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
Novembro/2011

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.

Resumo

Neste trabalho, propomos novos 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 dentro da análise, características do estágio de progressão do tumor, bem como a proporção de células iniciadas que foram "promovidas"a malignas , o problema de estimar a proporção de células malignas que "morrem"antes da indução do tumor e uma estrutura de dependência entre as células iniciadas, 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. Aplicações a conjuntos de dados reais mostraram a aplicabilidade dos modelos.

Palavras-chave: carcinogênese, modelos de sobrêvivência, estrutura de correlação, esquema de ativação hibrido.

Abstract

In this thesis we propose new models for the survival with surviving fraction to describe the biological mechanism of occurrence of events of interest (cancer) in studies of carcinogenesis in the presence of competitive independent latent causes or correlations. 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 tumor until a detectable cancer. These models allow a simple pattern of the dynamics of tumor growth, and incorporate into the analysis, characteristics of the stage of tumor progression, the proportion of initiated cells that have been "promoted"the malignant, the problem of estimating the proportion of malignant cells to "die"before the induction of the tumor and a structure dependence among the initiated cells, which is not possible in commonly used cure models. For the models proposed, discussed the inferential process, in terms of classical and Bayesian. Applications to real data sets showed the applicability of the models.

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

Sumário

1 Introdução **1**

2 Modelo com fração de cura destrutivo correlacionado **7**

2.1 Formulação do modelo 8

2.2 Casos especiais do modelo proposto 12

2.2.1 Modelo destrutivo correlacionado Poisson (DCP) 12

2.2.2 Modelo destrutivo correlacionado binomial (DCB) 13

2.2.3 Modelo destrutivo correlacionado binomial negativa (DCBN) 14

2.2.4 Modelo destrutivo correlacionado série logarítmica (DCSL) 15

2.3 Inferência 18

2.3.1 Estimação de máxima verossimilhança 18

2.3.2 Inferência Bayesiana 21

2.3.3 Critério para comparação de modelos 23

2.4 Dados de câncer de melanoma 24

2.5 Comentários finais 30

3 Modelo com fração de cura híbrido **32**

3.1 Formulação do modelo 33

3.2 Alguns modelos específicos 39

3.2.1 Modelo híbrido Poisson ponderada exponencialmente-Poisson (HPPEP) . 39

3.2.2 Modelo híbrido binomial negativa-Poisson (HBNP) 40

3.2.3 Modelo híbrido COM-Poisson-Poisson (HCPP) 41

3.3 Inferência 42
3.3.1 Função de verossimilhança 42
3.3.2 Distribuições a priori e a posteriori 46
3.4 Dados de câncer de melanoma 47
3.5 Comentários finais 56

4 Modelo com fração de cura híbrido correlacionado 58
4.1 Formulação do modelo 59
4.2 Alguns modelos específicos 60
4.2.1 Modelo híbrido correlacionado Poisson-Poisson (HCPP) 61
4.2.2 Modelo híbrido correlacionado binomial-Poisson (HCBP) 61
4.2.3 Modelo híbrido correlacionado binomial negativa-Poisson (HCBNP) 62
4.2.4 Modelo híbrido correlacionado série logarítmica-Poisson (HCSLP) 62
4.3 Inferência 65
4.3.1 Função de verossimilhança 65
4.3.2 Distribuições a priori e a posteriori 66
4.4 Dados de câncer de melanoma 67
4.5 Comentários finais 73

Lista de Figuras

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. 2

2.1 Representação do modelo proposto DCSPG em termos de um diagrama. 12
2.2 Painel esquerdo: gráfico TTT. Painel direito: curva Kaplan-Meier estratificada pelo estado de úlcera (superior: presente, inferior: ausente). 25
2.3 Função de sobrevivência sob o modelo DCG estratificado pelo estado 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. 27
2.4 Fração de cura para o modelo DCG *versus* espessura do tumor estratificada pelo estado de úlcera (superior: ausente, inferior: presente). 28
2.5 Densidades *a posteriori* marginais aproximadas dos parâmetros. 30

3.1 Representação do modelo proposto HPPPP em termos de um diagrama. 39
3.2 Painel esquerdo: gráfico TTT. Painel direito: curva Kaplan-Meier estratificada por categoria do nódulo (1 até 4, de cima para baixo). 48
3.3 Gráfico QQ do quantil aleatorizado residual normalizado com a reta identidade para o modelo HGP (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados). 49

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 iguais a (a) 29, (b) 47, e (c) 70 anos, respectivamente, e espessura do tumor 3,94 mm. 51
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. 52
3.6 Densidades *a posteriori* marginais aproximadas dos parâmetros. 54
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. 55

4.1 Gráfico QQ do quantil aleatorizado residual normalizado com a reta identidade para o modelo HCBNP (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados). 68
4.2 Função de sobrevivência sob o modelo HCBNP estratificado pelo estado de úlcera (superior: ausente, inferior: presente) para pacientes do sexo masculino com idades iguais a (a) 29, (b) 47, e (c) 70 anos, respectivamente, e para pacientes do sexo feminino com idades iguais a (d) 29, (e) 47, e (f) 70 anos, respectivamente. 70
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. 71
4.4 Densidades *a posteriori* marginais aproximadas dos parâmetros. 73

Lista de Tabelas

2.1 As escolhas de a_n, $g(\theta)$ e o parâmetro θ para alguns casos especiais da distribuição SPGI. 9

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. 17

2.3 Os valores do $\max \log L(\cdot)$ e as estatísticas AIC e BIC para os sete modelos ajustados: DCP, DCB, DCBN, DCG, DCSL, binomial negativa e geométrico. 26

2.4 Estimativas de máxima verossimilhança dos parâmetros do modelo DCG, seus desvios padrão e seus intervalos de confiança assintóticos de 95% (IC 95%). . . . 26

2.5 As estatísticas DIC, EAIC, EBIC e B para os sete modelos ajustados: DCP, DCB, DCBN, DCG, DCSL, binomial negativa e geométrico. 29

2.6 Médias *a posteriori*, os desvios padrão e os intervalos de credibilidade 95% (ICred 95%) para os parâmetros do modelo DCG. e o fator de redução de escala potencial estimado $\widehat{R}$. 29

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. 42

3.2 Os valores do $\max \log L(\cdot)$ e as estatísticas AIC e BIC para os quatros modelos ajustados: HPPEP, HBNP, HCPP e HGP. 49

3.3 Estimativas de máxima verossimilhança dos parâmetros do modelo HGP, seus desvios padrão e seus intervalos de confiança assintóticos de 95% (IC 95%). . . . 50

3.4 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. 50
3.5 As estatísticas DIC, EAIC, EBIC e B para os quatro modelos ajustados: HPPEP, HBNP, HCPP e HGP. 53
3.6 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 $\widehat{R}$. 53
3.7 Médias *a posteriori*, os desvios padrão e os 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. 54
3.8 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. 56

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. 64
4.2 Os valores do $\max \log L(\cdot)$ e as estatísticas AIC e BIC para os cinco modelos ajustados, HCPP, HCBP, HCBNP, HCGP e HCSLP. 68
4.3 Estimativas de máxima verossimilhança dos parâmetros do modelo HCBNP, seus desvios padrão e seus intervalos de confiança assintóticos de 95% (IC 95%). . . . 69
4.4 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. 69
4.5 As estatísticas DIC, EAIC, EBIC e B para os cinco modelos ajustados: HCPP, HCBP, HCBNP, HCGP e HCSLP. 72

4.6 Médias *a posteriori*, os desvios padrão e os intervalos de credibilidade 95% (ICred 95%) para os parâmetros do modelo HCBNP e o fator de redução de escala potencial estimado $\widehat{R}$. 72

Capítulo 1

Introdução

O câncer é definido como um tumor maligno, mas não é uma doença única e sim um conjunto de mais de 200 patologias, caracterizado pelo crescimento descontrolado de células anormais (malignas) e como conseqüência ocorre à invasão de órgãos e tecidos adjacentes envolvidos, podendo se disseminar para outras regiões do corpo, dando origem a tumores em outros locais. Essa disseminação é chamada de metástase (ver INCA (2011)).

O câncer ocorrer quando uma célula normal sofrer 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 visível. Esse processo passa por vários estágios 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 se encontram, geneticamente alteradas, porém ainda não é possível se 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, de forma lenta e gradual. 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 se caracteriza pela multiplicação descontrolada e irreversível das células alteradas. 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.

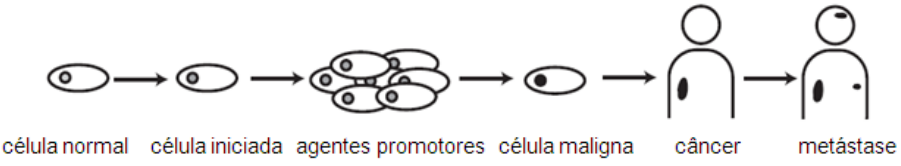

Figura 1.1: Evolução de uma célula normal em uma célula cancerosa. Os agentes cancerígenos conduzem a uma célula iniciada em cancerígena. Finalmente, células cancerígenas se espalham pelo corpo, formando os tumores.

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 subseqüente. 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 exatos mecanismos 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 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), ver 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 incoporarem 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, poder vir a não ser suscetível ao câncer (indivíduos fora de risco). O modelo clássico de Berkon-Gage (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.*, 1999b; Ibrahim *et al.*, 2001b; 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 dada por uma ou várias causas competitivas (células); ver Gordon (1990). O número de causas, assim como o tempo de sobrevivência associado a cada causa, não são observados (Cox & Oakes, 1984) e são chamados de fatores ou riscos latentes. O modelo proposto por Chen *et al.* (1999b) baseia-se na existência de fração de cura com fatores latentes, assim como, por exemplo, Yakovlev & Tsodikov (1996), Ibrahim *et al.* (2001b), Chen *et al.* (2002), Banerjee & Carlin (2004) e Yin & Ibrahim (2005). Uma outra abordagem é desenvolvida por Cooner *et al.* (2007) que modelam estocasticamente a sequência ordenada de tempos latentes, os quais induzem a ocorrência do evento em estudo. Para mais detalhes veja, por exemplo, Rodrigues *et al.* (2009b) e Balka *et al.* (2009). O cenário de causas competitivas permite longa duração quando a probabilidade de o número de causas latentes ser igual a zero é não nula.

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, COM-Poisson (Chen *et al.*, 1999b; 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 modelos de sobrevivência de dois estágios para dados da carcinogênese apresentam duas limitações básicas:

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

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

Para a limitação (i), nós podemos discutir que os modelos de sobrevivência de dois estágios para carcinogênese foca sobre eventos que precedem a ocorrência da primeira célula maligna em um tecido. Uma descrição explícita do estágio de progressão do tumor é evitado em modelos de dois estágios. Isto também é verdade com o modelo de radiação para carcinogênese proposto por Klebanov *et al.* (1993) e sua generalização por Yakovlev & Polig (1996). 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 proporcionou uma motivação para o presente trabalho. Para limitação (ii), Haynatzki *et al.* (2000) discutiram o problema 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 denvolvimento uma 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 matemáticos que possam incorporar adequadamente a dependência biológica, e isso que proporcionou a outra motivação para o presente trabalho.

Portanto, o objetivo principal deste trabalho é 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, propormos novos 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 extendem 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. Os modelos propostos foram aplicados a um conjunto de dados reais. Os resultados obtidos neste capítulo foram

condensados em um relatório técnico Borges *et al.* (2011a), aceito à publicação.

Nos Capítulos 3 e 4 propomos modelos de sobrivivência 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 de tumores. A diferença entre os dois Capítulos esta no fato que as células iniciadas (causas competitivas) definidas no Capítulo 3 são assumidas independentes, enquanto no Capítulo 4, incorporamos uma estrutura de dependência entre as mesmas. 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.* (2011b,c), submetidos à publicação.

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 a 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 tumor 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 tumorais após algum tratamento intensivo. As existentes provas rádio-biológico sobre as características temporais de reparação enzimática mencionado 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 biológica das células. Sugerimos ao leitor o artigo de (Klebanov *et al.*, 1993) para conhecer algumas referências específicas sobre este assunto. Além disso, os livros Maller & Zhou (1996) and Ibrahim *et al.* (2001a), bem como os artigos recentes Tsodikov *et al.* (2003), Cooner *et al.* (2007), Tournoud & Ecochard (2007), de Castro *et al.* (2009), Ortega *et al.* (2009) e Zhao *et al.* (2009) 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 de incoporamos uma estrutura de dependência entre as células iniciadas (Haynatzki *et al.*, 2000). Para criar a estrutura de dependência entre as células, nós usamos uma extensão da distribuição série de potência generalizada incluíndo um parâmetro adicional ρ (distribuição série de potência generalizada inflada (SPGI), estudada por Kolev *et al.* (2000)). O parâmetro ρ tem uma interpretação natural em termos de proporção de "'zero-inflado"e coeficiente de correlação. Em nossa abordagem, o número de células iniciadas segue uma distribuição SPGI. A distribuição SPGI é uma escolha natural para modelagem de dados de contagem correlacionados que apresentam superdispersão. 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. Isso significa que, o que está registrado é apenas a parte não danificada do número original de células iniciadas não eliminadas pelo tratamento ou reparadas pelo sistema de reparo de um indivíduo, que é representada por uma variável composta.

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, um conjunto de dados reais de câncer melanoma ilustra a utilidade do modelo proposto. Comentários finais são apresentados na Seção 2.5.

2.1 Formulação do modelo

Para um indivíduo na população, vamos denotar N o número de células iniciadas relacionados com a ocorrência de um tumor. Suponha que a variável não observada N (variável latente) segue uma distribuição série de potência generalizada inflada (SPGI) com função massa

de probabilidade ($f.m.p$) dada por

$$p_n = \mathbb{P}[N = n; \theta, \rho] = \frac{1}{g(\theta)} \sum_{n_1, n_2, \ldots} a_n \big[\theta(1-\rho)\big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \quad n = 0, 1, 2, \ldots, \ \rho \in [0, 1), \tag{2.1}$$

em que a_n 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 ∞) é tal que $g(\theta)$ é finita, e o somatório é sobre o conjunto de todos os inteiros não negativos $n_1, n_2, \ldots$, tal que $\sum_{i=1}^{\infty} i n_i = n$. Para mais detalhes sobre a distribuição SPGI, ver Kolev *et al.* (2000) e Minkova (2002). O parâmetro ρ é uma medida de associação entre as células tumorais. Valores de $\rho \to 1$ indicam forte associação entre as células, enquanto $\rho \to 0$ implica fraca associação entre as células. É interessante notar que quando $\rho = 0$ (isto é, quando há independência entre as células), a distribuição SPGI torna-se uma distribuição série de potência generalizada (Gupta, 1974; Consul, 1990). A Tabela 2.1 mostra as escolhas de a_n, $g(\theta)$ e o parâmetro θ correspondente 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 $m \in \mathbb{Z}^+$ (conjunto dos inteiros não negativos) e $\phi > -1$ devem ser tratados como parâmetros pertubadores.

Tabela 2.1: As escolhas de a_n, $g(\theta)$ e o parâmetro θ para alguns casos especiais da distribuição SPGI.

Distribuições	a_n	$g(\theta)$	θ	s
PI	$\frac{1}{n_1! n_2! \cdots}$	e^{θ}	η	∞
BI	$\binom{m}{m-n_1-n_2-\cdots, n_1, n_2, \ldots}$	$(1+\theta)^m$	$\frac{\pi}{1-\pi}$	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 série de potência generalizada inflada N é dada por

$$\mathbb{A}_{p_n}(z) = \frac{g\big(\theta z(1-\rho)(1-z\rho)^{-1}\big)}{g(\theta)} \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.2}$$

Agora, após um tratamento prolongado ("processo destrutivo"), temos como conseqüência imediata a formação ou não de lesões pré-cancerosas em um genoma das células. Seja $N = n$ o número de tais lesões, ou células alteradas, ou células iniciadas após o tratamento, e X_j, $j = 1, 2, \ldots, n$, são variáveis aleatórias independentes, independentemente de N, seguindo uma distribuição Bernoulli com probabilidade de sucesso p indicando a presença da j^{th} lesão and $f.g.p$

$$\mathbb{A}_{X_j}(z) = 1 - p(1-z), \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.3}$$

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

$$D = \begin{cases} \sum_{j=1}^{N} X_j & , \quad se \quad N > 0 \\ 0 & , \quad se \quad N = 0 \end{cases}. \tag{2.4}$$

Pela destruição ou o não reparo das células, temos que $D \leq N$. A distribuição condicional de D, dado $N = n$, será, portanto, referida como distribuição destrutiva.

Esta visão de (2.4), 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 (2.4) denota o número de células malignas vivas que são descendentes da j^{th} célula maligna iniciada durante algum intervalo de tempo . Neste contexto, D então denota o número total de células malignas que vivem em algum momento específico.

No cenário de causas competitivas (Cox & Oakes, 1984), o número de lesões não reparadas D em (2.4) e o tempo V para transformar essas lesões em um tumor detectável são ambos não observáveis (variáveis latentes). Vamos chamar V tempo de progressão. Assim, o tempo de início do tratamento até detecção do tumor (que é o evento de interesse) em um determinado indivíduo é definido pela variável aletória

$$Y = \min\{V_1, V_2, \ldots, V_D\} \tag{2.5}$$

para $D \geq 1$, e $Y = \infty$ se $D = 0$, o que leva a uma proporção p_0 da população cujas lesões são reparadas pelo tratamento, também chamada de "fração de cura". Nós assumimos que $V_1, V_2, \ldots$ são independentes de D. Assumimos também que, condicionada a D, as variáveis V_j são independentes e identicamente distribuídas (i.i.d).

De acordo com Rodrigues *et al.* (2009b, 2011), a função de sobrevivência de longa duração da variável aleatória Y em (2.5) 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 $S(\cdot)$ denota a função de sobrevivência comum dos tempos de vida não observáveis em (2.5) e $\mathbb{A}_D(\cdot)$ é a função geradora de probabilidade da variável composta D, a qual converge quando $z = S(y) \in [0, 1]$. Levando em conta (2.2) e (2.3), a função de sobrevivência de longa duração do tempo observado de um tumor detectável em (2.5) é expressada por

$$S_{pop}(y) = \frac{g\Big(\theta\big(1-\rho\big)\big(1-pF(y)\big)\big[1-\big(1-pF(y)\big)\rho\big]^{-1}\Big)}{g(\theta)}, \tag{2.6}$$

em que $F(y) = 1 - S(y)$. Se usarmos especificamente $\rho = 0$, obtemos a função de sobrevivência de longa duração série de potência generalizada.

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

$$\lim_{y\to\infty} S_{pop}(y) = p_0 = \frac{g\big(\theta\big(1-\rho\big)\big(1-p\big)\big[1-\big(1-p\big)\rho\big]^{-1}\big)}{g(\theta)}, \tag{2.7}$$

sendo p_0 denota a proporção de indivíduos "curados"ou "imunes"presentes na população a partir do qual os dados da amostra surgiu. Referimo-nos ao modelo definido em (2.6) por modelo destrutivo correlacionado série de potência generalizada inflada, ou simplesmente o modelo DCSPG. A Figura 2.1 ilustra o modelo DCSPG em termos de um diagrama.

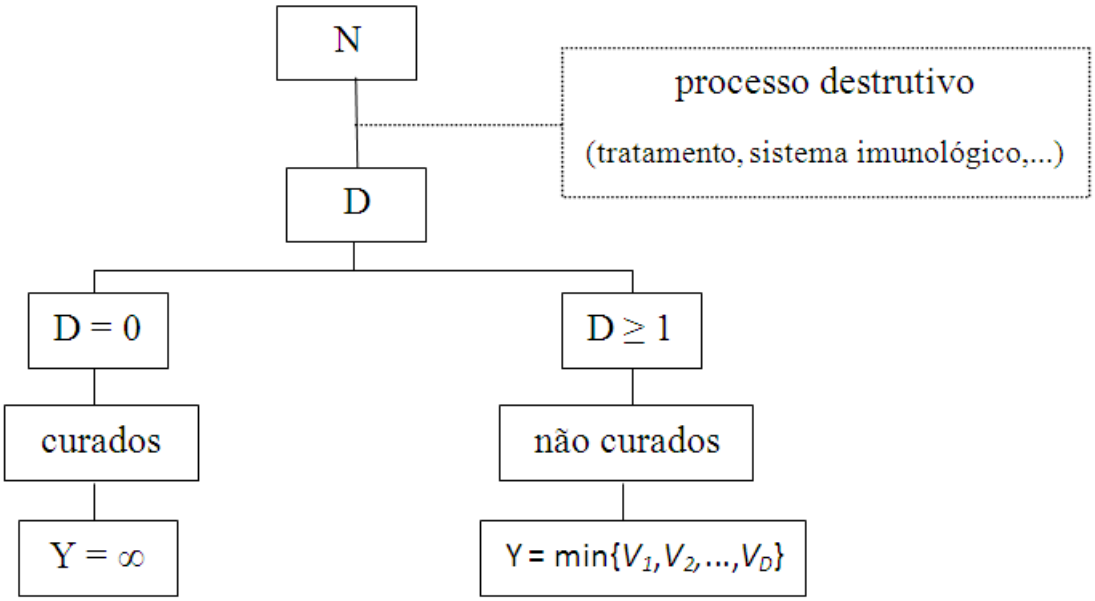

Figura 2.1: Representação do modelo proposto DCSPG em termos de um diagrama.

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 a escolha de $a_n = \frac{1}{n_1!n_2!\cdots}$, $g(\theta) = \exp\{\theta\}$ 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

$$\mathbb{P}_{Poi}[N = n] = \sum_{n_1,n_2,\cdots} \frac{e^{-\eta}}{n_1!n_2!\cdots}\Big[\eta(1-\rho)\Big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{2.8}$$

sendo $n = 0, 1, 2, \ldots$, 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 (2.8) é dada por

$$\mathbb{P}_{Poi}[N = n] = \begin{cases} e^{-\eta} & , \quad n = 0 \\ e^{-\eta}\sum_{i=1}^{n}\binom{n-1}{i-1}\frac{\left[\eta(1-\rho)\right]^i \rho^{n-1}}{i!} & , \quad n = 1, 2, \ldots. \end{cases} . \tag{2.9}$$

A média e a variância de N são

$$\mathbb{E}[N] = \frac{\eta}{1-\rho} \quad \text{e} \quad \mathbb{V}ar[N] = \frac{\eta(1+\rho)}{(1-\rho)^2}, \tag{2.10}$$

respectivamente. A $f.g.p$ é dada por

$$\mathbb{A}_N(z) = \exp\left\{-\frac{\eta(1-z)}{1-z\rho}\right\} \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.11}$$

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

$$S_{pop}(y) = \exp\left\{-\frac{\eta p F(y)}{1-\rho(1-pF(y))}\right\}. \tag{2.12}$$

Existem dois importantes casos especiais de (2.12). 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.*, 1999a).

2.2.2 Modelo destrutivo correlacionado binomial (DCB)

Para a escolha de $a_n = \binom{m}{m-n_1-n_2-\cdots,n_1,n_2,\ldots}$, $g(\theta) = (1+\theta)^m$ 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 $m \in \mathbb{Z}^+$, e sua $f.m.p$ é da forma

$$\mathbb{P}_{Bin}[N=n] = (1-\pi)^m \sum_{n_1,n_2,\ldots} \binom{m}{m-n_1-n_2-\cdots,n_1,n_2,\ldots} \rho^n \left\{\frac{\pi(1-\rho)}{\rho(1-\pi)}\right\}^{\sum_{i=1}^{\infty} n_i}, \tag{2.13}$$

sendo $n = 0,1,\ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, \ldots$, tal que $\sum_{i=1}^{\infty} i n_i = n$. Uma expressão alternativa para a $f.m.p$ em (2.13) é dada por

$$\mathbb{P}_{Bin}[N=n] = \begin{cases} (1-\pi)^m & , \quad n=0 \\ \sum_{i=1}^{\min(n,m)} \binom{m}{i}\binom{n-1}{i-1}\left[\pi(1-\rho)\right]^i (1-\pi)^{m-i}\rho^{n-i} & , \quad n=1,2,\ldots. \end{cases}. \tag{2.14}$$

A média e a variância de N são

$$\mathbb{E}[N] = \frac{m\pi}{1-\rho} \quad \text{e} \quad \mathbb{V}ar[N] = \frac{m\pi(1-\pi+\rho)}{(1-\rho)^2}, \tag{2.15}$$

respectivamente. A $f.g.p$ é dada por

$$\mathbb{A}_N(z) = \left[1 - \frac{\pi(1-z)}{1-z\rho}\right]^m \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.16}$$

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

$$S_{pop}(y) = \left[1 - \frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]^m. \tag{2.17}$$

Agora, fazendo $m \to \infty$ e $\pi \to 0$ em (2.17) tal que $m\pi = \eta p > 0$, obtemos no limite

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

que é de fato a função de sobrevivência de longa duração do modelo DCP apresentado anteriormente em (2.12). Se colocarmos $m = p = 1$ e $\rho = 0$, o modelo DCB coincide com o modelo de mistura padrão Boag (1949) e Berkson & Gage (1952).

2.2.3 Modelo destrutivo correlacionado binomial negativa (DCBN)

Para a escolha 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

$$\mathbb{P}_{NB}[N=n] = (1+\phi\eta)^{-\phi^{-1}} \sum_{n_1,n_2,\ldots} \frac{\Gamma\left(\phi^{-1}+\sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!}\left[\frac{\phi\eta(1-\rho)}{1+\phi\eta}\right]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{2.18}$$

sendo $n = 0, 1, \ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, \ldots$, tal que $\sum_{i=1}^{\infty} i n_i = n$, e $\Gamma(\cdot)$ denota a função gama. Uma expressão alternativa para a $f.m.p$ em (2.18) é dada por

$$\mathbb{P}_{NB}[N=n] = \begin{cases} (1+\phi\eta)^{-\phi^{-1}} & , \quad n = 0 \\ (1+\phi\eta)^{-\phi^{-1}} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{\Gamma(\phi^{-1}+i)}{\Gamma(\phi^{-1})i!}\left[\frac{\phi\eta(1-\rho)}{1+\phi\eta}\right]^i \rho^{n-i} & , \quad n = 1, 2, \ldots. \end{cases} \tag{2.19}$$

A média e a variância de N são

$$\mathbb{E}[N] = \frac{\eta}{1-\rho} \quad \text{e} \quad \mathbb{V}ar[N] = \frac{\eta(1+\rho+\phi\eta)}{(1-\rho)^2}, \tag{2.20}$$

respectivamente. A $f.g.p$ é dada por

$$\mathbb{A}_N(z) = \left[\frac{1-z\rho}{1+\phi\eta(1-z)-z\rho}\right]^{\phi^{-1}} \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.21}$$

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

$$S_{pop}(y) = \left[\frac{1-\rho(1-pF(y))}{1+\phi\eta pF(y)-\rho(1-pF(y))}\right]^{\phi^{-1}}. \tag{2.22}$$

Quando $\phi = 1$, obtemos a distribuição geométrica inflada com parâmetros $\theta = \frac{1}{1+\eta} \in (0,1)$ em (2.18) ou (2.19), neste caso $S_{pop}(\cdot)$ em (2.22) torna-se

$$S_{pop}(y) = \frac{1-\rho(1-pF(y))}{1+\eta pF(y)-\rho(1-pF(y))}, \tag{2.23}$$

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

2.2.4 Modelo destrutivo correlacionado série logarítmica (DCSL)

Para escolha 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 $p.m.f$ é da forma

$$\mathbb{P}_{LS}[N=n] = (-\log(\pi))^{-1} \sum_{n_1,n_2,\ldots} \frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots}\big[(1-\pi)(1-\rho)\big]^{\sum_{i=1}^{\infty} n_i}\rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{2.24}$$

sendo $n = 0,1,\ldots$, e o somatório é sobre todos inteiros não negativos $n_1, n_2, \ldots$, tal que $\sum_{i=1}^{\infty} in_i = n$. Uma expressão alternativa para a $f.m.p$ em (2.24) é dada por

$$\mathbb{P}_{LS}[N=n] = (-\log(\pi))^{-1} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{\big[(1-\pi)(1-\rho)\big]^i \rho^{n-i}}{i}, \quad n = 1,2,\ldots. \tag{2.25}$$

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 logarítima inflada modificada, cuja $f.m.p$ pode ser expressa como

$$\mathbb{P}_{LS}[N=n] = (-\log(\pi))^{-1} \sum_{i=1}^{n+1} \binom{n}{i-1} \frac{\left[(1-\pi)(1-\rho)\right]^i \rho^{n+1-i}}{i}, \quad n = 0, 1, 2, \ldots. \tag{2.26}$$

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

$$\mathbb{E}[N] = 1 - \frac{1-\pi}{\pi(1-\rho)\log(\pi)} \quad \text{e} \quad \mathbb{V}ar[N] = -\frac{(1-\pi)\left[\log(\pi)(1+\pi\rho)+1-\pi\right]}{\pi^2(1-\rho)^2(\log(\pi))^2}, \tag{2.27}$$

respectivamente. A $f.g.p$ é dada por

$$\mathbb{A}_N(z) = \frac{(-\log(\pi))^{-1}}{z} \log\left\{\frac{1-\rho z}{1-z(1-\pi(1-\rho))}\right\} \quad \text{para} \quad 0 \leq z \leq 1. \tag{2.28}$$

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

$$S_{pop}(y) = \frac{(-\log(\pi))^{-1}}{(1-pF(y))} \log\left\{\frac{1-\rho(1-pF(y))}{1-(1-pF(y))(1-\pi(1-\rho))}\right\}. \tag{2.29}$$

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

2.3 Inferência

2.3.1 Estimação de máxima verossimilhança

Para a formulação da função de verossimilhança considera-se as seguintes notações: Seja N_j, o número de células iniciadas (causas ou riscos) relacionada a ocorrência do câncer (evento de interesse), no j-ésimo indíviduo, $j = 1, 2, \ldots, m$, variáveis aleatórias independentes não observadas com distribuição de probabilidade SPGI com parâmetros θ e ρ. Seja D_j dado $N_j = n_j$, o número de células iniciadas não eliminadas pelo tratamento, no j-ésimo indíviduo, $j = 1, 2, \ldots, m$, 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 não eliminadas no j-ésimo indivíduo, com função distribuição indicada por $F(t_j|\boldsymbol{\gamma}) = 1 - S(t_j|\boldsymbol{\gamma})$ (no contexto bayesiano) ou $F(t_j;\boldsymbol{\gamma}) = 1 - S(t_j;\boldsymbol{\gamma})$ (no contexto clássico) e $\mathbb{P}[V_{j0} = \infty] = 1$, sendo que $\boldsymbol{\gamma}$ representa o vetor de parâmetros da distribuição. Seja Y_j como definido em (2.5) 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, os modelos DCP, DCB e DCBN das seções 2.2.1,2.2.2 e 2.2.3 não são identificáveis no sentido de Li *et al.* (2001), isto é, existem θ e θ^*, $\theta \neq \theta^*$, tais que $S_{pop}(y;\theta) = S_{pop}(y;\theta^*)$. Para evitar este problema, propomos relacionar os parâmetros p e η(ou π) dos modelos DCP, DCB e DCBN com os vetores de covariáveis $\boldsymbol{x'_j} = (x_{j1}, \ldots, x_{jk_1})$ e $\boldsymbol{w'_j} == (w_{j1}, \ldots, w_{jk_1})$, respectivamente, sem elementos comuns. Adotemos as funções de ligação

$$\log\left(\frac{p_j}{1-p_j}\right) = \boldsymbol{x'_j}\boldsymbol{\beta_1}, \quad \text{e} \quad \log(\eta_j) = \boldsymbol{w'_j}\boldsymbol{\beta_2} \quad \text{ou} \quad \log\left(\frac{\pi_j}{1-\pi_j}\right) = \boldsymbol{w'_j}\boldsymbol{\beta_2}, \quad j = 1, \ldots, m, \tag{2.30}$$

sendo $\boldsymbol{\beta'_1} = (\beta_{11}, \ldots, \beta_{1k_1})$ e $\boldsymbol{\beta'_2} = (\beta_{21}, \ldots, \beta_{2k_2})$ vetores com k_1 e k_2 coeficientes de regressão.

Uma questão crítica é a seleção de covariáveis a serem incluídas nas funções de ligação em (2.30). Mais precisamente, dada uma função de ligação e um conjunto de potenciais covariáveis, o problema é encontrar e adaptar o "melhor"modelo em um subconjunto "selecionado"de co-

variáveis. Infelizmente, este problema não será abordado aqui, pois no nosso exemplo aplicado as covariáveis são apenas selecionadas para resolver o problema de identificabilidade. 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 do processo destrutivo.

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*

$$f(t_j, \delta_j | n_j, d_j) = \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} I_{\{d_j \leq n_j\}}, \tag{2.31}$$

sendo I_A *a função indicadora do evento* A.

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

Em seguida apresentamos a função verossimilhança do processo destrutivo.

Teorema 2.1 *Supondo um processo destrutivo com censura não-informativa, a função de verossimilhança completa é dada por*

$$L(\boldsymbol{\vartheta}; \boldsymbol{D}_c) = \prod_{j=1}^{m} \big\{S(t_j; \boldsymbol{\gamma})\big\}^{d_j - \delta_j} \big\{d_j f(t_j; \boldsymbol{\gamma})\big\}^{\delta_j} \mathbb{P}[N_j = n_j, D_j = d_j] \tag{2.32}$$

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

Prova 2.2

$$f(\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{n}, \boldsymbol{d}) = \prod_{j=1}^{m} f(t_j, \delta_j, n_j, d_j) = \prod_{j=1}^{m} f(t_j, \delta_j | n_j, d_j) \mathbb{P}[N_j = n_j, D_j = d_j]$$

sendo $\boldsymbol{n}' = (n_1, \ldots, n_m)$ *e* $\boldsymbol{d}' = (d_1, \ldots, d_m)$. *O resultado segue diretamente de (2.31).*

Note que a função de verossimilhança (2.32) depende de $\boldsymbol{N}$ e $\boldsymbol{D}$ que são variáveis latentes. A função de verossimilhança marginal é dada por

$$\begin{aligned} L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) &= \prod_{j=1}^{m} \sum_{n_j=0}^{\infty} \sum_{d_j=0}^{n_j} \big\{S(t_j;\boldsymbol{\gamma})\big\}^{d_j-\delta_j} \big\{d_j f(t_j;\boldsymbol{\gamma})\big\}^{\delta_j} \mathbb{P}[N_j = n_j, D_j = d_j] \\ &= \prod_{j=1}^{m} \sum_{d_j=0}^{\infty} \big\{S(t_j;\boldsymbol{\gamma})\big\}^{d_j-\delta_j} \big\{d_j f(t_j;\boldsymbol{\gamma})\big\}^{\delta_j} \sum_{n_j=0}^{\infty} \cdots \sum_{n_j=d_j}^{\infty} \mathbb{P}[N_j = n_j, D_j = d_j] \\ &= \prod_{j=1}^{m} \sum_{d_j=0}^{\infty} \big\{S(t_j;\boldsymbol{\gamma})\big\}^{d_j-\delta_j} \big\{d_j f(t_j;\boldsymbol{\gamma})\big\}^{\delta_j} \mathbb{P}[D_j = d_j]. \end{aligned} \tag{2.33}$$

Recentemente De Castro *et al.* (2007) mostrou que a função de verossimilhança em (2.33) pode ser expressada pela seguinte expressão

$$L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) = \prod_{j=1}^{m} \big\{f_{pop}(t_j;\boldsymbol{\gamma})\big\}^{\delta_j} \big\{S_{pop}(t_j;\boldsymbol{\gamma})\big\}^{1-\delta_j}. \tag{2.34}$$

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

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

$$F(v;\boldsymbol{\gamma}) = 1 - \exp(-v^{\gamma_1} e^{\gamma_2}) \quad \text{e} \quad f(v;\boldsymbol{\gamma}) = \gamma_1 v^{\gamma_1 - 1} \exp(\gamma_2 - v^{\gamma_1} e^{\gamma_2}) \tag{2.35}$$

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 podem 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 vida. Sua taxa de falha 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; ver Johnson *et al.* (1994).

As estimativas de máxima verossimilhança de $\widehat{\boldsymbol{\vartheta}}$ são obtidas maximizando o logaritmo da função de verossimilhança em (2.34), $\ell(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) = log(L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}))$. A maximização é efetuada numericamente aplicando o método L-BFGS-B, implementado na função *optim* do sistema R (R Development Core Team, 2011). O programa computacional pode ser obtido mediante a solicitação ao autor. Sob certas condições de regularidades, 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

$$\boldsymbol{I(\vartheta)} = \mathbb{E}\left(-\frac{\partial^2 \log L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}\right) \tag{2.36}$$

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

Alternativamente, 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 $\widehat{\boldsymbol{\vartheta}_g}$ é a estimativa de máxima verossimilhança sobre o modelo g, q é o número de parâmetros estimados sobre o modelo g, and 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 considerando distribuições *a priori* próprias (Ibrahim *et al.*, 2001a). Embora não seja necessário, por simplicidade, assumiremos que os parâmetros $\boldsymbol{\beta_1'}$, $\boldsymbol{\beta_2'}$, γ_1, γ_2, ρ, ϕ e m são independentes *a priori*, isto é,

$$\pi(\boldsymbol{\vartheta}) = \prod_{j_1=1}^{k_1} \pi(\beta_{1j_1}) \prod_{j_2=1}^{k_2} \pi(\beta_{2j_2})\pi(\gamma_1)\pi(\gamma_2)\pi(\rho)\pi(\phi)\pi(m), \tag{2.37}$$

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 e $m \sim Uniforme\ discreta\big\{1, 2, \ldots, m_0\big\}$ para o modelo DCB. Todos os hiperparâmetros são especificados, a fim de garantir distribuições *a priori* vaga, sendo o valor de m_0 escolhido estritamente maior do que o estimador de máximo verossimilhança do parâmetro m.

Combinando a função de verossimilhança (2.34) com a distribuição *a priori* em (2.37), a distribuição *a posteriori* para $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \rho, \phi, m)$ 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 e o algoritmo de Metropolis-Hastings; ver Gamerman & Lopes (2006). Para a implementação do algoritmo são necessárias as distribuições condicionais completas *a posteriori* de todos os parâmetros, dadas por

$$\pi(\boldsymbol{\beta_1}|\boldsymbol{\beta_2}, \gamma_1, \gamma_2, \rho, \phi, m, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_1}), \quad \pi(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, \gamma_1, \gamma_2, \rho, \phi, m, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_2}),$$

$$\pi(\gamma_1|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \gamma_2, \rho, \phi, m, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\boldsymbol{\gamma_1}), \quad \pi(\gamma_2|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \gamma_1, \rho, \phi, m, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\gamma_2),$$

$$\pi(\rho|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \gamma_1, \gamma_2, \phi, m, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\rho), \quad \pi(\phi|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \gamma_1, \gamma_2, \rho, m, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(\phi),$$

e $\pi(m|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \gamma_1, \gamma_2, \phi, \boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})\pi(m)$. Todas estas distribuições condicionais não são avaliadas de forma fechada. Então faremos uso do algoritmo Metropolis-Hasting dentro do ciclo do algoritmo de Gibbs para simular amostras de $\boldsymbol{\vartheta}$. Tal algoritmo nos permite simular amostras de distribuições conjuntas, utilizando as distribuições condicionais completas dos parâmetros desconhecidos, como mostra o esquema a seguir:

1. Inicie com $\boldsymbol{\vartheta^{(0)}} = (\boldsymbol{\beta_1'^{(0)}}, \boldsymbol{\beta_2'^{(0)}}, \boldsymbol{\gamma'^{(0)}}, \rho^{(0)}, \phi^{(0)}, m^{(0)})$.

2. Gere ρ^* da distribuição *a priori* $\pi(\rho)$ descrita anteriormente.

3. Gere um valor u da distribuição uniforme $U(0, 1)$.

4. Se $u \leq \min\left\{1, \frac{\pi(\rho^*|\boldsymbol{\beta_1'^{(0)}}, \boldsymbol{\beta_2'^{(0)}}, \boldsymbol{\gamma'^{(0)}}, \phi^{(0)}, m^{(0)})}{\pi(\rho^{(0)}|\boldsymbol{\beta_1'^{(0)}}, \boldsymbol{\beta_2'^{(0)}}, \boldsymbol{\gamma'^{(0)}}, \phi^{(0)}, m^{(0)})}\right\}$, então atualize $\rho^{(1)}$ por ρ^*, caso contrário, permaneça com $\rho^{(0)}$, ou seja, $\rho^{(1)} = \rho^{(0)}$.

5. Proceda analogamente para obter $\boldsymbol{\gamma^{(1)}} = (\gamma_1^{(1)}, \gamma_2^{(1)})$, $\phi^{(1)}$, $m^{(1)}$, $\beta_{1j_1}^{(1)}$, $j_1 = 1, \ldots, k_1$ e $\beta_{2j_2}^{(1)}$, $j_2 = 1, \ldots, k_2$.

6. Repita os passos de 2 a 5 até obter uma amostra de uma distribuição estacionária.

O código computacional foi implementado no sistema OpenBUGS 3.0.3 (Thomas *et al.*, 2006).

2.3.3 Critério para comparação de modelos

Existem uma variedade de metodologias para comparar vários modelos competindo para o 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 densidade preditiva condicional ordinária (CPO); ver Gelfand *et al.* (1992). Denotamos $\boldsymbol{D}_{obs}^{(-j)}$ os dados observados com a j-ésima observação deletada. 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, $g(t_j;\boldsymbol{\vartheta}) = S_{pop}(t_j;\boldsymbol{\vartheta})$, sendo $f_{pop}(\cdot)$ 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 se expresso como

$$CPO_j = \int_\Theta g(t_j;\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs}^{(-j)})d\boldsymbol{\vartheta} = \left\{\int_\Theta \frac{\pi(\boldsymbol{\vartheta}|\boldsymbol{D}_{obs})}{g(t_j;\boldsymbol{\vartheta})}d\boldsymbol{\vartheta}\right\}^{-1}. \tag{2.38}$$

O modelo escolhido é que apresenta o maior valor CPO_j (em média). Para o modelo proposto, uma forma fechada do CPO_j não está disponível. No entanto, uma estimativa Monte Carlo da CPO_j pode ser obtida através de uma simples 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

$$\widehat{CPO_j} = \left\{\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{g(t_j;\boldsymbol{\vartheta}_q)}\right\}^{-1}. \tag{2.39}$$

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 os critérios com base na média *a posteriori* do *deviance*, que é em si uma medida de ajuste. O *deviance* pode ser aproximado por $\overline{D} = \sum_{q=1}^{Q}\frac{D(\boldsymbol{\vartheta}_q)}{Q}$, sendo $D(\boldsymbol{\vartheta}) = -2\sum_{j=1}^{m}\log\big(g(t_j;\boldsymbol{\vartheta})\big)$. Entre esses critérios, nós escolhemos o critério de informação *deviance* (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 $\widehat{DIC} = \overline{D} + \widehat{\rho}_D = 2\overline{D} - \widehat{D}$, sendo ρ_D o número efetivo de parâmetros definido como $\mathbb{E}[D(\boldsymbol{\vartheta})] - D(\mathbb{E}[\boldsymbol{\vartheta}])$, sendo $D(\mathbb{E}[\boldsymbol{\vartheta}])$ o *deviance* avaliado na

média *a posteriori*, que pode ser estimado por

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

Da mesma forma, o *EAIC* e *EBIC* podem também ser estimados utilizando as amostras MCMC por meio de $\widehat{EAIC} = \overline{D} + 2\#(\boldsymbol{\vartheta})$ e $\widehat{EBIC} = \overline{D} + \#(\boldsymbol{\vartheta})\log(m)$, sendo $\#(\boldsymbol{\vartheta})$ o número de parâmetros do modelo. 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 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, sendo 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, demonstramos uma aplicação dos modelos descritos na seção 2.2 em um conjunto de dados de melanoma maligno, que foi coletado no hospital universitário de Odense (dinamarca) por K. T. Drzewiecki. O conjunto de dados inclui 205 pacientes observados após operação 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 (T) 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 estado 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 identificação mencionada anteriormente na seção 2.3, nos modelos DCP, DCB e DCBN, o parâmetro p é ligada apenas a espessura do tumor, enquanto que o parâmetro η(ou π) está ligada apenas ao estado de úlcera.

Antes de ajustarmos os modelos devemos identificar o comportamento da função de risco dos tempos observados. Para isso utilizamos um método gráfico baseado no teste do tempo total (TTT) (Aarset, 1985). Na sua versão empírica o gráfico TTT é dado por $G(r/n) =$

$[(\sum_{j=1}^{r} Y_{j:n}) + (n-r)Y_{r:n}]/(\sum_{j=1}^{r} Y_{j:n})$, sendo $r = 1, \ldots, n$ e $Y_{j:n}$ representam as estatísticas de ordem da amostra. É provado que a função de risco cresce (decresce) se o gráfico TTT é côncavo (convexo), quando se aproxima de uma linha diagonal é constante e, se primeiramente sua curvatura é côncava e depois convexa, seu risco cresce e depois decresce. Embora o gráfico TTT seja apenas uma condição suficiente e não necessária para indicar o formato da função de risco, será utilizado como um indicador de seu comportamento. A Figura 2.2 (painel esquerdo) apresenta o gráfico TTT dos dados de câncer de melanoma que indica uma função de risco crescente, podendo então ser representada pela distribuição Weibull. A Curva Kaplan-Meier estratificada pelo estado de úlcera (ulc) na Figura 2.2 (painel direito) nivela 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.

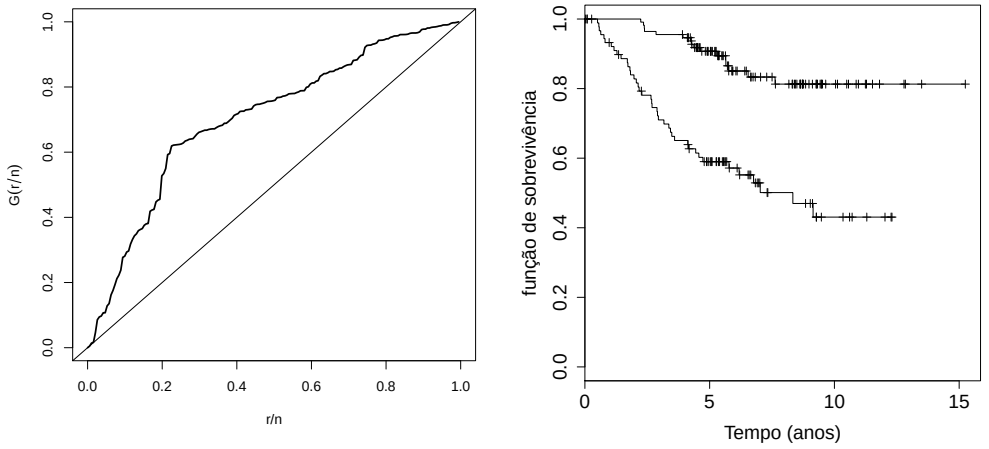

Figura 2.2: Painel esquerdo: gráfico TTT. Painel direito: curva Kaplan-Meier estratificada pelo estado de úlcera (superior: presente, inferior: ausente).

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. A Tabela 2.3 apresenta os valores do máximo da log-verossimilhança, $\max \log L(\cdot)$, e os valores das estatísticas AIC e BIC

para os modelos ajustados. As estatísticas AIC e BIC dão evidências a favor do modelo DCG. Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo DCG, seus desvios padrão e seus intervalos de confiança 95% baseados na teoria assintótica são apresentados na Tabela 2.4. A estimativa do parâmetro correlação ρ é 0,95, e como mencionado anteriormente na Seção 2.1, isso indica uma alta associação entre as células.

Tabela 2.3: Os valores do $\max \log L(\cdot)$ e as estatísticas AIC e BIC para os sete modelos ajustados: DCP, DCB, DCBN, DCG, DCSL, binomial negativa e geométrico.

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 2.4: 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)
$\beta_{2,ulc:presente}$	0,53	0,30	(-0,06 ; 1,12)
$\beta_{2,ulc:ausente}$	-0,48	0,41	(-1,28 ; 0,32)

A Figura 2.3 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 é vista a diminuir mais rapidamente para os pacientes com tumores mais espessos. Na Figura 2.3 (a), a função de sobrevivência não passa abaixo de 0,7.

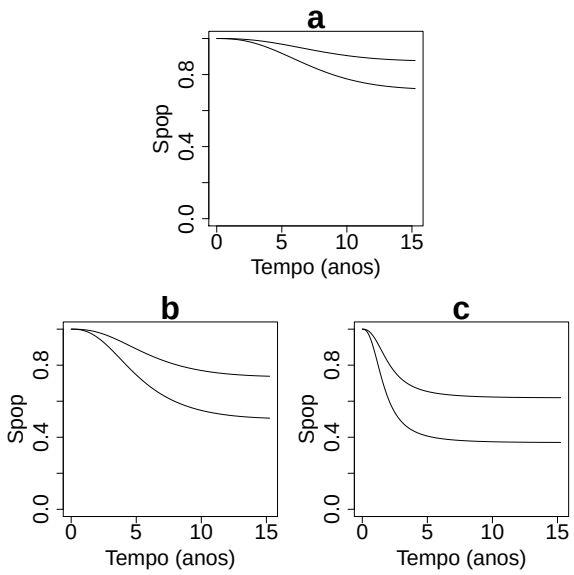

Figura 2.3: Função de sobrevivência sob o modelo DCG estratificado pelo estado de úlcera (superior: ausente, inferior: presente) para pacientes com espessura do tumor igual a (a) 0.32, (b) 1.94, e (c) 8.32 mm, respectivamente.

O modelo DCG foi ajustado com os parâmetros p e η associados à espessura do tumor e o estado de úlcera, respectivamente. Se trocarmos essas covariáveis, não há melhora no ajuste com relação aos critérios na Tabela 2.3, uma vez que, neste caso, obtermos os valores do ($\max \log L(\cdot)$; AIC; BIC) a ser (-204,61; 423,23; 446,49).

Finalmente, voltamos a nossa atenção para o papel das covariáveis sobre a fração de cura (ver Tabela 2.2). As estimativas dos coeficientes $\beta_{2,ulc}$ na Tabela 2.4 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 $\hat{\beta}_{2,espessura} > 0$ na Tabela 2.4, os valores maiores da espessura do tumor implica em uma menor estimativa da fração de cura. A Figura 2.4 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 estado de úlcera ausente e presente, respectivamente.

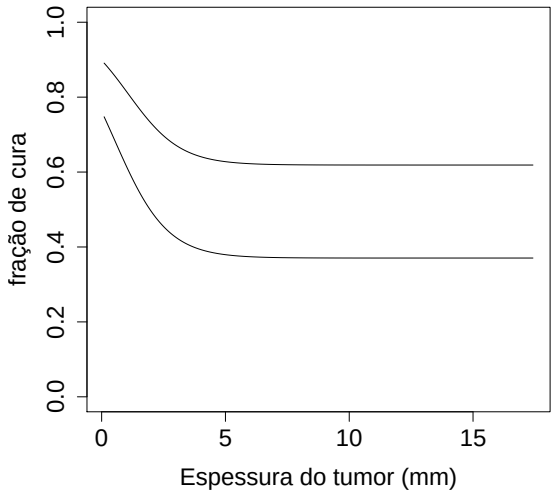

Figura 2.4: Fração de cura para o modelo DCG *versus* espessura do tumor estratificada pelo estado de úlcera (superior: ausente, inferior: presente).

Também obtemos os ajustes para os sete modelos da Tabela 2.3 através da inferência bayesiana. Utilizamos distribuições *a priori* independentes e não informativas, sendo $\beta_{1,intercepto} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{1,espessura} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{2,ulc:ausente} \sim \mathcal{N}\big(0, 10^3\big)$, $\beta_{2,ulc:presente} \sim \mathcal{N}\big(0, 10^3\big)$, $\gamma_1 \sim Gama\big(1, 0, 01\big)$, $\gamma_2 \sim \mathcal{N}\big(0, 10^3\big)$ e $\rho \sim Beta\big(1, 1\big)$, enquanto que $\phi \sim Gama\big(1, 0, 01\big)$ para o modelo DCBN e $m \sim Uniforme\ discreta\big\{1, 2, \ldots, 10^3\big\}$ para o modelo DCB. 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 2.5 foi aplicado os critérios de seleção de modelos definidos na seção 2.3.3 para os sete modelos ajustados: DCP, DCB, DCBN, DCG, DCSL, binomial negativa e geométrico. Os critérios dão evidências a favor do modelo DCG. A Tabela 2.6 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 2.5 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 2.5.

Tabela 2.5: As estatísticas DIC, EAIC, EBIC e B para os sete modelos ajustados: DCP, DCB, DCBN, DCG, DCSL, binomial negativa e geométrico.

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 2.6: Médias *a posteriori*, os desvios padrão e os intervalos de credibilidade 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

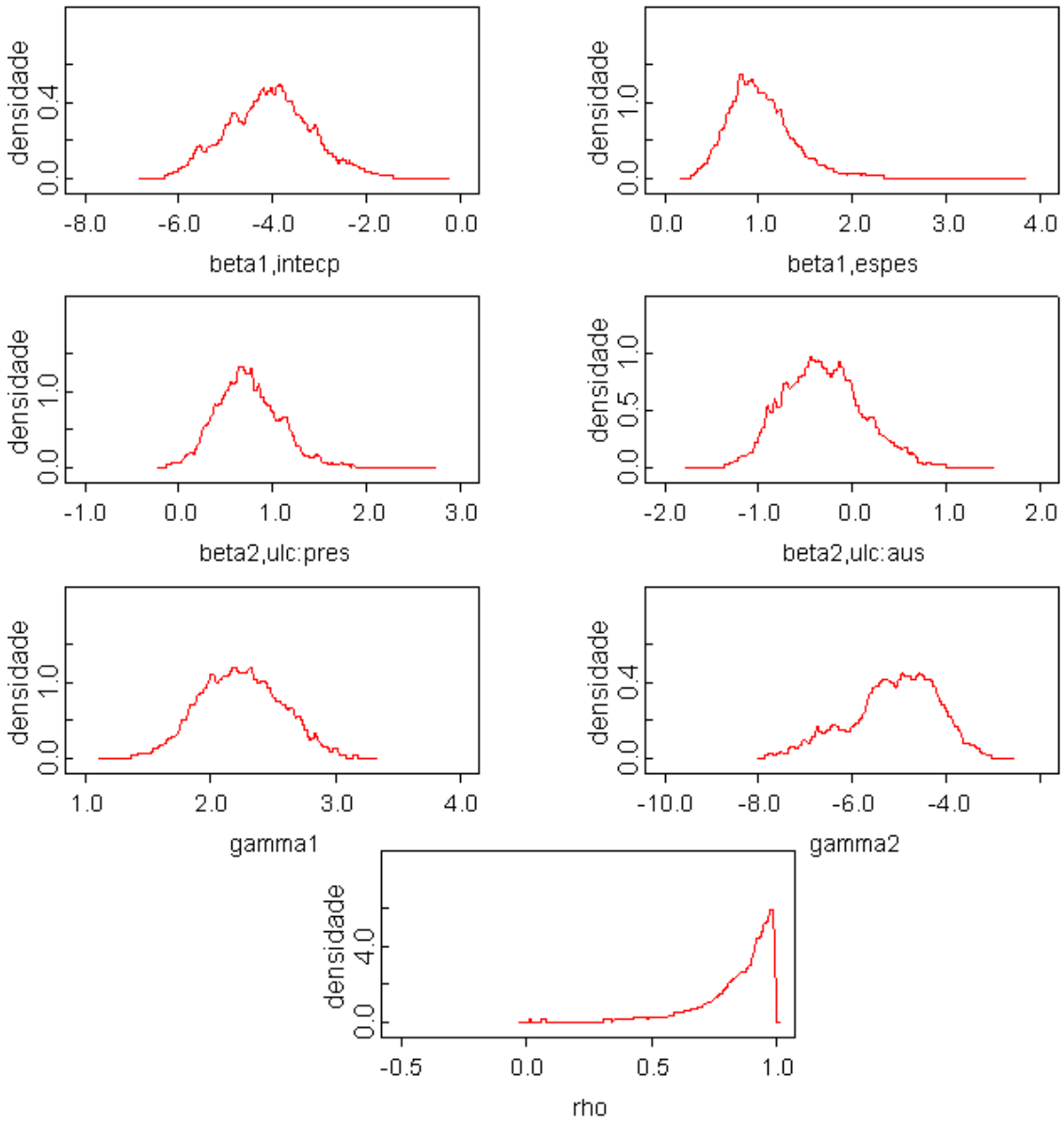

Figura 2.5: Densidades *a posteriori* marginais aproximadas dos parâmetros.

2.5 Comentários finais

Neste Capítulo, propomos um novo modelo de sobrevivência com fração de cura, que estende o modelo de Rodrigues *et al.* (2010, 2011), no sentido de incoporamos uma estrutura de dependência entre as células iniciadas (causas competitivas). Assumimos uma distribuição SPGI para

o número de células iniciadas e uma distribuição Weibull para os tempos de ocorrência do tumor, obtendo o modelo DCSPG. O modelo DCSPG incorpora na análise uma dependência biológica entre as células do tumor. A vantagem desta suposição é que podemos medir a interdependência entre as células de um tecido iniciado desenvolvendo um tumor maligno. 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. A relevância prática e a aplicabilidade do modelo foram demonstradas em um conjunto de dados reais de pacientes com câncer de melanoma. O modelo proposto, além de oferecer melhor interpretação para o mecanismo biológico da carcinogênese, oferece melhor ajuste do que os outros modelos de fração de cura comumente utilizados.

Capítulo 3

Modelo com fração de cura híbrido

Os modelos de sobrevivência de dois estágios para carcinogênese foca sobre eventos que precedem a ocorrência da primeira célula maligna em um tecido. Uma descrição explícita do estágio de progressão do tumor é evitado em modelos de dois estágios. Isto 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). Para esse fim, um modelo de sobrevivência geral para carcinogêneses espontânea baseado em um esquema híbrido latente de ativação para as células combinando o esquema de ativação máximo dentro de outro esquema de ativação mínimo (Cooner *et al.*, 2007) foi desenvolvido para permitir um padrão simples da dinâmica do crescimento do tumor. Assumimos o número de células iniciadas e o número de células malignas (causas competitivas) seguem distribuições Poisson ponderada. Supõe-se que o tumor (é monoclonal gerado durante o estágio de progressão) se tornar detectável quando seu tamanho atinge um certo nível limiar (proliferação de células tumorais geradas da célula maligna). A vantagem deste modelo é que ele incorpora dentro da análise 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, um conjunto de dados de câncer melanoma real ilustra a utilidade do modelo proposto. Comentários finais são apresentados na Seção 3.5.

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. Pensamos nessas 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, mas algumas delas ainda não serão reconhecidas pelo sistema de reparo ou tratamento não funcionou e, conseqüentemente, não reparadas;

(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 $F(\cdot) = 1 - S(\cdot)$, sendo $S(\cdot)$ função de sobrevivivência. Todas 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 feitas pela maioria dos modernos modelos de sobrevivência em dois estágios encontrados na literatura, ver por exemplo Rodrigues* et al. *(2009b); Chen* et al. *(1999a) e Cooner* et al. *(2007).*

Com base nas suposições acima, o modelo proposto pode ser desenvolvido da seguinte maneira. Para qualquer paciente 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 um conseqüência 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_3 o número de células tumorais originadas de cada célula maligna com com $f.m.p$ $p_{n_3} = \mathbb{P}[N_3 = n_3]$ para $n_3 = 0, 1, \ldots$. O tempo que a (i, j)-ésima célula maligna transformar-se 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 $X'_{ij}s$ são independentes com função distribuição $F(y) = 1 - S(y)$, independentes de N_k.

No cenário de causas competitivas (Cox & Oakes, 1984), o número de células iniciadas, malignas, tumorais e o tempo Z_{ij} são não observáveis (variáveis latentes). Assim, o tempo de

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

$$Y = \min\left\{\max\left\{Z_{ij}\right\}_{j=0}^{N_3}\right\}_{i=0}^{N_2}, \tag{3.2}$$

sendo $\mathbb{P}[Z_{0j} = \infty] = 1$, o que leva uma proporção p_0 da população não susceptível à ocorrência do tumor, também chamada de "fração de cura", e $\mathbb{P}[Z_{i0} = \infty] = 1$, o que leva 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 máximo dentro do esquema de ativação mínimo (ver 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.*

Em seguida, formulamos uma definição que será fundamental para encontramos a função de sobrevivência de longa duração da variável Y, a qual inclui os indivíduos que não são suceptíveis a ocorrência do tumor.

Definição 3.1 *Seja $a = \{a_n\}$ uma seqüência de números reais. Se*

$$\mathbb{A}_a(s) = a_0 + a_1 s + a_2 s^2 + \ldots = \sum_{n=0}^{\infty} a_n s^n$$

converge para valores de s no intervalo $[0, 1]$, então $\mathbb{A}_a(s)$ é definida como a função geradora da sequência (f.g.p) a (Feller, 1968).

Se $a = \{a_n\} = \{p_n\} = p$, temos que $\mathbb{A}_p(s) = E[s^N]$. Observe que a função geradora de probabilidades é a função geradora de momentos da variável aleatória N calculada no ponto $\log(s)$, isto é, $\mathbb{A}_p(s) = E[\exp\{\log(s)N\}]$. Mais informações sobre esta definição, sua relevância probabilística e sua expressão para algumas distribuições são encontradas no Capítulo XI de Feller (1968).

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 própria (suposição (iv)), $S(y) = 1 - F(y)$, dos tempos de progressão não observáveis em (3.2), a função de sobrevivência da variável aleatória Y em (3.2) é dada por*

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

sendo $\mathbb{A}_{p_{n_1}}(\cdot)$, é a f.g.p da seqüência $\{p_{n_1}\}$, que converge se $s = 1 - p\big(1 - S^*_{pop}(y)\big) \in [0,1]$, e

$$S^*_{pop}(y) = 1 + \mathbb{P}[N_3 = 0] - \mathbb{A}_{p_{n_3}}\Big(F(y)\Big), \tag{3.5}$$

que denotamos como a função de sobrevivência do estágio de progressão, em que $\mathbb{A}_{p_{n_3}}(\cdot)$, é a f.g.p da seqüência $\{p_{n_3}\}$, que converge se $s = F(y) \in [0,1]$.

Prova 3.1 *Temos que*

$$\begin{aligned}
S_{pop}(y) &= \sum_{l=0}^{\infty}\Big\{\mathbb{P}[N_2 = 0|N_1 = l] + \mathbb{P}[\max\{Z_{1j}\}_{j=0}^{N_3} > y, \ldots, \max\{Z_{N_2 j}\}_{j=0}^{N_3} > y; 1 \le N_2 \le l]\Big\}\mathbb{P}[N_1 = l] \\
&= \sum_{l=0}^{\infty}\Big\{\sum_{i=0}^{l}\big\{\mathbb{P}[\max\{Z_{ij}\}_{j=0}^{N_3} > y]\big\}^i \mathbb{P}[N_2 = i|N_1 = l]\Big\}\mathbb{P}[N_1 = l] \\
&= \sum_{l=0}^{\infty}\Big\{\sum_{i=0}^{l}\big\{1 - \mathbb{P}[Z_{i1} < y, \ldots, Z_{iN_3} < y; N_3 \ge 1]\big\}^i \mathbb{P}[N_2 = i|N_1 = l]\Big\}\mathbb{P}[N_1 = l] \\
&= \sum_{l=0}^{\infty}\Big\{\sum_{i=0}^{l}\big\{1 - \underbrace{\sum_{j=1}^{\infty} F(y)^j \mathbb{P}[N_3 = j]}_{\mathbb{A}_{p_{n_3}}\left(F(y)\right) - \mathbb{P}[N_3 = 0]}\big\}^i \mathbb{P}[N_2 = i|N_1 = l]\Big\}\mathbb{P}[N_1 = l] \\
&\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\left\{1 - p + p\left(1 + \mathbb{P}[N_3 = 0] - \mathbb{A}_{p_{n_3}}(F(y))\right)\right\}^l} \\
&= \sum_{l=0}^{\infty}\Big\{1 - p + p\Big(1 + \mathbb{P}[N_3 = 0] - \mathbb{A}_{p_{n_3}}(F(y))\Big)\Big\}^l \mathbb{P}[N_1 = l] \\
&= \mathbb{A}_{p_{n_1}}\Big(1 - p + pS^*_{pop}(y)\Big) \\
&= \mathbb{A}_{p_{n_1}}\Big(1 - p\big(1 - S^*_{pop}(y)\big)\Big)
\end{aligned} \tag{3.6}$$

A última expressão sintetiza de forma simples e objetiva os três estágios do processo da carcinogênise através de uma composição da função geradora de probilidade 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, $S(y) = 1 - F(y)$, então*

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

em que p_0 denota a proporção de indivíduos "curados"ou "imunes"que podem estar presentes na populaçao a partir do qual os dados são obtidos, e p_0^* denota a proporção de células iniciadas 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.*

Devido aos resultados em (3.4) e (3.7), definimos p_0 como a taxa de sobrevivência de longa duração ou fração de cura e $S_{pop}(y)$ como a função de sobrevivência de longa duração do processo da carcinogênise.

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 introduzido por Rodrigues* et al. *(2010, 2011).*

Agora, vamos supor que o número de células iniciadas, N_1, e número de células tumorais, N_3, seguem distribuições Poisson ponderada 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 $w(\cdot\,;\,\phi_k)$ 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 a média é tomada com relação

à variável N_k seguindo uma distribuição de Poisson com média η_k. Vamos denotar a distribuição Poisson ponderada em (3.8) por $PP_{\eta_k}(w_k)$, o que denota a distribuição Poisson ponderada com parâmetro η_k e função peso $w_k(\cdot\ ;\ \phi_k)$. Este conceito foi introduzido por Fisher (1934), mas foi Rao (1965) que estudou as distribuiçõess ponderadas em um caminho unificado. Ele destacou que em muitas situações, as observações registradas não podem ser considerados como uma amostra aleatória da distribuição original, por muitas razões, tais como não-observabilidade 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 comprimento tendencioso, e tem encontrado muitas aplicações importantes em biometria e meio ambiente.

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

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

Levando em conta (3.8) e (3.9), a função de sobrevivência de longa duração é expressa pelo 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

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

sendo $p_3(0) = w(0;\phi_3)e^{-\eta_3}/\mathbb{E}_{\eta_3}[w(N_3;\phi_3)]$. 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 o modelo HPPPP em termos de um diagrama.

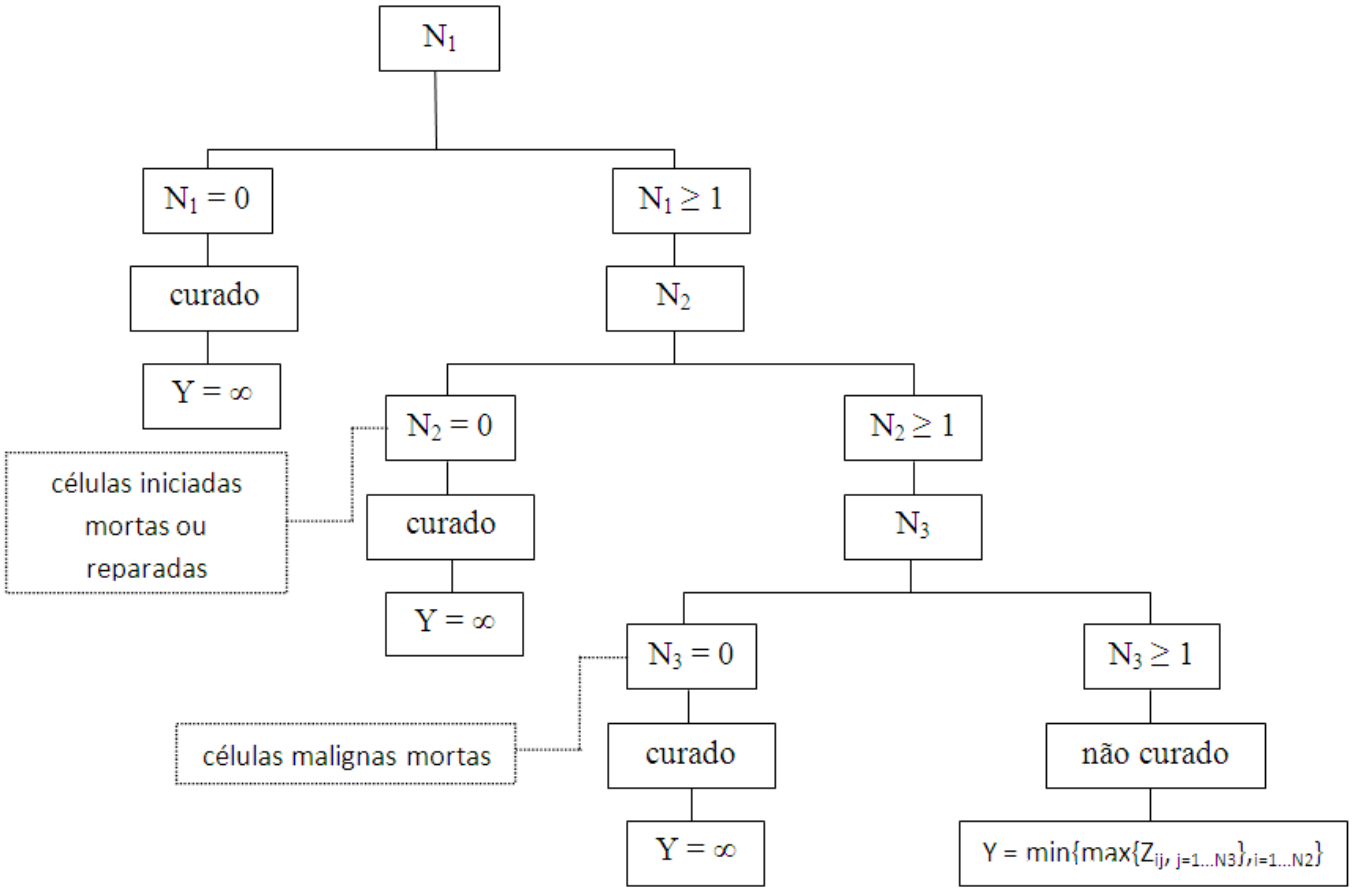

Figura 3.1: Representação do modelo proposto HPPPP em termos de um diagrama.

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, $w(n_1;\phi_1) = \exp\{n_1\phi_1\}$, então N_1 segue uma distribuição Poisson ponderada exponencialmente com parâmetros η_1 e ϕ_1, e sua $f.m.p$ é dada por

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

para $\eta_1,\ \phi_1 > 0$. 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 $\eta_3 > 0$; neste caso, a função peso é simplesmente constante. 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, sendo uma variável com distribuição binomial negativa com parâmetros ϕ_1 e η_1 (ver Piegorsch (1990) e 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 conhecidas por serem

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

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 $-1/\kappa$ fornece 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 fornece 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), se $-1/\eta_1 < \phi_1 < 0$, 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 céluas.

Seja o número de células tumorais, N_3, ser 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

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

O modelo 3.18 é não identificável (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)

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

$$p_1(n_1; \eta_1, \phi_1) = \frac{1}{Z(\eta_1, \phi_1)} \frac{\eta_1^{n_1}}{(n_1!)^{\phi_1}}, \quad n_1 = 0, 1, 2, \ldots, \tag{3.19}$$

sendo $Z(\eta_1, \phi_1) = \sum_{j=0}^{\infty} \eta_1^j/(j!)^{\phi_1}$. Em particular, quando $\phi_1 = 0$ e $\eta_1 < 1$, a distribuição COM-Poisson torna-se à distribuição geométrica com parâmetro $1-\eta_1$. A distribuição em (3.19) também pode ser vista 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

$$\mathbb{A}_{N_1}(s) = \frac{Z(\eta_1 s, \phi_1)}{Z(\eta_1, \phi_1)}. \tag{3.20}$$

Para os cálculos realizados na Seção 3.4, o truncamento da série $Z(\eta_1, \phi_1)$ é feita conforme descrito em Rodrigues *et al.* (2009a).

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

$$S_{pop}(y) = \frac{Z(\eta_1\{1 - pe^{-\eta_3}(e^{\eta_3 F(y)-1})\})}{Z(\eta_1, \phi_1)}. \tag{3.21}$$

Na Tabela 3.1, apresentamos a função de sobrevivência de longa duração, a função densidade imprópria $f_{pop}(y) = -S_{pop}(y)/dy$, a fração de cura e a proporção de células malignas que morrem antes da indução do tumor, correspondentes aos casos particulares apresentados nas Seções 3.3.1, 3.3.2 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	$\exp\left\{-\eta_1 e^{\phi_1} p e^{-\eta_3}\left(e^{\eta_3 F(y)} - 1\right)\right\}$	$\eta_1 e^{\phi_1} p e^{-\eta_3} \eta_3 f(y) e^{\eta_3 F(y)} S_{pop}(y)$	$\exp\left\{-\eta_1 e^{\phi_1} p e^{-\eta_3}\left(e^{\eta_3} - 1\right)\right\}$	$e^{-\eta_3}$
HBNP	$\left\{1 + \phi_1 \eta_1 p e^{-\eta_3}\left(e^{\eta_3 F(y)} - 1\right)\right\}^{-1/\phi_1}$	$\frac{\eta_1 p \eta_3 e^{-\eta_3} e^{\eta_3 F(y)}}{1+\phi_1 \eta_1 p e^{-\eta_3}\left(e^{\eta_3 F(y)} - 1\right)} S_{pop}(y)$	$\left\{1 + \phi_1 \eta_1 p e^{-\eta_3}\left(e^{\eta_3} - 1\right)\right\}^{-1\phi_1}$	$e^{-\eta_3}$
HCPP	$\frac{Z(\eta_1\{1-pe^{-\eta_3 F(y)}-1\},\phi_1)}{Z(\eta_1,\phi_1)}$	$\frac{p\eta_3 e^{-\eta_3} f(y) e^{\eta_3 F(y)}}{\left(1-pe^{-\eta_3}(e^{\eta_3 F(y)}-1)\right) Z(\eta_1,\phi_1)} \sum_{j=1}^{\infty} \frac{j[\eta_1\{1-pe^{-\eta_3}(e^{\eta_3 F(y)}-1)\}]^j}{(j!)^{\phi_1}}$	$\frac{Z(\eta_1\{1-pe^{-\eta_3}-1\},\phi_1)}{Z(\eta_1,\phi_1)}$	$e^{-\eta_3}$

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 considera-se as seguintes notações: seja $\boldsymbol{N} = (N_{1j}, N_{2j}, N_{3j})$ um vetor de variáveis aleatórias latentes, sendo 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 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 $F(t_j|\boldsymbol{\gamma}) = 1 - S(t_j|\boldsymbol{\gamma})$ (ou $F(t_j;\boldsymbol{\gamma}) = 1 - S(t_j;\boldsymbol{\gamma})$) 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 $\mathbb{P}[Z_{0hj} = \infty] = \mathbb{P}[Z_{i0j} = \infty] = 1$. $\boldsymbol{\gamma}$ representa o vetor de parâmetros da distribuição. 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, os modelos HPPEP e HBNP das seções 3.3.1 e 3.3.2 não são identificáveis no sentido de Li *et al.* (2001). Para evitar este problema, propormos 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

$$\log(\eta_{1j}) = \boldsymbol{x'_j}\boldsymbol{\beta_1} \quad , \quad \log\left(\frac{p_j}{1-p_j}\right) = \boldsymbol{\nu'_j}\boldsymbol{\beta_2} \quad \text{e} \quad \log(\eta_{3j}) = \boldsymbol{w'_j}\boldsymbol{\beta_3}, \quad j = 1, \ldots, m, \tag{3.22}$$

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 $\boldsymbol{W'} = (\boldsymbol{w'_1}, \boldsymbol{w'_2}, \ldots, \boldsymbol{w'_m})$.

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

Lema 3.1 *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*

$$f(t_j, \delta_j | n_{1j}, n_{2j}, n_{3j}) = \{1 - F^{n_{3j}}(t_j; \boldsymbol{\gamma})\}^{n_{2j} - \delta_j} \{n_{2j} n_{3j} f(t_j; \boldsymbol{\gamma}) F^{n_{3j}-1}(t_j; \boldsymbol{\gamma})\}^{\delta_j}. \tag{3.23}$$

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 ($\delta_j = 1$) :*

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

Combinando as duas situações, obtemos o resultado enunciado.

Em seguida, apresentamos a função verossimilhança do processo híbrido.

Teorema 3.3 *Supondo um processo híbrido com censura não-informativa, a função de verossimilhança completa é dada por*

$$\begin{aligned}
L(\boldsymbol{\vartheta}; \boldsymbol{D}_c) &= \prod_{j=1}^{m} \{1 - F^{n_{3j}}(t_j; \boldsymbol{\gamma})\}^{n_{2j} - \delta_j} \{n_{2j} n_{3j} f(t_j; \boldsymbol{\gamma}) F^{n_{3j}-1}(t_j; \boldsymbol{\gamma})\}^{\delta_j} \times \\
&\quad \mathbb{P}[N_{1j} = n_{1j}] \mathbb{P}[N_{2j} = n_{2j} | N_{1j} = n_{1j}] \{\mathbb{P}[N_{3j} = n_{3j}]\}^{n_{2j}} \qquad (3.24)
\end{aligned}$$

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*

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

sendo $\boldsymbol{n_1'} = (n_{11},\ldots,n_{1m})$, $\boldsymbol{n_2'} = (n_{21},\ldots,n_{2m})$ *e* $\boldsymbol{n_3'} = (n_{31},\ldots,n_{3m})$. *O resultado segue diretamente de (3.23).*

Note que a verossimilhança (3.24) depende de $\boldsymbol{N_1}$, $\boldsymbol{N_2}$ e $\boldsymbol{N_3}$, que são variáveis latentes.

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

$$L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs}) = \prod_{j=1}^{m}\big\{f_{pop}(t_j;\boldsymbol{\vartheta})\big\}^{\delta_j}\big\{S_{pop}(t_j;\boldsymbol{\vartheta})\big\}^{1-\delta_j}, \tag{3.25}$$

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:*

- $\delta_j = 0$*:*

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

- $\delta_j = 1$:

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

Agora, assumindo uma distribuição Weibull para o tempo até o tumor de cada célula maligna (Z), cuja funções distribuição e densidade são dadas, respectivamente, por

$$F(z;\boldsymbol{\gamma}) = 1 - \exp(-z^{\gamma_1} e^{\gamma_2}) \quad \text{e} \quad f(z;\boldsymbol{\gamma}) = \gamma_1 z^{\gamma_1 - 1} \exp(\gamma_2 - z^{\gamma_1} e^{\gamma_2}), \tag{3.26}$$

para $z > 0$, $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2)$, com $\gamma_1 > 0$, e $\gamma_2 \in \Re$.

As estimativas de máxima verossimilhança do parâmetro $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \phi_1)$ são obtidas de maneira análoga a 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 (3.25).

Para a implementação do algoritmo de Metropolis-Hastings 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

$$\pi(\boldsymbol{\beta_1}|\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\phi_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_1}), \quad \pi(\boldsymbol{\beta_2}|\boldsymbol{\beta_1},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\phi_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_2}),$$

$$\pi(\boldsymbol{\beta_3}|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\gamma_1,\gamma_2,\phi_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_3}), \quad \pi(\gamma_1|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_2,\phi_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\gamma_1}),$$

$$\pi(\gamma_2|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\phi_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\gamma_2) \text{ e } \pi(\phi_1|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\phi_1).$$

Novamente, estas distribuições condicionais não são avaliadas de forma fechada. Então faremos uso do algoritmo Metropolis-Hasting dentro do ciclo do algoritmo de Gibbs para simular amostras de $\boldsymbol{\vartheta}$.

3.4 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.* (2001a) (dados E1690, disponível em http://merlot.stat.uconn.edu/mhchen/survbook/). 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, ligamos os parâmetros η_1, p e η_3 em (3.22) para 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 expressa através de

$$\log(\eta_{1j}) = \beta_{1_1} x_{1j},\ \log\left(\frac{p_j}{1-p_j}\right) = \beta_{2_0} + \beta_{2_1} x_{2j} \text{ e } \log(\eta_{3j}) = \beta_{3_1} x_{3j},\ j = 1, \ldots, 417. \quad (3.29)$$

A Figura 3.2 (painel esquerdo) apresenta o gráfico TTT dos dados de câncer de melanoma que indica uma função de risco crescente, podendo então ser representada pela distribuição Weibull. A Curva Kaplan-Meier estratificada por categoria do nódulo na Figura 3.2 (painel direito) nivela entre 0,2 a 0,7. Este comportamento sugere claramente que os modelos que ignoram a possibilidade de taxa de cura não será adequado para analisar estes dados.

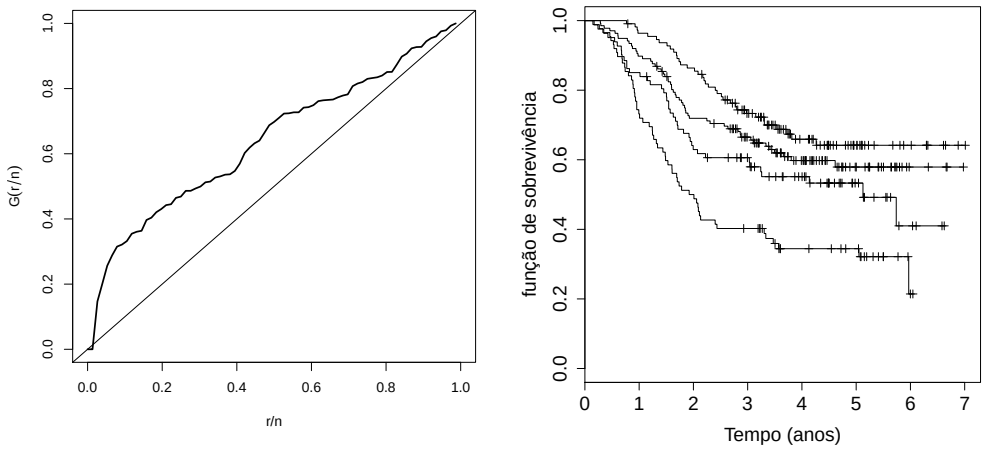

Figura 3.2: Painel esquerdo: gráfico TTT. Painel direito: curva Kaplan-Meier estratificada por categoria do nódulo (1 até 4, de cima para baixo).

Ajustamos os modelos da Tabela 3.1. Um caso particular do modelo HBNP também foi ajustado aos dados, a saber, o modelos híbrido geométrico-Poisson (HGP) ($\phi_1 = 1$). A Tabela (3.2) 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 quantil aleatorizado residual 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 3.2 e o gráfico QQ na Figura 3.3, selecionamos o modelo HGP como nosso modelo de trabalho. Estimativas de máxima verossimilhança dos coeficientes estão na Tabela 3.3. A estimativa do parâmetro de forma (γ_1) fornece uma evidência contra a distribuição exponencial ($\gamma_1 = 1$) para os tempos de vidas não observados, fortalecendo as informações do gráfico TTT.

Tabela 3.2: Os valores do $\max \log L(\cdot)$ e as estatísticas AIC e BIC para os quatros modelos ajustados: HPPEP, HBNP, HCPP e HGP.

Critério	HPPEP	HBNP	HCPP	HGP
$\max \log L(\cdot)$	-516,989	-509,068	-517,446	-509,479
AIC	1047,979	1032,138	1048,891	1030,959
BIC	1076,211	1060,369	1077,123	1055,157

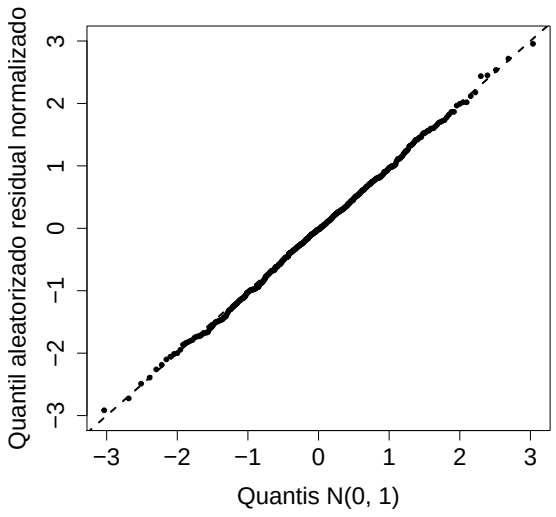

Figura 3.3: Gráfico QQ do quantil aleatorizado residual normalizado com a reta identidade para o modelo HGP (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados).

Tabela 3.3: 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	1,631	0,109	(1,417 ; 1,845)
γ_2	-1,298	0,163	(-1,617 ; -0,979)
β_{1_1}	0,023	0,005	(0,013 ; 0,0328)
β_{2_0}	-2,351	0,433	(-3,199 ; -1,502)
β_{2_1}	0,976	0,258	(0,470 ; 1,482)
β_{3_1}	0,079	0,024	(0,032 ; 0,126)

Usando as estimativas da Tabela 3.3, a função de ligação logarítmica em (3.22), e $\boldsymbol{I_0(\widehat{\beta_1})}$ extraída de (2.36), 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 3.3 para pacientes com espessura do tumor 0,7, 3,1 e 10.0 mm. Essas espessuras correspondem aos quantiles 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 quantiles 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 3.4: 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	0,348	0,041	(0,266 ; 0,429)
3,1	0,279	0,129	(0,026 ; 0,532)
10,0	0,112	0,175	(0,000 ; 0,454)

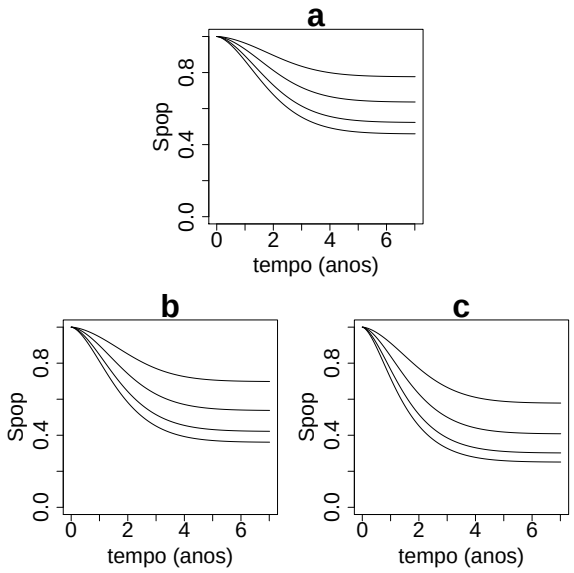

Figura 3.4: Função de sobrevivência sob o modelo HGP estratificado por categoria do nódulo (1 até 4, de cima para baixo) para pacientes com idades iguais a (a) 29, (b) 47, e (c) 70 anos, respectivamente, e espessura do tumor 3,94 mm.

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 3.3, 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.

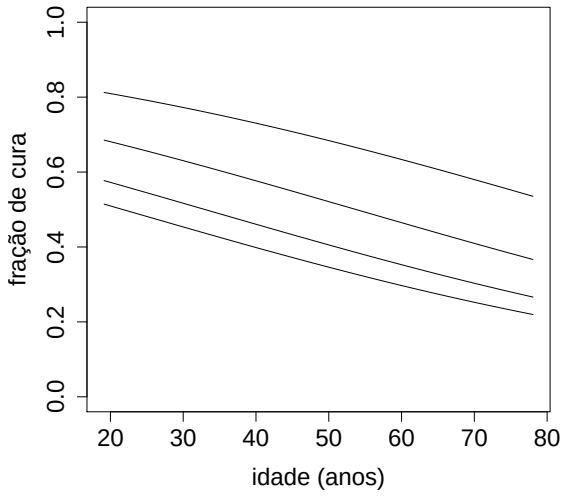

Figura 3.5: Fração de cura para o modelo HGP *versus* idade estratificada por categoria do nódulo (1 até 4, de cima para baixo) e espessura do tumor 3,94 mm.

Também obtemos os ajustes para os quatro modelos da tabela 3.2 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 3.5, 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 3.6 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 3.7 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 3.5.

Tabela 3.5: As estatísticas DIC, EAIC, EBIC e B para os quatro modelos ajustados: HPPEP, HBNP, HCPP e HGP.

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 3.6: 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	1,642	0,107	(1,430 ; 1,842)	1,002
γ_2	-1,350	0,165	(-1,684 ; -1,040)	1,003
β_{1_1}	0,022	0,005	(0,013 ; 0,032)	1,001
β_{2_0}	-2,357	0,464	(-3,266 ; -1,443)	1,003
β_{2_1}	1,095	0,324	(0,617 ; 1,887)	1,002
β_{3_1}	0,064	0,030	(0,001 ; 0,113)	1,001

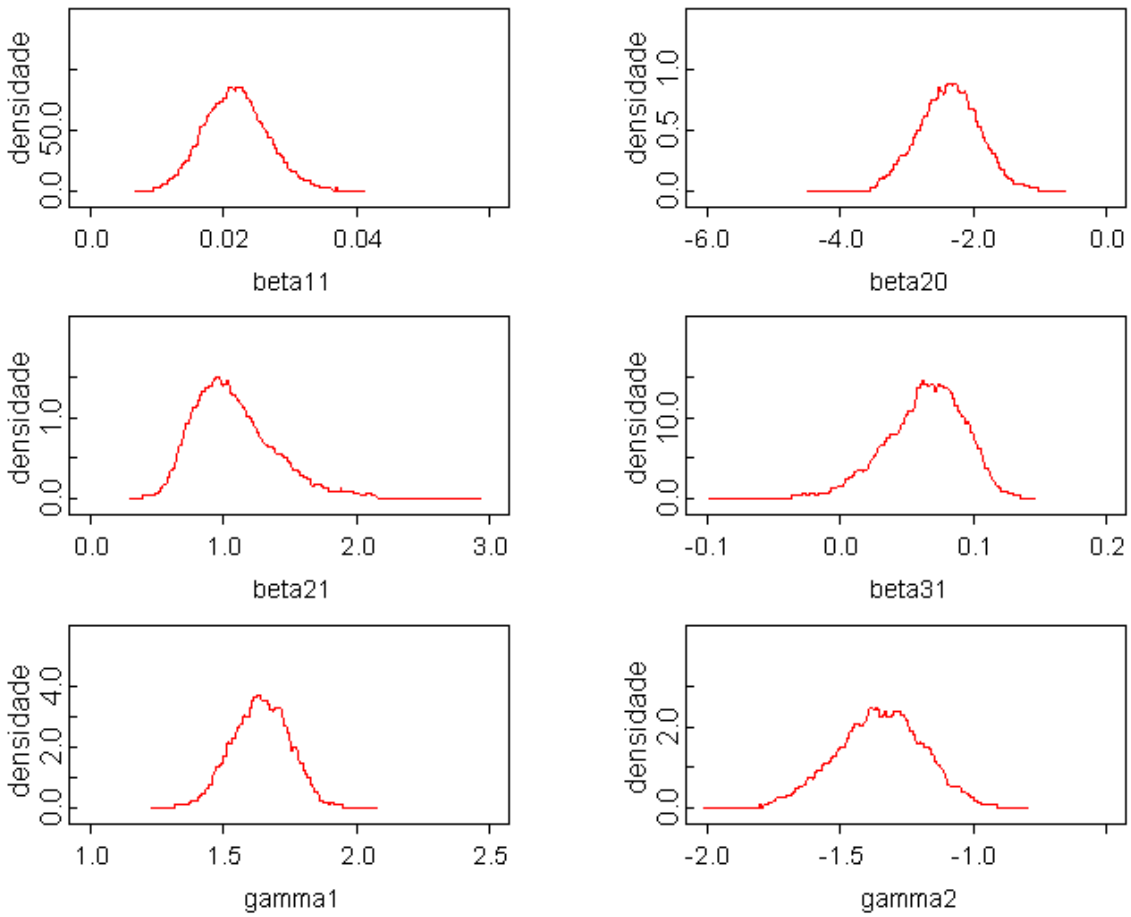

Figura 3.6: Densidades *a posteriori* marginais aproximadas dos parâmetros.

Tabela 3.7: Médias *a posteriori*, os desvios padrão e os 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 nódulo	**Média**	**desvio padrão**	**ICred 95%**
0,7	0.352	0.008	(0.339 ; 0.368)
3,1	0.296	0.033	(0.242 ; 0.369)
10,0	0.161	0.087	(0.045 ; 0.373)

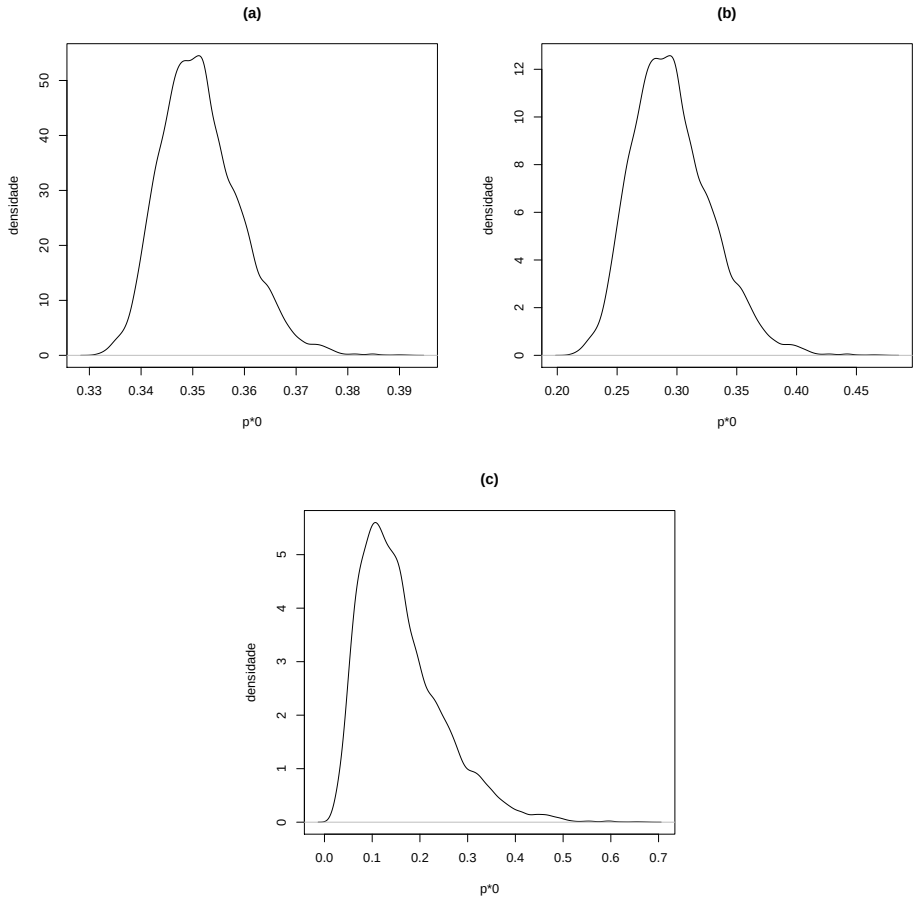

Figura 3.7: Densidade *a posteriori* marginal aproximada para a proporção de células malignas que morrem antes da indução do tumor (p_0^*) sob o modelo HGP para pacientes com espessura do tumor (a) 0,7, (b) 3,1 e (c) 10.0 mm.

A Tabela 3.8 contém os resumos *a posteriori* para a fração de cura estratificada por categoria do nódulo (1-4) e espessura do tumor 3,94 mm para pacientes com idades de 29, 47 e 70 anos de 3000 amostras retiradas do modelo HGP. Esta tabela nos permite avaliar o efeito combinado das covariáveis sob a fração de cura, notando que ambos agem para reduzir a fração de cura. As diferenças entre as idade 29 e 70 anos dos pacientes são significativas ao nível de 5% para todas as categorias do nódulo.

Tabela 3.8: 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	0,769	0,038	(0,689 ; 0,836)
	2	0,623	0,041	(0,542 ; 0,699)
	3	0,520	0,034	(0,460 ; 0,592)
	4	0,467	0,026	(0,418 ; 0,521)
47	1	0,692	0,041	(0,609 ; 0,693)
	2	0,527	0,035	(0,457 ; 0,528)
	3	0,422	0,032	(0,359 ; 0,485)
	4	0,371	0,035	(0,306 ; 0,440)
70	1	0,576	0,053	(0,471 ; 0,678)
	2	0,402	0,039	(0,327 ; 0,479)
	3	0,307	0,231	(0,228 ; 0,391)
	4	0,265	0,047	(0,181 ; 0,363)

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.

3.5 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 dentro da análise características do estágio de progressão, bem como a proporção de células iniciadas que foram "promovidas"a malignas e a proporção de células malignas que mor-

rem 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 foi 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 assumir que as células em um tecido pode 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 o problema 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 denvolvimento uma 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 matemáticos que possam incorporar adequadamente a dependência biológica, e isso que proporcionou a motivação para o presente capítulo.

Conseqüentemente, a finalidade principal deste Capítulo é propor um modelo de sobrevivência com fração de cura que estendem os modelos formulados no capítulo anterior, incorporando uma estrutura de dependência entre as células iniciadas ao tornar-se cancerosas. Para criar a estrutura de dependência entre as células, usamos uma extensão da distribuição série de potência

generalizada incluíndo um parâmetro adicional ρ (distribuição série de potência generalizada inflada (SPGI), Kolev *et al.* (2000)). O parâmetro ρ tem uma interpretação natural em termos de proporção de "zero-inflado"e coeficiente de correlação. Portanto, em nossa abordagem o número de células iniciadas segue uma distribuição SPGI. A distribuição SPGI são candidatas para a modelagem de dados de contagem correlacionados, os quais apresentam superdispersão. A vantagem desta suposição 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 as características dos estágios não observáveis (iniciação, promoção e progressão) do processo da carcinogênese, e a interdependência entre as células em um cenário de riscos competitivos.

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, um conjunto de dados de câncer melanoma real ilustra a utilidade do modelo proposto. Comentários finais são apresentados na Seção 4.5.

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 serem 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 $F(\cdot) = 1 - S(\cdot)$, sendo $S(\cdot)$ função de sobrevivivência. Todas

células malignas estão sujeitas a progressão dependentemente umas das outras.

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

Levando em conta (2.1), (2.2) e o Teorema 3.1, a função de sobrevivência de longa duração é expressa por

$$S_{pop}(y) = \frac{g\left(\dfrac{\theta_1\left[1-p\left(\frac{g\left(\frac{\theta_3 F(y)(1-\rho)}{1-\rho F(y)}\right)}{g(\theta_3)}-p_{n_3}(0)\right)\right](1-\rho)}{1-\rho\left[1-p\left(\frac{g\left(\frac{\theta_3 F(y)(1-\rho)}{1-\rho F(y)}\right)}{g(\theta_3)}-p_{n_3}(0)\right)\right]}\right)}{g(\theta_1)}, \tag{4.1}$$

em que $p_{n_3}(0) = \frac{w(0;\phi_3)e^{-\eta_3}}{\mathbb{E}_{\eta_3}[w(N_3;\phi_3)]}$. A fração de cura é determinada por $p_0 = \lim_{y\to\infty} S_{pop}(y)$. Assim, a partir de 4.1,

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

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

Referimo-nos ao modelo em (4.1) como modelo híbrido correlacionado série de potência 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 introduzido 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, introduzindo 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

$$\mathbb{P}_{Poi}[N_k = n_k] = \begin{cases} e^{-\eta_k} & , \quad n_k = 0 \\ e^{-\eta_k} \sum_{i=1}^{n_k} \binom{n_k-1}{i-1} \frac{\left[\eta_k(1-\rho)\right]^i \rho^{n_k-1}}{i!} & , \quad n_k = 1, 2, \ldots. \end{cases} \tag{4.2}$$

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

$$\mathbb{A}_{N_k}(z) = \exp\left\{-\frac{\eta_k(1-z)}{1-z\rho}\right\} \quad \text{para} \quad 0 \leq z \leq 1 \quad \text{e} \quad k = 1, 3. \tag{4.3}$$

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

$$S_{pop}(y) = \exp\left\{-\frac{\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)}{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]}\right\}. \tag{4.4}$$

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

$$\mathbb{P}_{Bin}[N_1 = n_1] = \begin{cases} (1-\pi_1)^{m_1} & , \quad n_1 = 0 \\ \sum_{i=1}^{\min(n_1, m_1)} \binom{m_1}{i}\binom{n_1-1}{i-1}\left[\pi_1(1-\rho)\right]^i (1-\pi_1)^{m_1-i}\rho^{n_1-i} & , \quad n_1 = 1, 2, \ldots. \end{cases} \tag{4.5}$$

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

$$\mathbb{A}_{N_1}(z) = \left[1 - \frac{\pi_1(1-z)}{1-z\rho}\right]_1^m \quad \text{para} \quad 0 \leq z \leq 1. \tag{4.6}$$

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

$$S_{pop}(y) = \left[1 - \frac{\pi_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)}{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]}\right]^{m_1}. \tag{4.7}$$

4.2.3 Modelo híbrido correlacionado binomial negativa-Poisson (HCBNP)

Quando as funções $a_{n_1} = \frac{\Gamma\left(\phi_1^{-1}+\sum_{i=1}^{\infty} n_{1i}\right)}{\Gamma(\phi_1^{-1})\left[\sum_{i=1}^{\infty} n_{1i}\right]!}$, $g(\theta_1) = (1-\theta_1)^{-\phi_1^{-1}}$, 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

$$\mathbb{P}_{NB}[N_1 = n_1] = \begin{cases} (1+\phi_1\eta_1)^{-\phi_1^{-1}} & , \quad n_1 = 0 \\ (1+\phi_1\eta_1)^{-\phi_1^{-1}} \sum_{i=1}^{n_1} \binom{n_1-1}{i-1} \frac{\Gamma(\phi_1^{-1}+i)}{\Gamma(\phi_1^{-1})i!} \left[\frac{\phi_1\eta_1(1-\rho)}{1+\phi_1\eta_1}\right]^i \rho^{n_1-i} & , \quad n_1 = 1,2,\ldots. \end{cases} \tag{4.8}$$

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

$$\mathbb{A}_{N_1}(z) = \left[\frac{1-z\rho}{1+\phi_1\eta_1(1-z)-z\rho}\right]^{\phi_1^{-1}} \quad \text{para} \quad 0 \leq z \leq 1. \tag{4.9}$$

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

$$S_{pop}(y) = \left[\frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}{1+\phi_1\eta_1 p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}\right]^{\frac{1}{\phi_1}}. \tag{4.10}$$

Quando $\phi_1 = 1$, obtemos a distribuição geométrica inflada com parâmetro $\theta_1 = \frac{1}{1+\eta_1} \in (0,1)$ em (4.8), e $S_{pop}(\cdot)$ em (4.10) reduz-se a

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

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

$$\mathbb{P}_{LS}[N_1 = n_1] = (-\log(\pi_1))^{-1} \sum_{i=1}^{n_1} \binom{n_1-1}{i-1} \frac{\left[(1-\pi_1)(1-\rho)\right]^i \rho^{n_1-i}}{i}, \quad n_1 = 1,2,\ldots. \tag{4.12}$$

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

$$\mathbb{P}_{LS}[N_1 = n_1] = (-\log(\pi_1))^{-1} \sum_{i=1}^{n_1+1} \binom{n_1}{i-1} \frac{\left[(1-\pi_1)(1-\rho)\right]^i \rho^{n_1+1-i}}{i}, \quad n_1 = 0, 1, 2, \ldots. \tag{4.13}$$

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

$$\mathbb{A}_{N_1}(z) = \frac{(-\log(\pi_1))^{-1}}{z} \log\left[\frac{1-\rho z}{1-z(1-\pi_1(1-\rho))}\right]. \tag{4.14}$$

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

$$S_{pop}(y) = \frac{(-\log(\pi_1))^{-1}}{1-p\left(\exp\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\} - e^{-\eta_3}\right)} \log\left[\frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]}{1-(1-\pi_1(1-\rho))\left(1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\} - e^{-\eta_3}\right)\right]\right)}\right]. \tag{4.15}$$

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	$\frac{(-\log(\pi_1))^{-1}}{1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)}\log\left[\frac{1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]}{1-(1-\pi_1(1-\rho))\left(1-\rho\left[1-p\left(\exp\left\{-\frac{\eta_3 S(y)}{1-\rho F(y)}\right\}-e^{-\eta_3}\right)\right]\right)}\right]$	$\frac{1}{\log(\eta_1)\left(1-\rho\left[1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right]\right)\left(1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)\right)}\left(\frac{(\rho+\eta_1(1-\rho)-1)p\eta_3 f(y)(1-\rho)e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{(1-\rho F(y))^2\left(p(1-\eta_1(1-\rho))\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)+\eta_1(1-\rho)\right)}\right)$ $+\frac{\left(\frac{p\eta_3(1-\rho)f(y)e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}}{(1-\rho F(y))^2}\right) S_{pop}(y)}{1-p\left(e^{-\frac{\eta_3 S(y)}{1-\rho F(y)}}-e^{-\eta_3}\right)}$	$\frac{(-\log(\pi_1))^{-1}}{1-p\left(1-e^{-\eta_3}\right)}\log\left[\frac{1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]}{1-(1-\pi_1(1-\rho))\left(1-\rho\left[1-p\left(1-e^{-\eta_3}\right)\right]\right)}\right]$	$e^{-\eta_3}$

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 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 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 $F(t_j|\boldsymbol{\gamma}) = 1 - S(t_j|\boldsymbol{\gamma})$ (ou $F(t_j;\boldsymbol{\gamma}) = 1 - S(t_j;\boldsymbol{\gamma})$) 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 $\mathbb{P}[Z_{0hj} = \infty] = \mathbb{P}[Z_{i0j} = \infty] = 1$. $\boldsymbol{\gamma}$ representa o vetor de parâmetros da distribuição. 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, os modelos HCPP, HCBP e HCBNP das Seções 4.2.1, 4.2.2 e 4.2.3 não são identificáveis no sentido de Li *et al.* (2001). Para evitar este problema, propomos relacionar os parâmetros η_1, p e η_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

$$\log(\eta_{1j}) = \boldsymbol{x_j'\beta_1} \ \left(ou \ \log\left(\frac{\pi_{1j}}{1-\pi_{1j}}\right) = \boldsymbol{x_j'\beta_1}\right) \ , \ \log\left(\frac{p_j}{1-p_j}\right) = \boldsymbol{\nu_j'\beta_2} \quad \text{e} \tag{4.16}$$

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

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 3.4, a função de verossimilhança marginal é dada por

$$L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs}) = \prod_{j=1}^{m} \big\{ f_{pop}(t_j; \boldsymbol{\gamma}) \big\}^{\delta_j} \big\{ S_{pop}(t_j; \boldsymbol{\gamma}) \big\}^{1-\delta_j}, \tag{4.17}$$

sendo $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma'}, \rho, \phi_1, \phi_3, m_1, m_3)$ 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.

Agora, assumimos uma distribuição Weibull para o tempo até o tumor de cada célula maligna (Z), cuja distribuição e função densidade são dadas, respectivamente, por

$$F(z; \boldsymbol{\gamma}) = 1 - \exp(-z^{\gamma_1} e^{\gamma_2}) \quad \text{e} \quad f(z; \boldsymbol{\gamma}) = \gamma_1 z^{\gamma_1 - 1} \exp(\gamma_2 - z^{\gamma_1} e^{\gamma_2}) \tag{4.18}$$

para $z > 0$, $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2)$, com $\gamma_1 > 0$, e $\gamma_2 \in \Re$.

As estimativas de máxima verossimilhança do parâmetro $\boldsymbol{\vartheta}$ são obtidas de maneira análoga a 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 e $m_1 \sim Uniforme\ discreta\ \big\{1, 2, \ldots, m_0\big\}$ para o modelo HCBP.

As distribuições *a priori* e *a posteriori* de $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\beta_3'}, \boldsymbol{\gamma'}, \rho, \phi_1, m_1)$ são

$$\pi(\boldsymbol{\vartheta}) = \prod_{j_1=1}^{k_1} \pi(\beta_{1j_1}) \prod_{j_2=1}^{k_2} \pi(\beta_{2j_2}) \prod_{j_3=1}^{k_3} \pi(\beta_{3j_3}) \pi(\gamma_1) \pi(\gamma_2) \pi(\rho) \pi(\phi_1) \pi(m_1), \tag{4.19}$$

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

respectivamente, sendo $L(\boldsymbol{\vartheta}; \boldsymbol{D}_{obs})$ dada por (4.17).

As distribuições condicionais completas *a posteriori* são dadas por

$$\pi(\boldsymbol{\beta_1}|\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\rho,\phi_1,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_1}) \text{ , } \pi(\boldsymbol{\beta_2}|\boldsymbol{\beta_1},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\rho,\phi_1,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_2}),$$

$$\pi(\boldsymbol{\beta_3}|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\gamma_1,\gamma_2,\rho,\phi_1,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\beta_3}) \text{ , } \pi(\gamma_1|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_2,\rho,\phi_1,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\gamma_1}),$$

$$\pi(\gamma_2|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\rho,\phi_1,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\gamma_2) \text{ , } \pi(\rho|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\phi_1,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\boldsymbol{\rho}),$$

$$\pi(\phi_1|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\rho,m_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(\phi_1) \text{ e } \pi(m_1|\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\beta_3},\gamma_1,\gamma_2,\rho,\phi_1,\boldsymbol{t},\boldsymbol{\delta}) \propto L(\boldsymbol{\vartheta};\boldsymbol{D}_{obs})\pi(m_1).$$

Novamente, estas distribuições condicionais não são avaliadas de forma fechada. Então faremos uso do algoritmo Metropolis-Hasting dentro do ciclo do algoritmo de Gibbs para simular amostras de $\boldsymbol{\vartheta}$.

4.4 Dados de câncer de melanoma

A metodologia apresentada neste capítulo será aplicada ao conjunto de dados da seção 2.4. O tempo observado (T) 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 que a morte do paciente ou o tempo de censura. Tendo em mente a questão da identificação 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 (4.16) 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

$$\log(\eta_{1j}) = \beta_{1_{pres}}x_{1j} + \beta_{1_{uls}}(1-x_{1j}) \ \left(ou \log\left(\frac{\pi_{1j}}{1-\pi_{1j}}\right) = \beta_{1_{pres}}x_{1j} + \beta_{1_{uls}}(1-x_{1j})\right), \quad (4.21)$$

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

Ajustamos os modelos da Tabela 4.1 e o modelo HCGP. A Tabela 4.2 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 quantil aleatorizado residual 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 4.2 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 4.3. 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 4.2: Os valores do $\max \log L(\cdot)$ e as estatísticas AIC e BIC para os cinco modelos ajustados, HCPP, HCBP, HCBNP, HCGP e HCSLP.

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

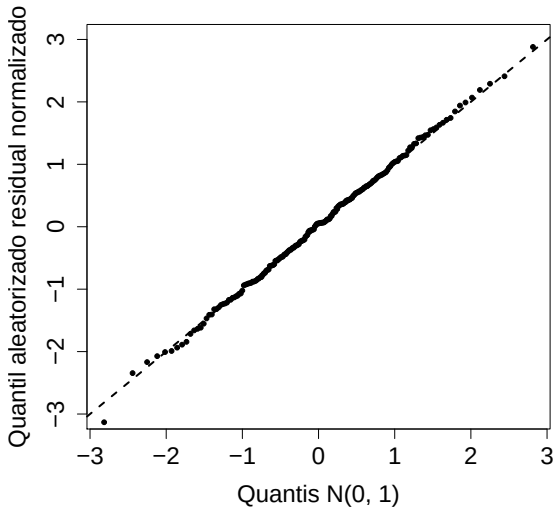

Figura 4.1: Gráfico QQ do quantil aleatorizado residual normalizado com a reta identidade para o modelo HCBNP (cada ponto corresponde à mediana de cinco conjuntos de resíduos ordenados).

Tabela 4.3: 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 4.3, a função de ligação logarítmica em (4.16), e $\boldsymbol{I_0(\widehat{\beta_1})}$ extraída de (2.36), 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 4.4. 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 4.4: 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	(0,00 ; 0,75)
feminino	0,20	0,28	(0,54 ; 1,00)

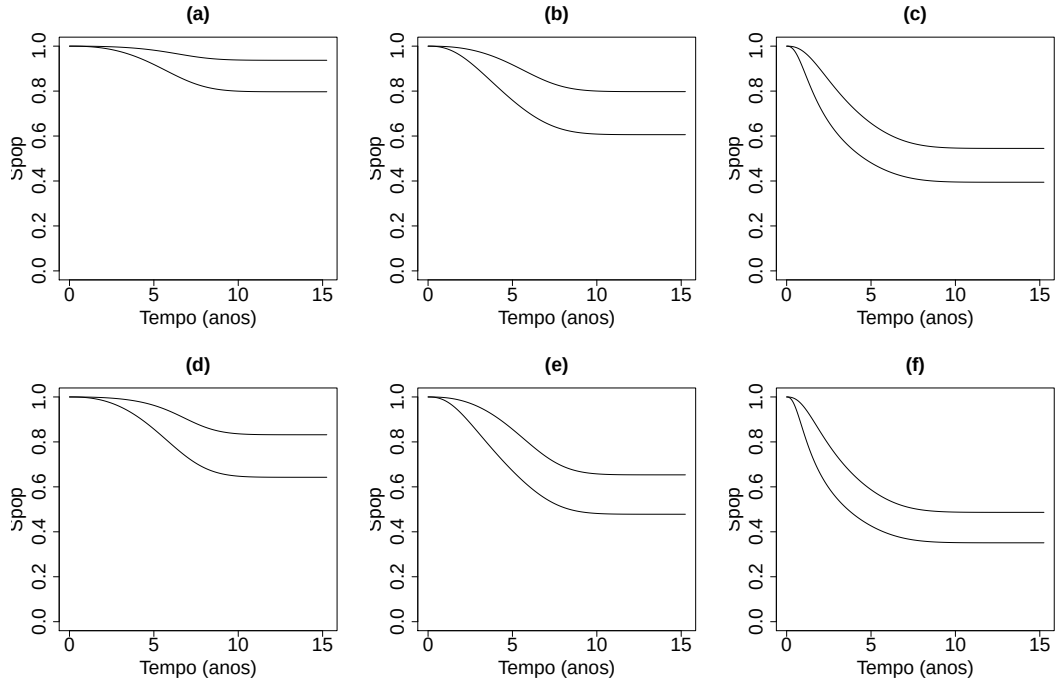

Figura 4.2: Função de sobrevivência sob o modelo HCBNP estratificado pelo estado de úlcera (superior: ausente, inferior: presente) para pacientes do sexo masculino com idades iguais a (a) 29, (b) 47, e (c) 70 anos, respectivamente, e para pacientes do sexo feminino com idades iguais a (d) 29, (e) 47, e (f) 70 anos, 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 4.3 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 4.3, 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.

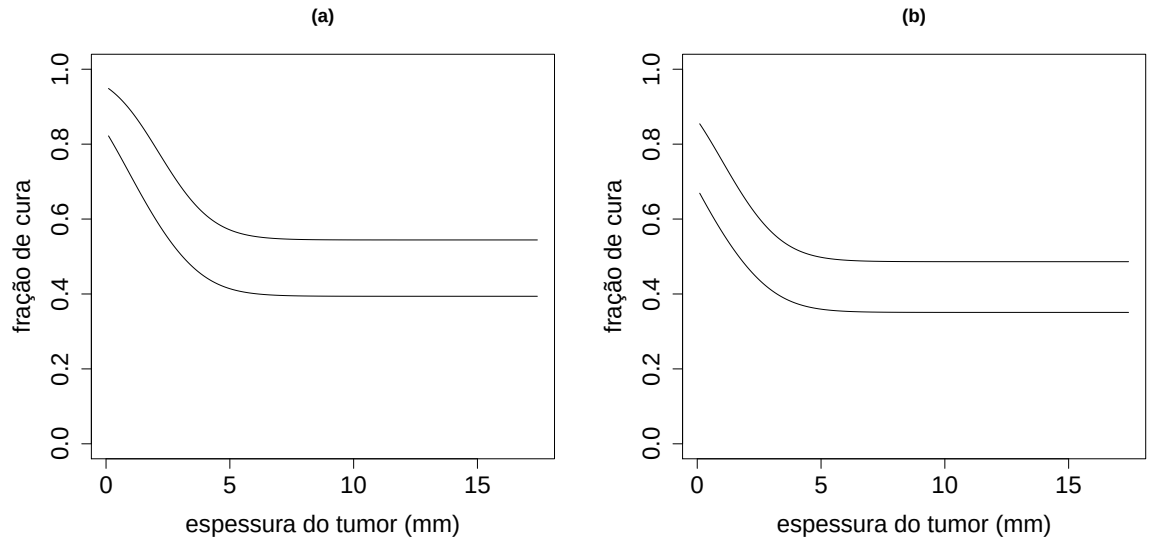

Figura 4.3: Fração de cura para o modelo HCBNP *versus* espessura do tumor estratificada pelo estado de úlcera (superior: ausente, inferior: presente) e sexo (a) masculino e (b) feminino, respectivamente.

Também obtemos os ajustes para os cinco modelos da tabela 4.2 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 e $m \sim Uniforme\ discreta\big\{1, 2, \ldots, 10^5\big\}$ para o modelo HCBP. 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 4.5, foi aplicado 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 4.6 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 4.5.

Tabela 4.5: As estatísticas DIC, EAIC, EBIC e B para os cinco modelos ajustados: HCPP, HCBP, HCBNP, HCGP e HCSLP.

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 4.6: Médias *a posteriori*, os desvios padrão e os 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

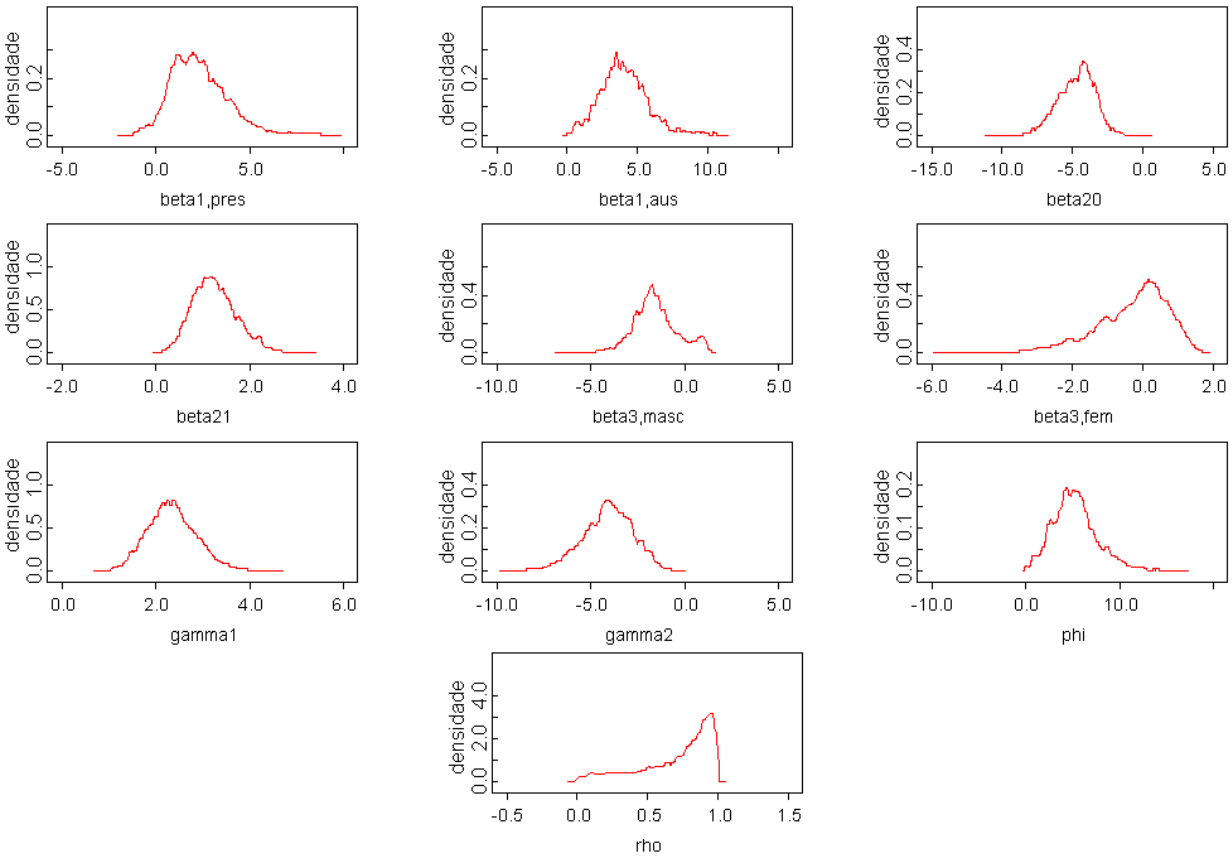

Figura 4.4: Densidades *a posteriori* marginais aproximadas dos parâmetros.

As estimativas das médias das distribuições *a posteriori* (Tabela 4.6) e de máxima verossimilhança (Tabela 4.3) pouco diferem, ao passo que os intervalos de credibilidade são mais precisos do que os intervalos de confiança assintóticos.

4.5 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 relevância prática e a aplicabilidade do modelo foram demonstradas em um conjunto de dados reais de pacientes com câncer de melanoma.

Referências

Aarset, M. V. (1985). The null distribution for a test of constant versus bathtub failure rate. *Scandinavian Journal of Statistics*, **12**(1), 55–68.

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.

Balka, J., Desmond, A. F. & McNicholas, P. D. (2009). Review and implementation of cure models based on first hitting times for wiener processes. *Lifetime Data Analysis*, **15**, 147–176.

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

Barral, A. M. (2001). *TImmunological Studies in Malignant Melanoma: Importance of TNF and the Thioredoxin System*. Doctorate Thesis - Linköping University, Linköping, 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. & Balakrishnan, N. (2011a). Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data. *Computational Statistics and Data Analysis*, **DOI: 10.1016/j.csda.2011.10.013**.

Borges, P., Rodrigues, J. & Louzada-Neto, F. (2011b). 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. (2011c). 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.

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. (1999a). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909–919.

Chen, M. H., Ibrahim, J. G. & Sinha, D. (1999b). 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.

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

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.

De Castro, A. F. M., Cancho, V. G. & Rodrigues, J. (2007). A flexible model forsurvival 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.

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.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications, third ed., vol. I*. New York.

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.

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. (2001a). *Bayesian Survival Analysis*. Springer, New York.

Ibrahim, J. G., Chen, M. H. & Sinha, D. (2001b). *Bayesian Survival Analysis*. New York: Springer.

INCA (2011). Instituto nacional do câncer. http://www.inca.gov.br/conteudoview.aspid=322. Último acesso em 30/05/2011.

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

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. 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.

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.

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.

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.

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.

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.