UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

NOVOS MODELOS DE SOBREVIVÊNCIA COM FRAÇÃO DE CURA BASEADOS NO PROCESSO DA CARCINOGÊNESE

Patrick Borges

UFSCar - São Carlos/SP
Maio/2012

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

NOVOS MODELOS DE SOBREVIVÊNCIA COM FRAÇÃO DE CURA BASEADOS NO PROCESSO DA CARCINOGÊNESE

Patrick Borges

Orientador: Prof. Dr. Josemar Rodrigues

Trabalho apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar como parte dos requisitos para obtenção do título de Doutor em Estatística.

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

B732nm

Borges, Patrick.
Novos modelos de sobrevivência com fração de cura baseados no processo da carcinogênese / Patrick Borges. -- São Carlos : UFSCar, 2012.
92 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2012.

1. Estatística. 2. Carcinogênese. 3. Modelos de sobrevivência. 4. Fração de cura. 5. Estrutura de correlação. 6. Esquema de ativação híbrido. I. Título.

CDD: 519.5 (20ª)

Patrick Borges

Novos Modelos de Sobrevivência com Fração de Cura Baseados no Processo da Carcinogênese

Tese apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Doutor em Estatística.

Aprovada em 03 de maio de 2012.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Josemar Rodrigues (DEs-UFSCar / Orientador)

1° Examinador _____
Prof. Dr. Gauss Moutinho Cordeiro (UFRPE)

2° Examinador _____
Prof. Dr. Heleno Bolfarine (IME-USP)

3° Examinador _____
Prof. Dr. Jorge Alberto Achcar (FMRP-USP)

4° Examinador _____
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP)

Agradecimentos

Alguém já disse que "**a gratidão é a lembrança do coração**". Faz sentido. Ao longo de nossas vidas sempre aparecem "anjos da guarda" que nos ajudam, e sem os quais nossos objetivos seriam muitos difíceis de alcançar, ou seriam até inatingíveis. Por isso essa parte da tese é tão especial. Quero aqui expressar de coração os meus agradecimentos às seguintes pessoas e instituições:

A Deus pelos momentos de felicidade, que iluminam e me dão força para seguir a minha caminhada, e pelos momentos de dificuldade que me moldam a cada instante para ser um ser humano mais digno a exemplo do Cristo.

À minha família, o alicerce de minha vida: meus pais, Geraldo Borges e Sandra Borges, pelo eterno cuidado, dedicação e amor; pelo apoio nos momentos difíceis e de inquietantes decisões; por estarem ao meu lado a cada passo, a cada pequena conquista e grandes realizações, pois estes não teriam valor se vocês não estivessem comigo. Agradeço a minha irmã, Daniela Borges, pelo companheirismo e amizade.

Ao meu amor, Wanderléia Aigner, pelo companheirismo em todos os momentos, pelos sorrisos, pelo cuidado carinhoso e por simplesmente ter aparecido na minha vida.

Ao meu grande amigo Julieverson Vasconcelos e à família Francisco Alves, que sempre me incentivaram a prosseguir meus estudos.

À professora Maria José Schuwartz Ferreira, que foi minha professora de probabilidade durante a graduação na Universidade Federal do Espírito Santo. A professora foi muito além das suas obrigações e, além da probabilidade, me ensinou a pensar de forma clara e organizada. De certo modo, eu acho que devo a ela grande parte do sucesso que venho obtendo em qualquer atividade "intelectual" que participe, os fracassos são devidos única e exclusivamente as minhas

limitações.

Ao professor Josemar Rodrigues por ser mais do que meu orientador, por acreditar na minha capacidade e no meu crescimento profissional e pessoal, pelo apoio em todos os momentos e, principalmente pela amizade.

Ao professor Narayanaswamy Balakrishnan, pelas preciosas sugestões, considerações, correções e incentivos que recebi durante a elaboração desta tese.

À Universidade Federal do Espírito Santo, incluindo os colegas do Departamento de Estatística, que incentivaram e permitiram a minha liberação para o Doutorado. Principalmente aos professores Edwards Cerqueira, o Chefão, e Mauro Campos, o Pesquisador, que pra mim é uma honra tê-los como amigos.

Aos professores Gutemberg Brasil e Renato Krohling, pela confiança demonstrada em suas cartas de recomendação.

À PPGEST/UFSCar por ter me recebido no curso de Doutorado, e, em especial ao apoio do professor Francisco Louzada-Neto.

Aos funcionários do Departamento de Estatística da UFSCar, especialmente à Isabel Araujo, pelos serviços gentilmente prestados.

Aos alunos do PPGEST, meus companheiros de vida acadêmica, meu muito obrigado. Quero aqui agradecer especialmente aos colegas Rubiane, Katiane, Silvana, Mari, Cynthia, Hugo e Vitor.

Finalmente, faço questão de agradecer a todas as pessoas que torceram ou intercederam por mim, mesmo que de forma anônima ou discreta. É como disse Vínicius de Moraes: "**Você não faz amigos, você os reconhece**". A todos esses amigos e amigas, meu muito obrigado.

Resumo

Neste trabalho propomos modelos de sobrevivência com fração de cura para descrever o mecanismo biológico da ocorrência do evento de interesse (câncer) em estudos da carcinogênese na presença de causas competitivas latentes independentes ou correlacionadas. A formulação dos novos modelos é baseada na modelagem estocástica da ocorrência dos tumores através de três estágios: iniciação de um tumor não detectável, promoção e a progressão do tumor até um câncer detectável. Estes modelos permitem um padrão simples da dinâmica de crescimento do tumor, além de incorporarem características do estágio de progressão do tumor, que não é possível na maioria dos modelos de sobrevivência com fração de cura comumente utilizados. Para os modelos propostos, discutimos o processo inferencial do ponto de vista clássico e bayesiano. Estudos de simulações foram feitos com o objetivo de analisar as propriedades assintóticas do processo de estimação clássico. Aplicações a conjuntos de dados reais mostraram a aplicabilidade dos modelos.

Palavras-chave: carcinogênese, modelos de sobrevivência, fração de cura, estrutura de correlação, esquema de ativação híbrido.

Abstract

In this dissertation we propose new models for survival with cure fraction to describe the biological mechanism of the event of interest (cancer) in studies of carcinogenesis in the presence of competing causes latent independent or correlated. The formulation of new models is based on stochastic modeling of the occurrence of tumors through three stages: initiation of a tumor not detectable, promotion and progression of the tumor to a detectable cancer. These models allow a simple pattern of the dynamics of tumor growth, and incorporate into the analysis features of the stage of tumor progression that is not possible in most survival models with cure fraction commonly used. For the proposed models, the inferential process was discussed in terms of classical and Bayesian point of view. Simulations studies were conducted in order to analyze the asymptotical properties of the classical estimation procedure. Real data applications demonstrate of use of the models.

Keywords: carcinogenesis, survival models, correlation structure, cured fraction, hybrid activation scheme.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Câncer, nome científico neoplasia, é o nome dado a um conjunto de mais de 200 doenças que têm em comum o crescimento desordenado de células que invadem tecidos e órgãos. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores malignos (podem também ser tumores benignos, mas estamos interessados na formação dos malignos), que podem disseminar-se para outras regiões do corpo. Essa disseminação é denominada de metástase (vide INCA, 2011).

O câncer ocorre quando uma célula normal sofre alterações no seu DNA (ácido desoxirribonucléico), sendo esse evento denominado mutação genética. As células cujo material genético foi modificado sofrem uma perda de sua função e multiplicam-se de maneira descontrolada, mais rapidamente do que as células normais do tecido à sua volta, invadindo-o. Geralmente, têm capacidade para formar novos vasos sanguíneos que as nutrirão e manterão as atividades de crescimento descontrolado. O acúmulo dessas células forma os tumores malignos. Invadem inicialmente os tecidos vizinhos, podendo chegar ao interior de um vaso sangüíneo ou linfático e, por meio desses, disseminar-se, chegando a órgãos distantes do local onde o tumor se iniciou, formando as metástases. As células cancerosas são geralmente menos especializadas nas suas funções do que as suas correspondentes normais. Conforme as células cancerosas vão substituindo as normais, os tecidos invadidos vão perdendo suas funções.

O processo de formação do câncer chama-se carcinogênese, em geral se dá lentamente, podendo levar vários anos para que uma célula cancerosa prolifere e dê origem a um tumor detec-

tável. Esse processo passa por vários estágios (vide INCA, 2011) antes de chegar ao tumor. São eles:

1. **Estágio de iniciação.** É o primeiro estágio da carcinogênese. Nele as células sofrem o efeito dos agentes cancerígenos ou carcinógenos que provocam modificações em alguns de seus genes. Nesta fase as células encontram-se, geneticamente alteradas, porém ainda não é possível detectar um tumor clinicamente. Encontram-se "preparadas", ou seja, "iniciadas" para a ação de um segundo grupo de agentes que atuará no próximo estágio.

2. **Estágio de promoção.** É o segundo estágio da carcinogênese. Nele, as células geneticamente alteradas, ou seja, "iniciadas", sofrem o efeito dos agentes cancerígenos classificados como oncopromotores. A célula iniciada é transformada em célula maligna gradualmente. Para que ocorra essa transformação, é necessário um longo e continuado contato com o agente cancerígeno promotor. A suspensão do contato com agentes promotores muitas vezes interrompe o processo nesse estágio. Alguns componentes da alimentação e a exposição excessiva e prolongada a hormônios são exemplos de fatores que promovem a transformação de células iniciadas em malignas.

3. **Estágio de progressão.** É o terceiro e último estágio e caracteriza-se pela multiplicação descontrolada e irreversível das células malignas. Nesse estágio o câncer já está instalado, evoluindo até o surgimento das primeiras manifestações clínicas da doença. Os fatores que promovem a iniciação ou progressão da carcinogênese são chamados agentes oncoaceleradores ou carcinógenos. O fumo é um agente carcinógeno completo, pois possui componentes que atuam nos três estágios da carcinogênese.

O processo de carcinogênese é representado esquematicamente na Figura 2.1.

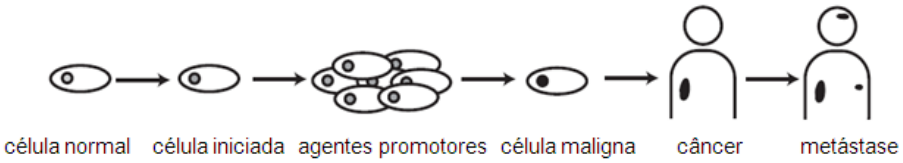

Figura 1.1: Evolução de uma célula normal em uma célula cancerosa. Os agentes cancerígenos conduzem a uma célula iniciada em cancerígena. Finalmente, células cancerígenas se espalham pelo corpo, formando os tumores.

No organismo humano existem mecanismos de defesa naturais que o protegem das agressões impostas por diferentes agentes que entram em contato com suas diferentes estruturas. Ao longo da vida são produzidas células alteradas, mas esses mecanismos de defesa possibilitam a interrupção desse processo, com sua eliminação subsequente. A capacidade de reparo do DNA danificado por agentes cancerígenos e a ação de enzimas responsáveis pela transformação e eliminação de substâncias cancerígenas introduzidas no corpo são exemplos de mecanismos de defesa. Esses mecanismos, próprios do organismo, são na maioria das vezes geneticamente pré-determinados, e variam de um indivíduo para outro. Esse fato explica a existência de vários casos de câncer numa mesma família, bem como o porquê de nem todo fumante desenvolver câncer de pulmão. Sem dúvida, o sistema imunológico desempenha um importante papel nesse mecanismo de defesa. Ele é constituído por um sistema de células distribuídas numa rede complexa de órgãos, como o fígado, o baço, os gânglios linfáticos, o timo e a medula óssea. Esses órgãos são denominados órgãos linfóides e estão relacionados ao crescimento, desenvolvimento e a distribuição das células especializadas na defesa do corpo. Dentre essas células, os linfócitos desempenham um papel muito importante nas atividades do sistema imune, relacionadas à produção de defesa deste processo da carcinogênese. Cabe aos linfócitos a atividade de atacar as células do corpo infectadas por vírus oncogênicos (capazes de causar câncer) ou as células em transformação maligna, bem como de secretar substâncias chamadas de linfocinas. As linfocinas regulam o crescimento e o amadurecimento de outras células e do próprio sistema imune. Acredita-se que distúrbios em sua produção ou em suas estruturas sejam causas de doenças, principalmente do câncer. Sem dúvida, a compreensão dos mecanismos exatos de ação do sistema imunológico muito contribuirá

para o entendimento da carcinogênese e, portanto, para novas estratégias de tratamento e de prevenção do câncer.

As primeiras tentativas de modelar a carcinogênese foram feitas nos anos 50 do século XX por Nordling (1953) e Armitage & Doll (1954), e os modelos sugeridos por estes autores são do tipo de multi-estágios. O modelo mais popular desse tipo na literatura é o modelo de dois estágios desenvolvidos por Dewanji *et al.* (1989), vide também Tan (1991) e as referências nele. Esta classe de modelos se ajusta aos dados experimentais muito bem, mas, devido à sua estrutura complexa, nem sempre são adequadas, além de não incorporarem na modelagem a possibilidade de cura dos indivíduos.

Recentemente, motivados pelos avanços dos tratamentos médicos (e o mecanismo defesas naturais do organismo) surgem entre os pesquisadores o interesse em proporem modelos de sobrevivência para carcinogênese que incorporam a possibilidade de indivíduos não serem suscetíveis ao câncer, ou seja, há uma parte da população que, devido a certa intervenção (tratamento e/ou defesas naturais do organismo) visando impedir a ocorrência do câncer, pode vir a não ser suscetível ao câncer (indivíduos fora de risco). O modelo clássico de Berkon-Gage (Boag, 1949; Berkson & Gage, 1952), estudado por Farewell (1982, 1986), Goldman (1984), Sy & Taylor (2000), Banerjee & Carlin (2004), entre muitos outros, assim como modelos mais recentes e abrangentes (Yakovlev & Tsodikov, 1996; Chen *et al.*, 1999; Ibrahim *et al.*, 2001; Chen *et al.*, 2002; Yin & Ibrahim, 2005) incorporam a possibilidade de avaliar a população curada de diversas formas.

A ocorrência do evento de interesse (câncer) pode ser provocada por uma ou várias causas competitivas (células); vide Gordon (1990). O número de causas, assim como o tempo de sobrevivência associado a cada causa, não são observados (Cox & Oakes, 1984) e são denominados de fatores ou riscos latentes. O modelo proposto por Chen *et al.* (1999) baseia-se na existência de fração de cura com fatores latentes, assim como, por exemplo, Yakovlev & Tsodikov (1996), Ibrahim *et al.* (2001), Chen *et al.* (2002), Banerjee & Carlin (2004) e Yin & Ibrahim (2005). Outra abordagem é desenvolvida por Kim *et al.* (2011), que modelam estocasticamente a sequência ordenada de tempos latentes, os quais induzem a ocorrência do evento em estudo. O cenário de causas competitivas permite longa duração quando a probabilidade do número de

riscos latentes ser igual a zero é não nula. Vale ressaltar que os modelos de cura podem ser propostos sem necessidade de modelar os riscos latentes.

O número de riscos latentes pode ser modelado por qualquer distribuição com média positiva e finita e suporte discreto, por exemplo, as distribuições de Poisson, binomial negativa, geométrica, Bernoulli e COM-Poisson (Chen *et al.*, 1999; Cooner *et al.*, 2007; Rodrigues *et al.*, 2011, 2009b; de Castro *et al.*, 2009). O modelo de Berkon-Gage (Berkson & Gage, 1952) pode ser considerado como um desses casos em que o número de riscos latentes tem distribuição de Bernoulli e há no máximo um risco latente.

Entretanto, a maioria dos modelos de sobrevivência com fração de cura encontrados na literatura para dados de carcinogênese apresentam duas limitações básicas:

(i) a suposição de que cada célula iniciada (causa competitiva ou fator de risco) torna-se maligna com probabilidade um e

(ii) a suposição de independência das células iniciadas ao tornarem-se malignas.

Para a limitação (i) nós encontramos poucos trabalhos na literatura sobre os modelos de fração de cura considerando a capacidade de reparo do DNA da célula iniciada, ou seja, a maioria dos modelos baseia-se sobre eventos que precedem a ocorrência da primeira célula maligna em um tecido e, portanto, o processo de reparo da célula é ignorado. Tendo como um limite de contrapartida a inclusão do processo de reparo da célula, isto nos levou à primeira motivação do presente trabalho. Para a limitação (ii), Haynatzki *et al.* (2000) discutiram que a suposição de independência pode não ser verdadeira quando a dinâmica da população de células de um tecido normal é considerada. Similarmente, há indícios de que as células pré-malignas (iniciadas) e malignas em um tecido influenciam no desenvolvimento umas das outras. Além disso, a interação entre as células saudáveis e pré-malignas no tecido devem ser levadas em consideração. Portanto, é desejável construir modelos estatísticos que possam incorporar adequadamente a dependência, e isto é que proporcionou a segunda motivação para o presente trabalho.

Portanto, o objetivo principal deste trabalho é apresentar alternativas para superar no mínimo uma das duas limitações básicas expostas acima dos modelos de sobrevivência com fração de cura para modelagem de dados de experimentos clínicos de câncer. Para esse fim, propomos

modelos de sobrevivência com fração de cura que podem acomodar características dos estágios não observáveis (iniciação, promoção e progressão) do processo da carcinogênise na presença de causas competitivas latentes independentes ou dependentes.

No Capítulo 2 propomos modelos de sobrevivência, denominados modelos de sobrevivência destrutivos correlacionados, os quais estendem os modelos formulados por Rodrigues *et al.* (2011) no sentido de incorporamos uma estrutura de dependência entre as células iniciadas. Pela inferência clássica e bayesiana obtivermos as estimativas dos parâmetros. Estudos de simulação foram realizados para analisar as propriedades frequentistas do processo de estimação clássico. Os modelos propostos foram aplicados a um conjunto de dados reais. Os resultados obtidos neste capítulo foram condensados no artigo aceito para publicação Borges *et al.* (2012).

Nos Capítulos 3 e 4 propomos modelos de sobrevivência baseados em um esquema de ativação latente híbrido para as células. A principal vantagem desta suposição é que podemos estimar as taxas de iniciação e proliferação de células cancerígenas. A diferença entre os dois capítulos está no fato de que as células iniciadas (causas competitivas) definidas no Capítulo 3 são assumidas independentes, enquanto no Capítulo 4 pressupormos que qualquer par de células são igualmente correlacionado. Realizarmos estudos de simulação para verificar as propriedades frequentistas do procedimento de estimação. Os modelos foram ajustados a um conjunto de dados reais para exemplificar a abordagem e a interpretação dos parâmetros. Resultaram destes capítulos, dois relatórios técnicos Borges *et al.* (2011a,b), foram submetidos para publicação. Finalmente, no Capítulo 5 apresentamos as considerações finais e listamos algumas linhas de pesquisas futuras.

A implementação computacional dos algoritmos e a elaboração dos gráficos foram desenvolvidas nos sistemas OpenBUGS 3.0.3 (Thomas *et al.*, 2006) e R (R Development Core Team, 2011). Os programas podem ser obtidos mediante solicitação ao autor.

Capítulo 2

Modelo com fração de cura destrutivo correlacionado

Rodrigues *et al.* (2010, 2011) propuseram um modelo estocástico para dados de sobrevivência com uma fração de cura (também conhecido como modelo com fração de cura destrutivo), que desempenha um papel importante em estudos biomédicos envolvendo um processo de reparação individual ou eliminação de células tumorais após um tratamento prolongado de câncer. Uma aplicação interessante é o modelo de irradiação prolongada para detectar tumores em um determinado período de tempo (Klebanov *et al.*, 1993). A literatura sobre os modelos de fração de cura está crescendo rapidamente, mas existem poucos trabalhos considerando a capacidade de reparar danos causados pela radiação ou eliminar as células cancerígenas após algum tratamento intensivo. As provas rádio-biológicas existentes sobre as características temporais de reparação enzimática mencionadas por Klebanov *et al.* (1993) motivaram Rodrigues *et al.* (2010, 2011) a considerarem o modelo com fração de cura destrutivo para descrever o processo biológico de eliminação de células alteradas (também chamadas de danificadas ou iniciadas) depois de algum tratamento específico, mas assumindo independência das células. Sugerimos ao leitor o artigo de Klebanov *et al.* (1993) para conhecer algumas referências específicas sobre este assunto. Além disso, os livros de Maller & Zhou (1996) e Ibrahim *et al.* (2001), bem como os artigos recentes de Tsodikov *et al.* (2003), Cooner *et al.* (2007), Tournoud & Ecochard (2007), Mizoi *et al.* (2007), de Castro *et al.* (2009), Ortega *et al.* (2009), Zhao *et al.* (2009) e Kim *et al.* (2011) podem ser

mencionados como alguns exemplos de modelos com fração de cura.

Neste capítulo propomos um novo modelo de sobrevivência com fração de cura, que estende o modelo de Rodrigues *et al.* (2010, 2011) no sentido que pressupormos que qualquer par de células são igualmente correlacionado (Haynatzki *et al.*, 2000). Para modelar a estrutura de dependência entre as células, nós usamos uma extensão da distribuição série de potência generalizada incluindo um parâmetro adicional ρ (distribuição *série de potências generalizada inflada*, SPGI, estudada por Kolev et al., 2000). O parâmetro ρ tem uma interpretação natural em termos de proporção de zeros adicionais e coeficiente de correlação. Em nossa abordagem, o número de células iniciadas segue uma distribuição SPGI. A principal vantagem desta distribuição é que a estrutura de correlação induzida pelo parâmetro adicional ρ resulta em uma caracterização natural da associação entre as células iniciadas. Além disso, fornece uma interpretação simples e realista do mecanismo biológico da ocorrência do evento de interesse (câncer), uma vez que inclui um processo de destruição das células tumorais após o tratamento inicial ou a capacidade de um indivíduo exposto à radiação para reparar células iniciadas que resulta em indução de câncer.

O Capítulo está organizado da seguinte forma. Na Seção 2.1 apresentamos a formulação do modelo. Alguns casos especiais do modelo proposto são apresentados na Seção 2.2. Na Seção 2.3 discutimos o processo inferencial clássico e bayesiano. Na Seção 2.4, apresentamos os resultados de um pequeno estudo de simulação que avalia a probabilidade de cobertura dos intervalos de confiança assintóticos. Na Seção 2.5 um conjunto de dados reais de câncer melanoma ilustra a utilidade do modelo proposto. Comentários finais são apresentados na Seção 2.6.

2.1 Formulação do modelo

Para um indivíduo na população, denotamos N o número de células iniciadas relacionados com a ocorrência de um tumor. A variável N é inobservada (variável latente). Em Rodrigues *et al.* (2010, 2011) N segue uma distribuição Poisson ponderada com parâmetros η e ϕ (Castillo & Pérez-Casany, 1998, 2005) e função massa de probabilidade ($f.m.p.$) da forma

$$p_n = \mathbb{P}[N = n; \eta, \phi] = \frac{w(n; \phi)p^*(n; \eta)}{\mathbb{E}_\eta[w(N; \phi)]},\ n = 0, 1, 2, \ldots, \tag{2.1}$$

em que $w(\cdot\ ;\ \phi)$ é uma função peso não negativa com parâmetro $\phi > 0$, $p^*(\cdot\ ;\ \eta)$ é a $f.m.p.$ de uma distribuição de Poisson com parâmetro $\eta > 0$, e $\mathbb{E}_\eta[\cdot]$ indica que o valor esperado é tomado com relação à variável Poisson com média η. Dependendo da escolha funcional de $w(\cdot\ ;\ \phi)$ obtemos importantes casos especiais de (2.1), incluindo as distribuições de Poisson, geométrica, binomial negativa, série logarítmica e COM-Poisson. Assim, o modelo proposto por Rodrigues *et al.* (2010, 2011) é mais flexível em termos de dispersão do que o modelo de tempo de promoção (Yakovlev & Tsodikov, 1996; Chen *et al.*, 1999), mas não incorpora uma estrutura de dependência entre as células iniciadas. Visando modelar a estrutura de dependência entre as células, a variável N seguirá uma distribuição SPGI com $f.m.p.$ dada por

$$p_n = \mathbb{P}[N = n; \theta, \rho] = \frac{1}{g(\theta)} \sum_{n_1, n_2, \ldots} a_n \big[\theta(1-\rho)\big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \quad n = 0, 1, 2, \ldots,\ \rho \in [0, 1), \tag{2.2}$$

em que $a_n > 0$ depende somente de n, $g(\theta) = \sum_{n=0}^{\infty} a_n \theta^n$ é uma função diferenciável, finita e positiva, e $\theta \in (0, s)$ (s pode ser ∞), e o somatório é sobre o conjunto de todos os inteiros não negativos $n_1, n_2, \ldots$, tais que $\sum_{i=1}^{\infty} i n_i = n$. O parâmetro $\rho \in [0, 1)$ tem uma interpretação natural em termos de proporção de zeros adicionais e coeficiente de correlação; para mais detalhes sobre a distribuição SPGI, vide Kolev *et al.* (2000) e Minkova (2002). Desta forma, utilizamos o parâmetro ρ como uma medida de associação entre as células. Precisamente, $\rho = Corr(W_r, W_s)$, $\forall\ r \neq s$, em que W_r é uma variável binária definida como

$$W_r = \begin{cases} 0 & , \text{ se a } r\text{-ésima célula é saudável} \\ 1 & , \text{ se a } r\text{-ésima célula é pré-maligna ou iniciada} \end{cases} . \tag{2.3}$$

A sequência de variáveis binárias $\{W_1, W_2, \ldots\}$, são utilizadas na construção do modelo SPGI; vide Kolev *et al.* (2000). O modelo SPGI permite apenas a presença de correlação positiva entre as células. Valores de $\rho \to 1$ indicam forte associação entre as células (isto é, as células em um tecido têm um alto grau de influência no desenvolvimento umas das outras), enquanto $\rho \to 0$ implica fraca associação entre as células (baixo grau de influência). É interessante notar que quando $\rho = 0$ (isto é, quando há independência entre as células), a distribuição SPGI torna-se uma distribuição série de potências generalizada (Gupta, 1974; Consul, 1990). A Tabela 2.1 mostra as escolhas de a_n, $g(\theta)$ e o parâmetro θ correspondentes a alguns casos especiais da distribuição SPGI, a saber,

distribuição Poisson inflada (PI), binomal negativa inflada (BNI), binomial inflada (BI) e série logarítmica inflada (SLI). Nos casos BI e BNI, os parâmetros adicionais $m_b \in \mathbb{Z}^+$ (conjunto dos inteiros não negativos) e $\phi > -1$ devem ser tratados como parâmetros perturbadores.

Tabela 2.1: Características da distribuição SPGI para algumas distribuições especiais.

Distribuições	a_n	$g(\theta)$	θ	s
PI	$\frac{1}{n_1!n_2!\cdots}$	e^{θ}	η	∞
BI	$\binom{m_b}{m_b-n_1-n_2-\cdots,n_1,n_2,\ldots}$	$(1+\theta)_b^m$	$\frac{\pi}{1-\pi}$	1
BNI	$\frac{\Gamma\left(\phi^{-1}+\sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!}$	$(1-\theta)^{-\phi^{-1}}$	$\frac{\phi\eta}{1+\phi\eta}$	∞
SLI	$\frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots}$	$-\log(1-\theta)$	$1-\pi$	1

A função geradora de probabilidade ($f.g.p.$) da variável aleatória SPGI N é dada por

$$\mathbb{A}_N(z) = \frac{g\big(\theta z(1-\rho)(1-z\rho)^{-1}\big)}{g(\theta)}, \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.4}$$

Após um tratamento prolongado ("processo destrutivo"), temos como consequência imediata a formação ou não de lesões cancerosas em um genoma das células. As células com lesões cancerosas serão denominadas malignas. Dado $N = n$, sejam X_j, $j = 1, 2, \ldots, n$, variáveis aleatórias independentes, independentemente de N, seguindo uma distribuição Bernoulli com probabilidade de sucesso p indicando a presença da j-ésima lesão e $f.g.p.$

$$\mathbb{A}_{X_j}(z) = 1 - p(1-z), \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.5}$$

A variável D representando o número total de células malignas dentre as N células iniciadas não eliminadas pelo tratamento é então dada por

$$D = \begin{cases} \sum_{j=1}^{N} X_j & , \quad se \quad N > 0 \\ 0 & , \quad se \quad N = 0 \end{cases}. \tag{2.6}$$

Notamos que $D \leq N$. A distribuição condicional de D, dado $N = n$ é Binomial(n;p). A variável D é não observável. Dado $D = d$, sejam V_j, $j = 1, \ldots, d$, variáveis aleatórias independentes, independentemente de D, com uma função de distribuição $F(y) = 1 - S(y)$. A variável aleatória V_j denota o tempo de progressão da j-ésima célula maligna em um tumor detectável, e $S(y)$ denota a função de sobrevivência.

Esta visão de (2.6) foi sugerida anteriormente por Yang & Chen (1991) no contexto de um estudo de bioensaio. Eles assumiram que os fatores de risco iniciais são células malignas iniciadas primárias, em que X_j em (2.6) denota o número de células malignas vivas que são descendentes da j-ésima célula maligna iniciada durante algum intervalo de tempo . Neste contexto, D denota o número total de células malignas que vivem em algum momento específico.

No cenário de causas competitivas (Cox & Oakes, 1984) das células malignas, o número de células iniciadas (N), malignas (D) e o tempo de progressão V_j são não observáveis (variáveis latentes). Assim, o tempo observável de início do tratamento até detecção do tumor (que é o evento de interesse) em um determinado indivíduo é definido pela variável aleatória

$$Y = \min\big(V_1, V_2, \ldots, V_D\big) \tag{2.7}$$

para $D \geq 1$, e $Y = \infty$ se $D = 0$, o que leva a uma proporção p_0 da população não susceptível à ocorrência do tumor, também denominada de fração de cura.

De acordo com Rodrigues *et al.* (2009b, 2011), a função de sobrevivência de longa duração da variável aleatória Y em (2.7) é dada por

$$S_{pop}(y) = P[Y \geq y] = \mathbb{A}_D(S(y)) = \sum_{d=0}^{\infty} P[D = d]\{S(y)\}^d = \mathbb{A}_N\Big(\mathbb{A}_{X_j}\big(S(y)\big)\Big),$$

sendo $\mathbb{A}_D(\cdot)$ é a função geradora de probabilidade da variável D, a qual converge quando $z = S(y) \in [0, 1]$. Levando em conta (2.4) e (2.5), a função de sobrevivência de longa duração do tempo observado de um tumor detectável em (2.7) é dada por

$$S_{pop}(y) = \frac{g\Big(\theta\big(1-\rho\big)\big[1 - pF(y)\big]\big\{1 - \big[1 - pF(y)\big]\rho\big\}^{-1}\Big)}{g(\theta)}. \tag{2.8}$$

Se usarmos especificamente $\rho = 0$, obtemos a função de sobrevivência de longa duração série de potências generalizada.

Dada uma função sobrevivência $S(\cdot)$, nós temos

$$\lim_{y\to\infty} S_{pop}(y) = p_0 = \frac{g\big(\theta(1-\rho)(1-p)\big[1-(1-p)\rho\big]^{-1}\big)}{g(\theta)}, \tag{2.9}$$

sendo que p_0 denota a proporção de indivíduos curados ou imunes presentes na população a partir do qual os dados da amostra foram obtidos. Referimo-nos ao modelo definido em (2.8) por modelo destrutivo correlacionado série de potências generalizada inflada, ou simplesmente o modelo DCSPGI. A Figura 2.1 ilustra o modelo DCSPGI em termos de um diagrama.

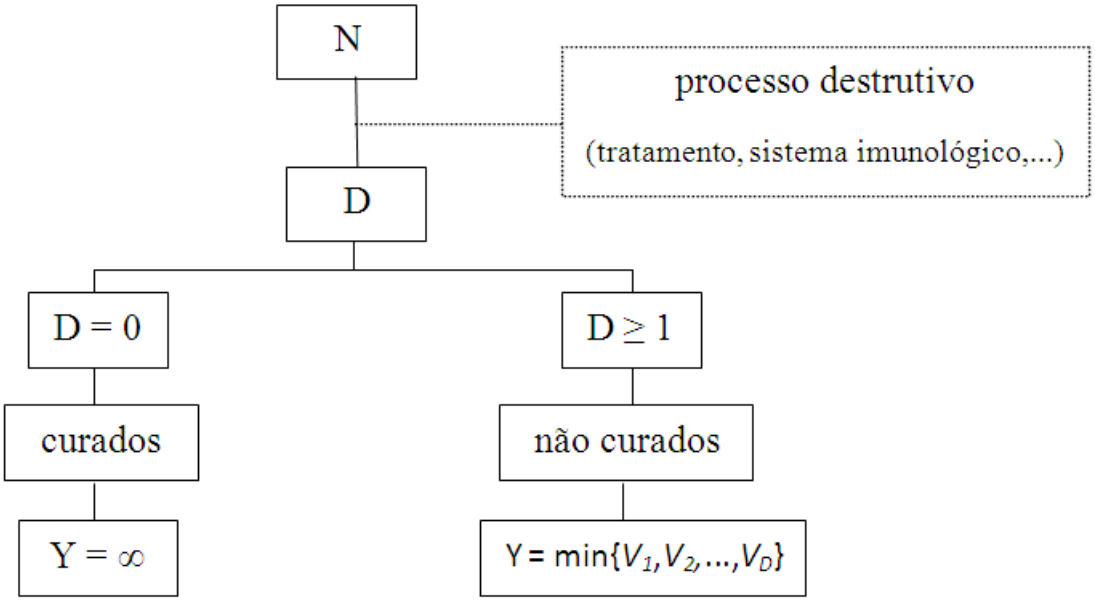

Figura 2.1: Representação do modelo DCSPGI.

2.2 Casos especiais do modelo proposto

Nesta seção, apresentamos alguns casos especiais do modelo DCSPG proposto na seção anterior.

2.2.1 Modelo destrutivo correlacionado Poisson (DCP)

Para as escolhas de $a_n = \frac{1}{n_1!n_2!\ldots}$, $g(\theta) = \exp(\theta)$ e o parâmetro $\theta = \eta$, dizemos que o número de células iniciadas N segue uma distribuição Poisson inflada com parâmetros $\eta > 0$ e $\rho \in [0, 1)$,

e sua *f.m.p.* é da forma

$$\mathbb{P}_{Poi}[N=n] = \sum_{n_1,n_2,\cdots} \frac{e^{-\eta}}{n_1!n_2!\cdots}\Big[\eta(1-\rho)\Big]^{\sum_{i=1}^{\infty} n_i}\rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{2.10}$$

em que $n = 0, 1, 2, \ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, n_3, \ldots$, tais que $\sum_{i=1}^{\infty} in_i = n$. Uma expressão alternativa para a *f.m.p.* em (2.10) (Kolev *et al.*, 2000; Minkova, 2002) é dada por

$$\mathbb{P}_{Poi}[N=n] = \begin{cases} e^{-\eta} & , \quad n=0 \\ e^{-\eta}\sum_{i=1}^{n}\binom{n-1}{i-1}\frac{\left[\eta(1-\rho)\right]^i\rho^{n-1}}{i!} & , \quad n=1,2,\ldots \end{cases}. \tag{2.11}$$

A média e a variância de N são

$$\mathbb{E}[N] = \frac{\eta}{1-\rho} \quad \text{e} \quad \mathbb{V}ar[N] = \frac{\eta(1+\rho)}{(1-\rho)^2}, \tag{2.12}$$

respectivamente. A *f.g.p.* é dada por

$$\mathbb{A}_N(z) = \exp\left\{-\frac{\eta(1-z)}{1-z\rho}\right\} \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.13}$$

Assim, a função de sobrevivência de longa duração do modelo DCP é dada por

$$S_{pop}(y) = \exp\left\{-\frac{\eta p F(y)}{1-\rho[1-pF(y)]}\right\}. \tag{2.14}$$

Existem dois importantes casos especiais de (2.14). Para $\rho = 0$, obtemos o modelo destrutivo Poisson (Rodrigues *et al.*, 2011), enquanto para $\rho = 0$ e $p = 1$, obtemos o modelo de tempo de promoção (Yakovlev & Tsodikov, 1996; Chen *et al.*, 1999).

2.2.2 Modelo destrutivo correlacionado binomial (DCB)

Para as escolhas de $a_n = \binom{m_b}{m_b-n_1-n_2-\cdots,n_1,n_2,\ldots}$, $g(\theta) = (1+\theta)_b^m$ e $\theta = \frac{\pi}{1-\pi}$, o número de células iniciadas N segue uma distribuição binomial inflada com parâmetros $\pi \in (0,1)$, $\rho \in [0,1)$ e $m_b \in \mathbb{Z}^+$, e sua *f.m.p.* é da forma

$$\mathbb{P}_{Bin}[N=n] = (1-\pi)_b^m \sum_{n_1,n_2,\ldots}\binom{m_b}{m_b-n_1-n_2-\cdots,n_1,n_2,\ldots}\rho^n\left\{\frac{\pi(1-\rho)}{\rho(1-\pi)}\right\}^{\sum_{i=1}^{\infty} n_i}, \tag{2.15}$$

em que $n = 0, 1, \ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, \ldots$, tais que $\sum_{i=1}^{\infty} i n_i = n$. Uma expressão alternativa para a *f.m.p.* em (2.15) (Kolev *et al.*, 2000; Minkova, 2002) é dada por

$$\mathbb{P}_{Bin}[N = n] = \begin{cases} (1-\pi)^m & , \quad n = 0 \\ \sum_{i=1}^{\min(n,m_b)} \binom{m_b}{i} \binom{n-1}{i-1} \left[\pi(1-\rho)\right]^i (1-\pi)^{m_b - i} \rho^{n-i} & , \quad n = 1, 2, \ldots \end{cases} . \quad (2.16)$$

A média e a variância de N são

$$\mathbb{E}[N] = \frac{m_b \pi}{1-\rho} \quad \text{e} \quad \mathbb{V}ar[N] = \frac{m_b \pi (1 - \pi + \rho)}{(1-\rho)^2}, \quad (2.17)$$

respectivamente. A *f.g.p.* é dada por

$$\mathbb{A}_N(z) = \left[1 - \frac{\pi(1-z)}{1 - z\rho}\right]_b^m \quad \text{para} \quad 0 \leq z \leq 1. \quad (2.18)$$

Assim, a função de sobrevivência de longa duração do modelo DCB é dada por

$$S_{pop}(y) = \left[1 - \frac{\pi p F(y)}{1 - \rho(1 - pF(y))}\right]_b^m . \quad (2.19)$$

Agora, fazendo $m_b \to \infty$ e $\pi \to 0$ em (2.19) tal que $m_b \pi = \eta p > 0$, obtemos no limite

$$\lim_{m_b \to \infty} \lim_{\pi \to 0} S_{pop}(y) = \lim_{m_b \to \infty} \left[1 - \frac{\eta p F(y)}{m_b(1 - \rho(1 - pF(y)))}\right]_b^m = \exp\left\{-\frac{\eta p F(y)}{1 - \rho(1 - pF(y))}\right\},$$

que é de fato a função de sobrevivência de longa duração do modelo DCP apresentado anteriormente em (2.14). Se tomarmos $m_b = p = 1$ e $\rho = 0$, o modelo DCB coincide com o modelo de mistura padrão (Boag, 1949; Berkson & Gage, 1952).

2.2.3 Modelo destrutivo correlacionado binomial negativa (DCBN)

Para as escolhas de $a_n = \frac{\Gamma\left(\phi^{-1} + \sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!}$, $g(\theta) = (1-\theta)^{-\phi^{-1}}$ e o parâmetro $\theta = \frac{\phi\eta}{1+\phi\eta}$, o número de células iniciadas N segue uma distribuiçao binomial negativa inflada com parâmetros $\eta > 0$, $\rho \in [0, 1)$, $\phi \geq -1$ e $\phi\eta > 0$, e sua *f.m.p.* é da forma

$$\mathbb{P}_{NB}[N = n] = (1 + \phi\eta)^{-\phi^{-1}} \sum_{n_1, n_2, \ldots} \frac{\Gamma\left(\phi^{-1} + \sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!} \left[\frac{\phi\eta(1-\rho)}{1+\phi\eta}\right]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \quad (2.20)$$

em que $n = 0, 1, \ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, \ldots$, tais que $\sum_{i=1}^{\infty} in_i = n$, e $\Gamma(\cdot)$ denota a função gama. Uma expressão alternativa para a *f.m.p.* em (2.20) (Kolev *et al.*, 2000; Minkova, 2002) é dada por

$$\mathbb{P}_{NB}[N = n] = \begin{cases} (1+\phi\eta)^{-\phi^{-1}} & , \quad n = 0 \\ (1+\phi\eta)^{-\phi^{-1}} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{\Gamma(\phi^{-1}+i)}{\Gamma(\phi^{-1})i!} \left[\frac{\phi\eta(1-\rho)}{1+\phi\eta}\right]^i \rho^{n-i} & , \quad n = 1, 2, \ldots \end{cases} \tag{2.21}$$

A média e a variância de N são

$$\mathbb{E}[N] = \frac{\eta}{1-\rho} \quad \text{e} \quad \mathbb{V}ar[N] = \frac{\eta(1+\rho+\phi\eta)}{(1-\rho)^2}, \tag{2.22}$$

respectivamente. A *f.g.p.* é dada por

$$\mathbb{A}_N(z) = \left[\frac{1-z\rho}{1+\phi\eta(1-z)-z\rho}\right]^{\phi^{-1}}, \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.23}$$

Assim, a função de sobrevivência de longa duração do modelo DCBN é dada por

$$S_{pop}(y) = \left[\frac{1-\rho(1-pF(y))}{1+\phi\eta pF(y)-\rho(1-pF(y))}\right]^{\phi^{-1}}. \tag{2.24}$$

Quando $\phi = 1$, obtemos a distribuição geométrica inflada com parâmetros $\theta = \frac{1}{1+\eta} \in (0, 1)$ em (2.20) ou (2.21). Neste caso $S_{pop}(\cdot)$ em (2.24) torna-se

$$S_{pop}(y) = \frac{1-\rho(1-pF(y))}{1+\eta pF(y)-\rho(1-pF(y))}, \tag{2.25}$$

dando origem ao modelo destrutivo correlacionado geométrico, ou simplesmente modelo DCG. Quando $\phi \to 0$, obtemos o modelo DCP.

2.2.4 Modelo destrutivo correlacionado série logarítmica (DCSL)

Para escolhas de $a_n = \frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots}$, $g(\theta) = -\log(1-\theta)$ e $\theta = 1-\pi$, o número de células iniciadas N segue uma distribuição série logarítmica com parâmetros $\pi \in (0, 1)$ e $\rho \in [0, 1)$, e sua *f.m.p.* é da forma

$$\mathbb{P}_{LS}[N = n] = (-\log(\pi))^{-1} \sum_{n_1, n_2, \ldots} \frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots} \big[(1-\pi)(1-\rho)\big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{2.26}$$

em que $n = 0, 1, \ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, \ldots$, tais que $\sum_{i=1}^{\infty} in_i = n$. Uma expressão alternativa para a *f.m.p.* em (2.26) (Kolev *et al.*, 2000; Minkova, 2002) é dada por

$$\mathbb{P}_{LS}[N = n] = (-\log(\pi))^{-1} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{\left[(1-\pi)(1-\rho)\right]^i \rho^{n-i}}{i}, \quad n = 1, 2, \ldots. \tag{2.27}$$

Em sua forma original, esta distribuição exclui o valor zero. Consequentemente, não pode ser usada para modelar o número de células iniciadas (no sentido de incluir a longa duração). Por esta razão, consideramos aqui uma distribuição série logarítmica inflada modificada, cuja *f.m.p.* pode ser escrita como

$$\mathbb{P}_{LS}[N = n] = (-\log(\pi))^{-1} \sum_{i=1}^{n+1} \binom{n}{i-1} \frac{\left[(1-\pi)(1-\rho)\right]^i \rho^{n+1-i}}{i}, \quad n = 0, 1, 2, \ldots. \tag{2.28}$$

A média e a variância da variável aleatória série logarítmica inflada modificada N são

$$\mathbb{E}[N] = 1 - \frac{1-\pi}{\pi(1-\rho)\log(\pi)} \quad \text{e} \quad \mathbb{V}ar[N] = -\frac{(1-\pi)\left[\log(\pi)(1+\pi\rho) + 1 - \pi\right]}{\pi^2(1-\rho)^2(\log(\pi))^2}, \tag{2.29}$$

respectivamente. A *f.g.p.* é dada por

$$\mathbb{A}_N(z) = \frac{(-\log(\pi))^{-1}}{z} \log\left\{\frac{1-\rho z}{1-z(1-\pi(1-\rho))}\right\}, \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.30}$$

Assim, a função de sobrevivência de longa duração do modelo DCSP modificado é dada por

$$S_{pop}(y) = \frac{(-\log(\pi))^{-1}}{(1-pF(y))} \log\left\{\frac{1-\rho(1-pF(y))}{1-(1-pF(y))(1-\pi(1-\rho))}\right\}. \tag{2.31}$$

Na Tabela 3.1, apresentamos a função de sobrevivência de longa duração e a fração de cura, bem como a função de densidade imprópria $f_{pop}(y) = -\frac{dS_{pop}(y)}{dy}$, correspondentes aos casos particulares apresentados nas Seções 2.2.1, 2.2.2, 2.2.3 e 2.2.4.

Tabela 2.2: Função de sobrevivência de longa duração ($S_{pop}(y)$), função de densidade ($f_{pop}(y)$) e fração de cura (p_0) para diferentes casos especiais.

Modelo	$S_{pop}(y)$	$f_{pop}(y)$	p_0
DCP	$\exp\left\{-\frac{\eta p F(y)}{1-\rho(1-pF(y))}\right\}$	$\left[\frac{\eta p f(y)\left[1-\rho(1-pF(y))\right]-\eta\rho p^2 f(y)F(y)}{\left[1-\rho(1-pF(y))\right]^2}\right] S_{pop}(y)$	$\exp\left\{-\frac{\eta p}{1-\rho(1-p)}\right\}$
DCB	$\left[1-\frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]_b^m$	$m_b\left[1-\frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]^{-1}\left[\frac{\pi p f(y)\left[1-\rho(1-pF(y))\right]-\pi p^2 F(y)\rho f(y)}{\left[1-\rho(1-pF(y))\right]^2}\right] S_{pop}(y)$	$\left[1-\frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]_b^m$
DCBN	$\left[\frac{1-\rho(1-pF(y))}{1+\phi\eta pF(y)-\rho(1-pF(y))}\right]^{\phi^{-1}}$	$\phi^{-1}\left[\frac{1-\rho(1-pF(y))}{1+\phi\eta pF(y)-\rho(1-pF(y))}\right]^{-1}\left[\frac{\left[1-\rho(1-pF(y))\right]\left[\phi\eta p f(y)+\rho p f(y)\right]-\rho p f(y)\left[1+\phi\eta pF(y)-\rho(1-pF(y))\right]}{\left[1+\phi\eta pF(y)-\rho(1-pF(y))\right]^2}\right] S_{pop}(y)$	$\left[\frac{1-\rho(1-p)}{1+\phi\eta p-\rho(1-p)}\right]^{\phi^{-1}}$
DCSL	$\frac{(-\log(\pi))^{-1}}{(1-pF(y))}\log\left[\frac{1-\rho(1-pF(y))}{1-(1-pF(y))(1-\pi(1-\rho))}\right]$	$\frac{\left[1-(1-pF(y))(1-\pi(1-\rho))\right]}{\log(\pi)(1-\rho(1-pF(y)))(1-pF(y))}\left[\frac{\rho p f(y)}{1-(1-pF(y))(1-\pi(1-\rho))}-\frac{(1-\rho(1-pF(y)))pf(y)(1-\pi(1-\rho))}{\left[1-(1-pF(y))(1-\pi(1-\rho))\right]^2}\right]-\frac{pf(y)S_{pop}(y)}{1-pF(y)}$	$\frac{(-\log(\pi))^{-1}}{(1-p)}\log\left[\frac{1-\rho(1-p)}{1-(1-p)(1-\pi(1-\rho))}\right]$

2.3 Inferência

2.3.1 Estimação de máxima verossimilhança

Para a formulação da função de verossimilhança consideram-se as notações a seguir. N_j é o número de células iniciadas relacionadas à ocorrência do câncer (evento de interesse) no j-ésimo indivíduo, $j = 1, 2, \ldots, m$, que são variáveis aleatórias independentes não observadas com distribuição de probabilidade SPGI com parâmetros θ e ρ. D_j dado $N_j = n_j$ é o número de células iniciadas não eliminadas pelo tratamento no j-ésimo indivíduo, $j = 1, 2, \ldots, m$, que são variáveis aleatórias independentes não observadas com distribuição binomial com n_j e probabilidade de sucesso p.

Sejam $V_{j1}, V_{j2}, \ldots, V_{jD_j}$ variáveis aleatórias independentes identicamente distribuídas que representam o tempo de ocorrência do câncer (evento de interesse) para as D_j células malignas no j-ésimo indivíduo, com função distribuição indicada por $F(t_j;\boldsymbol{\gamma}) = 1 - S(t_j;\boldsymbol{\gamma})$ e $\mathbb{P}[V_{j0} = \infty] = 1$, sendo que $\boldsymbol{\gamma}$ representa o vetor de parâmetros da distribuição. Seja Y_j como definido em (2.7) e sujeito a censura não informativa à direita. Assim, t_j é o tempo observado dado por $T_j = \min\big(Y_j, C_j\big)$, em que C_j é o tempo de censura, enquanto que δ_i é a variável indicadora de falha tal que $\delta_j = 1$ se $Y_j \leq C_j$, e $\delta_j = 0$, caso contrário, $j = 1, 2, \ldots, m$. Propomos relacionar os parâmetros p e η (ou π) dos modelos da Tabela 2.2 com os vetores de covariáveis $\boldsymbol{x}'_j = (x_{j1}, \ldots, x_{jk_1})$ e $\boldsymbol{w}'_j = (w_{j1}, \ldots, w_{jk_2})$, respectivamente. Adotemos as funções de ligação

$$\log\left(\frac{p_j}{1-p_j}\right) = \boldsymbol{x}'_{\boldsymbol{j}}\boldsymbol{\beta_1}, \quad \text{e} \quad \log(\eta_j) = \boldsymbol{w}'_{\boldsymbol{j}}\boldsymbol{\beta_2} \quad \text{ou} \quad \log\left(\frac{\pi_j}{1-\pi_j}\right) = \boldsymbol{w}'_{\boldsymbol{j}}\boldsymbol{\beta_2}, \quad j = 1, \ldots, m, \tag{2.32}$$

em que $\boldsymbol{\beta}'_{\mathbf{1}} = (\beta_{11}, \ldots, \beta_{1k_1})$ e $\boldsymbol{\beta}'_{\mathbf{2}} = (\beta_{21}, \ldots, \beta_{2k_2})$ vetores com k_1 e k_2 coeficientes de regressão. Além disso, para $\rho = 0$ os modelos DCP, DCB e DCBN são inidentificáveis no sentido de Li *et al.* (2001). Para evitar este problema, quando ajustarmos esses modelos, os vetores de covariáveis $\boldsymbol{x}'_j$ e $\boldsymbol{w}'_j$, não compartilham elementos comuns.

Uma questão crítica é a seleção de covariáveis a serem incluídas nas funções de ligação em (2.32). Infelizmente, este problema não será abordado aqui. Para os leitores interessados sugerimos os livros de Draper & Smith (1998) e Collet (1994) (contexto clássico) ou artigo de George & McCulloch (1993) (contexto bayesiano).

Os dados completos e observados são denotados por $\boldsymbol{D}_c = (m, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\delta}, \boldsymbol{N}, \boldsymbol{D})$ e $\boldsymbol{D}_{obs} = (m, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\delta})$, respectivamente, sendo que $\boldsymbol{t}' = (t_1, \ldots, t_m)$, $\boldsymbol{\delta}' = (\delta_1, \ldots, \delta_m)$, $\boldsymbol{N}' = (N_1, \ldots, N_m)$, $\boldsymbol{D}' = (D_1, \ldots, D_m)$, $\boldsymbol{X}' = (\boldsymbol{x_1'}, \boldsymbol{x_2'}, \ldots, \boldsymbol{x_m'})$ e $\boldsymbol{W}' = (\boldsymbol{w_1'}, \boldsymbol{w_2'}, \ldots, \boldsymbol{w_m'})$.

O próximo lema será fundamental para obter a função de verossimilhança dos parâmetros do modelo DCSPGI.

Lema 2.1 *Sob o modelo com fração de cura destrutivo, a densidade condicional de* (t_j, δ_j) *dado* $N_j = n_j$ *e* $D_j = d_j$, $j = 1, \ldots, m$, *é dada por*

$$f(t_j, \delta_j | n_j, d_j) = \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} I_{\{d_j \leq n_j\}}, \tag{2.33}$$

sendo I_A *a função indicadora do evento* $A = \{d_j \leq n_j\}$.

Prova 2.1 *Vide apêndice A em Mizoi (2004).*

A função de verossimilhança do modelo DCSPGI com censura não-informativa é dada por

$$L(\boldsymbol{\vartheta}; \boldsymbol{D}_c) = \prod_{j=1}^{m} \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} \mathbb{P}[N_j = n_j, D_j = d_j] \tag{2.34}$$

em que $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \rho, \phi)'$ denota o vetor de parâmetros do modelo DCSPGI.

Note que a função de verossimilhança (2.34) depende de $\boldsymbol{N}$ e $\boldsymbol{D}$ que são variáveis latentes. A função de verossimilhança marginal é dada por

$$\begin{aligned}
L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) &= \prod_{j=1}^{m} \sum_{n_j=0}^{\infty} \sum_{d_j=0}^{n_j} \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} \mathbb{P}[N_j = n_j, D_j = d_j] \\
&= \prod_{j=1}^{m} \sum_{d_j=0}^{\infty} \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} \sum_{n_j=0}^{\infty} \cdots \sum_{n_j=d_j}^{\infty} \mathbb{P}[N_j = n_j, D_j = d_j] \\
&= \prod_{j=1}^{m} \underbrace{\sum_{d_j=0}^{\infty} \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} \mathbb{P}[D_j = d_j]}_{\big\{f_{pop}(t_j;\boldsymbol{\gamma})\big\}^{\delta_j} \big\{S_{pop}(t_j;\boldsymbol{\gamma})\big\}^{1-\delta_j}, \text{ vide de Castro } et\ al. \text{ (2007)}} \\
&= \prod_{j=1}^{m} \big\{f_{pop}(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} \big\{S_{pop}(t_j; \boldsymbol{\gamma})\big\}^{1-\delta_j}.
\end{aligned} \tag{2.35}$$

Agora supondo uma distribuição Weibull para o tempo de progressão de cada célula (V_j), cuja distribuição e função densidade são dadas, respectivamente, por

$$F(v; \boldsymbol{\gamma}) = 1 - \exp(-v^{\gamma_1} e^{\gamma_2}) \quad \text{e} \quad f(v; \boldsymbol{\gamma}) = \gamma_1 v^{\gamma_1 - 1} \exp(\gamma_2 - v^{\gamma_1} e^{\gamma_2}), \tag{2.36}$$

para $v > 0$, $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2)$, com $\gamma_1 > 0$ e $\gamma_2 \in \Re$. Embora outras distribuições de tempos de vida pudessem ser usadas aqui, nossa escolha foi baseada no fato que a distribuição Weibull é uma das mais amplamente usadas para representar tempos de vida na análise de sobrevivência devido a sua versatilidade na captura de diferentes formas. Dependendo do valor de seu parâmetro de forma, γ_1, a distribuição Weibull é capaz de modelar uma variedade de comportamentos de tempos de vida. Sua função de risco é monótona decrescente para $\gamma_1 < 1$, para $\gamma_1 > 1$ é monótona crescente e para $\gamma_1 = 1$ é constante, equivalendo à distribuição exponencial; vide Johnson *et al.* (1994).

As estimativas de máxima verossimilhança de $\widehat{\boldsymbol{\vartheta}}$ são obtidas maximizando o logaritmo da função de verossimilhança em (2.35), $\ell(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) = \log(L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}))$. A maximização é efetuada numericamente aplicando o método L-BFGS-B, implementado na função *optim* do sistema R (R Development Core Team, 2012). Sob certas condições de regularidade, pode ser mostrado (Fahrmeir, 1988) que $\widehat{\boldsymbol{\vartheta}}$ têm distribuição assintótica normal multivariada, $\mathcal{N}\big(\boldsymbol{\vartheta}, \boldsymbol{I^{-1}(\vartheta)}\big)$, em que

$$\boldsymbol{I(\vartheta)} = \mathbb{E}\left(-\frac{\partial^2 \log L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'}\right) \tag{2.37}$$

é a matriz informação de Fisher. Além disso $\boldsymbol{I_0(\vartheta)} = -\frac{\partial^2 \log L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'}\Big|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}}$, denominada de matriz de informação observada, é um estimador consistente de $\boldsymbol{I(\vartheta)}$. Neste trabalho, o cálculo da matriz de informação observada é feito numericamente por meio da linguagem R.

Para comparar os modelos que surgem a partir da formulação geral apresentada na Seção 2.1, podemos considerar o AIC (critério de informação Akaike) e o BIC (critério de informação bayesiano), definidos, respectivamente, por $-2\log L(\widehat{\boldsymbol{\vartheta}_g}) + 2q$ e $-2\log L(\widehat{\boldsymbol{\vartheta}_g}) + q\log(m)$, sendo que $\widehat{\boldsymbol{\vartheta}_g}$ é a estimativa de máxima verossimilhança sob o modelo g, q é o número de parâmetros estimados sob o modelo g e m é o tamanho amostral. Os melhores modelos correspondem a menores valores de AIC e BIC.

2.3.2 Inferência Bayesiana

Como alternativa à inferência clássica dada pela maximização da função de verossimilhança, sugerimos a inferência bayesiana. Nesta abordagem, combinamos a função de verossimilhança com informações *a priori* obtendo a distribuição *a posteriori*. As estimativas dos parâmetros são

então dadas pelas médias das distribuições *a posteriori*.

Uma das formas de assegurarmos que a distribuição *a posteriori* seja própria é considerar distribuições *a priori* próprias (Ibrahim *et al.*, 2001). Embora não seja necessário, por simplicidade, assumiremos que os parâmetros $\boldsymbol{\beta_1'}$, $\boldsymbol{\beta_2'}$, γ_1, γ_2, ρ e ϕ são independentes *a priori*, isto é,

$$\pi(\boldsymbol{\vartheta}) = \prod_{j_1=1}^{k_1} \pi(\beta_{1j_1}) \prod_{j_2=1}^{k_2} \pi(\beta_{2j_2})\pi(\gamma_1)\pi(\gamma_2)\pi(\rho)\pi(\phi), \tag{2.38}$$

sendo $\beta_{1j_1} \sim\mathcal{N}\big(0, \sigma^2_{1j_1}\big)$, $j_1 = 1, \ldots, k_1$, $\beta_{2j_2} \sim\mathcal{N}\big(0, \sigma^2_{2j_2}\big)$, $j_2 = 1, \ldots, k_2$, $\gamma_1 \sim Gama\big(a_0, a_1\big)$, $\gamma_2 \sim\mathcal{N}\big(0, \sigma^2_{\gamma_2}\big)$ e $\rho \sim Beta\big(b_0, b_1\big)$, enquanto que $\phi \sim Gama\big(c_0, c_1\big)$ para o modelo DCBN. Todos os hiperparâmetros são especificados com o objetivo de garantir distribuições *a priori* vagas.

Combinando a função de verossimilhança (2.35) com a distribuição *a priori* em (2.38), a distribuição *a posteriori* para $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \rho, \phi)$ é obtida como $\pi\big(\boldsymbol{\vartheta}|\boldsymbol{t}, \boldsymbol{\delta}\big) \propto \pi(\boldsymbol{\vartheta})L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})$. Esta densidade *a posteriori* é analiticamente intratável. Como alternativa usamos os métodos de Monte Carlo com cadeias de Markov (MCMC), como por exemplo, o amostrador de Gibbs; vide Gamerman & Lopes (2006). Para a implementação do algoritmo são necessárias as distribuições condicionais completas de todos os parâmetros, dadas por

$$\pi(\boldsymbol{\beta_1}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_1}), \quad \pi(\boldsymbol{\beta_2}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_2}),$$

$$\pi(\gamma_1|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\gamma_1}), \quad \pi(\gamma_2|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\gamma_2) \quad \text{e}$$

$$\pi(\rho|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\rho), \quad \pi(\phi|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\phi).$$

Todas estas distribuições condicionais não são distribuições conhecidas. Então, precisamos usar algum algoritmo (por exemplo, Metropolis-Hasting) para simular amostras de $\boldsymbol{\vartheta}$. O código computacional foi implementado no sistema OpenBUGS 3.0.3 (Thomas *et al.*, 2006).

2.3.3 Critério para comparação de modelos

Existe uma variedade de metodologias para comparar vários modelos ajustados a um mesmo conjunto de dados e selecionar aquele que melhor se ajusta aos dados. Nestes casos é conveniente o uso de um critério de seleção de modelos. Um dos critérios comumente utilizados é baseado na ordenada da densidade preditiva condicional (CPO); vide Gelfand *et al.* (1992). Denotamos

$\boldsymbol{D}_{obs}^{(-j)}$ os dados observados com a j-ésima observação excluída. Em nosso modelo, para um tempo até a ocorrência do evento observado ($\delta_j = 1$), definimos $g(t_j;\boldsymbol{\vartheta}) = f_{pop}(t_j;\boldsymbol{\vartheta})$ e, para um tempo censurado, $g(t_j;\boldsymbol{\vartheta}) = S_{pop}(t_j;\boldsymbol{\vartheta})$, em que $f_{pop}(\cdot)$ e $S_{pop}(\cdot)$ são como na Tabela 2.2. Denotaremos a densidade *a posteriori* de $\boldsymbol{\vartheta}$ dado $\boldsymbol{D}_{obs}^{(-j)}$, por $\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs}^{(-j)})$, $j = 1,\ldots,m$. Para a j-ésima observação, CPO_j pode ser escrita como

$$CPO_j = \int_\Theta g(t_j;\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs}^{(-j)})d\boldsymbol{\vartheta} = \left\{\int_\Theta \frac{\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs})}{g(t_j;\boldsymbol{\vartheta})}d\boldsymbol{\vartheta}\right\}^{-1}. \tag{2.39}$$

O modelo escolhido é que apresenta o maior valor CPO_j (em média). Para o modelo proposto, uma forma fechada da CPO_j não está disponível. No entanto, uma estimativa Monte Carlo da CPO_j pode ser obtida por meio de uma amostra MCMC da distribuição *a posteriori* $\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs})$. Seja $\boldsymbol{\vartheta_1},\ldots,\boldsymbol{\vartheta_Q}$ uma amostra de tamanho Q de $\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs})$ após o aquecimento (burn-in). Uma aproximação Monte Carlo da CPO_j (Chen *et al.*, 2000) é dada por

$$\widehat{CPO_j} = \left\{\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{g(t_j;\boldsymbol{\vartheta}_q)}\right\}^{-1}. \tag{2.40}$$

Uma estatística resumo da $CPO_j's$ é $B = \sum_{j=1}^{m}\log\big(\widehat{CPO_j}\big)/m$. Quanto maior o valor de B, melhor o ajuste do modelo.

Há também critérios com base na média *a posteriori* do *desvio*, que é em si uma medida de ajuste. O *desvio* pode ser aproximado por $\overline{\mathbb{D}} = \sum_{q=1}^{Q}\frac{\mathbb{D}(\boldsymbol{\vartheta}_q)}{Q}$, sendo $\mathbb{D}(\boldsymbol{\vartheta}) = -2\sum_{j=1}^{m}\log\big(g(t_j;\boldsymbol{\vartheta})\big)$. Entre esses critérios, nós escolhemos o critério de informação do *desvio* (*DIC*) (Carlin & Louis, 2002), o critério de informação Akaike esperado (*EAIC*) (Brooks, 2002) e o critério de informação bayesiano esperado (*EBIC*) (Spiegelhalter *et al.*, 2002). O *DIC* pode ser estimado utilizando a amostra MCMC por $\widehat{DIC} = \overline{\mathbb{D}} + \widehat{\zeta}_{\mathbb{D}} = 2\overline{\mathbb{D}} - \widehat{\mathbb{D}}$, sendo $\zeta_{\mathbb{D}}$ o número efetivo de parâmetros definido como $\mathbb{E}[\mathbb{D}(\boldsymbol{\vartheta})] - \mathbb{D}(\mathbb{E}[\boldsymbol{\vartheta}])$, e $\mathbb{D}(\mathbb{E}[\boldsymbol{\vartheta}])$ o *desvio* avaliado na média *a posteriori*, que pode ser estimado por

$$\widehat{\mathbb{D}} = \mathbb{D}\left\{\frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\beta_{1_q}}, \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\beta_{2_q}}, \frac{1}{Q}\sum_{q=1}^{Q}\gamma_{1q}, \frac{1}{Q}\sum_{q=1}^{Q}\gamma_{2q}, \frac{1}{Q}\sum_{q=1}^{Q}\rho_q, \frac{1}{Q}\sum_{q=1}^{Q}\phi_q\right\}.$$

Da mesma forma, o *EAIC* e *EBIC* podem, também, ser estimados utilizando as amostras MCMC por meio de $\widehat{EAIC} = \overline{\mathbb{D}} + 2q$ e $\widehat{EBIC} = \overline{\mathbb{D}} + q\log(m)$, sendo que q é o número de parâmetros es-

timados sob o modelo g e m é o tamanho amostral. Na comparação de dois modelos alternativos, o modelo que tem o menor valor do critério utilizado é que se ajusta melhor aos dados.

2.4 Estudo de simulação

Com o intuito de verificar algumas propriedades frequentistas dos estimadores de máxima verossimilhança, realizamos um pequeno estudo de simulação. Neste estudo somente consideramos o modelo DCG da equação (2.25) (nosso modelo de trabalho na Seção 2.5). No processo de simulação, fixamos $\rho = 0,8$ e adotamos a distribuição de Weibull para os tempos de progressão com parâmetros $\gamma_1 = 5$ e $\gamma_2 = 2$. Assumimos para cada indivíduo duas covariáveis, x e w, sendo que estas foram consideradas fixas, mas tiveram seus valores gerados a partir de uma distribuição Bernoulli com parâmetro 0,5 e de uma distribuição normal com média 3 e variância 1, respectivamente. Relacionamos os parâmetros η, p do modelo DCG para covariáveis x e w, respectivamente. Adotamos as funções de ligação

$$\log(\eta_j) = \beta_{1_0} x_j + \beta_{1_1}(1 - x_j) \text{ e } \log\left(\frac{p_j}{1-p_j}\right) = \beta_{2_0} + \beta_{2_1} w_j,\ j = 1, \ldots, m, \tag{2.41}$$

sendo $\beta_{1_0} = 1$, $\beta_{1_1} = 1,5$, $\beta_{2_0} = -2,5$ e $\beta_{2_1} = 0,5$. A fração de cura é $p_{0j} = \frac{1-\rho(1-p_j)}{1+\eta_j p_j - \rho(1-p_j)}$ e a proporção de tempos censurados (φ_{cj}) é considerada como sendo igual a $(p_{0j} + 0,1)$. O intervalo de variação de p_{0j} nas simulações varia entre 18% e 60%. Os tempos observados e indicadores de censura são gerados por meio dos seguintes passos:

1. Gerar u_j ∼uniforme(0,1).

2. Se $u_j < p_{0j}$, então $y_j = \infty$; caso contrário,

$$y_j = \exp\left\{\frac{\log\left(-\log\left(\frac{u_j(1+\eta_j p_j - \rho + \rho p_j) + \rho(1-p_j) - 1}{p_j(u_j \eta_j - \rho(1-u_j))}\right)\right) - \gamma_2}{\gamma_1}\right\}.$$

3. Gerar c_j ∼exponencial(ξ_j), sendo o parâmetro ξ_j é escolhido de modo termos aproximadamente $\varphi_{cj}100\%$ de censura nos dados.

4. Fazer $t_j = \min(y_j, c_j)$.

5. Se $y_j < c_j$, então $\delta_j = 1$; caso contrário, $\delta_j = 0$, $j = 1, \ldots, m$.

Os tamanhos de amostras utilizados nas simulações foram m=50, 100, 200 e 400. Para cada conjunto de dados simulados, os parâmetros são estimados pelo método de máxima verossimilhança. A função log-verossimilhança foi maximizada numericamente usando o método L-BFGS-B, implementado na função *optim* do sistema R (R Development Core Team, 2012). Repetimos este processo 1000 vezes para cada configuração de amostras e calculamos a média e a raiz quadrada do erro quadrático médio (REQM) das estimativas dos parâmetros. Além disso, o intervalo de confiança de 95% foi obtido para cada parâmetro com base na teoria assintótica normal e observou-se se o intervalo de confiança continha o verdadeiro valor do parâmetro, determinando assim a probabilidade de cobertura (PC) dos intervalos de confiança para cada parâmetro. As simulações que não convergiram foram descartadas. Os resultados obtidos estão resumidos na Tabela 2.3. Podemos verificar que o REQM diminui com o aumento do tamanho da amostra e que as diferenças entre as estimativas médias e os valores verdadeiros, o denominado viés, são quase sempre menores que o REQM empírico, o que indica um bom desempenho dos estimadores de máxima verossimilhança. Em geral, as PCs empíricas parecem convergir para o nível nominal quando m aumenta. As conclusões deste estudo de simulação são limitados ao modelo DCG, mas nós acreditam que elas são semelhantes para outros modelos.

Tabela 2.3: Média, viés, REQM das estimativas de máxima verossimilhança e PC dos intervalos de confiança de 1000 repetições.

n	**parâmetro**	**média**	**viés**	**REQM**	**PC**
50	γ_1	5,66	0,66	0,74	0,93
	γ_2	1,94	-0,06	0,46	0,95
	ρ	0,78	-0,02	0,08	0,89
	β_{1_0}	0,94	-0,06	0,42	0,92
	β_{1_1}	1,49	-0,01	0,19	0,95
	β_{2_0}	-5,06	-2,56	2,39	0,89
	β_{2_1}	2,11	1,61	1,81	0,95
100	γ_1	5,32	0,32	0,44	0,94
	γ_2	1,93	-0,07	0,39	0,95
	ρ	0,77	-0,03	0,09	0,91
	β_{1_0}	0,94	-0,06	0,42	0,92
	β_{1_1}	1,43	-0,07	0,21	0,94
	β_{2_0}	-3,74	-1,24	0,77	0,95
	β_{2_1}	1,36	0,86	0,58	0,95
200	γ_1	5,16	0,16	0,32	0,95
	γ_2	1,95	-0,05	0,13	0,95
	ρ	0,78	-0,02	0,08	0,95
	β_{1_0}	0,93	-0,07	0,20	0,93
	β_{1_1}	1,42	-0,08	0,18	0,95
	β_{2_0}	-3,14	-0,64	0,66	0,95
	β_{2_1}	1,16	0,66	0,34	0,95
400	γ_1	5,07	0,07	0,12	0,95
	γ_2	1,97	-0,03	0,10	0,95
	ρ	0,76	-0,04	0,08	0,95
	β_{1_0}	0,95	-0,05	0,04	0,94
	β_{1_1}	1,48	-0,02	0,17	0,95
	β_{2_0}	-2,51	-0,01	0,55	0,95
	β_{2_1}	0,79	0,29	0,25	0,95

2.5 Dados de câncer de melanoma

A incidência de melanoma maligno cutâneo, um câncer comum da pele, está aumentando dramaticamente em pessoas com pele de cor clara em todas as partes do mundo. Este tipo de câncer é a segunda causa de perda de vida potencial nos últimos anos, afetando os indivíduos adultos mais jovens, atrás apenas da leucemia e causando um problema de saúde pública (Barral, 2001).

Nesta seção apresentamos uma aplicação dos modelos descritos na Seção 2.2 a um conjunto de dados de melanoma maligno, que foi coletado no hospital universitário de Odense, Dinamarca, por K. T. Drzewiecki. Ressaltamos que esse conjunto de dados não enfatiza o processo da carcinôgenese descrito no capítulo 1, entretanto ele pode ser modelado certamente pelos modelos descritos na Seção 2.2, contanto que pensamos nesses dados como sendo gerado por um processo de três estágios. O conjunto de dados inclui 205 pacientes observados após uma cirurgia para a remoção de melanoma maligno no período de 16 anos. Estes dados estão disponíveis no pacote *timereg* no R (Scheike, 2009). O tempo observado (Y) varia de 10 a 5565 dias (de 0,0274 a 15,25 anos, com média = 5,9 e desvio-padrão = 3,1 anos) e se refere ao tempo até a morte do paciente ou o tempo de censura. Pacientes que morreram de outras causas, bem como pacientes que ainda estavam vivos ao final do estudo são observações censuradas (72%). Tomamos o indicador de úlcera (ausente, $m = 115$; presente, $m = 90$) e espessura do tumor (em mm, média = 2,92 e desvio padrão = 2,96) como covariáveis. Tendo em mente a questão da identificabilidade mencionada anteriormente na Seção 2.3 nos modelos DCP, DCB e DCBN, o parâmetro p é ligado apenas à espessura do tumor, enquanto que o parâmetro η (ou π) está ligado apenas ao indicador de úlcera. A curva Kaplan-Meier estratificada pelo indicador de úlcera (ulc) na Figura 2.2 estabiliza acima de 0,4. Este comportamento sugere claramente que os modelos que ignoram a possibilidade de taxa de cura não serão adequados para analisar estes dados.

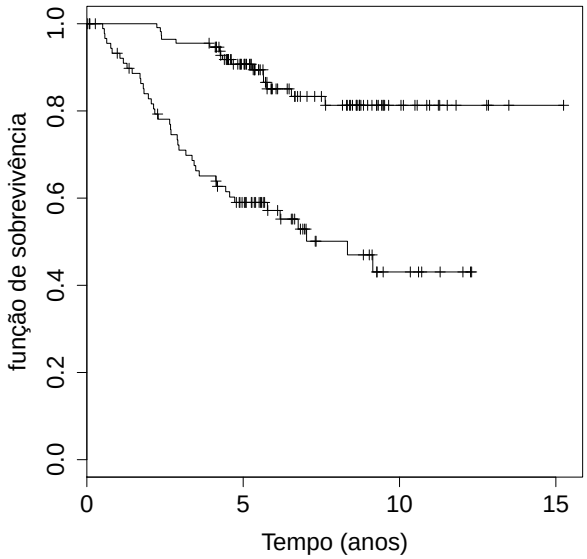

Figura 2.2: Curva de Kaplan-Meier estratificada pelo indicador de úlcera (superior: ausente, inferior: presente).

Ajustamos os modelos da Tabela 2.2 e o modelo DCG. Dois casos particulares do modelo DCBN também foram ajustados aos dados, a saber, os modelos binomial negativa ($p = 1$, $\rho = 0$) e geométrico ($p = 1$, $\phi = 1$ e $\rho = 0$). Desta forma, o mecanismo de destruição é ausente. Para estes modelos, o parâmetro η é ligado às duas covariáveis. Para o modelo DCB fixei o parâmetro $m_b = 15$. A Tabela 2.4 apresenta os valores do máximo da função log-verossimilhança, $\max \log L(\cdot)$, e os valores das estatísticas AIC e BIC para os modelos ajustados. As estatísticas AIC e BIC dão evidências a favor do modelo DCG e DCP. Utilizarei como modelo de trabalho o DCG. Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo DCG, seus desvios padrão e seus intervalos de confiança de 95% baseados na teoria assintótica são apresentados na Tabela 2.5. A estimativa do parâmetro correlação ρ é 0,95, e como mencionado anteriormente na Seção 2.1, isso indica uma forte associação entre as células. O gráfico QQ do resíduo dos quantis normalizado (Dunn & Smyth, 1996; Rigby & Stasinopoulos, 2005) na Figura 2.3 sugere que o modelo DCG é adequado.

Tabela 2.4: $Max \log L(\cdot)$ e as estatísticas AIC e BIC para os sete modelos ajustados.

	Modelo						
Critério	DCP	DCB	DCBN	DCG	DCSL	Binomial negativa	Geométrico
$\max \log L(\cdot)$	-198,60	-198,61	-198,12	-198,52	-197,96	-201,52	-205,42
AIC	411,21	413,21	412,24	411,06	413,92	415,04	420,83
BIC	434,47	439,80	438,82	434,32	443,83	435,00	437,45

Tabela 2.5: Estimativas de máxima verossimilhança dos parâmetros do modelo DCG, seus desvios padrão e seus intervalos de confiança assintóticos de 95% (IC 95%).

Parâmetro	**Estimativa**	**desvio padrão**	**IC** 95%
γ_1	2,46	0,34	(1,79 ; 3,12)
γ_2	-5,54	1,16	(-7,81 ; -3,26)
ρ	0,95	0,06	(0,83 ; 1,00)
$\beta_{1,intercepto}$	-4,84	0,95	(-6,70 ; -2,98)
$\beta_{1,espessura}$	0,95	0,27	(0,42 ; 1,48)
$\beta_{2,ulc:presente}$	0,63	0,30	(0,04 ; 1,22)
$\beta_{2,ulc:ausente}$	-0,48	0,41	(-1,28 ; 0,32)

A Figura 2.4 mostra a função sobrevivência para pacientes com espessura do tumor igual a 0,32, 1,94 e 8,32 mm, que correspondem aos quantis de 5%, 50% e 95%, respectivamente. A probabilidade de sobrevivência diminui mais rapidamente para os pacientes com tumores mais espessos. Na Figura 2.4 (a) a função de sobrevivência não é menor do que 0,7.

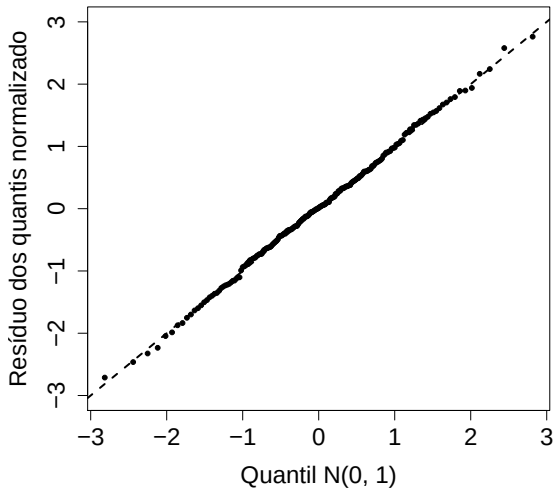

Figura 2.3: Gráfico QQ do resíduo dos quantis normalizado com a reta identidade para o modelo DCG (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados).

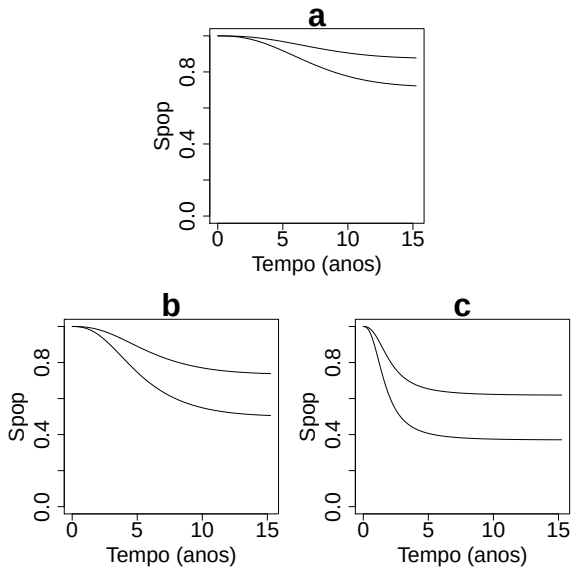

Figura 2.4: Função de sobrevivência sob o modelo DCG estratificado pelo indicador de úlcera (superior: ausente, inferior: presente) para pacientes com espessura do tumor igual a (a) 0,32, (b) 1,94, e (c) 8,32 mm, respectivamente.

O modelo DCG foi ajustado com os parâmetros p e η associados à espessura do tumor e ao indicador de úlcera, respectivamente. Se trocarmos essas covariáveis, não há melhora no ajuste com relação aos critérios na Tabela 2.4, uma vez que, neste caso, obtemos os valores do ($\max \log L(\cdot)$; AIC; BIC) iguais a (-204,61; 423,23; 446,49).

Finalmente, voltamos a nossa atenção para o papel das covariáveis sobre a fração de cura (vide Tabela 2.2). As estimativas dos coeficientes $\beta_{2,ulc}$ na Tabela 2.5 indicam que o número médio de células iniciadas é maior quando a úlcera está presente, de modo que a fração de cura diminui. Visto que $\hat{\beta}_{2,espessura} > 0$ na Tabela 2.5, os valores maiores da espessura do tumor implica em uma menor estimativa da fração de cura. A Figura 2.5 mostra o efeito combinado destas covariáveis sobre a fração de cura. As linhas correm quase paralelamente e as frações de cura, depois de uma queda acentuada, para espessura do tumor maior que 5mm, estão em 62,78% e 37,94% para o indicador de úlcera ausente e presente, respectivamente.

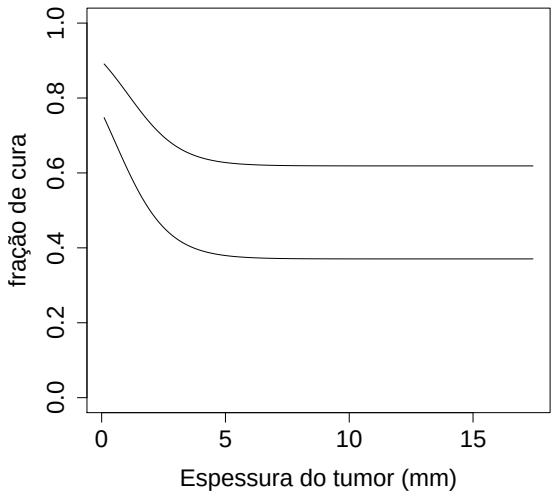

Figura 2.5: Fração de cura para o modelo DCG *versus* espessura do tumor estratificada pelo indicador de úlcera (superior: ausente, inferior: presente).

Também obtivemos os ajustes para os sete modelos da Tabela 2.4 através da inferência bayesiana. Utilizamos distribuições *a priori* independentes e não informativas, sendo $\beta_{1,intercepto} \sim \mathcal{N}\left(0, 10^3\right)$, $\beta_{1,espessura} \sim \mathcal{N}\left(0, 10^3\right)$, $\beta_{2,ulc:ausente} \sim \mathcal{N}\left(0, 10^3\right)$, $\beta_{2,ulc:presente} \sim \mathcal{N}\left(0, 10^3\right)$, $\gamma_1 \sim$

$Gama(1, 0, 01)$, $\gamma_2 \sim \mathcal{N}(0, 10^3)$ e $\rho \sim Beta(1, 1)$, enquanto que $\phi \sim Gama(1; 0, 001)$ para o modelo DCBN. Geramos duas cadeias paralelas de tamanho 35000 para cada parâmetro. Descartamos as primeiras 5000 e as restantes selecionadas de 10 em 10, resultando numa amostra de tamanho 3000. A convergência das cadeias foi monitorada empregando o método de Cowles & Carlin (1996).

Na Tabela 2.6 foram aplicados os critérios de seleção de modelos definidos na Seção 2.3.3 para os sete modelos ajustados. Os critérios dão evidências a favor do modelo DCG, seguido do modelo DCP. A Tabela 2.7 apresenta as médias *a posteriori*, os desvios padrão e os intervalos de credibilidade para os parâmetros do modelo DCG, incluindo o fator de redução de escala potencial estimado $\widehat{R}$ (Gelman & Rubin, 1992), que para todos os parâmetros está próximo de um, indicando a convergência das cadeias. A Figura 2.6 apresenta as densidades marginais *a posteriori* aproximadas para cada parâmetro.

Para avaliar a robustez do modelo com relação à escolha dos hiperparâmetros das distribuições *a priori*, um pequeno estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros não apresentam muita diferença e não alteram os resultados apresentados na Tabela 2.6.

Tabela 2.6: Critérios DIC, EAIC, EBIC e B para os sete modelos ajustados.

	Modelo						
Critério	DCP	DCB	DCBN	DCG	DCSL	Binomial negativa	Geométrico
DIC	406,21	407,73	407,01	406,56	415,52	413,63	416,31
EAIC	419,60	421,11	421,40	417,90	425,54	420,51	427,10
EBIC	442,86	447,68	447,98	441,16	448,76	440,44	443,72
B	-206,49	-205,92	-205,84	-206,33	-208,76	-206,97	-212,54

Tabela 2.7: Médias *a posteriori*, desvios padrão e intervalos de credibilidade de 95% (ICred 95%) para os parâmetros do modelo DCG e o fator de redução de escala potencial estimado $\widehat{R}$.

Parâmetro	**Média**	**desvio padrão**	**ICred 95%**	$\widehat{R}$
γ_1	2,25	0,33	(1,64 ; 2,89)	1,003
γ_2	-5,12	0,93	(-7,12 ; -3,56)	1,002
ρ	0,83	0,18	(0,52 ; 0,99)	1,004
$\beta_{1,intercepto}$	-4,05	0,90	(-5,72 ; -2,24)	1,001
$\beta_{1,espessura}$	0,53	0,38	(0,48 ; 1,99)	1,003
$\beta_{2,ulc:presente}$	0,74	0,34	(0,13 ; 1,49)	1,002
$\beta_{2,ulc:ausente}$	-0,31	0,43	(-1,07 ; 0,58)	1,001

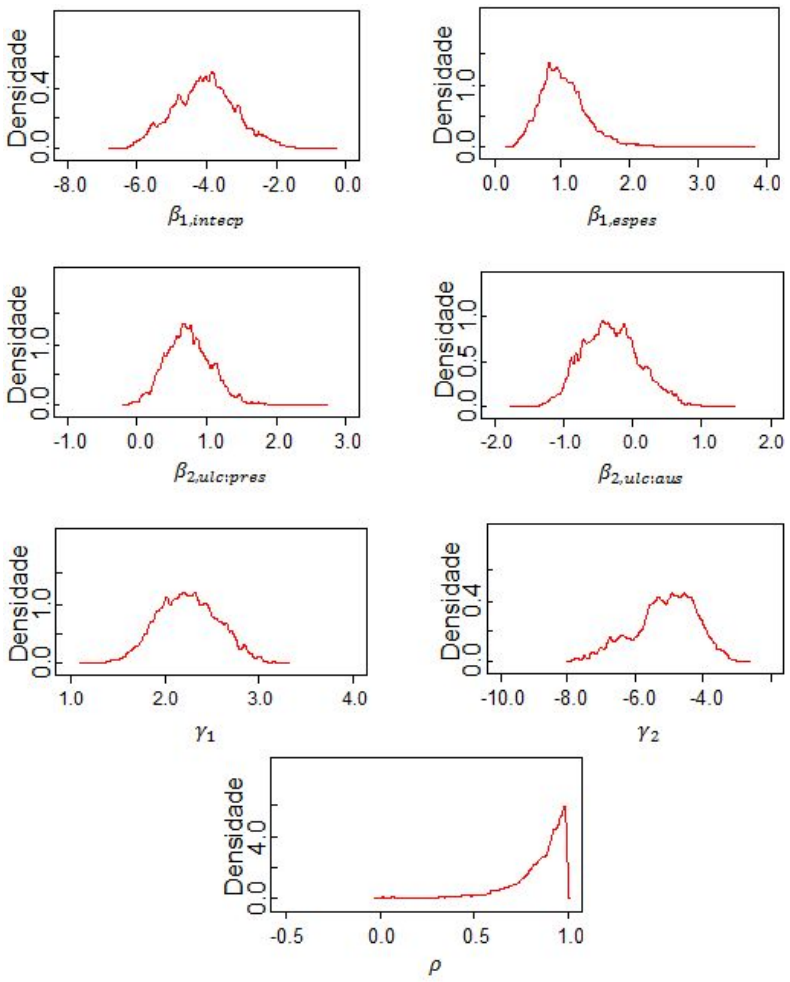

Figura 2.6: Densidades *a posteriori* aproximadas dos parâmetros.

2.6 Comentários finais

Neste capítulo propomos um modelo de sobrevivência com fração de cura, que estende o modelo de Rodrigues *et al.* (2010, 2011), no sentido de incorporarmos uma estrutura de dependência entre as células iniciadas. Assumimos uma distribuição SPGI para o número de células iniciadas e uma distribuição Weibull para os tempos de ocorrência do tumor, obtendo o modelo DCSPGI.

O modelo DCSPGI mostra explicitamente a contribuição para o tempo até o tumor ("tempo de falha") de três características distintas para o crescimento do tumor, o número médio de células iniciadas (θ), a proporção de células iniciadas "promovidas" a malignas (p) e a taxa de progressão ($F(y)$). Assim, o modelo incorpora parâmetros com claro significado biológico. Apesar da modelagem ser enfatizada pelo processo da carcinogênese, o modelo DCSPGI é satisfatório para qualquer tipo de dados de tempo de falha que têm uma fração de sobreviventes. Desta forma, dados de tempo de falha que não se "ajustam" a definição biológica dada no capítulo 1 pode ser modelada certamente pelo modelo DCSPGI, contanto que os dados tenham uma fração de sobreviventes e podem ser pensado como sendo gerado por um processo de três estágios. Assim o modelo pode ser útil para modelar vários tipos de dados de tempo de falha, incluindo o tempo para reincidência, tempo de morte, tempo para primeira infecção, e assim por diante. A aplicabilidade do modelo foi demonstrada em um conjunto de dados reais de pacientes com câncer de melanoma. Os dois processos de estimação apresentaram resultados próximos e implicam em conclusões similares a respeito do modelo a ser escolhido e das covariáveis a serem consideradas.

Capítulo 3

Modelo com fração de cura baseado em um esquema de ativação híbrido

Os modelos de sobrevivência para carcinogênese baseiam-se em eventos que precedem a ocorrência da primeira célula maligna em um tecido. Uma descrição explícita do estágio de progressão do tumor é omitida em modelos de dois estágios. Isso, também, é verdade com o modelo de radiação para carcinogênese proposto por Klebanov *et al.* (1993) e suas generalizações por Yakovlev & Polig (1996) e Rodrigues *et al.* (2010, 2011). Por esta razão, Yakovlev *et al.* (1996), Hanin *et al.* (1997) e Tsodikov *et al.* (1997) estabeleceram um limite de contrapartida do modelo de dois estágios da carcinogênese através da realização do estágio de progressão, que forneceu a motivação para o presente capítulo.

Portanto, o objetivo deste capítulo é descrever o mecanismo biológico da ocorrência do evento de interesse (tempo até um tumor detectável) levando em consideração os três estágios do processo da carcinogênese (iniciação, promoção e progressão). Com esse objetivo, um modelo de sobrevivência geral para carcinogênese espontânea baseado em um esquema híbrido latente de ativação para as células combinando o esquema de ativação máximo com o esquema de ativação mínimo (Cooner *et al.*, 2007) foi desenvolvido para permitir um padrão simples da dinâmica do crescimento do tumor. Assumimos que o número de células iniciadas e o número de células malignas (causas competitivas) seguem distribuições Poisson ponderadas. Supõe-se que o tumor (é monoclonal gerado durante o estágio de progressão) torna-se detectável quando seu tamanho

atinge certo nível limiar (proliferação de células tumorais geradas da célula maligna). A vantagem deste modelo é que ele incorpora características do estágio de progressão do tumor, bem como a proporção de células iniciadas que foram promovidas a malignas e a proporção de células malignas que morrem antes da indução de tumor.

O capítulo está organizado da seguinte forma. Na Seção 3.1 apresentamos a formulação do modelo. Alguns modelos específicos são apresentados na Seção 3.2. Na Seção 3.3 discutimos o processo inferencial, do ponto de vista clássico e bayesiano. Na Seção 3.4 apresentamos os resultados de um pequeno estudo de simulação que avalia a probabilidade de cobertura dos intervalos de confiança assintóticos. Na Seção 3.5 um conjunto de dados de câncer melanoma real ilustra a utilidade do modelo proposto. Comentários finais são apresentados na Seção 3.6.

3.1 Formulação do modelo

Na construção de nosso modelo geral, fazemos as seguintes suposições básicas:

(i) O evento de iniciação no processo da carcinogênese é a formação de uma lesão primária (ou pré-cancerosa) intracelular que, no longo prazo, é capaz de produzir um tumor evidente. Denotamos essas lesões pré-cancerosas como as células iniciadas. Tratamos o número de células iniciadas como uma variável aleatória N_1;

(ii) Todas as lesões primárias podem ser consideradas como estando sujeitas a processos de reparo (Ainsworth, 1982; Kopp-Schneider *et al.*, 1991) ou eliminadas depois de algum tratamento prolongado;

(iii) Uma lesão pré-cancerosa não reparada permanece dormente enquanto ela prossegue com o estágio de promoção de desenvolvimento do tumor. Todas as lesões estão sujeitas a promoção independentemente umas das outras;

(iv) Uma vez que a célula maligna ou clonogênica surge como resultado da promoção da célula iniciada, começa o estágio de progressão produzindo uma colônia de descendentes (células tumorais) , chamada de clone ou tumor. Tratamos o número de células malignas resultantes do estágio de promoção como uma variável aleatória N_2. O tempo que uma célula maligna

leva para se transformar em um tumor detectável é considerado como uma variável aleatória com função de distribuição $F(y) = 1 - S(y)$, sendo $S(y)$ função de sobrevivência. Todas as células malignas estão sujeitas a progressão independentemente umas das outras.

(v) Um tumor torna-se detectável quando o seu tamanho atinge um valor limite (proliferações de células tumorais). Tratamos o número de células tumorais como uma variável aleatória N_3.

Observação 3.1 *As suposições (i) e (iii) acima são suposições comuns presentes na maioria dos modelos modernos de sobrevivência em dois estágios encontrados na literatura, vide por exemplo, Chen* et al. *(1999), Cooner* et al. *(2007) e Rodrigues* et al. *(2009b).*

Com base nas suposições acima, o modelo proposto pode ser desenvolvido da seguinte maneira. Para um sujeito na população, seja N_1 o número de células iniciadas com função massa de probabilidade (*f.m.p.*) $p_{n_1} = \mathbb{P}[N_1 = n_1]$ para $n_1 = 0, 1, \ldots$. Após um tratamento prolongado (ou sistema de reparo) temos como uma consequência imediata a formação ou não de células malignas. Dado $N_1 = n_1$, sejam X_l, $l = 1, \ldots, n_1$, variáveis aleatórias independentes, independentemente de N_1, seguindo uma distribuição Bernoulli com probabilidade de sucesso p indicando que a l-ésima célula iniciada tornou-se maligna. Seja N_2 o número total de células malignas que surgem como resultado da promoção entre as $N_1 = n_1$ células iniciadas não eliminadas pelo tratamento, definida como

$$N_2 = \begin{cases} \sum_{l=1}^{N_1} X_l & , \quad se \quad N_1 > 0 \\ 0 & , \quad se \quad N_1 = 0 \end{cases} . \tag{3.1}$$

Notamos que $N_2 \leq N_1$. A distribução condicional de N_2, dado $N_1 = n_1$ é Binomial(n_1;p).

Agora, seja $N_{3i} = N_3$, $i = 1, 2, \ldots, N_2$, o número de células tumorais originadas da i-ésima célula maligna com *f.m.p.* $p_{n_3} = \mathbb{P}[N_3 = n_3]$ para $n_3 = 0, 1, \ldots$. O tempo para que a (i, j)-ésima célula maligna se transforme em um tumor detectável, denominado tempo de progressão, é denotado por Z_{ij}, para $i = 1, \ldots, N_2$ e $j = 1, \ldots, N_3$. Assumimos que, dado $N_k = n_k$, para $k = 1, 2, 3$, as variáveis $Z'_{ij}s$ são independentes com função distribuição $F(y) = 1 - S(y)$, independentes de N_k.

No cenário de causas competitivas (Cox & Oakes, 1984) das células malignas, o número de células iniciadas (N_1), malignas (N_2), tumorais (N_3) e o tempo Z_{ij} são inobserváveis. Assim, o

tempo observável de início do tratamento até a detecção do tumor (evento de interesse) para um dado indivíduo é definido como a variável aleatória

$$Y = \min\left\{\max\left\{Z_{ij}\right\}_{j=1}^{N_3}\right\}_{i=1}^{N_2}, \tag{3.2}$$

para $N_2 \geq 1$ e $N_3 \geq 1$, e $Y = \infty$ se $N_2 = 0$, o que leva uma proporção p_0 da população não susceptível à ocorrência do tumor, também, denominada de fração de cura, ou $Y = \infty$ se $N_3 = 0$, o que leva a uma proporção p_0^* de células malignas que morrem antes da indução do tumor.

Observação 3.2 *A variável Y é representada por um esquema híbrido latente de ativação para as células combinando o esquema de ativação pelo máximo com o esquema de ativação pelo mínimo (vide Cooner* et al. *(2007) para mais detalhes de esquemas de ativação), ou seja, Y representa o máximo dos tempos de progressão das células tumorais e o mínimo destes máximos gerando o tempo até um tumor detectável.*

A função de sobrevivência da variável aleatória Y será indicada por

$$S_{pop}(y) = \mathbb{P}[Y > y]. \tag{3.3}$$

Teorema 3.1 *Dada a função de sobrevivência (suposição (iv)), $S(y) = 1 - F(y)$, dos tempos de progressão não observáveis Z_{ij} , a função de sobrevivência da variável aleatória Y em (3.2) é dada por*

$$S_{pop}(y) = \mathbb{A}_{N_1}\Big(1 - p\big(1 - S^*_{pop}(y)\big)\Big) = \sum_{n_1=0}^{\infty} p_{n_1}\Big\{1 - p\big(1 - S^*_{pop}(y)\big)\Big\}^{n_1}, \tag{3.4}$$

sendo que $\mathbb{A}_{N_1}(\cdot)$ é a f.g.p. da variável N_1, que converge se $s = 1 - p\big(1 - S^*_{pop}(y)\big) \in [0,1]$, e

$$S^*_{pop}(y) = 1 + \mathbb{P}[N_3 = 0] - \mathbb{A}_{N_3}\Big(F(y)\Big), \tag{3.5}$$

a qual denotaremos como a função de sobrevivência do estágio de progressão, em que $\mathbb{A}_{N_3}(\cdot)$ é a f.g.p. da variável N_3, que converge se $s = F(y) \in [0,1]$.

Prova 3.1 *Temos que*

$$
\begin{aligned}
S_{pop}(y) &= \sum_{l=0}^{\infty}\left\{\mathbb{P}[N_2=0|N_1=l]+\mathbb{P}\left[\bigcap_{i=1}^{N_2}\max\{Z_{ij}\}_{j=0}^{N_3}>y; N_2\leq l\right]\right\}\mathbb{P}[N_1=l] \\
&= \sum_{l=0}^{\infty}\Big\{\sum_{i=0}^{l}\big\{\mathbb{P}[\max\{Z_{ij}\}_{j=0}^{N_3}>y]\big\}^i\mathbb{P}[N_2=i|N_1=l]\Big\}\mathbb{P}[N_1=l] \\
&= \sum_{l=0}^{\infty}\Big\{\sum_{i=0}^{l}\big\{1-\mathbb{P}[Z_{i1}<y,\ldots,Z_{iN_3}<y; N_3\geq 1]\big\}^i\mathbb{P}[N_2=i|N_1=l]\Big\}\mathbb{P}[N_1=l] \\
&= \sum_{l=0}^{\infty}\Big\{\sum_{i=0}^{l}\big\{1-\underbrace{\sum_{j=1}^{\infty}F(y)^j\mathbb{P}[N_3=j]}_{\mathbb{A}_{N_3}\left(F(y)\right)-\mathbb{P}[N_3=0]}\big\}^i\mathbb{P}[N_2=i|N_1=l]\Big\}\mathbb{P}[N_1=l] \\
&\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\left\{1-p+p\left(1+\mathbb{P}[N_3=0]-\mathbb{A}_{N_3}(F(y))\right)\right\}^l} \\
&= \sum_{l=0}^{\infty}\Big\{1-p+p\Big(1+\mathbb{P}[N_3=0]-\mathbb{A}_{N_3}(F(y))\Big)\Big\}^l\mathbb{P}[N_1=l] \\
&= \mathbb{A}_{N_1}\Big(1-p+pS^*_{pop}(y)\Big) \\
&= \mathbb{A}_{N_1}\Big(1-p\big(1-S^*_{pop}(y)\big)\Big). \qquad (3.6)
\end{aligned}
$$

A última expressão sintetiza de forma simples e objetiva os três estágios do processo da carcinogênise por meio de uma composição da função geradora de probabilidade do número de células iniciadas (N_1), a proporção de células iniciadas que foram promovidas a malignas (p) e a função de sobrevivência do estágio de progressão.

As funções de sobrevivência $S_{pop}(y)$ e $S^*_{pop}(y)$ em (3.4) e (3.5), respectivamente, não são próprias, isto é, $\lim_{y\to\infty} S_{pop}(y) > 0$ e $\lim_{y\to\infty} S^*_{pop}(y) > 0$, como mostra o próximo teorema.

Teorema 3.2 *Dada a função de sobrevivência própria, $S(y) = 1 - F(y)$, temos*

$$
\lim_{y\to\infty} S^*_{pop}(y) = \mathbb{P}[N_3=0] = p_0^* \qquad e \qquad \lim_{y\to\infty} S_{pop}(y) = \mathbb{A}_{p_{n_1}}\big(1-p(1-p_0^*)\big) = p_0, \qquad (3.7)
$$

em que p_0 denota a proporção de indivíduos curados ou imunes que podem estar presentes na população a partir do qual os dados são obtidos, e p_0^* denota a proporção de células malignas que morrem antes da indução do tumor.

Prova 3.2 *Os resultados são obtidos facilmente de (3.4) e (3.5), respectivamente.*

Observação 3.3 *O parâmetro p_0^* em (3.7) pode ser utilizado para avaliar a eficiência de um tratamento. Valores de $p_0^* \to 1$ indicam alta eficiência do tratamento, levando ao aumento de p_0, enquanto $p_0^* \to 0$ implica baixa eficiência do tratamento, p_0 diminui.*

Observação 3.4 *Se N_3 é uma variável aleatória degenerada em 1, isto é, $\mathbb{P}[N_3 = 1] = 1$, obtemos o modelo de sobrevivência destrutivo com fração de cura proposto por Rodrigues* et al. *(2010, 2011).*

Supomos agora que o número de células iniciadas, N_1, e número de células tumorais, N_3, seguem distribuições de Poisson ponderadas com parâmetros η_k e ϕ_k (Castillo & Pérez-Casany, 1998, 2005), $k = 1, 3$, respectivamente, com $f.m.p.$ da forma

$$p_k(n_k; \eta_k, \phi_k) = \mathbb{P}[N_k = n_k; \eta_k, \phi_k] = \frac{w(n_k; \phi_k) p^*(n_k; \eta_k)}{\mathbb{E}_{\eta_k}[w(N_k; \phi_k)]},\ n_k = 0, 1, 2, \ldots,\ k = 1, 3, \tag{3.8}$$

sendo que $w(\cdot\ ;\ \phi_k)$ é uma função peso não negativa com parâmetro $\phi_k > 0$, $p^*(\cdot\ ;\ \eta_k)$ é a $f.m.p.$ de uma distribuição de Poisson com parâmetro $\eta_k > 0$, e $\mathbb{E}_{\eta_k}[\cdot]$ indica que o valor esperado é tomada com relação à variável N_k seguindo uma distribuição de Poisson com média η_k. Denotamos a distribuição Poisson ponderada em (3.8) por $PP_{\eta_k}(w_k)$, o que representa a distribuição Poisson ponderada com parâmetro η_k e função peso $w_k(\cdot\ ;\ \phi_k)$. Este conceito foi proposto por Fisher (1934), mas foi Rao (1965) que estudou as distribuições ponderadas em um caminho unificado. Ele destacou que em muitas situações as observações registradas não podem ser consideradas como uma amostra aleatória da distribuição original, por muitas razões, tais como inobservabilidade de alguns eventos, danos causados às observações originais e a utilização de amostragem probabilística desigual. Muitas distribuições ponderadas são usadas na prática. Por exemplo, a distribuição ponderada com a função peso identidade é chamada de distribuição de tendenciosa pelo comprimento tem encontrado muitas aplicações importantes em biometria e meio ambiente (Zelen & Feinleib, 1969; Cnaan, 1985).

A $f.g.p.$ da variável aleatória Poisson ponderada N_k (Rodrigues *et al.*, 2009a) é dada por

$$\mathbb{A}_{N_k}(s) = \exp\{-\eta_k(1-s)\}\frac{\mathbb{E}_{\eta_k s}[w(N_k; \phi_k)]}{\mathbb{E}_{\eta_k}[w(N_k; \phi_k)]},\ \text{para } 0 \leq s \leq 1 \text{ e } k = 1, 3. \tag{3.9}$$

Levando em conta (3.8) e (3.9), a função de sobrevivência de longa duração é obtido do Teorema 3.1 por

$$S_{pop}(y) = \exp\Big\{-\eta_1 p\big(1 - S^*_{pop}(y)\big)\Big\}\frac{\mathbb{E}_{\eta_1\{1-p(1-S^*_{pop}(y))\}}[w(N_1;\phi_1)]}{\mathbb{E}_{\eta_1}[N_1;\phi_1]}, \tag{3.10}$$

sendo

$$S^*_{pop}(y) = 1 + p_{n_3}(0) - \exp\Big\{-\eta_3 S(y)\Big\}\frac{\mathbb{E}_{\eta_3 F(y)}[w(N_3;\phi_3)]}{\mathbb{E}_{\eta_3}[N_3;\phi_3]}, \tag{3.11}$$

em que $p_{n_3}(0) = w(0;\phi_3)e^{-\eta_3}/\mathbb{E}_{\eta_3}[w(N_3;\phi_3)]$. Pelo Teorema 3.2, a proporção de células malignas que morrem antes da indução do tumor $p_0^* = S^*_{pop}(+\infty) = p_{n_3}(0)$ e a fração de cura $p_0 = S_{pop}(+\infty) = \exp\Big\{-\eta_1 p(1-p_0^*)\Big\}^{\frac{\mathbb{E}_{\eta_1\{1-p(1-p_0^*)\}}[w(N_1;\phi_1)]}{\mathbb{E}_{\eta_1}[N_1;\phi_1]}}$.

Referimo-nos ao modelo em (3.10) como modelo híbrido Poisson ponderada-Poisson poderada, ou simplesmente, modelo HPPPP. A Figura 3.1 mostra um diagrama do modelo HPPPP.

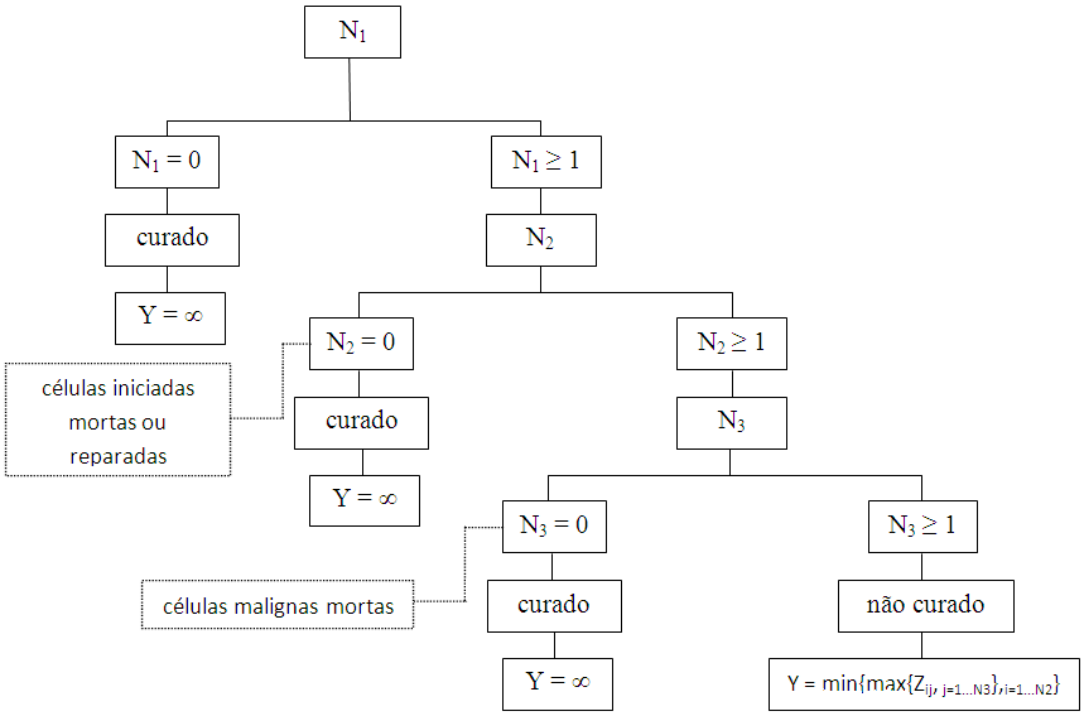

Figura 3.1: Representação do modelo proposto HPPPP.

3.2 Alguns modelos específicos

Nesta seção apresentamos alguns modelos específicos que surgem a partir da formulação geral apresentada na seção anterior.

3.2.1 Modelo híbrido Poisson ponderada exponencialmente-Poisson (HPPEP)

Quando a função peso do número de células iniciadas, N_1, é exponencial, isto é, $w(n_1;\phi_1) = \exp\big(n_1\phi_1\big)$, então N_1 segue uma distribuição Poisson ponderada exponencialmente com parâmetros η_1 e ϕ_1, e sua *f.m.p.* é dada por

$$p_1(n_1;\eta_1,\phi_1) = \frac{\eta_1^{n_1}\exp\big(\phi_1 n_1 - \eta_1 e^{\phi_1}\big)}{n_1!}, \quad n_1 = 0,1,2,\ldots, \tag{3.12}$$

para $\eta_1 > 0$ e $\phi_1 > 0$. Note que N_1 tem uma distribuição Poisson com parâmetro $\eta_1 e^{\phi_1}$.

Agora, supomos que o número de células tumorais, N_3, seguindo uma distribuição Poisson com parâmetro $\eta_3 > 0$. Assim, a partir de (3.10), a função de sobrevivência de longa duração do modelo HPPEP é dada por

$$S_{pop}(y) = \exp\{-\eta_1 p e^{\phi_1} e^{-\eta_3}(e^{\eta_3 F(y)} - 1)\}. \tag{3.13}$$

3.2.2 Modelo híbrido binomial negativa-Poisson (HBNP)

Seja o número de células iniciadas, N_1, com distribuição binomial negativa com parâmetros ϕ_1 e η_1 (Piegorsch, 1990; Saha & Paul, 2005), e sua *f.m.p.* é dada por

$$p_1(n_1;\eta_1,\phi_1) = \frac{\Gamma(\phi_1^{-1}+n_1)}{\Gamma(\phi_1^{-1})n_1!}\left(\frac{\phi_1\eta_1}{1+\phi_1\eta_1}\right)^{n_1}(1+\phi_1\eta_1)^{-\frac{1}{\phi_1}}, \quad n_1 = 0,1,2,\ldots. \tag{3.14}$$

para $\eta_1 > 0$, $\phi_1 \geq -1$ e $1+\phi_1\eta_1 > 0$. Ao compararmos esta forma com (3.8), percebemos imediatamente que (3.14) é uma distribuição Poisson ponderada com parâmetro $\phi_1\eta_1/(1+\phi_1\eta_1)$ e função peso $w(n_1;\phi_1) = \Gamma(\phi_1^{-1}+n_1)$. A média e a variância de N_1 são dadas por

$$\mathbb{E}[N_1] = \eta_1 \qquad \text{e} \qquad \mathbb{V}ar[N_1] = \eta_1(1+\phi_1\eta_1). \tag{3.15}$$

Também, a partir de (3.9), a *f.g.p.* é dada por

$$\mathbb{A}_{N_1}(s) = \{1+\phi_1\eta_1(1-s)\}^{-1/\phi_1}, \text{ para } 0 \leq s \leq 1. \tag{3.16}$$

Quando $\phi_1 = 1$ e $\phi_1 \longrightarrow 0$, obtemos as distribuições geométrica e Poisson, respectivamente. Em relação aos valores negativos de ϕ_1, Piegorsch (1990) destaca que se $\phi_1 = -1/\kappa$, sendo κ um inteiro positivo tal que $\kappa > \eta_1$, a distribuição binomial negativa com parâmetros η_1 e $-1/\kappa$ apresenta as mesmas probabilidades de uma distribuição binomial com parâmetros κ e η_1/κ. Ross & Preece (1985) provaram que mesmo se $\kappa = -1/\phi_1$ ($\phi_1 > 0$) não é um inteiro, a distribuição binomial negativa ainda apresenta valores positivos de $\mathbb{P}[N_1 = n_1], n_1 = 0, 1, \ldots, \kappa^*$, sendo que κ^* designa o maior inteiro menor do que κ. Portanto, ϕ_1 pode ser denominado de parâmetro de dispersão (Saha & Paul, 2005). Decorre de (3.15) que se $-1/\eta_1 < \phi_1 < 0$, que há subdispersão em relação à distribuição Poisson . Por outro lado, se $\phi_1 > 0$, há sobredispersão. O modelo binomial negativo, além de proporcionar bom ajuste em muitos casos práticos, também facilita as interpretações biológicas para os seus parâmetros (Tournoud & Ecochard, 2008). Em (3.15), η_1 é a média do número de células iniciadas, enquanto ϕ_1 fornece a variação inter-individual do número de células.

Seja o número de células tumorais, N_3, uma variável aleatória Poisson com parâmetro $\eta_3 > 0$, com *f.g.p.*

$$\mathbb{A}_{N_3}(s) = \exp\{-\eta_3(1-s)\}, \text{ para } 0 \leq s \leq 1. \tag{3.17}$$

Levando em conta (3.16) e (3.17), a função de sobrevivência de longa duração é dada por

$$S_{pop}(y) = \left\{1 + \phi_1\eta_1 p e^{-\eta_3}(e^{\eta_3 F(y)} - 1)\right\}^{-\frac{1}{\phi_1}}. \tag{3.18}$$

Quando $\phi_1 = 1$ em (3.18), obtemos o modelo híbrido geométrico-Poisson, denotado simplesmente por modelo HGP. Neste caso $S_{pop}(\cdot)$ torna-se

$$S_{pop}(y) = \left\{1 + \eta_1 p e^{-\eta_3}(e^{\eta_3 F(y)} - 1)\right\}^{-1}. \tag{3.19}$$

O modelo (3.18) é inidentificável (Li *et al.*, 2001), se os parâmetros η_1, p e η_3 são desconhecidos, isto é, existem $\boldsymbol{\vartheta} = (\phi_1, \eta_1, p, \eta_3, \boldsymbol{\gamma})$ e $\boldsymbol{\vartheta^*} = (\phi_1^*, \eta_1^*, p^*, \eta_3^*, \boldsymbol{\gamma^*})$, $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta^*}$, tais que $S_{pop}(y; \boldsymbol{\vartheta}) = S_{pop}(y; \boldsymbol{\vartheta^*})$, sendo $\boldsymbol{\gamma}$ o vetor de parâmetros da distribuição $F(\cdot)$.

3.2.3 Modelo híbrido COM-Poisson-Poisson (HCPP)

Supomos que o número de células iniciadas, N_1, segue uma distribuição COM-Poisson com parâmetros $\eta_1 > 0$ e $\phi_1 > 0$ (Shmueli *et al.*, 2005), com *f.m.p.*

$$p_1(n_1; \eta_1, \phi_1) = \frac{1}{Z(\eta_1, \phi_1)} \frac{\eta_1^{n_1}}{(n_1!)^{\phi_1}}, \quad n_1 = 0, 1, 2, \ldots, \tag{3.20}$$

sendo $Z(\eta_1, \phi_1) = \sum_{j=0}^{\infty} \eta_1^j/(j!)^{\phi_1}$. Em particular, quando $\phi_1 = 0$ e $0 < \eta_1 < 1$, a distribuição COM-Poisson torna-se igual a distribuição geométrica com parâmetro $1 - \eta_1$. A distribuição em (3.20), também, pode ser considerada como uma distribuição Poisson ponderada com função peso $w(n_1; \phi_1) = (n_1!)^{1-\phi_1}$. Portanto, usando (3.9), a *f.g.p.* é dada por

$$\mathbb{A}_{N_1}(s) = \frac{Z(\eta_1 s, \phi_1)}{Z(\eta_1, \phi_1)}. \tag{3.21}$$

Para os cálculos realizados na Seção 3.5, o truncamento da série $Z(\eta_1, \phi_1)$ é feito conforme descrito em Rodrigues *et al.* (2009a).

Agora suponhamos que o número de células tumorais, N_3, segue uma distribuição Poisson com parâmetro $\eta_3 > 0$. Assim, decorre de (3.10) que a função de sobrevivência de longa duração do modelo HCPP é dada por

$$S_{pop}(y) = \frac{Z(\eta_1\{1 - pe^{-\eta_3}(e^{\eta_3 F(y)-1})\})}{Z(\eta_1, \phi_1)}. \tag{3.22}$$

Na Tabela 3.1 apresentamos a função de sobrevivência de longa duração, a função densidade imprópria $f_{pop}(y) = -dS_{pop}(y)/dy$, a fração de cura e a proporção de células malignas que morrem antes da indução do tumor, correspondentes aos casos particulares apresentados nas Seções 3.2.1, 3.2.2 e 3.2.3.

Tabela 3.1: Função de sobrevivência de longa duração ($S_{pop}(y)$), função densidade ($f_{pop}(y)$), fração de cura (p_0), e proporção de células malignas que morrem antes da indução do tumor (p_0^*) para diferentes modelos.

Modelo híbrido	$S_{pop}(y)$	$f_{pop}(y)$	p_0	p_0^*
HPPEP	$\exp\{-\eta_1 e^{\phi_1} pe^{-\eta_3}(e^{\eta_3 F(y)} - 1)\}$	$\eta_1 e^{\phi_1} pe^{-\eta_3}\eta_3 f(y) e^{\eta_3 F(y)} S_{pop}(y)$	$\exp\{-\eta_1 e^{\phi_1} pe^{-\eta_3}(e^{\eta_3} - 1)\}$	$e^{-\eta_3}$
HBNP	$\{1 + \phi_1\eta_1 pe^{-\eta_3}(e^{\eta_3 F(y)} - 1)\}^{-1/\phi_1}$	$\frac{\eta_1 f(y) p\eta_3 e^{-\eta_3} e^{\eta_3 F(y)}}{1+\phi_1\eta_1 pe^{-\eta_3}(e^{\eta_3 F(y)}-1)} S_{pop}(y)$	$\{1 + \phi_1\eta_1 pe^{-\eta_3}(e^{\eta_3} - 1)\}^{-1\phi_1}$	$e^{-\eta_3}$
HCPP	$\frac{Z(\eta_1\{1-pe^{-\eta_3 F(y)}-1\},\phi_1)}{Z(\eta_1,\phi_1)}$	$\frac{p\eta_3 e^{-\eta_3} f(y) e^{\eta_3 F(y)}}{(1-pe^{-\eta_3}(e^{\eta_3 F(y)}-1))Z(\eta_1,\phi_1)} \sum_{j=1}^{\infty} \frac{j[\eta_1\{1-pe^{-\eta_3}(e^{\eta_3 F(y)}-1)\}]^j}{(j!)^{\phi_1}}$	$\frac{Z(\eta_1\{1-pe^{-\eta_3}-1\},\phi_1)}{Z(\eta_1,\phi_1)}$	$e^{-\eta_3}$

3.3 Inferência

Para a inferência adotamos os mesmos métodos clássico e bayesiano descritos na Seção 2.3. A função de verossimilhança do modelo HPPPP, as distribuições *a priori* dos parâmetros do modelo, assim como a distribuição *a posteriori* são descritas a seguir.

3.3.1 Função de verossimilhança

Para a formulação da função de verossimilhança consideram-se as seguintes notações. Seja $\boldsymbol{N} = (N_{1j}, N_{2j}, N_{3j})$ um vetor de variáveis aleatórias latentes, sendo que N_{1j} denota o número de células iniciadas no j-ésimo indivíduo, com distribuição $PP_{\eta_1}(w_1)$, N_{2j} denota o número de células malignas no j-ésimo indivíduo, em que N_{2j} dado N_{1j} segue uma distribuição binomial(N_{1j};p), e N_{3j} o número de células tumorais originadas de cada célula maligna no j-ésimo indivíduo, com distribuição $PP_{\eta_3}(w_3)$, $j = 1, 2, \ldots, m$.

Dado $N_{kj} = n_{kj}$, $k = 1, 2, 3$, sejam Z_{ihj} ($1 \leq i \leq n_{1j}$ e $1 \leq h \leq n_{3j}$), variáveis aleatórias contínuas (não-negativas) independentes com função distribuição $F(t_j;\boldsymbol{\gamma}) = 1 - S(t_j;\boldsymbol{\gamma})$ e independentes de N_{kj}, representando o tempo para a (i,h)-ésima célula maligna transformar-se em um tumor detectável no j-ésimo indivíduo e $\mathbb{P}[Z_{0hj} = \infty] = \mathbb{P}[Z_{i0j} = \infty] = 1$. Por sua vez, $\boldsymbol{\gamma}$ representa o vetor de parâmetros da distribuição do tempo de progressão (Z_{ij}). Seja Y_j como definido em (3.2) e sujeito a censura não informativa à direita. Assim, t_j é o tempo observado dado por $t_j = \min(Y_j, C_j)$, em que C_j é o tempo de censura, enquanto que δ_i é a variável indicadora de falha tal que $\delta_j = 1$ se $Y_j \leq C_j$, e $\delta_j = 0$, caso contrário, $j = 1, 2, \ldots, m$.

Além disso, os modelos HPPEP e HBNP das Seções 3.2.1 e 3.2.2 são inidentificáveis no sentido de Li *et al.* (2001). Para evitar este problema, propomos relacionar os parâmetros η_1, p e η_3 dos modelos HPPEP e HBNP com os vetores de covariáveis $\boldsymbol{x_j'} = (x_{j1}, \ldots, x_{jk_1})$, $\boldsymbol{\nu_j'} = (\nu_{j1}, \ldots, \nu_{jk_2})$ e $\boldsymbol{w_j'} = (w_{j1}, \ldots, w_{jk_3})$, respectivamente, sem elementos comuns. Adotemos as funções de ligação

$$\log(\eta_{1j}) = \boldsymbol{x_j'}\boldsymbol{\beta_1} \quad , \quad \log\left(\frac{p_j}{1-p_j}\right) = \boldsymbol{\nu_j'}\boldsymbol{\beta_2} \quad \text{e} \quad \log(\eta_{3j}) = \boldsymbol{w_j'}\boldsymbol{\beta_3}, \quad j = 1, \ldots, m, \tag{3.23}$$

em que $\boldsymbol{\beta_1'} = (\beta_{11}, \ldots, \beta_{1k_1})$, $\boldsymbol{\beta_2'} = (\beta_{21}, \ldots, \beta_{2k_2})$ e $\boldsymbol{\beta_3'} = (\beta_{31}, \ldots, \beta_{3k_3})$ são vetores com k_1, k_2 e k_3 coeficientes de regressão.

Os dados completos e observados são denotados por $\boldsymbol{D}_c = (m, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\delta}, \boldsymbol{N_1}, \boldsymbol{N_2}, \boldsymbol{N_3})$ e $\boldsymbol{D}_{obs} = (m, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\delta})$, respectivamente, sendo que $\boldsymbol{t'} = (t_1, \ldots, t_m)$, $\boldsymbol{\delta'} = (\delta_1, \ldots, \delta_m)$, $\boldsymbol{N_1'} = (N_{11}, \ldots, N_{1m})$, $\boldsymbol{N_2'} = (N_{21}, \ldots, N_{2m})$, $\boldsymbol{N_3'} = (N_{31}, \ldots, N_{3m})$, $\boldsymbol{X'} = (\boldsymbol{x_1'}, \boldsymbol{x_2'}, \ldots, \boldsymbol{x_m'})$, $\boldsymbol{V'} = (\boldsymbol{\nu_1'}, \boldsymbol{\nu_2'}, \ldots, \boldsymbol{\nu_m'})$ e $\boldsymbol{W'} = (\boldsymbol{w_1'}, \boldsymbol{w_2'}, \ldots, \boldsymbol{w_m'})$.

O próximo lema será fundamental para obter a função de verossimilhança do processo híbrido.

Teorema 3.3 *Sob o modelo com fração de cura híbrido e censura não-informativa, a densidade condicional de* (t_j, δ_j) *dado* $N_{1j} = n_{1j}$, $N_{2j} = n_{2j}$ *e* $N_{3j} = n_{3j}$, $j = 1, \ldots, m$ *é dada por*

$$f(t_j, \delta_j | n_{1j}, n_{2j}, n_{3j}) = \{1 - F^{n_{3j}}(t_j; \boldsymbol{\gamma})\}^{n_{2j} - \delta_j} \{n_{2j} n_{3j} f(t_j; \boldsymbol{\gamma}) F^{n_{3j}-1}(t_j; \boldsymbol{\gamma})\}^{\delta_j}. \tag{3.24}$$

Prova 3.3 *Consideramos duas situações:*

- *Observações censuradas* ($\delta_j = 0$) *:*

$$\begin{aligned}
\mathbb{P}[t_j = C_j, \delta_j = 0 | n_{1j}, n_{2j}, n_{3j}] &= \mathbb{P}[\delta_j = 0 | n_{1j}, n_{2j}, n_{3j}] \\
&= \mathbb{P}[Y_j > C_j | n_{1j}, n_{2j}, n_{3j}] \\
&= \mathbb{P}[\max\{Z_{1hj}\}_{h=1}^{n_{3j}} > t_j, \ldots, \max\{Z_{n_{2j}hj}\}_{h=1}^{n_{3j}} > t_j] \\
&= \{\mathbb{P}[\max\{Z_{1hj}\}_{h=1}^{n_{3j}} > t_j]\}^{n_{2j}} \\
&= \{1 - \mathbb{P}[Z_{11j} < t_j, \ldots, Z_{1n_{3j}<t_j}]\}^{n_{2j}} \\
&= \{1 - F^{n_{3j}}(t_j; \boldsymbol{\gamma})\}^{n_{2j}}.
\end{aligned}$$

- *Observações completas* ($\delta_j = 1$) *:*

$$\begin{aligned}
\mathbb{P}[t_j, \delta_j = 1 | n_{1j}, n_{2j}, n_{3j}] &= \mathbb{P}[t_j | Y_j < C_j, n_{1j}, n_{2j}, n_{3j}] \mathbb{P}[Y_j < C_j | n_{1j}, n_{2j}, n_{3j}] \\
&= \mathbb{P}[Y_j < C_j | n_{1j}, n_{2j}, n_{3j}] \times \\
&\quad \lim_{\Delta t_j \to 0} \frac{\mathbb{P}[t_j \leq Y_j \leq t_j + \Delta t_j | Y_j < C_j, n_{1j}, n_{2j}, n_{3j}]}{\Delta t_j} \\
&= \lim_{\Delta t_j \to 0} \frac{\mathbb{P}[t_j \leq Y_j \leq t_j + \Delta t_j | n_{1j}, n_{2j}, n_{3j}]}{\Delta t_j} \\
&= \frac{d}{dt_j} F_{Y_j}(t_j; \boldsymbol{\gamma}) = -\frac{d}{dt_j} \left\{1 - F^{n_{3j}}(t_j; \boldsymbol{\gamma})\right\}^{n_{2j}}.
\end{aligned}$$

Combinando as duas situações, obtemos o resultado enunciado.

Em seguida apresentamos a função verossimilhança dos parâmetros do modelo.

Teorema 3.4 *Supondo um processo híbrido com censura não-informativa, a função de verossimilhança é dada por*

$$\begin{aligned} L(\boldsymbol{\vartheta}; \boldsymbol{D}_c) &= \prod_{j=1}^{m} \{1 - F^{n_{3j}}(t_j; \boldsymbol{\gamma})\}^{n_{2j}-\delta_j} \{n_{2j} n_{3j} f(t_j; \boldsymbol{\gamma}) F^{n_{3j}-1}(t_j; \boldsymbol{\gamma})\}^{\delta_j} \times \\ &\quad \mathbb{P}[N_{1j} = n_{1j}] \mathbb{P}[N_{2j} = n_{2j} | N_{1j} = n_{1j}] \{\mathbb{P}[N_{3j} = n_{3j}]\}^{n_{2j}} \end{aligned} \tag{3.25}$$

em que $\boldsymbol{\vartheta}' = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \phi_1, \phi_2)$ *denota o vetor de parâmetros do modelo.*

Prova 3.4 *A função densidade conjunta é dada por*

$$\begin{aligned} f(\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{n_1}, \boldsymbol{n_2}, \boldsymbol{n_3}) &= \prod_{j=1}^{m} f(t_j, \delta_j, n_{1j}, n_{2j}, n_{3j}) \\ &= \prod_{j=1}^{m} f(t_j, \delta_j | n_{1j}, n_{2j}, n_{3j}) \mathbb{P}[N_{1j} = n_{1j}, N_{2j} = n_{2j}, N_{3j} = n_{3j}] \\ &= \prod_{j=1}^{m} f(t_j, \delta_j | n_{1j}, n_{2j}, n_{3j}) \mathbb{P}[N_{1j} = n_{1j}] \mathbb{P}[N_{2j} = n_{2j} | N_{1j} = n_{1j}] \times \\ &\quad \{\mathbb{P}[N_{3j} = n_{3j}]\}^{n_{2j}} \end{aligned}$$

em que $\boldsymbol{n_1'} = (n_{11}, \ldots, n_{1m})$, $\boldsymbol{n_2'} = (n_{21}, \ldots, n_{2m})$ *e* $\boldsymbol{n_3'} = (n_{31}, \ldots, n_{3m})$. *O resultado segue diretamente de (3.24).*

Note que a função de verossimilhança (3.25) depende de $\boldsymbol{N_1}$, $\boldsymbol{N_2}$ e $\boldsymbol{N_3}$, que são variáveis latentes.

Teorema 3.5 *Supondo um processo híbrido com censura não informativa, a função de verossimilhança marginal é dada por*

$$L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) = \prod_{j=1}^{m} \left\{ f_{pop}(t_j; \boldsymbol{\vartheta}) \right\}^{\delta_j} \left\{ S_{pop}(t_j; \boldsymbol{\vartheta}) \right\}^{1-\delta_j}, \tag{3.26}$$

sendo $f_{pop}(\cdot; \boldsymbol{\vartheta})$ *e* $S_{pop}(\cdot; \boldsymbol{\vartheta})$ *para os modelos da Seção 2.2 são dadas na Tabela 3.1.*

Prova 3.5 *A prova deste resultado é relativamente simples, apenas considerando as seguintes situações:*

- $\delta_j = 0$*:*

$$\begin{aligned}
L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) &= \prod_{j=1}^{m} \sum_{n_{1j}=0}^{\infty} \sum_{n_{2j}=0}^{n_{1j}} \Big\{1 - \sum_{n_{3j}=1}^{\infty} \{F(t_j;\boldsymbol{\gamma})\}^{n_{3j}} \mathbb{P}[N_{3j} = n_{3j}]\Big\}^{n_{2j}} \mathbb{P}[N_{2j} = n_{2j}|n_{1j}] \times \\
&\quad \mathbb{P}[N_{1j} = n_{1j}] \\
&= \prod_{j=1}^{m} \sum_{n_{1j}=0}^{\infty} \sum_{n_{2j}=0}^{n_{1j}} \big\{1 + \mathbb{P}[N_{3j} = 0] - \mathbb{A}_{N_{3j}}\big(F(t_j;\boldsymbol{\gamma})\big)\big\}^{n_{2j}} \mathbb{P}[N_{2j} = n_{2j}|n_{1j}] \times \\
&\quad \mathbb{P}[N_{1j} = n_{1j}] \\
&= \prod_{j=1}^{m} \sum_{n_{1j}=0}^{\infty} \big\{1 - p + p\big(1 + \mathbb{P}[N_{3j} = 0] - \mathbb{A}_{N_{3j}}(F(t_j;\boldsymbol{\gamma}))\big)\big\}^{n_{1j}} \mathbb{P}[N_{1j} = n_{1j}] \\
&= \prod_{j=1}^{m} \mathbb{A}_{N_{1j}}\Big(1 - p\big(1 - S^*_{pop}(t_j)\big)\Big) \\
&= \prod_{j=1}^{m} S_{pop}(t_j;\boldsymbol{\vartheta}).
\end{aligned}$$

- $\delta_j = 1$*:*

$$\begin{aligned}
L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) &= \prod_{j=1}^{m} \sum_{n_{1j}=0}^{\infty} \sum_{n_{2j}=0}^{n_{1j}} -\frac{d}{dt_j} \left(1 - \sum_{n_{3j}=1}^{\infty} \{F(t_j;\boldsymbol{\gamma})\}^{n_{3j}} \mathbb{P}[N_{3j} = n_{3j}]\right)^{n_{2j}} \times \\
&\quad \mathbb{P}[n_{2j}|n_{1j}]\mathbb{P}[N_{1j} = n_{1j}] \\
&= \prod_{j=1}^{m} -\frac{d}{dt_j} \sum_{n_{1j}=0}^{\infty} \sum_{n_{2j}=0}^{n_{1j}} \left(1 - \sum_{n_{3j}=1}^{\infty} \{F(t_j;\boldsymbol{\gamma})\}^{n_{3j}} \mathbb{P}[N_{3j} = n_{3j}]\right)^{n_{2j}} \times \\
&\quad \mathbb{P}[n_{2j}|n_{1j}]\mathbb{P}[N_{1j} = n_{1j}] \\
&= \prod_{j=1}^{m} -\frac{d}{dt_j} S_{pop}(t_j;\boldsymbol{\vartheta}) \\
&= \prod_{j=1}^{m} f_{pop}(t_j;\boldsymbol{\vartheta}).
\end{aligned}$$

As estimativas de máxima verossimilhança do parâmetro $\boldsymbol{\vartheta}' = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \phi_1)$ são obtidas de maneira análoga à Seção 2.3.1.

3.3.2 Distribuições a priori e a posteriori

Assumimos as seguintes distribuições *a priori* próprias e independentes para os parâmetros dos modelos: $\beta_{1j_1} \sim \mathcal{N}\big(0, \sigma^2_{1j_1}\big)$, $j_1 = 1, \ldots, k_1$, $\beta_{2j_2} \sim \mathcal{N}\big(0, \sigma^2_{2j_2}\big)$, $j_2 = 1, \ldots, k_2$, $\beta_{3j_3} \sim \mathcal{N}\big(0, \sigma^2_{3j_3}\big)$,

$j_3 = 1, \ldots, k_3$, $\gamma_1 \sim Gama\big(a_0, a_1\big)$ e $\gamma_2 \sim \mathcal{N}\big(0, \sigma^2_{\gamma_2}\big)$, enquanto que $\phi_1 \sim Gama\big(c_0, c_1\big)$ para os modelos HBNP e HCPP. Logo, as distribuições *a priori* e *a posteriori* de $\boldsymbol{\vartheta}' = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\beta_3'}, \boldsymbol{\gamma'}, \phi_1)$ são

$$\pi(\boldsymbol{\vartheta}) = \prod_{j_1=1}^{k_1} \pi(\beta_{1j_1}) \prod_{j_2=1}^{k_2} \pi(\beta_{2j_2}) \prod_{j_3=1}^{k_3} \pi(\beta_{3j_3})\pi(\gamma_1)\pi(\gamma_2)\pi(\phi_1)\pi(m) \quad \text{e} \tag{3.27}$$

$$\pi\big(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs}\big) \propto \pi(\boldsymbol{\vartheta})L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}), \tag{3.28}$$

respectivamente, sendo $L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})$ dada por (3.26).

Para a implementação do algoritmo de Gibbs na geração dos valores de $\boldsymbol{\vartheta}$, descrito na Seção 2.3.2, são necessárias as distribuições condicionais completas *a posteriori* de todos os parâmetros, dadas por

$$\pi(\boldsymbol{\beta_1}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_1}), \quad \pi(\boldsymbol{\beta_2}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_2}),$$

$$\pi(\boldsymbol{\beta_3}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_3}), \quad \pi(\gamma_1|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\gamma_1}),$$

$$\pi(\gamma_2|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\gamma_2) \text{ e } \pi(\phi_1|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\phi_1).$$

Novamente, estas distribuições condicionais não são avaliadas de forma fechada.

3.4 Estudo de simulação

Com os mesmos objetivos do estudo descrito na Seção 2.4 e de maneira análoga, realizamos um pequeno estudo de simulação. Neste estudo somente consideramos o modelo HGP da equação (3.19) (nosso modelo de trabalho na Seção 3.5) com distribuição Weibull para os tempos de progressão com parâmetros $\gamma_1 = 5$ e $\gamma_2 = 2$ e três covariáveis geradas a partir de uma distribuição normal com média 5 e variância 1, uma distribuição Bernoulli com parâmetro 0,5 e uma distribuição normal com média 0 e variância 1, as quais denotaremos por x, ν e w, respectivamente. Relacionamos os parâmetros η_1, p e η_3 do modelo HGP com as covariáveis x, ν e w, respectivamente. Adotamos as funções de ligação

$$\log(\eta_{1j}) = \beta_{1_1} x_j, \ \log\left(\frac{p_j}{1-p_j}\right) = \beta_{2_0} + \beta_{2_1}\nu_j \text{ e } \log(\eta_{3j}) = \beta_{3_1} w_j, \ j = 1, \ldots, m, \tag{3.29}$$

sendo $\beta_{1_1} = 1$, $\beta_{2_0} = -1$, $\beta_{2_1} = 1,5$ e $\beta_{3_1} = 0,5$. A fração de cura é $p_{0j} = \big\{1 + \eta_{1j} p_j e^{-\eta_{3j}} (e^{\eta_{3j}} - 1)\big\}^{-1}$ e a proporção de tempos censurados (φ_{cj}) é considerado como sendo igual a ($p_{0j} +$

0.1). O intervalo de variação de p_{0j} nas simulações varia entre 10% e 50%. Procedimento semelhante ao descrito na Seção 2.4 foi utilizado para a geração dos dados. A diferença entre o esquema utilizado e o apresentado na Seção 2.4 está no segundo item, que passa a ser o seguinte:

2 Se $u_j < p_{0j}$, então $y_j = \infty$; caso contrário,

$$y_j = \left(-\frac{\log\left\{1 - \frac{1}{\eta_{3j}} \log\{\frac{u_j^{-1}-1}{\eta_{1j} p_j e^{-\eta_{3j}}} + 1\}\right\}}{e^{\gamma_2}} \right)^{\frac{1}{\gamma_1}}.$$

Para cada tamanho amostral, mil simulações foram realizadas. As estimativas de máxima verossimilhança assim como as probabilidades de cobertura de cada parâmetro do modelo foram calculadas como o descrito na Seção 2.4. As simulações que não convergiram foram descartadas. Os resultados assim obtidos estão resumidos na Tabela 3.2. Podemos verificar que o REQM diminui com o aumento do tamanho da amostra e que as diferenças entre as estimativas médias e os valores verdadeiros, o denominado viés, são quase sempre menores que o REQM empírico, o que indica um bom desempenho dos estimadores de máxima verossimilhança. As PCs para alguns parâmetros são em torno de 0,89 e 0,93, sugerindo que o tamanho da amostra 400 não é ainda suficientemente grande para a normalidade assintótica dos MLEs, mas para os outros as PCs empíricas parecem a convergir para o nível nominal quando m aumenta. As conclusões deste estudo de simulação são limitados ao modelo HGP, mas nós acreditam que elas são semelhantes para outros modelos.

Tabela 3.2: Média, viés, REQM das estimativas de máxima verossimilhança e PC dos intervalos de confiança de 1000 repetições.

n	**parâmetro**	**média**	**viés**	**REQM**	**PC**
50	γ_1	5,37	0,36	0,25	0,93
	γ_2	2,35	0,35	0,29	0,88
	β_{1_1}	1,13	0,13	0,08	0,93
	β_{2_0}	-0,79	0,21	0,46	0,91
	β_{2_1}	2,42	0,92	0,58	0,84
	β_{3_1}	0,57	0,07	0,11	0,84
100	γ_1	5,14	0,14	0,21	0,94
	γ_2	2,25	0,25	0,27	0,90
	β_{1_1}	1,05	0,05	0,07	0,92
	β_{2_0}	-0,77	0,23	0,44	0,92
	β_{2_1}	2,37	0,87	0,53	0,85
	β_{3_1}	0,54	0,04	0,09	0,88
200	γ_1	5,04	0,04	0,13	0,94
	γ_2	2,16	0,16	0,17	0,90
	β_{1_1}	1,01	0,01	0,05	0,94
	β_{2_0}	-1,16	-0,16	0,32	0,95
	β_{2_1}	1,60	0,09	0,24	0,88
	β_{3_1}	0,51	0,01	0,07	0,93
400	γ_1	4,98	-0,02	0,05	0,94
	γ_2	2,13	0,13	0,06	0,91
	β_{1_1}	0,99	-0,01	0,02	0,92
	β_{2_0}	-1,12	-0,12	0,10	0,94
	β_{2_1}	1,51	0,03	0,08	0,89
	β_{3_1}	0,49	-0,01	0,03	0,93

3.5 Dados de câncer de melanoma

Nesta seção, apresentamos uma aplicação dos modelos descritos na Seção 3.2 em um conjunto de dados de melanoma maligno cutâneo. Os dados foram coletados em um estudo sobre melanoma com o objetivo de avaliar o desempenho da aplicação de uma dosagem alta de interferon alfa-2b como forma de prevenir recorrência de câncer. Os pacientes foram incluídos no estudo entre 1991 e 1995, tendo sido acompanhados até 1998. Uma descrição mais detalhada dos dados pode ser vista em Kirkwood *et al.* (2000) e Ibrahim *et al.* (2001) (dados E1690, disponível em http://merlot.stat.uconn.edu/∼mhchen/survbook/). Ressaltamos que esse conjunto de dados não enfatiza o processo da carcinôgenese descrito no capítulo 1, entretanto ele pode ser modelado certamente pelos modelos descritos na Seção 3.2, contanto que pensamos nesses dados como sendo gerado por um processo de três estágios. A amostra é composta por 417 pacientes sem valores faltantes, com 56% de observações censuradas. O tempo observado refere-se ao tempo em anos até a morte do paciente ou o tempo de censura (média=3,18 e desvio padrão = 1,69). Para fins ilustrativos, relacionamos os parâmetros η_1, p e η_3 em (3.23) com idade (x_1) (em anos; média =48,00 e desvio padrão=13,1), categoria do nódulo (x_2) (1, $m = 82$; 2, $m = 87$; 3, $m = 137$; 4, $m = 111$) e espessura do tumor (x_3) (em mm, média = 3,94 e desvio padrão = 3,20), respectivamente. A categoria do nódulo que vai de 1 até 4, respectivamente, é codificada a partir do número de linfonodos envolvidos na doença (0, 1, 2-3 e $\geq$4). Desta forma, a ligação entre os parâmetros e as covariáveis é dada por

$$\log(\eta_{1j}) = \beta_{1_1} x_{1j},\ \log\left(\frac{p_j}{1-p_j}\right) = \beta_{2_0} + \beta_{2_1} x_{2j} \text{ e } \log(\eta_{3j}) = \beta_{3_1} x_{3j},\ j = 1, \ldots, 417. \quad (3.30)$$

A Curva Kaplan-Meier estratificada por categoria do nódulo na Figura 3.2 estabiliza entre 0,2 a 0,7. Este comportamento sugere claramente que os modelos que ignoram a possibilidade de taxa de cura não serão adequados para analisar estes dados.

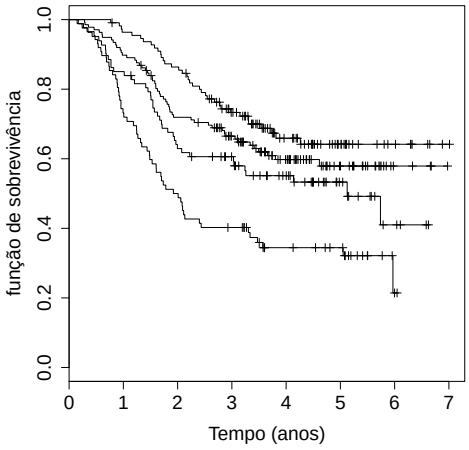

Figura 3.2: Curva Kaplan-Meier estratificada por categoria do nódulo (1 até 4, de cima para baixo).

Ajustamos os modelos da Tabela 3.1. Um caso particular do modelo HBNP, também, foi ajustado aos dados, a saber, o modelo híbrido geométrico-Poisson (HGP) ($\phi_1 = 1$). A Tabela 3.3 apresenta os valores do máximo da log-verossimilhança, $\max \log L(\cdot)$, e os valores das estatísticas AIC e BIC para os modelos ajustados. De acordo com os critérios AIC e BIC, o modelo HGP se destaca como o melhor. Ressaltamos que o modelo HCPP, mesmo com os parâmetros η_1, p e η_3 ligados a todas as covariáveis, não produz um ajuste tão bom quanto este. O gráfico QQ do resíduo dos quantis normalizado (Dunn & Smyth, 1996; Rigby & Stasinopoulos, 2005) na Figura 3.3 sugere que o modelo HGP é aceitável. Cada ponto na Figura 3.3 corresponde à mediana de cinco conjuntos de resíduos ordenados. Tendo em conta os critérios da Tabela 3.3 e o gráfico QQ na Figura 3.3, selecionamos o modelo HGP como nosso modelo de trabalho. Estimativas de máxima verossimilhança dos coeficientes e seus desvios padrão e intervalos de confiança assintóticos (IC) de 95% estão na Tabela 3.4.

Tabela 3.3: $Max \log L(\cdot)$ e as estatísticas AIC e BIC para os quatros modelos ajustados.

	Modelo			
Critério	HPPEP	HBNP	HCPP	HGP
$Max \log L(\cdot)$	-516,99	-509,07	-517,45	-509,48
AIC	1047,98	1032,14	1048,89	1030,96
BIC	1076,21	1060,37	1077,12	1055,16

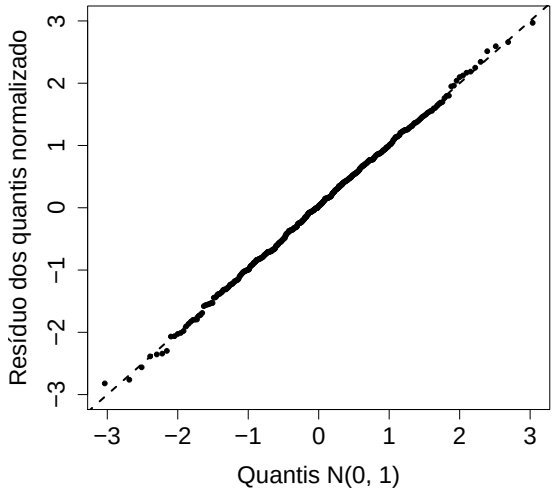

Figura 3.3: Gráfico QQ do resíduo dos quantis normalizado com a reta identidade para o modelo HGP (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados).

Tabela 3.4: Estimativas de máxima verossimilhança dos parâmetros do modelo HGP, seus desvios padrão e seus intervalos de confiança assintóticos de 95% (IC 95%).

Parâmetro	**Estimativa**	**desvio padrão**	**IC** 95%
γ_1	1,63	0,11	(1,42 ; 1,84)
γ_2	-1,29	0,16	(-1,62 ; -0,98)
β_{1_1}	0,02	0,01	(0,01 ; 0,034)
β_{2_0}	-2,35	0,43	(-3,19 ; -1,50)
β_{2_1}	0,98	0,26	(0,47 ; 1,48)
β_{3_1}	0,08	0,02	(0,03 ; 0,13)

Usando as estimativas da Tabela 3.4, e a função de ligação logarítmica em (4.22), obtemos as estimativas pontuais e intervalos de confiança assintótico de 95% (ICs) (os erros padrão necessários à construção dos ICs foram estimados aplicando o método delta (Sen & Singer, 1993)) para a proporção de células malignas que morrem antes da indução do tumor (p_0^*) na Tabela 3.5 para pacientes com espessura do tumor 0,7, 3,1 e 10.0 mm. Essas espessuras correspondem aos quantis de 5%, 50% e 95%. Notamos que os ICs são amplos. A Figura 3.4 mostra a função de sobrevivência para pacientes com idades 29, 47 e 70 anos e espessura do tumor 3,94 mm. As idades correspondem aos quantis de 5%, 50% e 95% e a espessura do tumor a média. A probabilidade de sobrevivência diminui mais rapidamente para os pacientes mais velhos. Na Figura 3.4 (a), a função de sobrevivência não desça abaixo de 0,4.

Tabela 3.5: Estimativas de máxima verossimilhança, desvios padrão e intervalos de confiança assintóticos de 95% (IC 95%) para a proporção de células malignas que morrem antes da indução do tumor para pacientes com espessura do tumor 0,7, 3,1 e 10.0 mm.

Espessura do tumor (mm)	$\widehat{p_0^*}$	**desvio padrão**	**IC** 95%
0,7	0,35	0,04	(0,27 ; 0,43)
3,1	0,28	0,13	(0,03 ; 0,53)
10,0	0,11	0,18	(0,00 ; 0,45)

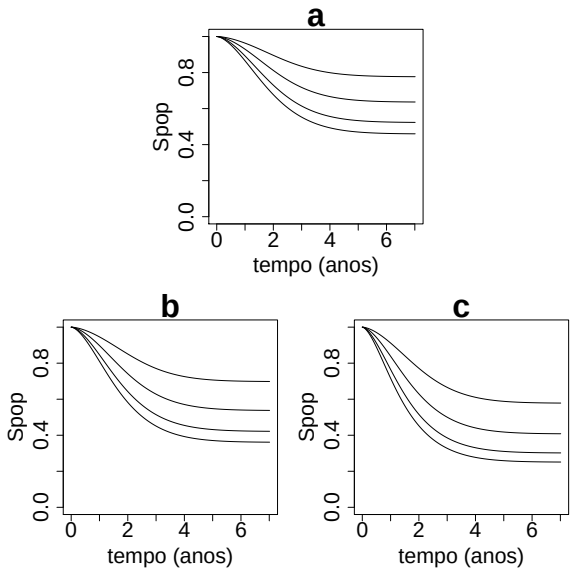

Figura 3.4: Função de sobrevivência sob o modelo HGP estratificado por categoria do nódulo (1 até 4, de cima para baixo) para pacientes com idades (a) 29, (b) 47, e (c) 70 anos, e espessura do tumor 3,94 mm.

Agora, voltamos a nossa atenção para o papel das covariáveis sobre a fração de cura p_0 (ver Tabela 3.1). O sinal positivo do coeficiente β_{1_1} significa que aumenta número médio de células iniciadas com o aumento da idade do paciente, de modo que a fração de cura diminui. Visto que $\beta_{2_1} > 0$ e $\beta_{3_1} > 0$ na Tabela 3.4, os valores mais elevados da categoria nódulo e espessura do tumor implicam em estimativas menores da fração de cura. A Figura 3.5 mostra o efeito combinado destas covariáveis sobre a fração de cura. As linhas correm quase paralelamente. A redução na fração de cura entre a idade mínima e máxima é de 35,2%, 47,7%, 55,0% e 58,4% para categoria do nódulo de 1 até 4 e espessura do tumor 3,94 mm, respectivamente.

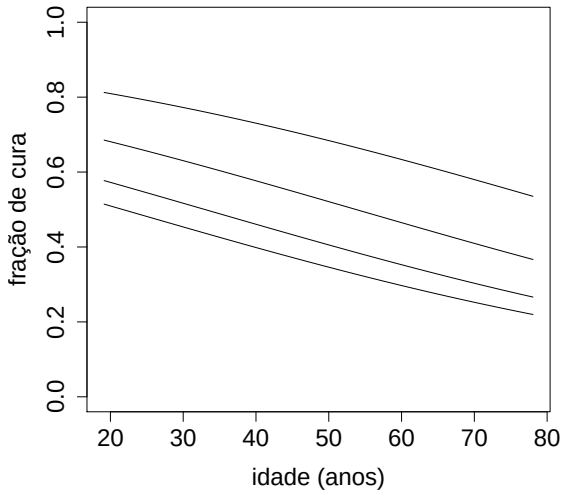

Figura 3.5: Fração de cura para o modelo HGP *versus* idade estratificada por categoria do nódulo (1 até 4, de cima para baixo) e espessura do tumor 3,94 mm.

Também obtemos os ajustes para os quatro modelos da Tabela 3.3 através da inferência bayesiana. Utilizamos distribuições *a priori* independentes e não informativas, sendo $\beta_{1_1} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{2_0} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{2_1} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{3_1} \sim \mathcal{N}\big(0, 10^3\big)$, $\gamma_1 \sim Gama\big(1, 0, 01\big)$ e $\gamma_2 \sim \mathcal{N}\big(0, 10^3\big)$, enquanto que $\phi \sim Gama\big(1, 0, 01\big)$ para os modelos HBNP e HCPP. Geramos duas cadeias paralelas de tamanho 35000 para cada parâmetro. Descartamos as primeiras 5000 e as restantes selecionadas de 10 em 10, resultando numa amostra de tamanho 3000. A convergência das cadeias foi monitorada empregando o método de Cowles & Carlin (1996).

Na Tabela 3.6, foi aplicado os critérios de seleção de modelos definidos na Seção 2.3.3 para os quatro modelos ajustados: HPPEP, HBNP, HCPP e HGP. O modelo HGP se destacar como o melhor. Portanto, selecionamos o modelo HGP como nosso modelo de trabalho. A Tabela 3.7 apresenta as médias *a posteriori*, os desvios padrão e os intervalos de credibilidade para os parâmetros do modelo HGP, incluindo o fator de redução de escala potencial estimado $\widehat{R}$ (Gelman & Rubin, 1992), que para todos os parâmetros está próximo de um, indicando a convergência das cadeias, enquanto a Figura 3.6 apresenta as densidades marginais a posteriori aproximadas para cada parâmetro. A Tabela 3.8 apresenta as médias *a posteriori*, os desvios padrão e os intervalos

de credibilidade para a proporção de células malignas que morrem antes da indução do tumor (p_0^*) para pacientes com espessura do tumor 0,7, 3,1 e 10.0 mm. Na Figura 3.7, mostramos a densidade *a posteriori* marginal aproximada de p_0^*.

Para avaliar a robustez do modelo com relação à escolha dos hiperparâmetros das distribuições *a priori*, um pequeno estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros não apresentam muita diferença e não alteram os resultados apresentados na Tabela 3.6.

Tabela 3.6: Critérios DIC, EAIC, EBIC e B para os quatro modelos ajustados.

	Modelo			
Critério	HPPEP	HBNP	HCPP	HGP
DIC	1035,58	1033,31	1036,01	1031,00
EAIC	1042,71	1040,06	1042,97	1037,17
EBIC	1070,94	1068,29	1071,20	1061,37
B	-515,63	-514,10	-515,88	-513,98

Tabela 3.7: Médias *a posteriori*, os desvios padrão e os intervalos de credibilidade 95% (ICred 95%) para os parâmetros do modelo HGP e o fator de redução de escala potencial estimado $\widehat{R}$.

Parâmetro	**Média**	**desvio padrão**	**ICred 95%**	$\widehat{R}$
γ_1	1,64	0,11	(1,43 ; 1,84)	1,002
γ_2	-1,35	0,17	(-1,68 ; -1,04)	1,003
β_{1_1}	0,02	0,01	(0,01 ; 0,03)	1,001
β_{2_0}	-2,36	0,46	(-3,27 ; -1,44)	1,003
β_{2_1}	1,09	0,32	(0,62 ; 1,89)	1,002
β_{3_1}	0,06	0,03	(0,00 ; 0,11)	1,001

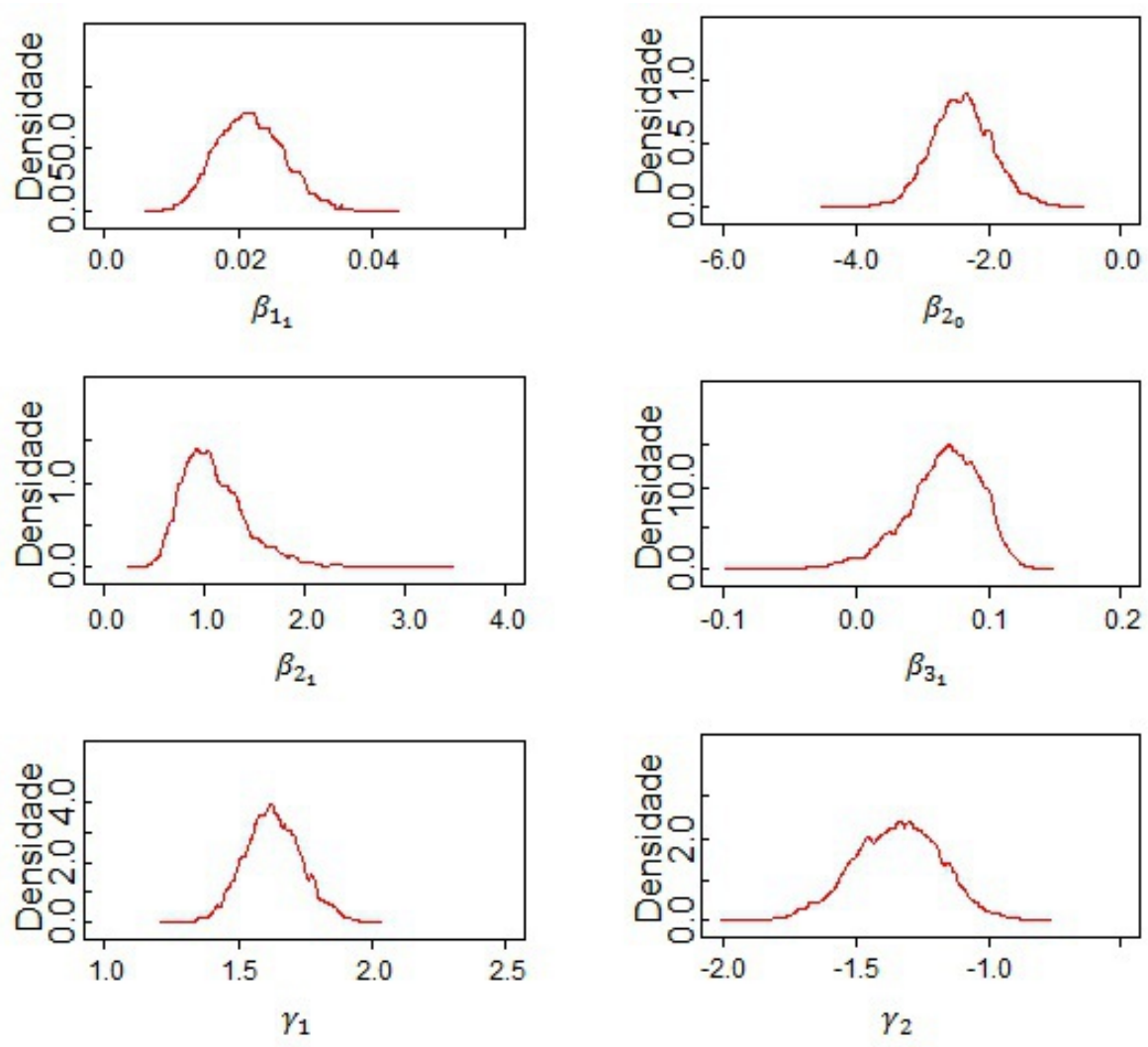

Figura 3.6: Densidades *a posteriori* aproximadas dos parâmetros.

Tabela 3.8: Médias *a posteriori*, desvios padrão e intervalos de credibilidade 95% (ICred 95%) para a proporção de células malignas que morrem antes da indução do tumor (p_0^*) para pacientes com espessura do tumor 0,7, 3,1 e 10.0 mm, sob o modelo HGP.

Categoria do nódulo	**Média**	**desvio padrão**	**ICred 95%**
0,7	0,35	0,01	(0,34 ; 0,37)
3,1	0,29	0,03	(0,24 ; 0,36)
10,0	0,16	0,09	(0,05 ; 0,37)

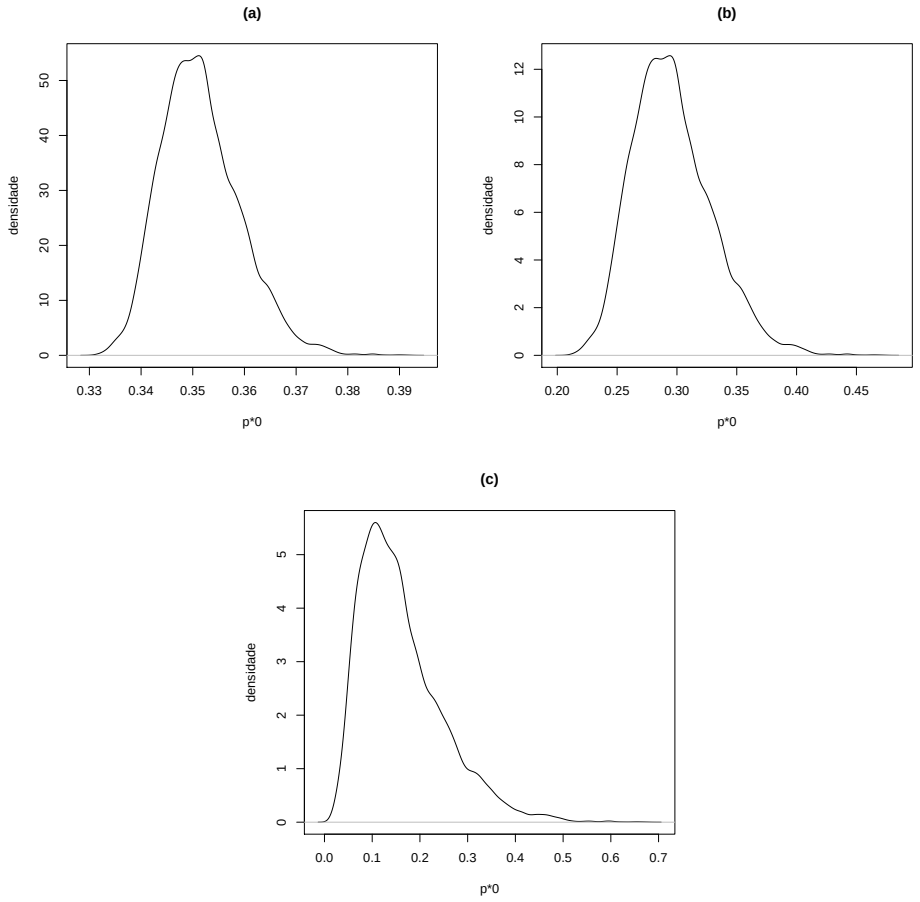

Figura 3.7: Densidade *a posteriori* marginal aproximada para a proporção de células malignas que morrem antes da indução do tumor (p_0^*) sob o modelo HGP para pacientes com espessura do tumor (a) 0,7, (b) 3,1 e (c) 10.0 mm.

A Tabela 3.9 contém os resumos *a posteriori* para a fração de cura estratificada por categoria do nódulo (1 a 4) e espessura do tumor 3,94 mm para pacientes com idades de 29, 47 e 70 anos de 3000 amostras obtidas com o modelo HGP. Esta tabela nos permite avaliar o efeito combinado das covariáveis sobre a fração de cura, notando que ambos agem para reduzir a fração de cura. As diferenças entre as idades 29 e 70 anos dos pacientes são significativas ao nível de 5% para todas as categorias do nódulo.

Tabela 3.9: Médias *a posteriori*, os desvios padrão e os intervalos de credibilidade 95% (ICred 95%) para a fração de cura (p_0) estratifica por categoria do nódulo (1-4) e espessura do tumor 3,94 mm, sob o modelo HGP.

Idade	**Categoria do nódulo**	**Média**	**desvio padrão**	**ICred 95%**
29	1	0,77	0,04	(0,69 ; 0,84)
	2	0,62	0,04	(0,54 ; 0,70)
	3	0,52	0,03	(0,46 ; 0,59)
	4	0,47	0,03	(0,42 ; 0,52)
47	1	0,69	0,04	(0,61 ; 0,69)
	2	0,53	0,04	(0,46 ; 0,53)
	3	0,42	0,03	(0,36 ; 0,49)
	4	0,37	0,04	(0,31 ; 0,44)
70	1	0,58	0,05	(0,47 ; 0,68)
	2	0,40	0,04	(0,33 ; 0,48)
	3	0,31	0,23	(0,23 ; 0,39)
	4	0,27	0,05	(0,18 ; 0,36)

Os resultados obtidos pela estimação de máxima verossimilhança e pela inferência bayesiana são próximos e implicam nas mesmas conclusões a respeito do modelo a ser escolhido e das covariáveis a serem consideradas.

3.6 Comentários finais

Neste capítulo propusemos um modelo de sobrevivência com fração de cura híbrido para acomodar características dos estágios não-observáveis da carcinogênese (iniciação, promoção e progressão) na presença de causas competitivas latentes. Nós assumimos uma distribuição Poisson ponderada para o número de causas competitivas dos estágios de iniciação e progressão, e um modelo Weibull para os tempos de vida, obtendo o modelo geral HPPPP. O modelo HPPPP incorpora características do estágio de progressão, bem como a proporção de células iniciadas que foram "promovidas"a malignas e a proporção de células malignas que morrem antes da indução

do tumor. A vantagem deste modelo é que podemos estimar a taxa de iniciação η_1 e a taxa de proliferação de células de tumor η_3, que não é possível na maioria dos modelos de fração de cura comumente utilizados. Os dois processos de estimação apresentaram resultados similares. A relevância prática e a aplicabilidade do modelo foram demonstradas em um conjunto de dados reais de pacientes com câncer de melanoma.

Apesar de apenas a distribuição Weibull ter sido considerada como a nossa distribuição do tempo de vida, em princípio, a metodologia não se restringe a ela e outras distribuições mais complexas podem ser consideradas. A questão inferencial pode tornar-se muito mais complexa neste caso.

Capítulo 4

Modelo com fração de cura híbrido correlacionado

No capítulo anterior foi proposto um modelo de sobrevivência com fração de cura utilizando um sistema híbrido para acomodar as características dos estágios não observáveis do processo da carcinogênese (iniciação, promoção e progressão). Este modelo supera a limitação que cada célula iniciada torna-se maligna com probabilidade um, mas assume que as células em um tecido podem dar origem a um tumor independentemente umas das outras, ou seja, elas são biologicamente independentes durante o processo da carcinogênese. Entretanto, Haynatzki *et al.* (2000) discutiram que a suposição de independência biológica pode não ser verdadeira quando a dinâmica da população de células de um tecido normal é considerada. Similarmente, há indícios de que as células pré-malignas (iniciadas) e malignas em um tecido influenciam no desenvolvimento umas das outras. Além disso, a interação entre as células saudáveis e pré-malignas no tecido devem ser levadas em consideração. Portanto, é desejável construir modelos estatísticos que possam incorporar adequadamente a dependência, e isso é que proporcionou a motivação para o presente capítulo.

Consequentemente, a finalidade principal deste capítulo é propor um modelo de sobrevivência com fração de cura que estende os modelos formulados no capítulo anterior, incorporando uma estrutura de dependência entre as células iniciadas ao tornarem-se cancerosas de maneira análoga ao capítulo 2.

O capítulo está organizado da seguinte forma. Na Seção 4.1 apresentamos a formulação do modelo. Alguns modelos específicos são apresentados na Seção 4.2. Na Seção 4.3 discutimos o processo inferencial. Na Seção 4.4 apresentamos os resultados de um pequeno estudo de simulação. Na Seção 4.5 um conjunto de dados de câncer melanoma ilustra a utilidade do modelo proposto. Comentários finais são apresentados na Seção 4.6.

4.1 Formulação do modelo

Na construção de nosso modelo geral, utilizamos as mesmas suposições básicas descritas na Seção 3.1, com exceção das suposições (iii) e (iv) que passarão a ser as seguintes:

(iii) Uma lesão pré-cancerosa não reparada permanece dormente enquanto ela prossegue com a fase de promoção do desenvolvimento do tumor. Todas as lesões estão sujeitas a promoção dependentemente umas das outras.

(iv) Uma vez que a célula maligna ou clonogênica surge como resultado da promoção da célula iniciada, começa o estágio de progressão produzindo uma colônia de descendentes (células tumorais) , chamada de clone ou tumor. Tratamos o número de células malignas resultantes do estágio de promoção como uma variável aleatória N_2. O tempo que uma célula maligna leva para se transformar em um tumor detectável é considerado como uma variável aleatória com função de distribuição $F(y) = 1 - S(y)$, sendo $S(y)$ função de sobrevivivência. Todas células malignas estão sujeitas a progressão dependentemente uma das outras.

Com base nessas novas suposições, o modelo proposto é desenvolvido de maneira análoga à Seção 3.1 (vide página 34). Entretanto, como o nosso objetivo é inserir uma estrutura de correlação entre as células (vide página 9), supomos agora que o número de células iniciadas, N_1, e o número de células tumorais, N_3, seguem distribuições série de potências generalizada inflada (SPGI) (vide Seção 2.1) com parâmetros $\theta_k \in (0, s)$ (s pode ser ∞) e $\rho_k = \rho \in [0, 1)$ (estamos supondo que correlação entre as células permanece a mesma independência do estágio) , $k = 1, 3$, respectivamente.

Levando em conta (2.2), (2.4) e o Teorema 3.1, a função de sobrevivência de longa duração é dada por

$$S_{pop}(y) = \frac{g\left(\frac{\theta_1(1-\rho)\left[1-p\left(1-S^*_{pop}(y)\right)\right]}{1-\rho\left[1-p\left(1-S^*_{pop}(y)\right)\right]}\right)}{g(\theta_1)}, \tag{4.1}$$

em que

$$S^*_{pop} = 1 + p_{n_3}(0) - \frac{g\left(\frac{\theta_3(1-\rho)F(y)}{1-\rho F(y)}\right)}{g(\theta_3)} \tag{4.2}$$

e $p_{n_3}(0) = \frac{1}{g(\theta_3)}$. A fração de cura é determinada por $p_0 = \lim_{y\to\infty} S_{pop}(y)$. Assim, a partir de (4.1),

$$p_0 = \frac{g\left(\frac{\theta_1(1-\rho)\left[1-p\left(1-p_{n_3}(0)\right)\right]}{1-\rho\left[1-p\left(1-p_{n_3}(0)\right)\right]}\right)}{g(\theta_1)}.$$

A proporção de células malignas que morrem antes da indução do tumor é determinada por $p_0^* = \mathbb{P}[N_3 = 0] = p_{n_3}(0) = \frac{1}{g(\theta_3)}$.

Referimo-nos ao modelo em (4.1) como modelo híbrido correlacionado série de potências generalizada inflada , ou simplesmente, modelo HCSPGI.

Observação 4.1 *Se N_3 é uma variável aleatória degenerada em 1, isto é, $\mathbb{P}[N_3 = 1] = 1$, obtemos o modelo com fração de cura destrutivo correlacionado proposto no Capítulo 2.*

4.2 Alguns modelos específicos

Nesta seção, apresentamos alguns modelos específicos que surgem a partir da formulação geral apresentada na seção anterior. As funções a_{n_k}, $g(\theta_k)$ e o parâmetro θ_k são dados na Tabela 2.1, acrescentando o índice k.

4.2.1 Modelo híbrido correlacionado Poisson-Poisson (HCPP)

Quando as funções $a_{n_k} = \frac{1}{n_{k1}!n_{k2}!\cdots}$, $g(\theta_k) = \exp\{\theta_k\}$ e o parâmetro $\theta_k = \eta_k$, $k = 1, 3$, dizemos que o número de células iniciadas N_1 e número de células tumorais N_3 têm distribuição Poisson inflada com parâmetros $\eta_k > 0$ e $\rho \in [0, 1)$, $k = 1, 3$, respectivamente, e sua $f.m.p.$ é da forma

$$\mathbb{P}_{Poi}[N_k = n_k] = \begin{cases} e^{-\eta_k} & , \quad n_k = 0 \\ e^{-\eta_k}\sum_{i=1}^{n_k}\binom{n_k-1}{i-1}\frac{\left[\eta_k(1-\rho)\right]^i\rho^{n_k-1}}{i!} & , \quad n_k = 1, 2, \ldots. \end{cases} \tag{4.3}$$

A *f.g.p.* é representada pela seguinte equação:

$$\mathbb{A}_{N_k}(z) = \exp\left\{-\frac{\eta_k(1-z)}{1-z\rho}\right\}, \quad \text{para} \quad 0 \leq z \leq 1 \quad \text{e} \quad k = 1, 3. \tag{4.4}$$

Assim, a partir de (4.1), a função de sobrevivência de longa duração do modelo HCPP é dada por

$$S_{pop}(y) = \exp\left\{-\frac{\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)}{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]}\right\}. \tag{4.5}$$

4.2.2 Modelo híbrido correlacionado binomial-Poisson (HCBP)

Quando $a_{n_1} = \binom{m_1}{m_1-n_{11}-n_{12}-\cdots, n_{11}, n_{12}, \ldots}$, $g(\theta_1) = (1+\theta_1)^{m_1}$ e $\theta_1 = \frac{\pi_1}{1-\pi_1}$, então o número de células iniciadas N_1 segue um distribuição binomial inflada com parâmetros $\pi_1 \in (0,1)$, $\rho \in [0,1)$ e $m_1 \in \mathbb{Z}^+$, e sua *f.m.p.* é da forma

$$\mathbb{P}_{Bin}[N_1 = n_1] = \begin{cases} (1-\pi_1)^{m_1} & , \quad n_1 = 0 \\ \sum_{i=1}^{\min(n_1, m_1)} \binom{m_1}{i}\binom{n_1-1}{i-1}\left[\pi_1(1-\rho)\right]^i (1-\pi_1)^{m_1-i}\rho^{n_1-i} & , \quad n_1 = 1, 2, \ldots. \end{cases} \tag{4.6}$$

A *f.g.p.* é representada pela seguinte equação:

$$\mathbb{A}_{N_1}(z) = \left[1 - \frac{\pi_1(1-z)}{1-z\rho}\right]^{m_1}, \quad \text{para} \quad 0 \leq z \leq 1. \tag{4.7}$$

Agora, supomos que o número de células tumorais, N_3, segue uma distribuição Poisson inflada com parâmetros $\eta_3 > 0$ e $\rho \in [0,1)$. Assim, a partir de (4.1), a função de sobrevivência de longa duração do modelo HCBP é dada por

$$S_{pop}(y) = \left[1 - \frac{\pi_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)}{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]}\right]^{m_1}. \tag{4.8}$$

4.2.3 Modelo híbrido correlacionado binomial negativa-Poisson (HCBNP)

Quando $a_{n_1} = \frac{\Gamma\left(\phi_1^{-1} + \sum_{i=1}^{\infty} n_{1i}\right)}{\Gamma(\phi_1^{-1})\left[\sum_{i=1}^{\infty} n_{1i}\right]!}$, $g(\theta_1) = (1-\theta_1)^{-\phi_1^{-1}}$, e parâmetro $\theta_1 = \frac{\phi_1\eta_1}{1+\phi_1\eta_1}$, dizemos que o número de células iniciadas N_1 segue uma distribuição binomial negativa inflada com parâmetros

$\eta_1 > 0$, $\rho \in [0,1)$, $\phi_1 \geq -1$ e $\phi_1\eta_1 > 0$, e sua *f.m.p.* é da forma

$$\mathbb{P}_{NB}[N_1 = n_1] = \begin{cases} (1+\phi_1\eta_1)^{-\phi_1^{-1}} & , \quad n_1 = 0 \\ (1+\phi_1\eta_1)^{-\phi_1^{-1}} \sum_{i=1}^{n_1} \binom{n_1-1}{i-1} \frac{\Gamma(\phi_1^{-1}+i)}{\Gamma(\phi_1^{-1})i!} \left[\frac{\phi_1\eta_1(1-\rho)}{1+\phi_1\eta_1}\right]^i \rho^{n_1-i} & , \quad n_1 = 1,2,\ldots. \end{cases} \tag{4.9}$$

A *f.g.p.* é representada pela seguinte equação:

$$\mathbb{A}_{N_1}(z) = \left[\frac{1-z\rho}{1+\phi_1\eta_1(1-z)-z\rho}\right]^{\phi_1^{-1}}, \quad \text{para} \quad 0 \leq z \leq 1. \tag{4.10}$$

Agora, suponhamos que o número de células tumorais, N_3, siga uma distribuição Poisson inflada com parâmetros $\eta_3 > 0$ e $\rho \in [0,1)$. Assim, a partir de (4.1), a função de sobrevivência de longa duração do modelo HCBNP é dada por

$$S_{pop}(y) = \left[\frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}{1+\phi_1\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}\right]^{\frac{1}{\phi_1}}. \tag{4.11}$$

Quando $\phi_1 = 1$, obtemos a distribuição geométrica inflada com parâmetro $\theta_1 = \frac{1}{1+\eta_1} \in (0,1)$ em (4.9), e $S_{pop}(\cdot)$ em (4.11) reduz-se a

$$S_{pop}(y) = \frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}{1+\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{(1-\rho F(y))}\right\}-e^{-\eta_3}\right)\right]}, \tag{4.12}$$

dando origem ao modelo híbrido correlacionado geométrico-Poisson, ou simplesmente, modelo HCGP.

4.2.4 Modelo híbrido correlacionado série logarítmica-Poisson (HCSLP)

Quando $a_{n_1} = \frac{(-1+n_{11}+n_{12}+\cdots)!}{n_{11}!n_{12}!\cdots}$, $g(\theta_1) = -\log(1-\theta_1)$ e $\theta_1 = 1-\pi_1$, então o número de células iniciadas N_1 segue uma distribuição série logarítmica inflada com parâmetros $\pi_1 \in (0,1)$ e $\rho \in [0,1)$, e sua *f.m.p.* é da forma

$$\mathbb{P}_{LS}[N_1 = n_1] = (-\log(\pi_1))^{-1} \sum_{i=1}^{n_1} \binom{n_1-1}{i-1} \frac{\left[(1-\pi_1)(1-\rho)\right]^i \rho^{n_1-i}}{i}, \quad n_1 = 1,2,\ldots. \tag{4.13}$$

Em sua forma original, esta distribuição exclui o valor zero. Consequentemente, não pode ser usada para modelar o número de células iniciadas (no sentido de incluir a longa duração). Para

os fins deste capítulo, consideramos uma série logarítima inflada modificada, cuja $f.m.p.$ pode ser escrita como

$$\mathbb{P}_{LS}[N_1 = n_1] = (-\log(\pi_1))^{-1} \sum_{i=1}^{n_1+1} \binom{n_1}{i-1} \frac{\left[(1-\pi_1)(1-\rho)\right]^i \rho^{n_1+1-i}}{i}, \quad n_1 = 0, 1, 2, \ldots. \tag{4.14}$$

A $f.g.p.$ é representada pela seguinte equação:

$$\mathbb{A}_{N_1}(z) = \frac{(-\log(\pi_1))^{-1}}{z} \log\left[\frac{1-\rho z}{1-z(1-\pi_1(1-\rho))}\right]. \tag{4.15}$$

Agora, supomos que o número de células tumorais, N_3, segue uma distribuição Poisson inflada com parâmetros $\eta_3 > 0$ and $\rho \in [0, 1)$. Assim, a partir de (4.1), a função de sobrevivência de longa duração do modelo HCSLP é dada por

$$\begin{aligned} S_{pop}(y) &= \frac{(-\log(\pi_1))^{-1}}{1 - p\left(\exp\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\} - e^{-\eta_3}\right)} \times \\ &\quad \log\left[\frac{1 - \rho\left[1 - p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]}{1 - (1-\pi_1(1-\rho))\left(1 - \rho\left[1 - p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]\right)}\right]. \end{aligned} \tag{4.16}$$

Na Tabela 4.1, apresentamos a função de sobrevivência de longa duração, a função densidade imprópria $f_{pop}(y) = -dS_{pop}(y)/dy$, a fração de cura e a propoção de células malignas que morrem antes da indução do tumor correspondentes aos casos particulares apresentados nas Seções 4.2.1, 4.2.2, 4.2.3 e 4.2.4.

Tabela 4.1: Função de sobrevivência de longa duração ($S_{pop}(y)$), função densidade ($f_{pop}(y)$), fração de cura (p_0), e propoção de células malignas que morrem antes da indução do tumor (p_0^*) para diferentes modelos.

Modelo	$S_{pop}(y)$	$f_{pop}(y)$	p_0	p_0^*
HCPP	$\exp\left\{-\frac{\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)}{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}\right\}$	$\left(\frac{\eta_1 p\eta_3 f(y)(1-\rho)^2 e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{(1-\rho F(y))^2\left(1-\rho\left[1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right]\right)^2}\right) S_{pop}(y)$	$\exp\left\{-\frac{\eta_1 p\left(1-e^{-\eta_3}\right)}{1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]}\right\}$	$e^{-\eta_3}$
HCBP	$\left[1-\frac{\pi_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)}{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}\right]^{m_1}$	$\left(-\frac{\eta_1\eta_3 p\phi_1 f(y)(1-\rho)^2 e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{(1-\rho F(y))^2\left(1-\rho\left[1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right]\right)\left(1-\rho+p(\rho-\eta)\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right)}\right) S_{pop}(y)$	$\left[1-\frac{\pi_1 p\left(1-e^{-\eta_3}\right)}{1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]}\right]^{m_1}$	$e^{-\eta_3}$
HCBNP	$\left[\frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}{1+\phi_1\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}\right]^{\frac{1}{\phi_1}}$	$\left(\frac{\eta_1 p\eta_3 f(y)(1-\rho)^2 e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{\left(1-\rho\left[1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right]\right)(1-\rho F(y))^2\left(1-\rho+(\phi_1\eta_1 p+\rho p)\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right)}\right) S_{pop}(y)$	$\left[\frac{1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]}{1+\phi_1\eta_1 p\left(1-e^{-\eta_3}\right)-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]}\right]^{\frac{1}{\phi_1}}$	$e^{-\eta_3}$
HCSLP	$\frac{(-\log(\pi_1))^{-1}}{1-p\left(\exp\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\}-e^{-\eta_3}\right)}\log\left[\frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}{1-(1-\pi_1(1-\rho))\left(1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]\right)}\right]$	$\frac{(\log(\eta_1))^{-1}\left(1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right)^{-1}}{\left(1-\rho\left[1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right]\right)}\left(\frac{(\rho+\eta_1(1-\rho)-1)p\eta_3 f(y)(1-\rho)e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{(1-\rho F(y))^2\left(p(1-\eta_1(1-\rho))\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)+\eta_1(1-\rho)\right)}\right)$ $+\left(1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right)^{-1}\left(\frac{p\eta_3(1-\rho)f(y)e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{(1-\rho F(y))^2}\right) S_{pop}(y)$	$\frac{(-\log(\pi_1))^{-1}}{1-p\left(1-e^{-\eta_3}\right)}\log\left[\frac{1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]}{1-(1-\pi_1(1-\rho))\left(1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]\right)}\right]$	$e^{-\eta_3}$

4.3 Inferência

4.3.1 Função de verossimilhança

Seja $\boldsymbol{N} = (N_{1j}, N_{2j}, N_{3j})$ um vetor de variáveis aleatórias latentes, sendo que N_{1j} denota o número de células iniciadas no j-ésimo indivíduo, com distribuição $PP_{\eta_1}(w_1)$, N_{2j} o número de células malignas no j-ésimo indivíduo, em que N_{2j} dado N_{1j} segue um distribuição binomial(N_{1j};p), e N_{3j} o número de células tumorais originadas de cada célula maligna no j-ésimo indivíduo, com distribuição $PP_{\eta_3}(w_3)$, $j = 1, 2, \ldots, m$.

Dado $N_{kj} = n_{kj}$, $k = 1, 2, 3$, sejam Z_{ihj}, $1 \leq i \leq n_{1j}$ e $1 \leq h \leq n_{3j}$, variáveis aleatórias contínuas (não-negativas) independentes com função distribuição $F(t_j;\boldsymbol{\gamma}) = 1 - S(t_j;\boldsymbol{\gamma})$, $\boldsymbol{\gamma}$ representa o vetor de parâmetros da distribuição, e independentes de N_{kj}, representando o tempo para a (i,h)-ésima célula maligna transformar-se em um tumor detectável no j-ésimo indivíduo e $\mathbb{P}[Z_{0hj} = \infty] = \mathbb{P}[Z_{i0j} = \infty] = 1$. Seja Y_j como definido em (3.2) e sujeito a censura à direita. Assim, t_j é o tempo observado dado por $t_j = \min\{Y_j, C_j\}$, com C_j é o tempo de censura, enquanto que δ_i é a variável indicadora de censura tal que $\delta_j = 1$ se $Y_j \leq C_j$, e $\delta_j = 0$, caso contrário, $j = 1, 2, \ldots, m$.

Além disso, para $\rho = 0$ os modelos HCPP, HCBP e HCBNP das Seções 4.2.1, 4.2.2 e 4.2.3 são inidentificáveis no sentido de Li *et al.* (2001). Para evitar este problema, propomos relacionar os parâmetros η_1 (ou π_1), p e η_3 (ou π_3) dos modelos HCPP, HCBP e HCBNP com os vetores de covariáveis $\boldsymbol{x_j'} = (x_{j1}, \ldots, x_{jk_1})$, $\boldsymbol{\nu_j'} = (\nu_{j1}, \ldots, \nu_{jk_2})$ e $\boldsymbol{w_j'} = (w_{j1}, \ldots, w_{jk_3})$, respectivamente, sem elementos comuns. Adotemos as funções de ligação

$$\log(\eta_{1j}) = \boldsymbol{x_j'}\boldsymbol{\beta_1} \left(\text{ou } \log\left(\frac{\pi_{1j}}{1-\pi_{1j}}\right) = \boldsymbol{x_j'}\boldsymbol{\beta_1}\right), \log\left(\frac{p_j}{1-p_j}\right) = \boldsymbol{\nu_j'}\boldsymbol{\beta_2} \quad \text{e} \tag{4.17}$$

$$\log(\eta_{3j}) = \boldsymbol{w_j'}\boldsymbol{\beta_3} \left(ou \ \log\left(\frac{\pi_{3j}}{1-\pi_{3j}}\right) = \boldsymbol{w_j'}\boldsymbol{\beta_3}\right), \quad j = 1, \ldots, m,$$

sendo $\boldsymbol{\beta_1'} = (\beta_{11}, \ldots, \beta_{1k_1})$, $\boldsymbol{\beta_2'} = (\beta_{21}, \ldots, \beta_{2k_2})$ e $\boldsymbol{\beta_3'} = (\beta_{31}, \ldots, \beta_{3k_3})$ vetores com k_1, k_2 e k_3 coeficientes de regressão.

Os dados completos e observados são denotados por $\boldsymbol{D}_c = (m, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\delta}, \boldsymbol{N_1}, \boldsymbol{N_2}, \boldsymbol{N_3})$ e $\boldsymbol{D}_{obs} = (m, \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\delta})$, respectivamente, sendo que $\boldsymbol{t'} = (t_1, \ldots, t_m)$, $\boldsymbol{\delta'} = (\delta_1, \ldots, \delta_m)$, $\boldsymbol{N_1'} = (N_{11}, \ldots, N_{1m})$, $\boldsymbol{N_2'} = (N_{21}, \ldots, N_{2m})$, $\boldsymbol{N_3'} = (N_{31}, \ldots, N_{3m})$, $\boldsymbol{X'} = (\boldsymbol{x_1'}, \boldsymbol{x_2'}, \ldots, \boldsymbol{x_m'})$,

$\boldsymbol{V}' = (\boldsymbol{\nu_1'}, \boldsymbol{\nu_2'}, \ldots, \boldsymbol{\nu_m'})$ e $\boldsymbol{W}' = (\boldsymbol{w_1'}, \boldsymbol{w_2'}, \ldots, \boldsymbol{w_m'})$.

Para m pares de tempos e indicadores de censura $(t_1, \delta_1), \ldots, (t_m, \delta_m)$ e, de acordo com o Teorema 3.5, a função de verossimilhança marginal é dada por

$$L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) = \prod_{j=1}^{m} \big\{ f_{pop}(t_j; \boldsymbol{\gamma}) \big\}^{\delta_j} \big\{ S_{pop}(t_j; \boldsymbol{\gamma}) \big\}^{1-\delta_j}, \tag{4.18}$$

sendo que $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\beta_3'}, \boldsymbol{\gamma'}, \rho, \phi_1)$ denota o vetor de parâmetros do modelo, enquanto que $f_{pop}(\cdot; \boldsymbol{\vartheta})$ e $S_{pop}(\cdot; \boldsymbol{\vartheta})$ para os modelos da Seção 4.1 são dadas na Tabela 4.1.

As estimativas de máxima verossimilhança do parâmetro $\boldsymbol{\vartheta}$ são obtidas de maneira análoga à Seção 2.3.1.

4.3.2 Distribuições a priori e a posteriori

As distribuições *a priori* dos parâmetros foram escolhidas de acordo com o espaço paramétrico de cada um deles, o que significa que $\beta_{1j_1} \sim \mathcal{N}\big(0, \sigma^2_{1j_1}\big)$, $j_1 = 1, \ldots, k_1$, $\beta_{2j_2} \sim \mathcal{N}\big(0, \sigma^2_{2j_2}\big)$, $j_2 = 1, \ldots, k_2$, $\beta_{3j_3} \sim \mathcal{N}\big(0, \sigma^2_{3j_3}\big)$, $j_3 = 1, \ldots, k_3$, $\gamma_1 \sim Gama\big(a_0, a_1\big)$, $\gamma_2 \sim \mathcal{N}\big(0, \sigma^2_{\gamma_2}\big)$ e $\rho \sim Beta\big(b_0, b_1\big)$, enquanto que $\phi_1 \sim Gama\big(c_0, c_1\big)$ para o modelo HCBNP.

As distribuições *a priori* e *a posteriori* de $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\beta_3'}, \boldsymbol{\gamma'}, \rho, \phi_1)$ são

$$\pi(\boldsymbol{\vartheta}) = \prod_{j_1=1}^{k_1} \pi(\beta_{1j_1}) \prod_{j_2=1}^{k_2} \pi(\beta_{2j_2}) \prod_{j_3=1}^{k_3} \pi(\beta_{3j_3}) \pi(\gamma_1) \pi(\gamma_2) \pi(\rho) \pi(\phi_1), \tag{4.19}$$

$$\pi\big(\boldsymbol{\vartheta} | \boldsymbol{D}_{obs}\big) \propto \pi(\boldsymbol{\vartheta}) L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}), \tag{4.20}$$

respectivamente, sendo $L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})$ dada por (4.18).

As distribuições condicionais completas *a posteriori* são dadas por

$$\pi(\boldsymbol{\beta_1}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\boldsymbol{\beta_1}) \ , \ \pi(\boldsymbol{\beta_2}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\boldsymbol{\beta_2}),$$

$$\pi(\boldsymbol{\beta_3}|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\boldsymbol{\beta_3}) \ , \ \pi(\gamma_1|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\boldsymbol{\gamma_1}),$$

$$\pi(\gamma_2|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\gamma_2) \ , \ \pi(\rho|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\boldsymbol{\rho}),$$

$$\pi(\phi_1|\cdot) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) \pi(\phi_1).$$

Novamente, estas distribuições condicionais não são avaliadas de forma fechada.

4.4 Estudo de simulação

Com o intuito de verificar algumas propriedades frequentistas dos estimadores de máxima verossimilhança, realizamos um pequeno estudo de simulação. Neste estudo somente consideramos o modelo HCBNP da equação (4.11) (nosso modelo de trabalho na Seção 4.5). No processo de simulação, fixamos $\rho = 0,7$, ou seja, uma alta associação entre as células iniciadas e $\phi_1 = 5$. Adotamos distribuição Weibull para os tempos de progressão com parâmetros $\gamma_1 = 2$ e $\gamma_2 = -3$. Assumimos para cada indivíduo três covariáveis, x, ν e w, sendo que estas foram consideradas fixas, mas tiveram seus valores gerados a partir de uma distribuição normal com média 5 e variância 1, uma distribuição Bernoulli com parâmetro 0,5 e uma distribuição normal com média 0 e variância 1, respectivamente. Relacionamos os parâmetros η_1, p e η_3 do modelo HCBNP para covariáveis x, ν e w, respectivamente. Adotamos as funções de ligação

$$\log(\eta_{1j}) = \beta_{1_1} x_j,\ \log\left(\frac{p_j}{1-p_j}\right) = \beta_{2_0}\nu_j + \beta_{2_1}(1-\nu_j) \text{ e } \log(\eta_{3j}) = \beta_{3_1} w_j,\ j = 1,\ldots,m, \quad (4.21)$$

sendo $\beta_{1_1} = 1$, $\beta_{2_0} = -1$, $\beta_{2_1} = 1,5$ e $\beta_{3_1} = 0,5$. A fração de cura é

$$p_{0j} = \left[\frac{1-\rho\Big(1-p_j(1-e^{-\eta_{3j}})\Big)}{1+\phi_1\eta_{1j}p_j(1-e^{-\eta_{3j}})-\rho\Big(1-p_j(1-e^{-\eta_{3j}})\Big)}\right]^{\frac{1}{\phi_1}}$$

e a proporção de tempos censurados (φ_{cj}) é considerada como sendo igual a $(p_{0j} + 0.1)$. O intervalo de variação de p_{0j} nas simulações varia entre 15% e 60%. Procedimento semelhante ao descrito na Seção 2.4 foi utilizado para a geração dos dados. A diferença entre o esquema utilizado e o apresentado na Seção 2.4 está no segundo item, que passa a ser o seguinte:

2 Se $u_j < p_{0j}$, então $y_j = \infty$; caso contrário,

$$y_j = \exp\left\{\frac{\log\left(-\log\left(\frac{(1-\rho)\left(\eta_{3j}-\log\left(\frac{-e^{\eta_{3j}}u_j^{\phi_1}+u_j^{\phi_1}\phi_1\eta_{1j}p_j+\rho e^{\eta_{3j}}u_j^{\phi_1}+u_j^{\phi_1}\rho p_j+e^{\eta_{3j}}(1-\rho)-\rho p_j}{p_j(u_j^{\phi_1}\phi_1\eta_{1j}+\rho u_j^{\phi_1}-\rho)}\right)\right)}{\eta_{3j}+\rho\log\left(\frac{-e^{\eta_{3j}}u_j^{\phi_1}+u_j^{\phi_1}\phi_1\eta_{1j}p_j+\rho e^{\eta_{3j}}u_j^{\phi_1}+u_j^{\phi_1}\rho p_j+e^{\eta_{3j}}(1-\rho)-\rho p_j}{p_j(u_j^{\phi_1}\phi_1\eta_{1j}+\rho u_j^{\phi_1}-\rho)}\right)-\rho\eta_{3j}}\right)\right)-\gamma_2}{\gamma_1}\right\}.$$

Para cada tamanho amostral, mil simulações foram realizadas. As estimativas de máxima verossimilhança assim como as probabilidades de cobertura de cada parâmetro do modelo foram calculadas como o descrito na Seção 2.4. As simulações que não convergiram foram descartadas. Os resultados assim obtidos estão resumidos na Tabela 4.2. Podemos verificar que o REQM diminui com o aumento do tamanho da amostra e que as diferenças entre as estimativas médias e os valores verdadeiros, o denominado viés, são quase sempre menores que o REQM empírico, o que indica um bom desempenho dos estimadores de máxima verossimilhança. As PCs para alguns parâmetros são em torno de 0,79 e 0,94, sugerindo que o tamanho da amostra 400 não é ainda suficientemente grande para a normalidade assintótica dos MLEs, mas para os outros as PCs empíricas parecem a convergir para o nível nominal quando m aumenta. As conclusões deste estudo de simulação são limitados ao modelo HCBNP, mas nós acreditam que elas são semelhantes para outros modelos.

Tabela 4.2: Média, viés, REQM das estimativas de máxima verossimilhança e PC dos intervalos de confiança de 1000 repetições.

n	**parâmetro**	**média**	**viés**	**REQM**	**PC**
50	γ_1	2,312	0,312	0,299	0,81
	γ_2	-2,486	0,514	0,530	0,82
	ρ	0,522	-0,178	0,700	0,65
	ϕ	5,817	0,817	2,385	0,87
	β_{1_1}	1,155	0,155	0,363	0,92
	β_{2_0}	-2,068	-1,068	0,642	0,91
	β_{2_1}	2,462	0,962	1,393	0,90
	β_{3_1}	0,280	-0,220	0,683	0,63
100	γ_1	2,132	0,132	0,289	0,88
	γ_2	-2,634	0,366	0,450	0,88
	ρ	0,580	-0,120	0,204	0,72
	ϕ	5,562	0,562	2,357	0,92
	β_{1_1}	1,146	0,146	0,259	0,94
	β_{2_0}	-1,943	-0,943	0,451	0,94
	β_{2_1}	2,252	0,752	1,249	0,95
	β_{3_1}	0,366	-0,134	0,247	0,72
200	γ_1	2,061	0,061	0,282	0,92
	γ_2	-2,876	0,124	0,340	0,90
	ρ	0,602	-0,098	0,158	0,83
	ϕ	5,250	0,250	1,525	0,94
	β_{1_1}	1,103	0,103	0,165	0,94
	β_{2_0}	-1,849	-0,849	0,343	0,94
	β_{2_1}	2,168	0,668	1,555	0,96
	β_{3_1}	0,439	-0,061	0,144	0,77
400	γ_1	2,002	0,002	0,171	0,94
	γ_2	-3,123	-0,123	0,743	0,94
	ρ	0,720	0,020	0,115	0,88
	ϕ	4,918	-0,082	1,110	0,95
	β_{1_1}	1,091	0,091	0,089	0,94
	β_{2_0}	-1,454	-0,454	0,310	0,94
	β_{2_1}	2,098	0,598	0,324	0,96
	β_{3_1}	0,476	-0,024	0,078	0,79

4.5 Dados de câncer de melanoma

A metodologia apresentada neste capítulo será aplicada ao conjunto de dados da Seção 2.5. Tendo em mente a questão da identificabilidade mencionada anteriormente na Seção 4.2, nos modelos HCPP, HCBP e HCBNP, ligamos os parâmetros η_1 (ou π_1), p e η_3 em (4.17) para estado de úlcera (x_1) (ausente, $m = 115$; presente, $m = 90$), espessura do tumor (x_2) (em mm, média = 2,92 e desvio padrão = 2,96) e sexo (x_3) (feminino, $m = 126$, masculino, $m = 79$), respectivamente. Desta forma, a ligação entre os parâmetros e as covariáveis é expressa através de

$$\log(\eta_{1j}) = \beta_{1_{pres}} x_{1j} + \beta_{1_{aus}}(1 - x_{1j}) \ \left(ou \log\left(\frac{\pi_{1j}}{1-\pi_{1j}}\right) = \beta_{1_{pres}} x_{1j} + \beta_{1_{aus}}(1 - x_{1j}) \right), \quad (4.22)$$

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_{2_0} + \beta_{2_1} x_{2j} \quad \text{e} \quad \log(\eta_{3j}) = \beta_{3_{mas}} x_{3j} + \beta_{3_{fem}}(1 - x_{3j}),\ j = 1, \ldots, 205.$$

Ajustamos os modelos da Tabela 4.1 e o modelo HCGP. Para o modelo DCB fixei o parâmetro $m_1 = 15$. A Tabela 4.3 apresenta os valores de máximo da log-verossimilhança, $\max \log L(\cdot)$, e os valores das estatísticas AIC e BIC para os modelos ajustados. De acordo com os critérios $\max \log L(\cdot)$, AIC e BIC, os modelos HCBNP e HCPP se destacam como os melhores. O gráfico QQ do resíduo dos quantis normalizado (Dunn & Smyth, 1996; Rigby & Stasinopoulos, 2005) na Figura 4.1 sugere que o modelo HCBNP é aceitável. Cada ponto na Figura 4.1 corresponde à mediana de cinco conjuntos de resíduos ordenados. Tendo em conta os critérios da Tabela 4.3 e o gráfico QQ na Figura 4.1, selecionamos o modelo HCBNP como nosso modelo de trabalho. Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo HCBNP, seus desvios padrão e seus intervalos de confiança 95% são apresentados na Tabela 4.4. A estimativa do parâmetro correlação (ρ) é 0,77, e como mencionado anteriormente na Seção 4.1, isso indica uma alta associação entre as células.

Tabela 4.3: $Max \log L(\cdot)$ e as estatísticas AIC e BIC para os cinco modelos ajustados.

	Modelo				
Critério	HCPP	HCBP	HCBNP	HCGP	HCSLP
$\max \log L(\cdot)$	-198,44	-209,31	-197,19	-199,90	-198,89
AIC	414,89	438,63	414,38	417,81	415,78
BIC	444,81	471,86	447,62	447,71	445,69

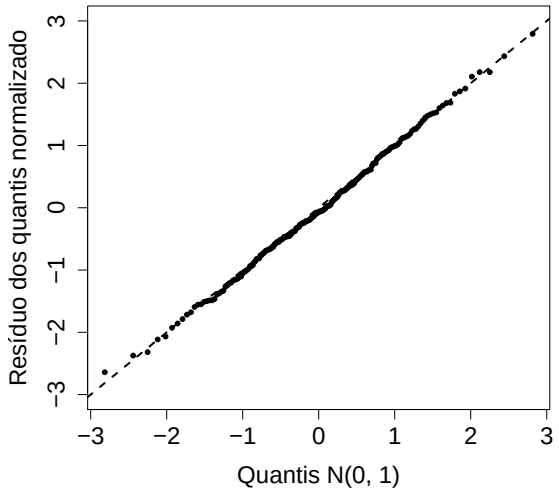

Figura 4.1: Gráfico QQ do resíduo dos quantis normalizado com a reta identidade para o modelo HCBNP (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados).

Tabela 4.4: Estimativas de máxima verossimilhança dos parâmetros do modelo HCBNP, seus desvios padrão e seus intervalos de confiança assintóticos de 95% (IC 95%).

Parâmetro	**Estimativa**	**desvio padrão**	**IC** 95%
γ_1	2,47	0,92	(0,67 ; 4,27)
γ_2	-4,03	2,29	(-8,52 ; 0,46)
ρ_1	0,77	0,09	(0.59 ; 0.95)
ϕ	5,23	3,33	(0,66 ; 9,80)
$\beta_{1_{pres}}$	2,15	2,32	(-2,40 ; 6,70)
$\beta_{1_{aus}}$	3,88	2,68	(-1,37 ; 9,13)
β_{2_0}	-4,89	1,65	(-8,12 ; -1,66)
β_{2_1}	1,12	0,40	(0,34 ; 1,90)
$\beta_{3_{mas}}$	-1,52	0,78	(-3,05 ; 0,01)
$\beta_{3_{fem}}$	0,49	0,89	(-1,25 ; 2,23)

Usando as estimativas da Tabela 4.4, a função de ligação logarítmica em (4.17), e $\boldsymbol{I_0(\widehat{\beta_1})}$ extraída de (2.37), obtemos as estimativas pontuais e intervalos de confiança assintótico de 95% (ICs) para a proporção de células malignas que morrem antes da indução do tumor (p_0^*) na Tabela 4.5. Notamos que os ICs são amplos. A Figura 4.2 mostra a função de sobrevivência para pacientes com espessura do tumor igual a 0,32, 1,94 e 8,32 mm, que correspondem aos quantis de 5%, 50% e 95%, respectivamente, e segundo o sexo. A probabilidade de sobrevivência diminui mais rapidamente para os pacientes do sexo feminino com tumores mais espessos. Na Figura 4.2 (f), a função de sobrevivência não desça abaixo de 0,35.

Tabela 4.5: Estimativas de máxima verossimilhança, desvios padrão e intervalos de confiança assintóticos de 95% (IC 95%) para a proporção de células malignas que morrem antes da indução do tumor estratificada pelo sexo.

Sexo	$\widehat{p_0^*}$	**desvio padrão**	**IC** 95%
masculino	0,80	0,14	(0,53 ; 1,00)
feminino	0,20	0,28	(0,00 ; 0,75)

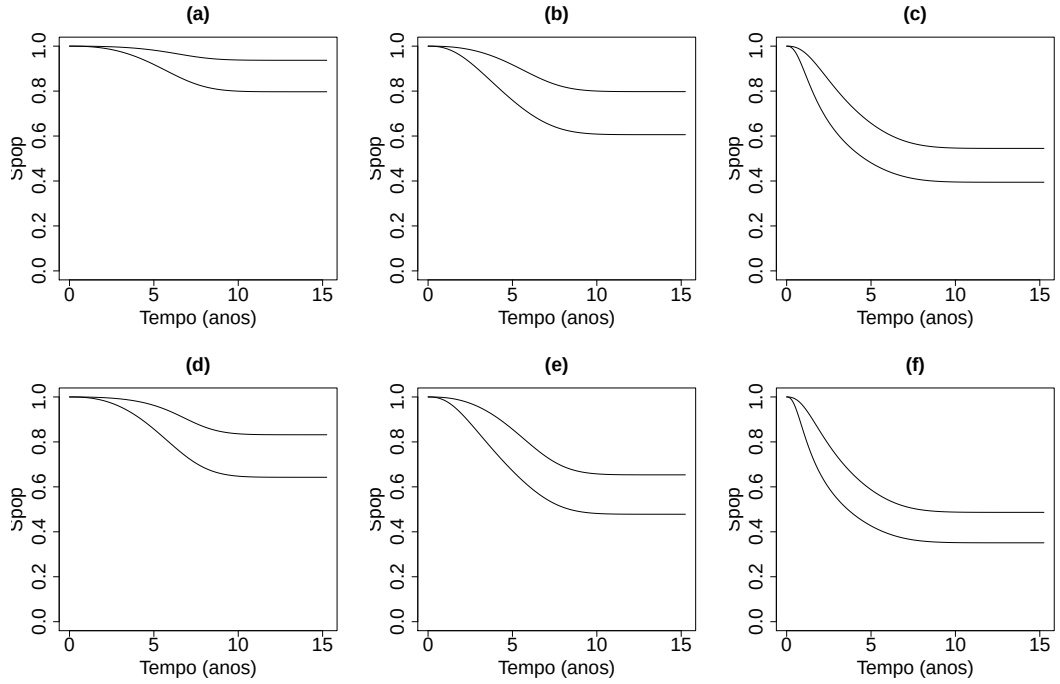

Figura 4.2: Função de sobrevivência sob o modelo HCBNP estratificado pelo estado de úlcera (superior: ausente, inferior: presente) para pacientes do sexo masculino com espessuras de tumor iguais a (a) 0.32, (b) 1.94, e (c) 8.32 mm, respectivamente, e para pacientes do sexo feminino com espessuras iguais a (d) 0.32, (e) 1.94, e (f) 8.32 mm, respectivamente.

Agora, voltamos a nossa atenção para o papel das covariáveis sobre a fração de cura p_0 (ver Tabela 4.1). As estimativas dos coeficientes β_1 na Tabela 4.4 indicam que o número médio de células iniciadas é maior quando a úlcera está presente, de modo que a fração de cura diminui. Visto que $\beta_{2_1} > 0$ e $\beta_{3_{fem}} > 0$ na Tabela 4.4, os valores mais elevados da espessura do tumor para pacientes do sexo feminino implicam em estimativas menores da fração de cura. A Figura 4.3 mostra o efeito combinado destas covariáveis sobre a fração de cura. As linhas correm quase paralelamente e as frações de cura, depois de uma queda acentuada, para espessura do tumor maior que 5mm e sexo feminino, estão em 49,79% e 35,94% (57,12% e 47,41% : sexo masculino) para o estado de úlcera ausente e presente, respectivamente.

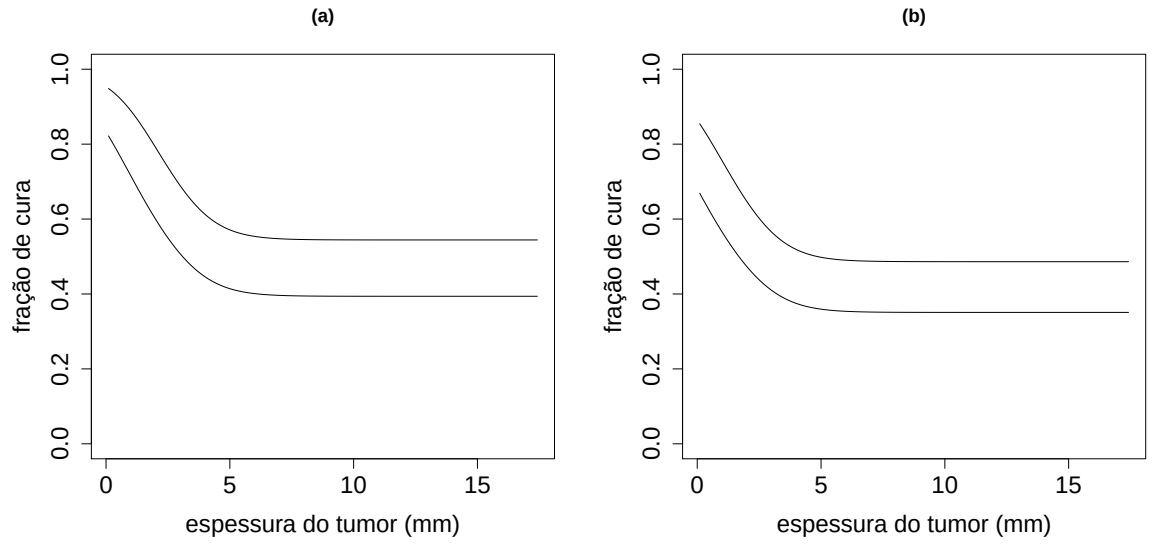

Figura 4.3: Fração de cura para o modelo HCBNP *versus* espessura do tumor estratificada pelo estado de úlcera (superior: ausente, inferior: presente) e sexo (a) masculino e (b) feminino, respectivamente.

Também obtemos os ajustes para os cinco modelos da Tabela 4.3 através da inferência bayesiana. Utilizamos distribuições *a priori* independentes e não informativas, sendo $\beta_{1_{pres}} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{1_{aus}} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{2_0} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{2_1} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{3_{mas}} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{3_{fem}} \sim \mathcal{N}\big(0, 10^3\big)$ $\gamma_1 \sim Gama\big(1, 0, 01\big)$, $\gamma_2 \sim \mathcal{N}\big(0, 10^3\big)$ e $\rho \sim Beta\big(1, 1\big)$, enquanto que $\phi \sim Gama\big(1, 0, 01\big)$ para o modelo HCBNP. Geramos duas cadeias paralelas de tamanho 35000 para cada parâmetro. Descartamos as primeiras 5000 e as restantes selecionadas de 10 em 10, resultando numa amostra de tamanho 3000. A convergência das cadeias foi monitorada empregando o método de Cowles & Carlin (1996).

Na Tabela 4.6 foram aplicados os critérios de seleção de modelos definidos na Seção 2.3.3 para os cinco modelos ajustados: HCPP, HCBP, HCBNP, HCGP e HCSLP. Os modelos HCPP e HCBNP se destacam como os melhores. Selecionarmos o modelo HCBNP como nosso modelo de trabalho. A Tabela 4.7 apresenta as médias *a posteriori*, os desvios padrão e os intervalos de credibilidade para os parâmetros do modelo HCBNP, incluindo o fator de redução de escala potencial estimado $\widehat{R}$ (Gelman & Rubin, 1992), que para todos os parâmetros está próximo de

um, indicando a convergência das cadeias. A Figura 4.4 apresenta as densidades marginais a posteriori aproximadas para cada parâmetro.

Para avaliar a robustez do modelo com relação à escolha dos hiperparâmetros das distribuições *a priori*, um pequeno estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros não apresentam muita diferença e não alteram os resultados apresentados na Tabela 4.6.

Tabela 4.6: Critérios DIC, EAIC, EBIC e B para os cinco modelos ajustados.

	Modelo				
Critério	HCPP	HCBP	HCBNP	HCGP	HCSLP
DIC	413,30	415,93	410,21	412,15	415,33
EAIC	427,61	428,64	423,81	426,71	428,15
EBIC	457,51	461,83	457,03	456,51	458,28
B	-206,96	-208,22	-205,11	-207,01	-207,36

Tabela 4.7: Médias *a posteriori*, desvios padrão e intervalos de credibilidade 95% (ICred 95%) para os parâmetros do modelo HCBNP e o fator de redução de escala potencial estimado $\widehat{R}$.

Parâmetro	**Média**	**desvio padrão**	**ICred 95%**	$\widehat{R}$
γ_1	2,36	0,52	(1,41 ; 3,45)	1,001
γ_2	-4,07	1,35	(-6,87 ; -1,66)	1,001
ρ	0,79	0,09	(0,66 ; 0,97)	1,003
ϕ	5,31	2,39	(1,15 10,64)	1,001
$\beta_{1_{pres}}$	2,35	1,58	(-0,23 ; 6,01)	1,002
$\beta_{1_{aus}}$	4,08	1,73	(0,87 ; 8,25)	1,003
β_{2_0}	-4,73	1,33	(-7,43 ; -2,49)	1,002
β_{2_1}	1,26	0,47	(0,45 ; 2,25)	1,002
$\beta_{3_{mas}}$	-1,55	1,19	(-3,88 ; 1,01)	1,001
$\beta_{3_{fem}}$	-0,29	1,03	(-2,75 ; 1,25)	1,001

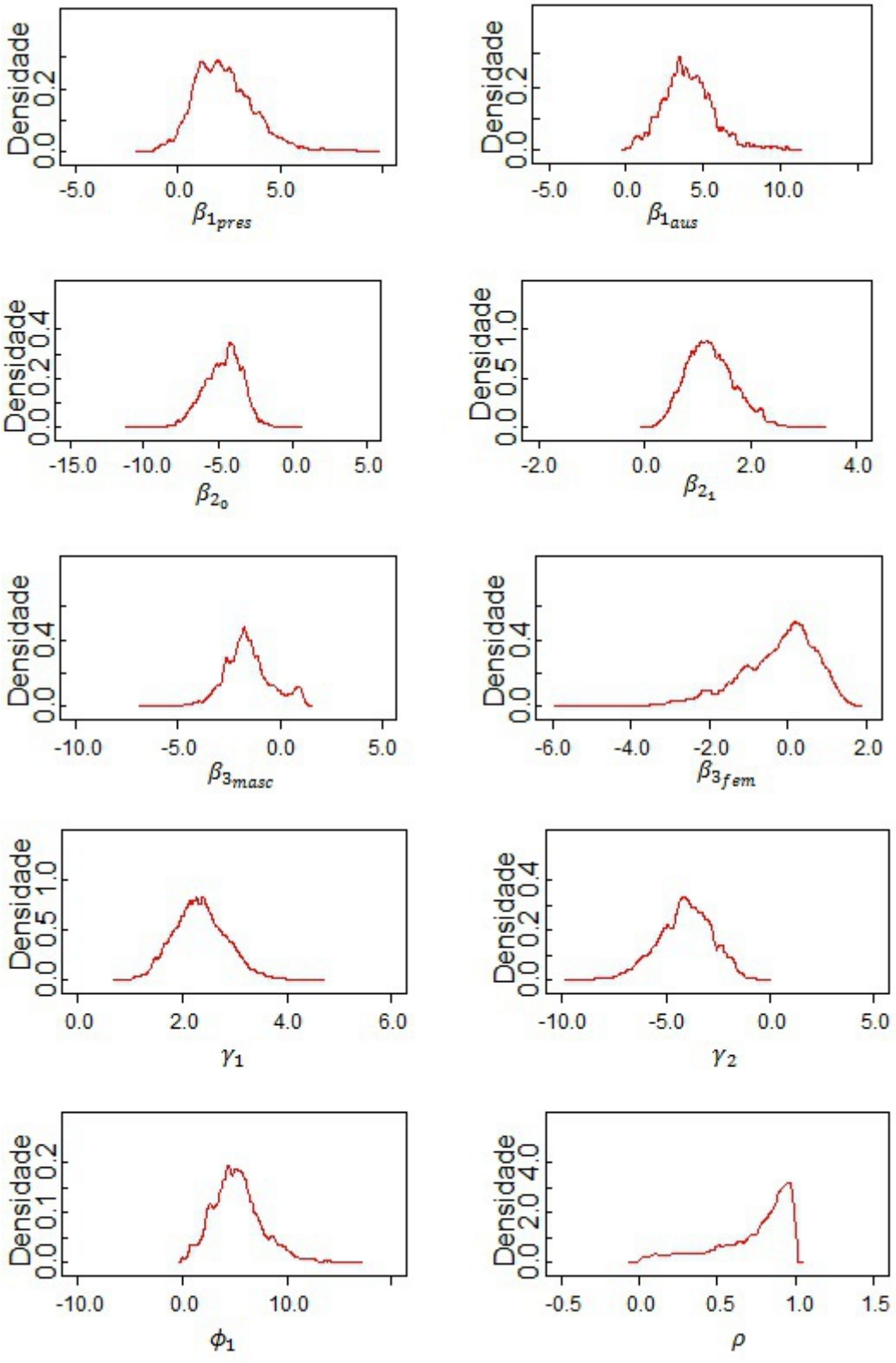

Figura 4.4: Densidades *a posteriori* aproximadas dos parâmetros.

As estimativas das médias das distribuições *a posteriori* (Tabela 4.7) e de máxima verossimilhança (Tabela 4.4) pouco diferem, ao passo que os intervalos de credibilidade são mais precisos do que os intervalos de confiança assintóticos.

4.6 Comentários finais

Neste capítulo propusemos um modelo de sobrevivência híbrido com fração de cura para acomodar características dos estágios não-observáveis da carcinogênese (iniciação, promoção e progressão) na presença de causas competitivas latentes dependentes, que estende o modelo do Capítulo 3. Assumimos uma distribuição SPGI para o número de células iniciadas e uma distribuição Weibull para os tempos de ocorrência do tumor, obtendo o modelo HCSPGI. O modelo HCSPGI incorpora dentro da análise características do estágio de progressão e a proporção de células malignas que morrem antes da indução do tumor, assumindo dependência biológica entre as células do tumor. A vantagem deste modelo é que se pode estimar a taxa de iniciação, a taxa de proliferação de células tumorais e a interdependência entre as células de um tecido iniciado desenvolvendo um tumor maligno, que não é possível na maioria dos modelos de fração de cura comumente utilizados. O processo de estimação bayesiana apresenta resultados mais precisos em termos de variabilidade das estimativas em relação ao processo clássico. A aplicabilidade do modelo foi demonstrada em um conjunto de dados reais de pacientes com câncer de melanoma.

Capítulo 5

Considerações Finais

Nesta tese foram apresentados modelos de sobrevivência com fração de cura baseados nos estágios inobserváveis do processo da carcinogênese (iniciação, promoção e progressão) na presença de causas competitivas latentes independentes ou dependentes, os quais estendem os modelos introduzidos por Rodrigues *et al.* (2010, 2011). As contribuições mais importantes desta tese dizem respeito à generalização e unificação dos modelos propostos por Rodrigues *et al.* (2010, 2011) com outros modelos já consagrados na literatura. Além disso, os novos modelos incorporam parâmetros com claro significado biológico.

As simulações dos modelos indicaram em geral um bom comportamento dos estimadores de máxima verossimilhança. A relevância prática e a aplicabilidade dos modelos foram demonstradas em conjuntos de dados reais de pacientes com câncer de melanoma, e além de oferecerem melhores interpretações para o mecanismo biológico da carcinogênese, proporcionaram bons ajustes.

Apesar de a tese ser enfatizada pela motivação biológica do processo da carcinogênese, os modelos propostos são satisfatórios para qualquer tipo de dados de tempo de falha que têm uma fração de sobreviventes. Portanto, acreditamos que esses modelos serão bastante úteis na compreensão global do processo biológico de uma variedade de infecções (por exemplo, HIV), experimentos quimiopreventivos de câncer, e assim por diante.

Propomos como possíveis pesquisas futuras que podem ser desenvolvidas com base nesta tese e nas suas referências.

1. Desenvolver os modelos destrutivos ou híbridos semiparamétricos (Ibrahim *et al.*, 2001)

2. Desenvolver os modelos destrutivos ou híbridos com tempo de vida acelerado (Yamaguchi, 1992; Sinha *et al.*, 2003)

3. Estudar os modelos destrutivos ou híbridos com outros esquemas de censura. Por exemplo, censura intervalar (Xiang *et al.*, 2011);

4. Estudar testes para comparar diferenças entre frações de cura (Gray & Tsiatis, 1989);

5. Estudar métodos para análise da qualidade do ajuste e das suposições necessárias ao adequado uso dos modelos com fração de cura.

Referências

Ainsworth, E. J. (1982). Radiation carcinogenesis-perspectives. *In Probability Models and Cancer*, **ed. L. Le Cam and L. Neyman. North-Holland, Amsterdam**, 99–169.

Armitage, P. & Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogesis. *British J. Cancer*, **8**, 1–12.

Banerjee, S. & Carlin, B. P. (2004). Parametric spatial cure rate model for interval-censored time-to-relapse data. *Biometrics*, **60**, 268–275.

Barral, A. M. (2001). *Immunological Studies in Malignant Melanoma: Importance of TNF and the Thioredoxin System*. Doctorate Thesis - Linkoping University, Linkoping, Sweden.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **42**, 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society B*, **11**(1), 15–53.

Borges, P., Rodrigues, J. & Louzada-Neto, F. (2011a). A correlated mechanistic cure rate survival model under a hybrid latent activation scheme. Technical Report TR-11-01, Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, BRASIL.

Borges, P., Rodrigues, J., Louzada-Neto, F. & Balakrishnan, N. (2011b). A cure rate survival model under a hybrid latent activation scheme: an application to malignant melanoma data. Technical Report TR-11-01, Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, BRASIL.

Borges, P., Rodrigues, J. & Balakrishnan, N. (2012). Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data. *Computational Statistics and Data Analysis*, **56**, 1703–1713.

Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, best, Carlin and Van der Linde. *Journal Royal Statistical Society, Series B*, **64**, 616–618.

Carlin, B. P. & Louis, T. A. (2002). *Bayes and Empirical Bayes Methods for Data Analyis*. Chapman & Hall, Boca Raton, second edition.

Castillo, J. & Pérez-Casany, M. (1998). Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, **50**, 567–585.

Castillo, J. & Pérez-Casany, M. (2005). Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference*, **134**, 486–500.

Chen, M. H., Ibrahim, J. G. & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909–919.

Chen, M. H., Shao, Q. M. & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Chen, M. H., Ibrahim, J. G. & Sinha, D. (2002). Bayesian inference for multivariate survival data with cure fraction. *Journal of Multivariate Analysis*, **89**, 101–126.

Cnaan, A. (1985). Survival models with two phases and length biased sampling. *Communications in Statistics - Theory and Methods*, **14**, 861–886.

Collet, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, New York.

Consul, P. C. (1990). New class of location-parameter discrete probability distributions and their characterizations. *Communications in Statistics: Theory and Methods*, **19**, 4653–4666.

Cooner, F., Banerjee, S., Carlin, B. & Sinha, D. (2007). Flexible cure rate modelling under latent activation schemes. *Journal American Statistics Association*, **102**, 560–572.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

de Castro, M., Cancho, V. G. & Rodrigues, J. (2007). A flexible model for survival data with a surviving fraction. Technical Report 245, Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, BRASIL.

de Castro, M., Cancho, V. G. & Rodrigues, J. (2009). A Bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal*, **51**, 443–455.

Dewanji, A., Venzon, D. J. & Moolgavkar, S. H. (1989). A stochastic two-stage model for cancer risk assessment. *Risk Analysis*, **9**, 179–187.

Draper, N. R. & Smith, H. (1998). *Applied Regression Analysis*. John Wiley and Sons, New York.

Dunn, P. K. & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Fahrmeir, L. (1988). A note on asymptotic testing theory for nonhomogeneous observations. *Stochastic Processes and Their Applications*, **28**, 267–273.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long term survivors. *Biometrics*, **38**, 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, **14**, 257–262.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, **6**, 13–25.

Gamerman, D. & Lopes, H. F. (2006). *Markov Chain Monte Carlo: stochastic simulation for bayesian inference*. 2nd edn. Boca Raton: Chapman & Hall.

Gelfand, A. F., Dey, D. K. & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *In: Bayesian statistics*, **4**, 147–167.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

George, E. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Goldman, A. I. (1984). Survivorship analysis when cure is a possibility: A Monte Carlo study. *Statistics in Medicine*, **3**, 153–163.

Gray, R. J. & Tsiatis, A. A. (1989). A linear rank test for use when the main interest is in differences in cure rates. *Biometrics*, **45**, 889–904.

Gupta, R. C. (1974). Modified power series distributions and some of its applications. *Sankhyã, Series B*, **35**, 288–298.

Hanin, L. G., Rachev, S. T., Tsodikov, A. D. & Yakovlev, A. Y. (1997). A stochastic model of carcionogenesis and tumor size at detection. *Advances in Applied Probability*, **29**, 607–628.

Haynatzki, G. R., Weron, K. & Haynatzka, V. R. (2000). A new statistical model of tumor latency time. *Mathematical and Computer Modelling*, **32**, 251–256.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, **57**, 383–388.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*. 2nd edition, New York: John Wiley & Sons.

Kim, S., Chen, M.-H. & Dey, D. (2011). A new threshold regression model for survival data with a cure fraction. *Lifetime Data Analysis*, **17**, 101–122.

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. (2000). High- and low-dose interferon alfa-2b

in high-risk melanoma: First analysis of Intergroup Trial E1690/S9111/C9190. *Journal of Clinical Oncology*, **18**, 2444–2458.

Klebanov, L. B., Rachev, S. T. & Yakovlev, A. (1993). A stochastic model of radiation carcinogenesis: Latent time distributions and their properties. *Mathermatical Biosciences*, **113**, 51–75.

Kolev, N., Minkova, L. & Neytchev, P. (2000). Inflated-parameter family of generalized power series distributions and their application in analysis of overdispersed insurance data. *ARCH Research Clearing House*, **2**, 295–320.

Kopp-Schneider, A., Portier, C. J. & Rippmann, F. (1991). The application of a multistage model that incorporates DNA demage and repair to the analysis of initiation/promotion experiments. *Mathermatical Biosciences*, **105**, 139–166.

Li, C. S., Taylor, J. & Sy, J. (2001). Identifiability of cure models. *Statistics and Probability Letters*, **54**, 389–395.

Maller, R. A. & Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Minkova, L. (2002). A generalization of the classical discrete distributions. *Communications in Statistics - Theory and Methods*, **31**(6), 871–888.

Mizoi, M., Lima, A. C. & Bolfarine, H. (2007). Cure rate models with measurement error. *Communications in Statistics - Simulation and Computation*, **36**, 185–196.

Mizoi, M. F. (2004). *Influência local em modelos de sobrevivência com fração de cura*. Ph.D. thesis, IME-USP.

Nordling, C. O. (1953). A new theory on the cancer inducing mechanism. *British J. Cancer*, **7**, 68–72.

Ortega, E. M. M., Cancho, V. G. & Paula, G. A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, 79–106.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**, 863–867.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyã, Series A*, **27**, 311–324.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, **54**, 507–554.

Rodrigues, J., de Castro, M., Cancho, V. & Balakrishnan, N. (2009a). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**, 3605–3611.

Rodrigues, J., de Castro, M., Cancho, V. G. & Louzada-Neto, F. (2009b). On the unification of the long-term survival models. *Statistics & Probability Letters*, **79**, 753–759.

Rodrigues, J., Cancho, V. G., de Castro, M. & Balakrishnan, N. (2010). A Bayesian destructive weighted Poisson cure rate model and an application to a cutaneous melanoma data. *Statistical Methods in Medical Research*, **doi: 10.1177/0962280210391443**.

Rodrigues, J., de Castro, M., Balakrishnan, N. & Cancho, V. G. (2011). Destructive weighted Poisson cure rate models. *Lifetime Data Analysis*, **17**, 333–346.

Ross, G. J. S. & Preece, D. A. (1985). The negative binomial distribution. *Statistician*, **34**, 323–336.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saha, K. & Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, **61**, 179–185.

Scheike, T. (2009). *timereg package, with contributions from T. Martinussen and J. Silver,*. R package version 1.1-6.

Sen, P. K. & Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society, Series C*, **54**, 127–142.

Sinha, D., Patra, K. & Dey, D. K. (2003). Modelling accelerated life test data by using a Bayesian approach. *Applied Statistics*, **52**, 249–259.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal Royal Statistical Society, Series B*, **64**, 583–639.

Sy, J. P. & Taylor, J. M. G. (2000). Estimation in a proportional hazards cure model. *Biometrics*, **56**, 227–336.

Tan, W. Y. (1991). *Stochastic Models of Carcinogenesis*. Marcel Dekker, New York.

Thomas, A., O'Hara, B., Ligges, U. & Sturtz, S. (2006). Making BUGS open. *R News*, **6**, 12–17.

Tournoud, M. & Ecochard, R. (2007). Application of the promotion time cure model with time-changing exposure to the study of hiv/aids and other infectious diseases. *Statistics in Medicine*, **26**, 1008–1021.

Tournoud, M. & Ecochard, R. (2008). Promotion time models with time-changing exposure and heterogeneity: application to infectious diseases. *Biometrical Journal*, **50**, 395–407.

Tsodikov, A. D., Asselain, B. & Yakovlev, A. Y. (1997). A distribution of tumor size at detection: An application to breast cancer data. *Biometrics*, **53**, 1495–1502.

Tsodikov, A. D., Ibrahim, J. G. & Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078.

Xiang, L., Ma, X. & Yau, K. K. W. (2011). Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in Medicine*, **30**, 995–1006.

Yakovlev, A. & Polig, E. (1996). A diversity of responses displayed by a stochastic model of radiation carcinogenesis allowing for cell death. *Mathermatical Biosciences*, **132**, 1–33.

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

Yakovlev, A. Y., Hannin, L. G., Rachev, L. G. & Tsodikov, A. D. (1996). A distribution of tumor size at detection and its limiting form. *Proceeding of the National Academy of Sciences, U.S.A*, **93**, 6671–6675.

Yamaguchi, K. (1992). Accelerated failure-time regression-models with a regression-model of surviving fraction - an application to the analysis of permanent employment in Japan. *Journal of the American Statistical Association*, **87**, 284–292.

Yang, G. L. & Chen, C. W. (1991). A stochastic two-stage carcinogenesis model: A new approach to computing the probability of observing tumor in animal bioassays. *Mathematical Biosciences*, **104**, 247–258.

Yin, G. & Ibrahim, J. G. (2005). Cure rate models: A unified approach. *Canadian Journal of Statistics*, **33**, 559–570.

Zelen, M. & Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, **56**, 601–614.

Zhao, Y., Lee, A. H., Yau, K. K. W. & Burke, V. (2009). A score test for assessing the cured proportion in the long-term survivor mixture model. *Statistics in Medicine*, **28**, 3454–3466.