



UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA APLICADA
E ESTATÍSTICA
MESTRADO EM MATEMÁTICA APLICADA E ESTATÍSTICA



Resíduos em Modelos de Sobrevivência com Fração de Cura

Cleanderson Romualdo Fidelis

Natal-RN
Maio 2019

Cleanderson Romualdo Fidelis

Resíduos em Modelos de Sobrevida com Fração de Cura

Trabalho apresentado ao Programa de Pós-Graduação em Matemática Aplicada e Estatística da Universidade Federal do Rio Grande do Norte, em cumprimento com as exigências legais para obtenção do título de Mestre.

Área de Concentração: Probabilidade e Estatística.

Linha de Pesquisa: Modelos de Regressão.

Orientadora

Prof^a Dra. Dione Maria Valença

Co-orientadora

Prof^a Dra. Mariana Correia de Araújo

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE – UFRN
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA APLICADA E ESTATÍSTICA –
PPGMAE

Natal-RN

Maio 2019

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Ronaldo Xavier de Arruda - CCET

Fidelis, Cleanderson Romualdo.

Resíduos em modelos de sobrevivência com fração de cura /
Cleanderson Romualdo Fidelis. - 2019.
76f.: il.

Dissertação (Mestrado) - Universidade Federal do Rio Grande
do Norte, Centro de Ciências Exatas e da Terra, Programa de Pós-
Graduação em Matemática Aplicada e Estatística. Natal, 2019.

Orientadora: Dione Maria Valença.

Coorientadora: Mariana Correia de Araújo.

1. Estatística - Dissertação. 2. Dados de tempo de vida -
Dissertação. 3. Diagnóstico - Dissertação. 4. Fração de cura -
Dissertação. 5. Resíduos de Cox-Snell - Dissertação. I. Valença,
Dione Maria. II. Araújo, Mariana Correia de. III. Título.

RN/UF/CCET

CDU 519.2

Dissertação de Mestrado sob o título *Resíduos em Modelos de Sobrevivência com Fração de Cura* apresentada por Cleanderson Romualdo Fidelis e aceita pelo Programa de Pós-Graduação em Matemática Aplicada e Estatística da Universidade Federal do Rio Grande do Norte, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Prof. Dra. Dione Maria Valença
Orientadora
Departamento de Estatística
Universidade Federal do Rio Grande do Norte - UFRN

Prof. Dra. Mariana Correia de Araújo
Co-orientadora
Departamento de Estatística
Universidade Federal do Rio Grande do Norte - UFRN

Prof. Dr. Bruno Monte de Castro
Departamento de Estatística
Universidade Federal do Rio Grande do Norte - UFRN

Prof. Dr. Diego Ignacio Gallardo Mateluna
Departamento de Matemáticas - Facultad de Ingeniería
Universidad de Atacama, Copiapó - Chile

Natal-RN, 31 de maio de 2019.

Ao Único, Eterno e Invisível Deus triúno, a minha família, minha querida Aline; a todos os membros da Primeira Evangélica Igreja Congregacional de Natal, Segunda Igreja Evangélica Congregacional de Esperança e a todos os amados amigos e irmãos em Cristo Jesus.

Agradecimentos

Ao Senhor Deus, revelado na pessoa Bendita de Seu Filho Jesus Cristo, por ter me feito seu filho e concedido sua maravilhosa graça há mais de 13 anos.

Agradeço especialmente a minha orientadora Prof. Dione, por todo esforço e dedicação, sem os quais este trabalho não teria chegado ao fim. A sua paciência e dedicação são virtudes que demonstram a ótima profissional e pessoa que és. Agradeço também a co-orientadora Prof. Mariana, por todo esforço, correções e paciência.

À minha amada família, especialmente a minha querida vó, Dona Maria de Lourdes por todo esforço e carinho todos esses anos, a minha mãe Merenice Fidelis por toda dedicação e preocupação, a minha tia Angélica Weber por todo incentivo e conselhos que foram de grande ajuda em todos os momentos complicados.

Ao meu amor, Aline Silva, por toda paciência enquanto estive sem paciência, todo suporte e ajuda na maior parte deste trabalho. Seu amor, seu cuidado, preocupação, foram combustíveis para eu poder prosseguir e chegar até aqui. Você é uma visível manifestação da graça de Deus em minha vida.

A família Santiago por todo acolhimento e carinho por todos esses anos, em especial a Gisele Amarante, Gilrenice Santiago (Tia Nice), Gilvonete Santiago (Tia Nete), Daniela Santiago, Fellipe Amarante, Thiago Santiago, Filipe Santiago, Ana Noemi, por toda paciência e crescimento durante todos esses anos. As famílias Carvalho e Soares, por toda felicidade proporcionada, Roseli de Carvalho, Pr. Antonio de Carvalho, Damaris, Susana e Anthony por tudo, assim como também a minha mãezona Inagi de Lira e ao Pr. Alcides de Lira, assim como a todos os seus filhos.

Aos meus amigos e amigas, Davi Cardoso, Rodolfo Santos, Pedro Batista, Arlindelson Caitano, Fernando Felipe, Daniel Félix e a todos os outros por tudo durante esses anos. Ao casal mais descolado que conheço Tiago e Thaís Felix, vocês moram no meu coração. Somente o céu irá pagar tudo que vocês fizeram e fazem por mim.

Aos meus amados pastores Weber Firmino e família e ao pastor Marcone Borba e família. Vocês são baluartes que eu me inspiro e realmente espero poder ser a metade

do que vocês são quando eu crescer. Deus em Cristo retribua cada um pelo esforço e preocupação.

A todos amigos que compõe a CRU Natal. Vocês foram um canal de edificação diária, aproximando-me do Senhor Jesus e me proporcionando grandes momentos de comunhão, treta, risadas e felicidade.

A todos que acreditaram em mim e fizeram/fazem parte desta vida maravilhosa que Deus me proporcionou, muito obrigado a todos.

Aos professores do PPGMAE, por toda paciência e ensinamentos.

Aos participantes da banca pelas contribuições.

À CAPES pelo apoio financeiro.

"Porque dele e por ele, e para ele, são todas as coisas; glória, pois, a ele eternamente.

Amém."

Romanos 11:36

Resíduos em Modelos de Sobrevivência com Fração de Cura

Autor: Cleanderson Romualdo Fidelis

Orientador: Prof. Dra. Dione Maria Valença

RESUMO

Modelos de sobrevivência com fração de cura representam uma generalização dos modelos de sobrevivência clássicos que permitem a ocorrência de uma fração de indivíduos curados na população estudada. Dois modelos paramétricos de sobrevivência com fração de cura bastante conhecidos são os modelos de mistura padrão e modelo de tempo de promoção. Após o ajuste de um modelo aos dados é importante a validação dos pressupostos. A análise de resíduos é uma importante técnica com este propósito. Embora análise de resíduos seja amplamente utilizada para avaliar o ajuste de modelos de sobrevivência usuais, existem poucos estudos sobre o uso de resíduos para dados de sobrevivência com fração de cura. Este trabalho tem como objetivo estudar uma recente proposta da literatura que avalia, através de resíduos tipo Cox-Snell, o ajuste global em modelos de mistura padrão e estender esta proposta para modelos de tempo de promoção. Consideramos um estudo de simulação para avaliar o desempenho dos resíduos. Uma aplicação é considerada para ilustrar os procedimentos estudados.

Palavras-chave: Dados de tempo de vida. Diagnóstico. Fração de cura. Resíduos de Cox-Snell.

Residuals in Survival Models with cure rate

Author: Cleanderson Romualdo Fidelis

Advisor: Prof. Dra. Dione Maria Valença

ABSTRACT

Cure fraction survival models represent a generalization of classical survival models which allow the occurrence of a fraction of cured individuals in the population studied. Two well-known parametric cure fraction survival models are the mixture cure models and the promotion time models. After the fitting a model to a dataset it is important validate the assumptions and the residuals analysis is an important technique for this purpose. Although residuals analysis is widely used to assess the fit of usual survival models, there are few residuals proposed for survival data with cure fraction. This work aims to study a recent proposal of the literature that evaluates, through Cox-Snell type residues, the global adjustment in mixture cure models and extends this proposal to promotion time model. We considered a simulation study to evaluate the performance of the proposed residuals. An application is considered to illustrate the studied procedures.

Keywords: Cox-Snell residuals. Cure rate. Diagnostic. Lifetime data.

Lista de figuras

1	Curvas de sobrevivências estimadas via Kaplan-Meier para dados hipotéticos.	p. 18
2	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Weibull sem covariáveis para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 38
3	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Weibull com covariáveis para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 39
4	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Log-Normal sem covariáveis (má especificação na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 40
5	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Log-Normal com covariáveis (má especificação na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 41
6	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de tempo de promoção Weibull sem covariáveis (má especificação na incidência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 42
7	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de tempo de promoção Weibull com covariáveis (má especificação na incidência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 43

8	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo tempo de promoção Log-Normal sem covariáveis (má especificação na incidência e na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 44
9	Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo tempo de promoção Log-Normal com covariáveis (má especificação na incidência e na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).	p. 45
10	Curva de sobrevivência estimadas via Kaplan-Meier para dados ovarian.	p. 46
11	Resíduos do tipo Cox-Snell para os ajustes dos dados ovarian.	p. 48

Sumário

1	Introdução	p. 13
1.1	Objetivos	p. 15
1.2	Organização do trabalho	p. 16
2	Análise de Sobrevivência com Fração de Cura	p. 17
2.1	Conceitos básicos	p. 18
2.1.1	Modelos paramétricos	p. 19
2.1.1.1	Distribuição de Weibull	p. 19
2.1.1.2	Distribuição log-normal	p. 20
2.1.2	Modelos de tempo de falha acelerado	p. 20
2.1.3	Representação dos dados de sobrevivência	p. 21
2.1.4	Resíduos de Cox-Snell	p. 22
2.2	Modelo de mistura padrão	p. 23
2.3	Modelo de tempo de promoção	p. 24
2.4	Modelo unificado	p. 24
2.4.1	Função de sobrevivência de longa duração	p. 25
2.4.2	Verossimilhança para o modelo unificado	p. 26
3	Resíduos em modelos de Sobrevivência com fração de cura	p. 29
3.1	Resíduo do tipo Cox-Snell para o Modelo de Mistura Padrão	p. 30
3.2	Resíduo do tipo Cox-Snell para o Modelo de Tempo de Promoção	p. 32
4	Estudos de Simulação	p. 35

4.1	Geração dos dados e obtenção das estimativas	p. 35
4.2	Ajuste dos dados e obtenção dos resíduos	p. 36
4.3	Resultados	p. 37
5	Aplicação	p. 46
5.1	Ajuste do modelo de regressão aos dados	p. 47
6	Considerações finais	p. 49
	Referências	p. 51
	Apêndice A – Obtenção da Função de Verossimilhança	p. 54
	Verossimilhança para os dados completos	p. 54
	Verossimilhança Marginal	p. 57
	Apêndice B – Aspectos Computacionais	p. 59
	Apêndice C – Códigos da aplicação	p. 67

1 Introdução

De acordo com Collett (2003), o termo análise de sobrevivência é utilizado para descrever a análise de dados envolvendo o tempo até a ocorrência de um evento, tendo este o tempo inicial e tempo final bem definidos. A presença de observações censuradas, a qual caracteriza-se por ser uma observação parcial da resposta, é uma característica fundamental neste tipo de dados. Segundo Colosimo e Giolo (2006), a análise de sobrevivência tem despertado interesse nas últimas décadas em função da evolução das técnicas estatísticas envolvidas e dos avanços computacionais.

Uma maneira de analisar dados que são provindos de estudos que possuem como variável resposta o tempo até a ocorrência de um evento (ou simplesmente “tempo de vida”) é utilizar a abordagem paramétrica. De acordo com Lawless (2003) destacam-se a utilização dos modelos probabilísticos exponencial, Weibull, log-normal, log-logístico, gama e gama generalizado, como principais modelos, devido a sua vasta utilização na literatura. Geralmente assume-se como análise inicial o pressuposto de homogeneidade nos tempos até a ocorrência do evento. Contudo, em muitas situações práticas, os dados estão associados com informações adicionais de maneira que os tempos até a ocorrência do evento dependem de alguns fatores (MALLER; ZHOU, 1996). Uma forma de mensurar essas informações adicionais é a inclusão de covariáveis através de um modelo de regressão que trate das relações envolvidas entre as covariáveis e o tempo de vida, rompendo assim o pressuposto da homogeneidade, uma vez que cada tempo depende de fatores/covariáveis. Em estudos da área médica, por exemplo, essas covariáveis podem surgir de delineamentos com o intuito de comparar algum tratamento ou informações como idade, sexo, peso de pacientes, entre outras.

Embora haja na literatura uma vasta utilização dos modelos de sobrevivência usuais, uma vez que estes carregam a suposição de que todos os indivíduos sofrerão o evento de interesse em algum momento, há situações que esses indivíduos não serão acometidos pelo evento de interesse mesmo que o tempo de acompanhamento seja suficientemente grande. Estes indivíduos são chamados “curados” ou “imunes” ao evento de interesse. Modelos que

tratam de dados deste tipo são denominados Modelos com Fração de Cura.

Modelos com fração de cura foram estudado inicialmente por Boag (1949) e Berkson e Gage (1952), os quais consideraram para incorporar indivíduos curados uma mistura de distribuições em um modelo que ficou conhecido como modelo de mistura padrão. Este modelo tem por característica assumir que uma parte da população está curada e uma outra está suscetível ao evento em estudo. Posteriormente, Yakovlev et al. (1993) propõem um modelo, estendido por Chen, Ibrahim e Sinha (1999), conhecido como modelo de tempo de promoção. Este tem por característica principal uma estrutura de riscos competitivos associada à suposição Poisson para a distribuição de uma variável latente representando o número de causas para a ocorrência do evento. Rodrigues et al. (2009) propõem uma abordagem geral que unifica os modelos relatados acima a partir da flexibilização da distribuição da variável aleatória representando o número de causas competindo para a ocorrência do evento (chamada **distribuição incidência**) e da distribuição associada ao tempo de vida dos indivíduos suscetíveis (chamada **distribuição de latência**). Distribuições de latência consideradas com frequência são a Weibull e log-normal, entre outras. Distribuição de incidência comumente consideradas são Bernoulli (modelo de mistura padrão), Poisson (modelo de tempo de promoção) e geométrica.

Os modelos de sobrevivência com fração de cura representam uma generalização do modelos de sobrevivência clássicos, uma vez que os modelos clássicos não tratam a eventual ocorrência de uma fração de curados na população estudada. Como exemplo de aplicação dos modelos de sobrevivência com fração de cura temos Araújo et al. (2004) que usam o modelo de mistura padrão para estudar os tempos até a ocorrência da reincidência de crimes praticados por ex-detentos e estimar a proporção dentre eles que não reincidiriam, ou seja, não voltariam a praticar crimes após serem libertados. Sposto (2002) utilizam modelos de mistura padrão paramétricos para analisar alguns conjuntos de dados oriundos de ensaios clínicos com grupos de crianças acometidas por câncer comparando os resultados com os obtidos com o ajuste do modelo de regressão de Cox. Tournoud e Ecochard (2007) ajustaram modelos de tempo de promoção a dois conjuntos de dados, sendo que o primeiro conjunto de dados diz respeito a um estudo realizado na África e Sul da Ásia que examinou a influência da amamentação de crianças no tempo de atraso até o diagnóstico de HIV do tipo 1 nestas crianças. Ajustando também um modelo de tempo de promoção, Varshney et al. (2018) buscaram estimar a proporção de curados dentre uma amostra de pacientes portadores de HIV/AIDS que se submeteram a um tratamento com terapia anti-retroviral.

Após o ajuste dos modelos aos dados, faz-se necessário uma validação dos pressupostos assumidos. A validação desses pressupostos acarreta em fortes indícios de que o problema está sendo abordado corretamente. A análise de resíduos é comumente utilizada para validar o ajuste de modelos de sobrevivência clássicos. Entre os resíduos mais usados citamos os resíduos padronizados, de Cox-Snell, martingale, deviance, entre outros (ver Lawless (2003) e Collett (2003)). Os resíduos de Cox-Snell são muito utilizados em análise de sobrevivência. Eles são baseados no método geral para a formulação de resíduos proposto originalmente em Cox e Snell (1968). O uso destes resíduos para modelos de sobrevivência com base em propriedades de uma transformação da função de sobrevivência ($-\log(S(t))$) foi proposto por Kalbfleisch (1974), Crowley e Hu (1977) e Lagakos (1979).

Existem poucos estudos sobre resíduos para dados de sobrevivência com fração de cura. Entre estes podemos citar Peng e Taylor (2017) que propõem um novo resíduo tipo martingale modificado para avaliar o ajuste com respeito à distribuição de latência de um modelo de mistura padrão de riscos proporcionais com base na verossimilhança parcial de Cox (COX, 1972). Scolas et al. (2018) estendem o uso dos resíduos tipo Cox-Snell em modelos de mistura padrão paramétricos com censura à direita e censura intervalar, para avaliar globalmente o ajuste da distribuição de latência e incidência do modelo. Estes autores também propõem um resíduo para do tipo Cox-Snell para avaliar apenas a distribuição de latência e uma proposta de resíduo chamado deviance para avaliar a suposição de linearidade na parte latente do modelo.

1.1 Objetivos

Considerando os modelos de sobrevivência com fração de cura em um contexto de censura à direita e com estrutura de regressão associada apenas à distribuição de incidência, este trabalho tem como objetivo principal propor e avaliar os resíduos do tipo Cox-Snell global para modelos de tempo de promoção, estendendo assim os resultados encontrados em Scolas et al. (2018).

Objetivos específicos:

- Descrever os resíduos do tipo Cox-Snell global proposto por Scolas et al. (2018) para o modelo de mistura padrão;
- Desenvolver os resíduos do tipo Cox-Snell global para modelos de tempo de promoção;

- Fazer um estudo de simulação avaliando conjuntamente a capacidade de detecção de violações em pressupostos com respeito às distribuições de incidência e/ou latência dos resíduos propostos assim como a possibilidade de falso alarme;
- Aplicar estes resíduos em um conjunto de dados reais.

1.2 Organização do trabalho

Este trabalho está dividido em seis capítulos. No Capítulo 2 são introduzidos os conceitos de Análise de Sobrevida com fração de cura, com ênfase nos modelos de mistura padrão e nos modelos de tempo de promoção. No Capítulo 3 são estudadas a metodologia de resíduos propostos por Scolas et al. (2018) para os modelos de mistura padrão e também a nossa proposta de resíduos para o modelo de tempo de promoção. No Capítulo 4 são apresentados resultados numéricos envolvendo os resíduos que foram estudados e propostos no Capítulo 3. No Capítulo 5 é apresentada uma aplicação que utiliza os resíduos descritos no Capítulo 3 para avaliar ajustes de modelos. Por fim, são apresentadas conclusões e possíveis pesquisas futuras envolvendo o que foi apresentado neste trabalho.

2 Análise de Sobrevivência com Fração de Cura

Quando tratamos da análise de dados sobre o tempo até a ocorrência de um evento de interesse é razoável supor que todos os indivíduos ou equipamentos, que são os objetos do estudo, sofrerão o evento quando o tempo de acompanhamento for “suficientemente grande”. Este é um pressuposto dos modelos de sobrevivência usuais (LAWLESS, 2003) que admitem, contudo, a ocorrência de informações incompletas (censuras) pelo fim do estudo ou pela perda do acompanhamento de alguns indivíduos. Essa suposição acompanha os modelos clássicos de análise de sobrevivência e tem sido utilizada por diversos pesquisadores ao longo dos anos. Porém, existem situações em que alguns, dentre o grupo estudado, podem não estar sujeitos ao evento de interesse mesmo que sejam acompanhados por um longo tempo. Por exemplo, na área da criminalística, suponha que o objetivo seja estudar o tempo até que um ex-presidiário sofra recidiva (seja preso novamente). Pode ocorrer que muitos dos que são ex-presidiários não sejam presos novamente, mesmo acompanhados por toda a vida. Nas áreas médicas, em casos que o objetivo do estudo é analisar o tempo até a recidiva de uma doença, alguns indivíduos podem estar curados. Indivíduos “curados” ou “imunes” não serão acometidos com o evento de interesse mesmo que o tempo de acompanhamento seja suficientemente grande.

A presença de um alto percentual de censura no fim do estudo remete a um forte indício da ocorrência de curados em um estudo. Aliado à avaliação do alto percentual de censura, deve-se atentar também para a necessidade de observar se o tempo de acompanhamento foi “suficientemente grande” para que tenha-se indicativos plausíveis da existência de curados na amostra estudada (MALLER; ZHOU, 1996).

A presença da fração de cura em dados de sobrevivência pode ser percebida a partir do gráfico das estimativas da função de sobrevivência pelo método de Kaplan-Meier (KAPLAN; MEIER, 1958). Uma cauda que estaciona acima de zero dá indícios da presença de curados no estudo. Como exemplo, considere a Figura 1. Esta apresenta a função de sobrevivência

estimada para dois grupos de indivíduos. Percebe-se que um dos grupos de pacientes (grupo A) tem uma menor quantidade de censura no fim do estudo, aquele cujo a curva mais aproxima-se de zero. Para o outro grupo (grupo B) a curva de sobrevivência se estabiliza próximo ao valor de 0,6, tendo em vista a presença de um alto percentual de censura no fim do estudo. Esta representação gráfica do grupo B fornece evidências de que há imunes nos dados.

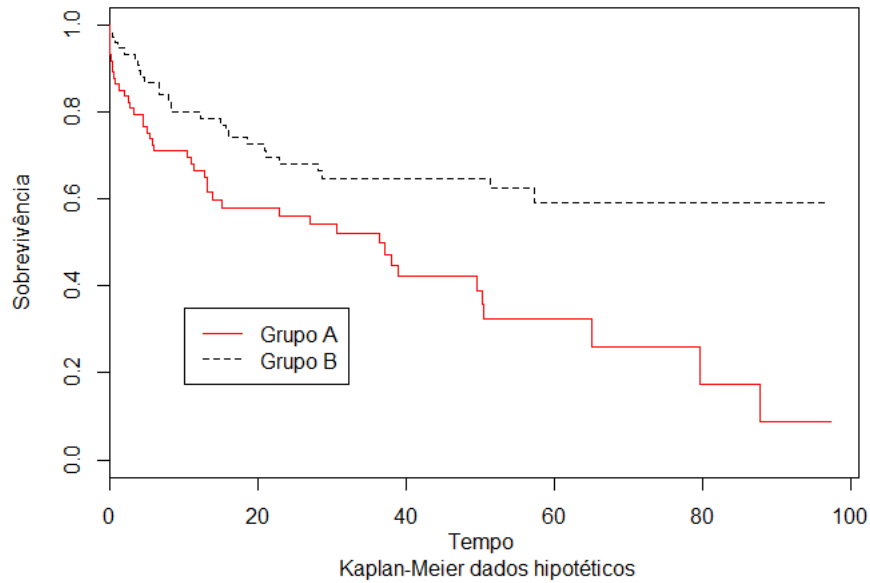


Figura 1: Curvas de sobrevivências estimadas via Kaplan-Meier para dados hipotéticos.

2.1 Conceitos básicos

Antes de apresentar o modelo de sobrevivência com fração de cura, faz-se necessário definir algumas funções e apresentar alguns conceitos importantes para a formulação desse modelo.

Seja T uma variável aleatória contínua e não negativa que representa o tempo até a ocorrência de um determinado evento, caracterizada por sua função de distribuição $F(t)$ e sua função densidade de probabilidade $f(t)$.

Definição 1 A função de sobrevivência da variável aleatória T é definida como

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx, t \in \mathbb{R}^+.$$

Esta função representa a probabilidade de um indivíduo não falhar antes do tempo t , ou seja, o indivíduo sobreviver pelo menos t unidades de tempo. Note que a função de sobrevivência é monótona decrescente, sendo $S(0) = 1$ e própria, uma vez que $\lim_{t \rightarrow \infty} S(t) = 0$.

Definição 2 A função risco de T , ou a função de taxa de falha é definida por

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

Esta por sua vez representa a taxa de falha instantânea no tempo t dado que o indivíduo não falhou até este tempo e pode ser definida também como o risco de falha instantânea de um indivíduo.

Definição 3 A função de risco acumulado é definida como

$$H(t) = \int_0^t h(x) dx.$$

Algumas relações matemáticas importantes entre as funções densidade de probabilidade, função risco, função risco acumulado e função de sobrevivência estão relatadas abaixo por

$$\begin{aligned} f(t) &= -\frac{\partial S(t)}{\partial t}, \\ h(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt} (\log S(t)) \text{ e} \\ H(t) &= -\log S(t) \text{ e } S(t) = \exp\{-H(t)\}. \end{aligned} \tag{2.1}$$

2.1.1 Modelos paramétricos

2.1.1.1 Distribuição de Weibull

Seja T uma variável aleatória com distribuição Weibull(α, γ). Temos que a função densidade de probabilidade e de sobrevivência, respectivamente, são:

$$f_T(t; \alpha, \gamma) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}$$

$$S_T(t; \alpha, \gamma) = \exp \left\{ - \left(\frac{t}{\alpha} \right)^\gamma \right\},$$

sendo $\gamma > 0$ é o parâmetro de forma e $\alpha > 0$ o parâmetro de escala.

2.1.1.2 Distribuição log-normal

Seja T uma variável aleatória com distribuição log-normal(μ, σ). Temos que a função densidade de probabilidade e de sobrevivência, respectivamente, são:

$$f_T(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi t\sigma}} \exp \left\{ -\frac{1}{2} \left(\frac{\log t - \mu}{\sigma} \right)^2 \right\}$$

e

$$S_T(t; \mu, \sigma) = \Phi \left(\frac{-\log t + \mu}{\sigma} \right)$$

sendo que $\sigma > 0$ é o parâmetro de escala, Φ é a função de distribuição de uma distribuição normal padronizada e $\mu \in \mathbb{R}$ é o parâmetro de forma da distribuição log-normal.

2.1.2 Modelos de tempo de falha acelerado

Seja T uma variável aleatória contínua e não negativa representando o tempo até a ocorrência de um evento. Considere os modelos caracterizados por $Y = \log T$, onde admite-se a seguinte relação:

$$Y = \log T = \mu + \sigma Z.$$

sendo μ o parâmetro de posição, σ parâmetro de escala e Z uma variável aleatória com função de densidade $f_0(z)$ e função de sobrevivência $S_0(z)$. Dizemos que Y pertence à família de posição e escala, com função densidade de probabilidade escrita da forma

$$f_Y(y; \mu, \sigma) = \frac{1}{\sigma} f_0 \left(\frac{y - \mu}{\sigma} \right), \quad -\infty < y < \infty, \quad (2.2)$$

e sua função de sobrevivência escrita por

$$S_Y(y; \mu, \sigma) = P(Y > y) = S_0 \left(\frac{y - \mu}{\sigma} \right), \quad -\infty < y < \infty, \quad (2.3)$$

com $\mu \in \mathbb{R}$ parâmetro de posição, $\sigma > 0$ parâmetro de escala, função densidade de probabilidade $f_0(\cdot)$ e função de sobrevivência $S_0(\cdot)$.

Se Y pertence à família de posição e escala, pode-se destacar duas propriedades importantes:

- a) $F(y) = P(Y \leq y) = F_0\left(\frac{y-\mu}{\sigma}\right) = 1 - S_0\left(\frac{y-\mu}{\sigma}\right),$
 b) $Y = \mu + \sigma Z.$

Se considerarmos a presença de $p - 1$ covariáveis $\mathbf{x} = (1, x_1, \dots, x_{p-1})^\top$ e um vetor de parâmetros desconhecidos caracterizado por $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^\top$ inseridos através do parâmetro de posição $\mu = \mu(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$ nas funções densidade de probabilidade e de sobrevivência, respectivamente, então

$$f_Y(y; \mu, \sigma) = \frac{1}{\sigma} f_0\left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma}\right) \quad -\infty < y < \infty \quad (2.4)$$

e

$$S_Y(y; \mu, \sigma) = P(Y > y) = S_0\left(\frac{\log y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma}\right) \quad -\infty < y < \infty. \quad (2.5)$$

Denominamos os modelos que atendem as características acima de **modelos de tempo de falha acelerado**, uma vez que as covariáveis inseridas no modelo através do parâmetro de posição têm a função de acelerar e desacelerar o tempo até a ocorrência do evento.

2.1.3 Representação dos dados de sobrevivência

Considerando uma amostra aleatória de tamanho n , seja T_i , uma variável aleatória representando o tempo de vida do indivíduo i com função de sobrevivência $S(t; \theta)$ e função de densidade $f(t; \theta)$ e C_i uma outra variável aleatória representando o tempo até a ocorrência de censura do indivíduo i , com $i = 1, 2, \dots, n$. Assumimos aqui censura à direita e o mecanismo de censura aleatória. Logo, o tempo observado até a ocorrência do evento para o indivíduo i será dado por

$$T_i^* = \min\{T_i, C_i\}.$$

Para indicar se o tempo observado foi o tempo até a falha ou tempo até a censura vamos utilizar a variável indicadora definida da seguinte forma

$$\delta_i = \begin{cases} 1, & T_i \leq C_i \\ 0, & \text{caso contrário.} \end{cases}$$

Se considerando um vetor de covariáveis x_i para cada indivíduo, a amostra observada de dados censurados é representada por (t_i^*, δ_i) , onde t_i^* é o valor observado de T^* para o i -ésimo indivíduo e associado a este tem-se a variável indicadora δ_i , sendo a indicadora de

falha. Métodos da máxima verossimilhança são comumente usados para obter a estimação $\hat{\theta}$ do vetor de parâmetros θ (ver Lawless (2003), Colosimo e Giolo (2006), entre outros).

2.1.4 Resíduos de Cox-Snell

Os resíduos de Cox-Snell são frequentemente utilizados na avaliação do ajuste de dados de sobrevivência, sendo assim denominados por derivarem de uma proposta mais geral de resíduos encontrada em Cox e Snell (1968). A proposta deste resíduo tem como base um resultado matemático sobre a distribuição da função de risco acumulada de uma variável aleatória conforme descrito a seguir.

Seja T uma variável aleatória contínua e $S(t)$ a sua respectiva função de sobrevivência. Tomando $S(T)$ como variável aleatória, é fácil mostrar que esta tem uma distribuição uniforme contínua em $[0, 1]$ e que a variável aleatória $Y = -\log S(T)$, tem distribuição exponencial de parâmetro 1 (exponencial padrão). Conseqüentemente, a função de sobrevivência de Y é dada por $P(Y > y) = \exp(-y)$ e, por (2.1), a função de risco acumulada $H(y) = y$.

O resíduo de Cox-Snell para o i -ésimo indivíduo é definido da seguinte forma (ver Collett (2003), por exemplo)

$$r_{CSi} = -\log(\hat{S}(t_i^*)). \quad (2.6)$$

em que $\hat{S} = S(\cdot; \hat{\theta})$ é a função de sobrevivência estimada e t_i^* é o valor observado de T^* . Logo, se o modelo é adequado, os resíduos observados (r_{CSi}, δ_i) com $i = 1, \dots, n$, devem se comportar como uma amostra censurada de uma distribuição exponencial padrão e tem por objetivo avaliar o ajuste aos dados.

Uma forma de avaliar graficamente se o ajuste do modelo é adequado, é a representação dos valores de r_{CSi} versus valores $\hat{H}(r_{CSi})$, sendo $\hat{H}(r_{CSi})$ as estimativas não paramétricas da função de risco acumulada dos resíduos observados (r_{CSi}, δ_i) , para $i = 1, \dots, n$, obtidas a partir do ajuste da função de sobrevivência pelo estimador de Kaplan-Meier. Se o modelo for corretamente especificado, os pontos do gráfico devem ficar em torno de uma reta com coeficiente angular 1 (COLOSIMO; GIOLO, 2006).

2.2 Modelo de mistura padrão

Os trabalhos iniciais que tratam dos modelos contendo frações de cura, denominados modelo de mistura padrão, são os trabalhos de Boag (1949) e Berkson e Gage (1952). Estes autores consideram uma mistura de distribuições, assumindo que existe uma fração (θ) da população que está curada, ou seja, que não sofrerá o evento de interesse, e uma outra parte dessa população ($1 - \theta$) que pode ser acometida do evento de interesse (ou censurada). Diversos pesquisadores estudaram extensões e aplicações destes modelos em várias áreas, tais como médica, financeira e industrial. Como exemplo podemos citar Farewell (1982), Ahmann et al. (1989), Kuk e (1992), Maller e Zhou (1996), Tamhane e Shi (2009), Yusuf e Bakar (2016), Scolas et al. (2016), entre outros.

Para o estudo deste modelo vamos considerar uma variável aleatória M com distribuição Bernoulli com parâmetro $(1 - \theta)$, com $\theta \in [0, 1]$, da seguinte forma:

$$M = \begin{cases} 1, & \text{se o indivíduo é suscetível ao evento} \\ 0, & \text{se o indivíduo é imune ao evento,} \end{cases}$$

sendo que $P_\theta(M = 0) = \theta$ representa a fração de curados na população e $P_\theta(M = 1) = 1 - \theta$ representa aqueles que são suscetíveis ao evento em estudo. Perceba que para cada indivíduo que assumir $M = 0$ resultará em $T = \infty$ o que implica $P(T > t | M = 0) = 1$. Já no caso dos indivíduos que assumirem $M = 1$, terão associado um tempo finito até a ocorrência do evento, seja esse censurado ou não. Sendo $P(T > t | M = 1) = S_0(t)$ a função de sobrevivência para o modelo de mistura padrão será dada por:

$$\begin{aligned} S_p(t) &= P(T > t) = P(T > t, M = 0) + P(T > t, M = 1) \\ &= P(T > t | M = 0)P_\theta(M = 0) + P(T > t | M = 1)P_\theta(M = 1) \\ &= \theta + (1 - \theta)S_0(t) \end{aligned} \tag{2.7}$$

em que θ representa a fração de cura induzida pelo modelo. As funções de sub-densidade e de sub-risco são dadas por:

$$f_p(t) = (1 - \theta)f(t) \quad \text{e} \quad h_p(t) = f(t) \frac{(1 - \theta)}{\theta + (1 - \theta)S_0(t)}.$$

2.3 Modelo de tempo de promoção

O modelo descrito acima tem como pressuposto que existe no máximo uma causa concorrendo para o tempo até a ocorrência do evento. Alternativamente, Yakovlev et al. (1993) e Chen, Ibrahim e Sinha (1999) propuseram um modelo mais geral, considerando a possibilidade da existência de várias causas competindo para a ocorrência de um evento de interesse em um indivíduo. Este modelo é referido em Rodrigues et al. (2009) como o modelo de tempo de promoção.

Seja M uma variável aleatória não observável com distribuição de $Poisson(\theta)$ que representa o número de causas ou fatores competindo para a ocorrência do evento de interesse. Definimos Z_1, Z_2, \dots, Z_M como variáveis aleatórias (latentes) representando os tempos até a ocorrência de um evento devido a cada uma das M causas. Quando $M = 0$, temos Z_0 de tal forma que $P(Z_0 = \infty) = 1$, representando o indivíduo que não é suscetível ao evento de interesse (pois quando não há causas competindo para a ocorrência do evento este indivíduo não sofrerá o evento de interesse).

O tempo observável até a ocorrência do evento é representado pela variável aleatória $T = \min\{Z_0, Z_1, \dots, Z_M\}$. Assumindo que as variáveis aleatórias Z_0, Z_1, \dots, Z_M são independentes e identicamente distribuídas com função de sobrevivência $S_0(\cdot)$, temos que a função de sobrevivência populacional é dada por

$$S_p(t) = \exp[-\theta(1 - S_0(t))],$$

em que $P_\theta(M = 0) = \exp(-\theta)$ é a fração de cura para o modelo de tempo de promoção.

As funções de sub-densidade e sub-risco são dadas por:

$$\begin{aligned} f_p(t) &= \frac{-d(S_p(t))}{dt} = \frac{-d[\exp(-\theta(1 - S(t)_0))]}{dt} \\ &= \theta f(t) \exp(-\theta F_0(t)) \end{aligned}$$

e

$$h_p(t) = \frac{f_p(t)}{S_p(t)} = \frac{\theta f_p(t) \exp(-\theta F_0(t))}{S_p(t)} = \theta f_0(t).$$

2.4 Modelo unificado

Proposto por Rodrigues et al. (2009), esta unificação teve como foco uma extensão dos modelos de longa duração propostos por Chen, Ibrahim e Sinha (1999). Como o próprio

nome sugere, o modelo unificado de longa duração tem como principal consequência a unificação de vários modelos, entre eles os modelos propostos Boag (1949), Berkson e Gage (1952) e Yakovlev et al. (1993), com a premissa de que todos podem ser descritos em torno desta única classe. Suponha que existam n indivíduos em um determinado estudo e que associado a estes indivíduos estão as seguintes variáveis aleatórias:

- $M \equiv$ número de causas ou fatores de risco competindo para a ocorrência do evento de interesse com distribuição de probabilidade $p_\theta(m) = P_\theta(M = m)$, sendo $\theta \in [0, 1]$ o parâmetro da distribuição.
- $Z_k \equiv$ tempo até à ocorrência do evento devido a k -ésima causa. Dado $M = m$ temos que Z_1, \dots, Z_m são variáveis aleatórias independentes e identicamente distribuídas com Z_k independente M , onde Z_k possuem função de distribuição comum caracterizada por $F_0(t) = 1 - S_0(t)$.
- $T \equiv$ tempo até a ocorrência do evento, definido como:

$$T = \min\{Z_0, Z_1, \dots, Z_M\},$$

com $P(Z_0 = \infty) = 1$, pois quando $M = 0$ não existem causas ou fatores competindo para a ocorrência do evento, ou seja, esta suposição permite um tempo de vida infinita para os imunes. As variáveis aleatórias Z_k e M são latentes, enquanto a variável aleatória T é observável.

2.4.1 Função de sobrevivência de longa duração

Para o modelo unificado, a função de sobrevivência de longa duração é dada por:

$$\begin{aligned} S_p(t) &= P(T > t) = P(T > t, M \geq 0) = \sum_{m=0}^{\infty} P(T > t, M = m) \\ &= \sum_{m=0}^{\infty} P(T > t | M = m) p_\theta(m) = P(T > t | M = 0) p_\theta(0) \\ &+ \sum_{m=1}^{\infty} P(T > t | M = m) p_\theta(m) \\ &= p_\theta(0) + \sum_{m=1}^{\infty} P(\min\{Z_1, Z_2, \dots, Z_m\} > t) p_\theta(m) \\ &= p_\theta(0) + \sum_{m=1}^{\infty} \frac{P(Z_1 > t, Z_2 > t, \dots, Z_m > t) p_\theta(m)}{p_\theta(m)} \\ &= p_\theta(0) + \sum_{m=1}^{\infty} P(Z_1 > t) \cdot \dots \cdot P(Z_m > t) p_\theta(m) = p_\theta(0) + \sum_{m=1}^{\infty} [S_0(t)]^m p_\theta(m), \end{aligned}$$

com $P(T > t | M = 0) = 1$, pois dado que não há nenhuma causa concorrendo para o evento, a probabilidade de o indivíduo sobreviver é 1. Considerando que a função de sobrevivência $S_0(t)$ é uma função de sobrevivência própria em que

$$\lim_{t \rightarrow \infty} S_0(t) = 0,$$

note que $S_p(t)$ é uma função de sobrevivência imprópria, uma vez que

$$\lim_{t \rightarrow \infty} S_p(t) = P_\theta(M = 0) = p_\theta(0) > 0,$$

em que $P_\theta(M = 0)$ representa a proporção de indivíduos curados. As funções densidade de probabilidade e de risco podem ser encontradas a partir de

$$f_p(t) = \frac{-dS_p(t)}{dt} = - \left[(-f_0(t)) \sum_{m=1}^{\infty} mp(m) S_0(t)^m \right]$$

e

$$h_p(t) = \frac{f_p(t)}{S_p(t)}.$$

Este modelo unifica os modelos de mistura padrão e de tempo de promoção (mais detalhes ver em Guedes (2011) e Rocha (2018)).

2.4.2 Verossimilhança para o modelo unificado

A partir de uma amostra obtida de tamanho n , desejamos fazer inferência sobre os parâmetros envolvidos no modelo unificado de longa duração. Neste caso, temos que definir algumas quantidades que possibilitem a realização da estimação desses parâmetros. Portanto, seja $Y_i = \min\{T_i, C_i\}$, que representa o tempo de sobrevivência para o indivíduo i , sendo que $T_i = \min\{Z_{i0}, Z_{i1}, \dots, Z_{iM_i}\}$ e C_i a variável aleatória que representa o tempo até a censura para o indivíduo i , sendo esta aleatória, não informativa e independente de T_i , $i = 1, \dots, n$ e δ_i a variável aleatória indicadora de censura quando $\delta_i = 0$ se $T_i > C_i$ e indicadora de falha quando $\delta_i = 1$ se $T_i \leq C_i$.

Dispomos das variáveis aleatórias latentes, M_i e Z_{ik} , $k = 1, \dots, M_i$ onde M_i representando o número de causas ou fatores que estão competindo para a ocorrência do evento no indivíduo i e Z_{ik} representando o tempo até a ocorrência do evento para o indivíduo i devido à k -ésima causa. Assumindo que $M_i = m_i$, temos que as variáveis aleatórias Z_{i1}, \dots, Z_{im_i} são *i.i.d.* (independentes e identicamente distribuídas), com função de sobrevivência dada por $S(\cdot | \boldsymbol{\xi})$ e função de distribuição acumulada dada por $F(\cdot | \boldsymbol{\xi}) = 1 - S(\cdot | \boldsymbol{\xi})$, sendo $\boldsymbol{\xi}$ o vetor de parâmetros relacionados à distribuição de Z_{ik} . Definimos também um

vetor de covariáveis $\mathbf{x}_i = (1, x_{i2}, \dots, x_{i(p-1)})^\top$ que estão relacionadas ao i -ésimo indivíduos. Denotamos os dados completos como $D_c = (n, \mathbf{Y}, \boldsymbol{\delta}, \mathbf{M}, \mathbf{X})$ sendo,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix} \quad \text{e} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}.$$

Outro fato que temos que lidar é como incluir as variáveis explicativas (covariáveis) no modelo de sobrevivência com fração de cura. Neste trabalho vamos considerar as covariáveis através da relação $\theta_i \equiv \theta(\mathbf{x}_i^\top \boldsymbol{\beta})$, em que θ_i é o parâmetro associado à distribuição de incidência para cada indivíduo $i = 1, \dots, n$.

Para o caso em que $M_i \sim \text{Bernoulli}(1 - \theta_i)$, ou seja, para o modelo de mistura padrão, vamos utilizar a função de ligação logística, dada por

$$\theta_i = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}. \quad (2.8)$$

Já para o caso em que $M_i \sim \text{Poisson}(\theta_i)$, ou seja, o modelo de tempo de promoção, a relação utilizada para incluir as covariáveis \mathbf{x}_i na distribuição de incidência do modelo é dada por

$$\theta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \quad (2.9)$$

e a fração de cura é dada por

$$p_{\theta_i}(0) = \exp(-\theta_i).$$

Neste trabalho consideraremos que as covariáveis influenciam apenas a distribuição de incidência, ou seja, considere que o parâmetro $\theta_i \equiv \theta_i(\mathbf{x}_i^\top \boldsymbol{\beta})$.

Seja $\boldsymbol{\lambda} = (\boldsymbol{\xi}^\top, \boldsymbol{\beta}^\top)^\top$ o vetor de parâmetros desconhecidos que queremos estimar. Conforme mostrado em Guedes (2011), a função de verossimilhança de $\boldsymbol{\lambda}$ que corresponde ao conjunto de dados completos D_c é dada por

$$L(\boldsymbol{\lambda}, D_c) = \prod_{i=1}^n [m_i f(y_i | \boldsymbol{\xi})]^{\delta_i} [S_0(y_i | \boldsymbol{\xi})]^{1 - \delta_i} p_{\theta_i}(m_i). \quad (2.10)$$

O logaritmo da função de verossimilhança fica dado por

$$\begin{aligned}\ell(\boldsymbol{\lambda}, \mathbf{D}_c) &= \log L(\boldsymbol{\lambda}, \mathbf{D}_c) \\ &= \sum_{i=1}^n \delta_i \log[m_i] + \sum_{i=1}^n \delta_i \log[f_0(y_i|\boldsymbol{\xi})] \\ &+ \sum_{i=1}^n (1 - \delta_i) \log(S_0(y_i|\boldsymbol{\xi})) + \sum_{i=1}^n \log(p_{\theta_i}(m_i)).\end{aligned}$$

Como \mathbf{M} é um vetor de variáveis latentes, não é possível estimar $\boldsymbol{\lambda}$ devido as funções de verossimilhança e log-verossimilhança dependerem de \mathbf{M} . Então, buscando contornar essa dificuldade, vamos utilizar a função de verossimilhança marginal, a qual é obtida a partir do somatório da distribuição conjunta nas variáveis M_i . Logo, a função de verossimilhança marginal para os dados observáveis denotado por $\mathbf{D} = (n, \mathbf{Y}, \boldsymbol{\delta}, \mathbf{X})$ será dada por

$$L^*(\boldsymbol{\lambda}; \mathbf{D}) = \prod_{i=1}^n [f_p(y_i|\boldsymbol{\xi}, \boldsymbol{\beta})]^{\delta_i} [S_p(y_i|\boldsymbol{\xi}, \boldsymbol{\beta})]^{1-\delta_i} \quad (2.11)$$

e o logaritmo da função de verossimilhança marginal será dado por

$$\ell(\boldsymbol{\lambda}, \mathbf{D}) = \sum_{i=1}^n \delta_i \log[f_p(y_i|\boldsymbol{\xi}, \boldsymbol{\beta})] + \sum_{i=1}^n (1 - \delta_i) \log[S_p(y_i|\boldsymbol{\xi}, \boldsymbol{\beta})]. \quad (2.12)$$

As demonstrações para obtenção das funções 2.10 e 2.11 estão dispostas no Apêndice A. As estimativas encontradas baseadas nas equações acima, em geral, são encontradas maximizando a (2.12) por um método numérico.

Um critério para avaliar os ajustes dos modelos aos dados é o AIC (*Akaike Criterion Information*, (AKAIKE, 1974)) definido como

$$AIC = -2\ell(\hat{\boldsymbol{\lambda}}, \mathbf{D}) + 2p, \quad (2.13)$$

sendo $\ell(\hat{\boldsymbol{\lambda}}, \mathbf{D})$ o logaritmo da função de verossimilhança definida acima e p a quantidade de parâmetros envolvido no modelo em estudo. Segundo Dirick, Claeskens e Baesens (2015) o valor resultante da estimação de AIC é o valor esperado da discrepância do modelo entre a informação de Kullback-Leibler (KULLBACK; LEIBLER, 1951) e o valor verdadeiro desconhecido do processo de geração dos dados, sem conhecer o verdadeiro valor deste processo. Para escolha deste modelo baseado neste critério serão preferíveis modelos que obtiverem os menores valores de AIC.

3 Resíduos em modelos de Sobrevivência com fração de cura

Segundo Cordeiro e Neto (2006), a análise de resíduos constitui uma etapa importante no processo de avaliação e seleção do modelo estatístico ajustado aos dados. No caso do ajuste de modelos de regressão envolvendo dados de sobrevivência com fração de cura (ver Capítulo 2), existe a necessidade de verificação das pressuposições e da adequação do modelo ajustado aos dados.

Nos modelos clássicos de análise de sobrevivência existem alguns resíduos para avaliar o ajuste do modelo, sendo o mais utilizado o resíduo de Cox-Snell (COX; SNELL, 1968), que tem como objetivo avaliar o ajuste global do modelo utilizado para analisar os dados.

Para o caso de se estudar resíduos para modelos envolvendo dados de sobrevivência com fração de cura, é necessário considerar que a população completa envolvida é composta de duas sub-populações: uma contendo a fração de curados e outra contendo aqueles que são suscetíveis ao evento em estudo. Devido às características contidas no modelo de sobrevivência com fração de cura, a função de sobrevivência associada a este não é uma sobrevivência própria e a proposta de resíduos para o caso de funções de sobrevivência própria não pode ser utilizada diretamente. Neste sentido, existem poucas propostas de resíduos para avaliar partes específicas do modelo de mistura padrão. Wileyto et al. (2013) propõem um resíduo para o modelo paramétrico de mistura padrão derivado do resíduo de Schoelfeld. Mais recentemente Peng e Taylor (2017) propuseram métodos para avaliar o ajuste de modelos de mistura padrão, investigar propriedades de resíduos existentes e propor novos resíduos para avaliar a parte latente do modelo. Scolas et al. (2018) propõem métodos de diagnósticos através de resíduos do tipo Cox-Snell para o modelo de mistura padrão com dados censurados à direita e para dados de censura intervalar.

Consideramos a seguir o estudo detalhado da proposta feita por Scolas et al. (2018), assim como a sua aplicação para os modelos de mistura padrão Weibull e modelo log-normal. Em seguida, é apresentada a extensão da proposta para o modelo de tempo

de promoção considerando as distribuições Weibull e log-normal para a distribuição de latência.

3.1 Resíduo do tipo Cox-Snell para o Modelo de Mistura Padrão

A literatura sobre modelos de mistura padrão é extensiva, não se limitando apenas as partes de extensões metodológicas, mas englobando também aplicação a problemas reais. Contudo, os estudos que focam na verificação do modelo ainda são pouco explorados (AMICO; KEILEGOM, 2018). De acordo com Peng e Taylor (2017), quando utiliza-se o modelo de mistura padrão, devido à distribuição teórica associada aos resíduos do tipo Cox-Snell ser do tipo mista, tais resíduos se comportam como uma amostra censurada de uma distribuição exponencial padrão no intervalo $[0, -\log(\theta))$ com $0 < \theta < 1$, fato este que pode ser visto a partir do que segue.

Conforme visto na Seção 2.2, seja Z uma variável aleatória contínua e não negativa que representa o tempo até a falha dos indivíduos suscetíveis com função de sobrevivência $S_0(z)$, M uma variável aleatória discreta com função de probabilidade $P_\theta(M = m) = \theta^m(1 - \theta)^{1-m}$. Podemos mostrar que a variável aleatória $S_0(Z) \sim U(0, 1)$, onde $U[0, 1]$ representa a uma distribuição Uniforme no intervalo $[0, 1]$. Ver Collett (2003), por exemplo.

Vimos que $S_p(t) = P(T > t) = \theta + (1 - \theta)S_0(t)$ é a função de sobrevivência de longa duração para o o modelo de mistura padrão mostrada em 2.7. Se vista como variável aleatória, a função de sobrevivência de longa duração $S_p(T) = \theta + (1 - \theta)S_0(T)$ terá distribuição associada conforme abaixo. Note que $S_p(t)$ é uma função de sobrevivência imprópria, compreendida no intervalo $\theta \leq S_p(t) \leq 1$. Sendo assim, tomando $x \in [\theta, 1]$, temos

$$\begin{aligned}
P(S_p(T) > x) &= P[\theta + (1 - \theta)S_0(T) > x] = P\left[S_0(T) > \left(\frac{x - \theta}{1 - \theta}\right)\right] \\
&= P\left[T < S_0^{-1}\left(\frac{x - \theta}{1 - \theta}\right)\right] = 1 - P\left[T > S_0^{-1}\left(\frac{x - \theta}{1 - \theta}\right)\right] \\
&= 1 - P\left[T > S_0^{-1}\left(\frac{x - \theta}{1 - \theta}\right) \middle| M = 1\right] P(M = 1) \\
&\quad - P\left[T > S_0^{-1}\left(\frac{x - \theta}{1 - \theta}\right) \middle| M = 0\right] P(M = 0) \\
&= 1 - P\left[Z > S_0^{-1}\left(\frac{x - \theta}{1 - \theta}\right)\right] (1 - \theta) - \theta \\
&= 1 - S_0\left[S_0^{-1}\left(\frac{x - \theta}{1 - \theta}\right)\right] (1 - \theta) - \theta \\
&= 1 - \left(\frac{x - \theta}{1 - \theta}\right) (1 - \theta) - \theta \\
&= 1 - x.
\end{aligned}$$

Logo, a distribuição da função de sobrevivência de longa duração vista como variável aleatória pode ser caracterizada por

$$\begin{aligned}
P(S_p(T) > x) &= \begin{cases} 1, & x < \theta \\ 1 - x, & \theta \leq x < 1 \\ 0, & x \geq 1 \end{cases} \Rightarrow \\
\Rightarrow P(S_p(T) \leq x) &= \begin{cases} 0, & x < \theta \\ x, & \theta \leq x < 1 \\ 1, & x \geq 1 \end{cases}
\end{aligned}$$

Além disso, se $v \in \mathbb{R}$, temos que

$$\begin{aligned}
P(-\log[S_p(T)] > v) &= P[S_p(T) < e^{-v}] \\
&= \begin{cases} 0, & e^{-v} < \theta \\ e^{-v}, & \theta \leq e^{-v} < 1 \\ 1, & e^{-v} \geq 1. \end{cases}
\end{aligned}$$

Sendo assim, temos que a função de distribuição associada a $H(T; \theta) = -\log(S_p(T))$ é

dado por

$$P(H > v) = \begin{cases} 0, & v \leq -\log(\theta) \\ e^{-v}, & 0 \leq v < -\log(\theta) \\ 1, & v < 0. \end{cases} \quad (3.1)$$

Note que H é uma variável aleatória do tipo mista cuja distribuição não segue uma distribuição exponencial padrão (RINCÓN, 2007). De fato, H tem distribuição do mínimo entre uma variável aleatória com distribuição exponencial de parâmetro 1 e a constante $-\log(\theta)$. Após o ajuste do modelo aos dados vamos utilizar a distribuição de referência de H a fim de que possamos avaliar este ajuste.

Os resíduos de Cox-Snell obtidos com base nos modelos clássicos (ver Seção 2.1.4) devem se comportar como uma amostra censurada de uma distribuição exponencial padrão quando o ajuste global é adequado. Para o nosso caso, se o modelo de mistura padrão adotado foi corretamente especificado, os resíduos de Cox-Snell definidos como

$$r_{CSi} = -\log(\hat{\theta}_i + (1 - \hat{\theta}_i)\hat{S}_0(y_i)), \quad i = 1, \dots, n, \quad (3.2)$$

devem se comportar como uma amostra censurada de uma distribuição exponencial padrão no intervalo $[0, -\log(\theta)]$.

3.2 Resíduo do tipo Cox-Snell para o Modelo de Tempo de Promoção

Análogo ao utilizado na Seção anterior para o que foi realizado para o modelo de mistura padrão, baseado nos resíduos propostos por Cox-Snell encontraremos uma distribuição adequada para comparar estes mesmos resíduos após o ajuste do modelo de tempo de promoção aos dados, de modo semelhante ao que acontece para os modelos clássicos.

A distribuição da função de sobrevivência de longa duração do modelo tempo de promoção $S_p(T) = \exp[-\theta(1 - S_0(T))] = \exp[-\theta(F_0(T))]$ vista como variável aleatória segue uma distribuição caracterizada por $G_\theta(\cdot)$, com função de distribuição da seguinte

forma: vamos tomar $x \in [e^{-\theta}, 1)$, então

$$\begin{aligned}
P[S_p(T) \leq x] &= P[e^{-\theta(F_0(T))} \leq x] = P[-\theta(F_0(T)) \leq -\log(x)] \\
&= P\left[F_0(T) > \frac{-\log(x)}{\theta}\right] = P\left[T > F_0^{-1}\left(\frac{-\log(x)}{\theta}\right)\right] \\
&= P\left[T > F_0^{-1}\left(\frac{-\log(x)}{\theta}\right) \mid M = 0\right] P_\theta[M = 0] \\
&\quad + \sum_{m=1}^{\infty} P\left[T > F_0^{-1}\left(\frac{-\log(x)}{\theta}\right) \mid M = m\right] P_\theta[M = m] \\
&= p_\theta(0) + \sum_{m=1}^{\infty} P\left[\min\{Z_1, \dots, Z_M\} > F_0^{-1}\left(\frac{-\log(x)}{\theta}\right) \mid M = m\right] p_\theta[m] \\
&= p_\theta(0) + \sum_{m=1}^{\infty} P\left[Z_1 > F_0^{-1}\left(\frac{-\log(x)}{\theta}\right)\right]^m p_\theta[m] \\
&= p_\theta(0) + \sum_{m=1}^{\infty} \left[S_0\left(F_0^{-1}\left(\frac{-\log(x)}{\theta}\right)\right)\right]^m p_\theta[m] \\
&= p_\theta(0) + \sum_{m=1}^{\infty} \left[1 - F_0\left(F_0^{-1}\left(\frac{-\log(x)}{\theta}\right)\right)\right]^m p_\theta[m] \\
&= p_\theta(0) + \sum_{m=1}^{\infty} \left[1 + \frac{\log(x)}{\theta}\right]^m p_\theta[m] = \sum_{m=0}^{\infty} \left[1 + \frac{\log(x)}{\theta}\right]^m p_\theta[m]. \quad (3.3)
\end{aligned}$$

Para a última parte da demonstração, utilizaremos as propriedades da distribuição de $M \sim Poisson(\theta)$. Neste caso, a função geradora de momentos de M é dada por

$$\mathcal{M}_M(t) = E[e^{tM}] = \sum_{m=0}^{\infty} e^{tm} p_\theta(m) = \exp[\theta(e^t - 1)]. \quad (3.4)$$

Além disso, podemos utilizar a função geradora de probabilidade para mostrar a seguinte relação:

$$E(a^M) = \mathcal{M}_M(\log(a)), \quad (3.5)$$

e com base em 3.4,

$$E(a^M) = \exp(\theta(a - 1)).$$

Assim, usando estes resultados em (3.3), podemos associar o resultado acima para auxiliar-nos, resultando em:

$$\begin{aligned}
P(S_p(T) \geq x) &= E\left[\left(1 + \frac{\log(a)}{\theta}\right)^M\right] \exp\left[\theta\left(1 + \frac{\log(x)}{\theta} - 1\right)\right] \\
&= \exp\left[\theta\left(\frac{\log(x)}{\theta}\right)\right] = \exp[\log(x)] = x.
\end{aligned}$$

A função de distribuição dos resíduos do tipo Cox-Snell visto como variável aleatória, uma vez que a relação assumida é dada por $H(T, \theta) = -\log(S_p(T))$, segue uma distribuição $F_H(T)$. Conforme dado a seguir pois se $v \in \mathbb{R}$, temos que

$$\begin{aligned} P[-\log(S_p(T)) > v] &= P[\log(S_p(T)) \leq -v] = P[S_p(T) \leq e^{-v}] \\ &= \begin{cases} 0, & e^{-v} < e^{-\theta} \\ e^{-v}, & e^{-\theta} \leq e^{-v} < 1 \\ 1, & e^{-v} \geq 1. \end{cases} \end{aligned} \tag{3.6}$$

Assim, temos

$$P[-\log(S_p(T)) > v] = \begin{cases} 1, & v \leq 0 \\ e^{-v}, & 0 < v \leq \theta \\ 0, & v > \theta. \end{cases}$$

Semelhante ao encontrado por Peng e Taylor (2017) e Scolas et al. (2018) quando estudaram a distribuição associada a $H(T; \theta)$ para o caso do modelo de mistura padrão, a distribuição associada a esta definição de $H(T; \theta)$ para o caso do modelo de tempo de promoção segue uma distribuição mista. Sendo que ainda que a distribuição associada a $H(T; \theta)$ não seja exponencial padrão, como no caso dos modelos clássicos terá comportamento exponencial padrão no intervalo $[0, \theta)$. Assim, é possível utilizar para avaliação do ajuste do modelo de tempo de promoção o resíduo tipo Cox-Snell dado por

$$r_{CS1i} = -\log \left[\exp(-\hat{\theta}_i(1 - \hat{S}_0(y_i))) \right] = \hat{\theta}_i(1 - \hat{S}_0(y_i)) \quad i = 1, \dots, n. \tag{3.7}$$

Se o ajuste estiver globalmente adequado, estes resíduos em (3.2) seguem o comportamento de uma amostra censurada de uma distribuição exponencial padrão no intervalo $[0, \theta)$.

4 Estudos de Simulação

Neste Capítulo vamos ilustrar, através de simulações, o desempenho de resíduos do tipo Cox-Snell no diagnóstico do ajuste de modelos de sobrevivência com fração de cura para avaliar capacidade de revelar violações com respeito à distribuição de incidência e/ou de latência assim como a influência do tamanho de amostra e da proporção de censura/fração de cura no diagnóstico.

Usamos para isso a proposta de resíduos do tipo Cox-Snell global dada em Scolas et al. (2018) para modelos de mistura padrão, assim como a extensão desta proposta para o modelo de tempo de promoção desenvolvida neste trabalho (Seção 3). As simulações foram realizadas no software R Core Team (2018), com base em uma adaptação do procedimento descrito em Loose (2016) e com o uso do pacote `flexcure` (Silva e Valença (2015), Silva (2015)). Os códigos são apresentados no Apêndice B.

4.1 Geração dos dados e obtenção das estimativas

Em todos os cenários os dados foram gerado com base em um modelo de mistura padrão Weibull. Foram considerados modelos sem covariáveis e com covariáveis apenas na distribuição de incidência. No caso com covariáveis, foram consideradas 3 covariáveis: x_1 , com os valores gerados de uma distribuição normal com média 0,5 e variância 1,5; x_2 , com valores gerados de uma distribuição uniforme no intervalo $[-3, 3]$ e x_3 , com valores gerados de uma distribuição Bernoulli com parâmetro 0,5. Os tamanhos amostrais assumidos foram $n = 100, 250$ e 500 . Foram gerados valores da variável M com base na distribuição de Bernoulli com parâmetro $(1 - \boldsymbol{\theta}(\mathbf{x}^\top \boldsymbol{\beta}))$, sendo $\mathbf{x}^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ e considerando para $\boldsymbol{\theta}(\mathbf{x}^\top \boldsymbol{\beta})$ a função de ligação dada em (2.8). A fração de cura desejada foi obtida com a escolha adequado do valor de β_0 (ver Tabela 1). Os valores fixados para os demais coeficientes foram $\beta_1 = 0,5$, $\beta_2 = -1$ e $\beta_3 = 1$.

Para $i = 1, \dots, n$, foram geradas censuras aleatórias c_i de uma distribuição uniforme

$U(0, \tau)$, em que a escolha do valor de τ afeta o percentual de censura desejada na amostra. Para cada indivíduo não imune ($M_i = 1$) foi gerado um tempo t_i de uma distribuição Weibull(α, γ) considerando $\alpha = 2$ e $\gamma = 0,5$ e os tempos observados y_i foram obtidos tomando $y_i = \min\{t_i, c_i\}$. Para cada indivíduo curado ($M_i = 0$), consideramos $y_i = \max\{c_1, \dots, c_n\}$. Associado a cada tempo observado foi gerado o indicador de falha, $\delta_i = 1$ se $y_i = t_i$ e $\delta_i = 0$ caso contrário (ou seja, se $t_i > c_i$, ou se o indivíduo é curado). Para definir a proporção de censura nos dados gerados é possível considerar duas quantidades: a proporção de censuras entre os indivíduos não curados (denotada por pc_1) e proporção de censurados e curados com respeito a todas as unidades em estudo (denotado por pc_2).

De acordo com Fonseca, Valença e Bolfarine (2013), se p_0 denota a proporção de curados, a relação entre estas quantidades é $pc_2 = pc_1(1 - p_0) + p_0$. Embora na prática seja possível observar apenas pc_2 , consideramos aqui pc_1 como a proporção de censura. Nesta simulação foram considerados três níveis (leve, moderado e baixo) de censuras/fração de cura, sendo estes níveis caracterizados pelas combinações dos percentuais de censura (pc_1) e de fração de cura (p_0), conforme descrito na Tabela 1 para o caso sem covariáveis.

Níveis de Censura	Percentuais de Censura	Percentuais de Fração de Cura
Leve	20%	10% ($\beta_0 = -2, 05$)
Moderado	30%	20% ($\beta_0 = -1, 36$)
Alto	40%	30% ($\beta_0 = -0, 9$)

Tabela 1: Percentuais de censura e fração de cura utilizadas para gerar os dados obtidos.

4.2 Ajuste dos dados e obtenção dos resíduos

Para ajustar cada amostra gerada nos casos com covariáveis e sem covariáveis foram considerados quatro modelos: i) o modelo de mistura padrão Weibull (o modelo correto), ii) o modelo de mistura padrão log-normal (ajusta de com má especificação na distribuição de latência), iii) o modelo de tempo de promoção Weibull (ajuste com má especificação na distribuição de incidência) e iv) o modelo de tempo de promoção log-normal, que representa um ajuste com má especificação tanto na distribuição de incidência quanto também na distribuição de latência. Note que ao ajustar modelos de mistura padrão utilizamos as estimativas geradas para obter os resíduos globais do tipo Cox-Snell dado em (3.2), que representa a proposta dada em Scolas et al. (2018). Quando ajustamos modelos de tempo de promoção, calculamos os resíduos globais com base em (3.7) utilizando, desta forma, os resíduos do tipo Cox-Snell propostos neste trabalho.

Para cada cenário, foram geradas 10.000 réplicas de Monte Carlo no caso sem covariáveis e no caso da geração dos dados com covariáveis foram utilizadas 2.000 réplicas.

4.3 Resultados

Foram realizadas inicialmente simulações com respeito à estimação dos parâmetros através dos dados gerados com a partir de alguns modelos escolhidos para validar nossos códigos. Para verificar o ajuste do modelo aos dados e avaliar o comportamento dos resíduos globais do tipo Cox-Snell foram considerados gráficos que incluem no eixo x os valores dos resíduos globais do tipo Cox-Snell e no eixo y uma estimativa não paramétrica da função risco acumulada com base nestes resíduos. Assim, se o ajuste for adequado, espera-se obter pontos em torno da reta com coeficiente angular 1 e intercepto 0 (ver Seção 2.1.4).

Considerando que cada uma das amostras geradas (10.000 no caso sem covariáveis e 2.000 com covariáveis) produzem um gráfico, apresentamos, de forma semelhante a Scolas et al. (2018), os pontos sobrepostos das réplicas em um único gráfico para cada cenário. Por exemplo, para cenários sem covariáveis e com $n = 100$ foram representados 10.000×100 pontos em cada gráfico.

A Figura 2 apresenta o ajuste do modelo de mistura padrão sem covariáveis para os três tamanhos de amostra e níveis de censura. Perceba que se tomarmos qualquer dos níveis de censura, para os tamanhos de amostra $n = 100$ temos uma dispersão maior dos pontos sobre a reta, mas a medida que aumentou-se o tamanho da amostra percebe-se melhorias na qualidade do ajuste, ou seja, a medida que dentro do mesmo nível de censura o tamanho da amostra aumenta, diminui a dispersão dos pontos sobre a reta, uma vez que esta figura representa o ajuste correto do modelo de mistura padrão Weibull sem covariáveis. Este gráfico pode ser uma opção de escolha do modelo ajustado para os dados quando não temos covariáveis no modelo e também servir de uma análise inicial do ajuste do modelo para um determinado conjunto de dados.

No caso da Figura 3, que apresenta o ajuste do modelo de mistura padrão com covariáveis na latência, deve ser analisada de forma similar a Figura 2, sendo que para esta as covariáveis inserem uma heterogeneidade para cada indivíduo i , com $i = 1, \dots, n$. Com isso, temos que cada gráfico para os níveis de censura e tamanhos de amostras apresentam uma dispersão maior do que no caso sem covariáveis. Vamos considerar o nível de censura moderado, por exemplo. Neste caso, como nos outros dois casos, percebeu-se que a medida

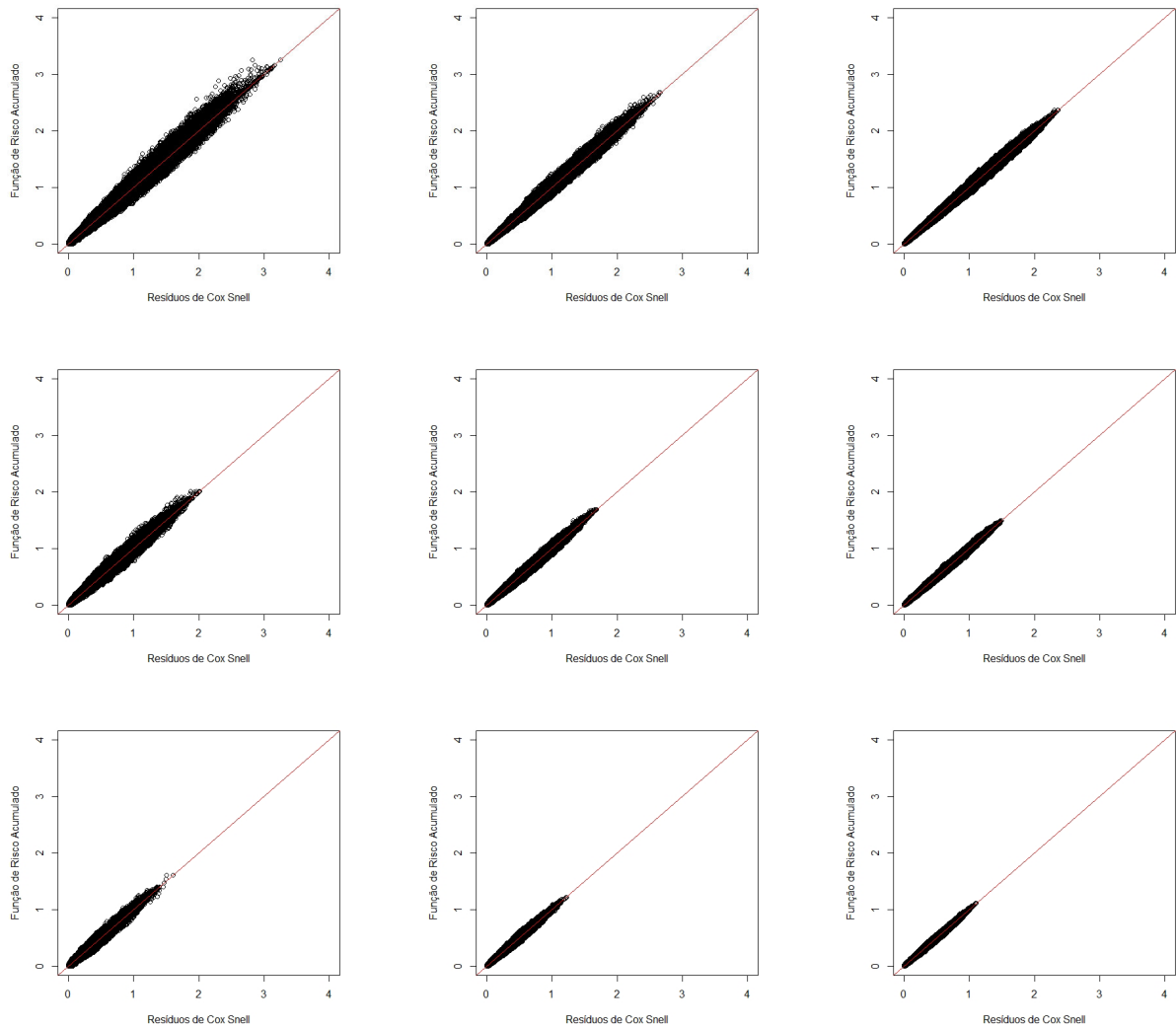


Figura 2: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Weibull sem covariáveis para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

que aumentou-se o tamanho da amostra a dispersão dos dados diminuiu, demonstrando assim que mesmo em tamanhos de amostra menores (neste caso $n = 100$) possa ser encontrada uma dispersão maior e uma impressão de um ajuste mal especificado. A medida que aumenta-se o tamanho da amostra, tal erro pode ser corrigido, chegando a indicar a especificação correta no ajuste do modelo.

As Figuras 4 e 5 dizem respeito ao ajuste do modelo de mistura padrão log-normal no caso sem covariáveis e com covariáveis, respectivamente. A Figura 4 apresenta uma clara indicação de violação no ajuste considerando o acentuado número de pontos acima da reta de coeficiente angular 1. Observa-se que a aparência de violação fica menos visível nos gráficos à medida que o nível de censura/fração de cura aumenta, sendo que um aumento

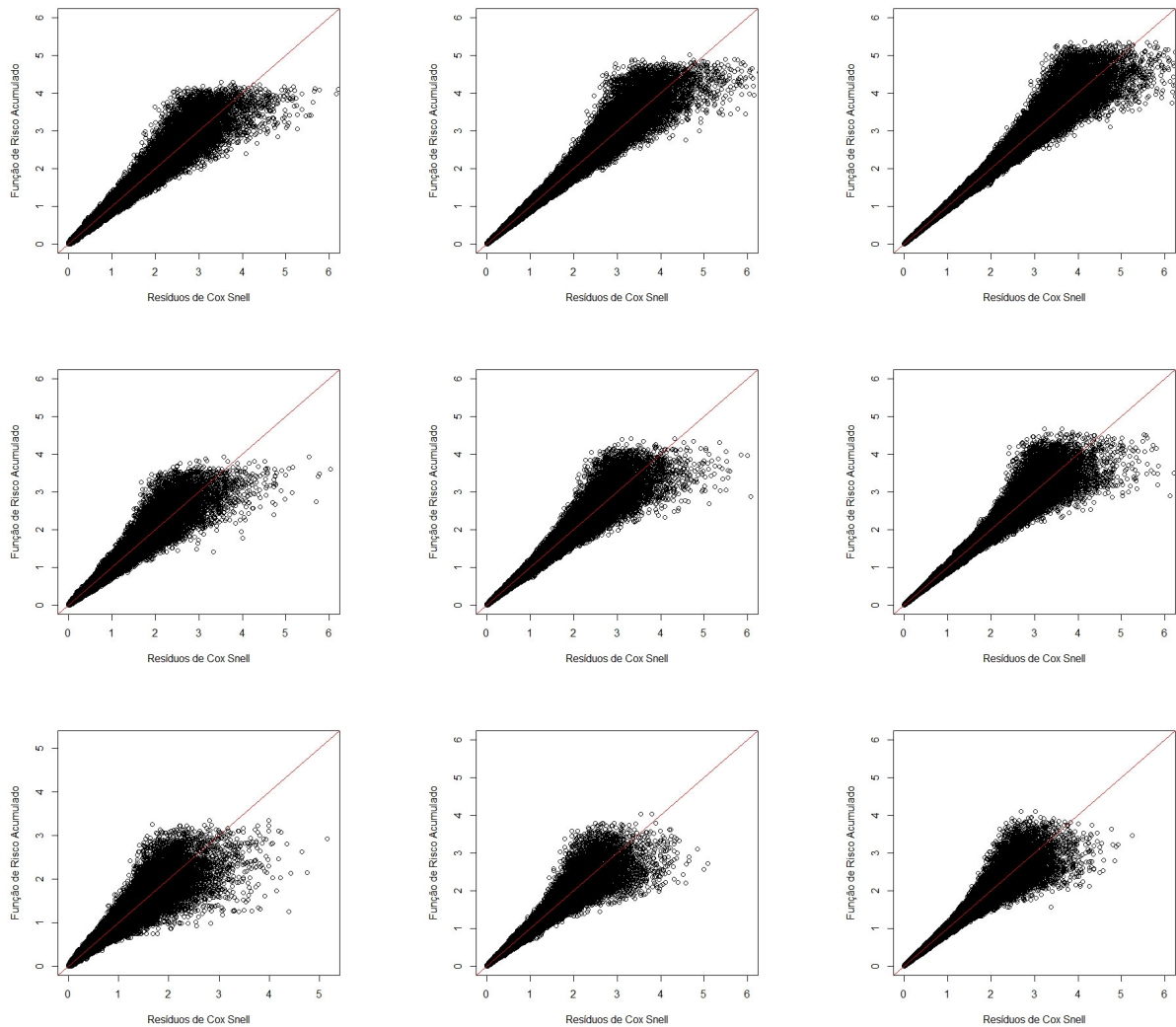


Figura 3: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Weibull com covariáveis para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

no percentual de censura pode levar a aparentes melhorias nos resultados obtidos, uma vez que acréscimo de informações incompletas (censuras) nos dados podem levar a essas conclusões. O comportamento que pode se esperar para este tipo de má especificação (na distribuição da latência) fica bem estabelecido com nível de censura leve para o maior tamanho de amostra ($n = 500$).

A Figura 5 apresenta o caso de um ajuste com má especificação na distribuição de latência na presença de covariáveis na distribuição de incidência. Como esperado, devido ao efeito causado pelas covariáveis temos que a dispersão nos pontos sobre os gráficos é bem maior se comparado à Figura 4 considerando o mesmo ajuste para o caso sem covariáveis. Comparado à Figura 3 pode-se observar um comportamento dos pontos com concentração

acima da reta de coeficiente angular 1. Vale ressaltar que a capacidade do gráfico de sinalizar violação é afetada pelo seu tamanho de amostra e nível de censura. Enquanto o aumento no tamanho de amostra parece confirmar o comportamento e violação, o aumento da censura/fração de cura o torna menos visível. Por exemplo, para o nível de censura alta, temos que se observarmos os três tamanhos de amostra ($n = 100, 250$ e 500), e compararmos com os 3 gráficos equivalentes na Figura 3 poderia-se inferir que o modelo ajustado aos dados é satisfatório, embora com o aumento no tamanho da amostra, a concentração dos pontos acima da curva de inclinação torna-se um pouco mais evidente.

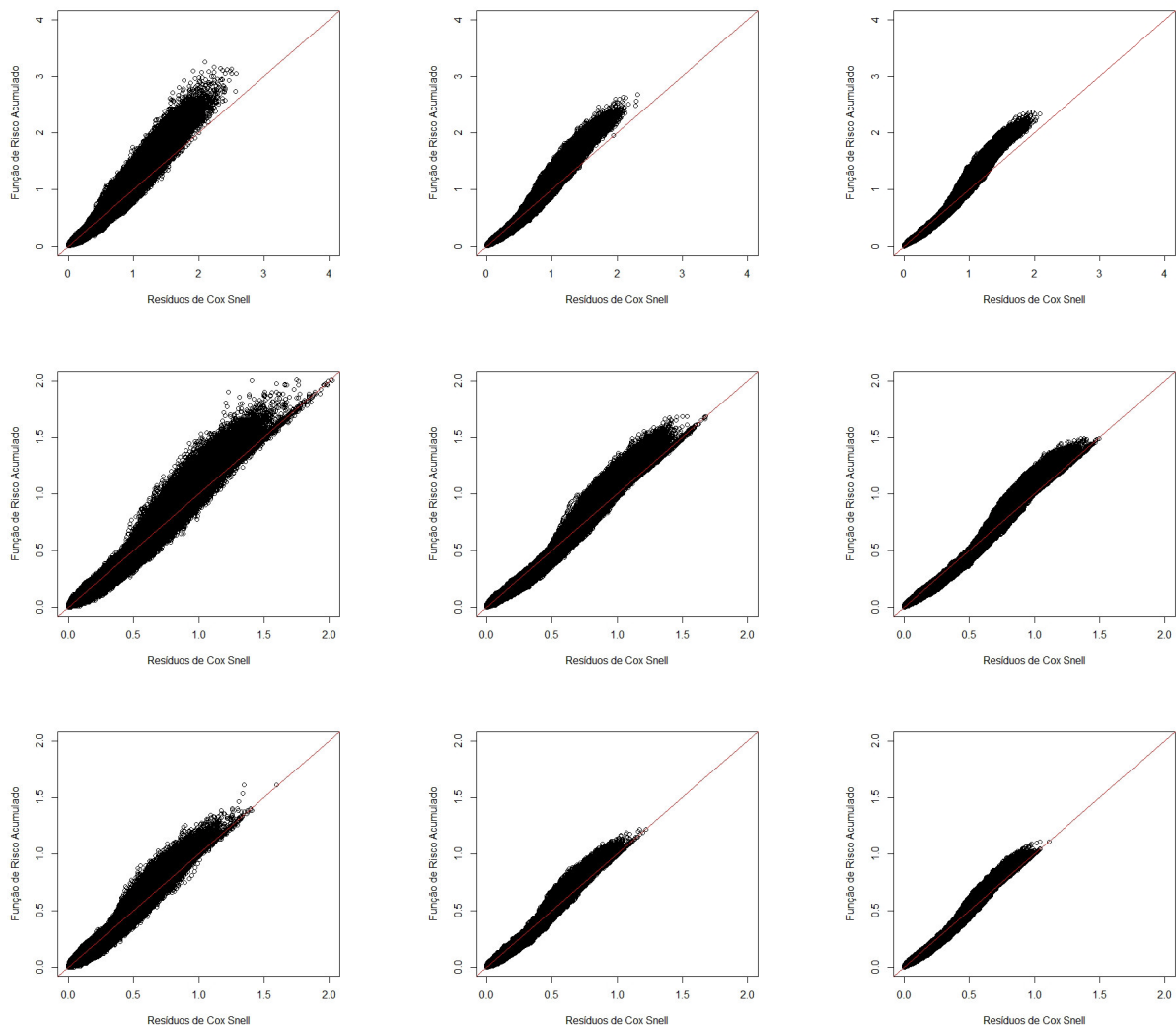


Figura 4: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Log-Normal sem covariáveis (má especificação na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

As Figuras 6 e 7 apresentam o ajuste do modelo de tempo de promoção Weibull

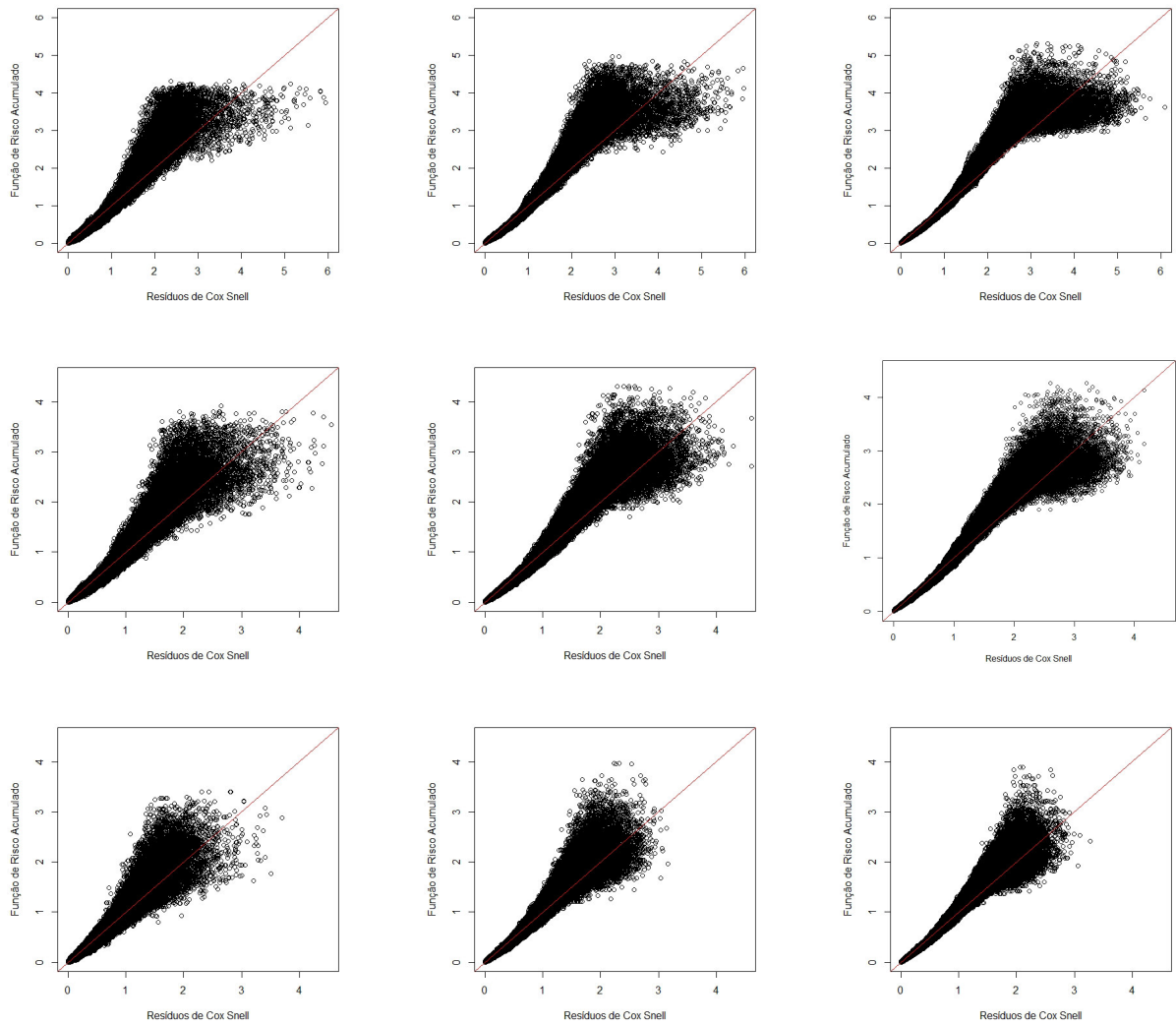


Figura 5: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de mistura padrão Log-Normal com covariáveis (má especificação na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

para os dados gerados nos casos sem covariáveis e com covariáveis. A Figura 6 apresenta um comportamento resultante da má especificação na distribuição de incidência em cada um dos gráficos. A forma resultante de cada gráfico, independente do nível de censura ou tamanho amostral, apresenta um comportamento bem diferente do caso onde não há má especificação no ajuste. A disposição nos pontos de cada gráfico das Figuras 6 e 7 demonstram a capacidade dos resíduos propostos neste trabalho (Seção 3.2) em captar a fuga do pressuposto assumido para a distribuição de incidência e latência.

As Figuras 8 e 9 apresentam o ajuste do modelo de tempo de promoção log-normal para os dados gerados nos casos sem covariáveis e com covariáveis, respectivamente. Os

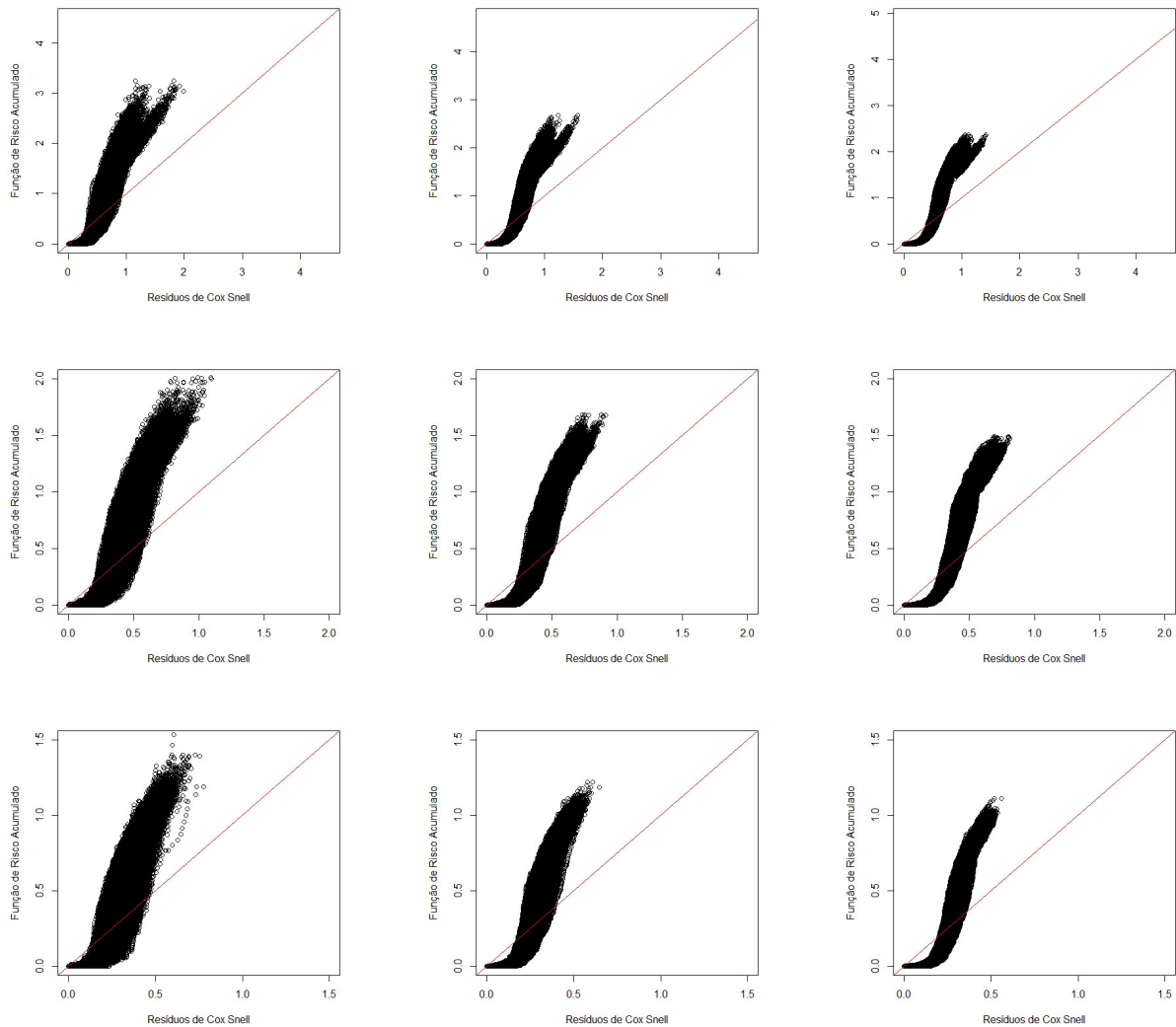


Figura 6: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de tempo de promoção Weibull sem covariáveis (má especificação na incidência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

ajustes apresentados na Figura 8 foram realizados considerando má especificação tanto na distribuição de incidência quanto também na distribuição de latência. O comportamento apresentado nos gráficos foi de pontos dispostos, em sua maioria, acima da reta de coeficiente de inclinação 1, para todos os tamanhos de amostra nos casos dos níveis de censura moderada e alta. Já no caso do nível de censura leve, os três gráficos para os tamanhos de amostra $n = 100$, 250 e 500 apresentam um comportamento com a maioria dos pontos se dispendo acima da reta de inclinação 1, mas também apresenta uma bifurcação, sendo esta mais acentuada no tamanho de amostra $n = 100$ e sendo suavizada a medida que aumentou-se o tamanho da amostra, indicando-nos que ajustes com má especificação deste tipo podem comportar-se tanto com uma concentração de pontos acima da reta de

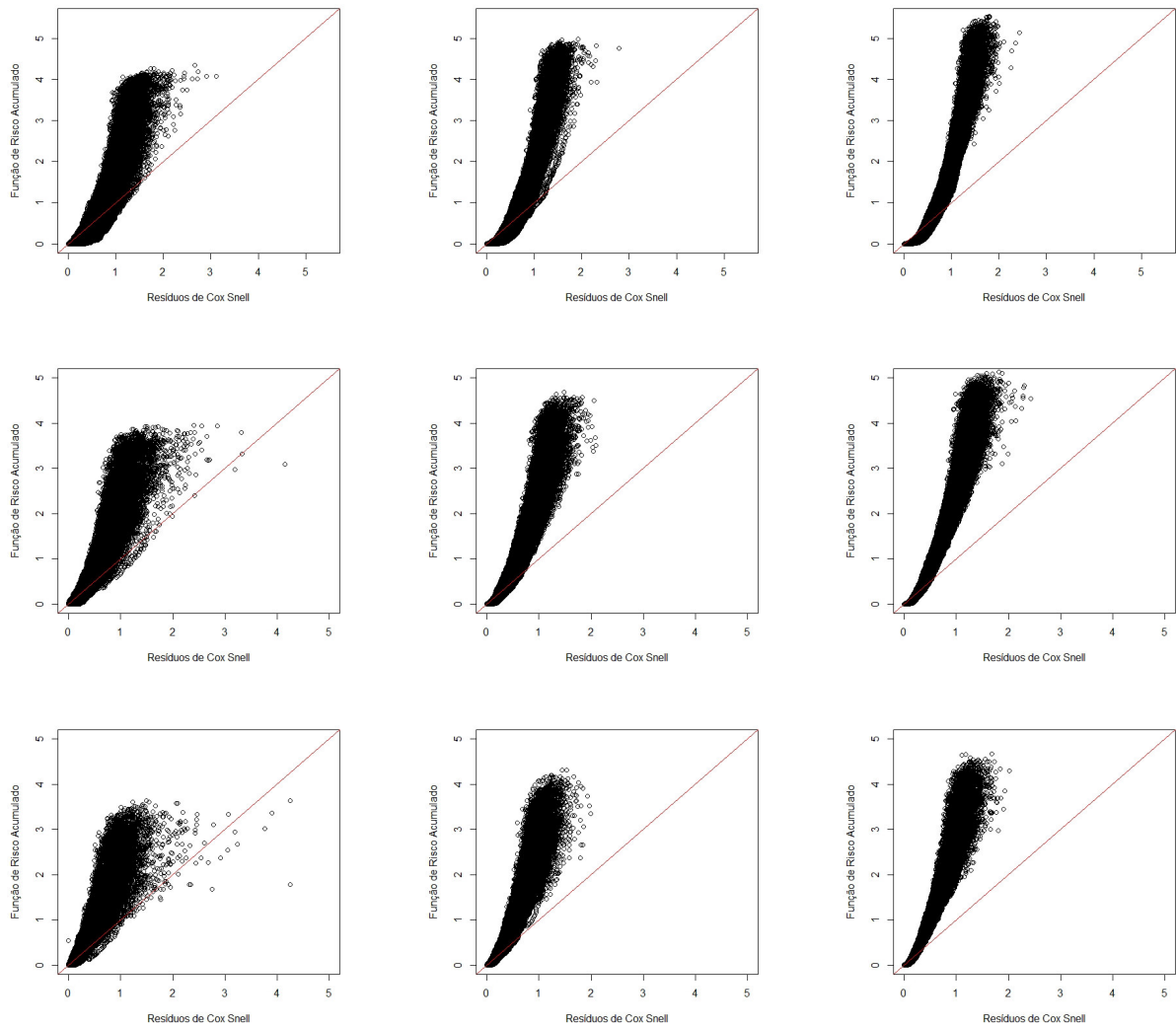


Figura 7: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo de tempo de promoção Weibull com covariáveis (má especificação na incidência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

inclinação 1 como também abaixo desta reta.

A Figura 9 apresenta os pontos dispostos nos gráficos de forma semelhante ao que ocorreu no cenário com má especificação apenas na latência (Figura 5). Para os tamanhos de amostra $n = 100$, independente dos níveis de censura, a má especificação nos ajustes ficou menos evidenciada que no caso quando aumentou-se o tamanho da amostra.

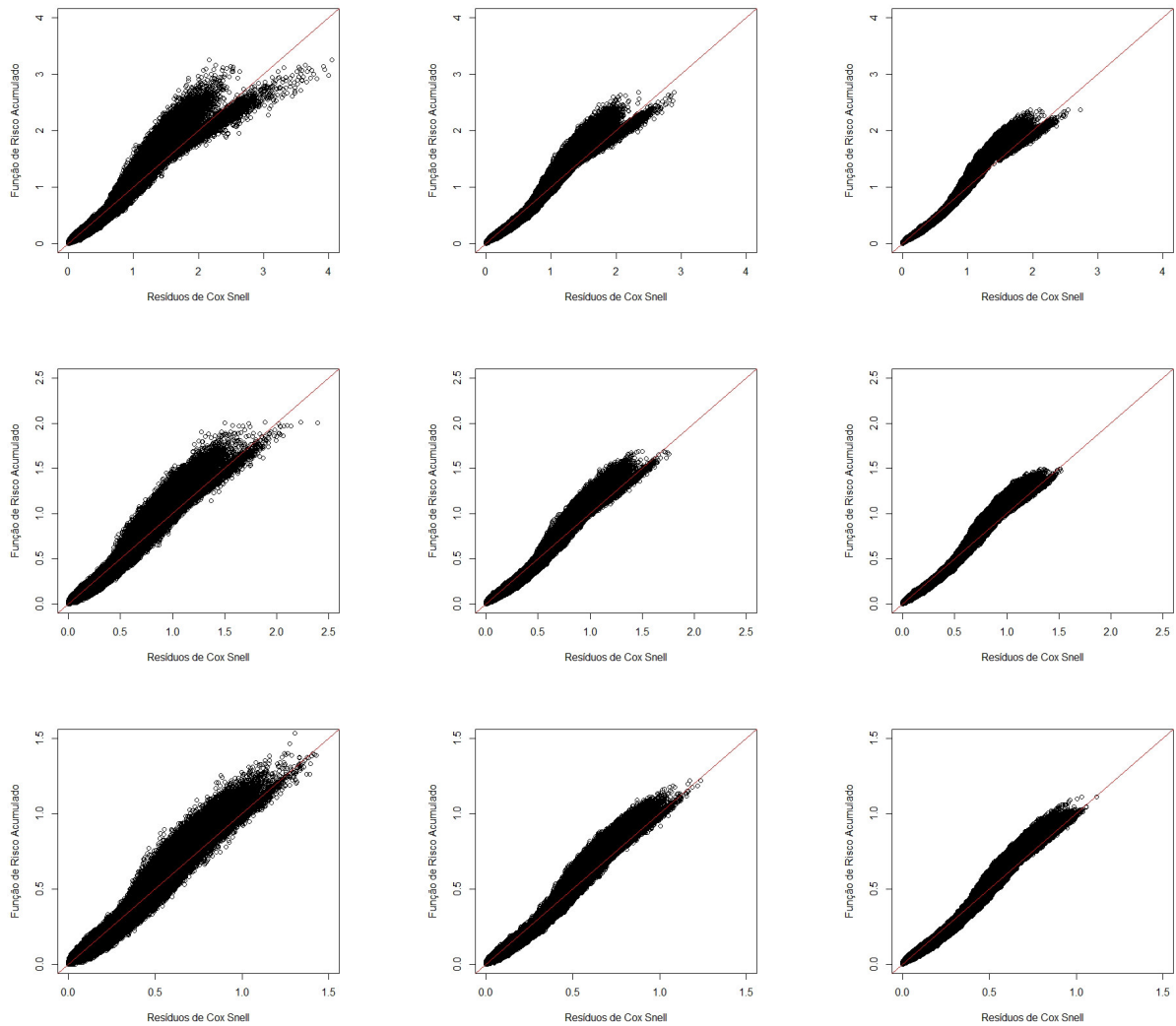


Figura 8: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo tempo de promoção Log-Normal sem covariáveis (má especificação na incidência e na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

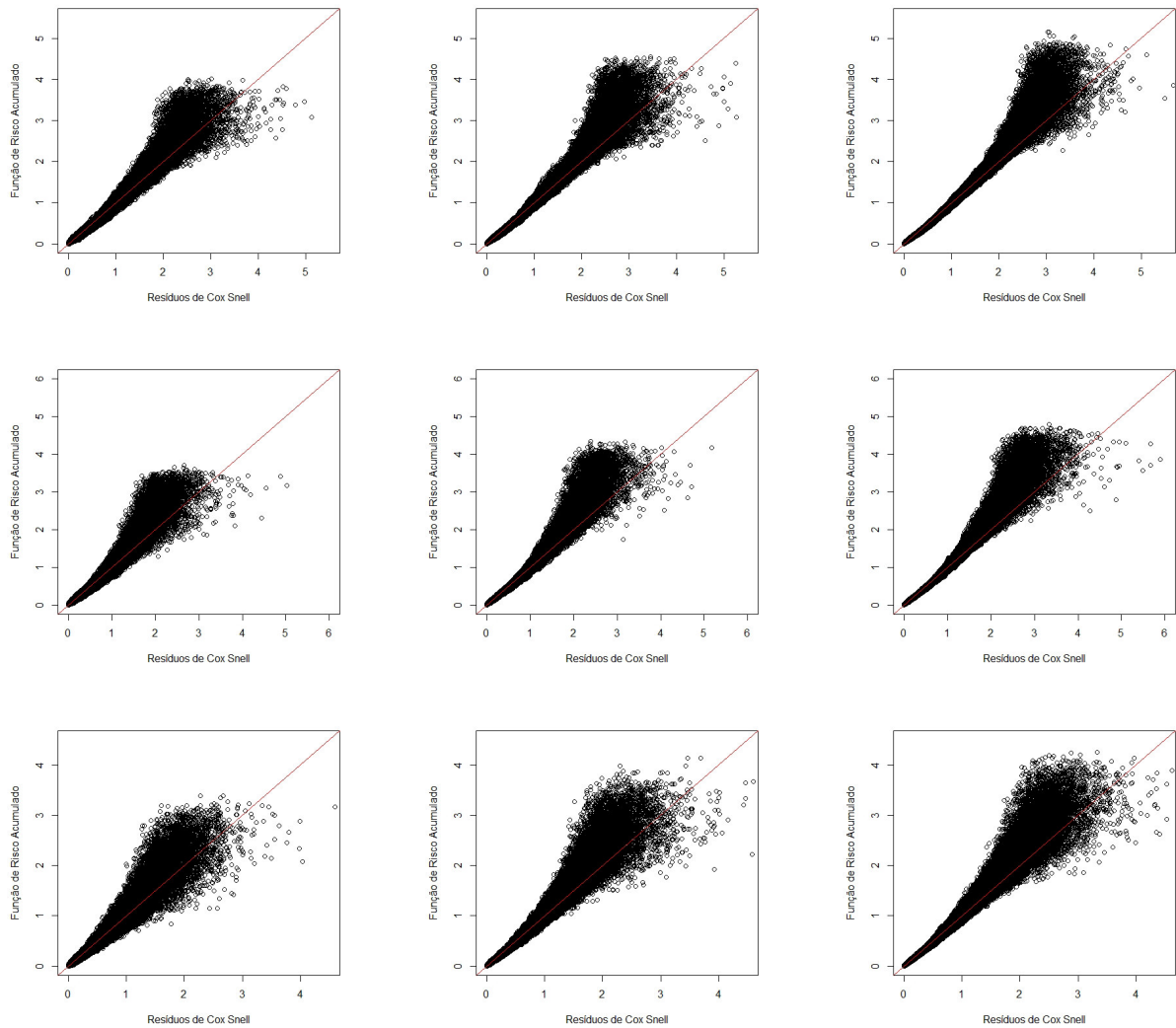


Figura 9: Resíduos do tipo Cox-Snell para avaliar o ajuste do modelo tempo de promoção Log-Normal com covariáveis (má especificação na incidência e na latência) para os níveis de censura leve, moderado e alto (de cima para baixo) e tamanho amostral 100, 250 e 500 (esquerda para a direita).

5 Aplicação

Neste Capítulo vamos apresentar uma aplicação a dados reais para utilização dos resíduos propostos no Capítulo 3. Os dados aqui utilizados encontram-se originalmente em Edmonson et al. (1979) e podem ser encontrados no pacote `survival` (THERNEAU, 2015) por nome `ovarian` do *software* R Core Team (2018). O objetivo original de Edmonson et al. (1979) foi o de utilizar um ensaio aleatorizado para comparar a eficácia de dois tratamentos para câncer de ovário em 26 pacientes.

A Figura 10 representa o gráfico da função de sobrevivência empírica estimada de forma não paramétrica através do estimador de Kaplan-Meier. Assim como foi comentado no Capítulo 2, uma vez que a curva de sobrevivência estimada estabilize-se bem acima de zero e o tempo de observação seja suficientemente grande, temos indícios para investigar a possível fração de curados existentes nos dados `ovarian`. No nosso caso, a curva estabilizou-se em torno de 0,5,0, nos fornecendo um indicativo da presença da fração de cura nos dados.

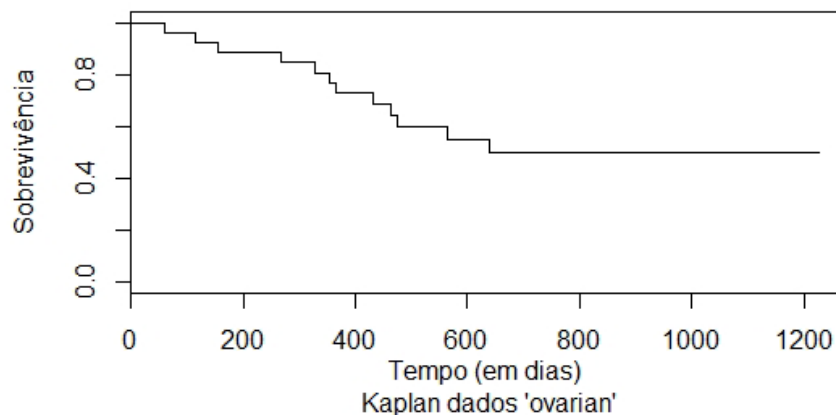


Figura 10: Curva de sobrevivência estimadas via Kaplan-Meier para dados ovarian.

5.1 Ajuste do modelo de regressão aos dados

Os tempos referentes aos dados *ovarian* são dispostos em dias até a morte do paciente por causa do câncer de ovário ou a ocorrência da censura, uma vez que esta é representada pela perda da informação do paciente ou morte deste por causa que não o câncer em estudo. As outras informações envolvidas são tipo de tratamento, a presença ou não de doença residual, uma medida de performance e idade.

Loose (2016) ajustou um modelo de tempo de promoção Weibull aos dados considerando as covariáveis inseridas no modelo através do parâmetro $\theta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$, $i = 1, \dots, 26$. Realizando testes de hipóteses constatou a significância apenas do parâmetro associado ao fator idade. Assim, o modelo final considerado por Loose (2016) é representado pela seguinte fração de cura

$$p_{\theta_i}(0) = \exp\{-\exp(x_i\beta)\}, \quad (5.1)$$

em que x_i assume valor 0 se idade < mediana (56,84) e valor 1 se idade \geq mediana.

A fim de avaliar o ajuste do modelo aos dados, além do modelo acima foram considerados outros ajustes, a saber: modelo de tempo de promoção log-normal, representado também por (5.1), e os modelos de mistura padrão Weibull e log-normal, representados pela fração de cura

$$p_{\theta_i}(0) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)}. \quad (5.2)$$

Observando a Figura 11, vamos escolher o ajuste que mais se aproxima da reta de coeficiente de inclinação 1. Baseados nisso, o ajuste escolhido foi o modelo de tempo de promoção Weibull. O critério AIC foi utilizado para escolha do ajuste do modelo aos dados (valores incluídos na Figura 11) e corrobora com a escolha gráfica feita acima.

A Tabela 2 apresenta as estimativas e os erros-padrão (entre parênteses) para os parâmetros dos quatro modelos ajustados aos dados de câncer de ovário.

Modelos	β_0	β_1	μ	σ
MTP-Weibull	-0,9054(0,4988)	1,1195(0,6119)	6,1891(0,1881)	0,8374(0,2621)
MMP-Weibull	0,7922(0,6315)	-1,6913(0,9474)	6,0347(0,1631)	0,7187(0,2710)
MTP-log-normal	-0,2825(1,8590)	1,2091(0,6290)	6,8190(2,3488)	0,0713(0,8484)
MMP-log-normal	0,5200(0,7377)	-1,9300(1,3791)	5,9615(0,3710)	-0,1669(0,3042)

Tabela 2: Estimativas (erros-padrão) para os parâmetros dos modelos ajustados aos dados de câncer de ovário.

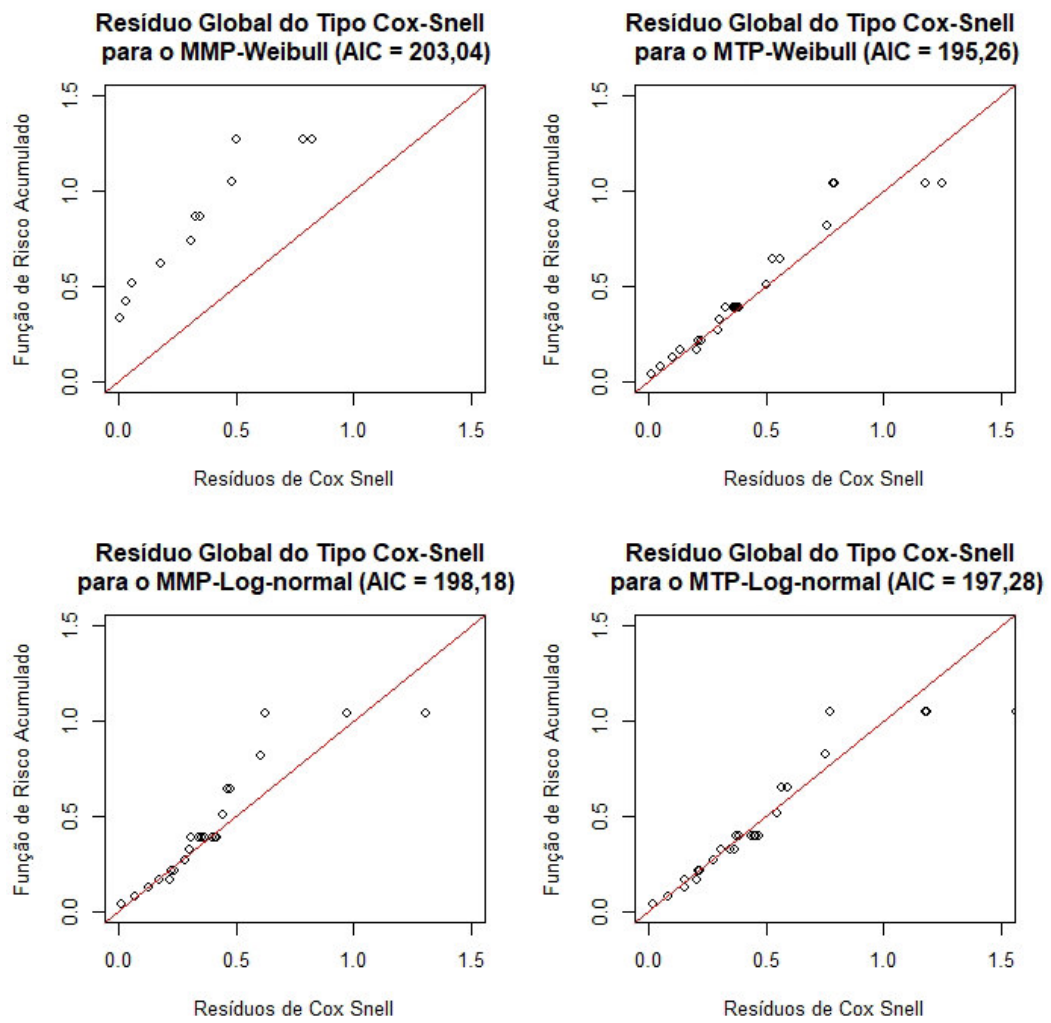


Figura 11: Resíduos do tipo Cox-Snell para os ajustes dos dados ovarian.

6 Considerações finais

Modelos de sobrevivência com fração de cura admitem que uma parte da população esteja curada e uma outra esteja suscetível ao evento em estudo. Em uma abordagem geral, estes modelos envolvem a suposição de duas distribuições: a distribuição da variável aleatória representando o número de causas competindo para a ocorrência do evento (distribuição incidência) e a distribuição associada ao tempo de vida dos indivíduos suscetíveis (distribuição de latência). Desta forma, a proposta de resíduos para estes modelos devem levar em consideração estas características além da presença de censura nos dados.

Neste trabalho, considerando modelos de sobrevivência com fração de cura em um contexto de censura à direita e com estrutura de regressão associada apenas à distribuição de incidência, descrevemos os resíduos do tipo Cox-Snell global proposto por Scolas et al. (2018) para o modelo de mistura padrão e estendemos este resíduos para modelos de tempo de promoção, apresentando para isso alguns aspectos teóricos. Realizamos um estudo de simulação que mostra a capacidade de detecção de violações de suposições com respeito à distribuição de latência e principalmente à distribuição de incidência. Como esperado, o aumento do tamanho amostral evidencia as possíveis violações existentes no ajuste do modelo aos dados, enquanto o aumento do nível de censura dificulta a identificação destas violações. Finalmente, aplicamos os resíduos aos ajustes de modelos de mistura padrão e tempo de promoção com distribuição de latência Weibull e log-normal de um conjunto de dados oriundos de um estudo de pacientes com câncer de ovário.

Como proposta futura para este estudo temos:

- Incluir covariáveis na distribuição de latência para avaliação dos resíduos do tipo Cox-Snell,
- Obter resultados gerais de resíduos para Modelos Log-simétrico com fração de cura (ROCHA, 2018),
- Propor resíduos para avaliar a forma funcional das covariáveis semelhante ao reali-

zado em Scolas et al. (2018),

- Comparar os resíduos propostos aqui o com a abordagem de quantil residual (DUNN; SMYTH, 1996).

Referências

- AHMANN, D. et al. Complete responses and long-term survivals after systemic chemotherapy for patients with advanced malignant melanoma. *Cancer*, Wiley Online Library, v. 63, n. 2, p. 224–227, 1989.
- AKAIKE, H. A new look at the statistical model identification. In: *Selected Papers of Hirotugu Akaike*. [S.l.]: Springer, 1974. p. 215–222.
- AMICO, M.; KEILEGOM, I. V. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, Annual Reviews, v. 5, p. 311–342, 2018.
- ARAÚJO, A. M. M. d. et al. Aplicação de modelos de mistura de longa duração em dados de reincidência ao crime. Florianópolis, SC, 2004.
- BERKSON, J.; GAGE, R. Survival curve for breast cancer patients following treatment. *J. her. Statis. Assoc.*, v. 47, p. 501–515, 1952.
- BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, v. 11, n. 1, p. 15–53, 1949.
- CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, Taylor & Francis, v. 94, n. 447, p. 909–919, 1999.
- COLLETT, D. *Modelling survival data in medical research*. 2nd. ed. [S.l.]: Chapman and Hall/CRC, 2003.
- COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. *Edgard Blucher*, 2006.
- CORDEIRO, G. M.; NETO, E. d. A. L. *Modelos paramétricos*. [S.l.]: Pernambuco: UFRPE, 2006.
- COX, D. Regression models and li-tables (with discussion). *J Royal Stat Soc, Series B*, p. 34–220, 1972.
- COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 248–275, 1968.
- CROWLEY, J.; HU, M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, Taylor & Francis, v. 72, n. 357, p. 27–36, 1977.
- DIRICK, L.; CLAESKENS, G.; BAESSENS, B. An akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, Elsevier, v. 241, n. 2, p. 449–457, 2015.

- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.
- EDMONSON, J. H. et al. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer treatment reports*, v. 63, n. 2, p. 241–247, 1979.
- FAREWELL, V. T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, JSTOR, p. 1041–1046, 1982.
- FONSECA, R. S.; VALENÇA, D. M.; BOLFARINE, H. Cure rate survival models with missing covariates: a simulation study. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 83, n. 1, p. 97–113, 2013.
- GUEDES, A. L. V. *Modelo de tempo de falha acelerado com fração de cura: uma abordagem unificada*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2011.
- KALBFLEISCH, J. Some extensions and applications of coxs regression and life model. In: INTERNATIONAL BIOMETRIC SOC 808 17TH ST NW SUITE 200, WASHINGTON, DC 20006-3910. *Biometrics*. [S.l.], 1974. v. 30, n. 3, p. 561–561.
- KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.
- KUK, A. Y.; , C.-H. . A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, Oxford University Press, v. 79, n. 3, p. 531–541, 1992.
- KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951.
- LAGAKOS, S. W. General right censoring and its impact on the analysis of survival data. *Biometrics*, JSTOR, p. 139–156, 1979.
- LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2003.
- LOOSE, L. H. *Avaliação e correções de testes de hipóteses em modelos de sobrevivência com fração de cura*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2016.
- MALLER, R.; ZHOU, X. *Survival Analysis with Long-Term Survivors*. [S.l.]: Wiley, 1996. (Wiley Series in Child Care and Protection). ISBN 9780471962014.
- PENG, Y.; TAYLOR, J. M. Residual-based model diagnosis methods for mixture cure models. *Biometrics*, Wiley Online Library, v. 73, n. 2, p. 495–505, 2017.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <<https://www.R-project.org/>>.
- RINCÓN, L. *Curso intermedio de PROBABILIDAD*. [S.l.]: Facultad de Ciencias, UNAM, 2007. Url<http://www.geocities.ws/mahilhm/rinconproba2.pdf>. ISBN 978-970-32-4269-6.

- ROCHA, J. B. *Modelos log-simétricos com fração de cura*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2018.
- RODRIGUES, J. et al. On the unification of long-term survival models. *Statistics & Probability Letters*, Elsevier, v. 79, n. 6, p. 753–759, 2009.
- SCOLAS, S. et al. Variable selection in a flexible parametric mixture cure model with interval-censored data. *Statistics in medicine*, Wiley Online Library, v. 35, n. 7, p. 1210–1225, 2016.
- SCOLAS, S. et al. Diagnostic checks in mixture cure models with interval-censoring. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 27, n. 7, p. 2114–2131, 2018.
- SILVA, R. P. d. *Modelos flexíveis de sobrevivência com fração de cura: implementação computacional*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2015.
- SILVA, R. P. da; VALENÇA, D. M. *flexcure: Flexible Parametric Survival Models with Cure Fraction*. [S.l.], 2015. R package version 0.1. Disponível em: <<https://github.com/Rumenick/flexcure>>.
- SPOSTO, R. Cure model analysis in cancer: an application to data from the children's cancer group. *Statistics in medicine*, Wiley Online Library, v. 21, n. 2, p. 293–312, 2002.
- TAMHANE, A. C.; SHI, J. Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of fdr. *Lecture Notes-Monograph Series*, JSTOR, p. 304–325, 2009.
- THERNEAU, T. M. *A Package for Survival Analysis in S*. [S.l.], 2015. Version 2.38. Disponível em: <<https://CRAN.R-project.org/package=survival>>.
- TOURNOUD, M.; ECOCHARD, R. Application of the promotion time cure model with time-changing exposure to the study of hiv/aids and other infectious diseases. *Statistics in medicine*, Wiley Online Library, v. 26, n. 5, p. 1008–1021, 2007.
- VARSHNEY, M. et al. Cure fraction model for the estimation of long-term survivors of hiv/aids patients under antiretroviral therapy. *J. Commun. Dis*, v. 50, n. 3, 2018.
- WILEYTO, E. P. et al. Assessing the fit of parametric cure models. *Biostatistics*, Oxford University Press, v. 14, n. 2, p. 340–350, 2013.
- YAKOVLEV, A. Y. et al. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et analyse de donnees spatio-temporelles*, Société Française de Biométrie, ENSA Rennes France, v. 12, p. 66–82, 1993.
- YUSUF, M. U.; BAKAR, M. R. A. Cure models based on weibull distribution with and without covariates using right censored data. *Indian Journal of Science and Technology*, v. 9, n. 28, 2016.

APÊNDICE A - Obtenção da Função de Verossimilhança

Neste seção, utilizando as informações dispostas no Capítulo 2, mostraremos os passos para a obtenção da função de verossimilhança para os dados completos dada em (2.10) e posteriormente para a função de verossimilhança marginal disposta em (2.11).

Verossimilhança para os dados completos

Na definição das variáveis aleatórias no Capítulo 2 percebemos que o tempo que pode ser observado para cada indivíduo poderia ser visto como $Y_i = \min\{T_i, C_i\}$, sendo $T_i = \min\{Z_{i0}, Z_{i1}, \dots, Z_{im_i}\}$ representando o tempo até a ocorrência do evento de interesse, em que Z_{ik} , $i = 1, \dots, n$ e $k = 0, 1, \dots, m_i$ com função de densidade $f_Z(t; \boldsymbol{\xi})$ e função de sobrevivência $S_Z(t; \boldsymbol{\xi})$, são variáveis aleatórias independentes e identicamente distribuídas, e C_i representa o tempo até a ocorrência da censura, sendo esta aleatória e não informativa, com função densidade dada por $g(t)$ e função de sobrevivência $G(t)$.

Sendo assim, temos que

$$\begin{aligned}
 S_T(t; \boldsymbol{\xi} | m_i) &= P(T > t | M_i = m_i) \\
 &= P(\min\{Z_{i0}, Z_{i1}, Z_{i2}, \dots, Z_{im_i}\} > t) \\
 &= P(Z_{i0} > t, Z_{i1} > t, Z_{i2}, \dots, Z_{im_i} > t) \\
 &= P(Z_{i0} > t)P(Z_{i1} > t) \cdots P(Z_{im_i} > t) \\
 &= \underbrace{S_Z(t; \boldsymbol{\xi}) S_Z(t; \boldsymbol{\xi}) \cdots S_Z(t; \boldsymbol{\xi})}_{m_i \text{ vezes}} \\
 \Rightarrow S_T(t; \boldsymbol{\xi} | m_i) &= [S_Z(t; \boldsymbol{\xi})]^{m_i}. \tag{A.1}
 \end{aligned}$$

Sendo assim, a partir de 2.1

$$\begin{aligned} f_T(t; \boldsymbol{\xi}|m_i) &= -\frac{d}{dt}S_T(t; \boldsymbol{\xi}|m_i) = -\frac{d}{dt}[S_Z(t; \boldsymbol{\xi})]^{m_i} \\ &= m_i[S_Z(t; \boldsymbol{\xi})]^{m_i-1} f_Z(t; \boldsymbol{\xi}). \end{aligned}$$

Dado $M_i = m_i$ a verossimilhança condicional de $\boldsymbol{\xi}$ 03 correspondente a um indivíduo i (ou uma observação, $D_i = (y_i, \delta_i)$) é dada por

$$\begin{aligned} L(\boldsymbol{\xi}; \mathbf{D}_i | M_i = m_i) &= [f_T(y_i; \boldsymbol{\xi}|m_i)]^{\delta_i} [S_T(y_i; \boldsymbol{\xi}|m_i)]^{1-\delta_i} \\ &= [m_i[S_Z(y_i; \boldsymbol{\xi})]^{m_i-1} f_Z(y_i; \boldsymbol{\xi})]^{\delta_i} [(S_Z(y_i; \boldsymbol{\xi}))^{m_i}]^{1-\delta_i} \\ &= [m_i f_Z(y_i; \boldsymbol{\xi})]^{\delta_i} [S_Z(y_i; \boldsymbol{\xi})]^{m_i-\delta_i}. \end{aligned}$$

Para obter a função de verossimilhança de $\boldsymbol{\lambda}$ correspondentes aos dados completos $\mathbf{D}_c = (n, \mathbf{y}, \boldsymbol{\delta}, \mathbf{M}, \mathbf{X})$ para um indivíduo i , dada por \mathbf{D}_{c_i} , temos

$$\begin{aligned} L(\boldsymbol{\lambda}, \mathbf{D}_{c_i}) &= \lim_{\Delta y_i \rightarrow} \frac{P(y_i \leq Y_i < y_i + \Delta y_i, \delta_i = \delta_{obs}, M_i = m_i)}{\Delta y_i} \\ &= \begin{cases} h_1, & \text{se } \delta_i = 1; \\ h_2, & \text{se } \delta_i = 0. \end{cases} \end{aligned}$$

Dessa forma, se

i) Se $\delta_i = 0$,

$$\begin{aligned} h_0 &= \lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq Y_i < y_i + \Delta y_i, \delta_i = 0, M_i = m_i)}{\Delta y_i} \\ &= \lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq C_i < y_i + \Delta y_i, T_i > C_i, M_i = m_i)}{\Delta y_i}, \end{aligned}$$

em que

$$\begin{aligned}
P(y_i \leq C_i < y_i + \Delta y_i, T > C_i, M_i = m_i) &\cong P(y_i \leq C_i < y_i + \Delta y_i, T_i > y_i, M_i = m_i) \\
&= P(y_i \leq C_i < y_i + \Delta y_i, T_i > y_i | M_i = m_i) \times \\
&\quad \times P_\theta(M_i = m_i) \\
&= P(y_i \leq C_i < y_i + \Delta y_i | M_i = m_i) \times \\
&\quad \times P(T_i > y_i | M_i = m_i) P_\theta(M_i = m_i) \\
&= P(y_i \leq C_i < y_i + \Delta y_i | M_i = m_i) \times \\
&\quad \times S_T(y_i | m_i) P_\theta(M_i = m_i).
\end{aligned}$$

Logo,

$$\begin{aligned}
h_0 &= \lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq C_i < y_i + \Delta y_i | M_i = m_i) S_T(y_i | m_i) P_\theta(M_i = m_i)}{\Delta y_i} \\
&= \left[\lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq C_i < y_i + \Delta y_i | M_i = m_i)}{\Delta y_i} \right] S_T(y_i | M_i) P_\theta(M_i = m_i) \\
&= g(y_i | m_i) S_T(y_i | m_i) P_\theta(M_i = m_i).
\end{aligned}$$

ii) Se $\delta_i = 1$,

$$\begin{aligned}
h_1 &= \lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq Y_i < y_i + \Delta y_i, \delta_i = 1, M_i = m_i)}{\Delta y_i} \\
&= \lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq R_i < y_i + \Delta y_i, T_i \leq C_i, M_i = m_i)}{\Delta y_i}.
\end{aligned}$$

em que

$$\begin{aligned}
P(y_i \leq T_i < y_i + \Delta y_i, T_i \leq C_i, M_i = m_i) &\cong P(y_i \leq T_i < y_i + \Delta y_i, C_i \geq y_i, M_i = m_i) \\
&= P(y_i \leq T_i < y_i + \Delta y_i, C_i \geq y_i | M_i = m_i) \\
&\quad \times P_\theta(M_i = m_i) \\
&= P(y_i \leq T_i < y_i + \Delta y_i | M_i = m_i) \\
&\quad \times P(C_i \geq y_i | M_i = m_i) P_\theta(M_i = m_i).
\end{aligned}$$

Logo,

$$\begin{aligned}
h_1 &= \lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq T_i < y_i + \Delta y_i | M_i = m_i) G(y_i | m_i) P(M_i = m_i)}{\Delta y_i} \\
&= \left[\lim_{\Delta y_i \rightarrow 0} \frac{P(y_i \leq T_i < y_i + \Delta y_i | M_i = m_i)}{\Delta y_i} \right] G(y_i | m_i) P(M_i = m_i) \\
&= f_T(y_i | m_i) G(y_i | m_i) P(M_i = m_i).
\end{aligned}$$

Portanto,

$$\begin{aligned}
L(\boldsymbol{\lambda}; \mathbf{D}_{ci}) &= \begin{cases} f_T(y_i | m_i) G(y_i | m_i) P_\theta(M_i = m_i), & \text{se } \delta_i = 1 \\ g(y_i | m_i) S_T(y_i | M_i) P_\theta(M_i = m_i), & \text{se } \delta_i = 0, \end{cases} \\
&= [f_T(y_i | m_i) P(M_i = m_i)]^{\delta_i} S_T(y_i | M_i) P_\theta(M_i = m_i)^{1-\delta_i} J \\
&\propto [f_T(y_i | m_i)]^{\delta_i} [S_T(y_i | M_i)]^{1-\delta_i} P_\theta(M_i = m_i) \\
&= L(\boldsymbol{\xi}; \mathbf{D}_i | M_i = m_i) P_\theta(M_i = m_i),
\end{aligned}$$

em que $J = g(y_i | m_i)^{1-\delta_i} G(y_i | m_i)^{\delta_i}$. Sendo assim, a função de verossimilhança conjunta para os dados completos ficará determinado da seguinte forma

$$\begin{aligned}
L(\boldsymbol{\lambda}; \mathbf{D}_c) &= \prod_{i=1}^n L(\boldsymbol{\xi}; \mathbf{D}_i | M_i = m_i) P_\theta(M_i = m_i) \\
&= [f_Z(y_i | m_i)]^{\delta_i} S_Z(y_i | M_i)^{1-\delta_i} P_\theta(M_i = m_i). \tag{A.2}
\end{aligned}$$

Verossimilhança Marginal

Semelhante ao que foi realizado na seção anterior, nesta seção mostraremos os passos e a obtenção da função de verossimilhança marginal apresentada em (2.11).

A função de verossimilhança marginal é obtida a partir da realização do somatório de todos os possíveis valores de M_i na função de verossimilhança para os dados completos em (A.2).

Sendo assim,

$$\begin{aligned}
L^*(\boldsymbol{\lambda}; \mathbf{D}_i) &= \sum_{m_i=0}^{\infty} L(\boldsymbol{\lambda}; \mathbf{D}_{c_i}) \\
&= \sum_{m_i=0}^{\infty} L(\boldsymbol{\xi}; \mathbf{D}_i | M_i = m_i) P(M_i = m_i) \\
&= \sum_{m_i=0}^{\infty} [m_i f_Z(y_i; \boldsymbol{\xi})]^{\delta_i} [S_Z(y_i; \boldsymbol{\xi})]^{m_i - \delta_i} P_{\theta}(M_i = m_i). \tag{A.3}
\end{aligned}$$

Se considerarmos os casos (i) e (ii), vistos na seção anterior, temos que

$$\begin{aligned}
L^*(\boldsymbol{\lambda}; \mathbf{D}_i) &= \begin{cases} \sum_{m_i=0}^{\infty} [S_Z(y_i; \boldsymbol{\xi})]^{m_i} P(M_i = m_i), & \text{se } \delta_i = 0 \\ \sum_{m_i=0}^{\infty} m_i f_Z(y_i; \boldsymbol{\xi}) [S_Z(y_i; \boldsymbol{\xi})]^{m_i} P(M_i = m_i), & \text{se } \delta_i = 1, \end{cases} \\
&= \begin{cases} S_p(y_i; \boldsymbol{\xi}), & \text{se } \delta_i = 0 \\ f_p(y_i; \boldsymbol{\xi}), & \text{se } \delta_i = 1, \end{cases} \\
&= [f_p(y_i; \boldsymbol{\xi})]^{\delta_i} [S_p(y_i; \boldsymbol{\xi})]^{1 - \delta_i}.
\end{aligned}$$

Logo, a função de verossimilhança para os dados completos será dada por:

$$\begin{aligned}
L^*(\boldsymbol{\lambda}; D) &= \prod_{i=1}^n L^*(\boldsymbol{\lambda}; D_i) \\
&= \prod_{i=1}^n [f_p(y_i; \boldsymbol{\xi})]^{\delta_i} [S_p(y_i; \boldsymbol{\xi})]^{1 - \delta_i}. \tag{A.4}
\end{aligned}$$

APÊNDICE B – Aspectos Computacionais

Exemplo de programa em R Core Team (2018) para gerar dados de acordo com o modelo de mistura padrão Weibull e avaliar via o resíduo do tipo Cox-Snell descrito em 3.1.

```
#####
```

```
#limpando todos os objetos
```

```
rm(list=ls())
```

```
#guardando os resultados
```

```
set.seed(126497)
```

```
#carregando pacote
```

```
require(survival)
```

```
#preliminares: gerando covariáveis e tamanhos de amostra
```

```
#tamanhos de amostra
```

```
n1 = 500
```

```
#covariáveis para o tamanho n = n1
```

```

x1n = rnorm(n1,0.5,1.5)
x2n = runif(n1,-3,3)
x3n = rbinom(n1,1,.5)

#A função D() é a que gera os dados

D =function(u,pforma,pscale,beta,x0,x1,x2,x3){#função para gerar dados
%com %fc e %cens (obs.: se u=0 => %cens = 0)
#1 passo
X=matrix(c(x0,x1,x2,x3),nrow=n)
eta=X%*%beta
theta=exp(eta)/(1+exp(eta))
M=rbinom(n, 1, (1-theta))
#cada valor i gerado (i=1,...n), corresponde a uma Bernoulli(thetai)
#passo 2
cens=runif(n,0,u)
#passos 3,4, 5 e 6
for(i in 1:n){
if(M[i]==0){
t[i]=max(cens)
y[i]=t[i]
d[i]=0
}else{
if(u!=0){
t[i]= rweibull(1,pforma,pscale)
y[i] = min(t[i], cens[i])
d[i] = ifelse(t[i] < cens[i], 1, 0)
}else{
t[i]= rweibull(1,pforma,pscale)
y[i] = t[i]
d[i] = 1
} } }
Dc=matrix(c(y,d,X,M),nrow=n)
Pcura=mean(M==0) #proporção de curados na amostra

```

```

#proporção de censuras dentre os susceptíveis
pc1 = sum(d==0 & M>0)/sum(M>0)
pc2 = pc1*(1-Pcura)+Pcura
return(list(Dc,Pcura,pc1,pc2))
}

#cenário A

#eta(X) = B0 - X2 + X3
#censura leve: u = 2.1, B0 = -3.5 e n = 100

#configurações

#iniciando valores necessários
y = d= M = cens = t = phi = pc1 = Pcura = NULL
#valores para gerar os dados
n = 100 #tamanho da amostra
x0<- rep(1,n)
x1<- x1n[1:n]
x2<- x2n[1:n]
x3<- x3n[1:n]
u = 2.1
pforma = 2
pscale= 0.5
beta = c(-4.5,.5,-1,1)

#gerando os dados
Dc = D(u,pforma,pscale,beta,x0,x1,x2,x3)
(fração_curados<- Dc[2])
(censura_sus<- Dc[3])
(cens_total<- Dc[4])

nloop = 0
r = 2000
erro = 0

```

```

ests<-matrix(,r,6) #matriz com as 10000 estimativas
guardar_tempos<- matrix(,n,r)
h1<- matrix(,n,r)
indicadora<- matrix(,n,r)

while(nloop<r){
#1 - gerando os dados
Dc = D(u=u,pforma=pforma,pscale=pscale,
beta=beta,x0=x0,x1=x1,x2=x2,x3=x3)
#saída dos dados: Dc[1]: dados;
#Dc[2]: fração de curados
#Dc[3]: proporção de censura dentre os suscetíveis

Dados = data.frame(Dc[1]) #transformando em data.frame
names(Dados)=c("tempo","status","x0","x1","x2","x3","M")
Dobs=Dados[,-7] #Dados observados (sem variáveis latentes)
guardar_tempos[,nloop+1] = as.numeric(Dc[[1]][,1])
indicadora[,nloop+1] = as.numeric(Dobs$status)

#3 - realizar estimativas com base nos dados gerados

#-->> densidade da Weibull
f = function(pforma,pscale,y){
fy = (pforma/(pscale^pforma))*(y^(pforma-1))*exp(-(y/pscale)^pforma)
return(fy)
}

#-->> sobrevivência da Weibull
s = function(pforma,pscale,y){
sy = exp(-(y/pscale)^pforma)
return(sy)
}

#-->> Função de Log-Verossimilhança Marginal
l = function(phi,D){
X = as.matrix(D[,3:(dim(D)[2])])

```



```

y = D[,1]
delta = D[,2]
pforma = phi[ncol(X)+1]
pscale = phi[ncol(X)+2]
b = matrix(phi[1:(ncol(X))], ncol=1)
eta=X%*%b
theta=exp(eta)/(1+exp(eta))
sp = theta+(1-theta)*s(pforma,pscale,y)
fp = (1 - theta)*f(pforma,pscale,y)
li= sum( delta*log(fp) + (1-delta)*log(sp) )
return(li)
}

#=====
# MAXIMIZANDO A FUNÇÃO DE LOG-VEROSSIMILHANÇA
#=====
# Utilizando o OPTIM (emv restrito e irrestrito)
phicc=c(beta,pforma,pscale)
phi.max=try(optim(phicc,l,D=Dobs,method='BFGS',hessian=T,
control=list(fnscale=-1)),T)

#parâmetros

gama = phi.max$par[5]
alpha = phi.max$par[6]
eta = phi.max$par[1]+phi.max$par[2]*x1+phi.max$par[3]*x2
+phi.max$par[4]*x3

#construção do resíduo tipo Cox-Snell
thetawei = exp(eta)/(1+exp(eta))
y = as.numeric(Dc[[1]][,1])
s0wei = exp(-(y/alpha)^gama)
sptwei = thetawei+(1-thetawei)*s0wei
h1[,nloop+1] = -log(sptwei)
indicadora[,nloop+1] = as.numeric(Dobs$status)

```

```

if(class(phi.max)=="try-error"){
nloop = nloop
erro = erro + 1
}else{
ests[nloop+1,1]<-round(phi.max$par[1],4)
ests[nloop+1,2]<-round(phi.max$par[2],4)
ests[nloop+1,3]<-round(phi.max$par[3],4)
ests[nloop+1,4]<-round(phi.max$par[4],4)
ests[nloop+1,5]<-round(phi.max$par[5],4)
ests[nloop+1,6]<-round(phi.max$par[6],4)
nloop = nloop + 1 #incremento do while
erro = erro
}
results<-data.frame(ests)
names(results)<-c("beta0","beta1","beta2","beta3","gama","alpha")
}

results<- as.data.frame(results)

write.table(results, file="C:/Users/clean/Dropbox/Dissertação/
Cronograma Semanas/Semana 21/MMP/Covariaveis/Weibull-Weibull/
Leve/estimativas_n100.txt")

(parametros<- c(beta,pforma,pscale))

#medidas avaliativas
#média
(media <- apply(results,2,"mean"))

#desvio padrão
(dp_beta0<- round(sqrt((1/(r-1))*sum((results$beta0-
mean(results$beta0))^2)),4))
(dp_beta1<- round(sqrt((1/(r-1))*sum((results$beta1-
mean(results$beta1))^2)),4))

```

```

(dp_beta2<- round(sqrt((1/(r-1))*sum((results$beta2-
mean(results$beta2))^2)),4))
(dp_beta3<- round(sqrt((1/(r-1))*sum((results$beta3-
mean(results$beta3))^2)),4))
(dp_gama<- round(sqrt((1/(r-1))*sum((results$gama-
mean(results$gama))^2)),4))
(dp_alpha<- round(sqrt((1/(r-1))*sum((results$alpha-
mean(results$alpha))^2)),4))

#raiz do eqm
(r_eqm_beta0<- round(sqrt((1/r)*sum((results$beta0-
parametros[1])^2)),4))
(r_eqm_beta1<- round(sqrt((1/r)*sum((results$beta1-
parametros[2])^2)),4))
(r_eqm_beta2<- round(sqrt((1/r)*sum((results$beta2-
parametros[3])^2)),4))
(r_eqm_beta3<- round(sqrt((1/r)*sum((results$beta3-
parametros[4])^2)),4))
(r_eqm_gama<- round(sqrt((1/r)*sum((results$gama-
parametros[5])^2)),4))
(r_eqm_alpha<- round(sqrt((1/r)*sum((results$alpha-
parametros[6])^2)),4))

medidas_ajuste<- rbind(media,
c(dp_beta0,dp_beta1,dp_beta2,dp_beta3,
dp_gama,dp_alpha),
c(r_eqm_beta0,r_eqm_beta1,r_eqm_beta2,r_eqm_beta3,
r_eqm_gama,r_eqm_alpha))
medidas_ajuste<- as.data.frame(medidas_ajuste, row.names = c("Media",
"Desvio padrão", "Raiz EQM"))

write.table(medidas_ajuste, file="C:/Users/clean/Dropbox/Dissertação/
Cronograma Semanas/Semana 21/MMP/Covariaveis/Weibull-Weibull/
Leve/medidas_ajuste_n100.txt")

```

```
#resíduo
resultado_time <- matrix(rep(NA, r*n), ncol=r)
resultado_surv <- matrix(rep(NA, r*n), ncol=r)
j = 1
for (j in 1:r){
  x <- survfit(Surv(h1[,j],indicadora[,j])~1,
  conf.type = "none")
  resultado_time[1:length(x$time), j] <- as.numeric(x$time)
  resultado_surv[1:length(x$surv), j] <- as.numeric(x$surv)
}
par(mfrow = c(1,1))
matplot(resultado_time, -log(resultado_surv), type="p",
col="black", pch=1,
xlab = "Resíduos de Cox Snell", ylab = "Função de Risco Acumulado",
xlim = c(0,6), ylim = c(0,6))
abline(c(0,3),c(1,0), col = "red")
```

APÊNDICE C - Códigos da aplicação

Os códigos retratados abaixo são referentes a aplicação para os dados `ovarian` encontrado no Capítulo 5.

```
#limpando todos os objetos
```

```
rm(list=ls())
```

```
require(flexcure)
```

```
require(flexsurv)
```

```
dados<- data.frame(ovarian)
```

```
#kaplan-meier para os dados completos
```

```
par(mfrow = c(1,1))
```

```
ekm<- survfit(Surv(futime,fustat)~1, data=dados, conf.type = "none")
```

```
plot(ekm, xlab = "Tempo (em dias) \n Kaplan dados 'ovarian' ",
```

```
ylab = "Sobrevivência")
```

```
#categorizando idade (age)
```

```
#para valores maiores ou iguais a mediana
```

```
#e para valores menores que a mediana
```

```
(med<- median(ovarian$age))
```

```
ovarian$age[ovarian$age<med]<-0
```

```
ovarian$age[ovarian$age>=med]<-1
```

```
ovarian$resid.ds[ovarian$resid.ds==1]<-0
```

```
ovarian$resid.ds[ovarian$resid.ds==2]<-1
```

```

#Exploração inicial
help(ovarian)
#fuptime: Institution code
#fustat: Survival time in days
#age: idade em anos
#resid.ds: doença residual presente (1 = no, 2 = yes)
#rx: grupo de tratamento
#ecog.ps: ECOG status de desempenho
# Source: Terry Therneau

# Explorando os dados

str(ovarian)
summary(ovarian)
table("censura=1, falha=2 "=ovarian$fustat)
table("present=1, no.present=2 "=ovarian$resid.ds)
table("better=1, bad=2 "=ovarian$ecog.ps)
table(ovarian$age)

#Graficos de KM para diferentes níveis de cada fator
par(mfrow=c(2,2))
#age
ekm<- survfit(Surv(fuptime,fustat)~age, data=dados)
plot(survfit(Surv(fuptime,fustat)~age,data=ovarian),xlab="Tempo",
ylab="Sobrevivencia",lty=c(1,2), col=c(2,4),
sub="Kaplan-Meier Dados `ovarian` - age ")
legend(10,.75,c("abaixo da med","maior ou = a med"),
col=c(2,4),lty=c(1,2), bty = "n", cex = 0.6)

#resid.ds
ekm<- survfit(Surv(fuptime,fustat)~resid.ds, data=dados)
plot(survfit(Surv(fuptime,fustat)~resid.ds,data=ovarian),xlab="Tempo",
ylab="Sobrevivencia",lty=c(1,2), col=c(2,4),
sub="Kaplan-Meier Dados `ovarian` - resid.ds ")

```

```

legend(10, .7, c("no", "yes"), col=c(2,4),
lty=c(1,2), bty = "n", cex = 0.6)

#rx
ekm<- survfit(Surv(futime,fustat)~rx, data=dados)
plot(survfit(Surv(futime,fustat)~rx,data=ovarian),xlab="Tempo",
ylab="Sobrevivencia",lty=c(1,2), col=c(2,4),
sub="Kaplan-Meier Dados ´ovarian´ - resid.ds ")
legend(10, .7, c("tratamento", "placebo"), col=c(2,4),
lty=c(1,2), bty = "n", cex = 0.6)

#ecog.ps
ekm<- survfit(Surv(futime,fustat)~ecog.ps, data=dados)
plot(survfit(Surv(futime,fustat)~ecog.ps,data=ovarian),
xlab="Tempo", ylab="Sobrevivencia",lty=c(1,2), col=c(2,4),
sub="Kaplan-Meier Dados ´ovarian´ - resid.ds ")
legend(10, .7, c("Bet", "Bad"), col=c(2,4), lty=c(1,2), bty = "n", cex = 0.6)
par(mfrow=c(1,1))

#####
# SELEÇÃO DO MODELO PARAMÉTRICO VIA RESÍDUO GLOBAL DO TIPO COX SNELL
#####
#mtp-weibull sem covariáveis

modelo.wei<- curereg(Surv(futime,fustat)~1,
cureformula =~1, timedist = "weibull", data=ovarian)

#estimação dos parâmetros associados ao modelo

gama = exp(modelo.wei$coefficients$coef.time[2])
alpha = exp(modelo.wei$coefficients$coef.time[1])
etawei = modelo.wei$coefficients$coef.cure[1]

#construção do resíduo tipo Cox-Snell
thetawei = exp(etawei)#/(1+exp(etawei))

```

```

y = ovarian$futime
s0wei = exp(-(y/alpha)^gama)
sptwei = exp(-thetawei*(1 - s0wei))
h1<- -log(sptwei)
indicadora<- ovarian$fustat

#estimação não paramétrica
x<- survfit(Surv(h1,indicadora)~1, conf.type = "none")
resultado_time <- x$time
resultado_surv <- x$urv

#construção do gráfico
par(mfrow = c(1,1))
matplot(resultado_time, -log(resultado_surv), type="p", col="black",
pch=1,
xlab = "Resíduos de Cox Snell", ylab = "Função de Risco Acumulado",
xlim = c(0,1.5), ylim = c(0,1.5))
abline(c(0,3),c(1,0), col = "red")
title("Resíduo Global do Tipo Cox-Snell \n para o MTP-Weibull")

#mmp-weibull com a covariável age

modelo.wei.mmp.cv<- curereg(Surv(futime,fustat)~1,
cureformula =~1+factor(age),
timedist = "weibull", ncausedist = "bernoulli", data=ovarian)

#estimação dos parâmetros associados ao modelo

gama = exp(modelo.wei.mmp.cv$coefficients$coef.time[2])
alpha = exp(modelo.wei.mmp.cv$coefficients$coef.time[1])
etawei = modelo.wei.mmp.cv$coefficients$coef.cure[1]+
ovarian$age*modelo.wei.mmp.cv$coefficients$coef.cure[2]

#construção do resíduo tipo Cox-Snell
thetawei = exp(etawei)/(1+exp(etawei))

```



```

y = ovarian$futime
s0wei = exp(-(y/alpha)^gama)
sptwei = thetawei+(1 - thetawei)*s0wei
h1<- -log(sptwei)
indicadora<- ovarian$fustat

x<- survfit(Surv(h1,indicadora)~1, conf.type = "none")
resultado_time <- x$time
resultado_surv <- x$surv

#gráfico
par(mfrow = c(2,2))
matplot(resultado_time, -log(resultado_surv), type="p", col="black",
pch=1,xlab = "Resíduos de Cox Snell",
ylab = "Função de Risco Acumulado",
xlim = c(0,1.5), ylim = c(0,1.5))
abline(c(0,3),c(1,0), col = "red")
title("Resíduo Global do Tipo Cox-Snell \n
para o MMP-Weibull com a covariável Age")

#mtp-weibull com a covariável age

modelo.wei.cv<- curereg(Surv(futime,fustat)~1,
cureformula =~1+factor(age), timedist = "weibull", data=ovarian)

modelo.wei.cv2<- curereg(Surv(futime,fustat)~1,
cureformula =~1+factor(age)+factor(resid.ds),
timedist = "weibull",data=ovarian)

#estimação dos parâmetros associados ao modelo

gama = exp(modelo.wei.cv$coefficients$coef.time[2])
alpha = exp(modelo.wei.cv$coefficients$coef.time[1])
etawei = modelo.wei.cv$coefficients$coef.cure[1]+
ovarian$age*modelo.wei.cv$coefficients$coef.cure[2]

```

```

#construção do resíduo tipo Cox-Snell
thetawei = exp(etawei)#/(1+exp(etawei))
y = ovarian$futime
s0wei = exp(-(y/alpha)^gama)
sptwei = exp(-thetawei*(1 - s0wei))
h1<- -log(sptwei)
indicadora<- ovarian$fustat

x<- survfit(Surv(h1,indicadora)~1, conf.type = "none")
resultado_time <- x$time
resultado_surv <- x$urv

#construção do gráfico
#par(mfrow = c(2,2))
matplot(resultado_time, -log(resultado_surv), type="p", col="black",
pch=1,
xlab = "Resíduos de Cox Snell", ylab = "Função de Risco Acumulado",
xlim = c(0,1.5), ylim = c(0,1.5))
abline(c(0,3),c(1,0), col = "red")
title("Resíduo Global do Tipo Cox-Snell \n
para o MTP-Weibull com a covariável Age")

#fração de cura média

cure<- curefraction(modelo.wei.cv)

mean(cure[, "curefraction"])

#mmp-log-normal com a covariável age

modelo.ln.mmp.cv<- curereg(Surv(futime,fustat)~1,
cureformula =~1+factor(age),
timedist = "lnorm", ncausedist = "bernoulli" ,data=ovarian)

```

```

#estimação dos parâmetros associados ao modelo

mu = modelo.ln.mmp.cv$coefficients$coef.time[1]
sigma = exp(modelo.ln.mmp.cv$coefficients$coef.time[2])
etaln = modelo.ln.mmp.cv$coefficients$coef.cure[1]+
ovarian$age*modelo.ln.mmp.cv$coefficients$coef.cure[2]

#construção do resíduo tipo Cox-Snell
thetaln = exp(etaln)/(1+exp(etaln))
y = ovarian$futime
s0ln = pnorm((-log(y)+mu)/sigma)
sptln = thetaln+(1-thetaln)*s0ln
h1<- -log(sptln)
indicadora<- ovarian$fustat

x<- survfit(Surv(h1,indicadora)~1, conf.type = "none")
resultado_time <- x$time
resultado_surv <- x$urv

#construção do gráfico
#par(mfrow = c(1,1))
matplot(resultado_time, -log(resultado_surv),
type="p", col="black", pch=1,
xlab = "Resíduos de Cox Snell", ylab = "Função de Risco Acumulado",
xlim = c(0,1.5), ylim = c(0,1.5))
abline(c(0,3),c(1,0), col = "red")
title("Resíduo Global do Tipo Cox-Snell \n
para o MMP-Log-normal com a covariável Age")

#mtp-log-normal com a covariável age

modelo.ln.mtp.cv<- curereg(Surv(futime,fustat)~1,
cureformula =~1+factor(age),

timedist = "lnorm", ncausedist = "poisson" ,data=ovarian)

```

```

#estimação dos parâmetros associados ao modelo

mu = modelo.ln.mtp.cv$coefficients$coef.time[1]
sigma = exp(modelo.ln.mtp.cv$coefficients$coef.time[2])
etaln = modelo.ln.mtp.cv$coefficients$coef.cure[1]+
ovarian$age*modelo.ln.mtp.cv$coefficients$coef.cure[2]

#construção do resíduo tipo Cox-Snell
thetaln = exp(etaln)/(1+exp(etaln))
y = ovarian$futime
s0ln = pnorm((-log(y)+mu)/sigma)
sptln = exp(-thetaln*(1 - s0ln))
h1<- -log(sptln)
indicadora<- ovarian$fustat

x<- survfit(Surv(h1,indicadora)~1, conf.type = "none")
resultado_time <- x$time
resultado_surv <- x$surv

#construção do gráfico
#par(mfrow = c(1,1))
matplot(resultado_time, -log(resultado_surv), type="p",
col="black", pch=1,
xlab = "Resíduos de Cox Snell", ylab = "Função de Risco Acumulado",
xlim = c(0,1.5), ylim = c(0,1.5))
abline(c(0,3),c(1,0), col = "red")
title("Resíduo Global do Tipo Cox-Snell \n
para o MTP-Log-normal com a covariável Age")

#medidas de ajuste e LRT

#aic

```

AIC(modelo.ln.mmp.cv,modelo.ln.mtp.cv,modelo.wei.mmp.cv,modelo.wei.cv)