UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Um novo modelo de sobrevivência Bell-Inversa Gaussiana com fração de cura

Renata Cristina Carregari

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Renata Cristina Carregari

Um novo modelo de sobrevivência Bell-Inversa Gaussiana com fração de cura

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
Maio de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C314n

Carregari, Renata Cristina
Um novo modelo de sobrevivência Bell-Inversa Gaussiana com fração de cura. / Renata Cristina Carregari; orientador Adriano Kamimura Suzuki. -- São Carlos, 2021.
90 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Análise de sobrevivência com fração de cura. 2. Modelo longa duração. 3. Distribuição Bell. 4. Distribuição Inversa Gaussiana. I. Kamimura Suzuki, Adriano , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Renata Cristina Carregari, realizada em 26/03/2021.

Comissão Julgadora:

Prof. Dr. Adriano Kamimura Suzuki (USP)

Prof. Dr. Paulo Henrique Ferreira da Silva (UFBA)

Prof. Dr. Cynthia Arantes Vieira Tojeiro (UFG)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Renata Cristina Carregari

A new Bell Inverse Gaussian cure rate survival model

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
May 2021

Dedico este trabalho ao meu pai, Orlando em sua memória, por ter desde sempre depositado fé e esperança em quem eu poderia me tornar e por ter me proporcionado as melhores oportunidades sempre. Dedico também este trabalho à minha mãe, que sempre batalhou para me dar tudo de melhor. Por sua paciência, amor e doação.

AGRADECIMENTOS

À Deus primeiramente por ser o que me mantém em pé todos os dias de minha vida.

Aos meus familiares, especialmente minha mãe, Eva que me deu apoio e forças quando tudo parecia não fazer mais sentido, ao meu irmão, Tiago que me apoiou e me estendeu a mão todos os dias, à minha cunhada, Cinthia que sempre me motivou e incentivou.

Às minhas amigas, Mariane e Joice, pela alegria e companheirismo.

Ao meu namorado e parceiro de vida, Rafael, por sempre me apoiar e trazer esperança para os meus dias.

Ao meu orientador, Adriano, por todos ensinamentos e paciência no decorrer do desenvolvimento deste trabalho.

Aos professores da USP e da UFSCar pelo empenho e ensinamentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal do Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes.”

(Marthin Luther King)

RESUMO

CARREGARI, R. C. **Um novo modelo de sobrevivência Bell-Inversa Gaussiana com fração de cura**. 2021. 90 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Neste trabalho propomos um novo modelo de sobrevivência denominado Bell-Inversa Gaussiana com fração de cura. Consideramos diferentes esquemas de ativação em que o número de fatores M tem a distribuição Bell e o tempo de ocorrência de um evento segue o modelo Inversa Gaussiana. Os parâmetros são estimados pelos métodos clássico e Bayesiano. Em um estudo de simulação, investigamos as médias das estimativas, os vieses, os erros quadráticos médios e as probabilidades de cobertura nos diferentes esquemas de ativação. Com o objetivo de detectar possíveis observações influentes ou extremas que podem causar distorções nos resultados da análise, utilizamos o método Bayesiano de análise de influência de deleção de casos baseado na divergência ψ. Por fim, mostramos a aplicabilidade do modelo proposto a um conjunto de dados reais.

Palavras-chave: Análise de sobrevivência; Distribuição Bell; Distribuição Inversa Gaussiana; Esquema de ativação latente; Modelo de sobrevivência com fração de cura.

ABSTRACT

CARREGARI, R. C. **A new Bell Inverse Gaussian cure rate survival model**. 2021. 90 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

In this work we propose a new survival model called the Bell-Inverse Gaussian cure rate. We consider different activation schemes in which the number of factors M has the Bell distribution and the time of occurrence of an event follows the Inverse Gaussian model. The parameters are estimated by the classical and Bayesian methods. In a simulation study, we investigate the mean estimates, biases, mean squared errors and coverage probabilities in different activation schemes. In order to detect possible influential or extreme observations that can cause distortions on the results of the analysis we use the Bayesian method of influence analysis of case deletion based on ψ-divergence. Finally, we show the applicability of the proposed model to a real dataset.

Keywords: Bell distribution; Cure rate survival model; Inverse Gaussian distribution; Latent activation schemes; Survival analysis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Em estudos de análise de sobrevivência, a variável reposta é o tempo até a ocorrência de um determinado evento de interesse cujas aplicações podem ser encontradas nas mais diferentes áreas do conhecimento, tais como: a Biologia, a Economia, as Engenharias e a Medicina. Este tempo é denominado tempo de falha, podendo ser o tempo até a falha de um equipamento, a morte de um indivíduo que possui uma doença ou até a recidiva da mesma, entre outros. Dados de análise de sobrevivência têm duas características principais: a presença de censura e o tempo de falha.

Em análise de sobrevivência padrão, a função de sobrevivência converge para zero quando o tempo tende a infinito, isto é, consideramos que todos os indivíduos em estudos são susceptíveis ao evento de interesse.

No entanto, em alguns estudos clínicos, os dados de tempo de falha de uma população são constituídos de dois grupos: os indivíduos que são susceptíveis e os não-susceptíveis ao evento de interesse. Os indivíduos do primeiro grupo podem eventualmente sofrer o evento de interesse enquanto os do segundo aparentam estarem livres de sinais do mesmo, e são denominados imunes ou curados.

No conjunto de dados de sobrevivência, a presença de uma grande quantidade de observações das quais os maiores tempos são censurados pode indicar que existe uma fração (proporção) de indivíduos curados na população em estudo. Caso o gráfico do estimador de Kaplan-Meier da função de sobrevivência apresente uma cauda direita em um nível aproximadamente constante e estritamente maior que zero para um período considerável, as observações censuradas determinam uma possível presença de uma fração de curados.

Um exemplo de possibilidade de cura é um conjunto de dados clínicos em que o evento de interesse é a recorrência de um determinado tipo de câncer após a administração de um tratamento, sendo que alguns dos pacientes podem não sofrer a recidiva.

Os modelos de sobrevivência com fração de cura, também conhecidos como modelos de longa duração (RODRIGUES; CANCHO; CASTRO, 2008; IBRAHIM; CHEN; SINHA, 2001), têm sido amplamente desenvolvidos e utilizados em várias aplicações, como em dados de estudos clínicos (KIRKWOOD *et al.*, 2004). Os principais objetivos do estudo é o estudo das funções de sobrevivência, o estudo da fração de curados, o efeito do tratamento sobre esta fração e também das covariáveis envolvidas nestes estudos.

Na literatura, nos últimos anos, têm surgido vários trabalhos que propõem novos modelos que contemplam a possibilidade de fração de curados. Alguns exemplos desta extensa literatura e em rápida evolução são: Cordeiro *et al.* (2016), Yiqi *et al.* (2016), Balakrishnan, Barui e Milienos (2017), Bao *et al.* (2017), Gallardo *et al.* (2017), Koutras e Milienos (2017), Ortega *et al.* (2017a), Ortega *et al.* (2017b), Rocha *et al.* (2017), Souza *et al.* (2017), Balakrishnan, Koutras e Milienos (2018), Cancho *et al.* (2018), Onicescu e Lawson (2018), Bao *et al.* (2019), Calsavara *et al.* (2019), He e Emura (2019), Bao *et al.* (2020), Cancho *et al.* (2020), Karamoozian, Baneshi e Bahrampour (2020), Ramires *et al.* (2020), Rodrigues *et al.* (2020), Silva, Cordeiro e Ortega (2020), Barriga *et al.* (2021), Cancho *et al.* (2021), Pal, Mondal e Kundu (2021) e Narisetty e Koenker (2021).

Sob uma abordagem que leva em conta uma sequência estocástica de causas latentes, Cooner *et al.* (2007) induzem a ocorrência do evento de interesse por meio de mecanismos de ativação latentes. Vários são os trabalhos sobre essas novas distribuições que levam em consideração diferentes mecanismos de ativação com causas latentes competitivas (TSODIKOV; IBRAHIM; YAKOVLEV, 2003), como por exemplo: Louzada, Cancho e Barriga Gladys D (2012), Cancho, Castro e Dey (2013), Louzada, Cancho e Yiqi (2015), Suzuki, Cancho e Louzada (2016), Suzuki *et al.* (2017), Gilavert, Suzuki e Saraiva (2018), Pescim *et al.* (2019) e Barriga *et al.* (2020).

Neste trabalho será levado em consideração os mecanismos de ativações latentes proposto por Cooner *et al.* (2007). O objetivo é propor um novo modelo para análise de dados de sobrevivência de longa duração denominado Bell-Inversa Gaussiana com proporção de cura (BIGcr). Assumimos que há diferentes mecanismos de ativação com causas latentes para o evento de interesse, cujo o número dessas causas é modelado por uma distribuição Bell (CASTELLARES; FERRARI; LEMONTE, 2018) e o tempo até a ocorrência do evento de interesse segue uma distribuição Inversa Gaussiana (IG).

Para estimação dos parâmetros, consideramos duas abordagens: uma clássica que foi realizada utilizando métodos de convergência e a Bayesiana com distribuições a *priori* não informativas, *i.e.*, com "grandes" variâncias. Como as distribuições a *posteriori* condicionais dos parâmetros não possuem forma analítica "fechada", obtemos as estimativas por meio do uso de métodos de Monte Carlo via Cadeias de Markov (MCMC).

Desenvolvemos um estudo de simulação para verificar o ajuste do modelo a conjuntos de dados artificiais, gerados sob três esquemas de ativação (primeiro, último e aleatório) do

evento de interesse. No estudo de simulação, também consideramos conjuntos de dados com a presença de dados perturbados. O interesse em conjuntos de dados com esta característica é que observações perturbadas podem ser influentes e causarem distorções nas estimativas dos parâmetros. Para estes casos, utilizamos o diagnóstico de influência de deleção de casos com base na medida de divergência ψ (PENG; DEY, 1995; WEISS, 1996) para identificar as observações influentes.

A divergência ψ inclui várias medidas de divergência como casos particulares, tais como: a divergência Kullback-Leibler, a distância J e a distância L_1, as quais utilizamos neste trabalho.

Como ilustração, ajustamos o modelo proposto a um conjunto de dados reais referente a um ensaio clínico de melanoma cutâneo de fase III, baixado do website <http://merlot.stat.uconn.edu/~mhchen/survbook/> (KIRKWOOD *et al.*, 2000)

Todas as implementações computacionais foram feitas utilizando o *software* R (R Development Core Team, 2018). Para a inferência Bayesiana fez-se o uso do sistema JAGS - Just Another Gibbs Sampler (PLUMMER *et al.*, 2006) por meio do pacote *runjags* (DENWOOD, 2016).

1.1 Organização do Trabalho

O Capítulo 2 apresenta uma breve revisão de alguns conceitos básicos de Análise de Sobrevivência bem como o conceito de ativação latente proposto por Cooner *et al.* (2007), Além disso, introduz a distribuição Bell e o modelo Inversa Gaussiana que serão utilizados para a construção do modelo proposto.

O Capítulo 3 apresenta a construção do modelo Bell Inversa Gaussiana com fração de cura (BIG$_{cr}$), sob os três esquemas de ativação latente. Também, por meio de uma abordagem clássica, realizamos um estudo de simulação e uma aplicação a dados reais.

No Capítulo 4 descrevemos uma abordagem Bayesiana e o procedimento de estimação dos parâmetros do modelo BIGcr. Também é apresentada uma análise de influência de deleção de casos baseada na divergência ψ bem como um estudo de simulação e aplicação a dados reais.

Por fim, no Capítulo 5 apresentamos as considerações finais juntamente com as nossas perspectivas de trabalhos futuros.

CAPÍTULO

2

ALGUNS CONCEITOS EM ANÁLISE DE SOBREVIVÊNCIA

2.1 Análise de Sobrevivência

Neste capítulo apresentamos alguns conceitos básicos utilizados em Análise de Sobrevivência e uma breve revisão sobre os modelos de sobrevivência de longa duração que foram utilizados no desenvolvimento do trabalho.

2.1.1 Tempo de falha

Em Análise de Sobrevivência, a variável reposta é o tempo até a ocorrência de um evento de interesse que é denominado **tempo de falha**. Podemos definir tempo de falha em várias áreas de estudo, como por exemplo na engenharia, o tempo de falha pode ser o tempo que um equipamento leva para apresentar algum defeito. Já na medicina, pode ser o tempo de morte ou recidiva de uma doença em um indivíduo.

Dados de sobrevivência têm duas características principais: a presença de dados censurados (censura) e o tempo de falha. O tempo de falha é definido como o tempo até a ocorrência do evento de interesse, e é constituído do evento de interesse (a falha), o tempo inicial e a escala de medida. Se por algum motivo o acompanhamento com o paciente é interrompido tem-se a ocorrência de censura (COLOSIMO; GIOLO, 2006).

O tempo de falha é definido pelo tempo inicial, a escala de medida e o evento de interesse. O tempo inicial do estudo deve ser definido com precisão. Por exemplo, na área médica pode-se considerar a data do início de um determinado tratamento ou do diagnóstico da doença. A escala de medida depende do problema em estudo, por exemplo na área médica, podendo ser o tempo real (em horas, dias, etc.), nas engenharias, pode ser dada pelo número de ciclos, de quilometragem de um veículo automotivo, entre outros.

O tempo de falha precisa ser definido de forma clara e precisa. Em algumas situações, o evento de interesse é simples de ser diagnosticado, tais como morte ou recidiva de uma doença. No entanto, por vezes pode ser mais complexo de ser demarcado, como por exemplo, saber quando um determinado produto alimentício fica inapropriado para o consumo.

Em Análise de sobrevivência temos um tipo de evento de interesse bem comum denominado eventos recorrentes, que acontecem mais de uma vez para um mesmo indivíduo ou equipamento. Por exemplo: gestações, cáries e danificação de equipamentos que podem ser consertados. Temos também os diferentes tipos de eventos decorrentes de um mesmo fator de risco em estudo, como efeitos adversos de medicamentos.

2.1.2 Censura

Em estudos clínicos muitas vezes o experimento termina antes que o evento de interesse de fato ocorra, então, pode-se observar por vezes estudos incompletos ou parciais. Tais observações são denominadas censuras podendo acontecer por inúmeros fatores, dentre eles a perda de acompanhamento do paciente no decorrer do estudo ou até mesmo a não-ocorrência do evento de interesse. Outro fator de censura é quando a falha ocorre por outras causas que não é a esperada no estudo.

Segundo Colosimo e Giolo (2006), podemos classificar a censura em:

- **Tipo I**: O estudo é terminado após um período de tempo já determinado no início do estudo, os indivíduos que não tiveram a falha são censurados.
- **Tipo II**: O estudo é terminado após um número pré-determinado de indivíduos atingir o evento de interesse.
- **Aleatória**: semelhante à censura do Tipo I, porém com os indivíduos sendo incorporados de maneira aleatória.

2.1.3 Função de Sobrevivência e Função de Risco

Seja Y uma variável aleatória contínua não-negativa com função densidade de probabilidade $f(y)$, descrevendo os tempos de vida de indivíduos de uma população. Em Análise de Sobrevivência a variável aleatória Y é geralmente especificada pela sua função de sobrevivência ou pela função taxa de falha. A probabilidade de um indivíduo sobreviver à um determinado tempo $S(y)$, ou seja, a probabilidade de um indivíduo viver um tempo maior que y é dada por: (COLOSIMO; GIOLO, 2006)

$$S(y) = P(Y > y) = \int_y^{+\infty} f(u)du.$$

Nota-se que $S(y)$ é monótona e decrescente, em que $S(0) = 1$ e $S(\infty) = \lim_{y\to\infty} S(y) = 0$.

Consequentemente, a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver até o tempo y, isto é, $F(y) = 1 - S(y)$. Logo, a probabilidade de ocorrer uma falha no intervalo $[y_1, y_2)$ é dada por:

$$P(y_1 \leq Y < y_2) = F(y_2) - F(y_1) = (1 - S(y_2)) - (1 - S(y_1)) = S(y_1) - S(y_2).$$

A taxa de falha no intervalo $[y_1, y_2)$ é definida como sendo a probabilidade de que a falha ocorra neste intervalo dado que não ocorreu antes de y_1 dividida pelo comprimento do mesmo (COLOSIMO; GIOLO, 2006), isto é,

$$\frac{P(Y \in [y_1, y_2) | Y \geq y_1)}{(y_2 - y_1)} = \frac{P(Y \in [y_1, y_2))}{(y_2 - y_1)P(Y \geq y_1)} = \frac{S(y_1) - S(y_2)}{\Delta(y)S(y_1)}, \tag{2.1}$$

em que $\Delta(y) = y_2 - y_1$.

Redefinindo o intervalo como $[y, y + \Delta y)$, a expressão (2.1) fica da forma:

$$\frac{S(y) - S(y + \Delta y)}{\Delta y S(y)}. \tag{2.2}$$

Fazendo Δy um valor baixo em (2.2), a função de risco $h(t)$ representa a taxa de falha instantânea no tempo y condicional à sobrevivência até o tempo y, que é dada por:

$$\begin{aligned} h(y) &= \lim_{\Delta y \to 0} \frac{P(y \leq Y < y + \Delta y | Y \geq y)}{\Delta y} \\ &= \frac{1}{S(y)} \lim_{\Delta y \to 0} \frac{F(y + \Delta y) - F(y)}{\Delta y} \\ &= \frac{F'(y)}{S(y)} = \frac{f(y)}{S(y)}. \end{aligned}$$

Uma outra expressão da função de risco é obtida da seguinte forma:

$$h(y) = \frac{F'(y)}{S(y)} = \frac{1}{S(y)} \frac{dF(y)}{dy} = \frac{1}{S(y)} \frac{d(1 - S(y))}{dy} = -\frac{1}{S(y)} \frac{dS(y)}{dy}.$$

Portanto,

$$h(y) = -\frac{d \log(S(y))}{dy}.$$

As taxas de falha são números positivos sem limite superior, podendo ser crescente, decrescente, constante ou em forma de banheira. Para maiores detalhes ver Colosimo e Giolo (2006).

2.1.4 Estimador de Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier (KAPLAN; MEIER, 1958), também chamado de estimador limite-produto é muito utilizado em Análise de Sobrevivência para estimação da função de sobrevivência. É uma adaptação da função de sobrevivência empírica que na ausência de censuras é definido como (COLOSIMO; GIOLO, 2006):

$$\widehat{S}(y) = \frac{\text{nº de observações que não falharam até o tempo } y}{\text{nº total de observações no estudo}}.$$

A função $\widehat{S}(y)$ é uma função do tipo escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Se existirem empates em um certo tempo y o tamanho do degrau fica multiplicado pelo número de empates.

Para obtermos a expressão geral do estimador de Kaplan-Meier, considere:

- $y_1 < y_2 < ... < y_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em $y_j,\ j = 1,...,k$;
- n_j o número de indivíduos sob risco em y_j, ou seja, os indivíduos que sobreviveram e não foram censurados até o instante imediatamente anterior a y_j.

O estimador de Kaplan-Meier é definido como:

$$\widehat{S}(y) = \prod_{j:y_j<y} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:y_j<y} \left(1 - \frac{d_j}{n_j}\right).$$

2.2 Modelos de longa duração

Uma forma de modelar dados de sobrevivência com fração de cura é considerar uma função de sobrevivência como uma mistura de pacientes curados e não-curados.

O modelo mais conhecido é o modelo paramétrico de mistura padrão proposto por Berkson e Gage (1952). Neste modelo é assumida uma probabilidade π para a população que está "curada"e $1-\pi$ para a população restante não-curada. A função de sobrevivência para toda a população no instante t ($S_1(t)$) é dada por:

$$S_1(t) = \pi + (1-\pi)S^*(t), \tag{2.3}$$

em que $S^*(t)$ representa a função de sobrevivência própria (ou seja, a integral no intervalo $[0,\infty)$ da função de densidade de probabilidade associada a esta função de sobrevivência tem valor igual a 1) associada aos indivíduos não-curados.

Uma outra forma de modelagem é baseada na estrutura de risco competitivo. Nestes modelos é definida uma variável aleatória como o número de causas competindo para a ocorrência do evento de interesse que está associado à variável resposta.

Por exemplo, temos o modelo proposto por Yakovlev *et al.* (1993), em que a função de sobrevivência é apresentada da seguinte forma:

$$S(y) = \exp\{-\theta F(y)\}, \tag{2.4}$$

em que o número de causas tem distribuição de Poisson com média θ. Uma característica importante é que quando o parâmetro θ é definido como uma função de covariáveis faz com que o modelo (2.4) tenha uma estrutura de risco proporcional, o que não ocorre no modelo de mistura dado em (2.3).

Rodrigues, Cancho e Castro (2008) propõem uma abordagem unificada que consiste em uma generalização que possibilita a obtenção da função de sobrevivência com fração de cura da população de qualquer modelo baseado em uma distribuição genérica do número de causas de ocorrência do evento por meio do uso da função geradora de sequência de números reais definida por Feller (1957). Esta abordagem considera um modelo dividido em dois estágios: o primeiro estágio é o estágio de iniciação, no qual definimos uma variável aleatória M, em que M representa o número de causas ou riscos de ocorrência de um particular evento de interesse (morte de um paciente, reincidência de um câncer, etc), com a função de probabilidade dada por:

$$p_m = P(M = m),\ m = 0, 1, 2, \ldots.$$

O segundo estágio é denominado estágio de maturação. Dado $M = m$, sejam $Z_j, j = 1, \ldots, m$, variáveis aleatórias contínuas (não-negativas) independentes com função de distribuição acumulada $F(z) = 1 - S(z)$ e independentes de M, representando o tempo de ocorrência do evento de interesse devido à j-ésima causa ou risco.

No cenário de competição, apenas o menor tempo de vida Z_j entre todas as causas é observado. Assim, com a suposição de inclusão de indivíduos não-susceptíveis ao evento de interesse, o tempo Y para a ocorrência do evento de interesse é definido como:

$$Y = \min\{Z_0, Z_1, \ldots, Z_M\}, \tag{2.5}$$

em que $P(Z_0 = +\infty) = 1$, sendo assim uma proporção p_0 da população não sujeita à ocorrência do evento de interesse.

Definição (Função geradora de sequência de números reais (FELLER, 1957)): Seja $\{a_m\}$ uma sequência de números reais. Se

$$A(s) = a_0 + a_1 s + a_2 s^2 + \ldots \tag{2.6}$$

converge para valores de s no intervalo $[0, 1]$, então $A(s)$ é definida como função geradora da sequência $\{a_m\}$.

Rodrigues, Cancho e Castro (2008) unificaram a teoria de análise de sobrevivência de longa duração, sendo possível obter a função de sobrevivência de longa duração $S_{pop}(y) = P(Y > y)$ da variável aleatória Y (dada em 2.5), apenas pelo conhecimento da função geradora de sequência $A_p(.)$ da variável latente M relacionada a $S(y)$.

Utilizando a definição anterior, Rodrigues, Cancho e Castro (2008) propuseram o seguinte Teorema:

Teorema.(RODRIGUES; CANCHO; CASTRO, 2008) Dada uma função de sobrevivência própria $S(y)$, a função de sobrevivência da variável aleatória Y dada em (2.5) pode se escrita como

$$S_{pop} = A_p[S(y)] = \sum_{m=0}^{+\infty} p_m[S(y)]^m, \tag{2.7}$$

em que $A_p(.)$ é a função geradora da sequência $\{p_m\}$ que converge no intervalo $0 \leq S(y) \leq 1$.
Demostração: A demonstração pode ser encontrada em Rodrigues, Cancho e Castro (2008).

Se $a = \{a_m\} = \{p_m\} = p$ então, $A_p(s) = E[s^M] = E[\exp(M\log(s))]$, isto é, $A_p(s)$ é função geradora de momentos de M calculada no ponto $\log(s)$.

Como observado por Rodrigues, Cancho e Castro (2008), a função de sobrevivência (2.7) não é própria, isto é,

$$\lim_{y\to+\infty} S_{pop}(y) > 0.$$

Além disso, a proporção p_0 de não-ocorrência do evento na população é dada por:

$$\lim_{y\to+\infty} S_{pop}(y) = P(M = 0) = p_0.$$

2.3 Processo carcinogênico

O processo carcinogênico em um indivíduo ocorre quando uma célula que já fora "iniciada" por algum fator externo como o fumo, a exposição exacerbada ao sol, dentre outros, se transforma em célula cancerígena, então esta se multiplica de forma desorganizada e assim, ocasiona a formação de um tumor. Alguns fatores devem ser levados em conta, pois dentro de cada indivíduo há a presença de milhões de células, sendo assim, não haveria a possibilidade de determinar o momento exato em que uma célula "iniciada" se transformou em célula cancerígena. Alguns autores como Yiqi *et al.* (2016), Suzuki *et al.* (2017), Gilavert, Suzuki e Saraiva (2018) e Barriga *et al.* (2020) realizaram estudos que levaram em conta os mecanismos de ativações propostos por Cooner *et al.* (2007), que utiliza três esquemas de ativações latentes considerando que as células estejam competindo entre si para de fato ocorrer o evento de interesse. Os esquemas são denominados: esquema de primeira ativação, ativação aleatória e última ativação. Para representar o tempo de ativação, assumimos que são necessárias a ocorrência de R das M causas, $R \leq M$, para que o evento de interesse ocorra. Assim, considerando $Z_{(1)} \leq \cdots \leq Z_{(R)} \leq \cdots \leq Z_{(M)}$ como sendo as M estatísticas de ordem, com $Z_{(R)}$ sendo a R-ésima estatística de ordem, então

o tempo até a ocorrência do evento de interesse é dado por $Y = Z_{(R)}$. O valor R pode ser uma constante fixa, uma função de M ou uma variável aleatória discreta com distribuição condicional em M. Se $M = 0$, o tempo para evento é assumido como infinito ($Y = +\infty$) com $P(Y = +\infty|M = 0) = 1$.

2.4 Mecanismos de ativação

Nesta seção apresentamos os três mecanismos de ativação: esquema de primeira ativação, última ativação e ativação aleatória.

2.4.1 Esquema de primeira ativação

Neste esquema, o evento de interesse ocorre devido a qualquer uma das possíveis causas. Neste cenário, temos que $R = 1$ e o tempo para o evento é dado por $Y = Z_{(1)} = min\{Z_1, Z_2, \ldots, Z_M\}$.

Assim, a função de sobrevivência populacional é dada por:

$$S_{pop}(y) = P(M = 0) + P(Z_1 > y, Z_2 > y, ..., Z_M > y, M \geq 1).$$

Após algumas álgebras, pode ser mostrado que a função de sobrevivência de Y fica da forma:

$$S_{pop} = A_M(S(y)),$$

em que A_M é a função geradora de probabilidade de M.

Logo, a função densidade é dada por $f_{pop}(y) = -S'_{pop}(y)$, tal que $f(y) = -S'(y)$ é a função de sobrevivência própria do tempo para o evento Z e a derivada negativa da função de sobrevivência é dada por:

$$-\frac{dS_{pop}(y)}{dy} = f(y) \sum_{m=1}^{\infty} m p_m \{S(y)\}^{m-1}. \tag{2.8}$$

2.4.2 Esquema de última ativação

No esquema denominado de última ativação, o evento de interesse ocorre somente após a ocorrência de todas as M causas. Neste caso, temos $R = M$ e o tempo de vida observado é dado por $Y = Z_{(M)} = \max\{Z_1, \ldots, Z_M\}$. A função de sobrevivência de Y é dada por:

$$S_{pop}(y) = 1 + A_M(0) - A_M(F(y)). \tag{2.9}$$

2.4.3 Esquema de ativação aleatória

No esquema de ativação aleatória, o evento de interesse ocorre após quaisquer M causas terem ocorrido, ou seja, $M \geq 1$, a distribuição condicional de R é Uniforme discreta em $\{1,2,...,M\}$, com função de sobrevivência populacional dada pela relação:

$$S_{pop} = P(Y \leq y) = P(M = 0) + \{1 - P(M = 0)\}S(y). \tag{2.10}$$

2.5 A distribuição Bell

Muitos problemas envolvendo dados de contagem ocorrem na prática como por exemplo o estudo do número de tempestades durante o ano, o número de acidentes de carro em um dado período de tempo, o número de sinistros de um veículo automotivo, dentre outros.

Um modelo probabilístico discreto muito utilizado na prática é a distribuição uniparamétrica de Poisson. A grande desvantagem do seu uso é que não acomoda sub e superdispersão devido ao fato da média ser igual à variância. Na literatura encontramos vários trabalhos propondo novas distribuições discretas por meio da discretização de uma variável contínua, como: Weibull discreta (Nakagawa; Osaki, 1975; Stein; Dattero, 1984), Gama discreta (YANG, 1994), Rayleigh discreta (Roy, 2004), Burr e Pareto discretas (KRISHNA; Singh Pundir, 2009), inversa Weibull discreta (JAZI; LAI; ALAMATSAZ, 2010), Lindley ponderada discreta (GHITANY *et al.*, 2011), Exponencial ponderada discreta (GUPTA; KUNDU, 2009), entre outras. No entanto, essas distribuições possuem mais de um parâmetro.

Castellares, Ferrari e Lemonte (2018) propuseram uma nova distribuição discreta uniparamétrica denominada distribuição de Bell, que foi obtida com base em uma expansão em série devido a Bell (BELL, 1934a; BELL, 1934b). Esta distribuição possui uma função massa de probabilidade de fácil manipulação. Além disso, possui várias propriedades interessantes como média, variância, função geradora de probabilidade e estimador de máxima verossimilhança sendo que todos possuem forma analítica fechada. É apresentada uma outra reparametrização em que a função massa de probabilidade pertence à família exponencial uniparamétrica em função da média da distribuição.

Seja M uma variável aleatória discreta com distribuição Bell de parâmetro $\theta > 0$, isto é, $M \sim Bell(\theta)$. A sua função massa de probabilidade é dada por:

$$P(M = m) = \frac{\theta^m e^{-e^\theta + 1} B_m}{m!},\ m = 0,1,2,\ldots, \tag{2.11}$$

em que os coeficientes B_m são números Bell, definidos por:

$$B_m = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^m}{k!}. \tag{2.12}$$

O número B_m dado na equação (2.12) é o n-ésimo momento da distribuição de Poisson com parâmetro igual a 1.

A função massa de probabilidade dada em (2.11) pode ser expressa da forma:

$$Pr(M=m) = \exp\left(m\log(\theta) + \log\left(\frac{B_m}{m!}\right) - e^{\theta} + 1\right) = \exp(\xi T(m) - A(\theta) + C(m)),$$

em que $m = 0,1,2,\ldots$. Assim, temos que a distribuição Bell pertence à família exponencial uniparamétrica com parâmetros naturais $\xi = \log(\theta), T(m) = m, A(\theta) = e^{\theta}$ e $C(m) = \log\left(\frac{B^m}{m!}\right) + 1$.

Adicionalmente, se $M_1, M_2, \ldots, M_n$ é uma amostra aleatória de tamanho n de $M \sim Bell(\theta)$, então $T(M_1, M_2, \ldots, M_n) = \sum_{i=1}^{n} M_i$ é uma estatística suficiente para θ.

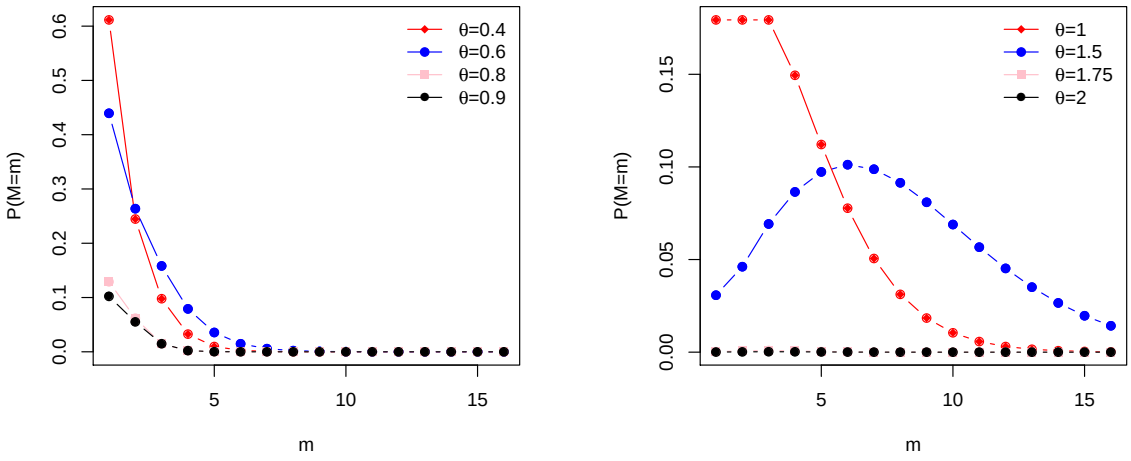
2.5.1 Algumas propriedades

Se $M \sim Bell(\theta)$, a sua função massa de probabilidade dada pela equação (2.11) é muito simples de se trabalhar e não envolve nenhuma função complicada. Por exemplo, temos:

$$P(M=0) = e^{-e^{\theta}+1}, P(M=1) = \theta e^{-e^{\theta}+1},$$
$$P(M=2) = \theta^2 e^{-e^{\theta}+1}, P(M=3) = \tfrac{5\theta^3}{3!} e^{-e^{\theta}+1},$$
$$P(M=4) = \tfrac{15\theta^4}{4!} e^{-e^{\theta}+1} \text{ e } P(M=5) = \tfrac{52\theta^4}{5!} e^{-e^{\theta}+1}.$$

As outras probabilidades podem ser facilmente obtidas. Na Figura 1 apresentamos a representação gráfica da distribuição Bell para diferentes valores do parâmetro θ.

Figura 1 – Gráficos da função massa de probabilidade da distribuição Bell para valores arbitrários do parâmetro θ.



Fonte: Elaborada pela autora.

Uma tabulação da função massa de probabilidade para $\theta = 0,05$, $0,1$, $0,2$, $0,3$, $0,4$, $0,5$, $0,6$, $0,7$, $0,8$, $0,9$, $1,0$, $1,25$, $1,5$, $1,75$ e 2 é apresentada na Tabela 1.

Tabela 1 – Probabilidades de $M \sim Bell(\theta)$

Parâmetro	$P(M=0)$	$P(M=1)$	$P(M=2)$	$P(M=3)$	$P(M=4)$	$P(M=5)$	$P(M>5)$
$\theta = 0,05$	0,9500	0,0475	0,0024	$9,8960 * 10^{-5}$	$3,7110 * 10^{-6}$	$1,2865 * 10^{-7}$	$4,3183 * 10^{-9}$
$\theta = 0,10$	0,9002	0,0900	0,0090	0,0007	$5,6260 * 10^{-5}$	$3,9007 * 10^{-6}$	$2,7044 * 10^{-7}$
$\theta = 0,20$	0,8013	0,1603	0,0320	0,0053	0,0008	0,0001	$1,6483 * 10^{-5}$
$\theta = 0,30$	0,7047	0,2114	0,0634	0,0158	0,0035	0,0007	0,0002
$\theta = 0,40$	0,6115	0,2446	0,0978	0,0326	0,0097	0,0027	0,0009
$\theta = 0,50$	0,5227	0,2613	0,1307	0,0544	0,0204	0,0071	0,0033
$\theta = 0,60$	0,4394	0,2637	0,1582	0,0791	0,0356	0,0148	0,0741
$\theta = 0,70$	0,3628	0,2531	0,1778	0,1037	0,0544	0,0264	0,0207
$\theta = 0,80$	0,2935	0,2349	0,1879	0,1253	0,0752	0,0417	0,0415
$\theta = 0,90$	0,2323	0.2090	0,1882	0,1411	0,0953	0,0594	0,0745
$\theta = 1,00$	0,1794	0,1794	0,1794	0,1495	0,1121	0,0777	0,1226
$\theta = 1,25$	0,0829	0,1036	0,1295	0,1349	0,1265	0,1096	0,3130
$\theta = 1,50$	0,0308	0,0461	0,0692	0,0865	0,0973	0,1012	0,5689
$\theta = 1,75$	0,0086	0,0150	0,0264	0,0385	0,0505	0,0612	0,7997
$\theta = 2,00$	0.0017	0.0034	0,0067	0,0112	0,0168	0,0233	0,9369

2.5.2 Esperança, variância e função geradora de probabilidades

Seja $M \sim Bell(\theta)$, em que $\theta > 0$. Então, sua função geradora de probabilidades é dada por:

$$A_M(s) = E(s^M) = \exp(e^{s\theta} - e^{\theta}),\ |s| < 1. \tag{2.13}$$

O valor esperado e a variância de $M \sim Bell(\theta)$ são dados por:

$$E(M) = \theta e^{\theta} \tag{2.14}$$

e

$$Var(M) = \theta(1+\theta)e^{\theta}, \tag{2.15}$$

respectivamente.

Na Tabela 2 apresentamos o valor esperado e a variância da distribuição Bell, para os mesmo valores de θ assumidos na Tabela 1. Observa-se que para maiores valores de θ, o seu valor esperado e variância aumentam.

Tabela 2 – Esperança e variância

Parâmetro	$E(M)$	$Var(M)$
$\theta = 0,05$	0,0101	0,0102
$\theta = 0,10$	0,1105	0,1216
$\theta = 0,20$	0,2443	0,2931
$\theta = 0,30$	0,4049	0,5264
$\theta = 0,40$	0,5967	0,8354
$\theta = 0,50$	0,8244	1,2365
$\theta = 0,60$	1,0932	1,7492
$\theta = 0,70$	1,4096	2,3964
$\theta = 0,80$	1,7804	3,2048
$\theta = 0,90$	2,2136	4.2059
$\theta = 1,00$	6,7225	5,4366
$\theta = 1,25$	4,3629	9,8166
$\theta = 1,50$	6,7225	16,8063
$\theta = 1,75$	10,0705	27,6940
$\theta = 2,00$	14,7781	44,3343

2.5.3 Outras propriedades

Seja $M \sim Bell(\theta)$, em que $\theta > 0$. Então, a variável aleatória M tem a mesma distribuição que a soma de N distribuições de Poisson zero truncadas independentes, com parâmetro $\theta > 0$ e $N \sim Poisson(e^{\theta} - 1)$.

Com base nesta propriedade, podemos construir a seguinte caracterização: Seja $X_1, X_2, \ldots$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, em que X_n tem distribuição de Poisson zero truncada de parâmetro $\theta > 0$ e N possui distribuição de Poisson com parâmetro $e^{\theta} - 1$ e independente da sequência $\{X_n, n \geq 1\}$. Então, a variável aleatória M dada por $M = X_1 + X_2 + \cdots + X_N$ tem distribuição $Bell(\theta)$ (CASTELLARES; FERRARI; LEMONTE, 2018).

Além disso, a distribuição Bell é identificável para todo $\theta > 0$, é fortemente unimodal e é infinitamente divisível (para outras propriedades bem como maiores detalhes ver Castellares, Ferrari e Lemonte (2018).

2.5.4 Estimador de Máxima Verossimilhança de θ

Seja $M_1, M_2, \ldots, M_n$ uma amostra aleatória de $M \sim Bell(\theta)$. A função de verossimilhança baseada na amostra observada $m_1, m_2, \ldots, m_n$, é dada por:

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} \frac{\theta^{m_i} e^{-e^{\theta}+1} B_{m_i}}{m_i!}. \tag{2.16}$$

Aplicando o logaritmo natural na equação (2.16) obtemos que a função de log-verossimilhança é dada por:

$$l(\theta) \propto -me^{\theta} + log(\theta)\sum_{i=1}^{n} m_i.$$

O estimador de máxima verossimilhança $\hat{\theta}$ de θ satisfaz a equação:

$$-e^{\hat{\theta}} + \frac{\bar{M}}{\hat{\theta}} = 0,$$

ou, equivalentemente, $\bar{M} = \hat{\theta}\exp(\hat{\theta})$, cuja solução é $\hat{\theta} = W_0(\bar{M})$, em que $W_0(.)$ é a função Lambert (CORLESS *et al.*, 1996) e $\bar{M} = n^{-1}\sum_{i=1}^{n} M_i$.

2.5.5 Uma outra reparametrização do modelo Bell

Em modelos de regressão, tipicamente é feita a modelagem da média da variável resposta. Então, para obter uma estrutura de regressão para a média da distribuição Bell, temos que trabalhar com uma parametrização diferente da função massa de probabilidade dada em (2.11). Seja $\mu = \theta e^{\theta}$ e, portanto, $\theta = W_0(\mu)$, em que $W_0(.)$ é a função Lambert (CORLESS *et al.*, 1996). A partir das equações (2.14) e (2.15), temos que a média e a variância são dadas por:

$$E(M) = \mu \quad \text{e} \quad Var(M) = \mu[1 + W_0(\mu)],$$

respectivamente.

A função massa de probabilidade da distribuição Bell sob a nova parametrização é:

$$P(M = m) = \exp\left(1 - e^{W_0(\mu)}\right)\frac{W_0(\mu)^m B_m}{m!},\ m = 0, 1, 2\ldots, \tag{2.17}$$

em que $\mu > 0$, B_m são os números Bell dados em (2.12). Analogamente como visto anteriormente, podemos observar que a função massa de probabilidade (2.17) pertence à família exponencial uniparamétrica.

2.6 Distribuição Inversa Gaussiana

A distribuição Inversa Gaussiana, também chamada de distribuição Wald, foi proposta por Schrödinger (1915) como a distribuição do tempo de primeira passagem que um movimento Browniano leva para atingir um *drift* positivo.

O nome da distribuição Inversa Gaussiana foi dado por Tweedie (1957). Essa distribuição tem chamado a atenção na literatura com suas aplicações em diversas áreas do conhecimento,

bem como as suas extensões e generalizações. Algumas referências são os livros: Chhikara e Folks (1989), Seshadri (1993), Seshadri (1999) e Folks (2007) e, os artigos recentes: Balakrishna e Rahul (2014), Choi *et al.* (2014), Ye e Chen (2014), Peng (2015), Hanagal e Bhambure (2016), Gómez–Déniz e Pérez–Rodríguez (2017), Nagamani e Tripathy (2018), Hanagal e Pandey (2020), Chen *et al.* (2019), Ghitany *et al.* (2019), Jayalath e Chhikara (2020), Punzo (2019), Shamany, Alobaidi e Algamal (2019), Upadhyay e Sen (2019), Wen *et al.* (2019), Kinat, Amin e Mahmood (2020), Xu, Hu e Wang (2020) e Morita *et al.* (2021).

De acordo com Folks e Chhikara (1978) e Chhikara e Folks (1989), a interpretação da variável aleatória Gaussiana Inversa como um tempo de primeira passagem sugere suas potenciais aplicações para o estudo de tempos de vida (como em Bacanli e Demirhan (2008), Stogiannis e Caroni (2012), Gunes *et al.* (1997), Nikulin e Solev (1999) e Lemeshko *et al.* (2010)) e também para uma ampla gama de campos (veja Johnson, Kotz e Balakrishnan (1994)).

Se Y é uma variável aleatória com distribuição Inversa Gaussiana (IG) ($Y \sim IG(\mu,\lambda)$) então, a sua função densidade de probabilidade é da forma:

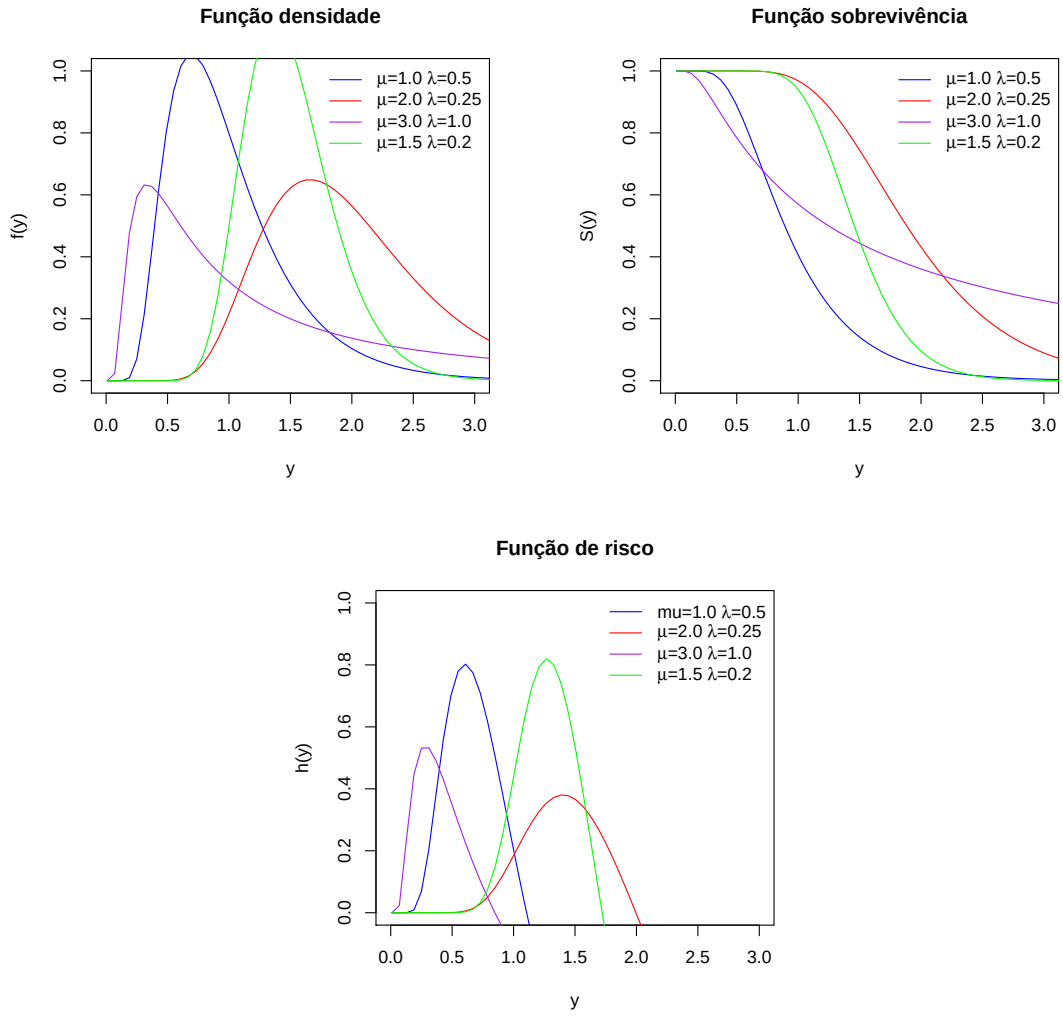
$$f(y) = (\lambda/2\pi y^3)^{1/2} \exp\left[-\lambda(y-\mu)^2/2\mu^2 y\right],\ y>0, \tag{2.18}$$

em que os parâmetros $\mu > 0$ e $\lambda > 0$. A média e a variância são dadas por $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \mu^3/\lambda$, respectivamente. A função de distribuição acumulada da distribuição Inversa Gaussiana é dada por:

$$F(y) = \Phi\left(\sqrt{\frac{\lambda}{y}}\left(\frac{y}{\mu}+1\right)\right) + e^{2\lambda/\mu}\Phi\left(-\sqrt{\frac{\lambda}{y}}\left(1+\frac{y}{\mu}\right)\right). \tag{2.19}$$

em que Φ é a distribuição de probabilidade acumulada da normal padrão.
A Figura 2, apresenta a ilustração gráfica das funções densidade, de sobrevivência e de risco da distribuição IG, em que a expressão da função de sobrevivência da distribuição Inversa Gaussiana é dada por $S(y) = 1 - F(y)$, sendo $F(.)$ a função de distribuição acumulada dada em (2.19), a função densidade dada por $f(y)$ apresentada em (2.18) e a função de risco pode ser obtida pela relação $\frac{f(y)}{S(y)}$.

Figura 2 – Funções de densidade, sobrevivência e risco da distribuição IG para valores arbitrários dos parâmetros.



Fonte: Elaborada pela autora.

Uma característica que faz com que a distribuição Inversa Gaussiana seja uma boa candidata para a modelagem de tempos de vida é o fato da sua função de risco ser unimodal. Folks e Chhikara (1978) mostraram que a função de risco cresce até um valor máximo e depois decresce assintoticamente para $0,5\lambda\mu^{-2}$.

Sanhueza, Leiva e Balakrishnan (2008), Kotz, Leiva e Sanhueza (2010) e Leiva *et al.* (2010) propuseram e discutiram novas classes de modelos de mistura baseados na distribuição Inversa Gaussiana. Balka, Desmond e McNicolas (2009) consideraram modelos com fração de cura com base na distribuição Inversa Gaussiana possuindo muitas extensões, incluindo os modelos de mistura. Uma distribuição de mistura Inversa Gaussiana (MIG) tem função de sobrevivência dada por:

$$S_{MIG}(y) = p_0 + (1-p_0)S_{IG}(y),\ y > 0, \tag{2.20}$$

em que $S_{IG}(y) = 1 - F_{IG}(y)$ é a função de sobrevivência da distribuição Inversa Gaussiana e p_0 é a fração de curados.

2.6.1 Estimador de máxima verossimilhança

Em contraste com outras distribuições utilizadas na análise de dados de sobrevivência, podemos obter analiticamente os estimadores de máxima verossimilhança para os parâmetros μ e λ.

Considere $Y_1, Y_2, \ldots, Y_n$ uma amostra aleatória de tamanho n de Y, em que $Y \sim IG(\mu, \lambda)$, com função de densidade de probabilidade definida em (**??**). Então, a função de máxima verossimilhança é dada por:

$$\mathscr{L}(\mu, \lambda) = \prod_{i=1}^{n} (\lambda / 2\pi y_i^3)^{1/2} \exp\left[-\lambda (y_i - \mu)^2 / 2\mu^2 y_i\right].$$

Logo, a função de log-verossimilhança é dada por:

$$l(\mu, \lambda) = \frac{n}{2} log\left(\frac{\lambda}{2\pi}\right) - \frac{3}{2} \log\left(\sum_{i=1}^{n} y_i\right) - \frac{\lambda}{2\mu^2} \sum_{i=1}^{n} \left(\frac{(y_i - \mu)^2}{y_i}\right). \tag{2.21}$$

Os estimadores de máxima verossimilhança $\hat{\mu}$ e $\hat{\lambda}$ são obtidos resolvendo as equações $U_\mu = 0$ e $U_\lambda = 0$, em que $U_\mu = \frac{\partial l(\mu,\lambda)}{\partial \mu}$ e $U_\lambda = \frac{\partial l(\mu,\lambda)}{\partial \lambda}$, sendo $l(\mu, \lambda)$ a função log-verossimilhança dada em (2.21).

Portanto, os estimadores são $\hat{\mu} = \bar{Y}$ e $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}\left(Y_i^{-1} - \frac{1}{\bar{Y}}\right)}$, em que $\bar{Y}$ é a média amostral dada por $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

CAPÍTULO

3

O MODELO BELL INVERSA GAUSSIANA COM FRAÇÃO DE CURA

Neste capítulo apresentamos o modelo Bell Inversa Gaussiana com fração de cura (BIGcr) para os três tipos de esquema de ativação.

Seja M as possíveis causas do evento de interesse para um indivíduo na população. Supondo que M segue distribuição Bell de parâmetro θ com função massa de probabilidade dada em (2.11) e função geradora de probabilidade apresentada em (2.13).

Considerando $Z_j, j = 1, ..., M$ o tempo de ocorrência da j-ésima causa competitiva, em que M e Z_j são variáveis latentes, ou seja, não são observáveis. Assumindo independência entre as variáveis Z_j e também entre Z_j e M, então, atribui-se a Z_j uma distribuição Inversa Gaussiana de parâmetros μ e λ, com função densidade de probabilidade dada em (**??**).

Com as definições dadas nas Seções 2.3 e 2.4, temos que as funções densidade e de sobrevivência impróprias do modelo BIGcr sob os esquemas de ativação latentes, são descritas por:

(1) *Esquema de primeira ativação:* Neste cenário assume-se que $R = 1$ e o tempo de vida observado é definido pela variável aleatória $Y = Z_{(1)} = \min\{Z_1, \ldots, Z_M\}$.

A função de sobrevivência de Y é dada por:

$$S_{\text{pop}}(y) = A_M(S_{IG}(y)) = \exp(e^{S_{IG}(y)\theta} - e^{\theta}), \tag{3.1}$$

em que $A_M(.)$ é dado na expressão (2.13) e $S_{IG}(y)$ é a função de sobrevivência da distribuição Inversa Gaussiana.

A função densidade de probabilidade associada a (3.1) é dada por:

$$f_{\text{pop}}(y) = \exp(e^{S_{IG}(y)\theta} - e^{\theta}) e^{S_{IG}(y)\theta} f_{IG}(y)\theta, \tag{3.2}$$

em que $f_{IG}(y) - \frac{dS_{IG}(y)}{dy}$ é a função densidade de probabilidade da distribuição Inversa Gaussiana.

(2) *Esquema de última ativação:* Neste cenário temos que $R = M$ e o tempo de vida observado é dado por $Y = Z_{(M)} = \max\{Z_1, \ldots, Z_M\}$. A função de sobrevivência de Y é dada por:

$$\begin{aligned} S_{\text{pop}}(y) &= 1 + A_M(0) - A_M(1 - S_{IG}(y)) \\ &= 1 + \exp(1 - e^{\theta}) - \exp(e^{F_{IG}(y)\theta} - e^{\theta}), \end{aligned} \tag{3.3}$$

em que $F_{IG}(y) = 1 - S_{IG}(y)$ é a função de distribuição acumulada da distribuição Inversa Gaussiana.

A função de sobrevivência dada em (3.3) leva à seguinte função densidade de probabilidade:

$$f_{\text{pop}}(y) = \exp(e^{F_{IG}(y)\theta} - e^{\theta}) e^{F_{IG}(y)\theta} f_{IG}(y)\theta. \tag{3.4}$$

(3) *Esquema de ativação aleatória*: Neste terceiro cenário, assume-se que $M \geq 1$ e considera para R uma distribuição Uniforme discreta em $\{1, 2, \ldots, M\}$. A função de sobrevivência de Y é dada por:

$$\begin{aligned} S_{\text{pop}}(y) &= P(Y > y) = P(M = 0) + \{1 - P(M = 0)\} S_{IG}(y) \\ &= \exp(1 - e^{\theta}) + [1 - \exp(1 - e^{\theta})] S_{IG}(y). \end{aligned} \tag{3.5}$$

Por meio da função de sobrevivência $S_{\text{pop}}(y)$ dada em (3.5), temos que a função densidade de probabilidade é dada por:

$$f_{\text{pop}}(y) = (1 - \exp(1 - e^{\theta})) f_{IG}(y). \tag{3.6}$$

Para os três esquemas, podemos reparametrizar as funções de sobrevivência e densidades de probabilidade em função da proporção de cura p, considerando:

$$p = P(M = 0) = \exp(-e^{-\theta} + 1) \Rightarrow \log(p) = -e^{\theta} + 1 \Rightarrow \theta = \log(1 - \log(p)). \tag{3.7}$$

A partir das expressões (3.1)-(3.6) e da expressão de θ em (3.7), temos que as funções de sobrevivência e de densidade de probabilidade para os três cenários ficam dadas por:

Esquema de primeira ativação:

$$\begin{aligned} &S_{pop}(y) = \exp((1 - \log(p))^{S_{IG}(y)} - 1 + \log(p)) \quad \text{e} \\ &f_{pop}(y) = \exp((1 - \log(p))^{S_{IG}(y)} - 1 + \log(p))(1 - \log(p))^{S_{IG}(y)} \log(1 - \log(p)) f_{IG}(y). \end{aligned} \tag{3.8}$$

Esquema de última ativação:

$$\begin{aligned} &S_{pop}(y) = 1 + p - \exp((1 - \log(p))^{F_{IG}(y)} - 1 + \log(p)) \quad \text{e} \\ &f_{pop}(y) = \exp((1 - \log(p))^{F_{IG}(y)} - 1 + \log(p))(1 - \log(p))^{F_{IG}(y)} \log(1 - \log(p)) f_{IG}(y). \end{aligned} \tag{3.9}$$

Esquema de ativação aleatória:

$$S_{pop}(y) = p + (1-p)S_{IG}(y) \text{ e } f_{pop}(y) = (1-p)f_{IG}(y). \tag{3.10}$$

As Figuras 3, 4 e 5, apresentam os gráficos dos diferentes comportamentos das funções de densidade, sobrevivência e de risco do modelo BIGcr sob os esquemas de primeira ativação, de ativação aleatória e de última ativação, respectivamente. Foram considerados diferentes valores arbitrários para os parâmetros.

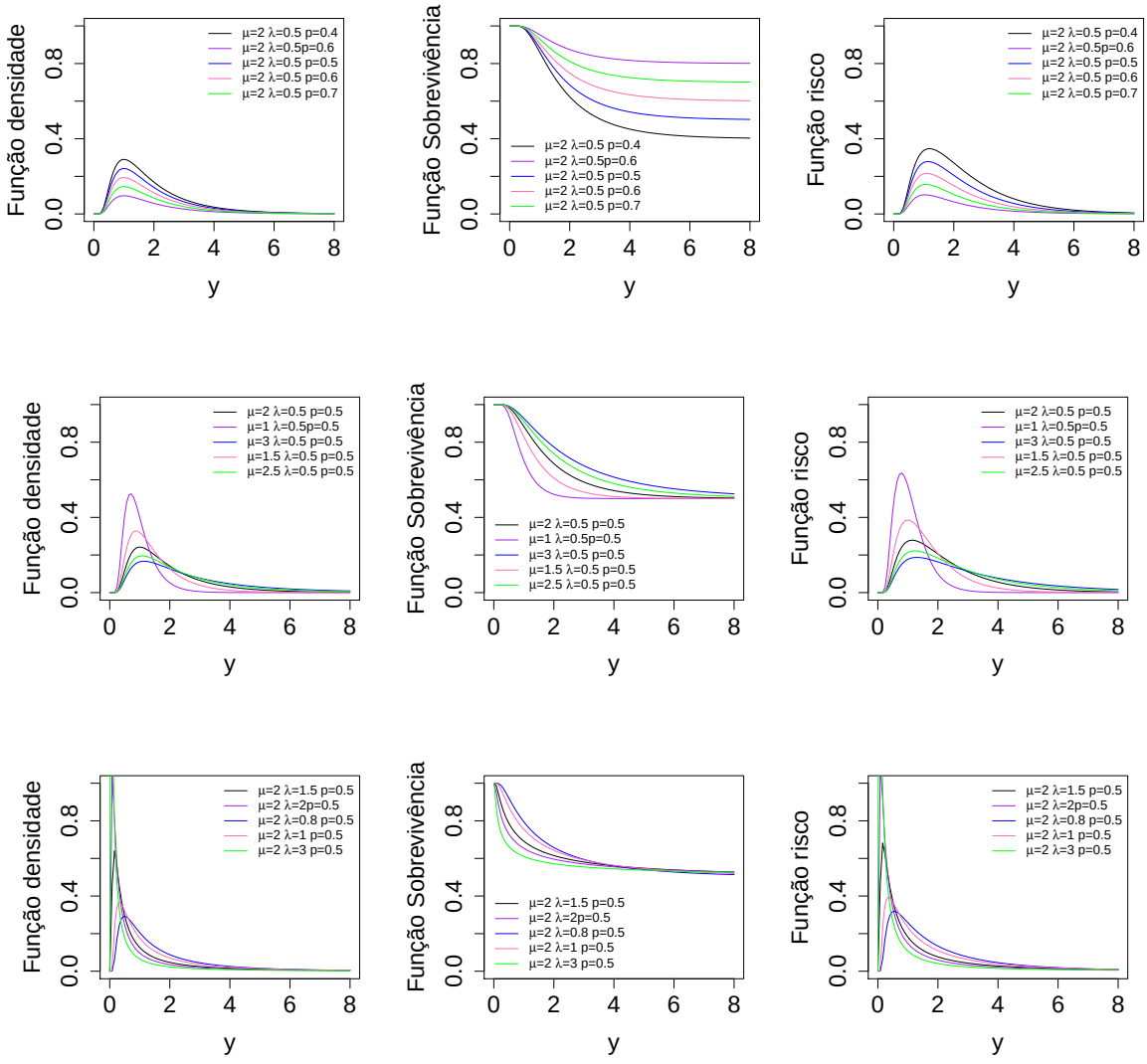


Figura 3 – Funções densidade, sobrevivência e risco sob o esquema de ativação aleatória para valores arbitrários dos parâmetros.

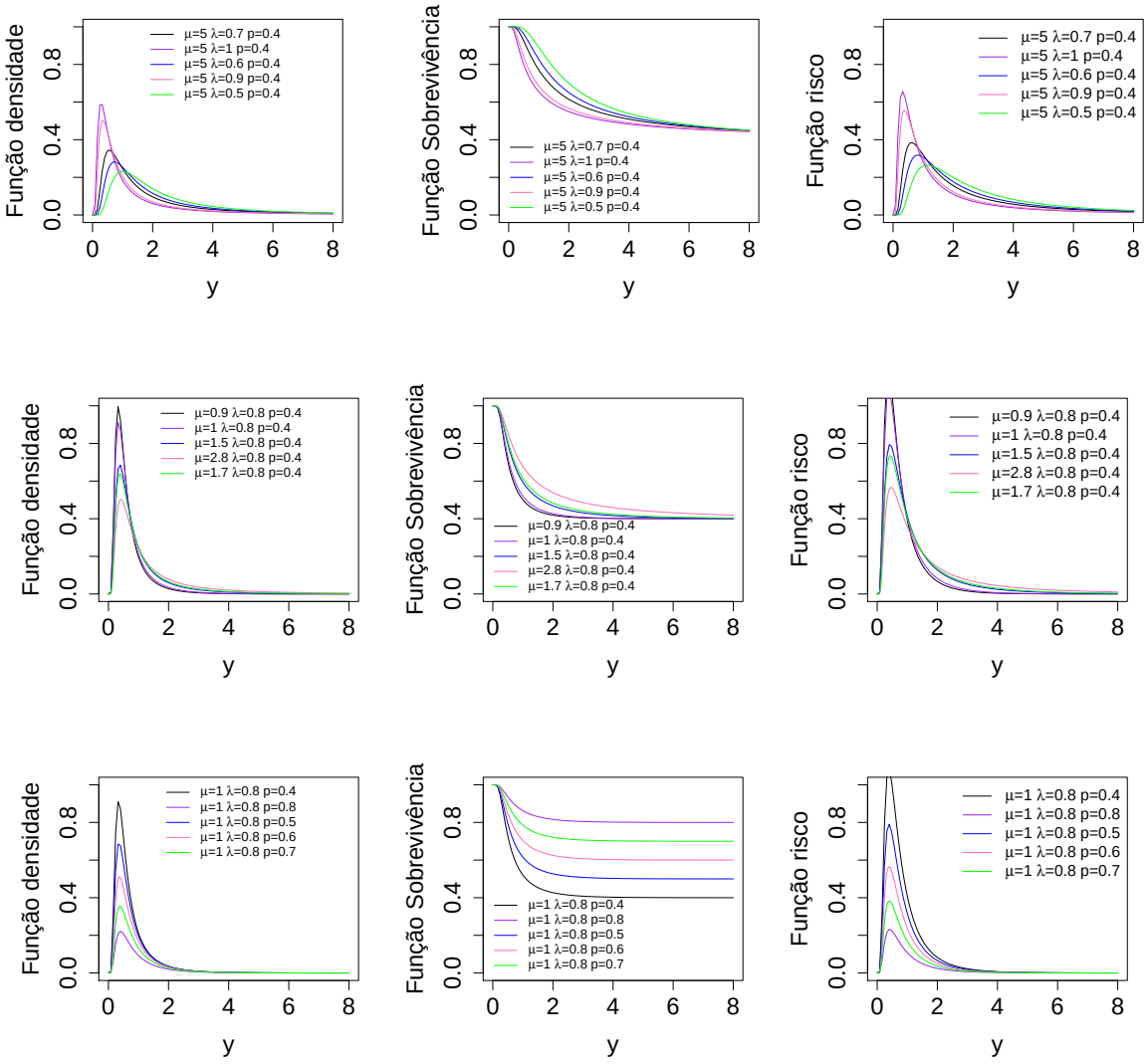


Figura 4 – Funções densidade, sobrevivência e risco sob o esquema de primeira ativação para valores arbitrários dos parâmetros.

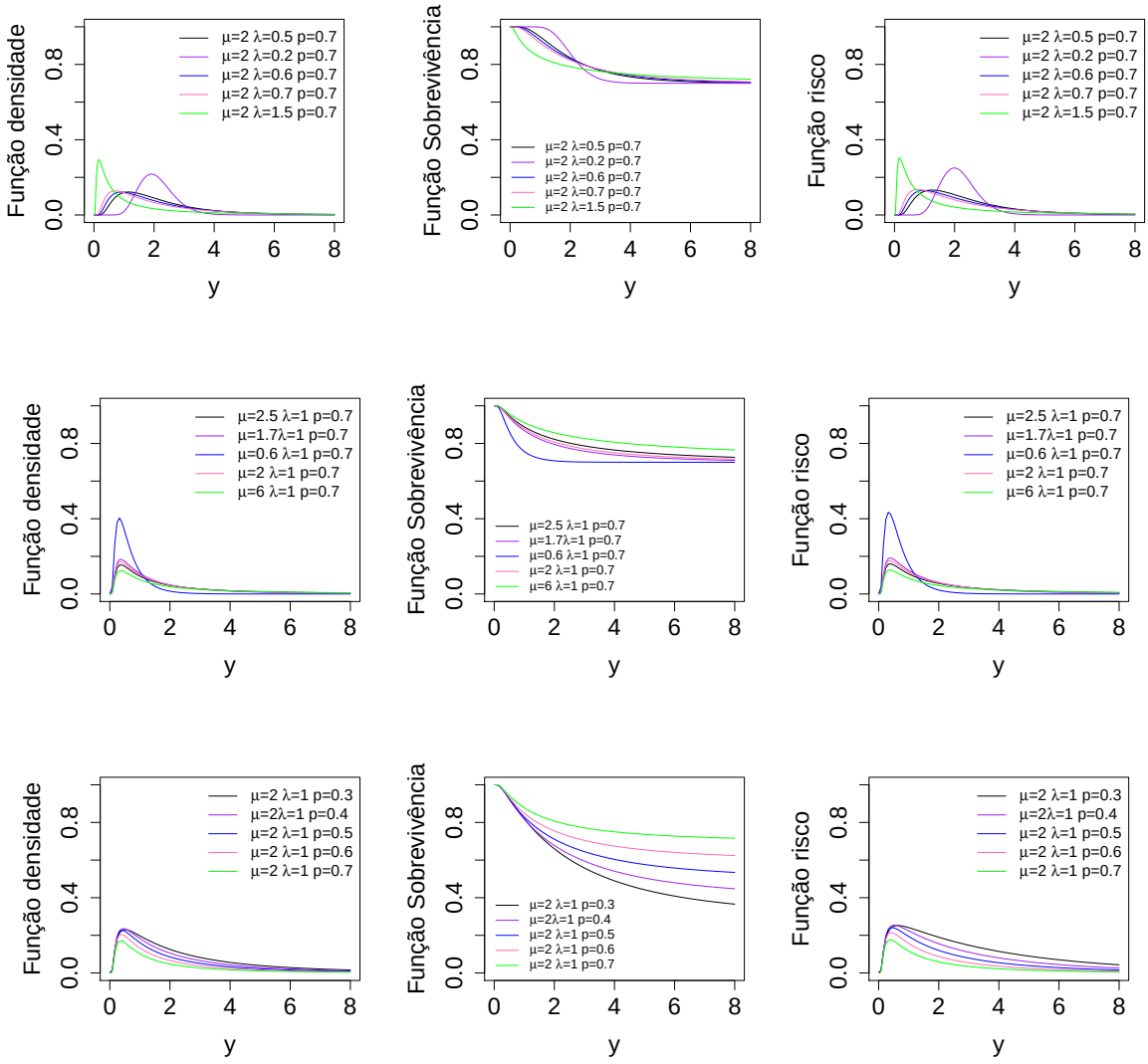


Figura 5 – Funções densidade, sobrevivência e risco sob o esquema de última ativação para valores arbitrários dos parâmetros.

3.1 Estimação por Máxima Verossimilhança

Considere que o tempo de vida denotado por T está sujeito a censura à direita. Seja C_i o i-ésimo tempo de censura. Assim, em uma amostra de tamanho n, o i-ésimo tempo observado Y_i é dado por $Y_i = \min\{T_i, C_i\}$. Além disso, há o interesse no indicador de censura $\delta_i = \mathrm{I}(T_i \leq C_i)$, em que $\delta_i = 1$ se Y_i é o tempo de falha e $\delta_i = 0$ se for o tempo de censura, para $i = 1, \ldots, n$.

Assumimos que temos disponível um vetor com g covariáveis e seja $x^\top = (1, x_1, \ldots, x_g)$. Relacionamos a proporção de curados com as covariáveis x_i, por meio da seguinte função de ligação:

$$logit(p_i) = \mathbf{x_i^T}\beta,$$

ou seja,

$$p_i = \frac{\exp(\mathbf{x_i^T}\beta)}{1+\exp(\mathbf{x_i^T}\beta)}, \tag{3.11}$$

para $i = 1, \ldots, n$, em que $\beta^\top = (\beta_0, \beta_1, \ldots, \beta_g)$ denota o vetor de parâmetros, de modo que para cada grupo de indivíduos, representado por x_i, temos uma fração de cura diferente.

A função de verossimilhança é dada por:

$$\mathscr{L}(\vartheta; \mathscr{D}) = \prod_{i=1}^{n} f_{\text{pop}}(y_i; \vartheta)^{\delta_i} S_{\text{pop}}(y_i; \vartheta)^{1-\delta_i},$$

em que $\vartheta = (\beta, \lambda)^\top$, $\mathscr{D} = (y, \delta, x)$, $y = (y_1, \ldots, y_n)^\top$ e $\delta = (\delta_1, \ldots, \delta_n)^\top$, enquanto que $f_{\text{pop}}(\cdot; \vartheta)$ e $S_{\text{pop}}(\cdot; \vartheta)$ estão descritas em (3.8), (3.9) e (3.10) para os modelos BIGcr sob os esquemas de primeira, última e ativação aleatória, respectivamente.

A estimativa de máxima verossimilhança (MLE) $\hat{\vartheta}$ de ϑ é obtida resolvendo as equações não-lineares $U_\mu = 0$, $U_\lambda = 0$ e $U_{\beta_j} = 0$ para $j = 0, 1$, em que $U_\mu = \frac{\partial l(\vartheta)}{\partial \mu}$, $U_\lambda = \frac{\partial l(\vartheta)}{\partial \lambda}$, $U_{\beta_j} = \frac{\partial l(\vartheta)}{\partial \beta_j}$ e $l(\vartheta) = \sum_{i=1}^{n} \delta_i \log(f_{pop}(y_i)) + \sum_{i=1}^{n} (1-\delta_i) \log(S_{pop}(y_i))$.

Podemos notar que essas equações não podem ser resolvidas analiticamente e *softwares* como o R, SAS e o Ox podem ser usados para resolvê-los numericamente. No Apêndice A apresentamos as derivadas para a construção das funções escore. Neste trabalho utilizamos o *software* R por meio da função *optim*.

O método de estimação por máxima verossimilhança para o vetor de parâmetros ϑ pode ser implementado por maximização numérica da função de log-verossimilhança $l(\vartheta)$. Além disso, intervalos de confiança e testes de hipóteses podem ser realizados usando a distribuição assintótica do estimador de máxima verossimilhança, que é uma distribuição normal multivariada com a matriz de variância-covariância como o inverso da informação esperada sob condições de regularidade. Mais especificamente, sob condições que são atendidas pelo vetor de parâmetro ϑ no interior do espaço paramétrico, mas não no limite, a distribuição assintótica de $\sqrt{n}(\widehat{\vartheta} - \vartheta)$ é normal multivariada $N_{p+3}(0, K(\vartheta)^{-1})$, em que $K(\vartheta)$ é a matriz de informação esperada. A

matriz de covariância assintótica $K(\vartheta)^{-1}$ de $\widehat{\vartheta}$ pode ser aproximada pelo inverso da matriz $(p+3)\times(p+3)$ de informação observada $-\ddot{L}(\vartheta)$. A distribuição normal multivariada aproximada $N_{p+3}(0,-\ddot{L}(\vartheta)^{-1})$ para $\widehat{\vartheta}$ pode ser utilizada para construir regiões de confiança aproximada para alguns parâmetros em ϑ.

3.2 Estudo de simulação

Neste trabalho realizamos um estudo de simulação com d = 1.000 réplicas para o modelo BIGcr, sob os três esquemas de ativação apresentados anteriormente. Consideramos quatro tamanhos amostrais $n = 100$, 200, 400 e 800, com uma proporção de dados censurados de 55%. Para obter os elementos da amostra, primeiramente geramos uma amostra de variáveis latentes M de uma distribuição Bell com parâmetro $\theta = \log(1-\log(p))$ (em que p é dado pela equação (3.11). Para gerar a $j-$ésima amostra dos tempos de falha, com $j = 1,\ldots,1.000$, gerou-se uma amostra de tamanho M_i de uma distribuição IG(μ,λ) de tal forma que o $i-$ ésimo tempo de falha fosse o máximo dos tempos gerados desta amostra no modelo de última ativação, o mínimo no modelo de primeira ativação e um valor aleatório no modelo de ativação aleatória. Também foram gerados os tempos de censura de uma distribuição Uniforme no intervalo (0,13). Assumimos uma covariável x que tem uma distribuição Bernoulli de parâmetro $0,5$. Os valores verdadeiros dos parâmetros μ,λ,β_0 e β_1 são: 0,5, 2, -0,5 e 0,7, respectivamente.

Em cada simulação foi calculada a média, o viés, o erro quadrático médio (EQM) e a probabilidade de cobertura (PC), que foi obtida por meio do intervalo assintótico de 95%.

Seja d o número de réplicas, $\vartheta = (\mu,\lambda,\beta_0,\beta_1)$, ϑ_i o i-ésimo parâmetro de ϑ e $\hat{\vartheta}_{ij}$ a estimativa do i-ésimo parâmetro na j-ésima réplica. A média, o viés e o EQM de ϑ_i são calculadas por: Média$(\widehat{\vartheta}_i) = \sum_{j=1}^{d}\frac{\widehat{\vartheta_{ij}}}{d}$, Viés$(\widehat{\vartheta}_i)$ = Média$(\widehat{\vartheta}) - \vartheta_{ij}$ e $EQM(\hat{\vartheta}_i) = \sum_{j=1}^{d}\frac{(\vartheta_i-\widehat{\vartheta_{ij}})^2}{d}$, respectivamente.

As Tabelas 3, 4 e 5 apresentam os valores das estimativas dos parâmetros obtidos pelo método de máxima verossimilhança para o modelo BIGcr sob o esquema de primeira ativação, última ativação e ativação aleatória, respectivamente. Para todos os cenários, podemos observar que à medida que aumenta o tamanho amostral n, diminui o viés e o EQM. Para a probabilidade de cobertura, os resultados obtidos mostraram que os valores das probabilidades estão próximos do valor nominal, exceto para o parâmetro μ no esquema de ativação aleatória e no caso de tamanho amostral $n = 100$ para o cenário de primeira ativação.

Tabela 3 – Estudo de simulação sob o esquema de primeira ativação.

n	Parâmetro	Média	Viés	EQM	PC
100	μ	0,6384	0,1384	0,4605	0,8890
	λ	1,9662	-0,0338	0,0324	0,9310
	β_0	-0,5236	-0,0236	0,0898	0,9700
	β_1	0,7324	0,0324	0,1647	0,9610
200	μ	0,5670	0,0670	0,0983	0,9240
	λ	1,9878	-0,0122	0,0155	0,9470
	β_0	-0,5130	-0,0130	0,0494	0,9560
	β_1	0,7086	0,0086	0,0787	0,9550
400	μ	0,5377	0,0377	0,0873	0,9330
	λ	1,9948	-0,0052	0,0079	0,9490
	β_0	-0,5100	-0,0100	0,0237	0,9430
	β_1	0,7114	0,0114	0,0401	0,9440
800	μ	0,5147	0,0147	0,0073	0,9330
	λ	1,9967	-0,0033	0,0040	0,9390
	β_0	-0,5027	-0,0027	0,0116	0,9410
	β_1	0,7002	0,0002	0,0186	0,9530

Tabela 4 – Estudo de simulação sob o esquema de última ativação.

n	Parâmetro	Média	Viés	EQM	PC
100	μ	0,5612	0,0612	0,1796	0,9360
	λ	1,9920	0,0080	0,0783	0,9310
	β_0	-0,5479	-0,0479	0,1617	0,9540
	β_1	0,7432	0,0432	0,2348	0,9590
200	μ	0,5150	0,0150	0,0110	0,9290
	λ	1,9971	-0,0029	0,0338	0,9470
	β_0	-0,5255	-0,0255	0,0708	0,9570
	β_1	0,7191	0,0191	0,1157	0,9470
400	μ	0,5101	0,0101	0,0144	0,9343
	λ	2,0000	0,0000	0,0189	0,9360
	β_0	-0,5202	-0,0202	0,0329	0,9500
	β_1	0,7236	0,0236	0,0572	0,9440
800	μ	0,5033	0,0033	0,0019	0,9530
	λ	1,9990	0,0010	0,0093	0,9530
	β_0	-0,5037	-0,0037	0,0159	0,9540
	β_1	0,6995	0,0005	0,0267	0,9480

Tabela 5 – Estudo de simulação sob o esquema de ativação aleatória.

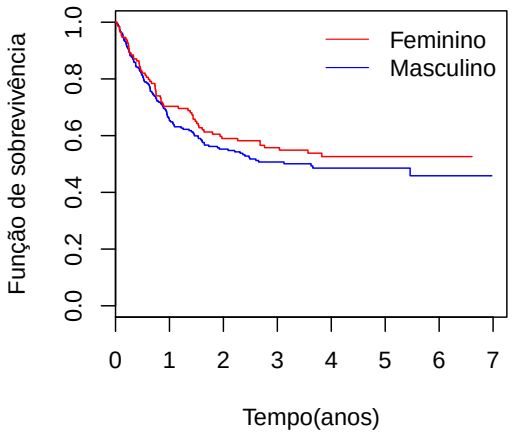
n	Parâmetro	Média	Viés	EQM	PC
100	μ	0,5445	0,0445	0,1125	0,9000
	λ	1,9694	-0,0306	0,0415	0,9350
	β_0	-0,5415	-0,0415	0,1205	0,9610
	β_1	0,7238	0,0238	0,1990	0,9660
200	μ	0,5054	0,0054	0,0223	0,8920
	λ	1,9712	-0,0288	0,0214	0,9370
	β_0	-0,5260	-0,0260	0,0585	0,9560
	β_1	0,7283	0,0283	0,1041	0,9540
400	μ	0,4935	-0,0065	0,0061	0,8930
	λ	1,9724	-0,0276	0,0112	0,9280
	β_0	-0,5107	-0,0107	0,0265	0,9580
	β_1	0,6992	0,0008	0,0502	0,9510
800	μ	0,4866	-0,0134	0,0029	0,9010
	λ	1,9771	-0,0229	0,0058	0,9350
	β_0	-0,5069	-0,0069	0,0142	0,9480
	β_1	0,7011	0,0011	0,0252	0,9580

3.3 Aplicação a dados reais

Nesta seção apresentamos uma aplicação do modelo proposto ao conjunto de dados reais de um ensaio clínico de melanoma cutâneo de fase III. Os dados foram obtidos pelo *Eastern Cooperative Oncology group*, disponível em http://merlot.stat.uconn.edu/ mhchen/survbook/ (KIRKWOOD *et al.*, 2000). Os pacientes foram submetidos à operação, fizeram um tratamento com uma alta dose da droga interferon alfa-2b para avaliar o desempenho do tratamento pós operatório para prevenir a recorrência da doença, sendo assim, fora analisado o comportamento da população que recebeu a administração da droga e outro que não a recebeu. O estudo teve início por meio da inclusão de pacientes no período de 1991 a 1995, acompanhando-os até 1998. Foi considerado como variável resposta o tempo de sobrevivência sem recidiva em anos. Após excluir indivíduos com dados incompletos e tempos de observação faltantes, obtivemos um subconjunto de $n = 408$ pacientes com aproximadamente 43% de censura. As seguintes variáveis foram coletadas de cada paciente: tempo observado (em anos, média = $2,31$, desvio padrão = $1,93$) e x : sexo (0, masculino ($n = 257$)), (1, feminino ($n = 151$)).

Como parte de uma análise preliminar dos dados, a Figura 6 exibe as curvas de sobrevivência Kaplan-Meier estratificadas por sexo, na qual podemos observar um platô bem acima de zero.

Figura 6 – Estimativa de Kaplan-Meier da função de sobrevivência estratificadas por sexo.



Fonte: Elaborada pela autora.

Isso é indicativo da presença de pacientes para as quais o melanoma maligno possivelmente não retornará (ou seja, eles foram curados), o que motivou o ajuste de um modelo de sobrevivência de taxa de cura para este conjunto de dados. Além disso, a presença de platô indica que modelos que ignoram a possibilidade de cura não serão adequados para esses dados. Nesta aplicação ajustamos o modelo BIG$_c r$ considerando os três mecanismos de ativações latentes: primeira, última e aleatória. Os valores obtidos para a log-verossimilhança foram 585,884, 546,953 e 560,049, respectivamente. Levando em consideração este critério, selecionamos o modelo BIG$_c r$ sob o esquema de última ativação como nosso modelo de trabalho.

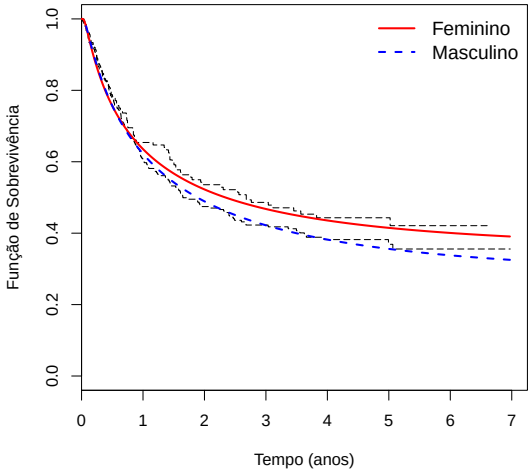
A Tabela 6 lista as estimativas de máxima verossimilhança (Estimativa), os erros padrões (EP) e os intervalos de confiança de 95% (IC (95%)) de cada parâmetro do modelo ajustado. Podemos observar que a covariável sexo não foi significativa a 5%.

Tabela 6 – Estimativas de máxima verossimilhança dos parâmetros para o modelo BIGcr sob o esquema de última ativação.

Parâmetro	Estimativa	EP	IC (95%)
$\log(\mu)$	0,078	0,251	(-0,413, 0,569)
$\log(\lambda)$	-1,389	0,161	(-1,705, -1,073)
β_0	-0,947	0,272	(-1,480, -0,414)
β_1	0,350	0,285	(-0,208, 0,908)

A fim de avaliar se o modelo é adequado, a Figura 7 mostra as curvas de Kaplan-Meier estratificadas pela covariável sexo juntamente com as curvas de sobrevivência do modelo BIGcr sob o esquema de última ativação estimada, na qual é possível observar o bom ajuste do modelo aos dados.

Figura 7 – Curvas de Kaplan-Meier e curvas de sobrevivência BIGcr estimadas para os dados de melanoma cutâneo.



Fonte: Elaborada pela autora.

CAPÍTULO

4

ABORDAGEM BAYESIANA

Neste capítulo será realizada a inferência dos parâmetros do modelo BIGcr por meio de uma abordagem Bayesiana.

De forma análoga ao Capítulo 3, considere que o tempo de vida denotado por T está sujeito a censura à direita. Seja C_i o i-ésimo tempo de censura. Assim, em uma amostra de tamanho n, o i-ésimo tempo observado Y_i é dado por $Y_i = \min\{T_i, C_i\}$. Além disso, há o interesse no indicador de censura $\delta_i = \mathrm{I}(T_i \leq C_i)$, em que $\delta_i = 1$ se Y_i é o tempo de falha e $\delta_i = 0$ se for o tempo de censura, para $i = 1, \ldots, n$.

Assumimos que temos disponível um vetor com r covariáveis e seja $\mathbf{x}^\top = (1, x_1, \ldots, x_r)$. Relacionamos a proporção de curados com as covariáveis $\mathbf{x_i}$, por meio da seguinte função de ligação:

$$logit(p_i) = \mathbf{x_i^T}\beta,$$

ou seja,

$$p_i = \frac{exp(\mathbf{x_i^T}\beta)}{1 + exp(\mathbf{x_i^T}\beta)}, \tag{4.1}$$

para $i = 1, \ldots, n$, em que $\beta^\top = (\beta_0, \beta_1, \ldots, \beta_g)$, denota o vetor de parâmetros de modo que para cada grupo de indivíduos, representado por x_i, temos uma fração de cura diferente.

A função de verossimilhança é dada por:

$$\mathscr{L}(\vartheta; \mathscr{D}) = \prod_{i=1}^{n} f_{\text{pop}}(y_i; \vartheta)^{\delta_i} S_{\text{pop}}(y_i; \vartheta)^{1-\delta_i}, \tag{4.2}$$

em que $\vartheta = (\beta, \lambda)^\top$, $\mathscr{D} = (y, \delta, x)$, $y = (y_1, \ldots, y_n)^\top$ e $\delta = (\delta_1, \ldots, \delta_n)^\top$, enquanto que $f_{\text{pop}}(\cdot; \vartheta)$ e $S_{\text{pop}}(\cdot; \vartheta)$ estão descritas em (3.8), (3.9) e (3.10) para os modelos BIGcr sob os esquemas de primeira, última e ativação aleatória, respectivamente.

4.1 Distribuições a *priori* e a *posteriori*

Para a realização da inferência Bayesiana devemos especificar uma distribuição a *priori* para os parâmetros μ, λ e β. Esta distribuição deve representar probabilisticamente o conhecimento que se tem sobre os parâmetros antes da realização do experimento.

Neste trabalho assumimos que μ, λ e β são independentes a *priori*, isto é,

$$\pi(\vartheta) = \pi(\mu)\pi(\lambda)\prod_{j=0}^{p}\pi(\beta_j), \tag{4.3}$$

em que $\vartheta = (\mu, \lambda, \beta)$.

Como os parâmetros μ e λ são definidos em $\Re_*^+$ e β_j em $\Re$, $j = 0, \ldots, g$, consideramos as seguintes distribuições a *priori*:

$$\mu \sim Gama(a,b), \quad \lambda \sim Gama(a,b) \quad \text{e} \quad \beta_j \sim N(0,\sigma^2), j = 0, \ldots, g,$$

em que $Gama(a,b)$ denota a distribuição Gama com parâmetro de forma a e de escala b e, $N(0,\sigma^2)$ denota a distribuição Normal com média 0 e variância σ^2.

Para os parâmetros μ e λ também consideramos as seguintes distribuições a *priori*:

$$\log(\mu) \sim N(0,\sigma^2) \quad \text{e} \quad \log(\lambda) \sim N(0,\sigma^2).$$

Para obtermos distribuições a *priori* pouco informativas, *i.e.*, com grandes variâncias, fixamos $a = 0,1$, $b = 0,01$ e $\sigma^2 = 10^3$.

Atualizando a distribuição a *priori* conjunta $\pi(\vartheta)$ dada em (4.3), via função de verossimilhança $L(\vartheta; \mathscr{D})$ dada em (4.2), obtemos a distribuição a *posteriori* conjunta para ϑ, que é analiticamente intratável para todos os modelos (primeira ativação, última ativação e ativação aleatória). Dessa forma, a inferência sobre os parâmetros foi baseada nos métodos de simulação de MCMC).

4.2 Critérios de comparação de modelos

Na literatura, diversas metodologias se propõem a analisar a adequabilidade de um modelo a um certo conjunto de dados, além de selecionar o melhor dentre uma coleção de modelos.

Neste trabalho utilizamos o critério LPML (*Logarithm of the Pseudo Marginal Likelihood*) que é obtido com base nas ordenadas da densidade preditiva condicional (CPO). O cálculo da CPO é feita da seguinte forma: considere $\mathscr{D}$ os dados completos, $\mathscr{D}^{(-i)}$ os dados com a i-ésima observação excluída e $\pi(\vartheta|\mathscr{D}^{(-i)})$ a densidade a *posteriori* de ϑ dado $\mathscr{D}^{(-i)}$, para $i = 1, \ldots, n$.

Para a i-ésima observação, a CPO_i é dada por:

$$CPO_i = \int_\Theta f(y_i|\vartheta)\pi(\vartheta|\mathscr{D}^{(-i)})d\vartheta = \left\{\int_\Theta \frac{\pi(\vartheta|\mathscr{D})}{f(y_i|\vartheta)}d\vartheta\right\}^{-1}, \tag{4.4}$$

em que $f(y_i|\vartheta)$ é a função densidade de probabilidade correspondente ao modelo estudado.

Na comparação de vários modelos, valores altos de CPO_i indicam um melhor ajuste do modelo. Para o modelo proposto neste trabalho, não é possível obter uma forma analítica fechada de CPO_i. No entanto, podemos obter uma estimativa de Monte Carlo de CPO_i por meio de uma simples amostra MCMC da distribuição a *posteriori* $\pi(\vartheta|\mathscr{D})$. Dessa forma, considere $\vartheta^{(1)}, \vartheta^{(2)}, ..., \vartheta^{(M)}$ uma amostra de tamanho M de $\pi(\vartheta|\mathscr{D})$ após o *burn-in*. Uma aproximação de Monte Carlo de CPO_i (CHEN; SHAO; IBRAHIM, 2000) é dada por:

$$\widehat{CPO_i} = \left\{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{f(y_i|\vartheta^{(m)})}\right\}^{-1}.$$

A estatística LPML é dada por LPML$= \sum_{i=1}^{n}\log(\widehat{CPO_i})$. Logo, o melhor modelo é aquele que corresponde ao maior valor da LPML.

Neste trabalho também consideramos outros três critérios de seleção de modelos: EAIC (*Expected Akaike Information Criterion*) proposto por Brooks (2002), EBIC (*Expected Bayesian*, ou *Schwarz* (CARLIN; LOUIS, 2001), *Information Criterion*) e DIC (*Deviance Information Criterion* (SPIEGELHALTER *et al.*, 2002)).

Os critérios EAIC, EBIC e DIC são critérios calculados com base na média a *posteriori* da *deviance*, $E\{D(\vartheta)\}$, que é uma medida de ajuste e que pode ser aproximada por:

$$\overline{D} = \frac{1}{M}\sum_{m=1}^{M} D(\vartheta^{(m)}),$$

em que $D(\vartheta) = -2\sum_{i=1}^{n}\log(f(y_i|\vartheta))$ é a deviance e $f(\cdot)$ é a função densidade de probabilidade correspondente ao modelo estudado.

Os critérios EAIC, EBIC e DIC podem ser calculados, respectivamente, por:

$$\widehat{EAIC} = \overline{D} + 2q, \quad \widehat{EBIC} = \overline{D} + q\log(n) \quad \text{e} \quad \widehat{DIC} = 2\overline{D} - \widehat{D},$$

em que q é o número de parâmetros do modelo e $\widehat{D} = D\left(\frac{1}{M}\sum_{m=1}^{m}\vartheta^{(m)}\right)$.

Comparando modelos alternativos, o modelo preferido é aquele com os menores valores para estes três critérios.

4.3 Método Bayesiano de análise de influência de deleção de casos

O método da deleção de casos (COOK; WEISBERG, 1982) é uma ferramenta muito utilizada quando o objetivo é avaliar a influência de uma observação no ajuste de um modelo. Várias técnicas de verificação de influência local têm sido amplamente utilizadas na literatura, como por exemplo em Suzuki, Cancho e Louzada (2016) e Suzuki *et al.* (2017).

Neste trabalho, consideramos a análise de influência de deleção de casos baseada na divergência ψ que é calculada da seguinte forma: seja $D_\psi(P;P_{(-i)})$ a divergência ψ entre P e $P_{(-i)}$, em que P indica a distribuição a *posteriori* de ϑ para os dados completos e $P_{(-i)}$ a distribuição a *posteriori* sem a i-ésima observação. A medida de divergência ψ é definida como:

$$D_\psi(P,P_{(-i)}) = \int_{\vartheta\in\Theta} \psi\left(\frac{\pi(\vartheta|\mathscr{D}^{(-i)})}{\pi(\vartheta|\mathscr{D})}\right)\pi(\vartheta|\mathscr{D})\,d\vartheta,$$

em que $\psi(\cdot)$ é uma função convexa com $\psi(1)=0$. Várias escolhas de ψ são dadas em Peng e Dey (1995). Neste trabalho consideramos as seguintes escolhas: $\psi(z)=-\log(z)$ define a divergência de Kullback-Leibler (K-L), $\psi(z)=(z-1)\log(z)$ a distância J (ou a versão simétrica da divergência de K-L) e $\psi(z)=0,5|z-1|$ a distância variacional ou norma L_1.

Por meio da expressão (4.4), a medida de divergência ψ pode ser escrita como:

$$D_\psi(P,P_{(-i)}) = E_{\vartheta|\mathscr{D}}\left[\psi\left(\frac{CPO_i}{f(y_i|\vartheta)}\right)\right]. \tag{4.5}$$

A partir da expressão (4.5) podemos obter, por exemplo, a divergência K-L:

$$\begin{aligned} D_{\text{K-L}}(P,P_{(-i)}) &= -E_{\vartheta|\mathscr{D}}\{\log(CPO_i)\} + E_{\vartheta|\mathscr{D}}\{\log[f(y_i|\vartheta)]\} \\ &= -\log(CPO_i) + E_{\vartheta|\mathscr{D}}\{\log[f(y_i|\vartheta)]\}. \end{aligned}$$

As estimativas de Monte Carlo de $D_\psi(P,P_{(-i)})$ são dadas por:

$$\widehat{D_\psi}(P,P_{(-i)}) = \frac{1}{Q}\sum_{q=1}^{Q}\psi\left(\frac{\widehat{CPO}_i}{f(y_i|\vartheta^{(q)})}\right).$$

Logo, obtemos a estimativa da divergência $D_{\text{K-L}}(P,P_{(-i)})$ da seguinte forma:

$$\widehat{D_{\text{K-L}}}(P,P_{(-i)}) = -\log(\widehat{CPO}_i) + \frac{1}{Q}\sum_{q=1}^{Q}\log\left[f(y_i|\vartheta^{(q)})\right].$$

A divergência $D_\psi(P,P_{(-i)})$ pode ser interpretada como sendo o efeito de excluir o i-ésimo caso dos dados completos sobre a distribuição a *posteriori* de ϑ. Como discutido por Peng e Dey (1995) e Weiss (1996), pode ser difícil para um pesquisador definir um ponto de corte

para o valor da medida de divergência, de modo a determinar se um pequeno subconjunto de observações é influente ou não.

Neste trabalho utilizamos a proposta de Peng e Dey (1995) e Weiss (1996), a qual considera uma moeda viesada com probabilidade de sucesso p. Assim, a divergência ψ entre as moedas viesadas é dada por:

$$D_{\psi}(g_0, g_1) = \int \psi\left(\frac{g_0(x)}{g_1(x)}\right) g_1(x)dx,$$

em que $g_0(x) = p^x(1-p)^{1-x}$ e $g_1(x) = 0,5$, $x = 0$ ou 1. Se $D_{\psi}(g_0, g_1) = d_{\psi}(p)$, pode ser verificado que d_{ψ} satisfaz a seguinte equação:

$$d_{\psi}(p) = \frac{\psi(2p) + \psi(2(1-p))}{2}.$$

É possível ver que d_{ψ} aumenta à medida que p se afasta de $0,50$. Além disso, $d_{\psi}(p)$ é simétrico em torno de $p = 0,50$ e d_{ψ} atinge seu mínimo em $p = 0,50$. Neste ponto, $d_{\psi}(0,5) = 0$ e $g_0 = g_1$. Por exemplo, se considerarmos $p > 0,80$ (ou $p < 0,20$) um forte viés em uma moeda, então, $d_{L_1}(0,80) = 0,30$. Esta equação implica que o caso i é considerado influente quando $d_{L_1} > 0,30$. Assim, se a divergência de Kullback-Leibler for usada, pode-se considerar uma observação influente quando $d_{\text{K-L}} > 0,2231$. Da mesma forma, se a distância J for utilizada, uma observação pode ser considerada influente quando $d_{\text{J}} > 0,4159$. Utilizamos estes valores de corte para determinar se uma observação é influente ou não.

4.4 Estudo de simulação

Realizamos um estudo de simulação com dois objetivos: *i*) fazer um estudo de recuperação de parâmetros, ou seja, avaliar as estimativas dos parâmetros para os modelos propostos; *ii*) examinar o desempenho da abordagem diagnóstica proposta, considerando conjuntos de dados simulados sem nenhum, um, dois ou três casos perturbados.

No primeiro estudo, consideramos o modelo BIGcr sob os três esquemas de ativação: primeira, última e aleatória.

Para gerar os dados utilizados no estudo bayesiano, os procedimentos foram os mesmos quando realizado o estudo de simulação sob uma abordagem clássica ja apresentados na seção 3.2.

Para cada conjunto de dados gerados simulamos duas cadeias de tamanho 60.000 para cada parâmetro, desconsiderando as primeiras 10.000 iterações para eliminar o efeito dos valores iniciais e, para evitar problemas de autocorrelação, consideramos um espaçamento de tamanho 10, obtendo uma amostra final de tamanho 10.000 sobre a qual a inferência *a posteriori* é baseada. Para cada tamanho amostral foi obtida a média *a posteriori*, o viés, o EQM e a PC dos parâmetros. As probabilidades de cobertura foram calculadas por meio do intervalo de credibilidade de 95%.

A convergência das cadeias foi monitorada de acordo com os métodos recomendados por Cowles e Carlin (1996). Em todos os casos, a convergência foi verificada por meio do diagnóstico de Gelman-Rubin (GELMAN; RUBIN, 1992) sendo muito próximo a 1 ($\leq 1,01$), bem como os gráficos de traços. As Tabelas 7, 8 e 9 apresentam as estimativas médias obtidas, o vício, o EQM e a PC para os parâmetros dos modelo BIGcr sob o esquema de primeira ativação, última ativação e ativação aleatória, respectivamente. Podemos notar que o vício e o EQM das estimativas diminuem conforme o aumento do tamanho das amostras. Além disso, para todos os tamanhos amostrais considerados, a probabilidade de cobertura está próxima do valor nominal.

Tabela 7 – Estudo de simulação Bayesiano sob o esquema de primeira ativação.

n	Parâmetro	Média	Viés	EQM	PC
100	μ	0,5227	0,0227	0,0118	0,9470
	λ	2,0600	0,0600	0,1567	0,9630
	β_0	-0,5384	-0,0288	0,0960	0,9640
	β_1	0,7320	0,0320	0,1769	0,9550
200	μ	0,5080	0,0080	0,0013	0,9460
	λ	2,0394	0,0393	0,0013	0,9540
	β_0	-0,5143	-0,0143	0,0484	0,9410
	β_1	0,7202	0,0202	0,0825	0,9420
400	μ	0,5107	0,0107	0,0013	0,9499
	λ	2,0024	0,0025	0,0457	0,9449
	β_0	-0,5132	-0,0132	0,0245	0,9589
	β_1	0,6988	-0,0012	0,0435	0,9609
800	μ	0,5046	0,0046	0,00055	0,9330
	λ	2,0077	0,0077	0,0235	0,9489
	β_0	-0,5050	-0,0050	0,0140	0,9449
	β_1	0,7072	0,0072	0,0229	0,9549

Já no segundo estudo, consideramos conjuntos de dados simulados com nenhum, um ou dois casos perturbados para examinar o desempenho das medidas de diagnóstico consideradas.

Uma amostra de tamanho 300 foi gerada do modelo BIGcr sob o esquema de ultima ativação.

Para criar uma observação inuente no conjunto de dados, escolhemos um, dois ou três casos selecionados de forma aleatória, e perturbamos a variável resposta da seguinte forma: e $y_i = y_i + 5S_y$, $i = 50$, 125 e 200, em que S_y é o desvio padrão dos y_i's. O modelo BIGcr sob o esquema de última ativação foi ajustado.

Na Tabela 10 apresentamos a estimativa média, o desvio padrão (DP) e a diferença relativa (DR, em porcentagem). O conjunto de dados (a) denota o conjunto de dados original simulado sem perturbação e os conjuntos de dados (b)-(h) denotam conjuntos de dados com casos perturbados (que estão apresentados na segunda coluna). A diferença relativa é a calculada pelas diferenças das estimativas obtidas nos casos (b) a (h) em comparação com a estimativas

no caso (a). Observamos que as inferências *a posteriori* são sensíveis após a perturbação do(s) caso(s) selecionado(s).

Tabela 8 – Estudo de simulação Bayesiano sob o esquema de última ativação.

n	Parâmetro	Média	Viés	EQM	PC
100	μ	0,5040	0,0040	0,0013	0,9500
	λ	2,0988	0,0988	0,2852	0,9540
	β_0	-0,5178	-0,0178	0,1044	0,9500
	β_1	0,7457	0,0457	0,1965	0,9480
200	μ	0,5008	0,0008	0,00061	0,9610
	λ	2,0577	0,0577	0,1264	0,9539
	β_0	-0,5082	-0,0082	0,0490	0,9599
	β_1	0,7113	0,0113	0,0915	0,9479
400	μ	0,4998	-0,0002	0,0003	0,9540
	λ	2,0269	0,0275	0,0623	0,9520
	β_0	-0,5042	-0,0042	0,0253	0,9450
	β_1	0,7052	0,0052	0,0502	0,9360
800	μ	0,5003	0,0003	0,0002	0,9540
	λ	2,0227	0,0219	0,0311	0,9500
	β_0	-0,5036	-0,0036	0,0124	0,9470
	β_1	0,7099	0,0099	0,0229	0,9520

Tabela 9 – Estudo de simulação Bayesiano sob o esquema de ativação aleatória.

n	Parâmetro	Média	Viés	EQM	PC
100	μ	0,5039	0,0039	0,0013	0,9580
	λ	2,0962	0,0962	0,2029	0,9530
	β_0	-0,5361	-0,0361	0,1025	0,9430
	β_1	0,7534	0,0534	0,2100	0,9350
200	μ	0,5009	0,0009	0,0006	0,9490
	λ	2,0655	0,0655	0,0902	0,9500
	β_0	-0,5220	-0,0220	0,0496	0,9350
	β_1	0,7324	0,0324	0,0943	0,9440
400	μ	0,4998	-0,0002	0,0003	0,9540
	λ	2,0269	0,0275	0,0623	0,9520
	β_0	-0,5042	-0,0042	0,0253	0,9450
	β_1	0,7052	0,0052	0,0502	0,9360
800	μ	0,4994	-0,0006	0,0001	0,9460
	λ	2,0303	0,0303	0,0223	0,9360
	β_0	-0,4949	0,0050	0,0124	0,9460
	β_1	0,6921	-0,0089	0.0208	0,9550

Tabela 10 – Média, DP e DR das estimativas dos parâmetros obtidas a partir do ajuste do modelo BIGcr sob o esquema de última ativação

		Parâmetros			
		$\log(\mu)$	$\log(\lambda)$	β_0	β_1
Nomes dos dados	Caso(s) perturbado(s)	Média (DP) DR(%)	Média (DP) DR(%)	Média (DP) DR(%)	Média (DP) DR(%)
a	Nenhum	-0,660 (0,038) -	0,826 (0,138) -	-0,550 (0,178) -	0,802 (0,240) -
b	50	-0,610 (0,048) 8,257	0,491 (0,145) 68,255	-0,538 (0,178) 2,218	0,762 (0,243) 5,266
c	125	-0,611 (0,047) 7,985	0,466 (0,140) 77,468	-0,553 (0,182) 0,618	0,784 (0,245) 2,288
d	200	-0,611 (0,047) 8,017	0,467 (0,142) 76,772	-0,545 (0,183) 0,974	0,775 (0,247) 3,535
e	{50,125}	-0,555 (0,054) 18,884	0,230 (0,146) 259,411	-0,545 (0,184) 0,881	0,750 (0,245) 6,887
f	{50,200}	-0,553 (0,055) 19,392	0,226 (0,147) 266,248	-0,548 (0,184) 0,435	0,754 (0,246) 6,377
g	{125,200}	-0,555 (0,056) 18,894	0,197 (0,148) 318,546	-0,557 (0,186) 1,221	0,774 (0,249) 3,621
h	{50,125,200}	-0,475 (0,066) 38,909	0,025 (0,148) 3.200,234	0,743 (0,763) 174,022	-0,134 (0,110) 697,16

A Tabela 11 mostra os valores obtidos dos critérios DIC, EAIC, EBIC e LPML para o caso sem perturbação (amostra original) e para cada versão perturbada. Como esperado, o conjunto de dados (a) foi o que melhor se ajustou ao modelo BIGcr sob o esquema de última ativação.

Tabela 11 – Critérios Bayesianos para conjunto de dados

Nomes	Critérios Bayesianos			
dos dados	EAIC	EBIC	DIC	LPML
a	568,696	583,511	564,562	-282,348
b	602,198	617,013	598,169	-302,021
c	608,372	623,187	604,338	-305,523
d	608,545	623,360	604,566	-306,344
e	627,813	642,628	623,784	-315,682
f	627,964	642,779	623,946	-316,288
g	633,681	648,496	629,689	-319,138
h	651,833	666,648	773,872	-327,436

Consideramos agora a amostra das distribuições a *posteriori* dos parâmetros do modelo BIGcr sob o esquema de última ativação para calcular as medidas de divergência ψ dadas em (4.3). Os resultados na Tabela 12 mostram que antes da perturbação (conjunto de dados (a)) os casos selecionados não são inuentes de acordo com todas as medidas de divergência ψ. No

entanto, após a perturbação (conjunto de dados (b)-(h)) as medidas aumentam, o que indica que os casos perturbados são influentes de acordo com o ponto de corte especificado.

Tabela 12 – Critérios Bayesianos para conjunto de dados

Nome dos dados	Caso(s) perturbado(s)	Medidas de divergência		
		K-L	distância J	norma L_1
a	50	0,005	0,010	0,040
	125	0,021	0,043	0,082
	200	0,024	0,048	0,087
b	50	3,339	6,234	0,813
c	125	3,771	7,194	0,831
d	200	4,489	8,672	0,875
e	50	1,882	3,639	0,674
	125	2,482	4,791	0,739
f	50	2,132	4,325	0,714
	200	2,792	5,553	0,778
g	125	2,344	4,978	0,735
	200	2,55	5,39	0,757
h	50	1,182	2,389	0,567
	125	1,556	3,13	0,632
	200	1,696	3,411	0,653

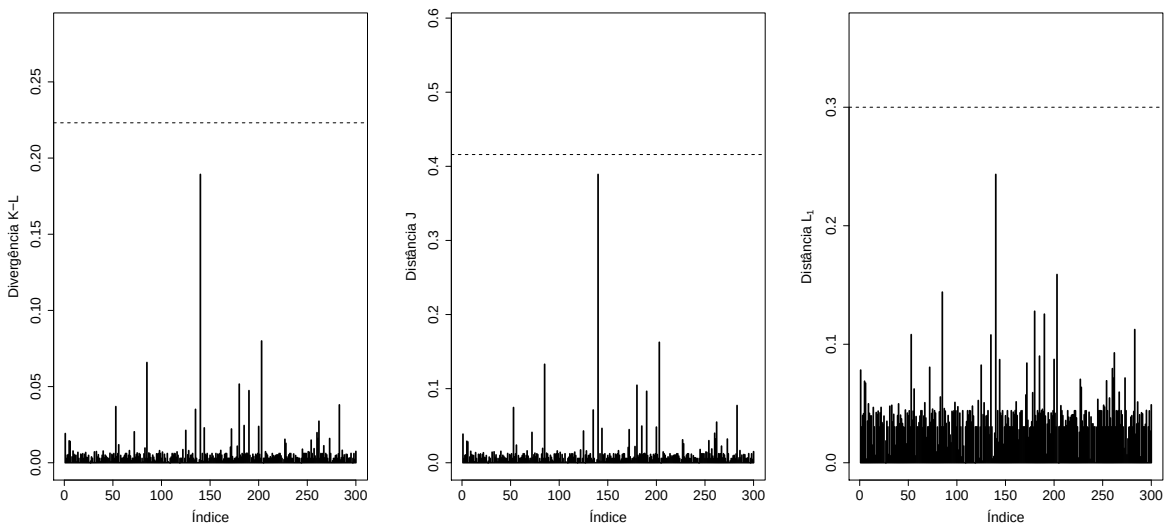
As Figuras 8, 9 e 10 mostram as medidas de divergência ψ para os casos (a), (b) e (h), respectivamente. Na Figura 8, observamos que não foram detectados pontos influentes como era esperado, pois nenhuma observação foi perturbada no conjunto de dados (a). Nas Figuras 9 e 10 as medidas de divergências detectaram a(s) observação(ões) perturbada(s) como ponto(s) influente(s). Os mesmos resultados foram obtidos para os casos (c) a (g) apresentados no Apêndice B nas Figuras 19 a 23, respectivamente.

4.5 Aplicação aos dados de melanoma

Nesta seção vamos novamente realizar a aplicação do modelo proposto a um conjunto de dados reais de um ensaio clínico de melanoma cutâneo de fase III descrito na Seção 3.3. A inferência será feita sob o ponto de vista Bayesiano. De forma análoga ao que foi feito na Seção 3.3 assumimos o modelo BIGcr sob o esquema de última ativação.

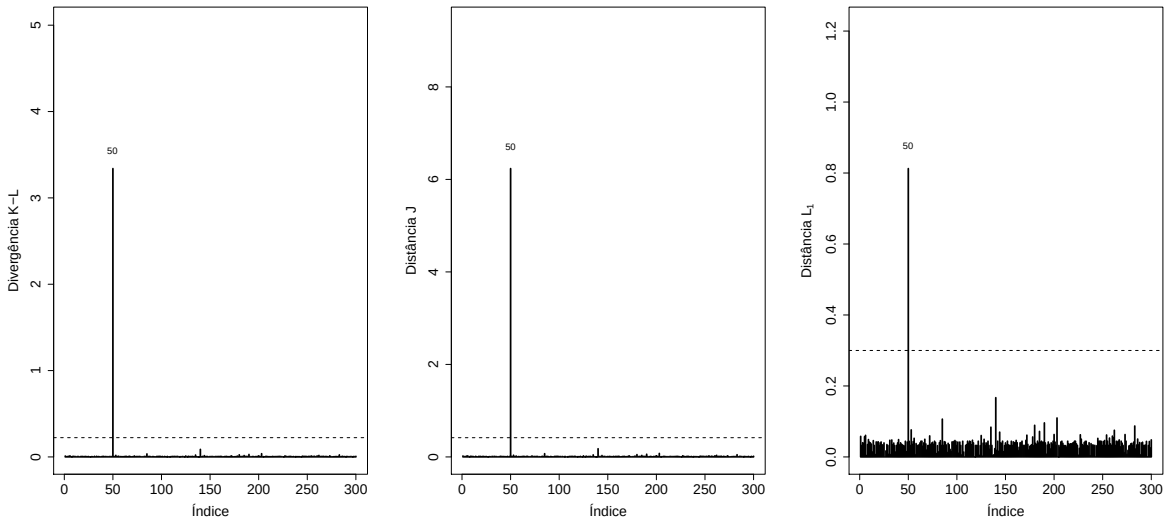
Simulamos duas cadeias de tamanho 40.000 para cada parâmetro, desconsiderando as primeiras 5.000 iterações para eliminar o efeito dos valores iniciais e, para evitar problemas de autocorrelação, consideramos um espaçamento de tamanho 10, obtendo uma amostra final de tamanho 7.000 sobre a qual a inferência *a posteriori* é baseada. Novamente, a convergência das cadeias foi monitorada de acordo com os métodos recomendados por (COWLES; CARLIN, 1996) . Em todos os casos, a convergência foi verificada por meio do diagnóstico de Gelman-

Figura 8 – Divergência ψ para o conjunto de dados (a).



Fonte: Elaborada pela autora.

Figura 9 – Divergência ψ para o conjunto de dados (b).

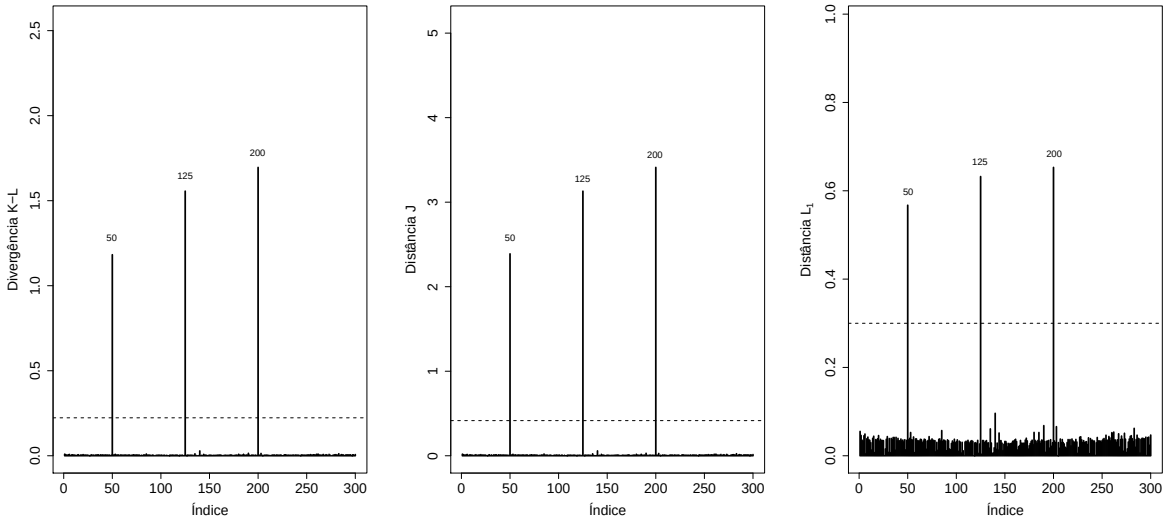


Fonte: Elaborada pela autora.

Rubin (GELMAN; RUBIN, 1992) sendo muito próximo a 1 ($\leq 1,01$), bem como os gráficos de traços.

Para cada parâmetro, na Tabela 13 apresentamos a média *a posteriori*, o desvio padrão e o intervalo de credibilidade de 95% (IC 95%). Podemos observar que as estimativas médias *a posteriori* foram próximas das obtidas por máxima verossimilhança. Também, a covariável sexo não foi significativa a 5%.

Figura 10 – Divergência ψ para o conjunto de dados (h).



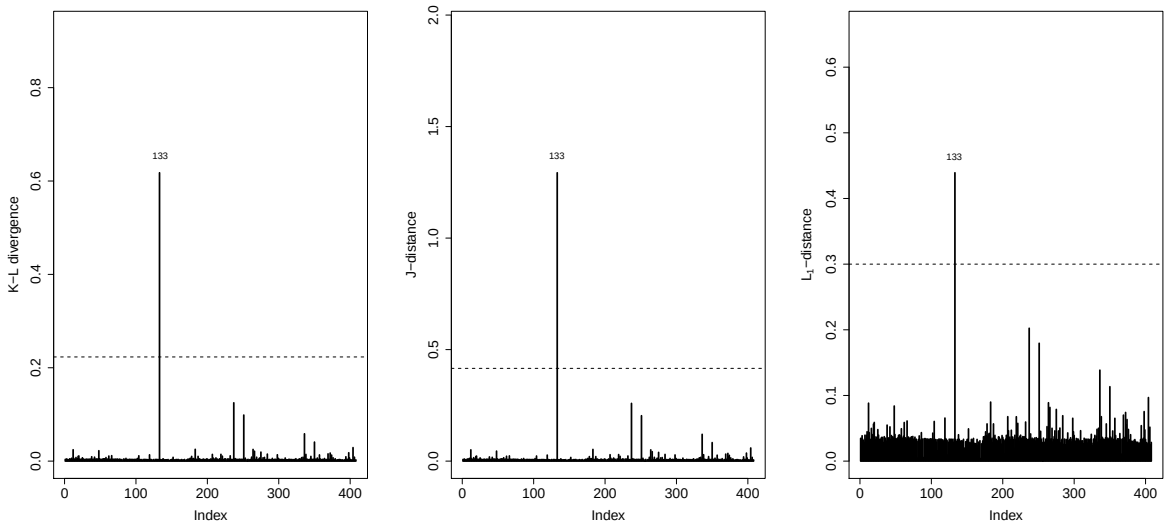
Fonte: Elaborada pela autora.

Tabela 13 – Resumo a *posteriori* dos parâmetros do modelo BIGcr sob o esquema de última ativação - dados melanoma

Parâmetro	Média	DP	IC (95%)
$\log(\mu)$	0,3943	0,3583	(-0,178, 1,183)
$\log(\lambda)$	-1,5482	0,1961	(-1,931, -1,1888)
β_0	-1,2448	0,3460	(-1,922, -0,6528)
β_1	0,3043	0,3890	(-0,554, 1,026)

Foram computadas as medidas de divergência ψ, apresentadas na Figura 11. Podemos observar que todas as três medidas detectaram a observação 133 como influente

Figura 11 – Divergência ψ para o conjunto de dados de melanoma.



Fonte: Elaborada pela autora.

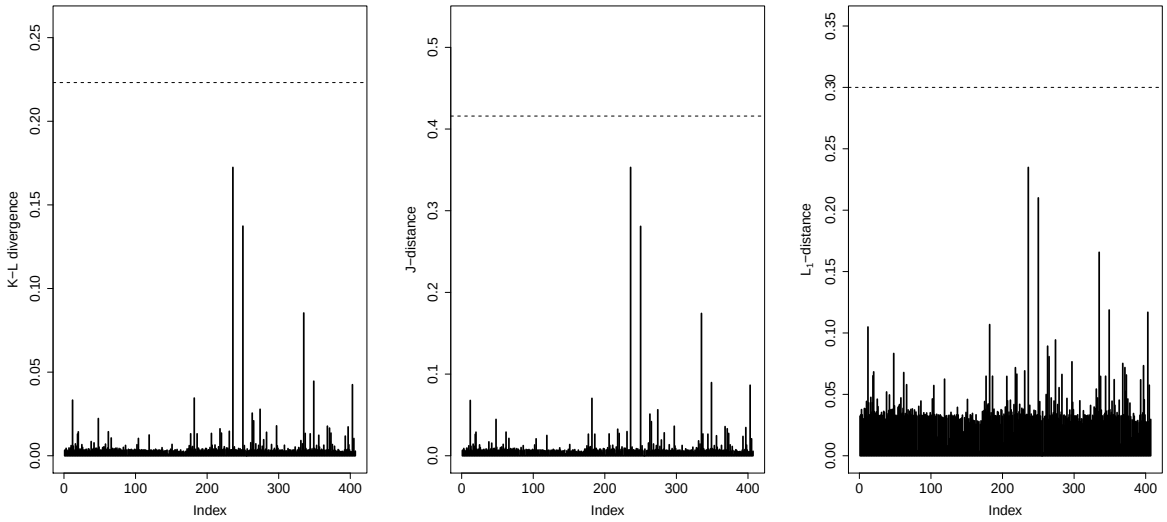
Assim, ajustamos novamente o modelo BIGcr sob o esquema de última ativação ao dados de melanoma, excluída a observação 133. A Tabela 14 apresenta a média *a posteriori*, o desvio padrão e o intervalo de credibilidade de 95% (IC 95%) do seus parâmetros. Comparando com os resultados apresentados na Tabela 13 podemos observar uma pequena diferença nas estimativas médias dos parâmetros, exceto para o parâmetro β_1.

Tabela 14 – Resumo a *posteriori* dos parâmetros do modelo BIGcr sob o esquema de última ativação - dados de melanoma excluída a observação 133.

Parâmetro	Média	DP	IC (95%)
$\log(\mu)$	0,181	0,287	(-0,231, 0,882)
$\log(\lambda)$	-1,332	0,183	(-1,758, -1,017)
β_0	-0,992	0,295	(-1,693, -0,537)
β_1	0,302	0,320	(-0,355, 0,901)

As três medidas divergência ψ foram novamente obtidas. Todas as medidas não detectaram observações influentes, como podemos observar no gráco de índices das três medidas de divergência ψ apresentado na Figura 12.

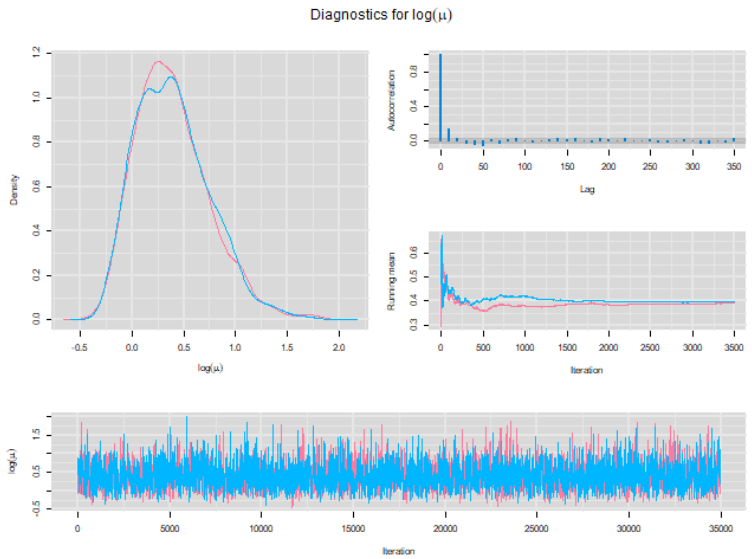
Figura 12 – Divergência ψ para o conjunto de dados de melanoma excluída a observação 133.



Fonte: Elaborada pela autora.

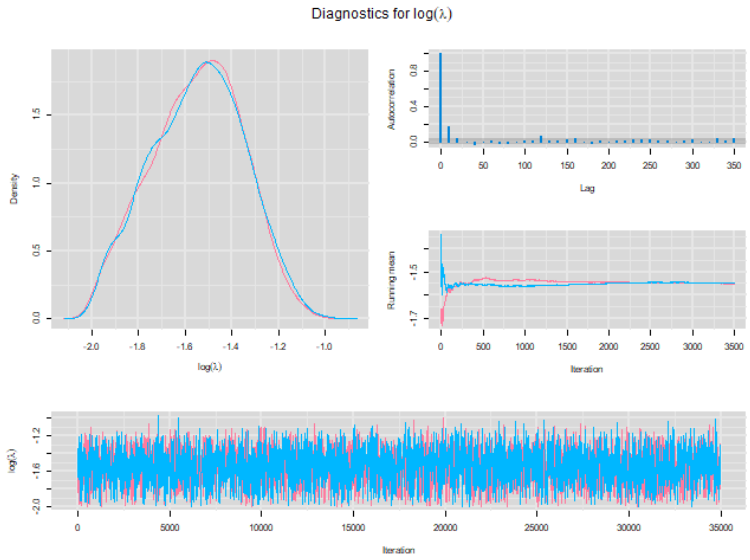
Podemos observar a convergência das cadeias nas Figuras 13, 14, 15 e 16 que apresentam a densidade estimada, o gráfico de autocorrelação e o gráfico de traços para $\log(\mu)$, $\log(\lambda)$, β_0 e β_1, respectivamente.

Figura 13 – Diagnóstico de convergência para $\log(\mu)$.



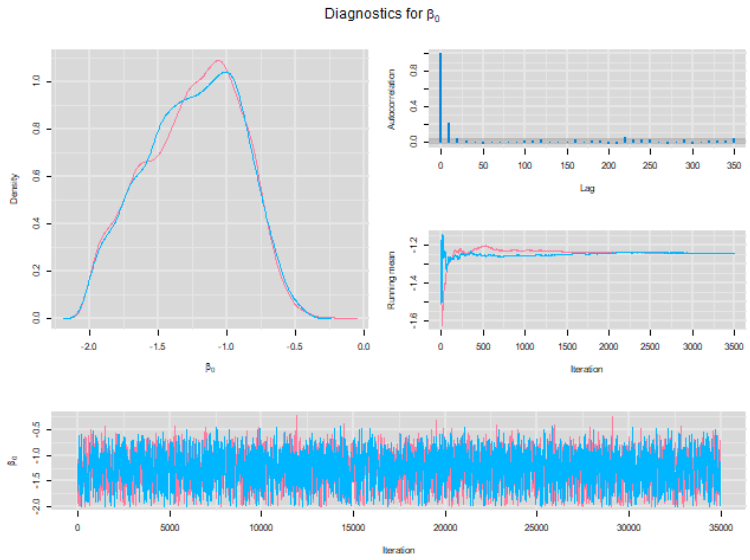
Fonte: Elaborada pela autora.

Figura 14 – Diagnóstico de convergência para $\log(\lambda)$.



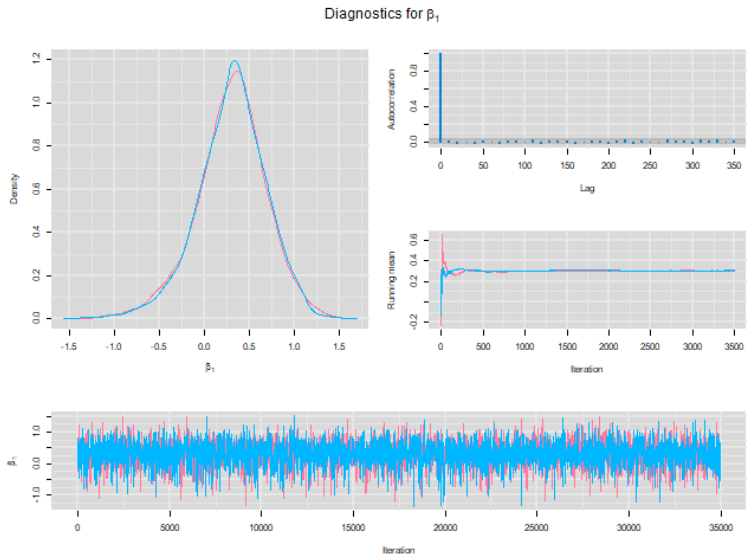
Fonte: Elaborada pela autora.

Figura 15 – Diagnóstico de convergência para $beta_0$.



Fonte: Elaborada pela autora.

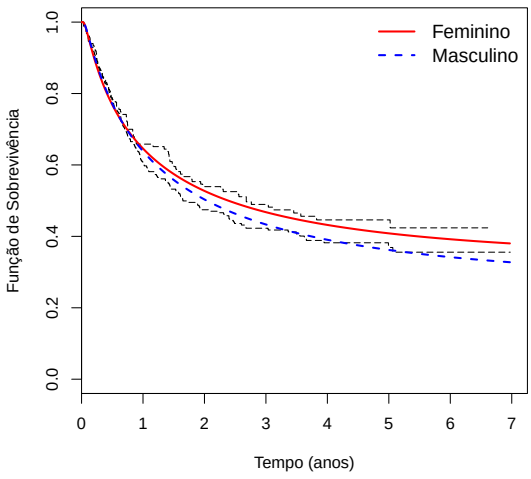
Figura 16 – Diagnóstico de convergência para $beta_1$.



Fonte: Elaborada pela autora.

A Figura 17 mostra as curvas de Kaplan-Meier e as curvas de sobrevivência BIG_{cr} estimadas, estratificadas pelo sexo para os dados de melanoma excluída a observação 133. Também, a curva de sobrevivência do modelo BIG_{cr} sob o esquema de última ativação aleatória estimada, na qual é possível observar o bom ajuste do modelo aos dados.

Figura 17 – Curvas de Kaplan-Meier e curvas de sobrevivência BIGcr estimadas para os dados de melanoma cutâneo excluída a observação 133.

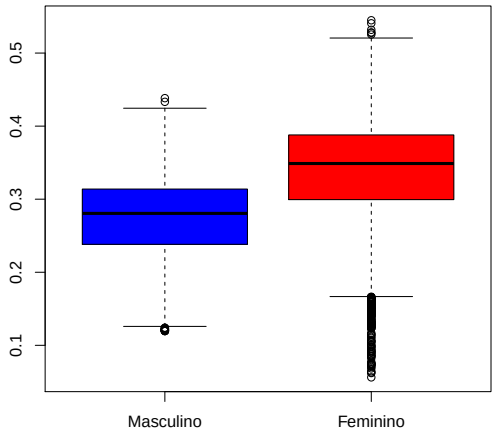


Fonte: Elaborada pela autora.

Por fim, a Figura 18 apresenta os *box plots* para as médias *a posteriori* das proporções

de curados dicotomizado pelo sexo, na qual observamos uma proporção um pouco maior para o sexo feminino em comparação com o sexo masculino.

Figura 18 – *box plots* para as médias a *posterioris* das proporções de curados dicotomizado pelo sexo.



Fonte: Elaborada pela autora.

CAPÍTULO

5

CONSIDERAÇÕES FINAIS

Neste trabalho propomos o modelo BIGcr para a modelagem de dados de sobrevivência com proporção de cura, considerando que o evento de interesse pode ser causado por três mecanismos de ativação latentes: primeira, última e ativação aleatória.

Assumimos que os tempos de ativação latentes são provenientes de um distribuição Inversa Gaussiana de parâmetros μ e λ e o número M de causas necessárias para produzir o evento de interesse segue uma distribuição Bell (CASTELLARES; FERRARI; LEMONTE, 2018) de parâmetro θ.

Para a estimação dos parâmetros de interesse assumimos as abordagens clássica e Bayesiana. Sob o ponto de vista clássico consideramos o método de Máxima Verossimilhança para a estimação dos parâmetros do modelo BIGcr e sob o ponto de vista bayesiano foram consideradas distribuições a *priori* pouco informativas e métodos de MCMC para a estimação dos parâmetro.

Na abordagem Bayesiana consideramos distribuições a *priori* não-informativas. O procedimento de estimação dos parâmetros foi feito utilizando métodos de MCMC. Além disso, utilizamos o diagnóstico de observações influentes, baseado na divergência ψ (PENG; DEY, 1995; WEISS, 1996). No Apêndice C apresentamos alguns códigos utilizados neste trabalho.

Ilustramos a performance dos procedimentos de estimação dos parâmetros do modelo proposto por meio de dados simulados e a um conjunto de dados reais de melanoma, no qual o modelo BIGcr sob o esquema de última ativação forneceu o melhor ajuste.

Como trabalho futuro, outras distribuições mais flexíveis podem ser consideradas para modelar os tempos de ativação latente. Por exemplo, o modelo Kumaraswamy complementary Weibull geometric (Kw–CWG) proposto por Afify *et al.* (2017), que contém 23 submodelos (listado na Tabela 1 em (AFIFY *et al.*, 2017)) como casos especiais.

Além disso, podemos obter uma extensão do modelo proposto para o caso multivariado

assumindo distribuições marginais BIGcr e função distribuição acumulada conjunta dada por uma função cópula (NELSEN, 2006), como em Suzuki *et al.* (2011), Suzuki, Louzada e Cancho (2013), Louzada, Suzuki e Cancho (2013), Oliveira, Suzuki e Saraiva (2014), Achcar, Moala e Tarumoto (2015), Biondo e Suzuki (2016), Cruz *et al.* (2017), Ribeiro, Suzuki e Saraiva (2017), Romeo, Meyer e Gallardo (2018) e Peres, Achcar e Martinez (2020).

Neste trabalho, a metodologia proposta é puramente paramétrica. Kim, Chen e Dey (2011) utilizaram, no entanto, o modelo Exponencial por partes para a distribuição cumulativa, que é semiparamétrica, levando a uma modelagem mais flexível. Tal abordagem pode ser investigada no contexto da modelagem aqui estudada.

REFERÊNCIAS

ACHCAR, J. A.; MOALA, F. A.; TARUMOTO, M. H. Z. A bivariate generalized exponential distribution derived from copula functions in the presence of censored data and covariates. **Pesquisa Operacional (Online)**, v. 35, p. 165–186, 2015. Citado na página 70.

AFIFY, A. Z.; CORDEIRO, G. M.; BUTT, N. S.; ORTEGA, E. M. M.; SUZUKI, A. K. A new lifetime model with variable shapes for the hazard rate. **Brazilian Journal of Probability and Statistics**, v. 31, p. 516–541, 2017. Citado na página 69.

BACANLI, S.; DEMIRHAN, Y. P. A group sequential test for the inverse gaussian mean. **Statistical Papers**, v. 49, n. 2, p. 377–386, 2008. Citado na página 37.

BALAKRISHNA, N.; RAHUL, T. Inverse gaussian distribution for modeling conditional durations in finance. **Communications in Statistics - Simulation and Computation**, Taylor Francis, v. 43, n. 3, p. 476–486, 2014. Disponível em: <https://doi.org/10.1080/03610918.2012.705938>. Citado na página 37.

BALAKRISHNAN, N.; BARUI, S.; MILIENOS, F. S. Proportional hazards under conway–maxwell-poisson cure rate model and associated inference. **Statistical Methods in Medical Research**, v. 26, p. 2055–2077, 2017. Citado na página 22.

BALAKRISHNAN, N.; KOUTRAS, M. V.; MILIENOS, F. S. A weighted poisson distribution and its application to cure rate models. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 47, n. 17, p. 4297–4310, 2018. Citado na página 22.

BALKA, J.; DESMOND, A.; MCNICOLAS, P. Review and implementation of cure models based on first hitting times for wiener processes. **Lifetime Data Analysis**, v. 15, p. 147–176, 2009. Citado na página 38.

BAO, Y.; CANCHO, V. G.; DEY, D. K.; BALAKRISHNAN, N.; SUZUKI, A. K. Power series cure rate model for spatially correlated interval-censored data based on generalized extreme value distribution. **Journal of Computational and Applied Mathematics**, v. 364, p. 112362, 2020. Citado na página 22.

BAO, Y.; CANCHO, V. G.; LOUZADA, F.; SUZUKI, A. K. Cure rate proportional odds models with spatial frailties for interval-censored data. **Communications for Statistical Applications and Methods**, v. 24, p. 605–625, 2017. Citado na página 22.

____. Cure rate proportional odds models with spatial frailties for interval-censored data. **Advances in Data Science and Adaptive Analysis**, v. 11, p. 1950005, 2019. Citado na página 22.

BARRIGA, G. D. C.; DEY, D. K.; CANCHO, V. G.; SUZUKI, A. K. Bayesian analysis of birnbaum-saunders survival model with cure fraction under a variety of activation mechanism. **Model Assisted Statistics and Applications**, v. 15, p. 35–51, 2020. Citado nas páginas 22 e 30.

BARRIGA, G. D. C.; SUZUKI, A. K.; CANCHO, V. G.; LOUZADA, F. A new class of cure rate survival models: Properties, inference and applications. **Advances in Data Science and Adaptive Analysis**, 2021. Disponível em: <https://doi.org/10.1142/S2424922X21500017>. Citado na página 22.

BELL, E. T. Exponential numbers. **The American Mathematical Monthly**, Taylor Francis, v. 41, n. 7, p. 411–419, 1934. Disponível em: <https://doi.org/10.1080/00029890.1934.11987615>. Citado na página 32.

____. Exponential polynomials. **Annals of Mathematics**, v. 35, n. 2, p. 258–277, 1934. Citado na página 32.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor & Francis, v. 47, n. 259, p. 501–515, 1952. Citado na página 28.

BIONDO, T. R.; SUZUKI, A. K. Modelos de sobrevivência bivariados derivados da cópula arquimediana de clayton: Uma abordagem bayesiana. **Matemática e Estatística em Foco**, v. 4, p. 87–102, 2016. Citado na página 70.

BROOKS, S. P. Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde (2002). **Journal of the Royal Statistical Society B**, v. 64, p. 616–618, 2002. Citado na página 55.

CALSAVARA, V. F.; RODRIGUES, A. S.; ROCHA, R.; TOMAZELLA, V.; LOUZADA, F. Defective regression models for cure rate modeling with interval-censored data. **Biometrical Journal**, v. 61, n. 4, p. 841–859, 2019. Citado na página 22.

CANCHO, V.; ZAVALETA, K. E. C.; MACERA, M. A. C.; SUZUKI, A. K.; LOUZADA, F. A bayesian cure rate model with dispersion induced by discrete frailty. **Communications for Statistical Applications and Methods**, v. 25, p. 471–488, 2018. Citado na página 22.

CANCHO, V. G.; BARRIGA, G. D. C.; CORDEIRO, G. M.; ORTEGA, E. M. M.; SUZUKI, A. K. Bayesian survival model induced by frailty for lifetime with long-term survivors. **Statistica Neerlandica**, 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/stan.12236>. Citado na página 22.

CANCHO, V. G.; CASTRO, M. de; DEY, D. K. Long-term survival models with latent activation under a flexible family of distributions. **Brazilian Journal of Probability and Statistics**, JSTOR, p. 585–600, 2013. Citado na página 22.

CANCHO, V. G.; MACERA, M. A. C.; SUZUKI, A. K.; LOUZADA, F.; ZAVALETA, K. E. C. A new long-term survival model with dispersion induced by discrete frailty. **Lifetime Data Analysis**, v. 26, p. 221–244, 2020. Citado na página 22.

CARLIN, B. P.; LOUIS, T. A. **Bayes and Empirical Bayes Methods for Data Analysis**. second. Boca Raton: Chapman & Hall/CRC, 2001. Citado na página 55.

CASTELLARES, F.; FERRARI, S. L. P.; LEMONTE, A. J. On the bell distribution and its associated regression model for count data. **Applied Mathematical Modelling**, Elsevier, v. 56, p. 172–185, 2018. Citado nas páginas 22, 32, 35 e 69.

CHEN, M.-H.; SHAO, Q.-M.; IBRAHIM, J. **Monte Carlo Methods in Bayesian Computation**. New York: Springer–Verlag, 2000. Citado na página 55.

CHEN, X.; JI, G.; SUN, X.; LI, Z. Inverse gaussian–based model with measurement errors for degradation analysis. **Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability**, v. 233, n. 6, p. 1086–1098, 2019. Citado na página 37.

CHHIKARA, R.; FOLKS, L. **The Inverse Gaussian Distribution: Theory, Methodology, and Applications**. New York: Marcel Dekker, 1989. Citado na página 37.

CHOI, S.; HUANG, X.; CORMIER, J. N.; DOKSUM, K. A. A semiparametric inverse-gaussian model and inference for survival data with a cured proportion. **Canadian Journal of Statistics**, v. 42, n. 4, p. 635–649, 2014. Citado na página 37.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. Edgard Blücher, 2006. (ABE - Projeto Fisher). ISBN 9788521203841. Disponível em: <https://books.google.com.br/books?id=g0-uOgAACAAJ>. Citado nas páginas 25, 26, 27 e 28.

COOK, R. D.; WEISBERG, S. **Residuals and Influence in Regression**. Boca Raton, FL: Chapman & Hall/CRC, 1982. Citado na página 56.

COONER, F.; BANERJEE, S.; CARLIN, B. P.; SINHA, D. Flexible cure rate modeling under latent activation schemes. **Journal of the American Statistical Association**, Taylor & Francis, v. 102, n. 478, p. 560–572, 2007. Citado nas páginas 22, 23 e 30.

CORDEIRO, G. M.; CANCHO, V. G.; ORTEGA EDWIN M, M.; BARRIGA GLADYS D, C. A model with long-term survivors: negative binomial birnbaum-saunders. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 45, n. 5, p. 1370–1387, 2016. Citado na página 22.

CORLESS, R. M.; GONNET, G. H.; HARE, D. E. G.; JEFFREY, D. J.; KNUTH, D. E. On the Lambert W function. **Advances in Computational Mathematics**, v. 5, p. 329–359, 1996. Citado na página 36.

COWLES, M. K.; CARLIN, B. P. Markov chain Monte Carlo convergence diagnostics: A comparative review. **Journal of the American Statistical Association**, v. 91, p. 883–904, 1996. Citado nas páginas 58 e 61.

CRUZ, J. N.; ORTEGA, E. M. M.; CORDEIRO, G. M.; SUZUKI, A. K.; MIALHE, F. L. Bivariate odd-log-logistic-weibull regression model for oral health-related quality of life. **Communications for Statistical Applications and Methods**, v. 24, p. 271–290, 2017. Citado na página 70.

DENWOOD, M. J. runjags: An r package providing interface utilities, model templates, parallel computing methods and additional distributions for mcmc models in jags. **Journal of Statistical Software**, v. 71, n. 9, p. 1–25, 2016. Citado na página 23.

FELLER, W. **An introduction to probability theory and its applications**. [S.l.]: Wiley, New York, 1957. v. 1. Citado na página 29.

FOLKS, J. L. **Inverse Gaussian distribution. The Encyclopedia of Statistical Sciences**. 6. ed. New York: Wiley, 2007. Citado na página 37.

FOLKS, J. L.; CHHIKARA, R. S. The inverse gaussian distribution and its statistical application - a review. **Journal of the Royal Statistical Society Series B**, v. 40, n. 3, p. 263–289, 1978. Citado nas páginas 37 e 38.

GALLARDO, D. I.; GóMEZ, Y. M.; ARNOLD, B. C.; GóMEZ, H. W. The pareto iv power series cure rate model with applications. **SORT-Statistics and Operations Research Transactions**, v. 1, p. 297–318, 2017. Citado na página 22.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, p. 457–511, 1992. Citado nas páginas 58 e 62.

GHITANY, M.; ALQALLAF, F.; AL-MUTAIRI, D.; HUSAIN, H. A two-parameter weighted lindley distribution and its applications to survival data. **Mathematics and Computers in Simulation**, v. 81, n. 6, p. 1190–1201, 2011. ISSN 0378-4754. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378475410003873>. Citado na página 32.

GHITANY, M. E.; MAZUCHELI, J.; MENEZES, A. F. B.; ALQALLAF, F. The unit-inverse gaussian distribution: A new alternative to two-parameter distributions on the unit interval. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 48, n. 14, p. 3423–3438, 2019. Disponível em: <https://doi.org/10.1080/03610926.2018.1476717>. Citado na página 37.

GILAVERT, P.; SUZUKI, A. K.; SARAIVA, E. F. O modelo liindely-weibull com proporção de cura: Uma abordagem bayesiana. **Revista Brasileira de Biometria**, v. 36, p. 998–1022, 2018. Citado nas páginas 22 e 30.

GUNES, H.; DIETZ, D. C.; AUCLAIR, P. F.; MOORE, A. H. Modified goodness-of-fit tests for the inverse gaussian distribution. **Computational Statistics Data Analysis**, v. 24, n. 1, p. 63–77, 1997. ISSN 0167-9473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947396000564>. Citado na página 37.

GUPTA, R. D.; KUNDU, D. A new class of weighted exponential distributions. **Statistics**, Taylor Francis, v. 43, n. 6, p. 621–634, 2009. Disponível em: <https://doi.org/10.1080/02331880802605346>. Citado na página 32.

GóMEZ–DéNIZ, E.; PéREZ–RODRíGUEZ, J. V. Mixture inverse gaussian for unobserved heterogeneity in the autoregressive conditional duration model. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 46, n. 18, p. 9007–9025, 2017. Disponível em: <https://doi.org/10.1080/03610926.2016.1200094>. Citado na página 37.

HANAGAL, D. D.; BHAMBURE, S. M. Modeling bivariate survival data using shared inverse gaussian frailty model. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 45, n. 17, p. 4969–4987, 2016. Disponível em: <https://doi.org/10.1080/03610926.2014.901380>. Citado na página 37.

HANAGAL, D. D.; PANDEY, A. Correlated inverse gaussian frailty models for bivariate survival data. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 49, n. 4, p. 845–863, 2020. Disponível em: <https://doi.org/10.1080/03610926.2018.1549256>. Citado na página 37.

HE, Z.; EMURA, T. The com-poisson cure rate model for survival data. **Computational Aspects Journal of the Chinese Statistical Association**, v. 57, p. 1–42, 2019. Citado na página 22.

IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. **Bayesian survival analysis**. Springer, 2001. (Springer series in statistics). ISBN 9783540952770. Disponível em: <https://books.google.com.br/books?id=flVcSQAACAAJ>. Citado na página 22.

JAYALATH, K. P.; CHHIKARA, R. S. Survival analysis for the inverse gaussian distribution with the gibbs sampler. **Journal of Applied Statistics**, Taylor Francis, v. 0, n. 0, p. 1–20, 2020. Disponível em: <https://doi.org/10.1080/02664763.2020.1828314>. Citado na página 37.

JAZI, M. A.; LAI, C.-D.; ALAMATSAZ, M. H. A discrete inverse weibull distribution and estimation of its parameters. **Statistical Methodology**, v. 7, n. 2, p. 121–132, 2010. ISSN 1572-3127. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1572312709000707>. Citado na página 32.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous Univariate Distributions**. 2nd. ed. New York: Wiley, 1994. Citado na página 37.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457–481, 1958. Citado na página 28.

KARAMOOZIAN, A.; BANESHI, M.; BAHRAMPOUR, A. Bayesian mixture cure rate frailty models with an application to gastric cancer data. **Statistical Methods in Medical Research**, 2020. Citado na página 22.

KIM, S.; CHEN, M. H.; DEY, D. K. A new threshold regression model for survival data with a cure fraction. **Lifetime data analysis**, Springer, v. 17, n. 1, p. 101–122, 2011. Citado na página 70.

KINAT, S.; AMIN, M.; MAHMOOD, T. Glm-based control charts for the inverse gaussian distributed response variable. **Quality and Reliability Engineering International**, v. 36, n. 2, p. 765–783, 2020. Citado na página 37.

KIRKWOOD, J. M.; IBRAHIM, J. G.; SONDAK, V. K.; RICHARDS, J.; FLAHERTY, L. E.; ERNSTOFF, M. S.; SMITH, T. J.; RAO, U.; STEELE, M.; BLUM, R. H. High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S9111/C9190. **Journal of Clinical Oncology**, v. 18, p. 2444–2458, 2000. Citado nas páginas 23 e 49.

KIRKWOOD, J. M.; MANOLA, J.; IBRAHIM, J.; SONDAK, V.; ERNSTOFF, M. S.; RAO, U. A pooled analysis of eastern cooperative oncology group and intergroup trials of adjuvant high-dose interferon for melanoma. **Clinical Cancer Research**, AACR, v. 10, n. 5, p. 1670–1677, 2004. Citado na página 22.

KOTZ, S.; LEIVA, V.; SANHUEZA, A. Two new mixture models related to the inverse gaussian distribution. **Methodology and Computing in Applied Probability**, v. 12, p. 199–212, 2010. Citado na página 38.

KOUTRAS, M. V.; MILIENOS, F. S. A flexible family of transformation cure rate models. **Statistics in Medicine**, v. 36, n. 16, p. 2559–2575, 2017. Citado na página 22.

KRISHNA, H.; Singh Pundir, P. Discrete burr and discrete pareto distributions. **Statistical Methodology**, v. 6, n. 2, p. 177–188, 2009. ISSN 1572-3127. Disponível em: <https://www.sciencedirect.com/science/article/pii/S157231270800052X>. Citado na página 32.

LEIVA, V.; SANHUEZA, A.; KOTZ, S.; ARANEDA, N. A unified mixture model based on the inverse gaussian distribution. **Pakistan Journal of Statistics**, v. 26, p. 445–460, 2010. Citado na página 38.

LEMESHKO, B. Y.; LEMESHKO S. B.AND AKUSHKINA, K. A.; NIKULIN, M. S.; SAAIDIA, N. **Inverse Gaussian Model and Its Applications in Reliability and Survival Analysis**. [S.l.]: In: Rykov V., Balakrishnan N., Nikulin M. (eds) Mathematical and Statistical Models and Methods in Reliability. Statistics for Industry and Technology, 2010. Citado na página 37.

LOUZADA, F.; CANCHO, V. G.; BARRIGA GLADYS D, C. The poisson-exponential regression model under different latent activation schemes. **Computational & Applied Mathematics**, SciELO Brasil, v. 31, n. 3, p. 617–632, 2012. Citado na página 22.

LOUZADA, F.; CANCHO, V. G.; YIQI, B. The log-weibull-negative-binomial regression model under latent failure causes and presence of randomized activation schemes. **Statistics**, Taylor & Francis, v. 49, n. 4, p. 930–949, 2015. Citado na página 22.

LOUZADA, F.; SUZUKI, A. K.; CANCHO, V. G. The fgm long-term bivariate survival copula model: Modeling, bayesian estimation, and case influence diagnostics. **Communications in Statistics - Theory and Methods**, v. 42, n. 4, p. 673–691, 2013. Citado na página 70.

MORITA, L. H. M.; TOMAZELLA, V. L. D.; RAMOS, P. L.; FERREIRA, P. H.; LOUZADA, F. The random deterioration rate model with measurement error based on the inverse gaussian distribution. **Braz. J. Probab. Stat.**, Brazilian Statistical Association, v. 35, n. 1, p. 187–204, 02 2021. Disponível em: <https://doi.org/10.1214/20-BJPS468>. Citado na página 37.

NAGAMANI, N.; TRIPATHY, M. R. Estimating common dispersion parameter of several inverse gaussian populations : A simulation study. **Journal of Statistics and Management Systems**, Taylor Francis, v. 21, n. 7, p. 1357–1389, 2018. Disponível em: <https://doi.org/10.1080/09720510.2018.1503406>. Citado na página 37.

Nakagawa, T.; Osaki, S. The discrete weibull distribution. **IEEE Transactions on Reliability**, R-24, n. 5, p. 300–301, 1975. Citado na página 32.

NARISETTY, N.; KOENKER, R. Censored quantile regression survival models with a cure proportion. **Journal of Econometrics**, 2021. ISSN 0304-4076. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304407620303997>. Citado na página 22.

NELSEN, R. **An introduction to Copulas**. [S.l.]: Springer, New York, 2006. Citado na página 70.

NIKULIN, M. S.; SOLEV, V. N. **Chi-squares goodness-of-fit test for doubly censored data with applications in survival analysis and reliability**. Birkhauser: Boston: In: Statistical and Probabilistic Models in Reliability. (Ionescu, D.C., Limnios, N. (Eds)), 1999. 101—112 p. Citado na página 37.

OLIVEIRA, M. A.; SUZUKI, A. K.; SARAIVA, E. F. Uma abordagem bayesiana para modelos de sobrevivência bivariados baseados em cópulas arquimedianas. **Revista Brasileira de Biometria**, v. 32, p. 390–411, 2014. Citado na página 70.

ONICESCU, G.; LAWSON, A. B. Bayesian cure-rate survival model with spatially structured censoring. **Spatial Statistics**, v. 28, p. 352–364, 2018. ISSN 2211-6753. One world, one health. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2211675317301926>. Citado na página 22.

ORTEGA, E. M. M.; CORDEIRO, G. M.; HASHIMOTO, E. M.; SUZUKI, A. K. Regression models generated by gamma random variables with long-term survivors. **Communications for Statistical Applications and Methods**, v. 24g, p. 43–65, 2017. Citado na página 22.

ORTEGA, E. M. M.; CORDEIRO, G. M.; SUZUKI, A. K.; RAMIRES, T. G. A new extended birnbaum-saunders model with cure fraction: classical and bayesian approach. **Communications for Statistical Applications and Methods**, v. 24, p. 397–419, 2017. Citado na página 22.

PAL, A.; MONDAL, S.; KUNDU, D. A. Cure rate model for exponentially distributed lifetimes with competing risks. **Journal of Statistical Theory and Practice**, v. 15, 2021. Citado na página 22.

PENG, C.-Y. Inverse gaussian processes with random effects and explanatory variables for degradation data. **Technometrics**, Taylor Francis, v. 57, n. 1, p. 100–111, 2015. Disponível em: <https://doi.org/10.1080/00401706.2013.879077>. Citado na página 37.

PENG, F.; DEY, D. Bayesian analysis of outlier problems using divergence measures. **Canadian Journal of Statistics**, Wiley Online Library, v. 23, n. 2, p. 199–213, 1995. Citado nas páginas 23, 56, 57 e 69.

PERES, M. V. O.; ACHCAR, J. A.; MARTINEZ, E. Z. Bivariate lifetime models in presence of cure fraction: a comparative study with many different copula functions. **Heliyon**, v. 6, p. 1–17, 2020. Citado na página 70.

PESCIM, R. R.; ORTEGA, E. M.; SUZUKI, A. K.; CANCHO, V. G.; CORDEIRO, G. M. A new destructive poisson odd log-logistic generalized half-normal cure rate model. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 48, n. 9, p. 2113–2128, 2019. Citado na página 22.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: convergence diagnosis and output analysis for mcmc. **R News**, v. 6, n. 1, p. 7–11, 2006. Citado na página 23.

PUNZO, A. A new look at the inverse gaussian distribution with applications to insurance and economic data. **Journal of Applied Statistics**, Taylor Francis, v. 46, n. 7, p. 1260–1287, 2019. Disponível em: <https://doi.org/10.1080/02664763.2018.1542668>. Citado na página 37.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <http://www.R-project.org>. Citado na página 23.

RAMIRES, T. G.; ORTEGA, E. M.; LEMONTE, A. J.; HENS, N.; CORDEIRO, G. M. A flexible bimodal model with long-term survivors and different regression structures. **Communications in Statistics - Simulation and Computation**, Taylor Francis, v. 49, n. 10, p. 2639–2660, 2020. Disponível em: <https://doi.org/10.1080/03610918.2018.1524902>. Citado na página 22.

RIBEIRO, T. R.; SUZUKI, A. K.; SARAIVA, E. F. Uma abordagem bayesiana para o modelo de sobrevivência bivariado derivado da cópula amh. **Revista da Estatística da Universidade Federal de Ouro Preto**, v. 6, p. 1–20, 2017. Citado na página 70.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F. A new class of defective models based on the marshall–olkin family of distributions for cure rate modeling. **Computational Statistics Data Analysis**, v. 107, p. 48–63, 2017. ISSN 0167-9473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947316302286>. Citado na página 22.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de. Teoria unificada de análise de sobrevivência. **ABE-Associaçao Brasileira de Estatıstica, 18o SINAPE-Sao Pedro-Sao Paulo**, 2008. Citado nas páginas 22, 29 e 30.

RODRIGUES, J.; INACIO, M. H. A.; SUZUKI, A. K.; SILVA, F. R. Bayesian superposition of pure-birth destructive cure processes for tumor latency. **Communications in Statistics-Simulation and Computation**, v. 49, p. 3240–3253, 2020. Citado na página 22.

ROMEO, J. S.; MEYER, R.; GALLARDO, D. I. Bayesian bivariate survival analysis using the power variance function copula. **Lifetime Data Analysis**, v. 24, p. 355–383, 2018. Citado na página 70.

Roy, D. Discrete rayleigh distribution. **IEEE Transactions on Reliability**, v. 53, n. 2, p. 255–260, 2004. Citado na página 32.

SANHUEZA, A.; LEIVA, V.; BALAKRISHNAN, N. A new class of inverse gaussian type distributions. **Metrika**, v. 68, p. 31–68, 2008. Citado na página 38.

SCHRöDINGER, E. Zur theorie der fall-und steigversuche und teilchen mit brownscher bewegung. **Physikalische Zeitschrift**, v. 16, n. 16, p. 289–295, 1915. Citado na página 36.

SESHADRI, V. **The Inverse Gaussian Distribution: A Case Study in Exponential Families**. New York: Claredon, 1993. Citado na página 37.

____. **The Inverse Gaussian Distribution: Statistical Theory and Applications**. New York: Springer, 1999. Citado na página 37.

SHAMANY, R.; ALOBAIDI, N.; ALGAMAL, Z. A new two-parameter estimator for the inverse gaussian regression model with application in chemometrics. **Electronic Journal of Applied Statistical Analysis**, v. 12, n. 2, 2019. Citado na página 37.

SILVA, G. O.; CORDEIRO, G. M.; ORTEGA, E. M. M. Surviving and non surviving fraction regression models based on the beta modified weibull distribution. **Model Assisted Statistics and Applications**, v. 15, p. 111–126, 2020. Citado na página 22.

SOUZA, D. de; CANCHO, V. G.; RODRIGUES, J.; BALAKRISHNAN, N. Bayesian cure rate models induced by frailty in survival analysis. **Statistical Methods in Medical Research**, v. 26, p. 2011–2028, 2017. Citado na página 22.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society B**, v. 64, p. 583–639, 2002. Citado na página 55.

Stein, W. E.; Dattero, R. A new discrete weibull distribution. **IEEE Transactions on Reliability**, R-33, n. 2, p. 196–197, 1984. Citado na página 32.

STOGIANNIS, D.; CARONI, C. Tests for outliers in the inverse gaussian distribution, with application to first hitting time models. **Journal of Statistical Computation and Simulation**, v. 82, n. 1, p. 73–80, 2012. Citado na página 37.

SUZUKI, A. K.; BARRIGA, G. D. C.; LOUZADA, F.; CANCHO, V. G. A general long-term aging model with different underlying activation mechanisms: Modeling, bayesian estimation, and case influence diagnostics. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 46, n. 6, p. 3080–3098, 2017. Citado nas páginas 22, 30 e 56.

SUZUKI, A. K.; CANCHO, V. G.; LOUZADA, F. The poisson–inverse-gaussian regression model with cure rate: a bayesian approach and its case influence diagnostics. **Statistical Papers**, Springer, v. 57, n. 1, p. 133–159, 2016. Citado nas páginas 22 e 56.

SUZUKI, A. K.; LOUZADA, F.; CANCHO, V. G. On estimation and influence diagnostics for a bivariate promotion lifetime model based on the fgm copula: A fully bayesian computation. **Trends in Computational and Applied Mathematics**, v. 14, n. 3, p. 441–461, 2013. Citado na página 70.

SUZUKI, A. K.; LOUZADA, F.; CANCHO, V. G.; BARRIGA, G. D. The fgm bivariate lifetime copula model: A bayesian approach. **Advances and Applications in Statistics**, v. 21, p. 55–76, 2011. Citado na página 70.

TSODIKOV, A. D.; IBRAHIM, J. G.; YAKOVLEV, A. Y. Estimating cure rates from survival data: an alternative to two-component mixture models. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 464, p. 1063–1078, 2003. Citado na página 22.

TWEEDIE, M. C. K. Statistical properties of the inverse gaussian distribution. **Annals of Mathematical Statistics**, v. 28, p. 362–377, 1957. Citado na página 36.

UPADHYAY, S. K.; SEN, R. A bayes analysis of inverse gaussian based accelerated test models. **International Journal of Reliability, Quality and Safety Engineering**, v. 26, n. 02, p. 1950010, 2019. Citado na página 37.

WEISS, R. An approach to bayesian sensitivity analysis. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 739–750, 1996. Citado nas páginas 23, 56, 57 e 69.

Wen, X.; Wang, Z.; Fu, H.; Wu, Q.; Liu, C. Blues and reliability analysis for general censored data subject to inverse gaussian distribution. **IEEE Transactions on Reliability**, v. 68, n. 4, p. 1257–1271, 2019. Citado na página 37.

XU, A.; HU, J.; WANG, P. Degradation modeling with subpopulation heterogeneities based on the inverse gaussian process. **Applied Mathematical Modelling**, v. 81, p. 177–193, 2020. ISSN 0307-904X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0307904X19307577>. Citado na página 37.

YAKOVLEV, A.; YU, A. B.; BARDOU, V.-J.; FOURQUET, A.; HOANG, T.; ROCHEFODIERE, A.; TSODIKOV, A. D. A simple stochastics model of tumor recurrence an its aplications to data on premenopausal breast cancer. In: BIOMETRIE, S. F. de (Ed.). **Biometrie et Analyse de Donnes Spatio-Temporelles No 12, B**. France: [s.n.], 1993. p. 33–82. Citado na página 29.

YANG, Z. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. **Journal of Molecular Evolution**, v. 39, p. 306–314, 1994. Citado na página 32.

YE, Z.-S.; CHEN, N. The inverse gaussian process as a degradation model. **Technometrics**, Taylor Francis, v. 56, n. 3, p. 302–311, 2014. Disponível em: <https://doi.org/10.1080/00401706.2013.830074>. Citado na página 37.

YIQI, B.; RUSSO, C. M.; CANCHO, V. G.; LOUZADA, F. Influence diagnostics for the weibull-negative-binomial regression model with cure rate under latent failure causes. **Journal of Applied Statistics**, Taylor & Francis, v. 43, n. 6, p. 1027–1060, 2016. Citado nas páginas 22 e 30.

APÊNDICE

A

APÊNDICE

A.1 Funções *score* para os esquemas de ativação

Neste apêndice vamos apresentar as expressões das funções score para os esquemas de ativação do modelo BIG_{cr}. Seja, $\vartheta = (\mu, \lambda, p)$, então a função de verossimilhança é dada por:

$$L(\vartheta) = \prod_{i=1}^{n} (f_{pop}(y_i))^{\delta_i} (S_{pop}(y_i))^{1-\delta_i},$$

sendo assim, temos que a log verossimilhança é dada por:

$$l(\vartheta) = \sum_{i=1}^{n} \delta_i \log(f_{pop}(y_i)) + \sum_{i=1}^{n} (1-\delta_i) \log(S_{pop}(y_i))$$

A.1.1 Esquema de primeira ativação

Seja $S_{pop}(y_i)$ e $f_{pop}(y_i)$, as funções de sobrevivência e densidade impróprias sob o esquema de primeira ativação apresentadas em (3.8).

$$\begin{aligned}
\frac{\partial f_{pop}(y_i)}{\partial \mu} &= exp((1-\log(p))^{S_{IG}(y_i)} - 1 + \log(p)) \Big(p \log(1-\log(p)) S_{IG}(y_i) \log(1-\log(p)) \log(1-\log(p))^{S_{IG}(y_i)} \\
&+ (1-\log(p))^{S_{IG}(y_i)} + \Big(1-\log(p))^{S_{IG}(y_i)}\Big) \log((1-\log(p))) \frac{\partial f_{IG}(y_i)}{\partial \mu} \log(1-\log(p)) (-f_{IG}(y_i) \\
& e^{(1-\log(p))^{S_{IG}(y_i)}-1} p f_{IG}(y_i) \log(-\log(p)+1) (-\log(p)+1)^{2S_{IG}(y_i)} + exp((1-\log(p)))^{S_{IG}(y_i)-1} p (S_{IG}(y_i) \\
& \log(1-\log(p)) (1-\log(p))^{S_{IG}(y_i)} + (1-\log(p))^{S_{IG}(y_i)}.
\end{aligned}$$

$$\frac{\partial f_{IG}(y_i)}{\partial \mu} = (y_i - \mu) \frac{exp\{(y_i-\mu)^2 / 2 y_i \lambda^2 \mu^2\}}{\lambda^2 \mu^3}.$$

$$\frac{\partial l(\vartheta)}{\partial \mu} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial \mu} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial \mu}.$$

$$\frac{\partial f_{IG}(y_i)}{\partial \lambda} = \frac{\left(\frac{1}{\sqrt{(2\pi\lambda^2 y_i^3}}\right)(y_i-\mu)^2 exp\left(-\frac{(y_i-\mu)^2}{2y_i\mu^2\lambda^2}\right)}{y_i\mu^2\lambda^3} + e^{\frac{-(y_i-\mu)^2}{(2\mu^2\lambda^2 y_i)}}\left(\frac{1}{\sqrt{(2\pi y_i^3}}\lambda^2\right).$$

$$\begin{aligned}\frac{\partial f_{pop}(y_i)}{\partial \lambda} &= exp((1-\log(p))^{S_{IG}(y_i)} - 1 + \log(p))\,(p\log(1-\log(p))\,S_{IG}(y_i)\log(1-\log(p))\\ &\quad (1-\log(p))^{S_{IG}(y_i)+(1-\log(p))^{S_{IG}(y_i)}} + (1-\log(p))^{S_{IG}(y_i)}\log(1-\log(p))\frac{\partial f_{IG}(y_i)}{\partial \lambda}\\ &\quad \log(1-\log(p))\left(-f_{IG}(y_i)e^{(1-\log(p))^{S_{IG}(y_i)}-1} p f_{IG}(y_i)\log(-\log(p)+1)\right.\\ &\quad (-\log(p)+1)^{2S_{IG}(y_i)} + exp((1-\log(p)))^{S_{IG}(y_i)-1} p\,(S_{IG}(y_i)\log(1-\log(p))\\ &\quad (1-\log(p))^{S_{IG}(y_i)} + (1-\log(p))^{S_{IG}(y_i)}.\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial \lambda} = \sum_{i=1}^{n}\delta_i\frac{1}{f_{pop}(y_i)}\frac{\partial f_{pop}(y_i)}{\partial \lambda} + \sum_{i=1}^{n}(1-\delta_i)\frac{1}{S_{pop}(y_i)} - f_{pop}(y_i)\frac{\partial f_{pop}(y_i)}{\partial \lambda}.$$

$$\begin{aligned}\frac{\partial f_{pop}(y_i)}{\partial p} &= exp((1-\log(p))^{S_{IG}(y_i)} - 1 + \log(p))(-S_{IG}(y_i)log(1-log(p))(1-\log(p))^{S_{IG}(y_i)}\\ &- (1-\log(p))^{S_{IG}(y_i)-1}) + ((1-\log(p))^{S_{IG}(y_i}\log(1-\log(p))exp(1-\log(p))\\ &\quad (-S_{IG}(y_i)(-\log(p)+1)).\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial p} = \sum_{i=1}^{n}\delta_i\frac{1}{f_{pop}(y_i)}\frac{\partial f_{pop}(y_i)}{\partial p} + \sum_{i=1}^{n}(1-\delta_i)\frac{1}{S_{pop}(y_i)} - f_{pop}(y_i)\frac{\partial f_{pop}(y_i)}{\partial p}.$$

$$\begin{aligned}\frac{\partial f_{pop}(y_i)}{\partial \beta_0} &= exp((1-\log(p_i))^{S_{IG}(y_i)} - 1 + \log(p_i))\left(S_{IG}(y_i)\log\left(-e^{-\beta_0}\right) - 1\right)\left(-e^{-\beta_0}\right)^{S_{IG}(y_i)}\\ &+ exp(e^{-\beta_0} + \left(-e^{-\beta_0}\right)^{S_{IG}(y_i)} - \beta_0\left(S_{IG}(y_i)\left(-e^{-\beta_0}\right)^{S_{IG}(y_i)-1} - 1\right)(1-\log(p_i))^{S_{IG}(y_i)}\\ &\quad \log(1-\log(p_i))f_{IG}(y_i).\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial \beta_0} = \sum_{i=1}^{n}\delta_i\frac{1}{f_{pop}(y_i)}\frac{\partial f_{pop}(y_i)}{\partial \beta_0} + \sum_{i=1}^{n}(1-\delta_i)\frac{1}{S_{pop}(y_i)} - f_{pop}(y_i)\frac{\partial f_{pop}(y_i)}{\partial \beta_0}.$$

$$\begin{aligned}\frac{\partial f_{pop}(y_i)}{\partial \beta_1} &= exp((1-\log(p_i))^{S_{IG}(y_i)} - 1 + \log(p_i))x_i S_{IG}(y_i)exp(-x_i\beta_1-\beta_0)\\ &\quad \log\left(-e^{-x_i\beta_1-\beta_0}\right)\left(-e^{-x_i\beta_1-\beta_0}\right)^{S_{IG}(y_i)-1} - x_i\,(-exp(-x_i\beta_1-\beta_0))^{S_{IG}(y_i)}\\ &+ (1-\log(p_i))^{S_{IG}(y_i)}\log(1-\log(p_i))f_{IG}(y_i)exp(e^{-x_i\beta_1-\beta_0}\\ &+ \left(-e^{-x_i\beta_1-\beta_0}\right)^{S_{IG}(y_i))}(x_i S_{IG}(y_i)exp(-x_i\beta_1-\beta_0)\left(-e^{-x_i\beta_1-\beta_0}\right)^{S_{IG}(y_i)-1}\\ &- x_i e^{-x_i\beta_1-\beta_0}).\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial \beta_1} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial \beta_1} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial \beta_1}.$$

A.1.2 Esquema de ativação aleatória

Seja $S_{pop}(y_i)$ e $f_{pop}(y_i)$, as funções de sobrevivência e densidade impróprias sob o esquema de ativação aleatória apresentadas em (3.10).

$$\frac{\partial l(\vartheta)}{\partial \mu} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} (1-p) \frac{\partial f_{IG}(y_i)}{\partial \mu} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i)(1-p) \frac{\partial f_{IG}(y_i)}{\partial \mu}.$$

$$\frac{\partial l(\vartheta)}{\partial \lambda} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} (1-p) \frac{\partial f_{IG}(y_i)}{\partial \lambda} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i)(1-p) \frac{\partial f_{IG}(y_i)}{\partial \lambda}.$$

$$\frac{\partial l(\vartheta)}{\partial p} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} - f_{IG}(y_i) + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) - f_{IG}(y_i).$$

$$\frac{\partial l(\vartheta)}{\partial \beta_0} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} f_{IG}(y_i)(exp(-\beta_0 + e^{-\beta_0} + 1) + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) f_{IG}(y_i)(exp(-\beta_0 + e^{-\beta_0} + 1).$$

$$\frac{\partial l(\vartheta)}{\partial \beta_1} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} e^{e^{-\beta_0-\beta_1 x_i} - \beta_0 - x_i\beta_0} x_i + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) e^{e^{-\beta_0-\beta_1 x_i} - \beta_0 - x_i\beta_0} x_i.$$

A.1.3 Esquema de última ativação

Seja $S_{pop}(y_i)$ e $f_{pop}(y_i)$, as funções de sobrevivência e densidade impróprias sob o esquema de primeira ativação apresentadas em (3.9).

$$\begin{aligned}
\frac{\partial f_{pop}(y_i)}{\partial \mu} &= exp((1-\log(p))^{F_{IG}(y_i)} - 1 + \log(p)) \, (p \log(1-\log(p)) F_{IG}(y_i) \log(1-\log(p)) \\
&\quad (1-\log(p))^{F_{IG}(y_i)} + (1-\log(p))^{F_{IG}(y_i)} + (1-\log(p))^{F_{IG}(y)} \log(1-\log(p)) \frac{\partial f_{IG}(y_i)}{\partial \mu} \\
&\quad \log(1-\log(p)) \, (f_{IG}(y_i) e^{(1-\log(p))^{F_{IG}(y_i)} - 1} p f_{IG}(y_i) \log(-\log(p)+1) \, (-\log(p)+1)^{2F_{IG}(y_i)} \\
&+ \quad exp((1-\log(p)))^{F_{IG}(y_i)-1} p (F_{IG}(y_i) \log(1-\log(p)) \left((1-\log(p))^{F_{IG}(y_i)} + (1-\log(p))^{F_{IG}(y_i)} \right).
\end{aligned}$$

$$\frac{\partial f_{IG}(y_i)}{\partial \mu} = (y_i - \mu) \frac{exp\{(y_i-\mu)^2 / 2 y_i \lambda^2 \mu^2\}}{\lambda^2 \mu^3}.$$

$$\frac{\partial l(\vartheta)}{\partial \mu} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial \mu} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial \mu}.$$

$$\frac{\partial f_{IG}(y_i)}{\partial \lambda} = \frac{\left(\frac{1}{\sqrt{(2\pi\lambda^2 y_i^3}} \right) (y_i - \mu)^2 exp\left(-\frac{(y_i-\mu)^2}{2 y_i \mu^2 \lambda^2} \right)}{y_i \mu^2 \lambda^3} + e^{\frac{-(y_i-\mu)^2}{(2\mu^2\lambda^2 y_i)}} \left(\frac{1}{\sqrt{(2\pi y_i^3}} \lambda^2 \right).$$

$$\frac{\partial f_{pop}(y_i)}{\partial \lambda} = exp((1-\log(p))^{F_{IG}(y_i)} - 1 + \log(p)) \, (p \log(1-\log(p)) F_{IG}(y_i) \log(1-\log(p))$$

$$(1-\log(p))^{F_{IG}(y_i)} + (1-\log(p))^{F_{IG}(y_i)} + (1-\log(p))^{F_{IG}(y_i)} \log(1-\log(p)) \frac{\partial f_{IG}(y_i)}{\partial \lambda}$$

$$\log(1-\log(p))\left(-f_{IG}(y_i) e^{(1-\log(p))^{F_{IG}(y_i)}-1} p f_{IG}(y_i) \log(-\log(p)+1)(-\log(p)+1)^{2F_{IG}(y_i)}\right.$$
$$+ \quad exp((1-\log(p)))^{F_{IG}(y_i)-1} p \left(F_{IG}(y_i) \log(1-\log(p))(1-\log(p))^{F_{IG}(y_i)} + (1-\log(p))^{F_{IG}(y_i)}\right).$$

$$\frac{\partial l(\vartheta)}{\partial \lambda} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial \lambda} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial \lambda}.$$

$$\begin{aligned}
\frac{\partial f_{pop}(y_i)}{\partial p} &= exp((1-\log p)^{F_{IG}(y_i)} - 1 + \log p)(-F_{IG}(y_i) log(1-log(p))(1-\log(p))^{F_{IG}(y_i)} \\
&- (1-\log(p))^{F_{IG}(y_i)-1}) + ((1-\log(p))^{F_{IG}(y_i)} \log(1-\log(p)) exp(1-\log(p)) \\
&\quad (-F_{IG}(y_i)(-\log(p)+1))).
\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial p} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial p} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial p}.$$

$$\begin{aligned}
\frac{\partial f_{pop}(y_i)}{\partial \beta_0} &= exp((1-\log(p_i))^{F_{IG}(y_i)} - 1 + \log(p_i)) \left(F_{IG}(y_i) \log\left(-e^{-\beta_0}\right) - 1\right) \left(-e^{-\beta_0}\right)^{F_{IG}(y_i)} \\
&+ exp(e^{-\beta_0} + \left(-e^{-\beta_0}\right)^{F_{IG}(y_i)} - \beta_0 \left(F_{IG}(y_i) \left(-e^{-\beta_0}\right)^{F_{IG}(y_i)-1} - 1\right) \\
&\quad (1-\log(p_i))^{F_{IG}(y_i)} \log(1-\log(p_i)) f_{IG}(y_i).
\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial \beta_0} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial \beta_0} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial \beta_0}.$$

$$\begin{aligned}
\frac{\partial f_{pop}(y_i)}{\partial \beta_1} &= exp((1-\log(p_i))^{F_{IG}(y_i)} - 1 + \log(p_i)) x_i F_{IG}(y_i) exp(-x_i\beta_1 - \beta_0) \\
&\quad \log\left(-e^{-x_i\beta_1-\beta_0}\right) \left(-e^{-x_i\beta_1-\beta_0}\right)^{F_{IG}(y_i)-1} - x_i \left(-exp - x_i\beta_1 - \beta_0\right)^{F_{IG}(y_i)} \\
&+ (1-\log(p_i))^{F_{IG}(y_i)} \log(1-\log(p_i)) f_{IG}(y_i) exp(e^{-x_i\beta_1-\beta_0} \\
&+ (-e^{-x_i\beta_1-\beta_0})^{F_{IG}(y_i)}) (x_i F_{IG}(y_i) exp(-x_i\beta_1 - \beta_0) \left(-e^{-x_i\beta_1-\beta_0}\right)^{F_{IG}(y_i)-1} \\
&- x_i e^{-x_i\beta_1-\beta_0}).
\end{aligned}$$

$$\frac{\partial l(\vartheta)}{\partial \beta_1} = \sum_{i=1}^{n} \delta_i \frac{1}{f_{pop}(y_i)} \frac{\partial f_{pop}(y_i)}{\partial \beta_1} + \sum_{i=1}^{n} (1-\delta_i) \frac{1}{S_{pop}(y_i)} - f_{pop}(y_i) \frac{\partial f_{pop}(y_i)}{\partial \beta_1}.$$

APÊNDICE

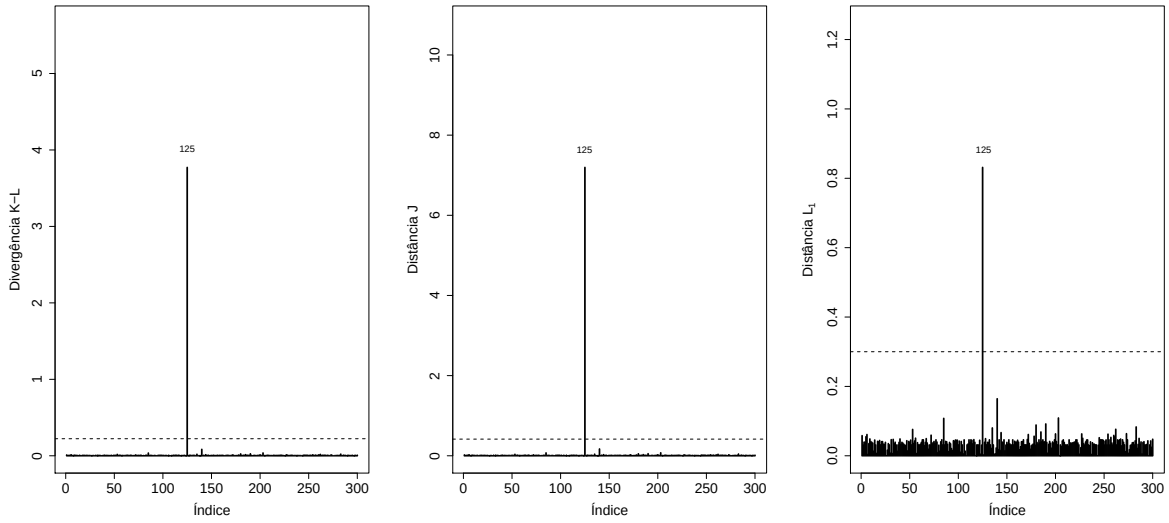
B

APÊNDICE

Neste apêndice apresentamos algumas medidas de divergência *psi* do estudo de simulação apresentado na Subseção 4.4.

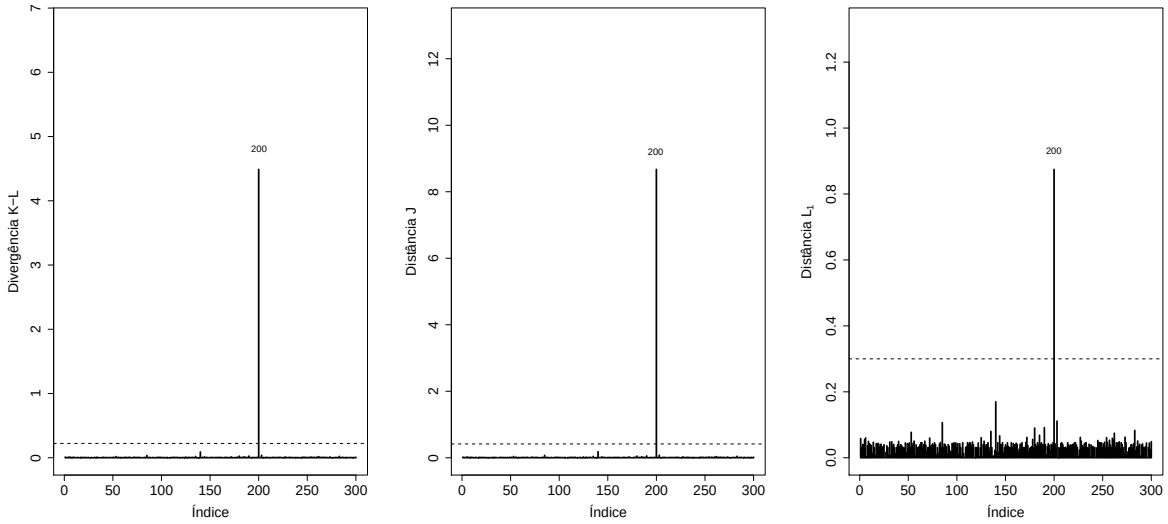
As Figuras 19 a 23 apresentam as medidas de divergência para os casos (c) a (g), respetivamente. Podemos observar em todas as figuras que as três medidas de divergência detectaram a(s) observação(ões) influente(s).

Figura 19 – Divergência ψ para o conjunto de dados (c).



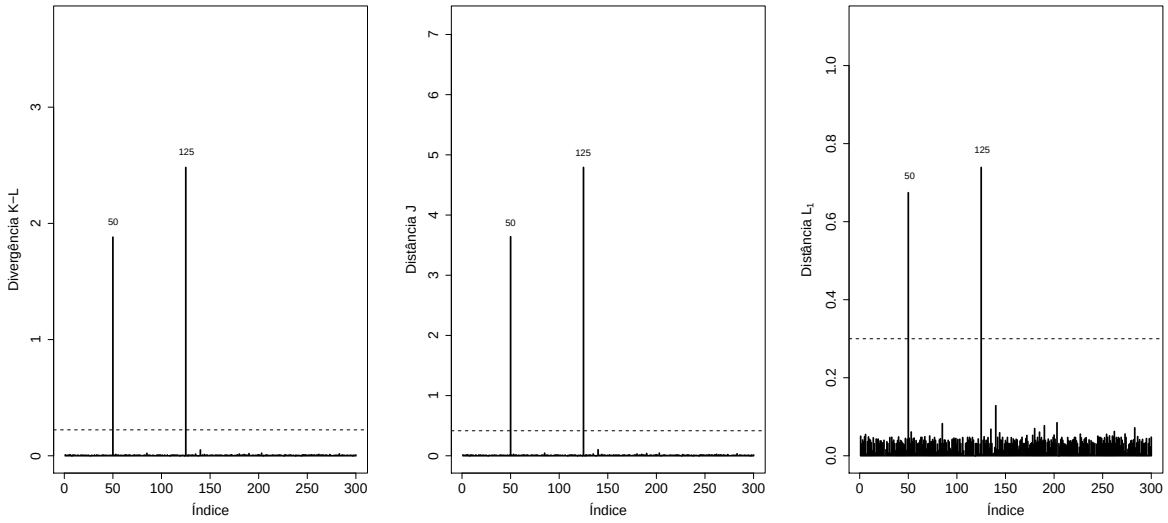
Fonte: Elaborada pela autora.

Figura 20 – Divergência ψ para o conjunto de dados (d).



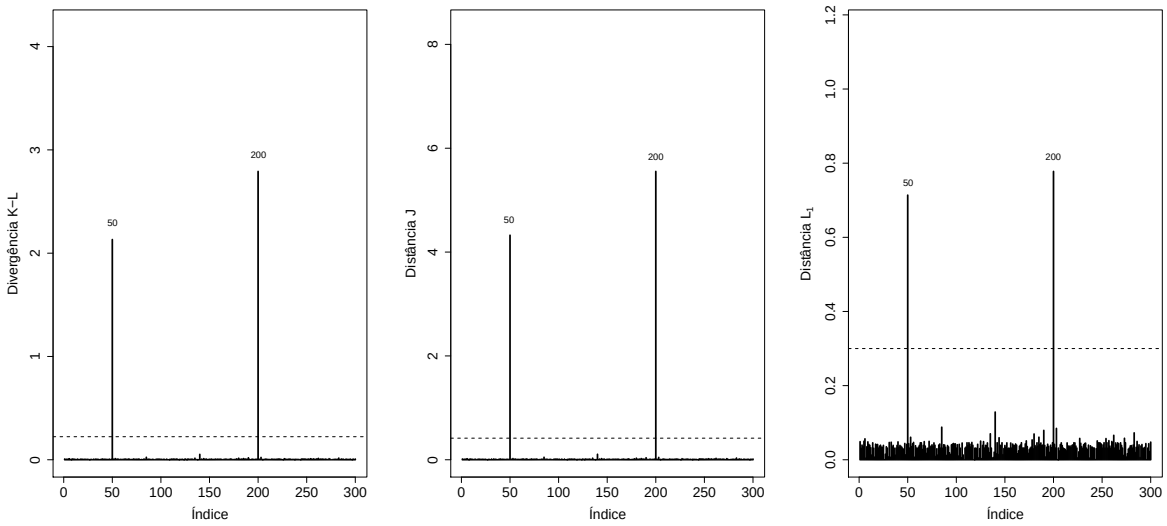
Fonte: Elaborada pela autora.

Figura 21 – Divergência ψ para o conjunto de dados (e).



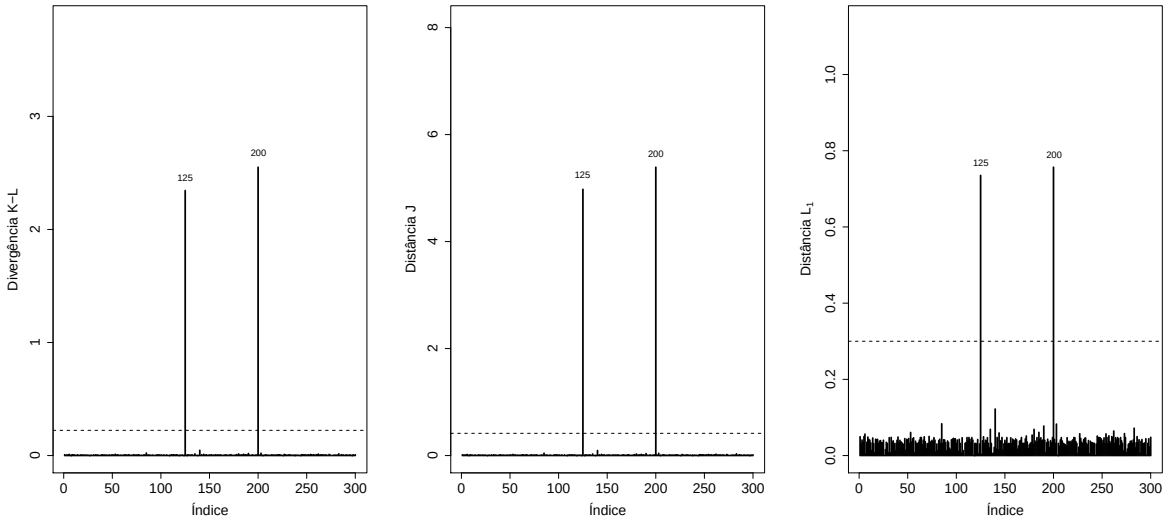
Fonte: Elaborada pela autora.

Figura 22 – Divergência ψ para o conjunto de dados (f).



Fonte: Elaborada pela autora.

Figura 23 – Divergência ψ para o conjunto de dados (g).



Fonte: Elaborada pela autora.

APÊNDICE

C

APÊNDICE

Código-fonte 1 – Função geradora de n valores da distribuição Bell de parâmetro θ.

```
1: rBell<-function(n,theta){
2: N=numeric(n)
3: a=rpois(n,exp(theta)-1)
4: for(i in 1:n) {
5: N[i]=sum(rztpois(a[i],theta))
6:                 }
7: return(N)
8:                              }
```

Código-fonte 2 – Função de verossimilhança considerando o modelo BIGcr sob o esquema de última ativação.

```
1: lv_lM = function(y,x,status,theta=c(1,1,1,1)) {
2: mu=exp(theta[1])
3: lambda=exp(theta[2])
4: vbeta=theta[3:4]
5: p=1/(1+exp(-vbeta[1]-vbeta[2]*x))
6: vf=log(sqrt(lambda/2*pi*y^3)*exp((-lambda*(y-mu)^2)/(2*mu^2*y))
     )
7: vF=pnorm(sqrt(lambda/y)*(y/mu-1),0,1)+exp(2*lambda/mu)*pnorm(-
     sqrt(lambda/y)*(y/mu+1),0,1)
8: aux1=1-log(p)
9: aux2=aux1^vF-aux1
10: lspop=log(1+p-exp(aux2))
11: lfpop =aux2+vF*log(aux1)+log(log(aux1))+vf
```

```
12: loglik = sum(status *lfpop + (1 - status) * lspop)
13: loglik
14:                                                    }
```

Código-fonte 3 – Abordagem Bayesiana - Código JAGS para ajuste do modelo BIGcr sob o esquema de última ativação.

```
 1: cat( "model
 2:                {
 3: mu<-exp(mu1)
 4: lambda<-exp(lambda1)
 5:
 6: for (i in 1:n) {
 7: p[i]<-1/(1+exp(-beta0-beta1*x[i]))
 8: vf[i]<-log(sqrt(lambda/2*3.14159265359*y[i]^3)*exp((-lambda*(y[
      i]-mu)^2)/(2*mu^2*y[i])))
 9: vF[i]<-pnorm(sqrt(lambda/y[i])*(y[i]/mu-1),0,1)+exp(2*lambda/mu
      )*pnorm(-sqrt(lambda/y[i])*(y[i]/mu+1),0,1)
10: S[i] <- 1-vF[i]
11: aux1[i] <-1-log(p[i])
12: aux2[i]=pow(aux1[i],vF[i])-aux1[i]
13: lspop[i] =log(1+p[i]-exp(aux2[i]))
14: lfpop[i] =aux2[i]+vF[i]*log(aux1[i])+log(log(aux1[i]))+vf[i]
15:
16: # Verossimilhança
17: L1[i]<-status[i]*lfpop[i]+(1-status[i])*lspop[i]
18: L[i]<-exp(L1[i])
19: phi[i]<-L[i]/C
20: zeros[i]~dbern(phi[i])
21:          }
22: # Priori
23: mu1~dnorm(0,0.001)
24: lambda1~dnorm(0,0.001)
25: beta0~dnorm(0,0.001)
26: beta1~dnorm(0,0.001)
27: C<-10000000000
28:                 }",
29: file="BIG_maximo.jags")
```


ICMC
USP
SÃO CARLOS