \right) \\ &= -S(t) \int_0^t \frac{\hat{S}(s^-)}{S(s)} \frac{I(Y(s) > 0)dM(s)}{Y(s)} + B(t), \end{aligned} \quad (3.2)$$

em que $B(t) = S(t) \int_0^t \frac{\hat{S}(s^-)}{S(s)} I(Y(s) = 0)d\Lambda(s)$.

Sejam $t_y = \inf\{s : Y(s) = 0\}$ e $\hat{S}(t) = \hat{S}(t_y)$ para $t \geq t_y$. Então,

$$\begin{aligned} B(t) &= S(t)I(t_y \leq t)\hat{S}(t_y) \int_{t_y}^t \frac{S(s^-)}{S(s)S(s^-)}d\Lambda(s) = S(t)I(t_y \leq t)\hat{S}(t_y) \int_{t_y}^t -\frac{dS(s)}{S(s)S(s^-)} \\ &= S(t)I(t_y \leq t)\hat{S}(t_y) \int_{t_y}^t d\left(\frac{1}{S(s)}\right) = S(t)\hat{S}(t_y)I(t_y \leq t) \left(\frac{1}{S(t)} - \frac{1}{S(t_y)} \right) \\ &= S(t)\hat{S}(t_y)I(t_y < t) \left(\frac{S(t_y) - S(t)}{S(t)S(t_y)} \right) = I(t_y < t)\hat{S}(t_y) \frac{S(t_y) - S(t)}{S(t_y)}. \end{aligned}$$

Jeong (2014) reescreveu o resultado (3.2) para n grande como

$$\hat{S}(t) - S(t) = -S(t) \int_0^t \frac{dM(\nu)}{Y(\nu)} + o_p(n^{-\frac{1}{2}}) = -S(t) \sum_{i=1}^n \int_0^t \frac{dM_i(\nu)}{Y(\nu)} + o_p(n^{-\frac{1}{2}}), \quad (3.3)$$

em que $o_p(n^{-\frac{1}{2}})$ é um termo que converge em probabilidade para 0 quando multiplicado por \sqrt{n} . Para justificar esse resultado, temos que pela desigualdade de Markov,

$$\begin{aligned} P(|\sqrt{n}B(t) - 0| \geq \epsilon) &\leq \frac{E(\sqrt{n}B(t))}{\epsilon} = \frac{\sqrt{n}}{\epsilon} E \left(I(t_y < t)\hat{S}(t_y) \frac{S(t_y) - S(t)}{S(t_y)} \right) \\ &\leq \frac{\sqrt{n}}{\epsilon} E \left(I(t_y < t) \frac{S(t_y) - S(t)}{S(t_y)} \right) \leq \frac{\sqrt{n}}{\epsilon} E(I(t_y < t)(1 - S(t))) \\ &= \frac{\sqrt{n}}{\epsilon} (1 - S(t))P(Y(t) = 0) = \frac{\sqrt{n}}{\epsilon} (1 - S(t))P(Y_1(t) = 0)^n \\ &= \frac{\sqrt{n}}{\epsilon} (1 - S(t))P(X_1 < t)^n \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Substituindo (3.3) em (3.1), temos

$$\begin{aligned}
u(\theta^*) &= -S(t^* + \theta_{t^*}) \int_0^{t^* + \theta_{t^*}} \frac{dM(v)}{Y(v)} + S(t^* + \theta_{t^*}) - \frac{1}{2} \left[-S(t^*) \int_0^{t^*} \frac{dM(v)}{Y(v)} + S(t^*) \right] + o_p(n^{-\frac{1}{2}}) \\
&= -S(t^* + \theta_{t^*}) \int_0^{t^* + \theta_{t^*}} \frac{dM(v)}{Y(v)} + \frac{1}{2} S(t^*) \int_0^{t^*} \frac{dM(v)}{Y(v)} + S(t^* + \theta_{t^*}) - \frac{1}{2} S(t^*) + o_p(n^{-\frac{1}{2}}) \\
&= - \left(S(t^* + \theta_{t^*}) - \frac{1}{2} S(t^*) \right) \int_0^{t^* + \theta_{t^*}} \frac{dM(v)}{Y(v)} - \frac{1}{2} \int_{t^*}^{t^* + \theta_{t^*}} \frac{dM(v)}{Y(v)} + 0 + o_p(n^{-\frac{1}{2}}) \\
&= -\frac{1}{2} S(t^*) \int_{t^*}^{t^* + \theta_{t^*}} \frac{dM(v)}{Y(v)} + o_p(n^{-\frac{1}{2}}) = \sum_{i=1}^n e_i(\theta_{t^*}) + o_p(n^{-\frac{1}{2}})
\end{aligned}$$

em que

$$e_i(x) = -\frac{1}{2} S(t^*) \int_{t^*}^{t^* + x} \frac{dM_i(v)}{Y(v)},$$

para $i = 1, \dots, n$.

Fleming e Harrington (1991) notaram a igualdade $E[dM_i(t)|\mathcal{F}_{s-}] = 0$, em que $\mathcal{F}_{s-} = \sigma\{N(u), I(X \leq u, \delta = 0) : 0 \leq u < s\}$ e Jeong (2014) usou esse fato para concluir que $E[e_i(\theta_{t^*})] = 0$ e $\text{Var}[e_i(\theta_{t^*})] = E[e_i^2(\theta_{t^*})]$. Portanto, o Teorema Central do Limite para martingais (Rebolledo, 1980) pode ser aplicado a $u(\theta_{t^*})$ e para um t^* fixado é válido o resultado:

$$\frac{u(\theta_{t^*})}{\sqrt{\sigma_{t^*}^2(\theta_{t^*})}} \xrightarrow{\mathcal{D}} N(0, 1),$$

em que $\sigma_{t^*}^2(\theta_{t^*}) = \sum_{i=1}^n e_i^2(\theta_{t^*})$.

A variância pode ser estimada, usando o estimador $\hat{e}_i(\theta_{t^*})$ para $e_i(\theta_{t^*})$, com

$$\begin{aligned}
\hat{e}_i(\theta_{t^*}) &= -\frac{1}{2} \hat{S}(t^*) \int_{t^*}^{t^* + \theta_{t^*}} \frac{d\hat{M}_i(v)}{Y(v)} \quad \text{e} \\
\hat{M}_i(t) &= N_i(t) - \int_0^t Y_i(v) d\hat{\Lambda}(v),
\end{aligned}$$

em que $\hat{\Lambda}(t)$ é o estimador de Nelson-Aalen da função de risco acumulada. Como $\hat{S}(t)$ e $\hat{\Lambda}(t)$ são estimadores consistentes para $S(t)$ e $\Lambda(t)$, respectivamente, $\hat{e}_i(\theta_{t^*})$ é um estimador consistente para $e_i(\theta_{t^*})$ (Jeong, 2014).

De modo a testar a hipótese $H_0 : \theta_{t^*} = m_0$ e construir um intervalo de confiança para θ_{t^*} , é definida a estatística $V(m_0)$ dada por

$$V(m_0) = \frac{\hat{u}^2(m_0)}{\hat{\sigma}_{t^*}^2(\hat{\theta}_{t^*})},$$

em que $\hat{u}(m_0) = \sum_{i=1}^n \hat{e}_i(m_0)$ e $\hat{\sigma}_{t^*}^2(\hat{\theta}_{t^*}) = \sum_{i=1}^n \hat{e}_i^2(\hat{\theta}_{t^*})$.

A estatística $\frac{\hat{u}(m_0)}{\hat{\sigma}_{t^*}(\hat{\theta}_{t^*})}$ tem o mesmo conceito da estatística de teste de Wald e segue a distribuição assintótica Normal padrão. Portanto, pelo teorema de Slutsky e sob H_0 , $V(m_0)$ tem uma distribuição assintótica qui-quadrado com 1 grau de liberdade.

Sendo assim, rejeita-se H_0 quando $V(m_0) > \chi_{1-\alpha,1}^2$ com um nível de significância α e um intervalo de confiança de $(1 - \alpha)$ para θ_{t^*} é dado por

$$\left\{ \theta_{t^*} : \frac{\hat{u}^2(m)}{\hat{\sigma}_{t^*}^2(\hat{\theta}_{t^*})} < \chi_{1-\alpha,1}^2 \right\}.$$

Para testar a igualdade entre dois grupos com relação à *vida mediana residual* em um tempo fixado t_0 , Jeong *et al.* (2007) propuseram um teste baseado na razão entre as medianas de vida residual.

Sejam $T_i^{(k)}$ o tempo de falha e $C_i^{(k)}$ o tempo de censura da observação i do grupo k , $i = 1, \dots, n$ e $k = 1, 2$. É observado o par $(X_i^{(k)}, \delta_i^{(k)})$, com $X_i^{(k)} = \min(T_i^{(k)}, C_i^{(k)})$ e $\delta_i^{(k)} = I(T_i^{(k)} \leq C_i^{(k)})$. Defina-se também $Y^{(k)}(t) = \sum_{i=1}^n Y_i^{(k)}(t) = \sum_{i=1}^n I(X_i^{(k)} \geq t)$, $N^{(k)}(t) = \sum_{i=1}^n N_i^{(k)}(t) = \sum_{i=1}^n I(X_i^{(k)} \leq t, \delta_i^{(k)} = 1)$, $\Lambda^{(k)}$ a função de risco acumulada do grupo k e $M_i^{(k)}(t) = N_i^{(k)}(t) - \int_0^t Y_i^{(k)}(v) d\Lambda^{(k)}(v)$.

Seja $\theta_{t_0,0}^{(k)}$ o verdadeiro valor da *vida mediana residual* no tempo t_0 para o grupo k . A hipótese de igualdade pode ser escrita como

$$H_0 : r_{t_0} = \frac{\theta_{t_0,0}^{(2)}}{\theta_{t_0,0}^{(1)}} = 1.$$

Para cada grupo k , escreve-se a função de estimação $u_k(\theta_{t_0}^{(k)})$ como

$$u_k(\theta_{t_0}^{(k)}) = \hat{S}_k(t_0 + \theta_{t_0}^{(k)}) - \frac{1}{2} \hat{S}_k(t_0),$$

em que \hat{S}_k é o estimador de Kaplan-Meier para S_k , a função de sobrevivência associada ao grupo k . Jeong *et al.* (2007) provaram o teorema que será apresentado a seguir.

Teorema 4 *Sob a hipótese nula $H_0 : r_{t_0} = r_{t_0,0}$, a estatística*

$$V_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0}) = \frac{u_1^2(\theta_{t_0}^{(1)})}{\hat{\sigma}^2(\hat{\theta}_{t_0}^{(1)})} + \frac{u_2^2(\theta_{t_0}^{(2)})}{\hat{\sigma}^2(\hat{\theta}_{t_0}^{(2)})} = \frac{u_1^2(\theta_{t_0}^{(1)})}{\hat{\sigma}^2(\hat{\theta}_{t_0}^{(1)})} + \frac{u_2^2(r_{t_0,0}\theta_{t_0}^{(1)})}{\hat{\sigma}^2(\hat{\theta}_{t_0}^{(2)})}.$$

A estatística $W_{t_0}(r_{t_0,0}) = \min_{\theta_{t_0}^{(1)}} V_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0})$ segue uma distribuição assintótica qui-quadrado com 1 grau de liberdade.

Para uma ideia da prova do Teorema 4, lembremos que assintoticamente $S_k(t_0 + \theta_{t_0}^{(k)}) - \frac{1}{2} S_k(t_0) = 0$ e pela definição da função de estimação $u_k(\hat{\theta}_{t_0}^{(k)}) = 0$. Então, para o grupo 1, em um valor $\theta_{t_0}^{(1)}$ em uma vizinhança em torno do verdadeiro valor $\theta_{t_0,0}^{(1)}$,

$$u_1(\theta_{t_0}^{(1)}) = \hat{S}_1(t_0 + \theta_{t_0}^{(1)}) - \frac{1}{2} \hat{S}_1(t_0) = \hat{S}_1(t_0 + \theta_{t_0}^{(1)}) - S_1(t_0 + \theta_{t_0}^{(1)}) - \frac{1}{2} (\hat{S}_1(t_0) - S_1(t_0)).$$

Como o estimador de Kaplan-Meier é uma função escada, a derivada com relação a $\theta_{t_0}^{(1)}$ da expressão acima é igual $-f_1(t_0 + \theta_{t_0}^{(1)})$, sendo f_k a função densidade associada ao grupo k . Usando a consistência do estimador $\hat{\theta}_{t_0}^{(k)}$ e a expansão de Taylor de primeira ordem de $u_1(\theta_{t_0}^{(1)})$ no ponto $\hat{\theta}_{t_0}^{(1)}$ resulta em

$$u_1(\theta_{t_0}^{(1)}) = -f_1(t_0 + \theta_{t_0,0}^{(1)}) (\hat{\theta}_{t_0}^{(1)} - \theta_{t_0,0}^{(1)}) + o_p(n^{-\frac{1}{2}}).$$

Sejam $r_{t_0} = \frac{\theta_{t_0,0}^{(2)}}{\theta_{t_0,0}^{(1)}}$ e $\hat{r}_{t_0} = \frac{\hat{\theta}_{t_0}^{(2)}}{\hat{\theta}_{t_0}^{(1)}}$. Para o grupo 2 temos

$$\begin{aligned} u_2(\theta_{t_0}^{(2)}) &= \hat{S}_2(t_0 + \theta_{t_0,0}^{(2)}) - \frac{1}{2}\hat{S}_2(t_0) = -f_2(t_0 + \theta_{t_0,0}^{(2)}) \left(\hat{\theta}_{t_0}^{(2)} - \theta_{t_0}^{(2)} \right) + o_p(n^{-\frac{1}{2}}) \\ &= -f_2(t_0 + \theta_{t_0,0}^{(2)}) \left(\hat{r}_{t_0} \hat{\theta}_{t_0}^{(1)} - \theta_{t_0}^{(2)} \right) + o_p(n^{-\frac{1}{2}}). \end{aligned}$$

Aplicando a expansão de Taylor de segunda ordem à função $\hat{r}_{t_0} \hat{\theta}_{t_0}^{(1)} - \theta_{t_0}^{(2)}$ no ponto $(r_{t_0,0}, \theta_{t_0,0}^{(1)})$, assintoticamente,

$$u_2(\theta_{t_0}^{(2)}) = -f_2(t_0 + \theta_{t_0,0}^{(2)}) \left[\theta_{t_0,0}^{(1)}(\hat{r}_{t_0} - r_{t_0,0}) + r_{t_0,0}(\theta_{t_0}^{(1)} - \theta_{t_0,0}^{(1)}) \right] + o_p(n^{-\frac{1}{2}}).$$

Podemos escrever $V_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0}) = K_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0}) + o_p(n^{-1})$, com

$$\begin{aligned} K_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0}) &= \frac{u_1^2(\theta_{t_0}^{(1)})}{\sigma_{t_0}^2(\theta_{t_0,0}^{(1)})} + \frac{u_2^2(\theta_{t_0}^{(2)})}{\sigma_{t_0}^2(\theta_{t_0,0}^{(2)})} = \frac{f_1^2(t_0 + \theta_{t_0,0}^{(1)}) \left(\hat{\theta}_{t_0}^{(1)} - \theta_{t_0}^{(1)} \right)^2}{\sigma_1^2(\theta_{t_0,0}^{(1)})} \\ &\quad + \frac{f_2^2(t_0 + \theta_{t_0,0}^{(2)}) \left[\theta_{t_0,0}^{(1)}(\hat{r}_{t_0} - r_{t_0,0}) + r_{t_0,0}(\theta_{t_0}^{(1)} - \theta_{t_0,0}^{(1)}) \right]^2}{\sigma_2^2(\theta_{t_0,0}^{(2)})}. \end{aligned}$$

Portanto, minimizar $V_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0})$ com relação a $\theta_{t_0}^{(1)}$ é equivalente a minimizar a forma quadrática $K_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0})$, o que ocorre assintoticamente quando $\theta_{t_0}^{(1)} = \hat{\theta}_{t_0}^{(1)}$. Quando $n \rightarrow \infty$,

$$\min_{\theta_{t_0}^{(1)}} V_{t_0}(\theta_{t_0}^{(1)}, r_{t_0,0}) - \frac{f_2^2(t_0 + \theta_{t_0,0}^{(2)}) \left(\theta_{t_0,0}^{(1)} \right)^2 (\hat{r}_{t_0} - r_{t_0,0})^2}{\sigma_2^2(\theta_{t_0,0}^{(2)})} \rightarrow 0,$$

que pelo método delta (Jeong *et al.*, 2007) é equivalente a $\frac{\hat{r}_{t_0} - r_{t_0}}{\sqrt{\text{Var}(\hat{r}_{t_0})}}$, que por sua vez tem distribuição assintótica qui-quadrado com 1 grau de liberdade.

Como resultado do Teorema 4, a hipótese $H_0 : r_{t_0} = r_{t_0,0}$ é rejeitada a um nível de significância α quando $W_{t_0}(r_{t_0,0}) > \chi_{1-\alpha,1}^2$ e um intervalo de confiança $100(1 - \alpha)\%$ para r_{t_0} é dado por

$$\left\{ r_{t_0} : \inf_{\theta_{t_0}^{(1)}} V_{t_0}(r_{t_0}, \theta_{t_0}^{(1)}) < \chi_{1-\alpha,1}^2 \right\}.$$

3.2 Abordagem via verossimilhança empírica

Owen (2001) define a função de verossimilhança empírica (EL) para dados censurados como

$$EL = \prod_{i=1}^n (p_i)^{\delta_i} \left(\sum_{x_j > x_i} p_j \right)^{1-\delta_i},$$

em que p_i é a probabilidade associada ao par (X_i, δ_i) . No contexto de verossimilhança, os tempos são vistos de forma discreta. A função de risco nesse caso se torna a probabilidade de o indivíduo falhar em um tempo t dado que ele sobreviveu até o instante de tempo anterior a t .

A função EL alcança seu valor máximo quando p_i for igual ao “salto” do estimador de Kaplan-Meier associado ao tempo de falha da observação i . Owen (2001) mostrou esse resultado escrevendo

a função de verossimilhança empírica com relação à função de risco na forma

$$EL = \prod_{j=1}^k \lambda_j^{d_j} (1 - \lambda_j)^{r_j - d_j}$$

com $t_1 < \dots < t_k$ os tempos de falha distintos, d_j sendo o número de observações que falharam no tempo t_j , r_j o número de indivíduos em risco em t_j e usando o estimador de máxima verossimilhança empírica para λ_j dado por $\hat{\lambda}_j = \frac{d_j}{r_j}$.

Seja $g(t)$ uma função tal que $\text{Var}[g(T)] < \infty$. Para o teste da razão de verossimilhanças, maximiza-se EL com a restrição

$$\sum_{i=1}^n g(X_i) p_i = \gamma \quad (3.4)$$

e também a condição

$$\int \frac{g^2(t)}{1 - G(t)} dF(t) < \infty, \quad (3.5)$$

em que G é a função de distribuição da censura. Para testar a hipótese $H_0 : E[g(T)] = \gamma$, [Bathke et al. \(2009\)](#) utilizaram o seguinte teorema.

Teorema 5 *Considere um conjunto de dados com censura à direita e sua função de verossimilhança como definida anteriormente. Suponha que $E[g(T)] = \gamma$ e a condição (3.5). Então,*

$$W(\gamma) = -2 \log \left[\frac{\max_0 EL}{EL(KM)} \right] \xrightarrow{\mathcal{D}} \chi_1^2, \quad (3.6)$$

em que o máximo do numerador é calculado sob todos os p_i que satisfaçam (3.4).

Seja $g_b(x) = I(x \leq t + b) - \frac{1}{2}I(x \leq t)$. Então

$$\begin{aligned} \int_0^\infty g(t) dF(t) &= \int_0^\infty \left[I(x \leq t + b) - \frac{1}{2}I(x \leq t) \right] dF(x) \\ &= \int_0^{t+b} dF(x) - \frac{1}{2} \int_0^t dF(x) = S(t + b) - \frac{1}{2}S(t). \end{aligned}$$

Sendo assim, testar a hipótese $H_0 : \theta(t) = b$ equivale a testar

$$H_0 : \int_0^\infty g_b(x) dF(x) = 0,$$

que é uma hipótese na forma $E[g(T)] = \gamma$. Portanto, o Teorema 5 pode ser usado para fazer inferências sobre $\theta(t)$ ([Zhou e Jeong, 2011](#)).

O método de máxima verossimilhança empírica pode ser usado para testar a razão entre a *vida mediana residual* de uma população com relação à outra, ou seja, $H_0 : \frac{\theta_1(t_1)}{\theta_2(t_2)} = c$. Para isso, definimos duas hipóteses:

$$H_{01} : \theta_1(t_1) = c\theta \quad e \quad H_{02} : \theta_2(t_2) = \theta.$$

São calculadas as estatísticas $W_1(c\theta, t_1)$ e $W_2(\theta, t_2)$, respectivamente, para as hipóteses acima. [Zhou e Jeong \(2011\)](#) mostraram que, sob a condição de $c = 1$, a estatística Q dada por

$$Q = \inf_{\theta} \{W_1(c\theta, t_1) + W_2(\theta, t_2)\}$$

tem distribuição qui-quadrado com 1 grau de liberdade.

Para provar esse resultado, seja $\eta = c\theta$. Define-se

$$\begin{aligned}\hat{\phi}_1(\eta) &= \hat{F}_{KM}(t + \eta) - \frac{1}{2}\hat{F}_{KM}(t) - \frac{1}{2} \quad \text{e} \\ \hat{\phi}_2(\theta) &= \hat{F}_{KM}(t + \theta) - \frac{1}{2}\hat{F}_{KM}(t) - \frac{1}{2},\end{aligned}$$

com $\hat{F}_{KM}(t) = 1 - \hat{S}(t)$, $\hat{S}(t)$ é o estimador de Kaplan-Meier.

Zhou e Jeong (2011) usam o resultado que $W_1(\eta, x) + W_2(\theta, x)$ pode ser expresso como

$$\frac{n \left[\hat{\phi}_1(\eta) - \phi_1(\eta) \right]}{\xi_1^2} + \frac{n \left[\hat{\phi}_2(\theta) - \phi_2(\theta) \right]}{\xi_2^2}, \quad (3.7)$$

em que ξ_1^2 e ξ_2^2 são as variâncias de $\hat{\phi}_1(\eta)$ e $\hat{\phi}_2(\theta)$, respectivamente. Sob H_{01} e H_{02} , $\phi_1(\eta) = \phi_2(\theta) = 0$ e os parâmetros η e θ podem ser estimados com $\hat{\phi}_1(\eta) = 0$ e $\hat{\phi}_2(\theta) = 0$.

Como o estimador de Kaplan-Meier é uniformemente consistente e assintoticamente normal, usando o método delta em (3.7) chega-se à expressão

$$U(\theta) \approx \frac{n(\hat{\eta} - \eta)^2}{\sigma_1^2} + \frac{n(\hat{\theta} - \theta)^2}{\sigma_2^2},$$

em que σ_1^2 e σ_2^2 são as variâncias de $\hat{\eta}$ e $\hat{\theta}$, respectivamente.

$U(\theta)$ é minimizada em θ_0 dado por

$$\theta_0 = \frac{\frac{c}{\sigma_1^2}\hat{\eta} + \frac{\hat{\theta}}{\sigma_2^2}}{\frac{c^2}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.$$

Avaliando a função $U(\theta)$ no ponto θ_0 e depois de uma extensa manipulação algébrica, resulta em

$$U(\theta_0) = \frac{(\hat{\eta} - c\hat{\theta})^2}{\sigma_1^2 + c^2\sigma_2^2},$$

que, sob H_0 , segue uma distribuição assintótica qui-quadrado com 1 grau de liberdade, pois $\text{Var}(\hat{\eta} - c\hat{\theta}) = \sigma_1^2 + c^2\sigma_2^2$ para duas amostras independentes (Zhou e Jeong, 2011).

A abordagem via verossimilhança empírica é flexível e pode ser usada também para testar a *vida média residual*. Para isso, o teste $H_0 : m(t) = \mu$ é equivalente a testar

$$H_0 : \frac{\int_t^\infty s dF(s)}{1 - F(t)} = t + \mu,$$

que por sua vez equivale a

$$H_0 : \int_t^\infty s dF(s) = (t + \mu)S(t).$$

Como $S(t) = \int_t^\infty dF(s)$, a hipótese nula é reescrita na forma

$$H_0 : \int_t^{\infty} (s - (t + \mu)) dF(s) = 0.$$

Com $g(s) = [s - (t + \mu)]I(s > t)$, o Teorema 5 pode ser usado para fazer inferências sobre a *vida média residual* de uma população e testar a igualdade entre as médias de vida residual de duas populações independentes.

3.3 Modelo de regressão para o tempo mediano residual

Jung *et al.* (2009) consideraram um modelo de regressão para a *vida quantílica residual* com uma transformação logarítmica no tempo de vida. Para o caso particular da *vida mediana residual* do tempo t_0 , o modelo resulta em

$$\text{mediana} \{ \log (T_i - t_0) | T_i \geq t_0, \mathbf{z}_i \} = \mathbf{z}'_i \boldsymbol{\beta}_{t_0}, \quad (3.8)$$

em que $\boldsymbol{\beta}'_{t_0} = (\boldsymbol{\beta}_{t_0}^0, \boldsymbol{\beta}_{t_0}^1, \dots, \boldsymbol{\beta}_{t_0}^p)$ é o vetor dos coeficientes de regressão e $\mathbf{z}'_i = (1, z_{1i}, \dots, z_{pi})$ é o vetor de covariáveis associado ao indivíduo i .

Para o caso sem censura, um modo de estimar $\boldsymbol{\beta}_{t_0}$ é minimizando a expressão $A(\boldsymbol{\beta}_{t_0})$ (Jung *et al.*, 2009) dada por

$$\begin{aligned} A(\boldsymbol{\beta}_{t_0}) &= \sum_{i=1}^n |\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}| = \sum_{i=1}^n [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) \\ &\quad - [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} < 0) \\ &= \sum_{i=1}^n [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) \\ &\quad - [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] [1 - I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} < 0)] \\ &= \sum_{i=1}^n 2 [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) - [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] \\ &= 2 \sum_{i=1}^n [\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}] \left[I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) - \frac{1}{2} \right]. \end{aligned}$$

Seja $\mathbf{U}(\boldsymbol{\beta}_{t_0})$ a primeira derivada de $A(\boldsymbol{\beta}_{t_0})$ com relação a $\boldsymbol{\beta}_{t_0}$. Jung *et al.* (2009) chamaram de *função de estimação* $\mathbf{U}(\boldsymbol{\beta}_{t_0})$ dada por

$$\mathbf{U}(\boldsymbol{\beta}_{t_0}) = \frac{\partial A(\boldsymbol{\beta}_{t_0})}{\partial \boldsymbol{\beta}_{t_0}} = -2 \sum_{i=1}^n \mathbf{z}_i \left[I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) - \frac{1}{2} \right],$$

e o estimador de $\boldsymbol{\beta}_{t_0}$ será a solução de $\mathbf{U}(\boldsymbol{\beta}_{t_0}) = \mathbf{0}$. Note que

$$E [I(\log (T_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0)] = \frac{1}{2}.$$

Se for considerado o caso com censura, assumindo independência entre T_i e C_i dado \mathbf{z}_i , temos

$$\begin{aligned}
E [I(\log(X_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) | \mathbf{z}_i] &= P(X_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}} | \mathbf{z}_i) \\
&= P(T_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}} | \mathbf{z}_i) P(C_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}} | \mathbf{z}_i) \\
&= \frac{P(T_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}} | \mathbf{z}_i)}{P(T_i > t_0 | \mathbf{z}_i)} P(C_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}} | \mathbf{z}_i) P(T_i > t_0 | \mathbf{z}_i) \\
&= \frac{1}{2} \frac{P(C_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}} | \mathbf{z}_i)}{P(C_i > t_0 | \mathbf{z}_i)} P(C_i > t_0 | \mathbf{z}_i) P(T_i > t_0 | \mathbf{z}_i) \\
&= \frac{1}{2} \frac{G(t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}})}{G(t_0)} P(X_i > t_0 | \mathbf{z}_i),
\end{aligned}$$

que resulta em

$$\begin{aligned}
E [I(\log(X_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0} \geq 0) | \mathbf{z}_i] &= E \left[\frac{1}{2} \frac{G(t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}})}{G(t_0)} I(X_i > t_0) \middle| \mathbf{z}_i \right] \quad \text{ou} \\
E [I(X_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}}) | \mathbf{z}_i] &= E \left[\frac{1}{2} \frac{G(t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}})}{G(t_0)} I(X_i > t_0) \middle| \mathbf{z}_i \right].
\end{aligned}$$

Jung et al. (2009) escrevem a equação de estimação $S(\boldsymbol{\beta}_{t_0})$ como

$$S(\boldsymbol{\beta}_{t_0}) = \sum_{i=1}^n \mathbf{z}_i \left[\frac{I(X_i > t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}})}{\hat{G}(t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0}})} - \frac{1}{2} \frac{I(X_i > t_0)}{\hat{G}(t_0)} \right] \approx \mathbf{0} \quad (3.9)$$

e definem o estimador $\hat{\boldsymbol{\beta}}_{t_0}$ como o valor de $\boldsymbol{\beta}_{t_0}$ que minimiza a função $\|S(\boldsymbol{\beta}_{t_0})\|$, em que $\|\cdot\|$ é a norma euclidiana. No Apêndice A de *Jung et al.* (2009) é mostrado que $\hat{\boldsymbol{\beta}}_{t_0}$ é consistente para $\boldsymbol{\beta}_{t_0}^*$, em que $\boldsymbol{\beta}_{t_0}^*$ é o verdadeiro valor de $\boldsymbol{\beta}_{t_0}$.

Suponha que temos o interesse em testar $H_0 : \boldsymbol{\beta}_{t_0} = \boldsymbol{\beta}_{t_0,0}$. *Jung et al.* (2009) mostraram que, sob H_0 ,

$$n^{-\frac{1}{2}} \Gamma_{t_0}^{-1} S(\boldsymbol{\beta}_{t_0,0}) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_{p+1}),$$

em que $\Gamma_{t_0} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \tau_{i,t_0} \tau'_{i,t_0}$, com

$$\begin{aligned}
\tau'_{i,t_0} &= \mathbf{z}'_i \left[\frac{I(X_i \geq t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0,0}})}{G(t_0 + e^{\mathbf{z}'_i \boldsymbol{\beta}_{t_0,0}})} - \frac{1}{2} \frac{I(X_i \geq t_0)}{G(t_0)} \right] \\
&+ \int_{-\infty}^{\infty} G^{-1}(s) \int_{-\infty}^s h^{-1}(v) \{dI(X_i \leq v, \delta_i = 0) - I(X_i \geq v) d\Lambda_G(v)\} dq_1(s) \\
&- q_2(t_0) \int_{-\infty}^{t_0} h^{-1}(s) [dI(X_i \leq s, \delta_i = 0) - I(X_i \geq s) d\Lambda_G(s)],
\end{aligned}$$

$\Lambda_G(s)$ é a função de risco acumulada associada à distribuição de censura,

$$h(s) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n I(X_i \geq s), \quad q_1(s) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n z'_i I\left(t_0 + e^{z'_i \beta_{t_0,0}} \geq \min(s, X_i)\right) \quad e$$

$$q_2(t_0) = \lim_{n \rightarrow \infty} \frac{1}{G(t_0)2n} \sum_{i=1}^n z'_i I(X_i \geq t_0).$$

Jung *et al.* (2009) mostraram que $\hat{\Gamma}_{t_0} = \frac{1}{n} \sum_{i=1}^n \hat{\tau}_{t_0} \hat{\tau}'_{t_0}$, em que

$$\begin{aligned} \hat{\tau}'_{t_0} = & z'_i \left[\frac{I\left(X_i \geq t_0 + e^{z'_i \hat{\beta}_{t_0}}\right)}{\hat{G}\left(t_0 + e^{z'_i \hat{\beta}_{t_0}}\right)} - \frac{1}{2} \frac{I(X_i \geq t_0)}{\hat{G}(t_0)} \right] + \sum_{i=1}^n z'_i \left[\frac{I\left(X_i \geq t_0 + e^{z'_i \hat{\beta}_{t_0}}\right)}{\hat{G}\left(t_0 + e^{z'_i \hat{\beta}_{t_0}}\right)} \right] \\ & \times \left[\frac{(1 - \delta_i) I\left(X_i \leq t_0 + e^{z'_i \hat{\beta}_{t_0}}\right)}{\sum_{m=1}^n I(X_m \geq X_i)} - \sum_{j=1}^n \frac{(1 - \delta_j) I\left(X_j \leq \min\left(t_0 + e^{z'_i \hat{\beta}_{t_0}}, X_i\right)\right)}{\left(\sum_{m=1}^n I(X_m \geq X_j)\right)^2} \right] \\ & - \sum_{i=1}^n z'_i \frac{I(X_i \geq t_0)}{2n \hat{G}(t_0)} \left[\frac{(1 - \delta_i) I(X_i \leq t_0)}{\sum_{m=1}^n I(X_m \geq X_i)} - \sum_{j=1}^n \frac{(1 - \delta_j) I\left(X_j \leq \min(t_0, X_i)\right)}{\left(\sum_{m=1}^n I(X_m \geq X_j)\right)^2} \right] \end{aligned}$$

é um estimador consistente para Γ_{t_0} . Pelo teorema de Slutsky, segue que sob H_0 ,

$$n^{-1} S'(\beta_{t_0,0}) \hat{\Gamma}_{t_0}^{-1} S(\beta_{t_0,0}) \xrightarrow{\mathcal{D}} \chi_{p+1}^2.$$

Agora considere $\beta'_{t_0} = (\beta_{t_0}^{(1)'}, \beta_{t_0}^{(2)'})$, com $\beta_{t_0}^{(1)}$ sendo um vetor $r \times 1$. Sejam $\hat{\beta}_{t_0} = (\hat{\beta}_{t_0}^{(1)}, \hat{\beta}_{t_0}^{(2)})$ e a hipótese $H_0 : \beta_{t_0}^{(1)} = \beta_{t_0,0}^{(1)}$. Seja $\beta_{t_0}^* = (\beta_{t_0,0}^{(1)}, \beta_{t_0}^{(2)})$. Jung *et al.* (2009) mostraram que, sob H_0 ,

$$V\left(\beta_{t_0,0}^{(1)}\right) = \min_{\beta_{t_0}^{(2)}} \left\{ n^{-1} S'(\beta_{t_0}^*) \hat{\Gamma}_{t_0} S(\beta_{t_0}^*) \right\} \xrightarrow{\mathcal{D}} \chi_r^2.$$

Kim *et al.* (2012) formularam um modelo semelhante ao (3.8) na forma

$$\log(T_i - t_0 | T_i > t_0) = z'_i \hat{\beta}_{t_0} + \epsilon_i,$$

em que ϵ_i são variáveis aleatórias independentes e com medianas iguais a zero.

Seja a função $\rho(u) = u \left[\frac{1}{2} - I(u \leq 0) \right]$. Kim *et al.* (2012) usaram os resultados de regressão quantílica para definir um estimador para β_{t_0} como o minimizador da expressão

$$\min_{\beta_{t_0}} \sum_{i=1}^n w_i I(X_i > t_0) \rho\left(\log(X_i - t_0) - z'_i \beta_{t_0}\right),$$

ou a solução da equação

$$\sum_{i=1}^n w_i I(X_i > t_0) z_i \psi\left(\log(X_i - t_0) - z'_i \hat{\beta}_{t_0}\right) = \mathbf{0},$$

em que $\psi(u) = \frac{1}{2} - I(u \leq 0)$. Se for considerado o método *IPCW* na estimação,

$$w_i = \frac{\delta_i}{1 - \hat{G}(X_i)}.$$

Kim *et al.* (2012) usaram a abordagem da verossimilhança empírica para fazer inferências sobre β_{t_0} . Sejam p_i e EL como definidos na Seção 3.2. Kim *et al.* (2012) consideraram $EL(p|\beta_{t_0})$ como o

valor máximo da função de verossimilhança empírica com relação a $\mathbf{p} = (p_1, \dots, p_n)$ sob a restrição

$$\sum_{i=1}^n p_i I(X_i > t_0) \mathbf{z}_i \psi(\log(X_i - t_0) - \mathbf{z}'_i \boldsymbol{\beta}_{t_0}) = \mathbf{0}.$$

Sendo assim, seja $R(\boldsymbol{\beta}_{t_0})$ definido por

$$R(\boldsymbol{\beta}_{t_0}) = \frac{EL(p|\boldsymbol{\beta}_{t_0})}{EL(KM)}.$$

Podemos aplicar a versão mais geral do Teorema 5, que considera múltiplas restrições. Seja $\boldsymbol{\beta}_{t_0,0}$ o verdadeiro valor de $\boldsymbol{\beta}_{t_0}$. Temos, sob H_0 ,

$$-2 \log\{R(\boldsymbol{\beta}_{t_0,0})\} \xrightarrow{\mathcal{D}} \chi_{p+1}^2.$$

Com o resultado acima é possível construir uma região de confiança de $100(1 - \alpha)\%$ para $\boldsymbol{\beta}_{t_0}$ como

$$\{\boldsymbol{\beta}_{t_0} : -2 \log R(p|\boldsymbol{\beta}_{t_0}) \leq \chi_{p+1,1-\alpha}^2\}.$$

Em muitos casos é de interesse testar hipóteses para uma parte de $\boldsymbol{\beta}_{t_0}$. Para isso, Kim *et al.* (2012) provaram o seguinte teorema.

Teorema 6 *Sob certas condições de regularidade (Kim et al., 2012), se $\boldsymbol{\beta}'_{t_0} = (\boldsymbol{\beta}'_{t_0,1}, \boldsymbol{\beta}'_{t_0,2})$, com $\boldsymbol{\beta}_{t_0,1} \in \mathbf{R}^{q_1}$, sob $H_0 : \boldsymbol{\beta}_{t_0,1} = \boldsymbol{\beta}_{t_0,1,0}$, seja $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{t_0,1,0}, \boldsymbol{\beta}_{t_0,2})$. Então,*

$$-2 \log \left(\sup_{\boldsymbol{\beta}_{t_0,2}} R(\boldsymbol{\beta}^*) \right) \xrightarrow{\mathcal{D}} \chi_{q_1}^2.$$

Agora serão apresentados dois modelos mais gerais para a *vida mediana residual*, porém eles não serão usados nas aplicações às bases simuladas e ao banco de dados do InCor.

Ma e Yin (2010) propuseram um modelo dado por

$$\text{mediana}(T_i - t | T_i > t, \mathbf{z}_i) = g(\eta(t), \mathbf{z}_i, \boldsymbol{\beta}),$$

em que $\eta(t)$ é a função de *vida mediana residual* basal. Ma e Yin (2010) notaram que para $t = 0$,

$$T_i = g(\eta(0), \mathbf{z}_i, \boldsymbol{\beta}) + \epsilon_i,$$

com $\text{mediana}(\epsilon_i | \mathbf{z}_i) = 0$.

Utilizando a técnica da probabilidade inversa ponderada (IPCW), supondo que os tempos de sobrevivência são contínuos e limitados a um valor L e os tempos de censura são contínuos com $P(C_i \geq L) > 0$, para $i = 1, \dots, n$, Ma e Yin (2010) calcularam os estimadores $\{\hat{\eta}(0), \hat{\boldsymbol{\beta}}\}$ minimizando a expressão

$$\sum_{i=1}^n \frac{\delta_i}{\hat{G}(X_i)} \rho(X_i - g\{\eta(0), \mathbf{z}_i, \boldsymbol{\beta}\}). \quad (3.10)$$

Pela definição do modelo, $\boldsymbol{\beta}$ não depende de t , Ma e Yin (2010) formularam, então, um método de se estimar $\hat{\eta}(t)$ para qualquer t . Para isso, fixa-se o parâmetro $\boldsymbol{\beta}$ no valor estimado $\hat{\boldsymbol{\beta}}$ e para um valor de t^* , estima-se $\eta(t^*)$ minimizando a expressão

$$\sum_{i=1}^n \frac{\delta_i}{\hat{G}(X_i)} I(X_i > t^*) \rho(X_i - t^* - g\{\eta(0), \mathbf{z}_i, \hat{\boldsymbol{\beta}}\}), \quad (3.11)$$

com relação a $\eta(t^*)$.

De acordo com [Ma e Yin \(2010\)](#), os estimadores $\hat{\eta}(0)$ e $\hat{\beta}$ oriundos de (3.10) têm distribuição assintótica normal, assim como $\hat{\eta}(t^*)$ resultante de (3.11). No entanto, quanto maior o valor de t^* , menos observações contribuem para a estimação de $\hat{\eta}(t^*)$; portanto, a performance do estimador piora com o aumento de t^* .

[Ma e Wei \(2012\)](#) formularam uma estimação para coeficientes que variam ao longo do tempo, ou seja,

$$\text{mediana}(T_i - t | \mathbf{z}_i, T_i \geq t) = g\{\mathbf{z}_i, \boldsymbol{\beta}(t)\}.$$

A estimação usa a técnica *spline* e para isso define-se $\mathbf{b}(t) = (\pi_1(t), \dots, \pi_k(t))'$, $\pi_i(t)$ como as funções base B-*spline*. Também são escolhidos $t_1 = 0$ e $t_{j+1} = X_j$ para $j = 1, \dots, J - 1$, em que J é um valor entre k e um terço dos valores distintos de X_j .

Para cada t_j , obtém-se $\tilde{\boldsymbol{\beta}}(t_j)$ resolvendo a equação

$$\sum_{i=1}^n \frac{\partial g\{\mathbf{z}_i, \boldsymbol{\beta}(t_j)\}}{\partial \boldsymbol{\beta}(t_j)} \left(\frac{I(X_i \geq t_j + g\{\mathbf{z}_i, \boldsymbol{\beta}(t_j)\})}{\hat{G}(t_j + g\{\mathbf{z}_i, \boldsymbol{\beta}(t_j)\})} - \frac{1}{2} \frac{I(X_i \geq t_j)}{\hat{G}(t_j)} \right) = 0.$$

Define-se o estimador de $\boldsymbol{\beta}(t)$ como $\hat{\boldsymbol{\beta}}(t) = \hat{\alpha} \mathbf{b}(t)$, com $\hat{\alpha}$ sendo obtido minimizando a expressão

$$\sum_{j=1}^J \{\alpha \mathbf{b}(t_j) - \tilde{\boldsymbol{\beta}}(t_j)\}' \{\alpha \mathbf{b}(t_j) - \tilde{\boldsymbol{\beta}}(t_j)\}.$$

O valor de α que minimiza a expressão acima pode ser encontrado através da fórmula

$$\hat{\alpha} = \left\{ \sum_{j=1}^J \tilde{\boldsymbol{\beta}}(t_j) \mathbf{b}'(t_j) \right\} \left\{ \sum_{j=1}^J \mathbf{b}(t_j) \mathbf{b}'(t_j) \right\}^{-1}.$$

[Ma e Wei \(2012\)](#) mostraram que $\hat{\boldsymbol{\beta}}(t)$ tem distribuição assintótica normal com média $\boldsymbol{\beta}(t)$ e matriz de variâncias-covariâncias que será omitida; para mais detalhes, vide [Ma e Wei \(2012\)](#).

Capítulo 4

Estudo de Simulação

Na primeira seção deste capítulo serão apresentados os estudos de simulação feitos para a estimação da *vida média residual* e da *vida mediana residual* utilizando a distribuição Weibull para a criação das réplicas. Na seção seguinte é apresentado um estudo de simulação com diferentes proporções de censura e que usa a distribuição Exponencial para a geração das amostras. Na terceira seção são apresentados os resultados das simulações dos testes em duas populações para a *vida média residual* e para a *vida mediana residual*.

4.1 Estimação da vida média e mediana residual em uma população

Utilizando a função *survreg* da linguagem R foi ajustado um modelo Weibull para o conjunto de dados do InCor, com a finalidade de se estimar a distribuição do tempo de vida. As estimativas dos parâmetros resultantes do ajuste do modelo foram $a = 0,89$ e $b = 139,62$ para uma função densidade escrita na forma

$$f(t) = \frac{a}{b} \left(\frac{t}{b} \right)^{a-1} e^{-\left(\frac{t}{b}\right)^a}, \quad t \geq 0.$$

Para a distribuição da censura, foi utilizado o seguinte procedimento: as observações censuradas foram consideradas “falhas” e as observações que falharam foram consideradas como “censuras”, ou em termos matemáticos, $\delta_i^* = 1 - \delta_i$. Utilizando o par (X_i, δ_i^*) foi ajustado um modelo Weibull e as estimativas dos parâmetros foram $a = 1,61$ e $b = 99,70$.

Para cada réplica foi gerada uma amostra da distribuição Weibull com parâmetros $a = 0,89$ e $b = 139,62$ representando os tempos de sobrevivência (T_i) e outra com parâmetros $a = 1,61$ e $b = 99,70$ representando as censuras (C_i). Os dados simulados são formados pelo mínimo entre T_i e C_i e o indicador de falha $I(T_i \leq C_i)$.

Os valores da *vida média residual* $m(t_0)$ e a *vida mediana residual* $\theta(t_0)$ são diferentes para cada t_0 fixado. As simulações foram feitas considerando os valores de $t_0 = 24, 42$ e 60 meses que correspondem a dois, três e meio e cinco anos. Foram geradas amostras com tamanho $n = 200, 500$ e 1000 . Os intervalos de confiança de 95% foram calculados utilizando o método de reamostragem de *bootstrap*.

Os resultados da simulação encontram-se nas Tabelas 4.1 e 4.2. Para cada cenário foram geradas 5000 réplicas; calculou-se a proporção de intervalos de confiança que continham o verdadeiro valor da vida média (ou mediana) residual e também a amplitude média do intervalo. Outras medidas apresentadas na tabela são a proporção média de censuras das réplicas (p^c) e a proporção média de censuras entre as observações com tempo maior que t_0 ($p_{t_0}^c$)

Como esperado, tanto para a *vida média residual* quanto para a *vida mediana residual*, a proporção de intervalos que contém o verdadeiro valor aumenta conforme o tamanho da amostra fica maior. Quanto maior o tempo em que a vida residual é calculada, menor é a proporção de intervalos

Proporção de intervalos contendo o verdadeiro valor						
t_0	Verdadeiro Valor	$n = 200$	$n = 500$	$n = 1000$	p^c	$p_{t_0}^c$
24	155,40	0,993(169,27)	1,000(170,01)	1,000(169,94)	55,22%	63,75%
42	158,55	0,941(160,47)	0,998(161,24)	1,000(161,30)	55,24%	67,50%
60	161,05	0,807(152,25)	0,957(153,38)	0,995(161,05)	55,21%	70,48%

Tabela 4.1: Resultados da simulação da vida média residual:
Proporção de cobertura (amplitude média dos intervalos)

Proporção de intervalos contendo o verdadeiro valor						
t_0	Verdadeiro Valor	$n = 200$	$n = 500$	$n = 1000$	p^c	$p_{t_0}^c$
24	103,29	0,987(165,90)	0,999(134,10)	0,999(89,49)	55,22%	63,75%
42	102,09	0,956(164,62)	0,995(154,19)	1,000(125,29)	55,24%	67,50%
60	99,89	0,912(157,44)	0,964(158,67)	0,995(146,92)	55,21%	70,48%

Tabela 4.2: Resultados da simulação da vida mediana residual:
Proporção de cobertura (amplitude média dos intervalos)

contendo o verdadeiro valor. Isso também é um resultado esperado, pois o aumento de t_0 faz com que menos informação da amostra seja usada para a inferência. A amplitude média do intervalo de confiança diminuiu com o aumento da amostra para a *vida mediana residual*, o que não ocorreu para a *vida média residual*.

Para a *vida mediana residual*, o verdadeiro valor diminui com o aumento de t_0 , porém isso não ocorreu para a vida média. A distribuição Weibull estimada para o conjunto de dados possui uma cauda pesada, o que influenciou o cálculo da *vida média residual* teórica.

4.2 Avaliação da censura na estimação da vida média e da vida mediana residual

Com a finalidade de estudar o efeito de diferentes proporções de censura nas amostras para a estimação da *vida média residual* e *vida mediana residual*, utilizamos a distribuição exponencial para a geração dos tempos de vida simulados e também para as censuras.

Se T segue uma distribuição exponencial de parâmetro λ , C segue uma distribuição exponencial de parâmetro μ e T é independente de C , então a probabilidade de T ser maior que C é calculada por

$$\begin{aligned} P(T > C) &= \int_0^\infty \int_c^\infty \lambda e^{-\lambda t} \mu e^{-\mu c} dt dc = \int_0^\infty \mu e^{-\mu c} (e^{-\lambda c}) dc \\ &= \frac{\mu}{\mu + \lambda} \int_0^\infty (\mu + \lambda) e^{-(\mu + \lambda)c} dc = \frac{\mu}{\mu + \lambda}. \end{aligned}$$

Utilizando esse resultado, é possível variar o valor de μ na geração das amostras simuladas, de modo que a proporção de censura seja a esperada. Neste estudo, o valor de λ foi fixado em 0,006, o que resulta em uma *vida média residual* no valor de 115,52 e *vida mediana residual* no valor de 166,66 para qualquer valor de t_0 . As proporções de censuras consideradas foram 30%, 50% e 70% para os tamanhos de amostra 50, 100, 200 e 300.

Na Tabela 4.3 são apresentados os resultados da estimação da *vida média residual*. Como esperado, quanto maior o tamanho da amostra, maior a proporção de cobertura do intervalo de confiança. Entretanto, para uma proporção de censura igual a 70% a proporção de cobertura não chegou a 0,500 mesmo com uma amostra de tamanho 300.

O aumento no valor de t_0 faz com que a amplitude do intervalo de confiança seja maior, ou seja, a estimativa é menos precisa. No caso da distribuição exponencial, em que tanto a *vida média residual* quanto a *vida mediana residual* são constantes em t_0 , é mais interessante fazer a estimação para $t_0 = 0$ para que ela seja mais precisa.

Proporção de intervalos contendo o verdadeiro valor				
t_0	n	30%	50%	70%
24	50	0,873(107,31)	0,719(107,25)	0,377(90,63)
	100	0,895(81,14)	0,794(87,86)	0,394(79,20)
	200	0,928(59,71)	0,815(67,54)	0,429(68,16)
	300	0,933(49,98)	0,835(60,33)	0,441(63,84)
42	50	0,863(115,13)	0,718(116,17)	0,339(92,82)
	100	0,899(86,27)	0,784(93,86)	0,381(84,50)
	200	0,927(64,24)	0,824(74,48)	0,421(75,48)
	300	0,936(53,61)	0,822(64,36)	0,430(68,72)
60	50	0,852(119,46)	0,715(121,05)	0,262(90,60)
	100	0,889(93,87)	0,761(101,69)	0,372(90,65)
	200	0,890(68,69)	0,770(80,54)	0,400(79,87)
	300	0,926(58,06)	0,832(71,00)	0,431(74,85)

Tabela 4.3: Resultados da simulação da vida média residual: Proporção de cobertura (amplitude média dos intervalos)

A Tabela 4.4 contém os resultados da estimação da *vida mediana residual*. Neste caso, com o tamanho de amostra 300 a proporção de cobertura ficou próxima dos 95%, mesmo com uma proporção de censura de 70%.

Para proporções de censura iguais a 30% e 50%, com o aumento do tamanho da amostra houve a diminuição da amplitude média dos intervalos de confiança e também quanto maior o valor de t_0 , maior a amplitude do intervalo. Para a proporção de censura de 70%, esse comportamento das amplitudes dos intervalos não ocorreu.

Proporção de intervalos contendo o verdadeiro valor				
t_0	n	30%	50%	70%
24	50	0,939(120,90)	0,937(169,28)	0,925(137,52)
	100	0,955(80,90)	0,939(106,41)	0,932(147,00)
	200	0,942(56,80)	0,947(67,66)	0,934(129,23)
	300	0,958(45,63)	0,945(55,66)	0,961(113,12)
42	50	0,933(130,51)	0,933(190,95)	0,858(125,20)
	100	0,949(88,18)	0,943(126,67)	0,934(141,87)
	200	0,953(62,09)	0,957(80,33)	0,940(146,26)
	300	0,945(49,66)	0,948(61,91)	0,938(130,73)
60	50	0,942(149,16)	0,944(196,81)	0,726(116,81)
	100	0,940(95,03)	0,943(144,60)	0,909(135,37)
	200	0,938(65,76)	0,937(89,04)	0,934(144,67)
	300	0,942(53,87)	0,945(70,99)	0,944(144,14)

Tabela 4.4: Resultados da simulação da vida mediana residual: Proporção de cobertura (amplitude média dos intervalos)

4.3 Teste da vida média e mediana em duas populações

O estudo de simulação para o teste entre duas populações usou o conjunto de dados do InCor como referência de distribuição, tal qual feito na Seção 4.1. Uma das variáveis que mais diferencia o tempo de sobrevivência é a taxa de creatinina, categorizada em baixa (taxa menor que 2,6 mg/dL) e alta (taxa maior ou igual a 2,6 mg/dL). A Figura 4.1 mostra as curvas de sobrevivência estimadas para esses dois grupos. Nota-se que a taxa de creatinina baixa está associada a maior sobrevivência dos pacientes. Entretanto, nem sempre uma diferença na curva de sobrevivência implica que as funções de *vida média residual* e *vida mediana residual* também sejam diferentes.

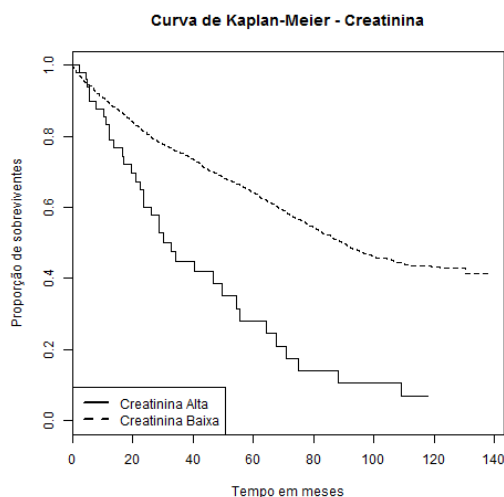


Figura 4.1: Estimador de Kaplan-Meier segundo categorias de taxa de creatinina

Na Figura 4.2 vemos que a *vida média residual* é menor para quem tem creatinina alta do que para as pessoas com creatinina baixa, para todos os valores de t_0 . A distância entre as duas curvas diminui conforme t_0 aumenta.

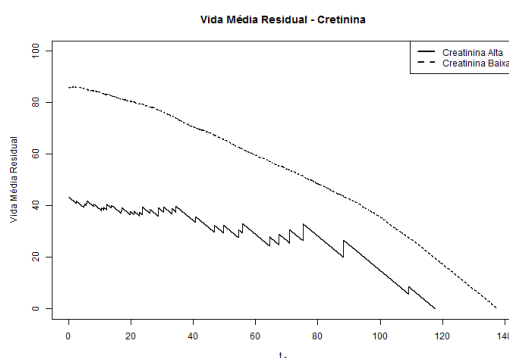


Figura 4.2: Vida média residual segundo categorias de taxa de creatinina

A Figura 4.3 mostra a *vida mediana residual* estimada para cada grupo da taxa de creatinina. O estimador da *vida mediana residual* tem o formato aproximado de uma reta em sua parte final por conta da natureza do estimador, que é decrescente e linear em t_0 quando o estimador de Kaplan-Meier se mantém com o mesmo valor. Assim como ocorreu para a *vida média residual*, a *vida mediana residual* é maior para as pessoas com taxa de creatinina baixa.

Seguindo o mesmo procedimento realizado na Seção 4.1, foi ajustado um modelo Weibull para os pacientes com creatinina alta e outro para aqueles com creatinina baixa. Os parâmetros das

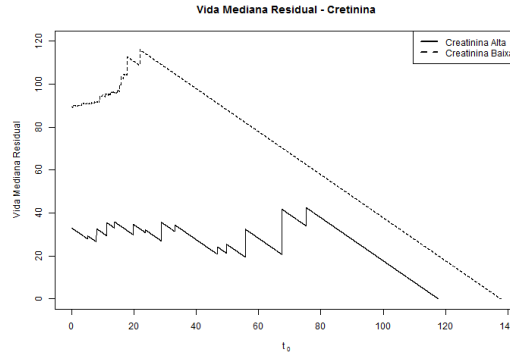


Figura 4.3: Vida mediana residual segundo categorias de taxa de creatinina

correspondentes distribuições Weibull da censura também foram estimados.

Para o teste de diferenças entre a *vida média residual* e *vida mediana residual* das populações simuladas foi considerada a razão de verossimilhanças empírica descrita na Seção 3.2. As diferenças foram testadas em três tempos diferentes ($t_0 = 24, 42$ e 60 meses) e para cada grupo de taxa de creatinina foram consideradas amostras de mesmo tamanho nos valores $n = 50, 100$ e 200 . Para cada cenário foram geradas 1000 réplicas. O teste realizado considerou um nível de significância igual a 5%.

Seja $r_{t_0}^{med}$ a razão entre a *vida média residual* de uma população sobre a outra, e $r_{t_0}^{md}$ a razão entre a *vida mediana residual* de uma população sobre a outra. Como o tempo de vida simulado segue uma distribuição Weibull, $r_{t_0}^{med}$ e $r_{t_0}^{md}$ variam com relação a t_0 . Nas Tabelas 4.5, 4.6 e 4.7 são mostrados os valores de t_0 , n , r_{t_0} e a porcentagem de vezes em que a hipótese nula do teste foi rejeitada refletindo o poder do teste.

Porcentagem de rejeição				
t_0	$r_{t_0}^{med}$	$n = 50$	$n = 100$	$n = 200$
24	4,08	97,5	99,9	100,0
42	4,41	91,6	98,7	100,0
60	4,68	86,0	96,5	99,6

Tabela 4.5: Resultados do teste de razão de verossimilhanças empíricas para a vida média residual

Podemos ver na Tabela 4.5 que, quanto maior o tamanho da amostra, maior é o poder do teste. Para um tamanho de amostra de 50 pessoas em cada população rejeitou-se a hipótese de igualdade da *vida média residual* em 97,5% das vezes considerando t_0 igual a 24 meses. Conforme o t_0 aumenta, o poder do teste diminui mesmo com $r_{t_0}^{med}$ ficando maior.

Na Tabela 4.6 temos o resultado encontrado no teste de igualdade da *vida mediana residual*. Ao contrário do que aconteceu na vida média, neste caso quanto maior o valor de t_0 , menor foi a razão verdadeira da *vida mediana residual*. Porém, o teste se mostra com poder crescente à medida que ocorre o aumento da amostra e menos poderoso com o aumento de t_0 . Com uma razão populacional de 2,04 foi necessária uma amostra de tamanho 200 em cada população para que a hipótese de igualdade fosse rejeitada em mais de 95% das vezes.

Na Seção 3.1 foi apresentado um teste para igualdade entre a *vida mediana residual* de duas populações independentes baseado na função de estimação $u(\theta_{t_0})$. A região de rejeição desse teste permite calcular um intervalo de confiança para $r_{t_0}^{md}$.

Porcentagem de rejeição				
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$
24	3,36	71,4	96,9	99,8
42	2,69	53,6	88,1	99,9
60	2,04	44,9	67,1	95,7

Tabela 4.6: Resultados do teste de razão de verossimilhanças empíricas para a vida mediana residual

Um estudo de simulação também foi realizado para esse caso. Os cenários considerados foram os mesmos do teste da razão de verossimilhanças empíricas sendo que neste caso também avaliamos o caso com $n = 300$, pois o tempo de processamento do teste baseado na função de estimação é mais rápido que o teste da razão de verossimilhanças empíricas. Para cada cenário foram geradas 5000 réplicas. A porcentagem de vezes em que a hipótese nula do teste foi rejeitada para cada cenário está na Tabela 4.7.

Porcentagem de rejeição					
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	3,36	65,1	94,6	100,0	100,0
42	2,69	51,4	82,6	99,4	99,9
60	2,04	46,6	65,3	91,9	99,2

Tabela 4.7: Resultados do teste da função de estimação para a vida mediana residual

A Tabela 4.8 mostra a amplitude média dos intervalos de confiança gerados nas simulações e também quantas vezes o verdadeiro valor da razão da *vida mediana residual* pertencia ao intervalo de confiança. Quanto maior o tamanho da amostra, menor a amplitude média do intervalo gerado, porém menor é a quantidade de vezes que o verdadeiro valor pertencia ao intervalo de confiança. O aumento de t_0 também afeta a assertividade do intervalo de confiança.

Porcentagem de cobertura e amplitude média					
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	3,36	98,3(8,2)	98,7(6,3)	98,7(4,6)	97,8(3,5)
42	2,69	95,1(11,9)	95,3(8,6)	89,8(6,6)	79,2(5,4)
60	2,04	87,8(19,3)	84,7(13,1)	72,2(9,3)	55,1(7,7)

Tabela 4.8: Porcentagem de vezes que $r_{t_0}^{md}$ pertencia ao intervalo de confiança (amplitude média do intervalo)

Uma variável de interesse dos pesquisadores é a fração de ejeção do coração, como explicado na Seção 1.5. Portanto, mais um estudo de simulação foi realizado, dessa vez a variável usada para a divisão da população foi a fração de ejeção categorizada entre valores menores do que 30% ou maiores do que 30%. A Figura 4.4 mostra a curva de Kaplan-Meier segundo essa categorização.

Os pacientes com fração de ejeção maior do que 30% têm sobrevida maior do que as pessoas com fração de ejeção abaixo de 30%. Considerando o horizonte de 80 meses a *vida média residual* é maior pra pacientes com fração de ejeção alta. Após isso, a situação se inverte e as pessoas com fração de ejeção baixa ficam com uma *vida média residual* maior, como pode ser visto na Figura 4.5.

Vemos na Figura 4.6 que para a *vida mediana residual* há uma diferença até os 40 meses. Depois disso, as curvas ficam muito próximas e de forma paralela. A fração de ejeção tem um comportamento diferente do que a taxa de creatinina. O fato de essa medida ser importante para os pesquisadores motivou o estudo de simulação usando o ponto de corte de 30% para dividir a população.

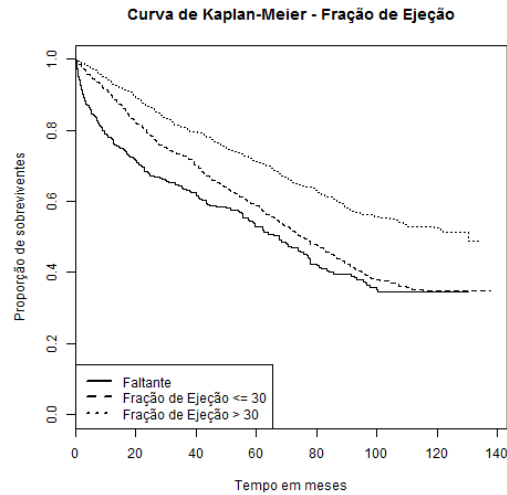


Figura 4.4: Estimador de Kaplan-Meier por faixa de fração de ejeção

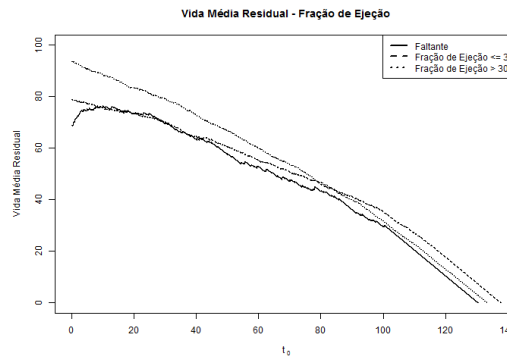


Figura 4.5: Vida média residual por faixa de fração de ejeção

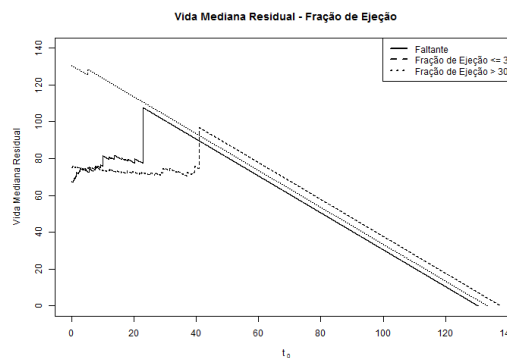


Figura 4.6: Vida mediana residual por faixa de fração de ejeção

Assim como foi feito para a taxa de creatinina, o banco de dados foi dividido e um modelo Weibull foi ajustado para os pacientes com fração de ejeção menor ou igual a 30%, e outro para as pessoas com fração de ejeção maior que 30%. As distribuições das respectivas censuras também foram estimadas.

O teste da razão de verossimilhanças empíricas foi estudado em três tempos diferentes ($t_0 = 24, 42$ e 60 meses) e quatro tamanhos de amostra diferentes ($n = 50, 100, 200$ e 300 para cada

população). Para cada cenário foram geradas 1000 réplicas considerando um nível de significância igual a 5%.

		Porcentagem de rejeição			
t_0	$r_{t_0}^{med}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	1,6	34,0	40,0	54,0	63,6
42	1,6	30,1	33,2	43,0	49,6
60	1,6	28,3	30,9	36,7	42,5

Tabela 4.9: Resultados do teste de razão de verossimilhanças empíricas para a vida média residual

Podemos ver na Tabela 4.9 que, quanto maior o tamanho da amostra, maior é o poder do teste. Porém, mesmo com uma amostra de tamanho 300 de cada população o teste rejeitou a igualdade em apenas 63,6% para t_0 igual a 24 meses, e menos da metade das vezes para os valores maiores de t_0 .

Na Tabela 4.10 temos o resultado encontrado no teste de igualdade da *vida mediana residual*. Assim como na simulação anterior, o teste para igualdade da mediana é menos poderoso que o teste de igualdade da *vida média residual*. Para t_0 igual a 60 meses e uma amostra de 300 indivíduos em cada população, o teste rejeitou a hipótese de igualdade em apenas 23,6% das vezes.

		Porcentagem de rejeição			
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	1,61	13,4	18,0	38,0	62,7
42	1,61	13,9	13,7	23,4	35,1
60	1,62	22,3	15,4	18,3	23,6

Tabela 4.10: Resultados do teste de razão de verossimilhanças empíricas para a vida mediana residual

O teste da função de estimação (Seção 3.1) foi aplicado para essa população simulada. Em cada cenário foram feitas 5000 simulações e os resultados encontram-se nas Tabelas 4.11 e 4.12. Em todos os cenários com tamanho de amostra maior que 100 em cada população, o intervalo de confiança para a razão entre as medianas da vida residual sobre o verdadeiro valor em mais de 95% das vezes. Porém, a amplitude do intervalo é grande e por isso o poder do teste é baixo. Para t_0 igual a 60 meses, mesmo com um tamanho de amostra de tamanho 1000 para cada população, o teste só rejeitou a igualdade em 45,9% das vezes.

		Porcentagem de rejeição				
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 1000$
24	1,61	4,9	8,1	21,9	42,4	98,8
42	1,61	5,1	4,4	8,7	17,7	86,1
60	1,62	5,2	4,2	4,7	8,4	45,9

Tabela 4.11: Resultados do teste da função de estimação para a vida mediana residual

Um último cenário de simulação foi considerado neste trabalho. Quando as populações são iguais com relação às suas distribuições do tempo de vida. A distribuição utilizada para a simulação foi a mesma da Seção 4.1. Foram considerados valores de t_0 iguais a 24, 42 e 60 meses e n iguais a 50, 100, 200 e 300 para cada população. O nível de significância do teste foi igual a 5%.

Nas Tabelas 4.13 e 4.14 encontram-se os resultados do teste da razão de verossimilhanças empíricas para a *vida média residual* e *vida mediana residual*, respectivamente. A Tabela 4.15 contém

Porcentagem de cobertura e amplitude média						
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 1000$
24	1,61	94,8(3,3)	97,9(2,7)	98,2(2,1)	98,6(1,9)	96,8(1,1)
42	1,61	97,5(3,1)	98,1(2,4)	98,2(1,9)	97,9(1,6)	95,2(0,8)
60	1,62	92,1(3,3)	96,2(2,9)	97,4(2,4)	98,6(2,1)	97,1(1,5)

Tabela 4.12: Porcentagem de vezes que $r_{t_0}^{md}$ estava dentro do intervalo de confiança (amplitude média do intervalo)

os resultados do teste que utiliza a função de estimação.

Porcentagem de rejeição					
t_0	$r_{t_0}^{med}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	1	19,5	20,4	19,8	21,7
42	1	23,4	20,9	22,4	23,0
60	1	27,5	25,5	24,1	24,4

Tabela 4.13: Resultados do teste da razão de verossimilhanças empíricas para a vida média residual

Porcentagem de rejeição					
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	1	5,5	2,0	2,4	1,2
42	1	9,9	4,5	2,1	2,3
60	1	17,2	8,7	4,7	3,2

Tabela 4.14: Resultados do teste da razão de verossimilhanças empíricas para a vida mediana residual

O teste para a *vida média residual* apresentou uma porcentagem de rejeição próxima de 20% em quase todos os cenários simulados, um valor mais alto que o teste da *vida mediana residual*.

Como o nível de significância utilizado no teste de hipótese foi igual a 5%, era esperado que a porcentagem de rejeição fosse próxima a esse valor. No caso da *vida média residual* o teste rejeitou mais do que o esperado e para a *vida mediana residual* o teste rejeitou menos do que 5% das vezes para tamanhos de amostra maiores.

Porcentagem de rejeição					
t_0	$r_{t_0}^{md}$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
24	1	1,54	1,26	1,13	0,90
42	1	3,40	0,58	0,56	1,15
60	1	5,50	1,81	1,07	0,72

Tabela 4.15: Resultados do teste da função de estimação para a vida mediana residual

4.4 Síntese dos resultados

Na Seção 4.1, a proporção de cobertura diminuiu com o aumento de t_0 tanto para a *vida média residual* quanto para a *vida mediana residual*. No caso da *vida mediana residual*, a amplitude média do intervalo aumentou com o aumento de t_0 , o que não ocorreu com a *vida média residual*.

Ao avaliar proporções de censura de 30%, 50% e 70%; para a *vida média residual* a proporção de cobertura não chegou aos 95% em nenhum dos cenários, no entanto, ela aumentou e a amplitude

média do intervalo diminuiu conforme o tamanho da amostra aumentava.

No caso da *vida mediana residual*, o aumento do tamanho da amostra não afetou a proporção de cobertura, mas a amplitude diminuiu para as censuras de 30% e 50%. Quando a censura foi de 70%, a amplitude não teve o mesmo comportamento com o aumento do tamanho da amostra.

Quando amostras simuladas de dois grupos separados segundo sua taxa de creatinina, tanto *vida média residual* quanto *vida mediana residual* eram bem diferentes entre si. A razão entre a *vida média residual* de um grupo e do outro grupo era maior que 4 e a razão entre a *vida mediana residual* era maior que 3 para $t_0 = 24$ e maior que 2 para t_0 igual a 42 ou 60 meses. Sendo assim, os três testes realizados rejeitaram hipótese de igualdade um grande número de vezes e para amostras de tamanho 200, por grupo, essa porcentagem de rejeição passou de 90%.

Para o caso em que os grupos foram separados segundo a fração de ejeção, a razão tinha valor próximo a 1,6 em todos os cenários. Nesse caso, o teste da razão de verossimilhanças empíricas apresentou porcentagens de rejeição maiores que o teste da função de estimação para a *vida mediana residual*.

Utilizando um nível de significância de 5% para os testes de hipóteses, era esperado que quando os grupos tivessem a mesma distribuição a porcentagem de rejeição ficasse próxima a esse valor. Porém, não foi o que ocorreu para o teste de razão de verossimilhanças empíricas da *vida média residual*, em torno de 20%. No caso da *vida mediana residual*, ambos os testes tiveram porcentagem de rejeição abaixo de 5% para amostras com tamanho maiores de 200.

Capítulo 5

Aplicação

Neste capítulo será feita uma análise estatística da *vida média residual* e da *vida mediana residual* no banco de dados do InCor. Iniciando com uma análise descritiva, depois utilizando os testes em duas populações apresentados nos Capítulos 2 e 3. Por fim, os modelos de regressão serão aplicados.

5.1 Análise descritiva

Um paciente de longo termo é aquele cujo tempo de vida será maior que o período da pesquisa realizada. Uma probabilidade de sobrevivência estimada maior do que 0 e a curva de Kaplan-Meier com pouca alteração na estimativa para tempos maiores sugerem a presença de pacientes de longo termo na amostra (Maller e Zhou, 1996). Na Figura 5.1 vemos que o estimador de Kaplan-Meier dos dados do InCor estima uma probabilidade de sobrevivência próxima a 0,4 para o maior tempo observado, o que nos leva a crer na existência de pacientes de longo termo. A proporção de censura para os dados do InCor foi de 54,5 %.

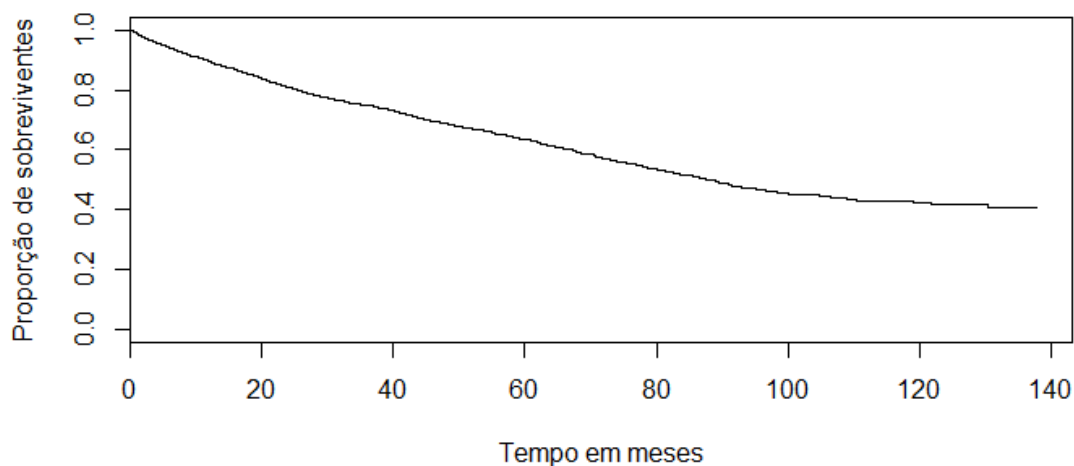


Figura 5.1: Estimador de Kaplan-Meier para o conjunto de dados

Nas Figuras 5.2 e 5.3 estão os estimadores não paramétricos da média e mediana de vida residual ao longo do tempo. Nesse caso, assim como para o tempo de sobrevivência, o estimador não paramétrico é usado como análise descritiva.

A *vida média residual* decresce quase linearmente ao longo do tempo. A *vida mediana residual* aumenta até o tempo de aproximadamente 24 meses e decresce linearmente a partir desse período. Os gráficos mostram que o estimador da *vida mediana residual* tem grandes “saltos” nos tempos de falha, os maiores desses “saltos” acontecem quando o tempo considerado é elevado e poucas

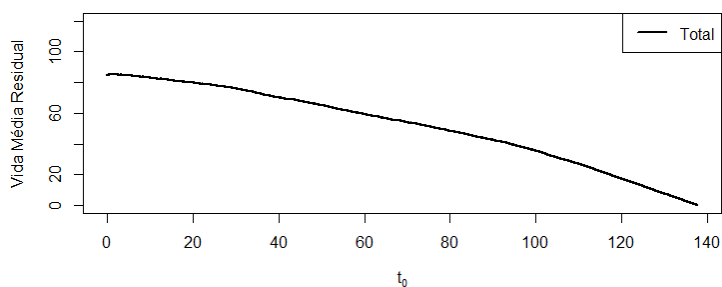


Figura 5.2: Vida média residual

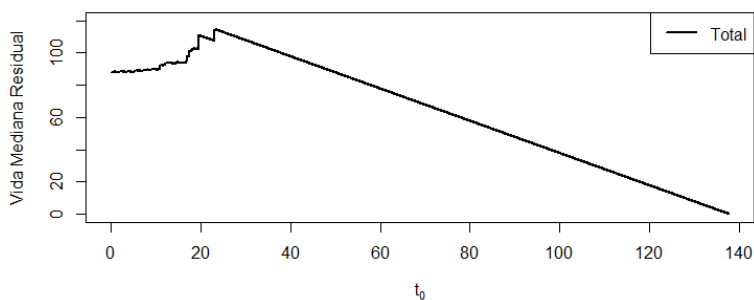


Figura 5.3: Vida mediana residual

observações que falharam têm um tempo de acompanhamento maior do que t_0 .

A estimação da *vida mediana residual* tem uma limitação decorrente do uso do estimador de Kaplan-Meier como estimador da função de sobrevivência. Na Figura 5.3 vemos que a partir de 25 meses, a estimativa é uma reta; isso acontece porque o menor valor que o estimador de Kaplan-Meier apresenta é 0,4, como visto no Gráfico 5.1. O estimador não paramétrico $\hat{\theta}(t) = \hat{S}^{-1}(\frac{1}{2}\hat{S}(t)) - t$ depende da inversa do estimador de Kaplan-Meier; portanto, para todos os valores de t em que $\frac{1}{2}\hat{S}(t) < 0,4$, a inversa do estimador de Kaplan-Meier terá o mesmo valor.

A Figura 5.4 mostra que os pacientes com taxa de creatinina maior do que 2,6 mg/dL têm uma vida média e mediana residual menor do que os outros pacientes em todos os tempos considerados. Os pacientes com fração de ejeção menor do que 30% têm uma *vida média residual* menor do que os pacientes com fração maior do que 30% até 60 meses e para a *vida mediana residual*, isso ocorre até 40 meses, como pode ser visto na Figura 5.5.

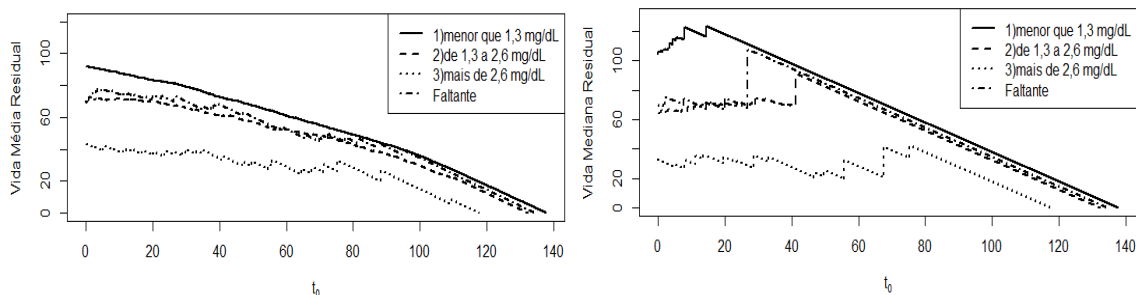


Figura 5.4: Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Taxa de Creatinina

Outra variável que explica o tempo de sobrevida dos pacientes com insuficiência cardíaca é a

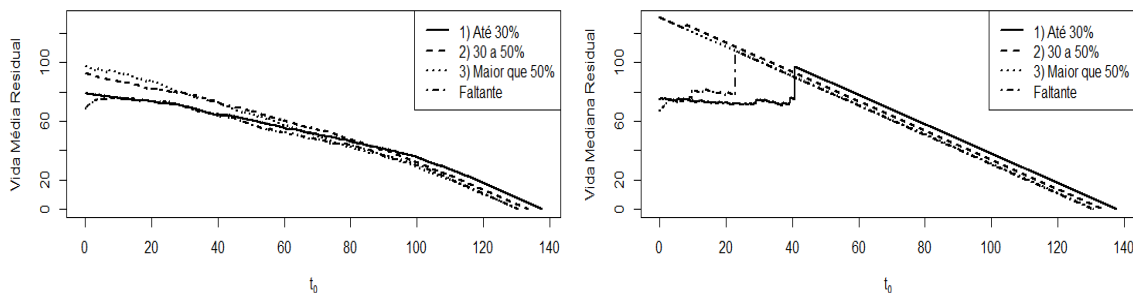


Figura 5.5: Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Fração de Ejeção

Classe Funcional. É uma classificação com categorias I, II, III e IV como dito na Seção 1.5; portanto, quanto maior o número, pior o estado de saúde do indivíduo. Na Figura 5.6 vê-se que as pessoas com a classe funcional IV têm uma *vida média residual* menor ao longo do tempo, já para a vida mediana é menor até quase 40 meses de acompanhamento. Os pacientes com classe funcional III também apresentam uma vida média e mediana residual menores do que as classes I e II.

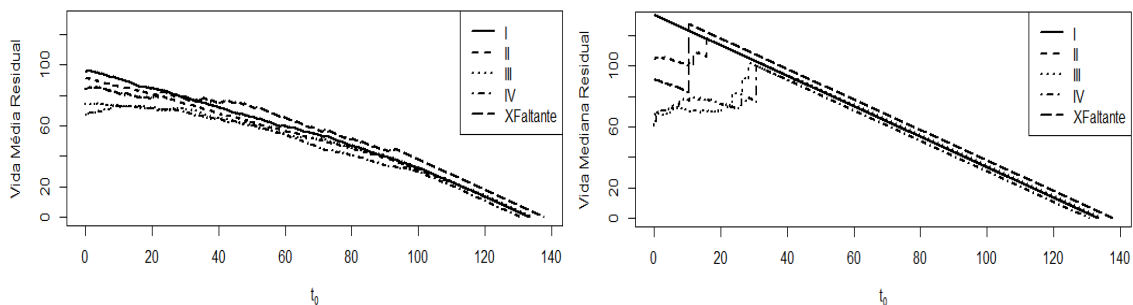


Figura 5.6: Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Classe Funcional

A Figura 5.7 descreve a vida residual de acordo com o diagnóstico realizado no paciente. Aqueles com doença de Chagas têm a menor *vida média residual* ao longo do tempo e têm a *vida mediana residual* menor até mais de 40 meses de acompanhamento.

Pacientes com diâmetro diastólico do ventrículo esquerdo maior do que 75mm têm uma vida média e mediana residual menor do que pessoas com o diâmetro menor do que 75mm, o que pode ser visto na Figura 5.8. Como esperado, a Figura 5.9 mostra que quanto maior a idade no início do acompanhamento, menor a vida média e mediana residual.

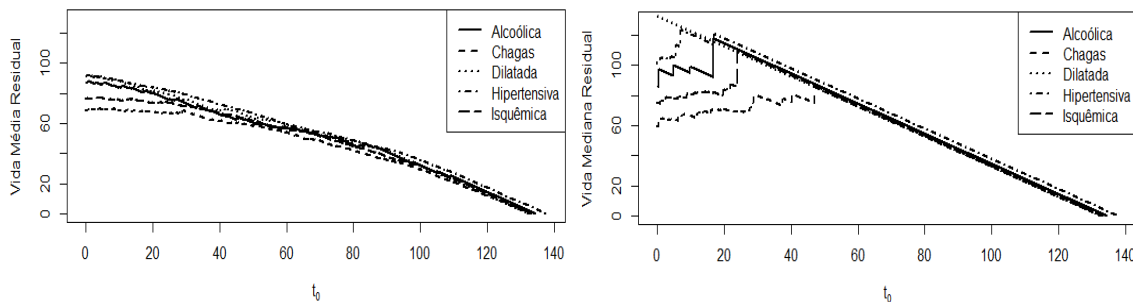


Figura 5.7: Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Etiologia

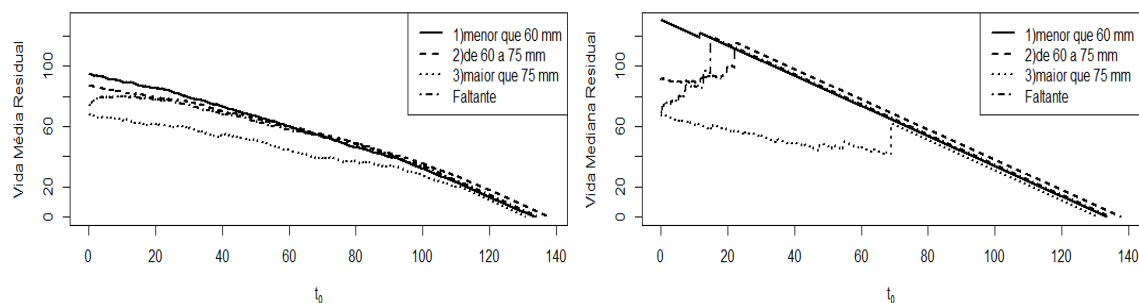


Figura 5.8: Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Diâmetro Diastólico do Ventrículo Esquerdo

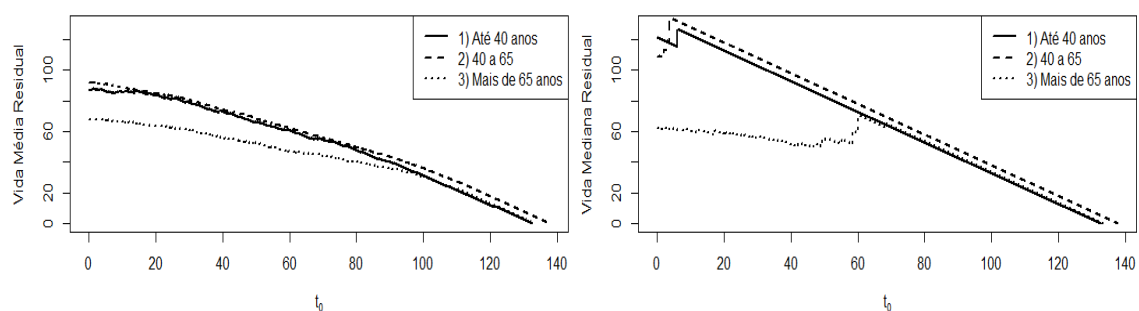


Figura 5.9: Estimativas da vida média residual (esquerda) e vida mediana residual (direita) para Idade

5.2 Teste dos efeitos associados com variáveis

Antes de ajustar os modelos de regressão mostrados nos Capítulos 2 e 3, serão feitos os testes apresentados anteriormente para verificar a igualdade das médias e medianas de vida residual de duas populações em dez tempos de acompanhamento: 0, 6, 12, 18, 24, 30, 36, 42, 48 e 60 meses. Os dados foram divididos de onze maneiras diferentes segundo as seis variáveis que tiveram a análise descritiva apresentada na Seção 5.1.

As tabelas com os resultados dos testes contêm a razão estimada entre as médias (medianas) de vida residual $r_{t_0}^{med}(r_{t_0}^{md})$ e o valor- p referente ao teste da razão de verossimilhanças empíricas (VE) e o teste que usa a função de estimação (FE).

Quando os dados são divididos com relação a pessoas com taxa de creatinina maior do que 2,6 mg/dL e menor do que 2,6, os testes baseados na verossimilhança empírica e na função de estimação rejeitam a hipótese de igualdade entre as médias e medianas de vida residual, como pode ser visto na Tabela 5.1.

A Tabela 5.2 mostra que quando a divisão da população é feita para taxa de creatinina igual a 1,3 mg/dL, as razões entre as médias e medianas de vida residual são mais próximas de 1, mas os três testes rejeitam a hipótese de igualdade.

Assim como foi feito para a taxa de creatinina, os dados foram divididos em dois grupos segundo a Fração de Ejeção do paciente. Foram feitos três diferentes cortes para essa variável. Na Tabela 5.3 estão os resultados dos testes quando a divisão é feita para a fração de ejeção de 25%.

Como visto na análise descritiva, quando olhamos os dados considerando apenas a fração de ejeção, em um determinado tempo t_0 as pessoas com fração de ejeção maior apresentam *vida mediana residual* menor do que os outros pacientes, o que não acontece na *vida média residual*. Outro ponto a se destacar é que os testes rejeitam a hipótese de igualdade das medianas de vida residual em

todos os cenários considerados. O teste para a *vida média residual* apresenta valor- p maior conforme a razão estimada se aproxima do 1.

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,5015	<0,0001	0,3632	<0,0001	<0,0001
6	0,4891	<0,0001	0,3101	<0,0001	<0,0001
12	0,4627	<0,0001	0,3638	<0,0001	<0,0001
18	0,4705	<0,0001	0,2808	<0,0001	0,0001
24	0,4928	<0,0001	0,2788	<0,0001	0,0002
30	0,4957	<0,0001	0,3194	<0,0001	0,0001
36	0,5234	<0,0001	0,3097	<0,0001	<0,0001
42	0,4941	<0,0001	0,2664	<0,0001	<0,0001
48	0,4670	<0,0001	0,2558	<0,0001	<0,0001
60	0,4830	<0,0001	0,3623	<0,0001	<0,0001

Tabela 5.1: Taxa de Creatinina $\geq 2,6$ mg/dL versus $< 2,6$ mg/dL

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,7422	<0,0001	0,5912	<0,0001	<0,0001
6	0,7687	<0,0001	0,5423	<0,0001	<0,0001
12	0,7916	<0,0001	0,5333	<0,0001	0,0001
18	0,8025	<0,0001	0,5434	<0,0001	0,0003
24	0,8130	<0,0001	0,5779	0,0010	0,0011
30	0,8188	<0,0001	0,6282	0,0083	0,0050
36	0,8170	<0,0001	0,6863	0,0164	0,0045
42	0,8400	<0,0001	0,7075	0,0272	0,0029
48	0,8343	<0,0001	0,9382	0,0106	0,0002
60	0,8464	<0,0001	0,9287	0,0002	<0,0001

Tabela 5.2: Taxa de Creatinina $\geq 1,3$ mg/dL versus $< 1,3$ mg/dL

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	1,1856	<0,0001	1,5229	<0,0001	<0,0001
6	1,1628	<0,0001	1,6452	<0,0001	<0,0001
12	1,1558	0,0001	1,7067	<0,0001	0,0008
18	1,1224	0,0014	1,6558	0,0005	0,0020
24	1,1285	0,0008	1,5965	0,0046	0,0036
30	1,1242	0,0013	1,4986	0,0183	0,0089
36	1,1372	0,0007	1,3917	0,0243	0,0056
42	1,1017	0,0117	1,2468	0,0365	0,0038
48	1,0778	0,0525	0,9521	0,0137	0,0004
60	1,0751	0,0742	0,9447	0,0005	<0,0001

Tabela 5.3: Fração de Ejeção $\geq 25\%$ versus $< 25\%$

Nas Tabelas 5.4 e 5.5 são apresentados resultados para os cortes em 30% e 35% da fração de ejeção. Para a *vida média residual*, o teste rejeitou a igualdade em todos os tempos considerados quando os dados foram divididos entre pessoas com fração de ejeção maior que 30% e menor que esse valor. No caso da *vida mediana residual*, ocorreu o mesmo comportamento observado na Tabela

5.3, em que a razão estimada das medianas é maior do que 1 até um certo tempo e menor do que 1 a partir dele e ainda assim o teste rejeitou a igualdade.

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	1,1881	<0,0001	1,7355	<0,0001	0,0001
6	1,1677	<0,0001	1,7017	<0,0001	0,0001
12	1,1698	<0,0001	1,6509	<0,0001	0,0002
18	1,1313	0,0001	1,5812	0,0002	0,0003
24	1,1363	<0,0001	1,5228	0,0038	0,0018
30	1,1280	0,0001	1,4019	0,0248	0,0065
36	1,1425	<0,0001	1,3648	0,0338	0,0041
42	1,1133	0,0009	0,9551	0,0335	0,0022
48	1,1009	0,0031	0,9521	0,0066	0,0002
60	1,0888	0,0115	0,9447	<0,0001	<0,0001

Tabela 5.4: Fração de Ejeção $\geq 30\%$ versus $< 30\%$

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	1,1583	<0,0001	1,6025	<0,0001	0,0024
6	1,1523	<0,0001	1,5472	<0,0001	0,0055
12	1,1502	<0,0001	1,5384	0,0001	0,0055
18	1,1094	0,0006	1,4690	0,0148	0,0152
24	1,0933	0,0031	1,3190	0,2189	0,1978
30	1,0728	0,0201	0,9601	0,2789	0,1952
36	1,0815	0,0109	0,9578	0,0975	0,0552
42	1,0658	0,0418	0,9551	0,0100	0,0049
48	1,0640	0,0469	0,9521	0,0003	0,0001
60	1,0454	0,1652	0,9447	<0,0001	<0,0001

Tabela 5.5: Fração de Ejeção $\geq 35\%$ versus $< 35\%$

Na Tabela 5.6 estão os resultados dos testes de hipóteses comparando pacientes de classe funcional III ou IV com pacientes de classe funcional I ou II.

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,7832	<0,0001	0,6180	<0,0001	<0,0001
6	0,8233	<0,0001	0,6244	<0,0001	<0,0001
12	0,8511	<0,0001	0,6054	<0,0001	0,0001
18	0,8739	<0,0001	0,6344	0,0009	0,0028
24	0,8973	0,0004	0,6975	0,0046	0,0137
30	0,9406	0,0439	0,9707	0,0053	0,0065
36	0,9383	0,0418	0,9689	0,0009	0,0006
42	0,9514	0,1198	1,0104	0,0001	<0,0001
48	0,9624	0,2352	1,0111	<0,0001	<0,0001
60	0,9610	0,2342	1,0129	<0,0001	<0,0001

Tabela 5.6: Classe Funcional III ou IV versus I ou II

O teste da razão de verossimilhanças empíricas não rejeita a hipótese de igualdade das médias

de vida residual nos tempo 42, 48 e 60 meses. O teste para a *vida mediana residual* rejeita a hipótese de igualdade mesmo quando a razão estimada é muito próxima de 1. Na Tabela 5.7 estão os resultados quando são comparadas pessoas de classe funcional IV com pessoas de classe funcional I, II ou III. O teste da *vida média residual* rejeita a igualdade até o tempo de 24 meses. Neste cenário os testes da razão de verossimilhanças empíricas e da função de estimação apresentaram resultados diferentes dependendo do tempo considerado.

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,7736	<0,0001	0,6659	<0,0001	0,0008
6	0,8386	0,0001	0,7895	0,0009	0,0244
12	0,8879	0,0066	0,8146	0,0012	0,0303
18	0,9084	0,0268	0,7384	0,0033	0,1357
24	0,9062	0,0236	0,6576	0,0019	0,3950
30	0,9240	0,0667	0,7396	0,0003	0,3330
36	0,9314	0,1062	0,9613	<0,0001	0,1117
42	0,9357	0,1355	0,9588	<0,0001	0,0146
48	0,9713	0,4974	0,9559	<0,0001	0,0004
60	0,9437	0,1861	0,9488	<0,0001	<0,0001

Tabela 5.7: Classe Funcional IV versus I, II ou III

Na Tabela 5.8 temos a comparação entre pessoas com e sem doença de Chagas. Em todos os cenários a hipótese de igualdade foi rejeitada.

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,7847	<0,0001	0,6522	<0,0001	0,0002
6	0,8019	<0,0001	0,6714	<0,0001	0,0011
12	0,8108	<0,0001	0,6876	<0,0001	0,0018
18	0,8284	<0,0001	0,6279	0,0005	0,0180
24	0,8346	<0,0001	0,6044	0,0050	0,0176
30	0,8789	0,0018	0,7397	0,0341	0,0196
36	0,8509	0,0002	0,7308	0,0238	0,0063
42	0,8633	0,0007	0,8268	0,0252	0,0023
48	0,8791	0,0027	0,9414	0,0136	0,0002
60	0,9040	0,0168	0,9324	0,0011	<0,0001

Tabela 5.8: Chagas versus Outras doenças

Na Tabela 5.9 são apresentados os resultados da comparação entre pessoas com diâmetro diastólico do ventrículo esquerdo maior do que 75 mm e menor do que 75 mm. A razão da *vida média residual* estimada varia pouco ao longo do tempo, em todos os cenários a hipótese de igualdade foi rejeitada.

A idade é uma variável que naturalmente explica o tempo de sobrevida. Na Tabela 5.10 vemos que a hipótese de igualdade é rejeitada em todos os cenários quando os pacientes são divididos entre indivíduos com mais de 65 anos e menos que 65 anos de idade.

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,7510	<0,0001	0,7325	<0,0001	<0,0001
6	0,7518	<0,0001	0,6557	<0,0001	<0,0001
12	0,7508	<0,0001	0,5560	<0,0001	<0,0001
18	0,7432	<0,0001	0,4939	<0,0001	<0,0001
24	0,7389	<0,0001	0,4881	0,0005	<0,0001
30	0,7515	<0,0001	0,4980	0,0097	<0,0001
36	0,7352	<0,0001	0,4978	0,0159	<0,0001
42	0,7652	<0,0001	0,4992	0,0343	0,0001
48	0,7595	<0,0001	0,5320	0,0173	<0,0001
60	0,7322	<0,0001	0,5971	0,0009	<0,0001

Tabela 5.9: Diâmetro Diastólico do ventrículo esquerdo ≥ 75 mm versus < 75 mm

t_0	Média		Mediana		
	$r_{t_0}^{med}$	valor- p VE	$r_{t_0}^{md}$	valor- p VE	valor- p FE
0	0,7402	<0,0001	0,5715	<0,0001	<0,0001
6	0,7446	<0,0001	0,4638	<0,0001	<0,0001
12	0,7520	<0,0001	0,4768	<0,0001	<0,0001
18	0,7467	<0,0001	0,5000	<0,0001	<0,0001
24	0,7512	<0,0001	0,5065	<0,0001	<0,0001
30	0,7531	<0,0001	0,5243	<0,0001	<0,0001
36	0,7555	<0,0001	0,5314	<0,0001	<0,0001
42	0,7562	<0,0001	0,5243	<0,0001	<0,0001
48	0,7510	<0,0001	0,5697	<0,0001	<0,0001
60	0,7556	<0,0001	0,7597	<0,0001	<0,0001

Tabela 5.10: Idade ≥ 65 anos versus < 65 anos

5.3 Modelos de regressão

De acordo com todas as análises feitas nas Seções 5.1 e 5.2, seis variáveis parecem explicar a vida média e mediana residual. Para a construção do modelo de regressão será utilizado o seguinte método para a inclusão das variáveis:

- Ajustar o modelo com apenas uma das variáveis, aquela cujo valor- p apresentar o menor valor permanecerá no modelo;
- Ajustar um modelo com duas variáveis, mantendo a que foi escolhida no passo 1 e testar as outras cinco variáveis, uma de cada vez, e no final, aquela que o valor- p apresentar o menor valor ficará no modelo;
- Os próximos passos são feitos de forma análoga até completar o modelo com as seis variáveis.

O modelo de regressão da *vida média residual* supõe a proporcionalidade da mesma ao longo do tempo (vide (2.5)), enquanto o modelo de regressão da *vida mediana residual* é ajustado para um determinado tempo t_0 . Sendo assim, os parâmetros para o modelo da *vida média residual* não incluem um termo para o intercepto, porém o modelo da *vida mediana residual* possibilita sua inclusão.

5.3.1 Vida média residual

O modelo de regressão da *vida média residual* que foi apresentado no Capítulo 2 é, de certa forma, semelhante ao modelo de riscos proporcionais de Cox.

A Tabela 5.11 apresenta os resultados da seleção de variáveis para o modelo de regressão da *vida média residual*. A fração de ejeção foi a variável com o maior valor- p no primeiro passo da seleção. O diâmetro diastólico do ventrículo esquerdo também foi significativo a um nível de 5%. Se fosse considerado um nível de significância de 10%, a etiologia cardiomiopatia isquêmica também seria considerada no modelo no passo 3.

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
FE < 25	-1,10(<0,01)	-1,08(<0,01)	-0,91(<0,01)	-0,83(<0,01)	-0,72(<0,01)	-0,71(<0,01)
FE Faltante	-1,61(<0,01)	-1,62(<0,01)	-1,43(<0,01)	-1,34(<0,01)	-1,23(<0,01)	-1,22(<0,01)
Diâmetro > 75		-0,99(<0,01)	-0,83(0,03)	-0,76(0,03)	-0,80(0,03)	-0,79(0,02)
Chagas			-0,52(0,20)	-0,52(0,20)	-0,55(0,15)	-0,56(0,13)
Isquêmica			-0,62(0,06)	-0,59(0,08)	-0,56(0,08)	-0,56(0,07)
Classe III e IV				-0,26(0,30)	-0,22(0,38)	-0,22(0,38)
Idade					-0,01(0,44)	-0,01(0,40)
Creatinina Alta						-0,23(0,71)

Tabela 5.11: Seleção de variáveis.
Modelo de regressão para o tempo médio residual

5.3.2 Vida mediana residual

O modelo de regressão que será considerado para modelar a *vida mediana residual* nos dados do InCor é o mostrado na Seção 3.3. O processo de seleção de variáveis foi o mesmo descrito no início desta seção. As variáveis escolhidas para explicar a *vida mediana residual* foram Classe Funcional, Etiologia, Diâmetro Diastólico do ventrículo esquerdo, Fração de Ejeção, Taxa de Creatinina e Idade.

Seguindo a análise descritiva e os testes realizados nas Seções 5.1 e 5.2, o efeito da variável Classe Funcional será testado criando-se uma variável indicadora para quem foi classificado nas categorias III ou IV, o efeito da Etiologia será testado com uma variável indicando quem tem doença de Chagas e outra indicando quem tem cardiomiopatia isquêmica. Para o diâmetro diastólico foi considerado como ponto de corte o valor 75mm, para a Fração de Ejeção foram atribuídas duas variáveis, uma indicando se a fração é menor do que 25% e outra indicando que essa informação está faltando. Uma variável binária foi criada para indicar pessoas que tem taxa de creatinina alta (maior do que 2,6 mg/dL). Como a variável Idade é contínua, foi definida uma nova variável dada por $Idade2 = Idade - 60$, a fim de facilitar a interpretação do efeito correspondente.

A Tabela 5.12 contém a sequência da seleção das variáveis do modelo de regressão quando o tempo t_0 foi fixado em 0 mês.

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
Intercepto	4,93(<0,01)	4,93(<0,01)	4,93(<0,01)	4,65(<0,01)	4,66(<0,01)	4,65(<0,01)
Classe III e IV	-1,41(<0,01)	-0,91(<0,01)	-0,85(0,01)	-0,53(0,07)	-0,53(0,06)	-0,52(0,06)
Chagas		-0,88(0,04)	-0,87(0,04)	-0,59(0,14)	-0,59(0,13)	-0,59(0,12)
Isquêmica		-0,91(0,01)	-0,88(0,02)	-0,55(0,10)	-0,53(0,11)	-0,51(0,11)
Diâmetro > 75			-0,52(0,14)	-0,58(0,10)	-0,58(0,10)	-0,61(0,07)
FE < 25%				-0,27(0,26)	-0,27(0,24)	-0,26(0,24)
FE Faltante				-0,99(0,03)	-0,98(0,04)	-0,98(0,04)
Creatinina Alta					-0,50(0,49)	-0,52(0,48)
Idade						-0,001(0,91)

Tabela 5.12: Seleção de variáveis para $t_0 = 0$.
Modelo de regressão para o tempo mediano residual

A Classe Funcional foi a variável com o menor valor- p quando foram testados modelos com apenas uma das variáveis. No segundo passo, fixando-se a Classe Funcional no modelo, a Etiologia foi a variável com o menor valor- p . Os passos seguintes seguiram o mesmo procedimento. Se for considerado que uma variável é significativa no modelo quando seu valor- p for menor que 5%, o modelo com Classe Funcional e Etiologia seria escolhido. A inclusão de mais variáveis no modelo fez com que o valor- p dessas variáveis ficasse maior. No modelo com todas as variáveis, por exemplo, somente para a Fração de Ejeção Faltante o valor- p foi menor que 0,05.

O mesmo método de seleção de variáveis foi usado quando t_0 foi fixado em 12, 24, 36, 48 e 60 meses. Para o tempo de vida mediano dado que o paciente sobreviveu 12 meses, se for considerado que uma variável for significativa quando o valor- p for menor que 10% o modelo com Classe Funcional, Fração de Ejeção e Diâmetro Diastólico tem todas as variáveis significantes. O resultado para cada passo está na Tabela 5.13.

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
Intercepto	4,83(<0,01)	4,83(<0,01)	4,83(<0,01)	4,83(<0,01)	4,83(<0,01)	4,80(<0,01)
Classe III e IV	-1,16(<0,01)	-0,69(0,04)	-0,58(0,08)	-0,44(0,15)	-0,45(0,15)	-0,43(0,14)
FE < 25		-0,82(<0,01)	-0,81(<0,01)	-0,64(<0,01)	-0,62(0,01)	-0,56(0,02)
FE Faltante		-0,92(0,03)	-0,92(0,03)	-0,81(0,04)	-0,82(0,04)	-0,78(0,06)
Diâmetro > 75			-0,68(0,09)	-0,66(0,09)	-0,66(0,09)	-0,71(0,07)
Chagas				-0,52(0,26)	-0,52(0,28)	-0,54(0,22)
Isquêmica				-0,57(0,14)	-0,55(0,13)	-0,56(0,11)
Creatinina Alta					-0,75(0,43)	-0,77(0,43)
Idade						-0,003(0,78)

Tabela 5.13: Seleção de variáveis para $t_0 = 12$.
Modelo de regressão para o tempo mediano residual

Na Tabela 5.14 são apresentados os resultados para t_0 igual a 24 meses, mais uma vez o valor- p da variável Classe Funcional aumenta conforme outras variáveis são incluídas no modelo. A fração de ejeção é significativa em todos os passos e o Diâmetro Diastólico também apresenta certa significância.

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
Intercepto	4,73(<0,01)	4,73(<0,01)	4,73(<0,01)	4,73(<0,01)	4,73(<0,01)	4,66(<0,01)
Classe III e IV	-1,03(<0,01)	-0,57(0,07)	-0,43(0,19)	-0,38(0,23)	-0,37(0,25)	-0,32(0,32)
FE < 25		-0,81(<0,01)	-0,80(<0,01)	-0,64(0,01)	-0,66(0,01)	-0,56(0,03)
FE Faltante		-0,95(0,01)	-0,96(0,01)	-0,87(0,01)	-0,85(0,02)	-0,75(0,04)
Diâmetro > 75			-0,75(0,09)	-0,76(0,10)	-0,76(0,09)	-0,72(0,12)
Chagas				-0,58(0,25)	-0,57(0,24)	-0,63(0,17)
Isquêmica				-0,45(0,25)	-0,43(0,27)	-0,42(0,27)
Creatinina Alta					-0,66(0,50)	-0,76(0,45)
Idade						-0,006(0,62)

Tabela 5.14: Seleção de variáveis para $t_0 = 24$.
Modelo de regressão para o tempo mediano residual

As Tabelas 5.15, 5.16 e 5.17 mostram os resultados dos modelos com t_0 igual a 36, 48 e 60 meses, respectivamente. A variável Classe Funcional e Etiologia (doença de Chagas e cardiomiopatia isquêmica) apresentaram menor significância com o aumento do tempo considerado. A Idade teve o comportamento inverso: quanto maior t_0 , mais significativa foi seu efeito. No entanto, em nenhum cenário a Idade teve um valor- p menor que 40%.

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
Intercepto	4,62(<0,01)	4,62(<0,01)	4,62(<0,01)	4,62(<0,01)	4,52(<0,01)	4,51(<0,01)
Classe III e IV	-1,07(<0,01)	-0,44(0,20)	-0,33(0,34)	-0,33(0,35)	-0,24(0,48)	-0,27(0,43)
FE < 25		-0,83(<0,01)	-0,82(<0,01)	-0,69(0,01)	-0,55(0,04)	-0,53(0,05)
FE Faltante		-1,10(0,01)	-1,10(0,01)	-0,91(0,03)	-0,79(0,06)	-0,77(0,08)
Diâmetro > 75			-0,78(0,14)	-0,65(0,21)	-0,69(0,20)	-0,58(0,26)
Chagas				-0,45(0,43)	-0,51(0,35)	-0,53(0,32)
Isquêmica				-0,50(0,24)	-0,51(0,23)	-0,51(0,23)
Idade					-0,01(0,51)	-0,01(0,49)
Creatinina Alta						-0,37(0,67)

Tabela 5.15: Seleção de variáveis para $t_0 = 36$.
Modelo de regressão para o tempo mediano residual

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
Intercepto	4,49(<0,01)	4,49(<0,01)	4,49(<0,01)	4,38(<0,01)	4,39(<0,01)	4,39(<0,01)
FE < 25	-1,02(<0,01)	-0,94(<0,01)	-0,84(0,01)	-0,71(0,03)	-0,62(0,04)	-0,61(0,05)
FE Faltante	-1,24(<0,01)	-1,24(<0,01)	-1,21(<0,01)	-1,05(0,02)	-0,92(0,06)	-0,90(0,06)
Diâmetro > 75		-0,77(0,14)	-0,55(0,33)	-0,71(0,20)	-0,59(0,29)	-0,59(0,30)
Chagas			-0,37(0,56)	-0,28(0,65)	-0,31(0,63)	-0,32(0,61)
Isquêmica			-0,65(0,21)	-0,56(0,27)	-0,56(0,25)	-0,56(0,26)
Idade				-0,01(0,55)	-0,01(0,56)	-0,01(0,54)
Classe III e IV					-0,19(0,62)	-0,20(0,61)
Creatinina Alta						-0,57(0,62)

Tabela 5.16: Seleção de variáveis para $t_0 = 48$.
Modelo de regressão para o tempo mediano residual

Variável	Estimativa do parâmetro (valor- p)					
	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5	Passo 6
Intercepto	4,35(<0,01)	4,35(<0,01)	4,35(<0,01)	4,17(<0,01)	4,19(<0,01)	4,19(<0,01)
FE < 25	-1,07(<0,01)	-1,01(<0,01)	-1,00(0,01)	-0,81(0,04)	-0,71(0,07)	-0,65(0,07)
FE Faltante	-1,49(<0,01)	-1,46(<0,01)	-1,46(<0,01)	-1,33(0,02)	-1,21(0,04)	-1,17(0,06)
Diâmetro > 75		-0,92(0,13)	-0,92(0,12)	-0,92(0,13)	-0,88(0,21)	-0,85(0,20)
Creatinina Alta			-0,87(0,27)	-0,59(0,54)	-0,75(0,48)	-0,79(0,47)
Idade				-0,02(0,51)	-0,02(0,53)	-0,02(0,48)
Chagas					-0,25(0,70)	-0,28(0,66)
Isquêmica					-0,22(0,73)	-0,26(0,69)
Classe III e IV						-0,14(0,76)

Tabela 5.17: Seleção de variáveis para $t_0 = 60$.
Modelo de regressão para o tempo mediano residual

Capítulo 6

Discussão e Conclusão

Neste trabalho foram apresentados estimadores não paramétricos para a *vida média residual* e para a *vida mediana residual*. Testes para igualdade dessas medidas em duas populações foram apresentados bem como modelos de regressão. Os estudos de simulação mostraram que a alta presença de censura faz com que os estimadores e testes precisem de um tamanho de amostra maior para uma melhor eficiência, em termos de amplitude dos intervalos de confiança.

Como discutido no início da Seção 5.1, o estimador de Kaplan-Meier leva a uma limitação na estimação da *vida mediana residual*. Para dados com característica de longo termo é interessante usar um estimador não paramétrico da função de sobrevivência mais apropriado do que o de Kaplan-Meier. A *vida média residual* possui as mesmas críticas que a média tradicional, é sensível a valores extremos. No contexto de análise de sobrevivência, em geral, as distribuições dos tempos de vida são assimétricas, o que tornam essa medida não muito adequada.

A aplicação do modelo de regressão da *vida média residual* ao conjunto de dados de pacientes com insuficiência cardíaca mostrou que a Fração de Ejeção menor que 25%, o Diâmetro Diastólico do ventrículo esquerdo maior que 75 mm e a Cardiomiopatia Isquêmica explicam uma *vida média residual* menor ao longo do tempo.

No modelo de regressão da *vida mediana residual* para o início do tratamento, a Classe Funcional, o Diâmetro Diastólico e a Etiologia foram significativos para explicar a *vida mediana residual*. Para tempos maiores, a Fração de Ejeção se mostrou significativo e a Classe Funcional perdeu sua significância, pois muitos dos pacientes em situação delicada (classes III e IV) não conseguiram viver por tantos anos. Assim como aconteceu com a Etiologia.

Os estimadores, testes e modelos apresentados neste trabalho foram originados através da perspectiva clássica da estatística. Gelfand e Kottas (2003) formularam um modelo de regressão semi-paramétrico Bayesiano para a *vida mediana residual* baseado no modelo de vida acelerada e ele foi aplicado em um conjunto de dados com apenas 19% de censuras.

No contexto de Análise de Sobrevivência, a mediana é mais utilizada do que a média por conta da natureza assimétrica da distribuição do tempo de vida. Os estudos de simulação feitos neste trabalho mostraram resultados que corroboram essa prática quando analisada a vida residual.

As simulações feitas neste trabalho mostraram instabilidades e resultados não esperados em alguns cenários. Pesquisas futuras podem ter um foco em estudos de simulações que considerem mais cenários de censura e tamanhos de amostra. Também seria interessante utilizar outras distribuições para a geração das amostras.

Na aplicação a dados reais, pode-se notar a presença de indivíduos de longo termo. Essa característica não foi explicitamente considerada na construção de estimadores e testes, constituindo-se, assim, importante tópico para pesquisas posteriores.

Referências Bibliográficas

- Aalen(1978)** Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726. Citado na pág. [3](#)
- Andersen e Gill(1982)** P. K. Andersen e R. Gill. Cox’s regression model for counting processes: A large sample study. *Annals of Statistics*, 10:1100–1200. Citado na pág. [6](#)
- Asadi e Bayramoglu(2005)** Majid Asadi e Ismihan Bayramoglu. A note on the mean residual life function of a parallel system. *Communications in Statistics - Theory and Methods*, 34:475–484. Citado na pág. [1](#)
- Bathke et al.(2009)** Arne Bathke, Mi-Ok Kim e Mai Zhou. Combined multiple testing by censored empirical likelihood. *Journal of Statistical Planning and Inference*, 139:814–827. Citado na pág. [30](#)
- Benda(2003)** B. B. Benda. Survival analysis of criminal recidivism of boot camp graduates using elements from general and developmental explanatory models. *International Journal of Offender Therapy and Comparative Criminology*, 47:89–110. Citado na pág. [1](#)
- Berger(1982)** Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300. Citado na pág. [15](#)
- Berger et al.(1988)** Roger L. Berger, Dennis D. Boos e Frank M. Guess. Tests and confidence sets for comparing two mean residual life functions. *Biometrics*, 44:103–115. Citado na pág. [14](#), [15](#), [16](#)
- Bohoris(1994)** George Bohoris. Comparison of the cumulative-hazard and kaplan-meier estimators of the survival function. *IEEE Transactions on Reliability*, 43:230–232. Citado na pág. [3](#)
- Breslow e Crowley(1974)** N. E. Breslow e J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2:437–453. Citado na pág. [3](#)
- Chen et al.(2005)** Y. Q. Chen, N. P. Jewell, X. Lei e S. C. Cheng. Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics*, 61:170–178. Citado na pág. [20](#), [21](#), [22](#)
- Colosimo e Giolo(2006)** Enrico Antônio Colosimo e Suely Ruiz Giolo. *Análise de Sobrevivência Aplicada*. Blucher. Citado na pág. [1](#), [2](#), [5](#)
- Cox(1975)** D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276. Citado na pág. [4](#)
- Cox(1972)** D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistics Society, Series B*, 34:187–220. Citado na pág. [4](#)
- Fleming e Harrington(1991)** Thomas R. Fleming e David P. Harrington. *Counting Processes and Survival Analysis*. Wiley. Citado na pág. [25](#), [27](#)
- Gelfand e Kottas(2003)** Alan E. Gelfand e Athanasios Kottas. Bayesian semiparametric regression for median residual life. *Scandinavian Journal of Statistics*, 30:651–665. Citado na pág. [59](#)

- Greenwood(1926)** Major Greenwood. The natural duration of cancer. *Reports On Public Health and Medical Subjects*, 33:1–26. Citado na pág. [3](#)
- Hall e Wellner(1981)** W. J. Hall e Jon A. Wellner. Mean residual life. *Statistics and Related Topics*, 1:169–184. Citado na pág. [10](#), [13](#)
- Harrington e Fleming(1982)** David P. Harrington e Thomas R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69:553–566. Citado na pág. [4](#)
- Jeong(2014)** Jong-Hyeon Jeong. *Statistical Inference on Residual Life*. Springer. Citado na pág. [7](#), [25](#), [26](#), [27](#)
- Jeong et al.(2007)** Jong-Hyeon Jeong, Sin-Ho Jung e Joseph P. Costantino. Nonparametric inference on median residual life function. *Biometrics*, 64:157–163. Citado na pág. [28](#), [29](#)
- Jung et al.(2009)** Sin-Ho Jung, Jong-Hyeon Jeong e Hanna Bandos. Regression on quantile residual life. *Biometrics*, 65:1203–1212. Citado na pág. [32](#), [33](#), [34](#)
- Kalbfleisch e Prentice(2002)** J. D. Kalbfleisch e R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Citado na pág. [3](#), [7](#)
- Kaplan e Meier(1958)** E. L. Kaplan e Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481. Citado na pág. [2](#), [3](#)
- Karlin e Taylor(1975)** Samuel Karlin e Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press. Citado na pág. [18](#)
- Kim et al.(2012)** Mi-Ok Kim, Mai Zhou e Jong-Hyeon Jeong. Censored quantile regression for residual lifetimes. *Lifetime Data Analysis*, 18:177–194. Citado na pág. [34](#), [35](#)
- Klein e Moeschberger(2003)** John P. Klein e Melvin L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Springer. Citado na pág. [6](#)
- Ma e Wei(2012)** Yanyuan Ma e Ying Wei. Analysis on censored quantile residual life model via spline smoothing. *Statistica Sinica*, 22:47–68. Citado na pág. [36](#)
- Ma e Yin(2010)** Yanyuan Ma e Guosheng Yin. Semiparametric median residual life model and inference. *Canadian Journal of Statistics*, 38:665–679. Citado na pág. [35](#), [36](#)
- Maguluri e Zhang(1994)** Gangaji Maguluri e Cun-Hui Zhang. Estimation in the mean residual life regression model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:477–489. Citado na pág. [18](#), [19](#), [20](#)
- Maller e Zhou(1996)** Ross Aa. Maller e Xian Zhou. *Survival Analysis with Long-Term Survivors*. Citado na pág. [47](#)
- Mantel(1966)** Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170. Citado na pág. [4](#)
- Nelson(1972)** Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–965. Citado na pág. [3](#)
- Oakes e Dasu(1990)** David Oakes e Tamraparni Dasu. A note on residual life. *Biometrika*, 77:409–410. Citado na pág. [16](#)
- Oakes e Dasu(2003)** David Oakes e Tamraparni Dasu. Inference for the proportional mean residual life model. *The Institute of Mathematical Statistics Lecture Notes-Monograph Series*, páginas 105–116. Citado na pág. [17](#)

- Owen(2001)** Art Owen. *Empirical Likelihood*. London: Chapman & Hall. Citado na pág. [29](#)
- Rebolledo(1980)** Rolando Rebolledo. Central limit theorems for local martingales. *Z. Wahrscheinlichkeitstheorie verw Gebiete*, 51:269–286. Citado na pág. [27](#)
- Sun e Zhang(2009)** Liuquan Sun e Zhigang Zhang. A class of transformed mean residual life models with censored survival data. *Journal of the American Statistical Association*, 104:803–815. Citado na pág. [22](#), [23](#), [24](#)
- Zhou e Jeong(2011)** Mai Zhou e Jong-Hyeon Jeong. Empirical likelihood ratio test for median and mean residual lifetime. *Statistics in Medicine*, 30:152–159. Citado na pág. [30](#), [31](#)