UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

Uso de métodos bayesianos na análise de dados de sobrevida para pacientes com câncer na mama na presença de censuras, fração de cura e covariáveis.

TATIANA REIS ICUMA

Ribeirão Preto - SP
2016

TATIANA REIS ICUMA

Uso de métodos bayesianos na análise de dados de sobrevida para pacientes com câncer na mama na presença de censuras, fração de cura e covariáveis.

Dissertação apresentada ao Programa de Pós-graduação em Saúde na Comunidade da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre.

Área de concentração: Saúde na comunidade.

Orientador: Prof. Dr. Jorge Alberto Achcar

Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD)

Ribeirão Preto - SP

2016

Ficha Catalográfica

Icuma, Tatiana Reis

Uso de métodos bayesianos na análise de dados de sobrevida para pacientes com câncer na mama na presença de censuras, fração de cura e covariáveis. Ribeirão Preto, 2016.

118 p. : il ; 30cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Saúde na Comunidade.

Orientador: Achcar, Jorge Alberto.

1. Análise de sobrevivência. 2. Fração de cura. 3. Inferência bayesiana. 4. Neoplasia de mama.

Folha de Aprovação

Tatiana Reis Icuma

Uso de métodos bayesianos na análise de dados de sobrevida para pacientes com câncer na mama na presença de censuras, fração de cura e covariáveis

Dissertação apresentada ao Programa de Pós-graduação em Saúde na Comunidade da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre.

Área de concentração: Saúde na Comunidade.

Aprovado em: ____/____/____

Banca Examinadora

Prof.(a) Dr.(a) _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof.(a) Dr.(a) _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof.(a) Dr.(a) _____ Instituição: _____

Julgamento: _____ Assinatura: _____

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

DEDICATÓRIA

À MINHA FAMILIA.

AGRADECIMENTOS

AOS MEUS PAIS, ELISA E ADILSON, PELO AMOR E TEMPO DEDICADOS NA MINHA CRIAÇÃO E DESENVOLVIMENTO, PELA EDUCAÇÃO DADA AO LONGO DA MINHA VIDA.

À ISABELA, QUE GENTILMENTE COMPARTILHOU O SEU BANCO DE DADOS PARA A REALIZAÇÃO DESSE ESTUDO.

AO PROFESSOR EDSON, PELO ACOLHIMENTO EM UM MOMENTO DE GRANDES MUDANÇAS NA MINHA VIDA, PELO INCENTIVO E OPORTUNIDADE DE INGRESSAR NA CARREIRA ACADÊMICA. VOCÊ É O MEU NORTE, SEMPRE ME ORIENTANDO COM SABEDORIA E DEDICAÇÃO.

AO PROFESSOR JORGE, MEU ORIENTADOR, POR ACREDITAR NO MEU TRABALHO E DESENVOLVER COM SÁBIAS ORIENTAÇÕES O MEU POTENCIAL, QUE NEM EU SABIA QUE EXISTIA.

À MINHA VÓ, DONA ROSA, PELOS DOIS ANOS DE MUITA DEDICAÇÃO, ME RECEBENDO E ACOLHENDO NA SUA CASA, COM MUITO AMOR E UMA DELICIOSA COMIDA CASEIRA.

AOS MEMBROS DA MINHA BANCA DE QUALIFICAÇÃO, FERNANDA, EMÍLIO E EDSON, PELA CONTRIBUIÇÃO COM O TRABALHO COM CORREÇÕES E SUGESTÕES QUE SEMPRE SERÃO BEM-VINDAS.

AO MEU NOIVO, VITOR, PELO COMPANHEIRISMO E SUPORTE, POR SEMPRE ACREDITAR EM MIM. AMOR, VOCÊ É A MELHOR DECISÃO QUE TOMEI NA MINHA VIDA. OBRIGADA

RESUMO

ICUMA, Tatiana Reis. **Uso de métodos bayesianos na análise de dados de sobrevida para pacientes com câncer na mama na presença de censuras, fração de cura e covariáveis.** 2016. 118 páginas. Dissertação (Mestrado) – Faculdade de Medicina de Ribeirão Preto – USP, Ribeirão Preto – SP – Brasil, 2016.

Introdução: Uma das maiores causas de mortes no mundo é devido ao câncer, cerca de 8,2 milhões em 2012 (World Cancer Report, 2014). O câncer de mama é a forma mais comum de câncer entre as mulheres e a segunda neoplasia mais frequente, seguida do câncer de pele não melanoma, representando cerca de 25% de todos os tipos de cânceres diagnosticados. Modelos estatísticos de análise de sobrevivência podem ser úteis para a identificação e compreensão de fatores de risco, fatores de prognóstico, bem como na comparação de tratamentos. **Métodos**: Modelos estatísticos de análise de sobrevivência foram utilizados para evidenciar fatores que afetam os tempos de sobrevida livre da doença e total de um estudo retrospectivo realizado no Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Ribeirão Preto, referente a 54 pacientes com câncer de mama localmente avançado com superexpressão do Her-2 que iniciaram a quimioterapia neoadjuvante associada com o medicamento Herceptin® (Trastuzumabe) no período de 2008 a 2012. Utilizaram-se modelos univariados com distribuição Weibull sem e com a presença de fração de cura sob o enfoque frequentista e bayesiano. Utilizaram-se modelos assumindo uma estrutura de dependência entre os tempos observados baseados na distribuição exponencial bivariada de Block Basu, na distribuição geométrica bivariada de Arnold e na distribuição geométrica bivariada de Basu-Dhar. **Resultados**: Resultados da análise univariada sem a presença de covariáveis, o modelo mais adequado às características dos dados foi o modelo Weibull com a presença de fração de cura sob o enfoque bayesiano. Ao incorporar nos modelos as covariáveis, observou-se melhor ajuste dos modelos com fração de cura, que evidenciaram o estágio da doença como um fator que afeta a sobrevida livre da doença e total. Resultados da análise bivariada sem a presença de covariáveis estimam médias de tempo de sobrevida livre da doença para os modelos Block e Basu, Arnold e Basu-Dhar de 108, 140 e 111 meses, respectivamente e de 232, 343, 296 meses para o tempo de sobrevida total. Ao incorporar as covariáveis, os modelos evidenciam que o estágio da doença afeta a sobrevida livre da doença e total. No modelo de Arnold a covariável tipo de cirurgia também se mostrou significativa. **Conclusões**: Os resultados do presente estudo apresentam alternativas para a análise de sobrevivência com tempos de sobrevida na presença de fração de cura, censuras e várias covariaveis. O modelo de riscos proporcionais de Cox nem sempre se adequa às características do banco de dados estudado, sendo necessária a busca de modelos estatísticos mais adequados que produzam inferências consistentes.

PALAVRAS-CHAVE: Análise de sobrevivência, Fração de cura, Inferência bayesiana, Neoplasia de mama.

ABSTRACT

ICUMA, Tatiana Reis. **Use of bayesian methods in the analysis of survival data for pacients with breast cancer in presence of censoring, cure fraction and covariates**. 2016. 118 páginas. Dissertação (Mestrado) – Faculdade de Medicina de Ribeirão Preto – USP, Ribeirão Preto – SP – Brasil, 2016.

Introduction: The leading worldwide cause of deaths is due to cancer, about 8.2 million in 2012 (World Cancer Report, 2014). Breast cancer is the most common form of cancer among women and the second most common cancer, followed by non-melanoma skin cancer, accounting for about 25% of all diagnosed types of cancers. Statistical analysis of survival models may be useful for the identification and understanding of risk factors, prognostic factors, and the comparison treatments. **Methods**: Statistical lifetimes models were used to highlight the important factors affecting the disease-free times and the total lifetime about a retrospective study conducted at the Hospital das Clinicas, Faculty of Medicine, University of São Paulo, Ribeirão Preto, referring to 54 patients with locally advanced breast cancer with Her-2 overexpression who started neoadjuvant chemotherapy associated with the drug Herceptin® (Trastuzumab) in the time period ranging from years 2008 to 2012. It was used univariate models assuming Weibull distribution with and without the presence of cure fraction under the frequentist and Bayesian approaches. It was also assumed models assuming a dependence structure between the observed times based on the bivariate Block-Basu exponential distribution, on the bivariate Arnold geometric distribution and on the bivariate Basu-Dhar geometric distribution. Results: From the results of the univariate analysis without the presence of covariates, the most appropriate model for the data was the Weibull model in presence of cure rate under a Bayesian approach. By incorporating the covariates in the models, there was best fit of models with cure fraction, which showed that the stage of the disease was a factor affecting disease-free survival and overall survival. From the bivariate analysis results without the presence of covariates, the estimated means for free survival time of the disease assuming the Block- Basu, Arnold and Basu-Dhar models were respectively given by 108, 140 and 111; for the overall survival times the means were given respectively by, 232, 343, 296 months. In presence of covariates, the models showed that the stage of the disease affects the disease-free survivals and the overall survival times. Assuming the Arnold model, the covariate type of surgery also was significant. **Conclusions**: The results of this study present alternatives for the analysis of survival times in the presence of cure fraction, censoring and covariates. The Cox proportional hazards model not always is apropriate to the database characteristics studied, which requires the search for more suitable statistical models that produce consistent inferences.

KEYWORDS: Bayesian inference, Breast neoplasms, Cure model, Survival analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. Introdução

Uma das maiores causas de mortes no mundo é devido ao câncer, cerca de 8,2 milhões em 2012 (Stewart e Christopher, 2014). O câncer de mama é a forma mais comum de câncer entre as mulheres e a segunda neoplasia mais frequente, seguida do câncer de pele não melanoma, representando cerca de 25% de todos os tipos de cânceres diagnosticados. Estima-se que cerca de um milhão e meio de novos casos são diagnosticados em todo o mundo, sendo a quinta forma de câncer com mais óbitos, 522 mil em 2012 (Ferlay et al., 2013). A mortalidade por câncer de mama tem decrescido em países desenvolvidos nas últimas duas décadas devido a melhorias nos diagnósticos e tratamentos (Boyle e Levin, 2008).

Nos Estados Unidos, é a segunda maior causa de óbitos por câncer, sendo estimado que uma em cada oito mulheres desenvolva a doença em sua vida (DeSantis et al., 2014). Sua incidência no Brasil em 2014 é de aproximadamente 56,20 casos para 100 mil mulheres (BRASIL, 2016). Ele representa cerca de 20% de todos os tipos de câncer e é o mais frequente em mulheres nas regiões Nordeste (38,74/100mil), Centro-Oeste (55,87 /100mil), Sudeste (68,08/100mil) e Sul (74,30/100mil), enquanto que na região Norte, é o segundo tumor mais incidente (22,26/100mil), após o tumor do colo do útero. O estado e a cidade de São Paulo possuem incidências acima da nacional (73,21/100mil e 91,21/100mil). Ver estimativas para o ano de 2016 na Tabela 1.

Tabela 1: Estimativas para o ano de 2016 do número de casos novos de câncer.

Região	Casos novos	Taxa bruta por 100 mil habitantes	Distribuição proporcional
Brasil	57960	56,20	19,26%
Norte	1810	22,26	17,35%
Nordeste	11190	38,74	20,53%
Centro-Oeste	4230	55,87	19,73%
Sudeste	29760	68,08	18,98%
Sul	10970	74,30	18,99%
Estado de São Paulo	15570	73,21	19,53%
Cidade de São Paulo	5550	91,21	22,60%

Fonte: INCA (ver: http://www.inca.gov.br/estimativa/2016/)

1.1. Alguns breves conceitos sobre o câncer de mama

O ciclo natural da vida se inicia quando ocorre a fecundação do óvulo pelo espermatozoide gerando a chamada célula-ovo. Em seguida, esta célula trata-se de fazer a divisão celular gerando duas células-filhas que repetem este processo até chegar aos 70 bilhões de células de um organismo adulto (Instituto Vencer o Câncer, 2013). O corpo

humano está em constante renovação celular, seja para repor células mortas ou para regenerar lesões, para isso, as células fazem cópias idênticas de si controladas pelo DNA.

No entanto, podem surgir as mutações, que produzem células-filhas não idênticas, podendo ser causadas por fatores externos (fatores ambientais) ou por fatores internos, sendo capazes de causar alterações na molécula de DNA. Essas mutações podem ser corrigidas por enzimas especializadas ou a estrutura afetada do DNA torna a célula incapaz de dividir-se, porém há situações em que a mutação não é eliminada e se elas ocorrem nos genes envolvidos nos mecanismos de divisão celular, podem causar uma multiplicação celular descontrolada.

Quando essas células começam a se multiplicar de forma desordenada produzem uma massa chamada tumor, que se interferir no funcionamento dos órgãos é chamado de maligno e consequentemente câncer. A metástase é o processo em que as células mutantes se desgarram da massa tumoral e penetram para dentro de vasos sanguíneos caindo na circulação e invadindo locais mais distantes da origem (Borges et al., 2007).

A mama é constituída por gordura, tecido conjuntivo, vasos sanguíneos, vasos linfáticos, lóbulos e ductos. Os lóbulos são responsáveis pela produção de leite e os ductos são pequenos canais que ligam os lóbulos aos mamilos. A maioria dos cânceres de mama tem início nos ductos, alguns nos lóbulos e os outros nos tecidos. O câncer de mama é derivado das células epiteliais que revestem o ducto terminal do lóbulo mamário. Quando a célula cancerosa não ultrapassa as camadas dos ductos, a neoplasia é classificada como in situ ou não invasiva e quando ocorre disseminação para todos os tecidos adjacentes, a neoplasia é classificada como invasiva e apresenta a possibilidade de desenvolver metástase, podendo migrar para outras partes do corpo. (Khatib; Modjtabai, 2006)

Pinho e Coutinho (2007) descrevem que como os demais cânceres, o câncer de mama ainda não tem uma etiologia totalmente esclarecida, sendo que a mesma está atribuída a uma interação de fatores que, de certa forma são considerados determinantes no desenvolvimento da doença.

1.2. Fatores de risco

Todos os cânceres de mama têm origem genética. Acredita-se que 90%-95% deles sejam esporádicos (não familiares) e decorram de mutações somáticas que se verificam durante a vida, e que 5%-10% sejam hereditários (familiares) devido à herança de uma mutação germinativa ao nascimento, que confere a estas mulheres suscetibilidade ao câncer de mama

(Bilmoria, 1995). O Projeto Diretrizes (Barros et al., 2001), iniciativa conjunta da Associação Médica Brasileira e Conselho Federal de Medicina, elaborado em 2001, listou os principais fatores que aumentam a chance de uma mulher vir a apresentar o câncer de mama (fatores de risco).

Os fatores com risco muito elevado (Risco Relativo>3) são: mãe ou irmã com câncer de mama na menopausa, antecedentes de neoplasia lobular “in situ”, suscetibilidade genética comprovada (mutação dos genes BRCA1 ou BRCA2). Os fatores com risco intermediário (1,5< Risco Relativo <3) são: mãe ou irmã com câncer de mama na pós-menopausa, nuliparidade (mulheres que nunca engravidaram) e antecedente de macrocistos apócrinos. Já os fatores com menor risco (Risco Relativo <1,5) e mais difundidos na população em geral são: menarca precoce (antes dos 12 anos), menopausa tardia (depois dos 55 anos), primeira gestação depois dos 34 anos, obesidade, dieta gordurosa, sedentarismo, terapia de reposição hormonal por mais de 5 anos e ingestão alcoólica excessiva.

Além desses, a idade continua sendo um dos mais importantes fatores de risco (INCA 2016). As taxas de incidência aumentam rapidamente até os 50 anos. Após essa idade, o aumento ocorre de forma mais lenta, o que reforça a participação dos hormônios femininos na etiologia da doença. Entretanto, o câncer de mama observado em mulheres jovens apresenta características clínicas e epidemiológicas bem diferentes das observadas em mulheres mais velhas. Geralmente são mais agressivos, apresentam uma alta taxa de presença da mutação dos genes BRCA1 e BRCA2, além de superexpressarem o gene do fator de crescimento epidérmico humano receptor 2 (HER-2).

Quando mencionada a influência hormonal no desenvolvimento do câncer de mama, deve ser destacada a importância do estrógeno (hormônio produzido primariamente pelo ovário). Beatson em 1896 reconhece o câncer de mama como hormônio dependente, quando provou através de seus experimentos que com a remoção dos ovários ocorre a regressão da disseminação do câncer de mama (Beatson, 1896).

A primeira gravidez em mulheres com idade igual ou inferior a 20 anos apresenta efeito de proteção contra o câncer de mama, pois se propõe que o desenvolvimento pleno da glândula mamária, quando ocorre em idade precoce, é fator de proteção contra o câncer de mama (Kelsey et al, 1993)

1.3. Classificação dos tipos de câncer de mama

Após a detecção do nódulo na mama é necessário fazer a biópsia gerando um relatório anatomopatológico capaz de caracterizar o tumor encontrado. O TNM - Classification of Malignant Tummours é o sistema mais usado para a classificação de tumores malignos e descrição de sua extensão anatômica (Compton, 2012), na prática ele caracteriza os casos de câncer em grupos de acordo com os estádios. Simplificadamente, os estádios classificam o câncer de acordo com a extensão da doença para auxiliar a escolha do tratamento. No câncer de mama é importante incluir também o status dos receptores de estrógeno (RE) e progesterona (RP) e, mais recentemente, o status de Receptor 2 do Fator de Crescimento Epidérmico Humano (HER-2) (Farante et al., 2010). A expressão aumentada de HER-2 ocorre em cerca de 20 a 30% das pacientes sendo responsável por estimular a proliferação celular e associado a um perfil mais agressivo da doença, um pior prognóstico e, por isso, o desenvolvimento de terapias alvo anti-Her-2 vem sendo extensamente estudadas (Slamon et al., 1989; Vu e Claret, 2012).

Atualmente, cada vez mais tem se tentado dividir o câncer de mama em várias doenças, porque é sabido que ele se comporta de várias formas. Os tumores não são iguais, existem casos que um responde bem ao tratamento e nunca mais volta e outros não respondem a tratamento nenhum e a paciente morre em menos de 1 ano. Evidentemente que esses casos são os dois extremos, o que a pesquisa médica tenta entender cada vez mais são os fatores que fazem com que esses tumores sejam diferentes. A individualização já é parcialmente possível, e o tratamento deve ser sempre planejado, levando em consideração os seguintes fatores:

- Expressão dos receptores hormonais
- Expressão e localização do Her-2
- Volume da doença
- Agressividade da doença
- Idade
- Co-morbidades associadas
- Perfil de eventos adversos de cada opção
- Tratamentos previamente utilizados
- Período livre de progressão após o último tratamento

1.4. Tratamento do câncer de mama

O tratamento do câncer de mama evoluiu muito nos últimos anos. O diagnóstico precoce e o uso da quimioterapia neoadjuvante (antes da cirurgia) nos tumores mais avançados têm proporcionado um maior número de cirurgias conservadoras da mama (Teixeira e Pinotti, 2000). Além disso, o surgimento de novas modalidades terapêuticas, como novos medicamentos e novas técnicas de radioterapia, tem levado a uma melhoria na sobrevida e na qualidade de vida, bem como uma diminuição nos índices de recidiva das mulheres portadoras de câncer de mama.

A quimioterapia neoadjuvante é considerada o tratamento padrão para pacientes com câncer de mama localmente avançado (EC II e III) e tem como objetivo principal reduzir o volume tumoral, melhorar as condições cirúrgicas e avaliar "in vivo" a resposta ao tratamento, além de obter respostas patológicas completas já que o prognóstico de sobrevida é dependente dessa remissão (Sanches-Munoz et al., 2013; Buzdar et al., 2007). Entretanto, os efeitos tóxicos da quimioterapia são bastante reconhecidos, representando um fator limitante ao seu uso e muitas vezes comprometendo a função de diversos órgãos (Teixeira e Pinotti, 2000). Dessa forma, tem sido uma área de pesquisa crescente a procura de veículos que direcionem as drogas antineoplásticas ao tumor, evitando o aporte delas aos tecidos normais.

Resposta patológica completa (pCR), se caracteriza pela ausência de tumor residual na mama e na axila após o tratamento neoadjuvante, é reconhecida como marcador prognóstico importante e está associada a maior sobrevida total e livre de doença, principalmente nas pacientes com tumores de comportamento mais agressivo como aquelas com receptor de estrogênio negativo e Her-2 positivo (Von Minckwitz, et al., 2012). Ou seja, de forma geral, pacientes que respondem bem à quimioterapia e que apresentam Resposta patológica completa tem maior tempo de sobrevida tanto livre da doença quanto total (Cortazar, 2014).

A primeira terapia alvo contra o câncer de mama aprovada pelo FDA em 1998 foi o Trastuzumabe (Herceptin®), um anticorpo monoclonal contra a porção extracelular no domínio IV do Her-2 (Vu e Claret, 2012). Desde então, estudos vêm sendo realizados para demonstrar o papel do Trastuzumabe no tratamento neoadjuvante, adjuvante e paliativo (Ver, por exemplo, Blackwell e Bullock, 2008; Slamon et al., 2001; Gelber et al., 2005). Esses trabalhos demonstraram redução do risco de recorrência, maior tempo livre de progressão, maiores taxas de resposta e ganho de sobrevida no grupo que associou Trastuzumabe à quimioterapia. O mecanismo de funcionamento do Trastuzumabe é o seguinte: esse anticorpo

tem atração pelo receptor Her-2, uma proteína que é abundante com funções vitais para as células tumorais, a sua ligação a esta proteína provoca uma série de distúrbios no funcionamento das células tumorais, causando sua morte.

No banco de dados que será introduzido adiante, todas as pacientes estudadas receberam a medicamento Herceptin®. A amostra foi coletada retrospectivamente no Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Ribeirão Preto, incluíram pacientes do sexo feminino com câncer da mama Her-2 positivo que foram atendidas no Ambulatório de Mastologia no período de 2008 a 2012.

Todas as pacientes tinham indicação de utilizar o anticorpo monoclonal anti-Her-2 (Herceptin®) durante o tratamento, no entanto, essa medicação não era padronizada pelo SUS e necessitava de liberação prévia à utilização, na ocasião do estudo. Durante os primeiros ciclos de quimioterapia neoadjuvante era realizada a solicitação do Trastuzumabe como medicação não padronizada. Quando o medicamento estava disponível na ocasião da neoadjuvância ele era prontamente iniciado, no entanto, quando estava disponível apenas no pós-operatório, ele era utilizado na adjuvância por um ano.

A partir de 2007, através de processo administrativo do HCFMRP-USP, a medicação começou a ser fornecida. Em 2009 a medicação foi incorporada pela Secretaria de Estado da Saúde do Governo do Estado de São Paulo, sendo liberada perante protocolo para as instituições cadastradas. E apenas em 2012 o Trastuzumabe foi incorporado pelo SUS para tratamento do câncer de mama inicial mediante o Decreto 7.646. A droga é oferecida no SUS por decisão da Comissão Nacional de Incorporação de Tecnologia (CONITEC) que analisou o custo-efetividade da droga por mais de um ano e também colocou o assunto em consultas públicas.

Em Maio de 2012 a CONITEC publicou os Relatórios de recomendação do Trastuzumabe para tratamento de câncer de mama inicial (Relatório 07) e para câncer de mama avançado (Relatório 08). Nestes relatórios são apresentadas as evidências científicas da eficácia deste medicamento após revisão sistemática da literatura. As consultas públicas que foram realizadas e os preços internacionais também são discutidos. Por fim, a CONITEC decide por recomendar a incorporação do Trastuzumabe para o tratamento do câncer de mama, condicionada à exigência de exame molecular (FISH ou CISH) para confirmação do status Her-2 em tumores com expressão imunohistoquímica com resultado de 2 a 3 cruzes, monitoramento dos resultados clínicos da utilização do medicamento nos hospitais integrantes

do SUS habilitados na alta complexidade em oncologia, e conforme diretrizes diagnósticas e terapêuticas do Ministério da Saúde.

As Portarias 18 e 19 de Julho de 2012 tornam pública a decisão de incorporar a medicamento Trastuzumabe no SUS para o tratamento de câncer de mama localmente avançado e inicial. E em janeiro de 2013 a Portaria 73 estabelece protocolo de uso do Trastuzumabe na quimioterapia de câncer de mama Her-2 positivo inicial e localmente avançado.

1.5. Análise de sobrevivência e apresentação de um conjunto de dados de câncer de mama

A análise estatística dos tempos de sobrevivência (tempos até recidiva, tempos até óbito, tempos até cura) tem o diferencial de permitir a presença de observações censuradas. Os dados censurados são relacionados a indivíduos perdidos ou que não se observa a ocorrência do evento de interesse durante o tempo de seguimento. Esta situação pode ocorrer em diferentes áreas, em estudos de pacientes com câncer, por exemplo, os pesquisadores podem estar interessados na proporção de pacientes curados, pacientes com recidiva, pacientes que morreram devido a doença.

Na oncologia a análise de sobrevivência é muito utilizada na identificação de fatores de riscos, fatores de prognósticos, bem como na comparação de tratamentos (ver, por exemplo, Cox, 1972; Cox e Oakes, 1984; Colosimo e Giolo, 2006), e é de grande utilidade também para a compreensão do modo que os fatores de interesse afetam a sobrevida dos pacientes. Tendo em vista o impacto do câncer de mama na população e a importância do avanço do conhecimento a seu respeito, neste presente trabalho será utilizado um banco de dados de um estudo realizado no Hospital das Clínicas de Ribeirão Preto como motivação para explorar o uso e a aplicação de algumas técnicas estatísticas específicas de análise de sobrevivência mais apropriadas à análise desses dados.

É um estudo retrospectivo realizado no Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Ribeirão Preto, referente a 54 pacientes do sexo feminino com câncer de mama localmente avançado (Estágio II e III) com superexpressão do Her-2 (Her-2 positivo) que iniciaram a quimioterapia neoadjuvante no período de 2008 a 2012, atendidas no Ambulatório de Mastologia do HCFMRP-USP (dados introduzidos na Tabela A.1,

Apêndice A). Todas as pacientes receberam o medicamento Herceptin®. Cada paciente foi acompanhada desde a data de entrada no estudo (início da quimioterapia neoadjuvante) até a data de encerramento do estudo (01/01/2014).

Os dados apresentam duas variáveis resposta de interesse para cada paciente: o tempo de sobrevida livre da doença (SLD) (a paciente pode apresentar recidiva ou não) e o tempo de sobrevida total (ST) (óbito por câncer de mama ou sobrevida até o último tempo de seguimento), dados em meses. As colunas "Recidiva" e "Óbito" dadas na Tabela A.1 introduzida no Apêndice A no final deste trabalho, contêm as informações de censuras associadas respectivamente aos tempos de sobrevida livre da doença e aos tempos de sobrevida total. A recidiva ocorreu em 29% das pacientes e 13% vieram a óbito durante o acompanhamento do estudo, todos os óbitos foram precedidos a recidiva.

Duas observações foram excluídas por conter dados faltantes e as sete covariáves de interesse observadas foram: idade (≤40 anos; >40 anos), uso do medicamento Herceptin® (≥4 ciclos; <4 ciclos na neoadjuvância), estágio da doença (2 ou 3), tipo de cirurgia realizada na paciente (radical; conservadora), resposta patológica completa (sim; não), receptor de estrogênio (positivo; negativo), receptor de progesterona (positivo; negativo).

Tabela 2: Descrição das covariáveis observadas.

Covariáveis Observadas	Todas as pacientes (n=52)	Pacientes com recidiva (n=15)	Pacientes que vieram a óbito (n=7)
40 anos ou mais	40 (76%)	9 (60%)	5 (71%)
4 ou mais ciclos completos do medicamento	37 (75%)	13 (86%)	6 (85%)
Estágio 3 da doença	43 (82%)	14 (93%)	7 (100%)
Cirurgia Radical	35 (67%)	11 (73%)	7 (100%)
Resposta patológica completa	24 (46%)	5 (33%)	2 (29%)
Positivo para receptor de Estrogênio	24 (46%)	5 (33%)	3 (42%)
Positivo para receptor de Progesterona	18 (34%)	3 (7%)	1 (14%)

Na Tabela 2, está a descrição das covariáveis. A maioria das pacientes possui 40 anos ou mais (76%). Em relação ao uso do medicamento Herceptin®, 75% das pacientes receberam pelo menos 4 ciclos do medicamento antes da cirurgia. Pacientes do estágio 3 representam 82% da amostra. Foi realizada cirurgia do tipo radical em 67% das pacientes. O índice de resposta patológica completa corresponde a 46% da amostra e os índices de receptor de

estrogênio e progesterona positivos são 46% e 34% respectivamente. Das pacientes que tiveram recidiva e que vieram a óbito a maioria tem 40 anos ou mais, receberam 4 ou mais ciclos do medicamento, são do estágio 3 da doença e passaram por cirurgia radical. A resposta patológica completa foi observada em 33% das pacientes que tiveram recidiva e em 29% das que vieram a óbito.

Algumas técnicas estatísticas de análise de sobrevivência são, portanto, comumente utilizadas para análise de dados de câncer, com grande destaque ao estimador não-paramétrico produto-limite de Kaplan-Meier (Kaplan e Meier, 1958) para a curva de sobrevivência, o modelo de riscos proporcionais de Cox (Cox, 1972), aos modelos paramétricos baseados na distribuição de Weibull (ver, por exemplo, Lawless, 1982) e testes não paramétricos para comparações entre curvas de sobrevida, como os populares testes de Wilcoxon e do log-rank (ver, por exemplo, Lee e Wenyuwang, 2003). Entretanto, em algumas situações é possível e necessário melhores análises estatísticas com novos modelos introduzidos na literatura. Um caso especial, observado em muitas aplicações é a não verificação do pressuposto de riscos proporcionais, como suposição básica no modelo de Cox. E em outros casos pode ocorrer que uma parte dos indivíduos não seja suscetível ao evento de interesse, como assumidos em alguns modelos paramétricos; nesses casos, modelos que incluem fração de cura são mais adequados à estrutura dos dados.

Modelos de fração de cura, também conhecidos como modelos de mistura de longa duração, assumem que a população em estudo é uma mistura de indivíduos suscetíveis a um evento de interesse, e indivíduos não suscetíveis, em que nunca é observado o evento de interesse. Esses indivíduos não estão em risco com respeito ao evento de interesse e são considerados imunes, não suscetíveis ou curados (Maller e Zhou, 1996).

Em alguns casos, podem-se ter dois tempos de sobrevida associados a cada unidade amostral. Usualmente assume-se independência entre esses tempos, mas em alguns casos o tempo de sobrevida observado para um evento de interesse pode afetar o tempo de sobrevida observado para outro evento de interesse. Neste sentido, modelos paramétricos baseados em distribuições bivariadas podem ser utilizados. Uma distribuição muito popular assumindo dados contínuos é a distribuição exponencial bivariada proposta por Block e Basu (Block e Basu, 1974). Podemos citar outros modelos contínuos que estão presentes na literatura como Freund, 1961; Marshall e Olkin, 1967; Hougaard, 1986; Downton, 1970; Arnold e Strauss,

1988. Outra possibilidade são as distribuições bivariadas discretas (ver, por exemplo, Arnold, 1975 ou Basu e Dhar, 1995).

Todos esses modelos são candidatos para a análise dos dados de câncer de mama da tabela A.1, dado a estrutura dos dados.

Como uma análise preliminar e exploratória dos dados de câncer de mama introduzidos na Tabela A.1, temos na Figura 1, os gráficos dos estimadores não paramétricos de Kaplan-Meier (1958) para as funções de sobrevivência dos tempos de sobrevida livre da doença e tempos de sobrevida total. Por esses gráficos observa-se que a função de sobrevivência decresce ao longo do período de seguimento, mas este decréscimo torna-se mais lento até tornar-se constante. A presença deste "platô" à direita das curvas de sobrevida sugere que em uma parte dos indivíduos amostrados não haverá recidiva da doença, enquanto uma parte dos mesmos indivíduos não deverá ir a óbito devido ao câncer de mama (não necessariamente os mesmos indivíduos). Este comportamento da curva de Kaplan-Meier sugere a presença de uma fração de cura, ou seja, uma proporção de indivíduos em que o evento de interesse não ocorrerá.

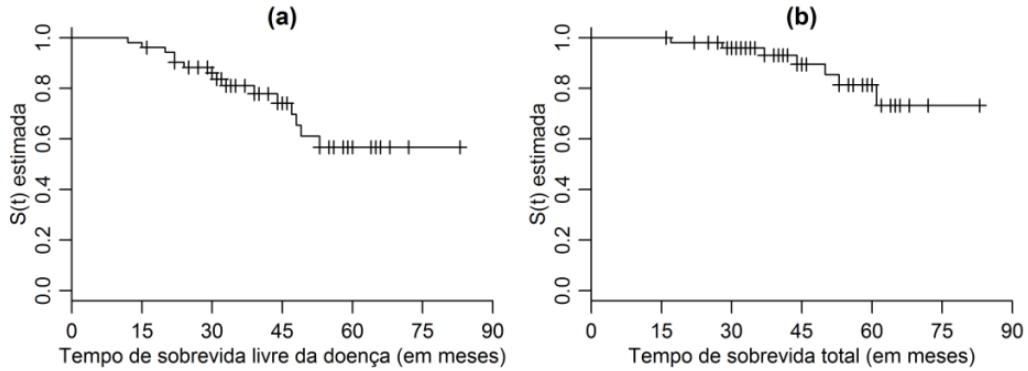

Figura 1: Estimadores de Kaplan-Meier: (a) Tempos de sobrevida livre da doença, (b) Tempos de sobrevida total.

Sendo assim, neste presente trabalho serão explorados modelos de fração de cura, no cenário univariado e também modelos que incorporam uma estrutura de dependência entre os tempos de sobrevida, no cenário bivariado na análise dos dados de câncer de mama.

1.6. Modelo de riscos proporcionais de Cox

Em uma segunda etapa da análise preliminar dos dados de câncer de mama, dado a presença de covariáveis, consideramos o modelo de riscos proporcionais de Cox (Cox, 1972). A grande popularidade desse modelo na análise de dados de sobrevivência na área médica é devido a não necessidade de suposição de uma distribuição paramétrica para os tempos de sobrevivência, uma tarefa nem sempre fácil. Especificamente Cox (1972) assumiu que a distribuição de sobrevivência satisfaz a seguinte condição

$$h(t|x) = h_0(t)\, exp\{\beta x\}, t > 0 \qquad (1)$$

sendo que X é uma covariável (pode também ser um vetor de covariáveis) e h_0 é uma função não-negativa não especificada.

Este modelo é composto pelo produto de dois componentes, um componente não-paramétrico e outro componente paramétrico e por isso denominado como um modelo semi-paramétrico. O componente não-paramétrico, $h_0(t)$, não é especificado e é uma função não-negativa do tempo t. Ele é usualmente chamado de função de risco basal, pois $h(t) = h_0(t)$ quando $x = 0$. Quando a covariável é especificada na forma

$$\theta = exp\{\beta_0 + \beta_1 x\} \qquad (2)$$

β_0 é incorporado na função de risco basal $h_0(t)$. Quando x é modificado, a função de riscos condicional se modifica proporcionalmente. Este modelo é também denominado de modelo de riscos proporcionais, pois a razão das taxas de falha de dois indivíduos diferentes é constante no tempo. Isto é, a razão das funções de taxa de falha para os indivíduos i e j dada por,

$$\frac{h_0(t)exp(\beta x_i)}{h_0(t)exp(\beta x_j)} = exp\{\beta(x_i - x_j)\}, para\ i \neq j \qquad (3)$$

não depende do tempo t.

A suposição básica para o uso do modelo de regressão de Cox é, portanto, que as taxas de falha sejam proporcionais. Este modelo é bastante utilizado em estudos médicos principalmente pela sua flexibilidade devido ao componente não-paramétrico, (ver por exemplo, Kalbfleisch e Prentice, 1980).

A violação da suposição básica, que é a de taxas de falha proporcionais, pode acarretar em sérios vícios na estimação dos coeficientes do modelo (Struthers e Kalbfleisch, 1986). Uma proposta introduzida na literatura para avaliar a suposição de riscos proporcionais no modelo de Cox é a de analisar os resíduos de Schoenfeld (Schoenfeld, 1982). Para definir tais resíduos, considere que o i-ésimo indivíduo com vetor de covariáveis $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})'$ observado falhar (apresentar o evento de interesse), tem-se para esse indivíduo um vetor de resíduos de Schoenfeld $r_i = (r_{i1}, r_{i2}, \dots, x_{ip})$ em que cada componente r_{iq}, para $q = 1, \dots, p$, é definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} exp\{x_j' \hat{\beta}\}}{\sum_{j \in R(t_i)} exp\{x_j' \hat{\beta}\}} \quad (4)$$

Os resíduos são definidos para cada falha e não são definidos para as censuras.

Como usual para resíduos, $\sum_i r_i = 0$. Para permitir que a estrutura de correlação dos resíduos seja considerada, uma forma padronizada dos resíduos de Schoenfeld é definida por,

$$s_i^* = [I(\hat{\beta})]^{-1} \times \boldsymbol{r_i} \quad (5)$$

sendo que $I(\hat{\beta})$ a matriz de informação observada.

O uso dos resíduos padronizados de Schoenfeld para avaliar a suposição de riscos proporcionais é baseado em um resultado apresentado por Grambsch e Therneau (Grambsch e Therneau, 1994) que considera,

$$\lambda(t) = \lambda_o \exp\{x'\beta(t)\} \quad (6)$$

Com a restrição de que $\beta(t) = \beta$, como uma forma alternativa de representar o modelo de Cox. Observe que a restrição $\beta(t) = \beta$ implica na proporcionalidade dos riscos. Quando $\beta(t)$ não é constante, o impacto de uma ou mais covariáveis no risco pode variar com o tempo. Logo, se a suposição de riscos proporcionais é válida, o gráfico de $\beta_q(t)$ versus t deve ser uma linha horizontal. Inclinação zero mostra evidências a favor da proporcionalidade dos riscos.

As técnicas gráficas envolvem conclusões subjetivas, pois dependem da interpretação dos gráficos. Medidas estatísticas bem como a realização de testes de hipóteses são desse modo, de grande utilidade. O coeficiente de correlação de Pearson (r) entre os resíduos padronizados de Schoenfeld e $g(t)$ para cada covariável é uma dessas medidas. No software

livre R, a função $g(t)$ é definida como uma versão contínua à esquerda da curva de sobrevivência de Kaplan-Meier. Valores de r próximos de zero mostram evidências a favor da suposição de riscos proporcionais.

Para testar a hipótese global de proporcionalidade de riscos sobre todas as covariáveis no modelo de Cox, assumindo que $g_q(t) = g(t)$, tem-se a estatística de teste: $T = \frac{(g-\bar{g})' S^* I S^{*\prime} (g-\bar{g})}{d \sum_k (g_k - \bar{g})^2} \sim \chi^2_{(p,1-\alpha)}$ sendo que, I é a matriz de informação observada, d é o número de falhas e $S^* = dRI^{-1}$, sendo R a matriz $d\ x\ p$ dos resíduos de Schoenfeld não padronizados. Sob a hipótese nula de proporcionalidade dos riscos, T tem aproximadamente distribuição qui-quadrado com p graus de liberdade (Grambsch e Therneau, 1994).

Para testar a hipótese de riscos proporcionais para a q-ésima covariável $(q = 1, \dots, p)$ utiliza-se a estatística de teste: $T_q = \frac{d \left(\sum_k (g_k - \bar{g}) S^*_{qk}\right)^2}{I_q^{-1} \sum_k (g_k - \bar{g})^2}$, em que I_q^{-1} é o q-ésimo elemento da diagonal do inverso da matriz de informação observada. Sob a hipótese nula de riscos proporcionais para a q-ésima covariável, T_q tem aproximadamente distribuição qui-quadrado com 1 grau de liberdade. Valores de $T_q > \chi^2_{(1,1-\alpha)}$ mostram evidências contra a suposição de riscos proporcionais para a covariável q.

1.7. Aplicação do modelo de riscos proporcionais de Cox aos dados de câncer de mama

O modelo de riscos proporcionais de Cox abrange um grande número de situações práticas onde pode ser utilizado, aqui ele será ajustado ao conjunto de dados de câncer de mama para evidenciar o efeito das covariáveis sobre o tempo de sobrevida e será verificada a sua adequabilidade a este conjunto de dados.

Este conjunto de dados possui dois tempos de sobrevida para cada paciente e sete covariáveis, os tempos de sobrevida serão tratados independentes. Uma primeira visualização do comportamento das covariáveis nos tempos de sobrevida livre da doença é dada pelo gráfico dos estimadores não paramétricos de Kaplan-Meier (1958), na Figura 2. As curvas de Kaplan-Meier para diferentes níveis da covariável Idade se cruzam, indicando que possivelmente o pressuposto de riscos proporcionais não é verificado.

Os estimadores de máxima verossimilhança (EMV) para os parâmetros do modelo de regressão de Cox considerando o tempo de sobrevida livre da doença são dados na Tabela 3,

onde se observa que nenhuma covariável traz evidências de efeitos nos tempos de sobrevida livre da doença (valor p > 0,05 para testes de hipóteses de que os parâmetros de regressão sejam iguais à zero). O pressuposto de riscos proporcionais precisa ser verificado antes de qualquer interpretação do modelo ajustado, para isso se optou pelo método gráfico (Figura 3) e pelo teste de hipóteses dos resíduos de Schoenfeld (Tabela 4).

Tabela 3: EMV para os parâmetros do modelo de regressão de riscos proporcionais de Cox - Tempos de sobrevida livre da doença.

Covariável	**Coeficiente**	**HR**	**Erro Padrão**	**Valor p**	**Intervalo de Confiança 95% de HR**	
					Limite Inferior	**Limite Superior**
Idade	-0,62	0,54	0,57	0,27	0,18	1,64
Herceptin	-0,27	0,76	0,82	0,74	0,15	3,84
Estágio	0,56	1,75	1,10	0,61	0,20	15,26
Cirurgia	0,28	1,33	0,62	0,65	0,39	4,49
Resposta patológica completa	-0,42	0,65	0,59	0,47	0,21	2,08
Receptor de Estrogênio	-0,40	0,67	0,72	0,58	0,16	2,73
Receptor de Progesterona	-0,27	0,77	0,86	0,76	0,14	4,14

Na Figura 3, observa-se que novamente a covariável Idade mostra indícios de não proporcionalidade nos riscos devido a inclinação da reta não ser nula. A confirmação da não proporcionalidade dos riscos na covariável Idade se dá pela rejeição da hipótese nula $(H_0: r = 0)$ com um nível de significância igual à 0,05 (valor p < 0,05 na Tabela 4).

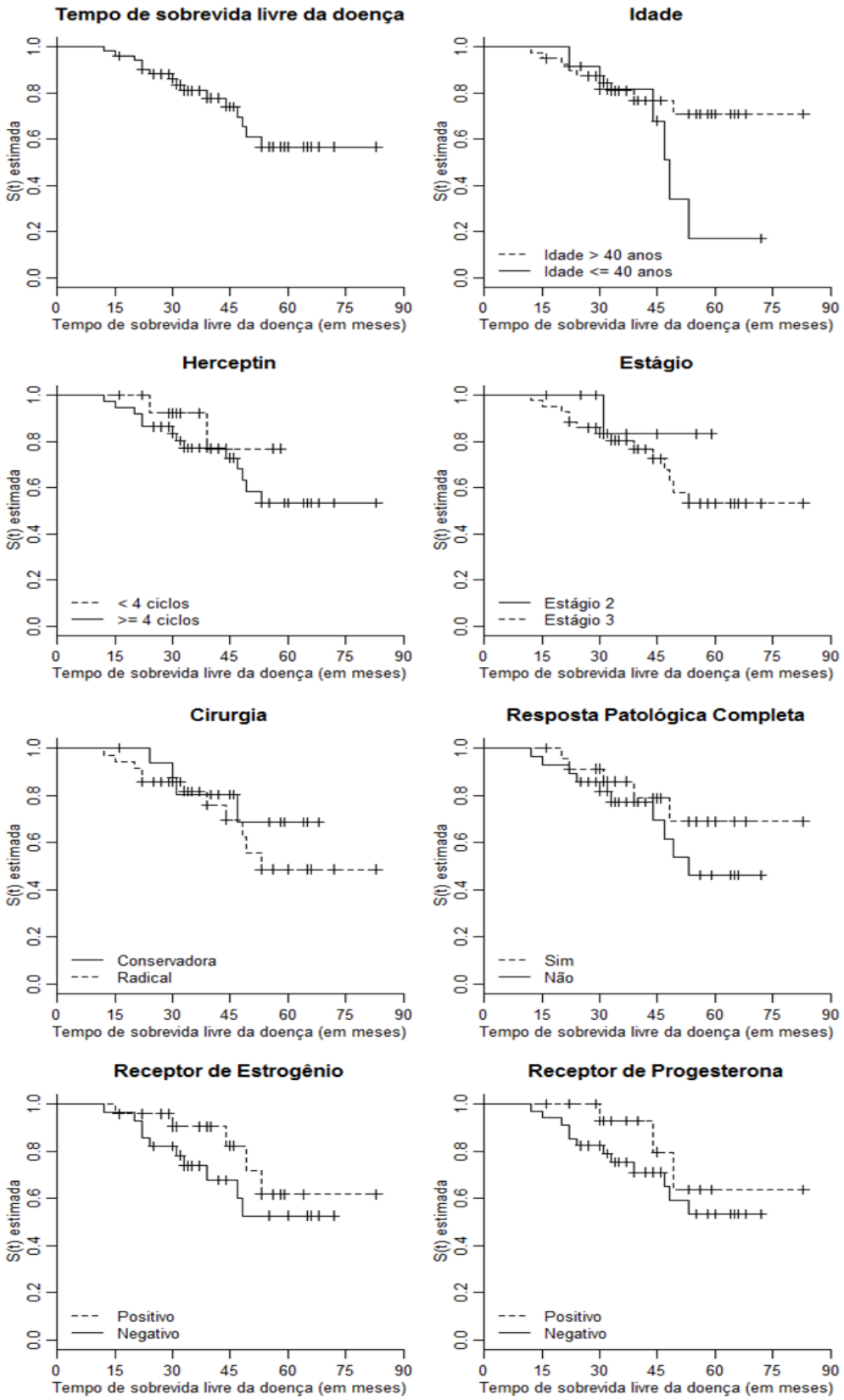

Figura 2: Estimadores de Kaplan-Meier das covariáveis nos tempos de sobrevida livre da doença.

Tabela 4: Testes de proporcionalidade dos riscos no modelo de Cox para o tempo de sobrevida livre da doença.

Covariável	r	χ^2	Valor p
Idade	-0,51	4,11	0,04
Herceptin	0,11	0,20	0,66
Estágio	0,01	0,00	0,98
Cirurgia	-0,01	0,00	0,98
Resposta patológica completa	0,04	0,02	0,88
Receptor de Estrogênio	0,10	0,11	0,74
Receptor de Progesterona	0,24	0,95	0,33
GLOBAL	NA	6,24	0,51

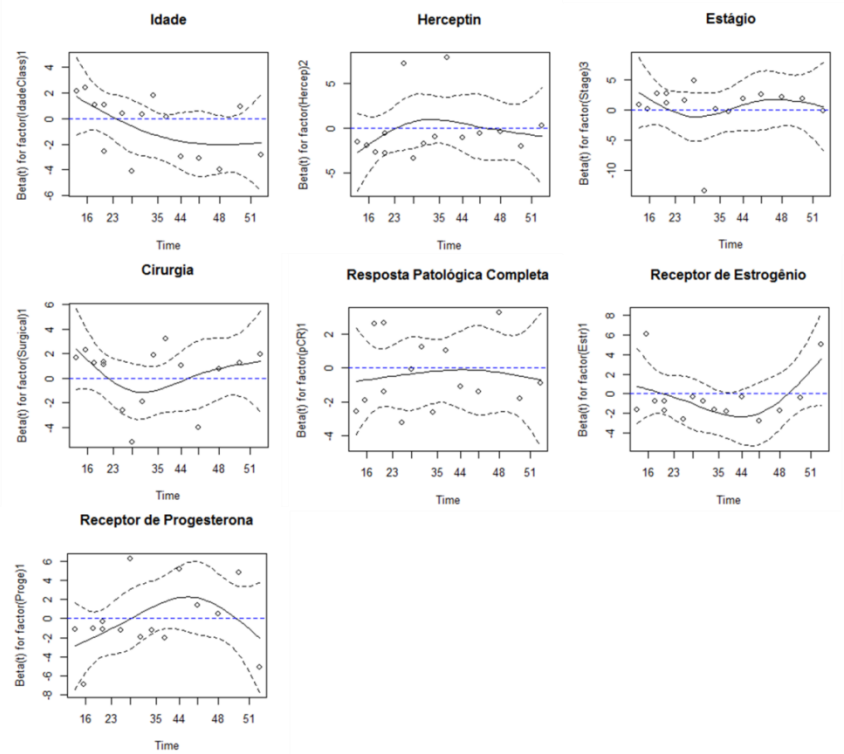

Figura 3: Gráfico de resíduos de Schoenfeld das covariáveis nos tempos de sobrevida livre da doença.

Dessa forma, o modelo de riscos proporcionais de Cox não se adequa aos tempos de sobrevida livre da doença, pois a covariável Idade não possui riscos proporcionais. Esta covariável é de extrema importância para o pesquisador, sendo assim não podendo ser deixada de fora do modelo estatístico.

Para o tempo de sobrevida total também foi observado o comportamento das covariáveis pelo gráfico dos estimadores não paramétricos de Kaplan-Meier (1958), na Figura 4. Nas covariáveis Idade, Herceptin, Resposta Patológica Completa e Receptor de Estrogênio as

curvas de Kaplan-Meier se cruzam, indicando que possivelmente o pressuposto de riscos proporcionais não é verificado.

Os estimadores de máxima verossimilhança (EMV) para os parâmetros do modelo de regressão de Cox para os tempos de sobrevida total são dados na Tabela 5, nas covariáveis Estágio e Cirurgia o algoritmo computacional (método iterativo) usado para encontrar os EMV do modelo não convergiu porque todas as pacientes com o tempo completo (que vieram a óbito) são do estágio 3 da doença e passaram por cirurgia radical (ver Tabela 2), condenando todas as inferências calculadas neste modelo. Para verificar o pressuposto de riscos proporcionais se optou pelo método gráfico (Figura 5) e pelo teste de hipóteses dos resíduos de Schoenfeld (Tabela 6).

Tabela 5: EMV para os parâmetros do modelo de regressão de riscos proporcionais de Cox - Tempos de sobrevida total.

Covariável	**Coeficiente**	**HR**	**Erro Padrão**	**Valor p**	**Intervalo de Confiança 95% de HR**	
					Limite Inferior	**Limite Superior**
Idade	0,46	1,58	0,91	0,62	0,26	9,43
Herceptin	0,60	1,82	1,24	0,63	0,16	20,74
Estágio	18,42	10×10^{7}	24×10^{3}	1	0	Inf
Cirurgia	20,93	$12{,}25 \times 10^{8}$	16×10^{3}	1	0	Inf
Resposta patológica completa	-0,73	0,48	0,99	0,46	0,07	3,37
Receptor de Estrogênio	0,44	1,56	0,96	0,64	0,24	10,12
Receptor de Progesterona	-1,64	0,19	1,32	0,21	0,01	2,57

Tabela 6: Testes de proporcionalidade dos riscos no modelo de Cox para o tempo de sobrevida total.

Covariável	r	χ^2	**Valor p**
Idade	-0,46	1,48	0,22
Herceptin	0,29	0,58	0,45
Estágio	-0,68	0,00	1,00
Cirurgia	0,39	0,00	1,00
Resposta patológica completa	0,18	0,18	0,67
Receptor de Estrogênio	0,86	5,25	0,02
Receptor de Progesterona	-0,23	0,53	0,47
GLOBAL	NA	7,63	0,37

O modelo proposto por Cox também não se adequa aos tempos de sobrevida total, pois o algoritmo computacional (método iterativo) usado para encontrar os EMV dos parâmetros do modelo não convergiu, nas covariáveis Estágio e Cirurgia e a covariável Receptor de Estrogênio não possui riscos proporcionais.

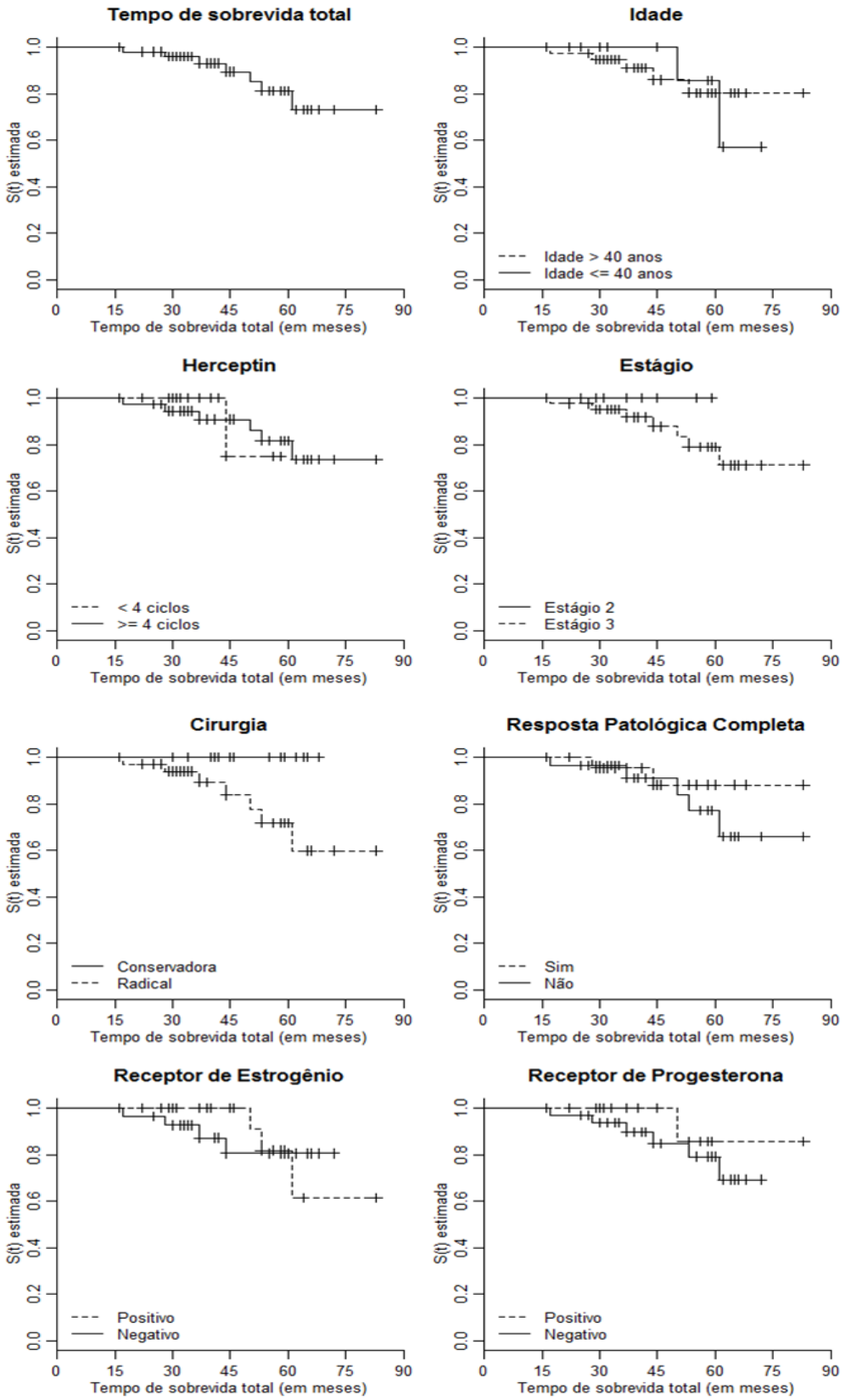

Figura 4: Estimadores de Kaplan-Meier das covariáveis nos tempos de sobrevida total.

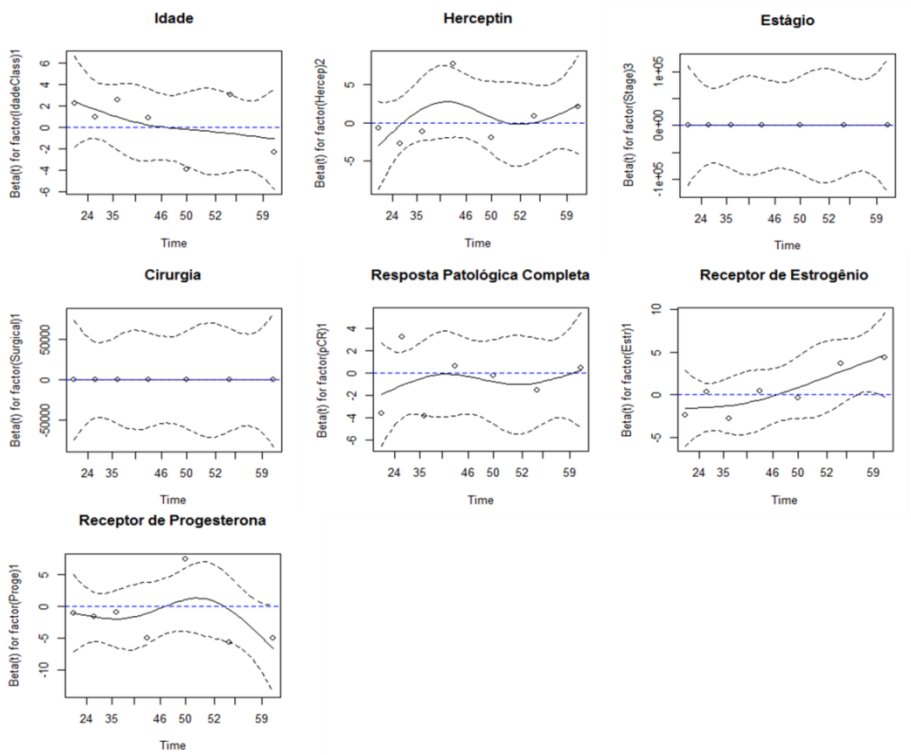

Figura 5: Gráficos de resíduos de Schoenfeld das covariáveis nos tempos de sobrevida total.

É importante salientar que apesar do modelo de riscos proporcionais de Cox ser o modelo mais utilizado na análise de sobrevivência de dados médicos, Efron (1977) mostrou que se consegue mais eficiência na obtenção dos estimadores de parâmetros de regressão em modelos paramétricos, sob certas circunstâncias, do que no modelo de Cox.

1.8. Uso de modelos de sobrevivência paramétricos

Como observado anteriormente, o uso do modelo de riscos proporcionais de Cox não é adequado para a análise dos dados de câncer de mama introduzidos na Tabela A.1. Assim serão explorados nesse trabalho alguns modelos paramétricos de sobrevivência para os dados apresentados, considerando todas as suas características: várias covariáveis, presença de censuras, fração de cura, tempos independentes e tempos com alguma estrutura de dependência.

Dois casos especiais serão explorados na análise dos dados: univariado e bivariado. Para isso serão considerados modelos baseados em distribuições paramétricas.

No caso univariado, os modelos serão baseados na distribuição de Weibull. No caso bivariado, serão considerados modelos baseados na distribuição exponencial bivariada para

dados contínuos proposta por Block e Basu (1974) e nas distribuições geométricas bivariadas para dados discretos propostas respectivamente, por Arnold (1975) e Basu-Dhar (1995).

As inferências para os modelos propostos de regressão com dados de sobrevivência na presença de censuras serão obtidas usando métodos de inferência frequentista e métodos de inferência bayesiana (ver, por exemplo, Paulino et al, 2003).

Sob o enfoque bayesiano, vamos usar métodos MCMC (Monte Carlo em Cadeias de Markov) para a obtenção das quantidades a posteriori de interesse (ver, por exemplo, Gelfand e Smith, 1990; Casela e George, 1992; Chib e Greenberg, 1995).

2. Objetivos

O objetivo principal do presente estudo é a aplicação de modelos estatísticos adequados às características do banco de dados introduzido na Tabela A.1, na busca de evidências de fatores relevantes que possam afetar os tempos de sobrevida livre da doença e total das mulheres participantes do estudo que geraram o banco de dados utilizado. O uso desses modelos ajustados aos dados tem como finalidade levar o pesquisador a obter informações importantes que possam auxiliar o desenvolvimento de metodologias e terapias mais eficientes contra o câncer de mama.

Serão considerados modelos para a análise dos dados de sobrevivência na presença de fração de cura, censuras e várias covariáveis sob uma abordagem frequentista e bayesiana. Nessa direção, serão explorados modelos univariados, supondo a independência entre os tempos de sobrevida, onde cada tempo será analisado separadamente e também serão explorados modelos bivariados, onde os tempos possuem uma estrutura de dependência entre si.

2.1. Caso univariado

Para o caso univariado vários modelos baseados na distribuição de Weibull serão considerados para os tempos de sobrevida livre da doença e total:

Modelos sem a presença de covariáveis:

- Modelo de Weibull sob o enfoque frequentista
- Modelo de Weibull sob o enfoque bayesiano
- Modelo de Weibull bayesiano na presença de fração de cura

Modelos com a presença de covariáveis:

- Modelo de Weibull sob o enfoque frequentista
- Modelo de Weibull sob o enfoque bayesiano
- Modelo de Weibull bayesiano na presença de fração de cura afetando o parâmetro de escala
- Modelo de Weibull bayesiano na presença de fração de cura afetando o parâmetro de escala e a probabilidade de cura.

2.2. Caso bivariado

Para o caso bivariado serão considerados modelos que assumem distribuições de probabilidade para dados contínuos ou discretos, todos sob o enfoque bayesiano:

- Modelo com distribuição exponencial bivariada de Block e Basu
- Modelo com distribuição geométrica bivariada de Arnold
- Modelo com distribuição geométrica bivariada de Basu-Dhar

3. Material e Métodos

3.1. Conceitos básicos em análise de Sobrevivência

A análise de sobrevivência é uma técnica estatística aplicada a situações quando se pretende analisar dados relacionados ao tempo de ocorrência de algum evento de interesse, isto é, ao tempo transcorrido entre um evento inicial, no qual o indivíduo entra em um estado particular e um evento final, que modifica este estado.

Em análise de sobrevivência, a variável resposta é, geralmente, o tempo de sobrevida. Define-se sobrevida como o intervalo de tempo desde a entrada do indivíduo no estudo até a ocorrência do evento de interesse, podendo este evento ser o tempo de falha ou óbito, ou o tempo até o término do estudo. O diferencial das técnicas de análise de sobrevivência em relação á outras técnicas estatísticas é a possibilidade de considerar dados censurados, ou seja, indivíduos que apresentam apenas informação parcial da resposta. Isto se refere às situações em que por alguma razão houve a perda de seguimento durante o estudo, ou seja, o acompanhamento do paciente foi interrompido, seja porque o paciente mudou de cidade ou o paciente morreu por uma causa que não seja a estudada. Sem a presença de censuras, as técnicas estatísticas clássicas, como a análise de regressão e planejamento de experimentos, poderiam ser utilizadas na análise desses tipos de dados (Colosimo e Giolo, 2006).

Os dados censurados, resultados provenientes de um estudo de sobrevivência devem ser usados na análise, pois fornecem informações sobre o tempo de sobrevida de pacientes e a sua omissão no cálculo das estatísticas de interesse pode acarretar conclusões viciadas. Existem várias formas de censuras, sendo a mais usual a censura à direita, que ocorre quando o evento de interesse não é observado até o término do estudo ou até o ultimo instante em que o individuo é acompanhado. Censuras aleatórias são frequentes na área médica; elas acontecem quando um paciente é retirado no decorrer do estudo sem ter ocorrido o evento de interesse ou também, podem ocorrer caso o paciente apresente a falha devido à outra doença diferente da doença estudada.

Na análise de sobrevivência, o tempo de vida ou tempo de sobrevida é denotado por uma variável aleatória não negativa $T \geq 0$ que pode ser expressa através da função densidade de probabilidade $f(t)$, da função de sobrevivência $S(t) = P(\ T > \ t)$ ou a função de risco, $h(t)$.

A função densidade de probabilidade é definida como o limite da probabilidade de observar o evento de interesse em um indivíduo no intervalo de tempo $[t, t + \Delta t]$ por unidade de tempo, expressa por,

$$f(t) = lim_{\Delta t \to 0} \frac{P(t \leq t + \Delta t)}{\Delta t} \quad (7)$$

em que $f(t) \geq 0$, para todo t, e tem área abaixo da curva igual a 1 para $t > 0$.

A função de sobrevivência S(t) é definida como a probabilidade de um indivíduo sobreviver pelo menos até um tempo *t* qualquer, isto é, a probabilidade de ocorrer o evento além de t, e é dada por,

$$S(t) = P(T > t) = 1 - F(t) \quad (8)$$

em que $F(t) = P(T \leq t)$ é a função distribuição acumulada em t.

Da função de sobrevivência $S(t)$ é possível obter a função densidade de probabilidade $f(t)$, da relação,

$$f(t) = -\frac{d}{dt} S(t) = \frac{d}{dt} F(t) \quad (9)$$

em que $\frac{d}{dt}$ denota a derivada da função em relação à t.

A função de risco é utilizada para descrever como o risco do evento muda com o tempo t. Essa função é definida como a probabilidade do evento ocorrer no intervalo de tempo $[t, t + \Delta t]$, dado que o individuo tenha sobrevivido pelo menos até o tempo t, e é dada por,

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (10)$$

A função de risco também pode ser obtida da relação entre a função densidade de probabilidade $f(t)$ e a função de sobrevivência $S(t)$,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} log\, S(t) \quad (11)$$

3.1.1. Estimador não paramétrico de Kaplan-Meier para a função de sobrevivência

O passo inicial de qualquer análise estatística consiste em uma descrição ou estudo preliminar dos dados. A presença de observações censuradas impede o uso das técnicas convencionais de descrição, como médias, histogramas e Box-plots, entre outros. O estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), também chamado de estimador produto-limite de Kaplan-Meier, permite estimar a função de sobrevivência e, a partir dela, estimar as quantidades de interesse que usualmente são o tempo médio ou mediano, alguns percentis ou certas frações de falhas em tempos fixos de acompanhamento (Colosimo e Giolo, 2006).

Sejam $t_1 < t_2 < \cdots < t_k$, os k tempos distintos e ordenados de falhas; d_j denotando o número de falhas em t_j, $j = 1,2,\ldots,k$ e n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não apresentaram o evento de interesse e não foram censurados até o instante imediatamente anterior a t_j. O estimador produto-limite de Kaplan-Meier (EKM) é, então, definido por:

$$\hat{S}(t) = \prod_{j:t_j<t}\left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t}\left(1 - \frac{d_j}{n_j}\right) \qquad (12)$$

O EKM possui as seguintes propriedades estatísticas: é um estimador não viciado para amostras grandes, é fracamente consistente, converge assintóticamente para um processo gaussiano e é estimador de máxima verossimilhança de $S(t)$. Usualmente o EKM é representado em um gráfico mostrando o comportamento da curva de sobrevivência.

Além de descrever os tempos de sobrevida, o EKM é utilizado para identificar o comportamento dos tempos de acordo com categorias de covariáveis de interesse, produzindo assim, evidências de possíveis fatores que possam afetar os tempos de sobrevida estudados.

3.1.2. Técnicas paramétricas em análise de sobrevivência

Modelos paramétricos assumem que os dados seguem uma distribuição de probabilidade conhecida. Algumas das principais distribuições de probabilidade usadas em análise de sobrevivência são apresentadas a seguir.

Distribuição exponencial

Seja T uma variável aleatória denotando o tempo de falha com função densidade de probabilidade dada por,

$$f(t) = \frac{1}{\alpha} exp\left\{-\left(\frac{t}{\alpha}\right)\right\},\ t > 0 \qquad (13)$$

em que α é o tempo médio de sobrevida ($\propto > 0$).

Esta distribuição apresenta um único parâmetro e se caracteriza por ter uma função de risco constante, também chamada de taxa de falha instantânea; nesta distribuição na linguagem de confiabilidade industrial, tanto uma unidade velha quanto uma unidade nova, que ainda não falhou, têm o mesmo risco de falhar em um tempo futuro. Esta propriedade á chamada de falta de memória. A função de risco é dada por,

$$h(t) = \frac{1}{\alpha}, \qquad t \geq 0 \qquad (14)$$

A função de sobrevivência é dada por,

$$S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \qquad (15)$$

Denota-se a distribuição exponencial por $T \sim Exp(\alpha)$.

Distribuição de Weibull

A distribuição de Weibull foi proposta originalmente por Weibull (1951). Sua popularidade em aplicações práticas se deve ao fato dela apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua função de riscos pode ser monótona crescente, decrescente e constante. A função densidade de probabilidade é dada por,

$$f(t_i) = \frac{\alpha\, t_i^{\alpha-1} \exp[-\left(\frac{t_i}{\lambda}\right)^{\alpha}]}{\lambda^{\alpha}} \qquad (16)$$

em que, $t_i > 0$ denota os tempos de sobrevida. Os parâmetros $\lambda > 0$ e $\alpha > 0$ denotam respectivamente, os parâmetros de escala e de forma para a distribuição. Diferentes valores de α levam a diferentes formas para a distribuição o que a torna muito flexível na análise de

dados para tempos de sobrevida. Na análise de sobrevivência o grande interesse é focado na função de sobrevivência $S(t^*) = P(T > t^*)$ em que t^* é um tempo qualquer fixado. Assumindo a distribuição de Weibull com f.d.p. (16), a função de sobrevivência é dada por,

$$S(t^*) = exp\left\{-\left(\frac{t^*}{\lambda}\right)^{\alpha}\right\} \tag{17}$$

A função de risco $h(t)$ ou taxa instantânea de falha, da distribuição de Weibull (ver, por exemplo, Lawless, 1982) é dada, de $h(t) = f(t)/S(t)$, por:

$$h(t) = \alpha \frac{t^{\alpha-1}}{\lambda^{\alpha}} \tag{18}$$

Observar que se $\alpha = 1$, temos a distribuição exponencial, isto é, a distribuição exponencial é um caso especial da distribuição de Weibull. A função de risco $h(t)$ dada por (18) é estritamente crescente para $\alpha > 1$, estritamente decrescente para $\alpha < 1$ e constante para $\alpha = 1$. Assim, observa-se uma grande flexibilidade de ajuste aos dados. A média e a variância da distribuição de Weibull com densidade dada por (16) são dadas respectivamente por:

$$\mu = E(T) = \lambda\Gamma\left(1 + \frac{1}{\alpha}\right) \tag{19}$$

$$\sigma^2 = Var(T) = \lambda^2\left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma\left[1 + \frac{1}{\alpha}\right]^2\right\} \tag{20}$$

em que *Γ(.)* denota uma função gama, $\Gamma(z) = \int_0^{\infty} e^{-t}\ t^{z-1}\, dt$.

Distribuição Log-normal

A função densidade de probabilidade de uma variável aleatória T com distribuição log-normal é dada por,

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} exp\left\{-\frac{1}{2}\left(\frac{log\, t - \mu}{\sigma}\right)^2\right\},\ t > 0 \tag{21}$$

em que $\mu > 0$ e $\sigma > 0$ são respectivamente a média e o desvio-padrão para os logaritmos dos tempos de sobrevida.

As funções de sobrevivência e função de risco neste caso, não apresentam uma forma analítica explícita, sendo expressas por,

$$S(t) = \Phi\left(\frac{-\log t+\mu}{\sigma}\right) \quad e \quad h(t) = \frac{f(t)}{S(t)} \tag{22}$$

em que $\Phi(.)$ é a função distribuição acumulada de uma distribuição normal padrão (distribuição normal com média zero e variância igual a um). A função de risco não é monótona como a da distribuição Weibull, ou seja, ela cresce, atinge um valor máximo e depois decresce.

Distribuição Log-logística

Se T é uma variável aleatória, tal que $ln(T)$ tem distribuição logística, então T segue uma distribuição Log-logística, com função de densidade de probabilidade dada por,

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}\left(1+\left(\frac{t}{\alpha}\right)^{\gamma}\right)^{-2}, t > 0 \tag{23}$$

em que $\alpha > 0$ é o parâmetro de forma e $\gamma > 0$ o de escala. As funções de sobrevivência e de risco são dadas, respectivamente por,

$$S(t) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} \quad e \quad h(t) = \frac{\gamma\left(\frac{t}{\alpha}\right)^{\gamma-1}}{\alpha\left[1+\left(\frac{t}{\alpha}\right)^{\gamma}\right]} \tag{24}$$

em que, para $\gamma > 1$, tem-se padrão similar ao da distribuição log-normal, isto é, o risco é crescente alcançando um pico e a partir daí começa a declinar; para $\gamma < 1$, o risco é decrescente, similar a função de risco da distribuição Weibull.

3.2. Estimação dos parâmetros dos modelos probabilísticos

Os modelos probabilísticos apresentados na seção anterior possuem quantidades desconhecidas, denominados parâmetros. Os parâmetros devem ser estimados a partir das observações amostrais, para que o modelo fique determinado e, assim, seja possível responder às perguntas de interesse.

Existem alguns métodos de estimação conhecidos na literatura (Colossimo e Giolo, 2006) sendo que o mais apropriado para dados com censuras é o método de máxima verossimilhança. A metodologia de estimação incorpora os dados censurados, é relativamente simples em termos de interpretação e possui propriedades ótimas para grandes amostras.

3.2.1. Método de máxima verossimilhança em modelos de sobrevivência

Supor uma amostra de observações não censuradas $t_1, \dots, t_n$ de uma população, onde os tempos de sobrevida tenham uma densidade $f(t;\ \theta)$, onde θ é um parâmetro desconhecido. A função de verossimilhança para o parâmetro θ é dada por

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \qquad (25)$$

Na expressão (25), θ pode estar representando um único parâmetro ou um vetor de parâmetros (ver, por exemplo, Colosimo e Giolo, 2006).

Para definir a verossimilhança para dados censurados, considere T uma variável aleatória representando o tempo de falha de um paciente e C uma variável aleatória, independente de T, representando o tempo de censura. Para um dado paciente temos como dado observado, $t = min(T, C)$ e

$$\delta = \begin{cases} 1\ se\ T \leq C \\ 0\ se\ T > C \end{cases} \qquad (26)$$

sendo que δ é uma variável indicadora de falha.

Supor que os pares (T_i, C_i), para $i = 1, \dots, n$ formam uma amostra aleatória de tamanho n. As observações podem ser divididas em duas partes: as r primeiras observações ordenadas são as observações não censuradas $(1, 2, \dots, r)$ e as $n - r$ seguintes são observações censuradas $(r\ +1,\ r\ +2, \dots,\ n)$.

Para todos os mecanismos de censura (censuras de tipo I onde o tempo de seguimento é fixado, censuras de tipo II onde o número de falhas é fixado no início do experimento ou censuras aleatórias) a expressão para a função de verossimilhança é dada por,

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta) \qquad (27)$$

ou equivalentemente por,

$$L(\theta) = \prod_{i=1}^{n}[f(t_i;\theta)]^{\delta_i}\ [S(t_i;\theta)]^{1-\delta_i} = \prod_{i=1}^{n}[h(t_i;\theta)]^{\delta_i}\ S(t_i;\theta) \qquad (28)$$

em que δ_i é a variável indicadora de falha dada em (26).

Na prática é sempre conveniente considerar o logaritmo da função de verossimilhança. Os valores que maximizam $L(\theta)$ ou equivalentemente $l(\theta) = logL(\theta)$ são os estimadores de máxima verossimilhança. Eles são encontrados resolvendo-se o seguinte sistema de equações,

$$\mathrm{U}(\theta) = \frac{\partial \log \mathrm{L}(\theta)}{\partial \theta} = 0 \qquad (29)$$

sendo θ um vetor de parâmetros. Um caso particular é dado quando temos apenas um parâmetro.

3.3. Modelos de regressão paramétrica em análise de sobrevivência

A construção de modelos de regressão em análise de sobrevivência busca ajustar os dados a modelos paramétricos existentes com finalidade de obter inferências para quantidades populacionais de interesse e também conhecer como o tempo de sobrevida está relacionado com uma ou mais covariáveis de interesse. Com o uso de modelos de regressão paramétricos, é possível a identificação de quais covariáveis afetam o tempo de sobrevida bem como a intensidade e a direção de cada uma delas em explicar a ocorrência do evento estudado (Hougaard, 1999; Colossimo e Giolo, 2006; Louzada, Mazucheli e Achcar, 2002).

Um modelo de regressão bastante utilizado na análise de sobrevivência na presença de covariáveis como foi enfatizado nas seções 1.6 e 1.7 é o modelo de regressão de riscos proporcionais de Cox (Cox, 1972). Enquanto os modelos paramétricos assumem uma distribuição conhecida de probabilidade para os tempos de sobrevida, o modelo semi-paramétrico de Cox, tem como característica principal, o pressuposto de proporcionalidade dos riscos entre as categorias de uma determinada covariável sem assumir uma distribuição de probabilidade específica para o tempo de sobrevida T o que caracteriza um modelo não-paramétrico. Este modelo é também denominado modelo de riscos proporcionais, pois a razão das taxas de falha de dois indivíduos diferentes é constante no tempo, ou seja, se o risco de um indivíduo for duas vezes o risco de outro indivíduo no início do estudo, esta razão entre os riscos permanecerá constante para todo o período de acompanhamento. Entretanto, em algumas aplicações não se verifica o pressuposto de riscos proporcionais, como assumido no

modelo de Cox como foi observado na análise preliminar apresentada na seção 1.7 para os dados de câncer de mama introduzidos na tabela A.1.

Sendo assim, quando conhecemos a distribuição dos tempos de sobrevida, o ajuste de um modelo paramétrico pode trazer mais informações sobre a natureza da distribuição do comportamento da função de risco ao longo do tempo. Além disso, é um modelo mais flexível, dado sua facilidade em incorporar o efeito das covariáveis em seus parâmetros.

Do ponto de vista paramétrico, os modelos de sobrevivência são constituídos por dois componentes: um aleatório e outro determinístico (ver, por exemplo, Louzada, Mazuchelli e Achcar, 2002), onde o componente determinístico é dado por,

$$\eta = g(ax) \tag{30}$$

onde η é um dado parâmetro de uma distribuição de probabilidade; $g(.)$ é uma função positiva e contínua, geralmente assumida igual a $\exp(\beta x)$, $\beta = (\beta_0, \beta_1, \dots, \beta_k)^t$ é um vetor de parâmetros de regressão a serem estimados e associados a um vetor k covariáveis $x = (x_1, x_2, \dots, x_k)^t$. Note que $x = (x_1, x_2, \dots, x_k)^t$ estabelece um efeito multiplicativo no parâmetro η, e é responsável pela aceleração ou desaceleração do tempo de sobrevida.

Desse modo, uma função log-linear é convenientemente utilizada para escrever a relação entre η e o vetor de covariáveis x, de tal maneira que para o $i-$ésimo indivíduo temos,

$$ln[\eta(x_i)] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{31}$$

Em geral, é comum assumir que as covariáveis afetam apenas o parâmetro de locação de uma determinada distribuição, porém, em muitas aplicações, assumir também que o parâmetro de escala seja afetado pelas covariáveis o pode ser mais apropriado na análise dos dados (Louzada, Mazuchelli e Achcar, 2002).

3.4. Modelos de fração de curas

De acordo com Maller e Zhou (1996), em um modelo de fração de cura assume-se que uma fração p de indivíduos na população é curada ou nunca observou o evento de interesse; logo, $(1 - p)$ é a fração de indivíduos não curados (susceptíveis). Esse tipo de modelo possui grande vantagem em relação aos modelos paramétricos usuais por incorporarem a heterogeneidade das duas subpopulações; indivíduos susceptíveis e indivíduos curados.

Portanto, a função de sobrevivência nesse caso pode ser escrita considerando uma mistura na forma,

$$S(t) = p + (1-p)S_0(t) \qquad (32)$$

em que $p \in (0,1)$ é o parâmetro de mistura (proporção de imunes) e $S_0(t)$ é a função de sobrevivência basal para a população de indivíduos não curados (indivíduos suscetíveis).

Considerando uma amostra aleatória de tempos de sobrevida (t_i, δ_i), $i = 1, \dots, n$, a contribuição do $i-$ésimo indivíduo para a função de verossimilhança é dada por (28).

A partir da função de sobrevivência definida em (32), é possível obter a função densidade de probabilidade, utilizando o resultado $f(t_i) = -\frac{d}{dt}S(t_i)$, dada por:

$$f(t_i) = (1-p)\, f_0(t_i) \qquad (33)$$

em que $f_0\ (t_i)$ é a função densidade de probabilidade para os indivíduos suscetíveis.

Substituindo a função de densidade (33) e a função de sobrevivência (32) na função de verossimilhança (28) obtêm-se a seguinte função de verossimilhança para o modelo de mistura com fração de cura (ou longa duração):

$$L_i = \prod_{i=1}^{n}[(1-p)f_0(t_i)]^{\delta_i}\,[p + (1-p)\, S_0(t_i)]^{1-\delta_i} \qquad (34)$$

Portanto, a função log-verossimilhança considerando todas as observações é dada por:

$$l_i = r\, log(1-p) + \sum_{i=1}^{n}\delta_i\, log f_0(t_i) + \sum_{i=1}^{n}(1-\delta_i)\, log[p + (1-p)\, S_0(t_i)] \qquad (35)$$

em que, $r = \sum_{i=1}^{n}\delta_i$ é o número de observações não censuradas.

3.5. Uso de métodos Bayesianos em análise de sobrevivência: alguns conceitos básicos

A estatística bayesiana tem sido cada vez mais utilizada como uma alternativa a estatística clássica ou frequentista. Os métodos bayesianos têm se mostrado muito eficazes e poderosos na análise de dados, principalmente na área da saúde, onde em muitos casos o tamanho amostral é pequeno, nessas condições, teorias assintóticas (presentes na frequentista) podem não ser recomendadas.

Na prática, a maior diferença entre as duas estatísticas é que a bayesiana tenta medir o grau de incerteza que se tem sobre a ocorrência de um determinado evento do espaço amostral, utilizando distribuições de probabilidades a priori e a informação amostral (verossimilhança). A inferência bayesiana se caracteriza por calcular uma função densidade de probabilidade conjunta (densidade a posteriori) sobre todos os possíveis vetores de parâmetros (espaço dos parâmetros). Na inferência bayesiana, a incerteza sobre os parâmetros desconhecidos associa-se uma distribuição de probabilidade (Gianola e Fernando, 1986), enquanto que, na inferência frequentista, os parâmetros são valores fixos ou constantes, aos quais não se associam a qualquer distribuição (Blasco, 2001). No contexto bayesiano, o objetivo é, condicionalmente aos dados y observados, descrever a incerteza sobre o valor de algum parâmetro θ não observado, em termos de probabilidades ou densidades (Box e Tiao, 1992). O parâmetro θ pode ser um escalar ou um vetor de parâmetros.

A informação acerca de um parâmetro θ, também chamada de distribuição a priori, é incorporada ao estudo através do uso do teorema de Bayes, que combina a informação contida nos dados, resultando na distribuição a posteriori. Dessa forma é possível incorporar na análise de dados o conhecimento de um pesquisador ou especialista, quando disponível. A fundamentação da teoria de inferência bayesiana é baseada na fórmula de Bayes.

3.5.1. Fórmula de Bayes

Sejam os eventos $A_1, A_2, \dots, A_k$ formando uma sequência de eventos mutualmente exclusivos e exaustivos formando uma partição do espaço amostral Ω, isto é, $\bigcup_{j=1}^{k} A_j = \Omega$ e $A_i \cap A_j = \emptyset$ (conjunto vazio) para $i \neq j$ tal que $P\left(\bigcup_{j=1}^{k} A_j\right) = \sum_{j=1}^{k} A_j = 1$. Então para qualquer outro evento $B(B \subset \Omega)$, temos

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} \quad (36)$$

para todo i variando de 1 até k.

Seja θ um vetor de parâmetros s serem estimados. Logo, pelo teorema de Bayes, tem-se a seguinte distribuição de probabilidade a posteriori para θ.

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta})f(y|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})f(y|\boldsymbol{\theta})d\boldsymbol{\theta}} \quad (37)$$

assumindo que θ seja contínuo, $\pi(\theta)$ é a distribuição a priori conjunta para θ e $f(y|\theta) = L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$ a função de verossimilhança de θ.

Assim, a partir da fórmula de Bayes, temos,

$$\pi(\boldsymbol{\theta}|y) \propto L(\boldsymbol{\theta}|Y)\pi(\boldsymbol{\theta}) \quad (38)$$

Assim temos distribuição a posteriori $\propto$ verossimilhança x distribuição a priori, sendo que o símbolo $\propto$ representa proporcional.

A função de probabilidade a priori representa o conhecimento prévio a respeito dos elementos de θ antes da observação dos dados, refletindo a incerteza em relação aos possíveis valores de θ antes do vetor de dados y ser selecionado. A função a posteriori incorpora o estado de incerteza do conhecimento prévio a respeito do parâmetro θ após a observação dos dados em y e a função de verossimilhança representa a contribuição de y para o conhecimento sobre θ.

3.5.2. Distribuições a priori

Uma distribuição a priori para um parâmetro pode ser elicitada de várias formas:

(a) Podemos assumir distribuições a priori definidas no domínio de variação do parâmetro de interesse. Como caso particular, poderíamos considerar uma distribuição a priori Beta que é definida no intervalo (0, 1) para proporções que também são definidas no intervalo (0, 1) ou considerar uma priori normal para parâmetros definidos em toda reta;

(b) Podemos construir uma priori baseada em informações de um ou mais especialistas;

(c) Podemos considerar métodos estruturais de elicitação de distribuições a priori (ver, por exemplo, Paulino et al, 2003);

(d) Podemos considerar distribuições a priori não informativas quando temos total ignorância sobre os parâmetros de interesse;

(e) Podemos usar métodos bayesianos empíricos em dados ou experimentos prévios para construir a priori de interesse.

3.5.3. Métodos de simulação para amostras da distribuição a posteriori

Na obtenção de sumários a posteriori é necessário resolver integrais múltiplas, muitas vezes, complicadas, o que exige o uso de métodos numéricos ou de aproximações de integrais, especialmente quando a dimensão do vetor de parâmetros é grande.

Daí surge a necessidade do uso de métodos computacionais poderosos, como os métodos de Monte Carlo em cadeias de Markov (MCMC) que incluem alguns algoritmos de simulação de amostras da distribuição a posteriori conjunta de interesse, como os algoritmos de Metropolis-Hastings e o amostrador de Gibbs. É importante salientar que os métodos com base em simulação de amostras da distribuição a posteriori conjunta de interesse, como, por exemplo, o método de Monte Carlo em cadeias de Markov (MCMC), passaram a ser muito utilizados com o avanço dos recursos computacionais em termos de harware e software. Esses métodos consistem na simulação de uma variável aleatória através de uma cadeia de Markov, no qual a sua distribuição assintóticamente se aproxima da distribuição a posteriori de interesse (ver, por exemplo, Bernardo e Smith, 1994).

A cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia depende somente do estado atual e dos dados. No entanto, como existe certa dependência com os valores iniciais fixados no processo de simulação, na prática uma amostra simulada inicial é descartada após um período de aquecimento, chamada “Burn-in- sample”.

As formas mais usuais de simulação dos métodos MCMC são dadas pelo amostrador de Gibbs e o algoritmo de Metropolis-Hasting. Essas duas formas simulam amostras da distribuição a posteriori conjunta a partir das distribuições condicionais (ver, por exemplo, Gelfand e Smith, 1990; Chib e Greenberg, 1995).

O amostrador de Gibbs nos permite gerar amostras da distribuição a posteriori conjunta desde que as distribuições condicionais completas possuam formas fechadas ou conhecidas. Por outro lado, o algoritmo de Metropolis-Hasting permite gerar amostras da distribuição a posteriori conjunta com distribuições condicionais completas possuindo ou não uma forma conhecida ou fechada.

O amostrador de Gibbs

Suponha que $\theta = (\theta_1, \dots, \theta_k)$ é um vetor de parâmetros aleatórios e y é o vetor dos dados observados; tem-se como objetivo, obter inferências sobre a distribuição a posteriori conjunta $\pi(\theta|y) = \pi(\theta_1, \dots, \theta_k|y)$ (Bernardo e Smith, 1994).

Dado um vetor arbitrário de valores iniciais $\theta_1^{(0)}, \dots, \theta_k^{(0)}$ para as quantidades desconhecidas, implementa-se o seguinte procedimento iterativo:

Obtém-se $\theta_1^{(1)}$ de $\pi\left(\theta_1 \middle| \boldsymbol{y}, \theta_2^{(0)}, \dots, \theta_k^{(0)}\right)$

Obtém-se $\theta_2^{(1)}$ de $\pi\left(\theta_2 \middle| \boldsymbol{y}, \theta_1^{(1)}, \ \theta_3^{(0)}, \dots, \theta_k^{(0)}\right)$

Obtém-se $\theta_3^{(1)}$ de $\pi\left(\theta_3 \middle| \boldsymbol{y}, \theta_1^{(1)}, \ \theta_2^{(1)}, \theta_4^{(0)}, \dots, \theta_k^{(0)}\right)$

$\vdots$

Obtém-se $\theta_k^{(1)}$ de $\pi\left(\theta_k \middle| \boldsymbol{y}, \theta_1^{(1)}, \ \dots, \theta_{k-1}^{(1)}\right)$

Obtém-se $\theta_1^{(2)}$ de $\pi\left(\theta_1 \middle| \boldsymbol{y}, \theta_2^{(1)}, \ \dots, \theta_k^{(1)}\right)$

$\vdots$

e assim por diante.

Agora, suponha que este processo é continuado através de t iterações e é independentemente replicado m vezes para que ao final se tenha m replicações do vetor amostrado $\theta^t = (\theta_1^{(t)}, \dots, \theta_k^{(t)})$, onde θ^t é uma realização de uma cadeia de Markov com probabilidade de transição dada por,

$$p(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}) = \prod_{l=1}^{k} \pi(\theta_{kl}^{t+1} | \boldsymbol{y}, \theta_1^{t+1}, \dots, \theta_{l-1}^{t+1}, \theta_{l+1}^{t}, \dots, \theta_k^{t}) \qquad (39)$$

Como, como $t \to \infty$, $(\theta_1^{(t)}, \dots, \theta_k^{(t)})$ tende em distribuição a um vetor aleatório cuja densidade conjunta é $\pi(\theta|y)$, ou seja, a distribuição a posteriori de interesse. Em particular, θ_i^t tende em distribuição a uma quantidade aleatória cuja densidade é $\pi(\theta_i|y)$, também chamada de densidade marginal a posteriori de θ_i. Desta maneira, para t grande, as replicações $(\theta_{i1}^{(t)}, \dots, \theta_{im}^{(t)})$ são aproximadamente uma amostra aleatória de $\pi(\theta_i|y)$.

Após a geração de amostras da distribuição a posteriori de interesse, utilizamos essas amostras para obter estimadores de Monte Carlo para sumários a posteriori de interesse como a média a posteriori, o desvio-padrão a posteriori e intervalos de credibilidade de interesse.

O algoritmo Metropolis-Hastings

Supor que se deseja simular uma densidade a posteriori $\pi(\theta|y)$. Um algoritmo de Metropolis-Hastings se inicia com um valor inicial θ^0 e especifica uma regra para a simulação do $t-$ é$simo$ valor da sequência θ^t dado o $(t-1)-$ é$simo$ valor da sequência θ^{t-1}. Esta regra consiste em uma densidade proposta (ou densidade geradora) a qual simula um valor candidato θ^* e o cálculo da uma probabilidade de aceitação P, que indica a probabilidade do valor candidato ser aceito para ser o próximo valor na sequência. Especificamente, esse algoritmo pode ser descrito da seguinte forma (ver, por exemplo, Albert, 2007),

1. Simular um valor candidato θ^* de uma densidade proposta $p(\theta^*|\theta^{t-1})$.
2. Calcular a razão

$$R = \frac{\pi(\theta^*|\boldsymbol{y})p(\theta^{t-1}|\theta^*)}{\pi(\theta^{t-1}|\boldsymbol{y})p(\theta^*|\theta^{t-1})} \qquad (40)$$

3. Calcular a probabilidade de aceitação $P = min\{R, 1\}$
4. Amostrar um valor θ^t tal que $\theta^t = \theta^*$ com probabilidade P, caso contrário $\theta^t = \theta^{t-1}$.

Sob certas condições de regularidade facilmente satisfeitas na densidade proposta $p(\theta^*|\theta^{t-1})$, a sequência simulada $\theta^1, \theta^2, \ldots$ convergirá a uma variável aleatória que é distribuída de acordo com a distribuição a posteriori $\pi(\theta|y)$ (ver, por exemplo, Bernardo e Smith, 1994; Chib e Greenberg, 1995).

4. Modelos para análise univariada dos dados de câncer de mama

Nesta seção serão apresentados alguns modelos univariados dos tempos de sobrevida das pacientes com câncer de mama dados na Tabela A.1. Todos os modelos assumem uma distribuição de Weibull apresentada na seção 3.1.2.

4.1. Modelos sem a presença de covariáveis

Sob o enfoque Frequentista

Para a análise sob o enfoque frequentista, os estimadores para os parâmetros λ e α da equação (17) foram obtidos usando o método de máxima verossimilhança, maximizando a função de verossimilhança obtida a partir das equações (16) e (17) dada por,

$$L(\alpha,\lambda) = \prod_{i=1}^{n}\left[\frac{\alpha}{\lambda^{\alpha}} t_i^{\alpha-1} exp\left(-\frac{t_i}{\lambda}\right)^{\alpha}\right]^{\delta_i}\left\{exp\left[-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right]\right\}^{1-\delta_i} =$$

$$= \prod_{i=1}^{n}\left[\frac{\alpha}{\lambda^{\alpha}} t_i^{\alpha-1}\right]^{\delta_i} exp\left[-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right] \qquad (41)$$

na presença de dados censurados usando métodos numéricos implementados em softwares estatísticos.

Sob o enfoque Bayesiano

Para a análise sob o enfoque bayesiano, foi considerada a densidade da distribuição Weibull em uma forma reparametrizada de (16) e para a obtenção dos sumários a posteriores de interesse utilizou-se métodos MCMC (Monte Carlo em Cadeias de Markov) (ver, por exemplo, Gelfand e Smith,1990; Casela e George, 1992; Chib e Greenberg, 1995) com o uso do software OpenBugs (Spiegelhalter et al, 2003). Assim considera-se a densidade,

$$f(t_i) = \alpha\theta {t_i}^{\alpha-1} exp\{-\theta {t_i}^{\alpha}\} \qquad (42)$$

em que $\theta = 1/\lambda^{\alpha}$.

Distribuição de Weibull para os indivíduos suscetíveis assumindo um modelo de fração de cura

Um caso especial, é quando se assume uma distribuição de Weibull para indivíduos suscetíveis com função de densidade de probabilidade dada por (42) e função de sobrevivência,

$$S_0(t) = exp[-\theta t^{\alpha}] \quad (43)$$

Assumindo o modelo de misturas (32), o logaritmo da função de verossimilhança para p, α e θ é dado por:

$$l(p,\theta,\alpha) = r\,ln(1-p) + r\,ln(\alpha) + r\,ln(\theta) + (\alpha-1)v - \theta A_1(\theta) + A_2(p,\theta,\alpha) \quad (44)$$

sendo que $A_1(\theta) = \sum_{i=1}^{n} \delta_i\, t_i^{\alpha}$, $A_2(p,\theta,\alpha) = \sum_{i=1}^{n}(1-\delta_i)\, ln[p + (1-p)e^{-\theta t_i^{\alpha}}]$, $r = \sum_{i-1}^{n} \delta_i$ e $v = \sum_{i=1}^{n} \delta_i\, ln(t_i)$

Na presença de um vetor de covariáveis $x = (x_1, \ldots, x_k)$ que afeta os parâmetros p e θ, mas não afeta o parâmetro de forma α, vamos assumir o seguinte modelo de regressão:

$$\theta_i = \beta_0\, exp(\beta_1 x_{1i} + \cdots + \beta_k x_{ki}) \text{ e } ln\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_k x_{ki} \quad (45)$$

4.2. Modelos com a presença de covariáveis

Sob o enfoque Frequentista

Considerando os dados de câncer de mama introduzidos na Tabela A.1, assumir o modelo de regressão de Weibull definido por:·.

$$log(t_i) = \beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i + \sigma^* \varepsilon_i \quad (46)$$

sendo que, t_i denotam os tempos de sobrevida, $i = 1, \ldots, n$; $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ e β_7, são parâmetros de regressão.

O parâmetro σ^* está relacionado com o parâmetro de forma da distribuição de Weibull com densidade (16) pela relação $\sigma^* = 1/\alpha$. O termo ε_i em (46) é uma quantidade aleatória com distribuição de valor extremo (ver Nelson, 2004 ou Lawless, 1982) também definida

como distribuição de valor extremo de tipo I (mínimo) ou distribuição de Gumbel (ver, Gumbel, 1954) com função densidade de probabilidade dada por:

$$f(\varepsilon) = exp(\varepsilon - exp(\varepsilon)), -\infty < \varepsilon < \infty \quad (47)$$

Também observar que o parâmetro de escala λ definido em (16) está relacionado com as covariáveis a partir da relação,

$$\lambda_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i) \quad (48)$$

Isto é, o modelo de regressão definido por (46) define um modelo de regressão no parâmetro de escala (ver, por exemplo, Colosimo e Giolo, 2006) assumindo mesmo parâmetro de forma.

Para o modelo de regressão (46), estimamos os parâmetros de regressão $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \; e \; \beta_7$, e o parâmetro σ^* usando métodos de máxima verossimilhança (ver, por exemplo, Mood, Graybill e Boes, 1974). Estimadores de máxima verossimilhança para os parâmetros $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ e β_7, e σ^* são obtidos maximizando-se a função de verossimilhança, $L(\theta) = \Pi f(\varepsilon_i)$ onde $f(\varepsilon_i) = exp[\varepsilon_i - exp(\varepsilon_i)]$, $i = 1, \ldots, n$, $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \; e \; \beta_7, \sigma^*)$ e,

$$\sigma^* \varepsilon_i = log(t_i) - [\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i] \quad (49)$$

Na prática, em geral maximiza-se o logaritmo da função de verossimilhança na determinação dos estimadores de máxima verossimilhança usando algum método numérico (por exemplo, método de Newton-Raphson), usualmente disponível em softwares estatísticos existentes, como o software Minitab®.

Sob o enfoque Bayesiano

Assumindo uma distribuição de Weibull com densidade dada em (42), na presença de covariáveis, o modelo de regressão é dado por,

$$\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i \quad (50)$$

Assumindo uma distribuição de Weibull (42) na presença de covariáveis, fração de cura (32) e o modelo de regressão afetando o parâmetro de escala da distribuição Weibull, temos que o modelo de regressão é dado por,

$$\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i) \quad (51)$$

Assumindo uma distribuição de Weibull (42) na presença de covariáveis, fração de cura (32) e os modelos de regressão afetando o parâmetro de escala da distribuição Weibull e a fração de cura p, os modelos de regressão são dados respectivamente por,

$$\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i) \quad (52)$$

e

$$logito(p_i) = \gamma_0 + \gamma_1 idade_i + \gamma_2 hercep_i + \gamma_3 estágio_i + \gamma_4 cirur_i + \gamma_5 pCR_i + \gamma_6 estrog_i + \gamma_7 progest_i \quad (53)$$

5. Resultados da análise univariada dos dados de câncer de mama

Nesta aplicação serão considerados os dois tempos de sobrevida disponíveis no conjunto de dados, o tempo de sobrevida livre da doença e o tempo de sobrevida total (dados na Tabela A.1 no Apêndice A). Para a análise univariada dos dados, primeiramente será feita uma análise com os tempos de sobrevida livre da doença, em seguida, com os tempos de sobrevida total. Serão considerados modelos baseados na distribuição Weibull, sem covariáveis e com covariáveis, sem a presença de fração de cura e com fração de cura; sob o enfoque frequentista e bayesiano.

5.1. Análise estatística dos tempos de sobrevida livre da doença (SLD)

5.1.1. Distribuição de Weibull sem a presença de covariáveis sob o enfoque Frequentista

Usando o software Minitab®, encontramos os estimadores de máxima verossimilhança (EMV) e algumas estatísticas de interesse dadas na Tabela 7. Dos resultados da Tabela 7, observa-se que o tempo médio estimado de sobrevida livre da doença é de 73,05 meses. O tempo mediano é de 68,26 meses. A distribuição Weibull é uma distribuição assimétrica, sendo assim, para futuras conclusões iremos considerar os tempos medianos, em vez das médias que podem não ser apropriadas como medidas de centralidade, que serão obtidos a partir da relação: $\widehat{S(t)} = 0{,}5$.

Tabela 7: EMV para os parâmetros da distribuição de Weibull - Tempos de sobrevida livre da doença.

Parâmetro	Estimativa	Erro Padrão	Intervalo de Confiança (95%)	
			Limite Inferior	Limite Superior
Forma	1,95	0,4189	1,2795	2,9709
Escala	82,38	14,0980	58,9076	115,2130
Média	73,05	12,7516	51,8838	102,850

Nos tempos de sobrevida livre da doença, 37 pacientes não apresentaram o evento de interesse (recidiva), aproximadamente 71% da amostra. Este é um indicativo de uma possível necessidade de complementar o modelo com fração de cura. Outra possibilidade na procura de possíveis melhores inferências é reanalisar os dados sob o enfoque bayesiano (ver, por exemplo, Paulino, Turkman e Murteira, 2003).

5.1.2. Distribuição de Weibull sem a presença de covariáveis sob o enfoque Bayesiano

Para uma análise bayesiana, consideramos distribuições a priori gama $G(0{,}1;\ 0{,}1)$ aproximadamente não informativas para α e θ, onde $G(a, b)$ denota uma distribuição gama com média igual à a/b e variância igual à a/b^2. Na simulação de amostras da distribuição a posteriori para α e θ, consideramos uma amostra de aquecimento "burn-in sample" de tamanho 1.000 para eliminar o efeito do valor inicial no processo iterativo; após essa amostra de aquecimento, geramos outras 600.000 amostras tomando amostras de 100 em 100, totalizando uma amostra final de tamanho 6.000 que será utilizada para obter as quantidades a posteriori de interesse (uso do software OpenBugs). Na Tabela 8, temos os sumários a posteriori de interesse.

Utilizando distribuições a priori não informativas o tempo médio estimado de sobrevida livre da doença é de 81,71 meses e o tempo mediano é de 74,51 meses, sendo assim, o modelo sob o enfoque bayesiano estimou um tempo mediano maior do que o modelo sob o enfoque frequentista.

Tabela 8: Sumários a posteriori de interesse - Tempos de sobrevida livre da doença.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
Forma	1,84	0,3717	1,12	2,64
Escala	90,84	20,2915	65,79	144,10
Média	81,71	20,2226	58,27	139,62

5.1.3. Modelo de Weibull com fração de cura sem a presença de covariáveis sob o enfoque Bayesiano

Para uma segunda análise bayesiana dos tempos de sobrevida livre da doença vamos agora assumir uma distribuição de Weibull (16) sem a presença de covariáveis e na presença de fração de cura. Para a análise bayesiana consideramos as seguintes distribuições a priori: $\alpha \sim Gama(1{,}1)$, $p \sim Beta(70{,}30)$ e $\theta \sim U(0{,}300)$, onde $U(a, b)$ denota uma distribuição uniforme no intervalo (a, b) e $\mathrm{Beta}(\mathrm{a}, \mathrm{b})$ denota uma distribuição beta com média igual à $a/\ (a + b)$ e variância igual á $ab/[(a + b)2(a + b + 1)]$. Observar que os hiper-parâmetros da distribuição beta dados por $a = 70$ e $b = 30$ foram escolhidos levando a uma priori informativa (uso de métodos bayesianos empíricos, ver, por exemplo, Carlin e Louis, 2002) para p, com média igual à 0,70 (um valor próximo da proporção observada de dados censurados, interpretados como pacientes imunes ou curados).

Na simulação de amostras da distribuição a posteriori de interesse, consideramos uma amostra de aquecimento "burn-in sample" de tamanho 1.000, foram geradas outras 600.000 amostras tomadas de 100 em 100 totalizando uma amostra final de tamanho 6.000. Na Tabela 9, temos os sumários a posteriori de interesse. Dos resultados da Tabela 9, observa-se que a proporção estimada de indivíduos "curados" é de 67%, resultado próximo do valor observado nos dados.

O tempo mediano estimado por esse modelo com a presença de fração de cura foi de 51,19 meses, o menor tempo mediano estimado dentre os três modelos apresentados até aqui.

Tabela 9: Sumários a posteriori de interesse modelo com fração de cura sem covariáveis - Tempos de sobrevida livre da doença.

Parâmetro	**Média**	**Desvio Padrão**	**Intervalo de Credibilidade (95%)**	
			Limite Inferior	**Limite Superior**
Forma	2,52	0,5685	1,508	3,771
Escala	42,94	6,5655	32,97	58,51
p (fração de cura)	0,67	0,0409	0,5825	0,7411

Sendo assim, as funções de sobrevivência considerando modelos baseados na distribuição Weibull sob uma abordagem frequentista, bayesiana sem e com fração de cura (ver estimadores em Tabela 7, Tabela 8, Tabela 9) são dadas, respectivamente por:

- Frequentista sem fração de cura: $S(t) = exp[-(\frac{t}{82,38})^{1,95}]$
- Bayesiano sem fração de cura: $S(t) = exp[-(\frac{t}{90,84})^{1,84}]$
- Bayesiano com fração cura: $S(t) = 0,67 + (1 - 0,67)exp[-(\frac{t}{42,94})^{2,52}]$

Na Figura 6, temos os gráficos das funções de sobrevivência estimadas considerando os estimadores Kaplan-Meier e os modelos Weibull sob uma abordagem frequentista, bayesiano sem e com presença de fração de cura. Observa-se que o modelo com fração de cura acompanha melhor o estimador de Kaplan-Meier inclusive em sua curvatura que após 45 meses tende a diminuir o ritmo de decaimento. Enquanto que os outros dois modelos sem a presença de fração de cura apresentam um decaimento bem acentuado especialmente para tempos de sobrevida grandes, o que não corresponde com o comportamento e a realidade dos dados estudados.

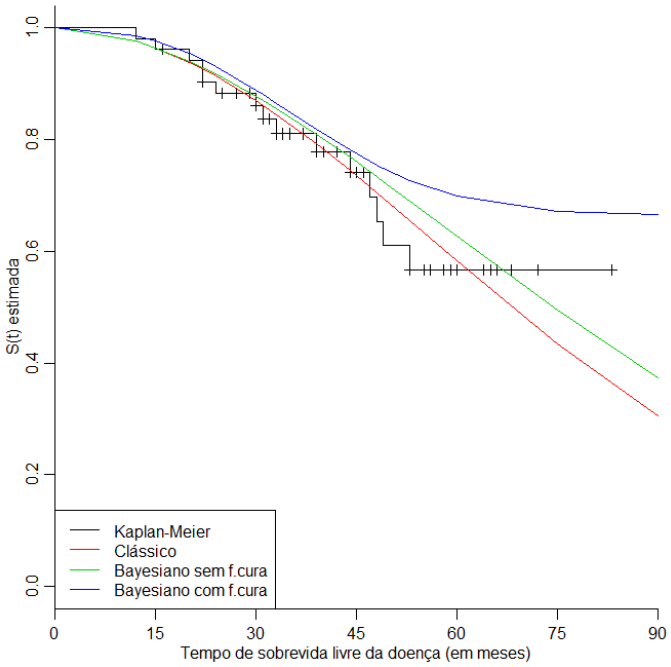

Figura 6: Gráficos da função de sobrevivência estimada - Kaplan e Meier, Weibull frequentista, Weibull Bayesiano sem e com fração de cura (tempos de sobrevida livre da doença).

A seguir, incluiremos as covariáveis observadas a fim de identificar possíveis fatores que afetem o tempo de sobrevida livre da doença, ou seja, fatores que possam influenciar o tempo até a recidiva do câncer de mama nas pacientes após a cirurgia.

5.1.4. Modelo de Weibull na presença de covariáveis sob o enfoque Frequentista

Para iniciar a investigação de possíveis fatores que afetam o tempo de sobrevida livre da doença, vamos assumir o modelo de regressão dado em (46). Dos resultados da Tabela 10, observa-se que todas as covariáveis não mostram efeitos significativos, pois todos os intervalos de confiança para os parâmetros de regressão correspondentes contém o valor zero. Além disso, nenhum valor-p é inferior do que 0,05 (nível de significância usual) evidenciando a não significância de todas as covariáveis neste modelo.

Neste conjunto de dados há uma grande proporção de censuras (71%), o que pode dificultar a descoberta de possíveis covariáveis significativas afetando os tempos de sobrevida livres da doença. Na análise de sobrevivência com dados médicos é comum essa dificuldade, pois estes dados geralmente possuem uma grande proporção de censuras e diversas

covariáveis de interesse. Por isso a necessidade cada vez maior de modelos e técnicas estatísticas mais adequadas para analisar dados com estas características.

Tabela 10: EMV para os parâmetros de regressão de Weibull - Tempos de sobrevida livre da doença.

Parâmetro	Estimativa	Erro Padrão	Z	P	Intervalo de Confiança (95%)	
					Limite Inferior	Limite Superior
β_0 (intercepto)	4,60	1,6835	2,73	0,006	1,2992	7,8985
Idade	0,33	0,303	1,09	0,277	-0,2651	0,9238
Herceptin	0,08	0,425	0,19	0,846	-0,7508	0,9154
Estágio	-0,20	0,577	-0,35	0,725	-1,3341	0,9277
Cirurgia	-0,18	0,3207	-0,55	0,581	-0,8056	0,4518
Resposta patológica completa	0,23	0,3063	0,74	0,462	-0,3752	0,8255
Receptor de Estrogênio	0,14	0,3774	0,36	0,715	-0,6021	0,8775
Receptor de Progesterona	0,19	0,4567	0,41	0,679	-0,7065	1,0839
Forma	1,94	0,4233			1,2663	2,9766

(Z:estatística Z; P: valor-p)

Na procura de possíveis melhores inferências, vamos reanalisar os dados sob o enfoque bayesiano.

5.1.5. Modelo de Weibull na presença de covariáveis sob o enfoque Bayesiano

Assumindo distribuições a priori não-informativas normais $\mathrm{N}(0{,}1)$ para todos os parâmetros de regressão β_r , $r = 0{,}1{,}2{,}\ldots{,}7$; uma priori $\mathrm{Gama}(1{,}1)$ para o parâmetro de forma α e usando o software OpenBugs (burn-in sample =1.000 e 6.000 amostras finais tomadas de 100 em 100) e o modelo de regressão dado em $\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i$ (50) , temos na Tabela 11 os sumários a posteriori de interesse.

Dos resultados da Tabela 11, a covariável estágio tem efeito significativo, isto é, o intervalo de credibilidade para o parâmetro de regressão β_3 não inclui o valor zero. Sendo assim, observa-se que o modelo bayesiano detectou covariáveis significativas mesmo assumindo distribuições a priori não informativas para os parâmetros do modelo, sendo que o modelo de regressão sob o enfoque frequentista mostrou não significância para todas as covariáveis (ver Tabela 10).

Tabela 11: Sumários a posteriori de interesse - Tempos de sobrevida livre da doença - Modelo de regressão.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
β_0 (intercepto)	-1,4	0,9062	-3,1800	0,3691
Idade	-0,62	0,4736	-1,5390	0,3062
Herceptin	-0,86	0,6658	-2,2810	0,3722
Estágio	-1,01	0,4422	-1,8710	-0,1220
Cirurgia	0,18	0,4889	-0,7601	1,1610
Resposta patológica completa	-0,61	0,4891	-1,5890	0,3473
Receptor de Estrogênio	-0,34	0,5584	-1,4510	0,7302
Receptor de Progesterona	-0,49	0,6271	-1,7700	0,7180
Forma	1,26	0,2569	0,7973	1,8050

5.1.6. Modelo de Weibull com fração de cura e com covariáveis afetando o parâmetro de escala da distribuição Weibull

Considerar agora uma análise bayesiana dos dados assumindo uma distribuição de Weibull na presença de covariáveis, fração de cura e o modelo de regressão afetando o parâmetro de escala da distribuição Weibull $\theta = \frac{1}{\lambda^{\alpha}}$ dado em $\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i)$ (51).

Assumindo distribuições a priori não-informativas normais $\mathrm{N}(0,1)$ para todos os parâmetros de regressão βr , $r = 0,1,2,\ldots,7$; $\theta \sim Gama(1,1)$, $p \sim Beta(70,30)$ e baseado em uma amostra (burn-in sample =1.000 e 6.000 amostras finais tomadas de 100 em 100), temos na Tabela 12 os sumários a posteriori de interesse.

Tabela 12: Sumários a posteriori de interesse - Tempos de sobrevida livre da doença - Modelo de regressão na presença de fração de curas afetando o parâmetro de escala.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
β_0 (intercepto)	0,93	0,8956	-0,8246	2,6940
Idade	-0,08	0,3585	-0,7525	0,6769
Herceptin	0,40	0,5216	-0,5865	1,5180
Estágio	0,79	0,3697	0,0622	1,5090
Cirurgia	-0,32	0,4163	-1,1710	0,5006
Resposta patológica completa	0,54	0,3811	-0,1921	1,3150
Receptor de Estrogênio	0,48	0,4443	-0,3829	1,3940
Receptor de Progesterona	0,23	0,4823	-0,6975	1,2290
Escala	1,94	0,5117	1,0960	3,0770
p (fração de cura)	0,66	0,0422	0,5743	0,7381

Dos resultados da Tabela 12, observa-se que a covariável estágio tem um efeito significativo (intervalo de credibilidade para o parâmetro de regressão β_3 correspondente não inclui o valor zero).

5.1.7. Modelo de Weibull com fração de cura e com covariáveis afetando o parâmetro de escala da distribuição Weibull e a probabilidade de cura

Considerar agora uma análise bayesiana dos tempos de sobrevida livre da doença assumindo uma distribuição de Weibull na presença de covariáveis, fração de cura e os modelos de regressão afetando o parâmetro de escala da distribuição Weibull e a fração de cura p, dados respectivamente por $\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i)$ (52) e $logito(p_i) = \gamma_0 + \gamma_1 idade_i + \gamma_2 hercep_i + \gamma_3 estágio_i + \gamma_4 cirur_i + \gamma_5 pCR_i + \gamma_6 estrog_i + \gamma_7 progest_i$ (53).

Assumindo distribuições a priori não-informativas normais $N(0,1)$ para todos os parâmetros de regressão β_r, γ_r ; $r = 0,1,2,\ldots,7$ e $\theta \sim Gama(1,1)$ e usando o software OpenBugs (burn-in sample =1.000 e 6.000 amostras finais tomadas de 100 em 100), temos na Tabela 13 os sumários a posteriori de interesse.

Tabela 13: Sumários a posteriori de interesse - Tempos de sobrevida livre da doença - Modelos de regressão afetando parâmetro de escala da distribuição Weibull e a fração de cura.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
Modelo de regressão afetando o parâmetro de escala				
γ_0 (intercepto)	0,05	0,9623	-1,8350	1,8970
Idade	0,22	0,9528	-1,7280	2,0480
Herceptin	-0,12	0,9283	-1,9680	1,6770
Estágio	-0,61	0,6212	-1,9360	0,4957
Cirurgia	-0,06	0,9323	-1,9330	1,7260
Resposta patológica completa	0	0,9300	-1,9020	1,7820
Receptor de Estrogênio	-0,15	0,9509	-1,9850	1,7190
Receptor de Progesterona	0,04	0,9526	-1,8600	1,8500
Escala	1,54	0,4563	0,8375	2,6090
Modelo de regressão afetando o parâmetro de fração de cura				
β_0 (intercepto)	1,11	0,9333	-0,7087	2,9540
Idade	0,32	0,4848	-0,5951	1,3140
Herceptin	0,52	0,5931	-0,5760	1,7580
Estágio	0,77	0,3861	0,0097	1,5350
Cirurgia	-0,24	0,4839	-1,1760	0,7276
Resposta patológica completa	0,52	0,4599	-0,3784	1,4320
Receptor de Estrogênio	0,39	0,5120	-0,6171	1,3970
Receptor de Progesterona	0,32	0,5483	-0,7373	1,4100

Dos resultados da Tabela 13, observa-se que a covariável estágio tem efeito significativo neste modelo (intervalo de credibilidade para o parâmetro de regressão β_3 correspondente não inclui o valor zero).

5.2. Análise estatística dos tempos de sobrevida total (ST)

Da mesma forma como foi considerado para os tempos de sobrevida livre da doença, vamos assumir a distribuição de Weibull para os tempos de sobrevida total.

5.2.1. Distribuição de Weibull sem a presença de covariáveis sob o enfoque Frequentista

Usando o software Minitab® encontramos os estimadores de máxima verossimilhança (EMV) dados na Tabela 14. Observa-se que o tempo médio estimado de sobrevida total é de 96,61 meses. O tempo mediano estimado é de 94,34 meses.

Tabela 14: EMV para os parâmetros da distribuição de Weibull - Tempos de sobrevida total.

Parâmetro	Estimativa	Erro Padrão	Intervalo de Confiança (95%)	
			Limite Inferior	Limite Superior
Forma	2,57	0,78450	1,4119	4,6741
Escala	108,81	26,8791	67,0495	176,579
Média	96,61	23,2029	60,3397	154,69

Nos tempos de sobrevida total, 45 pacientes não apresentaram o evento de interesse, aproximadamente 86% da amostra. Este é um indicativo de uma possível necessidade de complementar o modelo com a fração de cura e reanalisar estes dados sob o enfoque bayesiano poderá também trazer melhores inferências.

5.2.2. Distribuição de Weibull sem a presença de covariáveis sob o enfoque Bayesiano

Para uma análise bayesiana consideramos distribuições a priori gama $G(0{,}1;\ 0{,}1)$ não-informativas para α e θ. Na simulação de amostras da distribuição a posteriori para α e θ, (burn-in sample=1.000 e 6.000 amostras finais tomadas de 100 em 100).

Tabela 15: Sumários a posteriori de interesse - Tempos de sobrevida total.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
Forma	2,12	0,607	1,117	3,48
Escala	153,20	86,34	84,56	344,1
Média	139,10	92,61	75,5	327,3

Na Tabela 15, temos os sumários a posteriori de interesse. Utilizando distribuições a priori não informativas o tempo médio estimado de sobrevida total é de 139,10 meses e o tempo mediano é de 132,95 meses, um pouco maior do que o tempo mediano estimado pelo modelo sob o enfoque frequentista.

5.2.3. Modelo Weibull com fração de cura sem a presença de covariáveis sob o enfoque Bayesiano

Para uma análise bayesiana consideramos as seguintes distribuições a priori: $\alpha \sim Gama(1{,}1)$, $p \sim Beta(86{,}14)$ e $\theta \sim U(0{,}300)$. Observar que os hiperparâmetros da distribuição beta dados por $a = 86$ e $b = 14$ foram escolhidos como uma priori informativa para p com média igual à 0,84 (um valor próximo da proporção observada de dados censurados, interpretados como pacientes imunes ou curados).

Com um burn-in sample =1.000 e 6.000 amostras finais tomadas de 100 em 100 para obter as quantidades a posteriori de interesse. Na Tabela 16, temos os sumários a posteriori de interesse. Dos resultados da Tabela 16, observa-se que a proporção estimada de "curados" é de 83% resultado próximo do valor observado nos dados.

O tempo mediano estimado por esse modelo com a presença de fração de cura foi de 80,92 meses, sendo o menor tempo mediano estimado dentre os três modelos apresentados para o tempo de sobrevida total.

Tabela 16: Sumários a posteriori de interesse modelo com fração de cura sem covariáveis - Tempos de sobrevida total.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
Forma	2,52	0,850	1,095	4,359
Escala	60,34	19,650	39,850	109,300
p (fração de cura)	0,83	0,035	0,760	0,896

Observar que as funções de sobrevivência considerando modelos baseados na distribuição Weibull sob uma abordagem frequentista, bayesiana sem e com fração de cura (ver em Tabela 14, Tabela 15, Tabela 16) são dadas, respectivamente por:

- Frequentista sem fração de cura: $S(t) = exp[-(\frac{t}{108,81})^{2,57}]$
- Bayesiano sem fração de cura: $S(t) = exp[-(\frac{t}{153,20})^{2,12}]$
- Bayesiano com fração cura: $S(t) = 0,83 + (1 - 0,83)exp[-(\frac{t}{60,34})^{2,52}]$

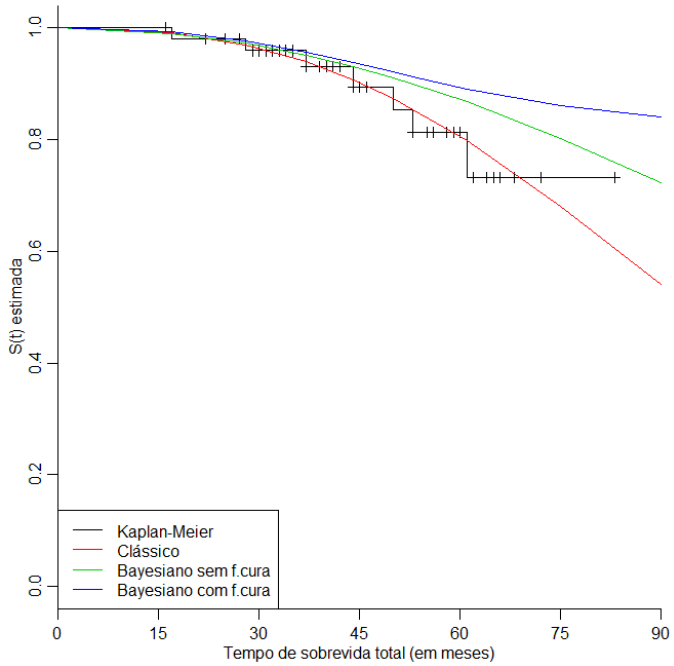

Figura 7 - Gráficos da função de sobrevivência estimada - Kaplan e Meier, Weibull Bayesiano sem e com fração de curas (Tempos de sobrevida total).

Na Figura 7, temos os gráficos das funções de sobrevivência estimadas considerando os estimadores de Kaplan-Meier e os modelos Weibull sob uma abordagem frequentista, bayesiano sem e com presença de fração de cura. Observa-se pelo gráfico um bom ajuste do modelo na presença de fração de cura, o que não ocorre para os modelos sem a presença de fração de cura, pois as curvas estimadas dos modelos sem fração de cura possuem um decaimento muito acentuado fazendo com que para tempos de sobrevida grandes as curvas sem fração de cura se distanciem cada vez mais da realidade dos dados aqui representado pela curva de Kaplan-Meier.

A partir daqui, as covariáveis observadas serão inclusas nos modelos com o objetivo de identificar possíveis fatores que afetem o tempo de sobrevida total das pacientes, independente se houve ou não a recidiva do câncer.

5.2.4. Modelo de Weibull na presença de covariáveis sob o enfoque Frequentista

Assumir agora um modelo de regressão de Weibull (46), isto é, $log(t_i) = \beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i + \sigma^* \varepsilon_i$ para os tempos de sobrevida total com as mesmas covariáveis consideradas no modelo de regressão para os tempos de sobrevida livre da doença.

Dos resultados da Tabela 17, observa-se que todas as covariáveis não mostram efeitos significativos, pois todos os intervalos de confiança dos parâmetros contém o valor 0 (valor-p maior do que 0,05 para os testes de hipóteses sobre os parâmetros de regressão serem iguais a zero para todas as covariáveis). Sendo assim, não existem evidências de que essas covariáveis afetem o tempo de sobrevida total das pacientes.

Tabela 17: EMV para os parâmetros de regressão de Weibull - Tempos de sobrevida total.

Parâmetro	Estimativa	Erro Padrão	Z	P	Intervalo de Confiança (95%)	
					Limite Inferior	Limite Superior
β_0 (intercepto)	25,96	3511,26	0,01	0,9940	-6855,98	6907,89
Idade	-0,16	0,32	-0,51	0,6120	-0,79	0,47
Herceptin	-0,30	0,43	-0,71	0,4780	-1,14	0,53
Estágio	-5,03	1136,11	0,00	0,9960	-2231,78	2221,71
Cirurgia	-6,10	843,87	-0,01	0,9940	-1660,05	1647,85
Resposta patológica completa	0,27	0,36	0,74	0,4580	-0,44	0,98
Receptor de Estrogênio	-0,24	0,34	-0,69	0,4890	-0,91	0,43
Receptor de Progesterona	0,75	0,47	1,57	0,1150	-0,18	1,68
Forma	2,82	0,90			1,51	5,27

(Z:estatística Z; P: valor-p)

5.2.5. Modelo de Weibull na presença de covariáveis sob o enfoque Bayesiano

Considerar uma análise bayesiana dos tempos de sobrevida total assumindo um modelo de regressão de Weibull (50), $\theta_i = exp(\beta_0 + \beta_1 idade_i + \beta_2 hercep_i + \beta_3 estágio_i + \beta_4 cirur_i + \beta_5 pCR_i + \beta_6 estrog_i + \beta_7 progest_i$ na presença das mesmas covariáveis consideradas para os tempos livres da doença.

Assumindo distribuições a priori não-informativas normais $N(0,1)$ para todos os parâmetros de regressão βr , $r = 0,1,2,\ldots,7$; uma priori $Gama(1,1)$ para o parâmetro de

forma α. Com um burn-in sample = 1.000 e 6.000 amostras finais tomadas de 100 em 100, temos na Tabela 18 os sumários a posteriori de interesse.

Tabela 18: Sumários a posteriori de interesse - Tempos de sobrevida total - Modelo de regressão.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
β_0 (intercepto)	-5,88	2,3360	-10,3100	-1,2930
Idade	-0,07	0,6505	-1,3240	1,2430
Herceptin	-0,43	0,7706	-2,0050	1,0360
Estágio	-0,70	0,7007	-1,9590	0,7388
Cirurgia	1,00	0,6971	-0,3052	2,4290
Resposta patológica completa	-0,57	0,6555	-1,8560	0,6831
Receptor de Estrogênio	0,02	0,6662	-1,2950	1,3150
Receptor de Progesterona	-0,72	0,7373	-2,2180	0,6894
Forma	1,53	0,4528	0,7313	2,5030

Dos resultados da Tabela 18, observa-se que todas as covariáveis não mostram efeitos significativos, pois os intervalos de credibilidade 95% para todos os parâmetros de regressão incluem o valor zero. O modelo de regressão Weibull sob o enfoque bayesiano não detectou nenhuma covariável que afete o tempo de sobrevida total das pacientes.

5.2.6. Modelo Weibull com fração de cura e com covariáveis afetando o parâmetro de escala da distribuição Weibull

Considerar agora uma análise bayesiana dos tempos de sobrevida total assumindo uma distribuição de Weibull na presença de covariáveis, fração de curas e o modelo de regressão afetando o parâmetro de escala da distribuição Weibull.

Assumindo distribuições a priori não-informativas normais $\mathrm{N}(0,1)$ para todos os parâmetros de regressão βr , $r = 0,1,2,\ldots,7$; $\gamma \sim Gama(1,1)$, $p \sim Beta(86,14)$ e tomando 6.000 amostras finais, de 100 em 100, burn-in sample =1.000, temos na Tabela 19 os sumários a posteriori de interesse.

Dos resultados da Tabela 19, observa-se que todas as covariáveis não mostram efeitos significativos, pois os intervalos de credibilidade 95% para todos os parâmetros de regressão incluem o valor zero.

Tabela 19: Sumários a posteriori de interesse - Tempos de sobrevida total - Modelo de regressão na presença de fração de cura afetando o parâmetro de escala.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
β_0 (intercepto)	0,48	0,9666	-1,3980	2,3230
Idade	0,05	0,5617	-1,0700	1,1780
Herceptin	0,51	0,6053	-0,6691	1,7740
Estágio	0,97	0,4926	-0,0120	1,9210
Cirurgia	-0,33	0,9468	-2,0480	1,7200
Resposta patológica completa	0,10	0,5777	-0,9722	1,3220
Receptor de Estrogênio	0,62	0,5320	-0,4528	1,7130
Receptor de Progesterona	0,19	0,6149	-0,9521	1,5190
Escala	2,25	0,9291	0,8718	4,5130
p (fração de cura)	0,84	0,0330	0,7677	0,8965

5.2.7. Modelo Weibull com fração de cura e com covariáveis afetando o parâmetro de escala da distribuição Weibull e a probabilidade de cura

Considerar agora uma análise bayesiana dos tempos de sobrevida total assumindo uma distribuição de Weibull na presença de covariáveis, fração de curas e os modelos de regressão afetando parâmetro de escala da distribuição Weibull e a fração de cura.

Assumindo distribuições a priori não-informativas normais $\mathrm{N}(0,1)$ para todos os parâmetros de regressão γ_r, β_r; $r = 0,1,2,\ldots,7$ e $\theta \sim Gama(1,1)$ e usando o software OpenBugs (burn-in sample =1.000 e 6.000 amostras finais tomadas de 100 em 100).

Tabela 20: Sumários a posteriori de interesse - Tempos de sobrevida total - Modelos de regressão afetando parâmetro de escala da distribuição Weibull e a fração de cura.

Parâmetro	Média	Desvio Padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Superior
Modelo de regressão afetando o parâmetro de escala				
γ_0 (intercepto)	0,23	0,9730	-1,698	2,114
Idade	-0,03	0,9333	-1,891	1,776
Herceptin	0,03	0,8987	-1,689	1,734
Estágio	-0,17	0,7217	-1,782	1,082
Cirurgia	-0,51	0,9942	-2,411	1,479
Resposta patológica completa	0,32	0,9498	-1,603	2,076
Receptor de Estrogênio	-0,36	0,9344	-2,146	1,523
Receptor de Progesterona	0,19	0,9811	-1,785	2,076
Modelo de regressão afetando o parâmetro de escala				
β_0 (intercepto)	0,87	0,9749	-1,044	2,706
Idade	0,11	0,6219	-1,072	1,365
Herceptin	0,54	0,6835	-0,7076	1,984
Estágio	1,13	0,4875	0,1646	2,077
Cirurgia	-0,65	0,8564	-2,1670	1,276
Resposta patológica completa	0,37	0,6786	-0,9687	1,737
Receptor de Estrogênio	0,33	0,6548	-0,9529	1,649
Receptor de Progesterona	0,51	0,6867	-0,8443	1,908
Escala	1,80	0,7722	0,7507	3,738

Temos na Tabela 20, os sumários a posteriori de interesse. Dos resultados da Tabela 20, observa-se que a covariável estágio tem efeito significativo (intervalo de credibilidade para o parâmetro de regressão β_3 não inclui o valor zero).

5.3. Discussão dos resultados obtidos

O uso de modelos de fração de cura pode ser de grande interesse na análise de dados de sobrevida para pacientes com câncer de mama, dado que novas terapias levam a tempos de sobrevida livre da doença maiores ou mesmo a cura de muitas pacientes, significando que em uma grande parcela da amostra não ocorre o evento de interesse e consequentemente uma baixa proporção de dados completos. Dessa forma, modelos tradicionais sem a presença de fração de cura podem não ser apropriados.

O uso de métodos bayesianos tem crescido de forma substancial na análise de dados médicos de sobrevivência na presença de censuras e covariáveis, especialmente usando métodos MCMC de simulação de amostras da distribuição a posteriori de interesse para determinar os sumários a posteriori de interesse. Isso tem se tornado rotina na análise de dados médicos.

Foram utilizados modelos baseados na distribuição Weibull, devido a flexibilidade destes modelos e também por ser a distribuição que mais se adequou aos dados. Nos três modelos sem covariável para os tempos de sobrevida livre da doença obtivemos valores dos parâmetros de forma e de escala semelhantes nos modelos sem fração de cura, sendo que estes foram bem diferentes dos valores do modelo com fração de cura. Nos tempos de sobrevida total, os três modelos sem covariável apresentaram os parâmetros de forma semelhantes e os parâmetros de escala bem distintos.

Em ambos os tempos de sobrevida analisados o modelo com fração de cura demonstra ser o modelo mais adequado ao comportamento dos dados devido a presença de uma grande proporção de dados censurados. Tempo mediano é uma estimativa do tempo em que 50% das pacientes permanecem vivas, os menores tempos medianos em ambos os tempos de sobrevida foram os tempos estimados pelo modelo com fração de cura, devido a presença de grande proporção de censuras.

Os modelos Weibull sob o enfoque frequentista com e sem covariável obtiveram desempenho inferior quanto ao ajuste e predição em relação aos modelos bayesianos

apresentados nesta aplicação. É importante salientar que os resultados obtidos sob o enfoque frequentista são menos precisos em relação aos resultados sob o enfoque bayesiano por se utilizarem de métodos assintóticos para os estimadores de máxima verossimilhança, por apresentarem uma grande proporção de dados censurados e por serem dependentes do tamanho amostral. Na aplicação apresentada, a presença de uma grande proporção de dados censurados pode levar a inferências assintóticas não muito precisas.

Somente os modelos bayesianos conseguem incorporar a informação do especialista, no caso, do médico. Dessa forma têm-se inferências mais precisas sob o enfoque bayesiano.

Os modelos bayesianos com fração de cura demonstraram serem bastante sensíveis para detectar covariáveis significativas e são muito úteis, pois as probabilidades de cura podem ser estimadas para cada paciente (valores fixados das covariáveis), possibilitando, pelo médico, uma classificação de pacientes com maiores ou menores chances de cura.

Em ambos os tempos de sobrevida (SLD e ST) a covariável estágio se mostrou significativa, na Figura 2 e na Figura 4 com os gráficos das curvas de Kaplan-Meier estimadas para as covariáveis, observa-se que na covariável estágio a curva referente às pacientes do estágio 3, em ambos os tempos de sobrevida, tem um decaimento mais acentuado do que a curva referente as pacientes do estágio 2.

Nos tempos de sobrevida livre da doença, esse covariável foi significativa nos modelos Weibull (bayesiano) sem a presença de fração de cura, no modelo Weibull com fração de cura afetando o parâmetro de escala e no modelo Weibull bayesiano com fração de cura afetando o parâmetro de escala e a probabilidade de cura. No modelo Weibull (bayesiano) sem a presença de fração de cura, o tempo mediano de sobrevida livre da doença de pacientes do estágio 2 é 2,77 vezes maior do que o tempo mediano de pacientes do estágio 3, os outros modelos estimaram esta razão de tempos medianos como 2,20 vezes e 2,15 vezes respectivamente.

Dos resultados da análise do tempo de sobrevida total, somente o modelo Weibull bayesiano com fração de cura afetando o parâmetro de escala e a probabilidade de cura traz evidências de que a covariável estágio afeta o tempo de sobrevida total das pacientes, estimando que o tempo mediano de sobrevida de pacientes do estágio 2 é 3,09 vezes maior do que pacientes do estágio 3. Todas as outras covariáveis do estudo não apresentaram evidências de influência nos tempos de sobrevida das pacientes.

É interessante observar que outras distribuições paramétricas para dados de sobrevivência poderiam ser usadas na análise desses dados, mas a distribuição Weibull apresentou um bom ajuste para os dados. Nessa direção podemos mencionar várias distribuições exponenciais generalizadas (ver, por exemplo, Mudholkar e Srivastava, 1993; Gupta e Kundu, 1999, 2007; Raqab e Ahsanullah, 2001; Raqab, 2002; Sarhan, 2007; Carrasco et al, 2008; Achcar e Boleta, 2009).

A proporção de dados completos (ocorreu o evento de interesse) nos tempos de sobrevida livre da doença é de 29% o que corresponde a 15 pacientes e essa proporção diminui para 14% (7 pacientes) nos tempos de sobrevida total. Das 15 pacientes em que ocorreu a recidiva da doença apenas 7 morreram; e das pacientes que não apresentaram recidiva nenhuma morreu, insinuando que os tempos de sobrevida podem ter uma estrutura de dependência entre si. Portanto, para um estudo nesse sentido, prosseguimos com a reanalise destes mesmos dados considerando modelos de sobrevivência baseados em distribuições bivariadas que incorporem a estrutura de dependência que possa existir entre os tempos de sobrevida observados.

6. Modelos para análise bivariada dos dados de câncer de mama

6.1. Tempos de sobrevida dependentes assumindo uma distribuição exponencial bivariada de Block e Basu

Em muitas aplicações de análise de sobrevivência, usualmente temos dois tempos de vida T_1 e T_2 associados para cada unidade. Nestas aplicações, os modelos mais populares e mais aplicados em tempos de vida são dados pelas distribuições exponenciais bivariadas. Dentre essas distribuições exponenciais bivariadas, alguns modelos foram extensivamente usados por engenheiros de confiabilidade e pesquisadores médicos: o modelo exponencial bivariado Block & Basu (1974); o modelo exponencial bivariado Gumbel (1960); o modelo exponencial bivariado Freund (1961) e o modelo exponencial bivariado Marshall & Olkin (1967 a,b). Outras distribuições paramétricas exponenciais bivariadas são introduzidas na literatura (ver, por exemplo, Hougaard, 1986; Downton, 1970; Arnold & Strauss, 1988).

A distribuição exponencial bivariada proposta por Block e Basu (1974) é uma generalização da distribuição exponencial para dados bivariados, ou seja, a estrutura de dependência entre os tempos de sobrevida é incorporada ao modelo. Sua função densidade com parâmetros $\lambda_1 > 0$, $\lambda_2 > 0$ *e* $\lambda_3 > 0$ para tempos de sobrevida $T_1 > 0$ *e* $T_2 > 0$ é dada por,

$$f(t_1, t_2) = \begin{cases} f_1(t_1, t_2) = \frac{\lambda\lambda_1\lambda_{23}}{\lambda_{12}} exp\{-\lambda_1 t_1 - \lambda_{23} t_2\} & , se\ t_1 < t_2 \\ f_2(t_1, t_2) = \frac{\lambda\lambda_2\lambda_{13}}{\lambda_{12}} exp\{-\lambda_{13} t_1 - \lambda_2 t_2\} & , se\ t_1 \geq t_2 \end{cases} \qquad (54)$$

sendo que $\lambda_{12} = \lambda_1 + \lambda_2$, , $\lambda_{13} = \lambda_1 + \lambda_3$, $\lambda_{23} = \lambda_2 + \lambda_3$ e $\lambda = \lambda_1 + \lambda_2 + \lambda_3$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ e $\lambda_3 \geq 0$.

A função de sobrevivência conjunta para a distribuição Block e Basu é dada por,

$$S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2) = \begin{cases} S_1(t_1, t_2) & , se\ t_1 < t_2 \\ S_2(t_1, t_2) & , se\ t_1 \geq t_2 \end{cases} \qquad (55)$$

em que,

$S_1(t_1, t_2) = \frac{\lambda}{\lambda_{12}} exp(-\lambda_1 t_1 - \lambda_{23} t_2) - \frac{\lambda_3}{\lambda_{12}} exp(-\lambda t_2)$ e

$S_2(t_1, t_2) = \frac{\lambda}{\lambda_{12}} \exp(-\lambda_{13} t_1 - \lambda_2 t_2) - \frac{\lambda_3}{\lambda_{12}} \exp(-\lambda t_1)$

As médias e as variâncias para T_1 *e* T_2 são dadas por,

$$\mu_1 = E(T_1) = \frac{1}{\lambda_{13}} + \frac{\lambda_2\lambda_3}{\lambda\lambda_2\lambda_{13}}$$

$$\mu_2 = E(T_2) = \frac{1}{\lambda_{23}} + \frac{\lambda_1\lambda_3}{\lambda\lambda_{12}\lambda_{23}}$$

$$\sigma_1^2 = Var(T_1) = \frac{1}{\lambda_{13}^2} + \frac{\lambda_2\lambda_3(2\lambda_1\lambda + \lambda_2\lambda_3)}{\lambda^2\lambda_{12}^2\lambda_{13}^2}$$

$$\sigma_2^2 = Var(T_2) = \frac{1}{\lambda_{23}^2} + \frac{\lambda_1\lambda_3(2\lambda_2\lambda + \lambda_1\lambda_3)}{\lambda^2\lambda_{12}^2\lambda_{23}^2}$$

O coeficiente de correlação para T_1 *e* T_2 é dado por,

$$\rho_{12} = \frac{\lambda_3[(\lambda_1^2 + \lambda_2^2)\lambda + \lambda_1\lambda_2\lambda_3]}{\phi_1\phi_2}$$

sendo que:

$$\phi_1 = [\lambda_{12}^2\lambda_{13}^2 + \lambda_2(\lambda_2 + 2\lambda_1)\lambda^2]^{1/2}$$

$$\phi_2 = [\lambda_{12}^2\lambda_{23}^2 + \lambda_1(\lambda_1 + 2\lambda_2)\lambda^2]^{1/2}$$

A covariância entre T_1 e T_2 é dada por:

$$Cov(T_1, T_2) = \frac{(\lambda_1^2 + \lambda_2^2)\lambda_3\lambda + \lambda_1\lambda_2\lambda_3^2}{\lambda^2\lambda_{12}\lambda_{13}\lambda_{23}}$$

Suponha que ambos, T_1 *ou* T_2 podem ser censurados e que a censura é independente dos tempos de sobrevida. Neste caso, podemos sub-dividir as n observações em quatro classes:

C_1: ambos t_{1i} e t_{2i} são tempos de sobrevivência observados;

C_2: t_{1i} é um tempo de sobrevivência e t_{2i} é um tempo de censura (ou seja, sabemos apenas que $T_{2i} \geq t_{2i}$);

C_3: t_{1i} é um tempo de censura e t_{2i} é um tempo de sobrevivência;

C_4: ambos t_{1i} e t_{2i} são tempos de censura,

onde i $=$ 1, … , n.

A função de verossimilhança para um modelo contínuo (ver, por exemplo, Lawless, 1982, página 479) é dada por,

$$L = \prod_{i\in C_1} f(t_{1i}, t_{2i}) \prod_{i\in C_2} \left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}}\right) \prod_{i\in C_3} \left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}}\right) \prod_{i\in C_4} S(t_{1i}, t_{2i}) \qquad (56)$$

sendo que, $f(t_{1i}, t_{2i})$ como definida em (54) e $S(t_{1i}, t_{2i})$ definida em (55) e

- $-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}} = \begin{cases} S'_{1t_1}(t_{1i}, t_{2i}) & , se\ t_{1i} < t_{2i} \\ S'_{2t_1}(t_{1i}, t_{2i}) & , se\ t_{1i} \geq t_{2i} \end{cases}$

$S'_{1t_i}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_1}{\lambda_{12}} exp\{-\lambda_1 t_1 - \lambda_{23} t_{2i}\},$

$S'_{2t_i}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_{13}}{\lambda_{12}} exp\{-\lambda_{13} t_{1i} - \lambda_2 t_{2i}\} - \frac{\lambda\lambda_3}{\lambda_{12}} exp\{-\lambda t_{1i}\}$

e

- $-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}} = \begin{cases} S'_{1t_2}(t_{1i}, t_{2i}) & , se\ t_{1i} < t_{2i} \\ S'_{2t_2}(t_{1i}, t_{2i}) & , se\ t_{1i} \geq t_{2i} \end{cases}$

$S'_{1t_2}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_{23}}{\lambda_{12}} exp\{-\lambda_1 t_{1i} - \lambda_{23} t_{2i}\} - \frac{\lambda\lambda_3}{\lambda_{12}} exp\{-\lambda t_{2i}\},$

$S'_{2t_2}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_2}{\lambda_{12}} exp\{-\lambda_{13} t_{1i} - \lambda_2 t_{2i}\}$

Para uma análise bayesiana da distribuição Block e Basu na presença de observações censuradas, assumimos distribuições a priori Gama independentes para os parâmetros λ_k, isto é,

$$\lambda_k \sim Gamma\ (a_k, b_k)$$

para $k = 1,2\ e\ 3$; $a_k\ e\ b_k$ são hiperparâmetros conhecidos; $Gamma\ (a_k, b_k)$ denota uma distribuição gamma com média $a_k\ /\ b_k$ e variância $a_k\ /\ b_k^2$.

Na presença do vetor de covariáveis x, vamos considerar o seguinte modelo de regressão:

$$\lambda_{1i} = \alpha_1 exp\{\boldsymbol{\beta}'_1 \boldsymbol{x}_i\} \qquad (57)$$

$$\lambda_{1i} = \alpha_2 exp\{\boldsymbol{\beta}_2' \boldsymbol{x}_i\}$$

sendo que $\beta_j = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jp})'$; $j = 1,2$ é o vetor dos parâmetros de regressão e $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$, $i = 1,2,\dots,n$.

Neste caso, vamos assumir as seguintes distribuições a priori para os parâmetros $\alpha_1, \alpha_2, \beta_{1l}, \beta_{2l}$ e λ_3:

$$\alpha_k \sim Gamma\ (c_k, d_k)$$
$$\lambda_3 \sim Gamma(e, f)$$
$$\beta_{kl} \sim N(0, \sigma_{kl}^2)$$

para $k = 1,2$; $l = 1,2,..\ p$; c_k; d_k; e; f; σ_{kl}^2 são hiperparâmetros conhecidos e $N(0, \sigma_{kl}^2)$ denota uma distribuição normal com média igual a zero e variância σ_{kl}^2. Além disso, assumimos independência a priori entre todos os parâmetros.

Sob o enfoque bayesiano, usamos métodos de Monte Carlo em cadeias de Markov (MCMC) (ver, por exemplo, Casella e George, 1992; Chib e Greenberg, 1995; Gelfand e Smith, 1990) e o software OpenBugs (Spiegelhalter, et al, 2003) para simular amostras da distribuição a posteriori conjunta de interesse. Usando o software OpenBugs não é preciso especificar todas as distribuições a posteriori condicionais necessárias para o amostrador de Gibbs; só precisamos especificar a função de verossimilhança e as distribuições a priori para os parâmetros do modelo. A partir das amostras simuladas de Gibbs, encontramos estimativas de Monte Carlo para os sumários a posteriori de interesse.

6.2. Tempos de sobrevida dependentes assumindo uma distribuição geométrica bivariada de Arnold

Uma alternativa para o uso de uma distribuição contínua para tempos de sobrevida bivariados é admitir os tempos T_1 e T_2 como variáveis aleatórias discretas, que podem tomar valores em qualquer número inteiro positivo, para isso, aproxima-se a parte decimal do tempo de sobrevida para o inteiro mais próximo.

Dessa forma, a literatura apresenta diferentes distribuições discretas bivariadas que poderiam ser utilizadas para analisar os dados da Tabela A.1. Uma distribuição discreta

multivariada foi proposta por Arnold (1975) motivada da distribuição exponencial multivariada de Marshall-Olkin (1967 a,b). Em 1988 Nair e Nair (Nair e Nair, 1988) estudaram as características de algumas distribuições exponenciais bivariadas geométricas. A distribuição geométrica bivariada proposta por Arnold (1975) tem função de probabilidade dada por:

$$P(T_1 = t_1, T_2 = t_2) = \begin{cases} P_1(t_1, t_2) = \theta_1\theta_2(1-\theta_1-\theta_2)^{t_1-1}(1-\theta_2)^{t_2-t_1-1}, t_1 < t_2 \\ 0, t_1 = t_2 \\ P_2(t_1, t_2) = \theta_1\theta_2(1-\theta_1-\theta_2)^{t_2-1}(1-\theta_1)^{t_1-t_2-1}, t_1 > t_2 \end{cases} \quad (58)$$

sendo que, as funções de probabilidade marginais para T_1 *e* T_2 são distribuições geométricas padrão que iniciam em 1, dadas, respectivamente por,

$$p(t_1) = (1-\theta_1)^{t_1-1}\theta_1 \ , t_1 = 1,2,3, \ldots$$

e

$$p(t_2) = (1-\theta_2)^{t_2-1}\theta_2 \ , t_2 = 1,2,3, \ldots$$

As médias, variâncias, covariância e correlação são dadas por,

$$\mu_1 = E(T_1) = \frac{1}{\theta_1}, \mu_2 = E(T_2) = \frac{1}{\theta_2}$$

$$\sigma_1^2 = Var(T_1) = \frac{1-\theta_1}{\theta_1^2}, \sigma_2^2 = Var(T_2) = \frac{1-\theta_2}{\theta_2^2}$$

$$Cov(T_1, T_2) = \frac{-1}{1-r}$$

$$\rho_{12} = Corr(T_1, T_2) = -\frac{\theta_1\theta_2}{(1-r)[(1-\theta_1)(1-\theta_2)]^{0.5}}$$

sendo que $r = 1-\theta_1-\theta_2$, $0 < \theta_1 < 1$ *e* $0 < \theta_2 < 1$.

Sejam $\{(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})\}$ amostras aleatórias independentes de tamanho n derivadas de uma distribuição geométrica bivariada com função de probabilidade dada em (11). Assumir Y_1 *e* Y_2 como o vetor de censuras de T_1 *e* T_2 e que as censuras são

independentes dos tempos de sobrevida. Vamos subdividir as n observações nas seguintes quatro classes:

C_1: $T_{1i} < Y_{1i}\ e\ T_{2i} < Y_{2i}$, então ambos, $t_{1i}\ e\ t_{2i}$ são os tempos de sobrevida;
C_2: $T_{1i} < Y_{1i}\ e\ Y_{2i} < T_{2i}$, então se observa t_{1i} e y_{2i} ;
C_3: $Y_{1i} < T_{1i}\ e\ T_{2i} < Y_{2i}$, então se observa y_{2i} e $\mathrm{t_{1i}}$;
C_4: $Y_{1i} < T_{1i}\ e\ Y_{2i} < T_{2i}$, então se observa y_{1i} e y_{2i}.

Dadas as definições acima, a função de verossimilhança para $\theta_1\ e\ \theta_2$ assumindo a distribuição geométrica bivariada com função de massa de probabilidade dada por (58) e com dados censurados à direita é dada por,

$$L(\theta_1,\theta_2) = \prod_{i\in C_1} P(t_{1i},t_{2i}) \prod_{i\in C_2}\left(\sum_{t_{2i}=y_{2i+1}}^{\infty} P(t_{1i},t_{2i})\right) \prod_{i\in C_3}\left(\sum_{t_{1i}=y_{1i+1}}^{\infty} P(t_{1i},t_{2i})\right) \prod_{i\in C_4}\left(\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty} P(t_{1i},t_{2i})\right) \quad (59)$$

sendo que,

- $\sum_{t_{2i}=y_{2i+1}}^{\infty} P(t_{1i},t_{2i}) = \begin{cases}\sum_{t_{2i}=y_{2i+1}}^{\infty} P_1(t_{1i},t_{2i}) \ , if\ t_{1i} < t_{2i} \\ \sum_{t_{2i}=y_{2i+1}}^{\infty} P_2(t_{1i},t_{2i}) \ , if\ t_{1i} \geq t_{2i}\end{cases}$

  $$\sum_{t_{2i}=y_{2i+1}}^{\infty} P_1(t_{1i},t_{2i}) = \theta_1(1-\theta_1-\theta_2)^{t_{1i}-1}(1-\theta_2)^{y_{2i}-t_{1i}-1}$$

  $$\sum_{t_{2i}=y_{2i+1}}^{\infty} P_2(t_{1i},t_{2i}) = \theta_1(1-\theta_1-\theta_2)^{y_{2i}-1}(1-\theta_1)^{t_{1i}-y_{2i}-1}$$

- $\sum_{t_{1i}=y_{1i+1}}^{\infty} P(t_{1i},t_{2i}) = \begin{cases}\sum_{t_{1i}=y_{1i+1}}^{\infty} P_1(t_{1i},t_{2i}) \ , if\ t_{1i} < t_{2i} \\ \sum_{t_{1i}=y_{1i+1}}^{\infty} P_2(t_{1i},t_{2i}) \ , if\ t_{1i} \geq t_{2i}\end{cases}$

  $$\sum_{t_{1i}=y_{1i+1}}^{\infty} P_1(t_{1i},t_{2i}) = \theta_2(1-\theta_2)^{t_{2i}-y_{1i}-1}(1-\theta_1-\theta_2)^{y_{1i}}$$

  $$\sum_{t_{1i}=y_{1i+1}}^{\infty} P_2(t_{1i},t_{2i}) = \theta_2(1-\theta_1)^{y_{1i}-t_{2i}}(1-\theta_1-\theta_2)^{t_{2i}-1}$$

- $\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty} P(t_{1i},t_{2i}) = \begin{cases}\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty} P_1(t_{1i},t_{2i}) \ , if\ t_{1i} < t_{2i} \\ \sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty} P_2(t_{1i},t_{2i}) \ , if\ t_{1i} \geq t_{2i}\end{cases}$

$$\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty} P_1(t_{1i},t_{2i}) = (1-\theta_2)^{y_{2i}-y_{1i}}(1-\theta_1-\theta_2)^{y_{1i}}$$

$$\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty} P_2(t_{1i},t_{2i}) = (1-\theta_1)^{y_{1i}-y_{2i}}(1-\theta_1-\theta_2)^{y_{2i}}$$

Para uma análise bayesiana, vamos assumir a seguinte distribuição a priori conjunta para $\theta_1 \, e \, \theta_2$:

$$\pi(\theta_1,\theta_2) \propto \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}(1-\theta_1-\theta_2)^{\alpha_0-1}, \theta_1+\theta_2<1 \qquad (60)$$

sendo que a função dada em (60) é a função de probabilidade de uma distribuição Dirichlet $Dir_2(\alpha_0,\alpha_1,\alpha_2)$ com hiperparâmetros $\alpha_0,\alpha_1 e \, \alpha_2$.

Combinando-se a distribuição a priori de Dirichlet (60) com a função de verossimilhança (59), obtemos a partir da fórmula de Bayes, a distribuição a posteriori conjunta para $\theta_1 \, e \, \theta_2$.

$$\pi(\theta_1,\theta_2|\boldsymbol{z}) \propto \theta_1^{m_1+\alpha_1-1}\theta_2^{m_2+\alpha_2-1}(1-\theta_1)^{z_1}(1-\theta_2)^{z_2}(1-\theta_1-\theta_2)^{z_{12}+\alpha_0-1}$$

Na presença de covariáveis $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$ associadas a cada tempo de sobrevida bivariado $T_{1i} \, e \, T_{2i}$, podemos assumir o modelo de regressão logística dado por,

$$\theta_{1i} = \frac{exp\{\boldsymbol{\beta_1'x}_i\}}{1-exp\{\boldsymbol{\beta_1'x}_i\}}$$

$$\theta_{2i} = \frac{exp\{\boldsymbol{\beta_2'x}_i\}}{1-exp\{\boldsymbol{\beta_2'x}_i\}} \qquad (61)$$

sendo que $\beta_j = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jp})'$; $j = 1,2$ é o vetor dos paramêtros de regressão $i = 1,2,\dots,n$.

6.3. Tempos de sobrevida dependentes assumindo uma distribuição geométrica bivariada de Basu-Dhar

A distribuição geométrica bivariada de Basu-Dhar (1995) tem função de sobrevivência dada por:

$$P(T_1 > t_1, T_2 > t_2) = p_1^{t_1}\, p_2^{t_2}\, p_{12}^{\max(t_1,t_2)} \qquad (62)$$

sendo que $0 < p_1 < 1$, $0 < p_2 < 1$ e $0 < p_{12} \leq 1$. Observa-se que a função de sobrevivência (62) satisfaz a propriedade de perda de memória sem quaisquer restrições adicionais nos parâmetros, a saber,

$$P(T_1 > s_1 + t, T_2 > s_2 + t \,/\, T_1 > s_1, T_2 > s_2) = P(T_1 > t, T_2 > t) = (p_1 p_2 p_{12})^t \quad (63)$$

A função de probabilidade da distribuição geométrica bivariada de Basu-Dhar é dada por,

$$P(T_1 = t_1, T_2 = t_2) = \begin{cases} (p_1)^{t_1-1}(1-p_1)(p_2 p_{12})^{t_2-1}(1-p_2 p_{12}) & para\ T_1 < T_2 \\ (p_1 p_2 p_{12})^{t_1-1}(1-p_1 p_{12}-p_2 p_{12}+p_1 p_2 p_{12}) & para\ T_1 = T_2 \\ (p_2)^{t_2-1}(1-p_2)(p_1 p_{12})^{t_1-1}(1-p_1 p_{12}) & para\ T_1 > T_2 \end{cases} \quad (64)$$

As distribuições marginais de T_1 e T_2 são dadas respectivamente por,

$$P(T_1 = t_1) = P(T_1 > t_1 - 1) - P(T_1 > t_1) = (1-p_1 p_{12})(p_1 p_{12})^{t_1-1}$$
$$P(T_2 = t_2) = P(T_2 > t_2 - 1) - P(T_2 > t_2) = (1-p_2 p_{12})(p_2 p_{12})^{t_2-1}$$

sendo que $t_1, t_2 = 1,2,3,\ldots$ e as médias são dadas respectivamente por,

$$E(T_1) = \sum_{t1=1}^{\infty} t_1\, P(T_1 = t_1) = \left(1 - p_1 p_{12}\right)^{-1}$$
$$E(T_2) = \sum_{t2=1}^{\infty} t_2\, P(T_2 = t_2) = \left(1 - p_2 p_{12}\right)^{-1} \qquad (65)$$

A função de verossimilhança para p_1, p_2, p_{12} é dada por,

$$L(p_1, p_2, p_{12}) = \begin{cases} \prod_{i\in c1} P(T_{1i} = t_{1i}, T_{2i} = t_{2i}) \prod_{i\in c2} P(T_{1i} = t_{1i}, T_{2i} > t_{2i}) \\ \prod_{i\in c3} P(T_{1i} > t_{1i}, T_{2i} = t_{2i}) \prod_{i\in c4} P(T_{1i} > t_{1i}, T_{2i} > t_{2i}) \end{cases} \qquad (66)$$

onde,

- $P(T_1 = t_1, T_2 = t_2) = \begin{cases} (p_1)^{t_1-1}(1-p_1)(p_2p_{12})^{t_2-1}(1-p_2p_{12}) & for\ T_1 < T_2 \\ (p_1p_2p_{12})^{t_1-1}(1-p_1p_{12}-p_2p_{12}+p_1p_2p_{12}) & for\ T_1 = T_2 \\ (p_2)^{t_2-1}(1-p_2)(p_1p_{12})^{t_1-1}(1-p_1p_{12}) & for\ T_1 > T_2 \end{cases}$

Observar que,

- $P(T_1 > t_1, T_2 > t_2) = p_1^{t_1}p_2^{t_2}p_{12}^{\max(t_1,t_2)}$

- $P(T_1 = t_1, T_2 > t_2) = \begin{cases} (p_1)^{t_1-1}(1-p_1)(p_2p_{12})^{t_2} & para\ T_1 \leq T_2 \\ (p_2)^{t_2}(p_1p_{12})^{t_1-1}(1-p_1p_{12}) & para\ T_1 > T_2 \end{cases}$

- $P(T_1 > t_1, T_2 = t_2) = \begin{cases} (p_1)^{t_1}(p_2p_{12})^{t_2-1}(1-p_2p_{12}) & para\ T_1 < T_2 \\ (p_2)^{t_2-1}(p_1p_{12})^{t_1}(1-p_2) & para\ T_1 \geq T_2 \end{cases}$

7. Resultados da análise bivariada dos dados de câncer de mama

Nessa seção será apresentada a análise bayesiana bivariada dos tempos de sobrevida dados na Tabela A.1. Supondo os modelos que foram apresentados na seção 6.

7.1. Análise Bayesiana dos tempos de sobrevida da Tabela A.1 assumindo a distribuição exponencial bivariada Block e Basu.

Inicialmente assumindo a distribuição exponencial bivariada proposta por Block e Basu sob um enfoque bayesiano sem a presença de covariáveis. Assumimos uma distribuição a priori $\text{Gama}(1{,}100)$ para os parâmetros $\lambda r, r = 1,2,3$ para os dados de sobrevida bivariados T_1 (sobrevida livre doença) e T_2 (sobrevida total), com um "burn-in sample" de 10.000 amostras e 1.000 amostras finais tomadas de 100 em 100, temos na Tabela 21 os sumários a posteriori de interesse. A convergência do algoritmo Gibbs sampling foi verificada a partir de gráficos de séries temporais das amostras simuladas de Gibbs.

Tabela 21: Sumários a posteriori de interesse - Distribuição exponencial bivariada Block e Basu - sem a presença de covariáveis.

Parâmetro	Média	Desvio padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Inferior
λ_1	0,0025	0,0016	0,3830	0,0063
λ_2	0,1850	0,2510	0,0029	0,9230
λ_3	0,0075	0,0021	0,0038	0,0119
Média 1 (SLD)	108,4000	22,9000	71,8800	158,1000
Média 2 (ST)	232,6000	56,4400	147,0000	358,6000
ρ_{12}	0,0013	0,0006	0,2480	0,0026
Desvio Padrão 1 (SLD)	107,9000	22,5300	71,7200	156,5000
Desvio Padrão 2 (ST)	173,5000	44,4700	111,0000	282,7000

A partir dos resultados da Tabela 21, observa-se que as estimativas de Monte Carlo para as médias com base na função de perda de erro quadrático, isto é, as médias a posteriori para μ_1 e μ_2 (ver seção 6.1) do tempo de sobrevida livre de doença e do tempo de sobrevida total são dadas, respectivamente, por 108,4 meses e 232,6 meses.

Considerar agora uma análise sob o enfoque bayesiano dos tempos de sobrevida bivariados na presença de covariáveis, assumindo o seguinte modelo de regressão,

$$\lambda_{vi} = \alpha_v exp(\beta_{v1} idade_i + \beta_{v2} hercep_i + \beta_{v3} estágio_i + \beta_{v4} cirurg_i + \beta_{v5} pCR_i + \beta_{v6} estrog_i + \beta_{v7} progest_i) \quad (67)$$

sendo que $v = 1$ (sobrevida livre da doença) e $v = 2$ (sobrevida total).

Assumindo distribuições a priori não-informativas normais $N(0,1)$ para todos os parâmetros de regressão β_{1r} e β_{2r} , $r = 0,1,2,\ldots,7$; $\alpha_1 \sim Gama(1,1)$, $\alpha_2 \sim Gama(1,1)$ e $\alpha_3 \sim Gama(1,100)$ usando o software OpenBugs com um "burn-in sample" de 10.000 amostras e 1.000 amostras finais tomadas de 100 em 100, temos na Tabela 22, os sumários a posteriori de interesse

Dos resultados da Tabela 22, conclui-se que só a covariável estágio tem efeito significativo (intervalo de credibilidade para o parâmetro de regressão da idade não incluem o valor zero) para o tempo de sobrevida total.

Tabela 22: Sumários a posteriori de interesse – Assumindo a distribuição exponencial bivariada Block e Basu – na presença de covariáveis.

Parâmetro	**Média**	**Desvio padrão**	**Intervalo de Credibilidade (95%)**	
			Limite Inferior	**Limite Superior**
α_1	0,3488	0,4942	0,0026	1,7590
α_2	0,2162	0,3995	0,0002	1,3720
λ_3	0,0082	0,0020	0,0046	0,0126
Sobrevida livre da doença				
Idade	-0,6166	0,8221	-2,2070	1,0280
Herceptin	-0,8113	0,8532	-2,5280	0,8380
Estágio	-1,1060	0,6792	-2,4190	0,2833
Cirurgia	-0,1591	0,8108	-1,7840	1,4300
Resposta patológica completa	-0,4599	0,8124	-2,0720	1,1180
Receptor de Estrogênio	-0,3456	0,8354	-1,9780	1,2760
Receptor de Progesterona	-0,3651	0,8592	-2,0650	1,2980
Sobrevida total				
Idade	-0,5067	0,9379	-2,3420	1,3150
Herceptin	-0,9139	0,9402	-2,7780	0,9161
Estágio	-2,1560	0,7888	-3,6690	-0,5626
Cirurgia	-0,4386	0,9410	-2,2940	1,4040
Resposta patológica completa	-0,3345	0,9299	-2,1650	1,4840
Receptor de Estrogênio	-0,2966	0,9527	-2,1770	1,5520
Receptor de Progesterona	-0,2269	0,9544	-2,1060	1,6370

7.2. Análise Bayesiana dos tempos de sobrevida assumindo a distribuição geométrica bivariada proposta por Arnold.

Considerando a distribuição geométrica bivariada proposta por Arnold sob um enfoque bayesiano sem a presença de covariáveis, assumindo uma distribuição priori $\text{Dirichlet}(1,1,1)$ com função de probabilidade (60) para os parâmetros θ_1 e θ_2 onde $r = 1 - \theta_1 - \theta_2$ da distribuição geométrica bivariada de Arnold para os tempos T_1 (tempos de sobrevida livre de doença) e T_2 (tempo de sobrevida total), apresentados na Tabela A.1. No software OpenBugs,

foram geradas 10.000 amostras de aquecimento e outras 1.000 amostras finais tomadas de 100 em 100, temos na Tabela 23 os sumários a posteriori de interesse.

A partir dos resultados da Tabela 23, as médias a posteriori para os tempos T_1 (tempos de sobrevida livre de doença) e T_2 (tempo de sobrevivência global), são estimadas, respectivamente, por 140,4 e 343,6 meses, isto é, resultados semelhantes aos obtidos usando a distribuição de Block e Basu para o tempo T_1 (tempos de sobrevida livre de doença) (108,4 meses), mas muito diferente para o tempo de sobrevida global (232,6 meses).

Tabela 23: Sumários a posteriori de interesse – Distribuição geométrica bivariada Arnold - sem a presença de covariáveis.

Parâmetro	**Média**	**Desvio padrão**	**Intervalo de Credibilidade (95%)**	
			Limite Inferior	**Limite Inferior**
Média 1 (SLD)	140,4000	36,2600	87,7500	222,1000
Média 2 (ST)	343,6000	144,1000	160,1000	720,6000
r	0,9891	0,0022	0,9845	0,9931
θ_1	0,0076	0,0018	0,0045	0,0113
θ_2	0,0033	0,0012	0,0014	0,0062

Assumindo tempos de sobrevida discretos na presença de covariáveis, inicialmente consideramos distribuições geométricas independentes para os dois tempos de sobrevida. A distribuição geométrica tem função de probabilidade dada por,

$$P(T = t) = \theta(1-\theta)^t \, , t = 0,1,2,3, \ldots \tag{68}$$

sendo que a média é dada por, $\frac{(1-\theta)}{\theta}$.

A função de verossimilhança da i-ésima contribuição é dada por,

$$L_i = [P(T_i = t_i)]^{\delta_i} \, [P(T_i \geq t_i)]^{1-\delta_i} \tag{69}$$

onde $\delta_i = 1$ para uma observação completa e $\delta_i = 0$ para uma observação censurada e $\mathrm{P(T_i \geq t_i) = 1 - P(T_i < t_i)}$, isto é, $\mathrm{P(T_i < t_i) = \sum_{u=0}^{t_i-1} \theta(1-\theta)^u = \theta + \theta(1-\theta) + \theta(1-\theta)^2 + \theta(1-\theta)^3 + \cdots + \theta(1-\theta)^{t_i-1}}$.

Resultado:

$$\sum_{k=0}^{n} ar^k = \frac{a(1-r^{n-1})}{1-r}$$

com a $= \theta$, r $= 1-\theta$ e n $= t_i - 1$,

$$P(T_i < t_i) = \sum_{u=0}^{t_i-1} \theta(1-\theta)^u = \frac{\{\theta[1-(1-\theta)^{t_i}]\}}{\theta} = [1-(1-\theta)^{t_i}]$$

isto é,

$$P(T_i \geq t_i) = 1 - P(T_i < t_i) = 1 - [1-(1-\theta)^{t_i}] = (1-\theta)^{t_i} \quad (70)$$

Assim, a verossimilhança da i-ésima contribuição é dada por,

$$L_i = [\theta(1-\theta)^{t_i}]^{\delta_i} [P(T_i \geq t_i)]^{1-\delta_i} = [\theta(1-\theta)^{t_i}]^{\delta_i} [\theta(1-\theta)^{t_i}]^{1-\delta_i} \quad (71)$$

Na presença de covariáveis assumindo um modelo de regressão logístico dado por,

$$\text{logit}(\theta_{vi}) = \beta_{v0} + \beta_{v1}\text{age}_i + \beta_{v2}\text{hercep}_i + \beta_{v3}\text{stage}_i + \beta_{v4}\text{surgical}_i + \beta_{v5}\text{pCR}_i + \beta_{v6}\text{estrog}_i + \beta_{v7}\text{progest}_i \quad (72)$$

sendo que *v=1* (sobrevida livre da doença) e *v=2* (sobrevida total).

Assumindo distribuições a priori não-informativas normais $N(0,1)$ para todos os parâmetros de regressão βr , $r = 0,1,2,\ldots,7$ e usando o software OpenBugs com "burn-in" de 10.000 amostras e 1000 amostras finais tomadas de 50 em 50, temos na Tabela 24 os sumários a posteriori de interesse.

A partir dos resultados da Tabela 24, é possível observar que a covariável estágio tem um efeito significativo (intervalo de credibilidade de 95% para os parâmetros da regressão correspondentes não incluem o valor zero) para os tempos de sobrevida livre de doença e da sobrevida total.

Tabela 24: Sumários a posteriori de interesse – Distribuição geométrica bivariada Arnold – na presença de covariáveis.

Parâmetro	Média	Desvio padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Inferior
Sobrevida livre da doença				
β_{10}	-1,1500	0,8690	-2,8740	0,5450
Idade	-0,6106	0,4773	-1,5210	0,3398
Herceptin	-0,7497	0,6248	-2,0410	0,4093
Estágio	-0,7921	0,3799	-1,5490	-0,0418
Cirurgia	0,1646	0,4980	-0,7901	1,1480
Resposta patológica completa	-0,5464	0,4869	-1,5190	0,3899
Receptor de Estrogênio	-0,3122	0,5590	-1,4280	0,7668
Receptor de Progesterona	-0,4517	0,6064	-1,6620	0,7179
Sobrevida Total				
β_{20}	-1,1310	0,8939	-2,9280	0,6033
Idade	0,0109	0,0160	-0,0180	0,0446
Herceptin	-0,8977	0,7486	-2,4380	0,5125
Estágio	-1,3640	0,4298	-2,2120	-0,5195
Cirurgia	0,8478	0,6536	-0,4066	2,1540
Resposta patológica completa	-0,7906	0,6164	-2,0110	0,4031
Receptor de Estrogênio	-0,0024	0,6621	-1,3180	1,2740
Receptor de Progesterona	-0,7040	0,7083	-2,1410	0,6498

Para uma segunda análise com a distribuição geométrica bivariada proposta por Arnold sob um enfoque bayesiano com a presença de covariáveis e o modelo de regressão dado em (72), vamos assumir distribuições a priori informativas, uso de métodos bayesianos empíricos (ver, por exemplo, Carlin and Louis, 2002) para os parâmetros, baseando-se nos resultados da Tabela 24: $\beta_{10} \sim N(-1.15,1)$, $\beta_{20} \sim N(-1.13,1)$, $\beta_{11} \sim N(-0.61,1)$, $\beta_{12} \sim N(-0.74,1)$, $\beta_{13} \sim N(-0.79,1)$, $\beta_{14} \sim N(0.16,1)$, $\beta_{15} \sim N -0.54,1)$, $\beta_{16} \sim N(-0.31,1)$, $\beta_{17} \sim N(-0.45,1)$, $\beta_{21} \sim N(0.02,1)$, $\beta_{22} \sim N(-0.89,1)$, $\beta_{23} \sim N(-1.36,1)$, $\beta_{24} \sim N(0.84,1)$, $\beta_{25} \sim N(-0.80,1)$, $\beta_{26} \sim N(-0.002,1)$ e $\beta_{27} \sim N(-0.70,1)$. Na simulação de amostras da distribuição a posteriori de interesse, consideramos uma amostra de aquecimento de tamanho 1.000 e mais 1.000 amostras tomadas de 100 em 100.

É importante salientar que, neste caso, a convergência do algoritmo de Gibbs utilizando o OpenBugs só foi obtida usando as distribuições a priori informativas.

A partir dos resultados da Tabela 25, observa-se que a covariável estágio tem efeito significativo sobre o parâmetro θ_1 relacionado com a distribuição marginal para os tempos de sobrevida livre de doença (intervalos de credibilidade de 95% para todos os parâmetros da regressão incluem o valor zero); da mesma forma as covariáveis estágio e tipo de cirurgia tem efeitos significativos sobre o parâmetro θ_2 relacionado com a distribuição marginal para os tempos de sobrevida global (os intervalos de credibilidade 95% para os parâmetros da regressão não inclui o valor zero).

Tabela 25: Sumários a posteriori de interesse – Distribuição geométrica bivariada Arnold – na presença de covariáveis – utilizando distribuições a priori informativas.

Parâmetro	Média	Desvio padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Inferior
Sobrevida livre da doença				
β_{10}	-1,7140	0,8822	-3,3870	0,0032
Idade	-0,0071	0,0101	-0,0268	0,0137
Herceptin	-0,8801	0,5837	-2,0080	0,3045
Estágio	-0,7286	0,3690	-1,4780	-0,0542
Cirurgia	0,2633	0,4948	-0,7073	1,2620
Resposta patológica completa	-0,6618	0,5073	-1,6390	0,2715
Receptor de Estrogênio	-0,3258	0,5475	-1,4120	0,7353
Receptor de Progesterona	-0,5954	0,6393	-1,8410	0,5772
Sobrevida total				
β_{20}	-1,6740	0,9003	-3,5940	-0,0651
Idade	0,0126	0,0156	-0,0182	0,0437
Herceptin	-0,9547	0,7826	-2,5520	0,4777
Estágio	-1,3460	0,4422	-2,2710	-0,4310
Cirurgia	1,4810	0,7212	0,1432	2,9430
Resposta patológica completa	-0,9938	0,6679	-2,2630	0,3098
Receptor de Estrogênio	0,1817	0,6863	-1,1220	1,5980
Receptor de Progesterona	-1,0780	0,7662	-2,6040	0,3299

7.3. Análise Bayesiana dos tempos de sobrevida assumindo a distribuição geométrica bivariada proposta por de Basu-Dhar

Assumindo distribuições a priori uniformes $U(0,1)$ para os $p_1, p_2 \; e \; p_{12}$ da distribuição geométrica bivariada Basu-Dhar para os tempos de sobrevida livre de doença e os tempos de sobrevida total da Tabela A.1, não considerando a presença de covariáveis, também utilizando o software OpenBugs (amostra "burn-in" de 10.000 e amostra final de tamanho 100, tomando de 10 em 10 amostras de Gibbs entre 10.000 amostras simuladas) para encontrar os sumários a posteriori de interesse (ver Tabela 26).

A partir dos resultados da Tabela 26, as estimativas de Monte Carlo das médias a posteriori para o tempo de sobrevida livre de doença e o tempo de sobrevida total, são respectivamente, 111,8 meses e 296,8 meses, isto é, resultados semelhantes aos obtidos usando a distribuição de Block e Basu para o tempo de sobrevida livre de doença (108,4 meses), mas um pouco diferente para o tempo de sobrevida total (232,6 meses).

Tabela 26: Sumários a posteriori de interesse - Distribuição geométrica bivariada Basu-Dhar - sem a presença de covariáveis.

Parâmetro	Média	Desvio padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Inferior
Média 1 (SLD)	111,8000	29,2600	69,9700	177,9000
Média 2 (ST)	296,8000	121,6000	151,7000	559,6000
p_1	0,9924	0,0019	0,9882	0,9957
p_{12}	0,9981	0,0013	0,9952	0,9999
p_2	0,9981	0,0013	0,9949	0,9999

Agora, considere uma análise bayesiana dos dados de sobrevida discretos bivariados T_1 (tempo de sobrevida livre da doença) e T_2 (tempo de sobrevida total) na presença de covariáveis com uma distribuição geométrica bivariada Basu-Dhar e os seguintes modelos de regressão:

$$\text{logit}(p_{1i}) = \beta_{10} + \beta_{11}(\text{idade}_i - 48.29) + \beta_{12}\text{hercep}_i + \beta_{13}\text{estágio}_i + \beta_{14}\text{cirur}_i + \beta_{15}\text{pCR}_i + \beta_{16}\text{estrog}_i + \beta_{17}\text{progest}_i$$

$$\text{logit}(p_{2i}) = \beta_{20} + \beta_{21}(\text{idade}_i - 48.29) + \beta_{22}\text{hercep}_i + \beta_{23}\text{estágio}_i + \beta_{24}\text{cirur}_i + \beta_{25}\text{pCR}_i + \beta_{26}\text{estrog}_i + \beta_{27}\text{progest}_i$$

$$\text{logit}(p_{12i}) = \beta_{30} + \beta_{31}(\text{idade}_i - 48.29) + \beta_{32}\text{hercep}_i + \beta_{33}\text{estágio}_i + \beta_{34}\text{cirur}_i + \beta_{35}\text{pCR}_i + \beta_{36}\text{estrog}_i + \beta_{37}\text{progest}_i$$

Assumindo distribuições a priori normais $N(0,1)$ para todos os parâmetros de regressão e usando o software OpenBugs (amostra "burn-in" de 2.000 e 1.000 amostras finais tomadas de 10 em 10), temos na Tabela 27, os sumários a posteriori de interesse.

A partir dos resultados da Tabela 27, observa-se que a covariável estágio tem um efeito significativo (intervalo de credibilidade de 95% para os parâmetros da regressão não incluem o valor zero) para o tempo de sobrevida livre da doença e total. Daí, conclui-se que o estágio afeta tempos de sobrevida livre da doença e total.

Tabela 27: Sumários a posteriori de interesse - Distribuição geométrica bivariada Basu e Dhar - na presença de covariáveis.

Parâmetro	Média	Desvio padrão	Intervalo de Credibilidade (95%)	
			Limite Inferior	Limite Inferior
β_{10}	1,3400	0,8967	-0,3826	3,1100
Idade	0,0068	0,0101	-0,0127	0,0265
Herceptin	0,8981	0,6641	-0,2601	2,4000
Estágio	0,8167	0,3660	0,0935	1,5210
Cirurgia	-0,1487	0,4878	-1,0990	0,7388
Resposta patológica completa	0,6196	0,4829	-0,2563	1,6430
Receptor de Estrogênio	0,4131	0,5548	-0,6165	1,5260
Receptor de Progesterona	0,4341	0,5914	-0,7340	1,6240
β_{20}	0,9537	0,9497	-0,8931	2,8510
Idade	-0,0255	0,0527	-0,1516	0,0572
Herceptin	1,0080	0,8677	-0,6321	2,8090
Estágio	1,7860	0,6333	0,6847	3,2380
Cirurgia	-0,3264	0,8780	-1,9660	1,4400
Resposta patológica completa	0,7544	0,8104	-0,8618	2,3520
Receptor de Estrogênio	0,0949	0,8715	-1,6010	1,8400
Receptor de Progesterona	0,4968	0,9086	-1,2900	2,2490
β_{30}	1,0840	0,8819	-0,6912	2,8430
Idade	-0,0245	0,0518	-0,1481	0,0517
Herceptin	1,0030	0,8217	-0,6065	2,5550
Estágio	1,7230	0,6428	0,6972	3,2140
Cirurgia	-0,3913	0,9357	-2,0500	1,8130
Resposta patológica completa	0,7146	0,7643	-0,7982	2,2340
Receptor de Estrogênio	0,1289	0,8498	-1,4520	1,8830
Receptor de Progesterona	0,5573	0,8718	-1,2300	2,2090

É importante salientar que, para este modelo, a convergência do algoritmo de simulação MCMC considerando distribuições a priori não informativas foi facilmente obtida sem a necessidade de distribuições a priori informativas como foi assumido usando a distribuição geométrica bivariada de Arnold (uma vantagem da distribuição geométrica de Basu- Dhar, quando comparado com a distribuição geométrica de Arnold). Além disso, observa-se que o modelo de regressão assumindo uma distribuição geométrica Basu-Dhar é mais sensível para identificar os efeitos significativos das covariáveis.

7.4. Discussão dos resultados obtidos

A identificação de modelos apropriados para analisar dados de sobrevivência bivariadas na presença de censuras e covariáveis é de grande importância e interesse para muitas áreas de aplicação, tais como engenharia e medicina. Na presença de uma grande parte dos dados censurados, poderíamos ter grandes dificuldades para obter as inferências de interesse assumindo distribuições bivariadas contínuas apresentadas na literatura. Desta forma, a

utilização de distribuições discretas bivariadas poderia ser uma boa alternativa para analisar dados de sobrevida com alguma estrutura de dependência.

A utilização de métodos bayesianos e técnicas de simulação MCMC também abre um novo horizonte na análise desses dados, como observado nos resultados da análise dos dados de câncer de mama apresentados anteriormente. Além disso, observou-se que os modelos bivariados considerando dados discretos podem ser mais sensíveis e eficientes na obtenção de inferências de interesse. Inferências importantes foram obtidas para a nossa aplicação, considerando os dados de câncer de mama introduzidas na Tabela A.1.

Assumindo as três distribuições de sobrevivência bivariadas, vemos que não há diferenças significativas entre os tempos de sobrevida para os pacientes que receberam pelo menos quatro e menos de quatro ciclos de Herceptin® antes da cirurgia.

As covariáveis que mostraram evidências de afetar os tempos de sobrevida observados foram: estágio e tipo de cirurgia (não simultaneamente).

Sob um modelo de regressão para os parâmetros da distribuição de Block e Basu, vemos que somente a covariável estágio tem efeito significativo.

Sob um modelo de regressão para os parâmetros da distribuição Arnold assumindo distribuições a priori não informativas e tempos de sobrevida independentes T_1 e T_2, vemos que a covariável estágio tem efeito significativo. E para os parâmetros de regressão assumindo distribuições a priori informativas com o modelo bivariado Arnold (T_1 e T_2 dependentes), vemos que as covariáveis estágio e tipo de cirurgia têm efeitos significativos.

Finalmente, de acordo com um modelo de regressão para os parâmetros da distribuição bivariada de Basu-Dhar, apenas a covariável estágio tem efeito significativo (zero não incluso no intervalo de credibilidade 95% para os parâmetros da regressão associados no modelo).

A partir desses resultados de inferência considerando os três modelos, temos que,

- As pacientes com estágios avançados em geral, têm mais recidivas e morrem mais, como se observa nas estimativas não-paramétricas de Kaplan e Meier indicadas na Figura 2 e na Figura 4.
- O tipo de cirurgia é um fator de confusão. A cirurgia radical não afeta diretamente os tempos de sobrevida, na verdade, as pacientes em estágio mais avançado, no estágio 3, se submetem mais à cirurgia radical (72%). Por isso, tem-se a impressão de que

aquelas que fazem a cirurgia radical vivem menos, mas na verdade as pacientes mais suscetíveis são do estágio 3.

As estimativas de Monte Carlo para as médias a posteriori μ_1 e μ_2, médias de T_1 (sobrevida livre da doença) e T_2 (sobrevida total) são muito semelhantes assumindo a distribuição exponencial bivariada de Block e Basu (108,4 e 232,6 meses) e a distribuição Basu-Dhar (111,8 e 296,8 meses), mas os intervalos de credibilidade 95% são diferentes (maior para a distribuição Basu-Dhar). Assim, as estimativas de Monte Carlo para as médias a posteriori μ_1 e μ_2 são muito diferentes assumindo a distribuição bivariada de Arnold (140,4 e 343,6 meses), com intervalos de credibilidade 95% maiores.

É importante salientar, que métodos de discriminação devem ser desenvolvidos para a comparação dos diferentes modelos de sobrevida bivariados assumidos na análise dos dados de câncer de mama da Tabela A.1, na presença de um grande número de observações censuradas, pois alguns métodos de discriminação existentes como o critério DIC (Deviance Information Criterion) introduzido por Spiegelhalter et al (2002) pode não ser confiável para discriminar os modelos propostos.

Como uma forma empírica para comparar os modelos propostos, poderíamos comparar as estimativas de Monte Carlo obtidas das médias a posteriori para os tempos de sobrevida livre de doença e os tempos de sobrevida total com uma estimativa não-paramétrica (estimativas de Kaplan-Meier). A partir das estimativas apresentadas na Tabela 28, observa-se que as estimativas baseadas na distribuição de Block e Basu estão mais próximas das estimativas de Kaplan-Meier para as médias, sendo assim, uma possível indicação de melhor ajuste dos dados. Observe que o conjunto de dados apresenta uma grande proporção de observações censuradas e os modelos bivariados propostos são mais sensíveis para incorporar esse fato. Outra possibilidade em um trabalho futuro: uso de modelos bivariados com fração cura.

Tabela 28: Estimativas para as médias dos tempos de sobrevida livre de doença e os tempos de sobrevida global assumindo os modelos bivariados propostos.

Método	Sobrevida livre da doença	Sobrevida total
Kaplan-Meier	63,0	73,5
Block and Basu	108,4	232,6
Arnold	140,4	343,6
Basu-Dhar	111,8	296,8

8. Considerações Finais

O interesse do médico pesquisador no estudo que gerou o banco de dados aqui utilizado (Tabela A.1) foi de caracterizar as pacientes com câncer de mama localmente avançado com superexpressão do Her-2 que foram submetidas a quimioterapia neoadjuvante associada com o Herceptin® (Buzatto, 2015). Entender quais os fatores fazem com que algumas pacientes se beneficiem do tratamento enquanto que outras não e acabam vir a óbito. Uma primeira descrição dos dados foi apresentada na seção 1.5, Tabela 2.

O Herceptin® (Trastuzumabe) é uma medicação de alto custo, que enfrenta muitas dificuldades práticas para a sua obtenção e causa alguns efeitos colaterais (principalmente cardíacos). Devido a isso, a grande importância de estudos que possam evidenciar as características das pacientes que mais se beneficiam da medicação, tornando o tratamento de câncer de mama cada vez mais individualizado.

Na seção 1.7 foi mostrado o banco de dados utilizado (Tabela A.1) em particular não pode ser analisado utilizando o modelo de Cox de riscos proporcionais, que são usualmente utilizados na literatura médica. Piccart-Gebhart et. al. (2007) conduziram um estudo aleatorizado que acompanhou mulheres com câncer de mama que receberam Trastuzumabe no tratamento adjuvante por 1 ou 2 anos com mulheres que não receberam a medicação. Esse estudo contou com 1701 mulheres que tomaram a medicação por 2 anos, 1703 por 1 anos e 1698 controles (dessas, 861 optaram posteriormente em receber a medicação). O modelo de Cox de riscos proporcionais foi utilizado neste estudo para estimar os riscos relativos. Outro estudo recente que também utilizou o modelo de Cox foi o estudo de Gianni et. al. (2010) que compara pacientes que receberam Trastuzumabe por 1 ano (neoadjuvante e adjuvante; n=117) com paciente que não o receberam (controle; n=118). Esses estudos contam com tamanhos amostrais grandes, característica não presente no banco de dados utilizado nesse estudo, que pode prejudicar os resultados de análises dependentes de teorias assintóticas.

Os resultados do presente estudo apresentam alternativas para a análise de sobrevivência com tempos de sobrevida na presença de fração de cura, censuras e várias covariáveis. O modelo de riscos proporcionais de Cox nem sempre se adequa às características do banco de dados estudado, sendo necessária a busca de modelos estatísticos mais adequados que produzam inferências consistentes.

Usualmente na análise de dados de sobrevivência tem-se a presença de fração de cura, quando em certa proporção de indivíduos não ocorre o evento de interesse. Dessa forma, modelos tradicionais sem a presença de fração de cura podem não ser apropriados. Através das aplicações pode-se observar que a distribuição de Weibull é uma boa opção quando comparada a outras distribuições utilizadas em análise de sobrevivência, pois apresenta uma boa flexibilidade no ajuste e também por ser a distribuição que mais se adequou aos dados.

Em alguns casos, além da presença de fração de cura, podem-se ter dois ou mais tempos de sobrevida associados a cada unidade amostral. Sendo muito importante utilizar um parâmetro de dependência entre os tempos, se utilizando de distribuições bivariadas. Pela aplicação considerada, o modelo bivariado, permitiu aprimorar os resultados para a tomada de decisão e a utilização de distribuições discretas bivariadas poderia ser uma boa alternativa para analisar os dados com tempos de sobrevida bidimensional.

É importante salientar que os resultados obtidos sob o enfoque frequentista são menos precisos em relação aos resultados sob o enfoque bayesiano por se utilizarem de métodos assintóticos para os estimadores de máxima verossimilhança, por apresentarem uma grande proporção de dados censurados e por serem dependentes do tamanho amostral. Na aplicação apresentada, a presença de uma grande proporção de dados censurados pode levar a inferências assintóticas não muito precisas.

Um diferencial da técnica bayesiana em relação a frequentista, se dá devido a possibilidade de incorporar a informação do especialista, no caso, do médico. Dessa forma têm-se inferências mais precisas sob o enfoque bayesiano.

9. Algumas Perspectivas Futuras

A partir dos resultados obtidos neste trabalho, observa-se várias perspectivas promissoras para o desenvolvimento de trabalhos futuros considerando modelos paramétricos discretos e contínuos para os dados bivariados especialmente sob o enfoque bayesiano.

Na situação univariada é possível conduzir um estudo mais detalhado a respeito das prioris com o objetivo de obter resultados mais precisos com menores erro padrão.

Outros conjuntos de dados de sobrevivência com dados médicos podem ser considerados.

Uma possibilidade de estudo é considerar funções cópulas para capturar a dependência entre dados bivariados. Considerar também frações de curas para os novos modelos estudados.

Métodos de discriminação e técnicas de verificação de ajuste para os modelos de sobrevivência bivariados podem ser desenvolvidos com o objetivo de comparar diferentes modelos e definir o mais adequado a cada banco de dados utilizado.

10. Referências

ACHCAR, J. A.; BOLETA, J. Distribuição exponencial generalizada: uso de métodos Bayesianos. **Rev. Bras. Biom.**, São Paulo, v.27, n.4, p.644-658, 2009.

ALBERT, J. **Bayesian Computation with R**. New York: Springer-Verlag, 2007. 300p.

ARNOLD, B.C. A characterisation of the exponential distribution by multivariate geometric compounding. **Sankhya**, Series A, v.37, n.1, p.164-173, 1975.

ARNOLD, B.C.; STRAUSS, D. Bivariate distributions with exponential conditionals. **J. Amer. Statist. Assoc.**, v.83, p.522-527, 1988.

BARROS, A.C.S.D.; BARBOSA E.M.; GEBRIM L.H. **Diagnóstico e Tratamento de Câncer de Mama**. Associação Médica Brasileira e Conselho Federal de Medicina (Projeto Diretrizes). 15 Ago. 2001.

BASU, A. P.; DHAR, S. Bivariate geometric distribution. **Journal Applied Statistical Science**, v.2, n.1, p.33-44, 1995.

BEATSON, G. T. On the treatment of inoperable cases of carcinoma the mamma: suggestions for a new method of treatment, with illustrative cases. **Lancet**, v.2, p.104-107, 1896.

BERNARDO, J. M.; SMITH, A. F. M. **Bayesian theory**. New York: Wiley, 1994.

BILMORIA, M. M. The woman at increased risk for breast cancer: evaluation and management strategies. **Cancer**, n.45, p.263-78, 1995.

BLACKWELL K.; BULLOCK K. (2008), Clinical Efficacy of Taxane-Trastuzumab Combination Regimens for HER-2 Positive Metastatic Breast Cancer. **The Oncologist**, v.13, n.5, p.515-25, 2008.

BLASCO, A. The Bayesian controversy in animal breeding. **Journal of Animal Science**, v.79, p.2023-2046, 2001.

BLOCK, H.W.; BASU, A.P. A continuous bivariate exponential extension. **J. Amer. Statist. Assoc.**, v.69, n.348, p.1031-1037, 1974.

BORGES, E.C.; CAMARGO, G.C.; SOUZA, M.O.; PONTUAL, N.A.; NOVATO, T.S. Qualidade de vida em pacientes ostomizados: uma comparação entre portadores de câncer colorretal e outras patologias. **Rev. Inst. Ciênc. Saúde**, v.25, n.4, p.357-63, 2007.

BOX, G.E.P.; TIAO, G.C. **Bayesian Inference in Statistical Analysis**. New York: J. Wiley Interscience, 1992. 588p.

BOYLE, P.; LEVIN, B. **World Cancer Report**: 2008. Lyon: International Agency for Research on Cancer, 2008.

BRASIL. Ministério da Saúde. Instituto Nacional De Câncer José Alencar Gomes Da Silva. **Incidência de câncer no Brasil:** estimativa 2016. Rio de Janeiro: INCA, 2016.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. **Portaria n. 73,** de 30 de janeiro de 2013. Inclui procedimentos na Tabela de Procedimentos, Medicamentos, Órteses/Próteses e Materiais Especiais do SUS e estabelece protocolo de uso do trastuzumabe na quimioterapia do câncer de mama HER-2 positivo inicial e localmente avançado. Disponível em:

<http://bvsms.saude.gov.br/bvs/saudelegis/sas/2013/prt0073_30_01_2013.html>. Acesso em: 14 abr. 2016.

BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. **Trastuzumabe para tratamento do câncer de mama inicial:** relatório de recomendação da comissão nacional de incorporação de Tecnologia no SUS – CONITEC-07. Brasília, 2012. 30p.

BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. **Trastuzumabe para tratamento do câncer de mama inicial:** relatório de recomendação da comissão nacional de incorporação de Tecnologia no SUS – CONITEC-08. Brasília, 2012. 40p.

BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. **Portaria n. 18**, de 25 de julho de 2012. Torna pública a decisão de incorporar o medicamento trastuzumabe no Sistema Único de Saúde (SUS) para o tratamento do câncer de mama localmente avançado. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/sctie/2012/prt0018_25_07_2012.html>. Acesso em: 14 abr. 2016.

BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. **Portaria n. 19**, de 25 de julho de 2012. Torna pública a decisão de incorporar o medicamento trastuzumabe no Sistema Único de Saúde (SUS) para o tratamento do câncer de mama inicial. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/sctie/2012/prt0019_25_07_2012.html>. Acesso em: 14 abr. 2016.

BUZDAR, A. U. et al. Neoadjuvant therapy with paclitaxel followed by 5-fluorouracil, epirubicin, and cyclophosphamide chemotherapy and concurrent trastuzumab in human epidermal growth factor receptor 2-positive operable breast cancer: an update of the initial randomized study. **Clinical Cancer Research**, v.13, n.1, p.228–233, 2007.

CARLIN, B. P.; LOUIS, T. A. **Bayes and Empirical Bayes Methods for Data Analysis**. London: Chapman Hall, 2002.

CARRASCO, J.M.; ORTEGA, E.M.M.; CORDEIRO, G.M.A. Generalized Modified Weibull Distribution for Lifetime Modelling. **Computational Statistics and Data Analysis**, v.53, p.450–462, 2008.

CASELLA G.; GEORGE, E. I. Explaining the Gibbs sampler. **The American Statistician**, v.46, p.167–174, 1992.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings algorithm. **The American Statistician**, v. 49, 327–335, 1995.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. São Paulo: Edgard Blucher Ltda., 2006. 205 p.

COMPTON, C. C. et al. **AJCC Cancer staging atlas**: a companion to the seventh editions of the ajcc cancer staging manual and handbook. New York: Springer, 2012. 637p.

CORTAZAR, P. et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. **Lancet**, v.384, n.9938, p.164-172, 2014.

COX, D. R. Regression models and life tables. **Journal of the Royal Statistical Society B**, v.34, n.2, p.187–220, 1972.

COX, D. R.; OAKES, D. **Analysis of Survival Data**. London: Chapman & Hall, 1984. 198p.

DESANTIS, C.; MA, J.; BRYAN, L. J. A. Breast cancer statistics, 2013. **Cancer J Clin**, v.64, n.1, p.52-62, 2014.

DOWNTON, F. Bivariate exponential distributions in reliability theory. **Journal of the Royal Statistical Society B**, v.32, p.408–417, 1970.

EFRON, B. The Efficiency of Cox's Likelihood Function for Censored Data. **Journal of the American Statistical Association**, v.72, n.359, p.557-565, 1977.

FARANTE, G. et al. Novo TNM: Classificação do câncer de mama proposta pelo Instituto Europeu de Oncologia de Milão, Itália. **Rev. Bras. Mastologia**, v.20, n.2, p.61-65, 2010.

FERLAY, J. et al. **Cancer Incidence and Mortality Worldwide**: No. 11. Lyon: International Agency for Research on Cancer, 2013. Disponível em: <http://globocan.iarc.fr>. Acesso em: 04 out. 2015.

FREUND, J. E. A bivariate extension of the exponential distribution. **Journal of the American Statistical Association**, v.56, p.971–977, 1961.

GELBER R.D. et al. Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. **New England Journal of Medicine**, v.353, p.1659-72, 2005.

GELFAND, A. E.; SMITH, A. F. M. Sampling based approaches to calculating marginal densities. **Journal of the American Statistical Association**, v.85, p.398–409, 1990.

GIANNI, L. Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with HER2-positive locally advanced breast cancer (the NOAH trial): a randomised controlled superiority trial with a parallel HER2-negative cohort. **Lancet**, v.375, p.377-84, 2010.

GIANOLA, D.; FERNANDO, R.L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v.63, p.217-244, 1986.

GRAMBSCH, P. M.; THERNEAU, T. M. Proportional Hazards Tests and Diagnostics based on Weighted Residuals. **Biometrika**, v.81, n.3, p.515-526, 1994.

GUMBEL, E. J. Bivariate exponential distributions. **Journal of the American Statistical Association**, v.55, p.698–707, 1960.

GUMBEL, E. J. Statistical theory of extreme values and some practical applications. **Applied Mathematics Series**, v.33, 1955.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions. **Australian and New Zealand Journal of Statistics**, v.41, p.173-188, 1999.

HOUGAARD, P. A class of multivariate failure time distributions. **Biometrika**, v.3, n.73, p.671–678, 1986.

HOUGAARD, P. Fundamentals of survival data. **Biometrics**, v.55, n.1, p.13-22, 1999.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data**. 2 ed. New York: John Wiley and Sons, 1980. 447p.

KAPLAN, E.L.; MEIER, P. Nonparametric estimation from incomplete observations. **J. Amer. Statist. Ass.**, v.53, n.282, p.457–48, 1958.

KELSEY, J.L.; GAMMON, M.D.; JOHN, E.M. Reproductive factors and breast cancer. **Epidemiol Rev**, v.15, n.1, p.36-47, 1993.

KHATIB, O. M. N.; MODJTABAI, A. (Ed.). **Guidelines for the early detection and screening of breast cancer.** [S.l.]: World Health Organization, 2006. 57 p. (EMRO Technical Publications Series; 30). Disponível em: <http://applications.emro.who.int/dsaf/dsa696.pdf>. Acesso em: 14 abr. 2006.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data**. New York: John Wiley, 1982. 580p.

LEE, E.T.; WENYUWANG, J. **Statistical methods for survival data analysis**. 3. ed. New York: John Wiley & Sons, 2003. 535 p.

LOUZADA, F.; MAZUCHELI, J.; ACHCAR, J. A. **Introdução à análise de sobrevivência e confiabilidade**. São Carlos: IMCA, 2002.

MALLER, R. A.; ZHOU, X. (1996), **Survival analysis with long-term survivors**. Chinchester: John Wiley & Sons, 1996. 278 p. (Wiley Series in Probability and Statistics: Applied Probability and Statistics, book 16).

MARSHALL, A. W.; OLKIN, I. A multivariate exponential distribution. **Journal of the American Statistical Association**, v.62, p.30–44, 1967b.

MARSHALL, A. W.; OLKIN. I. A generalized bivariate exponential distribution. **Journal of Applied Probability**, v.4, p.291–302, 1967a.

MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. **Introduction to the Theory of Statistics**. [S.l.]: McGraw-Hill, 1974. 577p.

MUDHOLKAR, G.S.; SRIVASTAVA, D.K. Exponentiated Weibull family for analyzing bathtub failure-rate data. **IEEE Transactions on Reliability**, v.42, n.2, p.299–302, 1993.

NAIR, K. R. M.; NAIR, N. U. On characterizing a bivariate geometric distribution. **Ann. Inst. Statist. Math**. v.40, n.2, p.267-71, 1988.

NELSON, W. **Applied life data analysis**. New Jersey: John Wiley & Sons, 2004. 662p.

O que é o Cancêr?: genes. Instituto Vencer o Cancer, 2013. Disponível em: <http://vencerocancer.com.br>. Acesso em: 06 abr. 2016.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446p.

PICCART-GEBHART, M.J. et al. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomized controlled trial. **Lancet**, v.369, n. 9555, p.29-36, 2007.

PINHO, V. F. S.; COUTINHO, E. S. F. Variáveis associadas ao câncer de mama em usuárias de unidades básicas de saúde. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 23, n. 5p. 1061-1069, 2007.

RAQAB, M. Z. Inferences for generalized exponential distribution based on record statistics. **Journal of Statistical Planning and Inference**, v.104, p.339-350, 2002.

RAQAB, M. Z.; AHSANULLAH, M. Estimation of the location and scale parameters of generalized exponential distribution based on order statistics. **Journal of Statistical Computation and Simulation**, v.69, p.109-124, 2001.

SANCHEZ-MUNOZ, A. et al. The role of immunohistochemistry in breast cancer patients treated with neoadjuvant chemotherapy: an old tool with an enduring prognostic value. **Clinical Breast Cancer**, v.13, n.2, p.146-52, 2013.

SARHAN, A. M. Analysis of incomplete, censored data in competing risks models with generalized exponential distributions. **IEEE Transactions on Reliability**, v.56, p.132-138, 2007.

SCHOENFELD, D. Partial Residuals for the Proportional Hazard Regression Model. **Biometrika**, v.69, n.1, p.239-241, 1982.

SLAMON, D.J. et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. **Science**, v.244, n.4905, p.707-12, 1989.

SLAMON, D.J. et al. Use of chemotherapy plus a monoclonal antibody against Her 2 for metastatic breast cancer that overexpresses Her2. **New England Journal of Medicine**, v.344, n.11, p.783-92, 2001.

SPIEGELHALTER, D. et al. **WinBUGS User Manual**: version 1.4. Cambridge: MRC Biostatistics Uni, 2003.

STEWART, W.; CHRISTOPHER, PW. **World Cancer Report**: 2014. Lyon: International Agency for Research on Cancer, 2014.

STRUTHERS, C. A.; KALBFLEISCH, J. D. Misspecified Proportional Hazards Models. **Biometrika**, v.73, n.2, p.363-369,1986.

TEIXEIRA, L.C.; PINOTTI, J.A. Câncer na mama: Quimioterapia. In: HALBE, H.W. **Tratado de Ginecologia**. São Paulo: Rocca, 2000.

VON MINCKWITZ, G. et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. **Journal of Clinical Oncology**, v.30, n.15, p.1796-804, 2012.

VU, T.; CLARET, F.X. Trastuzumab: updated mechanisms of action and resistance in breast cancer. **Frontiers in Oncology**, v.2, n.62, 2012.

WEIBULL, W. A Statistical distribution function of wide applicability. **Journal of Applied Mechanics**, p.293–7, dec. 1952.

A. Conjunto de dados de pacientes com Câncer de mama

Tabela A.1 - Dados de 54 pacientes com câncer de mama.

Idade	Hercep	Est	Cirur	pCR	Estr	Proge	Recidiva	SLD	Óbito	ST
50	1	3	1	1	0	0	0	60	0	60
24	1	2	0	1	1	1	0	45	0	45
44	1	3	1	1	1	1	0	83	0	83
43	1	3	1	1	1	1	0	53	0	53
29	1	3	0	0	1	1	1	30	0	58
40	1	3	1	0	0	0	0	72	0	72
48	1	3	1	1	1	1	0	30	0	30
62	1	3	0	0	0	0	0	65	0	65
51	1	3	0	1	0	0	0	68	0	68
48	1	3	1	1	0	0	1	20	0	60
50	1	3	0	0	1	0	0	64	0	64
44	1	3	1	0	0	1	0	33	0	33
63	1	2	1	1	0	0	0	37	0	37
52	1	3	1	0	1	0	0	27	0	27
35	1	3	1	0	0	0	1	22	0	59
41	1	3	1	0	0	0	0	34	0	34
57	1	3	1	1	0	0	0	34	0	34
57	1	3	0	1	1	0	0	46	0	46
32	1	3	1	0	0	0	0	32	0	32
71	1	3	1	0	0	0	0	66	0	66
37	1	3	1	0	1	0	1	53	1	61
62	1	2	0	1	0	0	0	55	0	55
42	1	3	1	1	0	0	0	65	0	65
30	1	3	1	0	1	1	1	44	1	50
60	1	3	1	0	1	0	1	15	1	53
51	1	3	1	0	0	0	1	33	1	37
62	1	3	1	0	0	0	1	12	1	17
47	1	2	0	0	1	1	0	59	0	59
42	1	3	1	0	1	0	0	39	0	39
42	1	3	1	0	1	1	0	29	0	29
63	1	3	1	0	0	0	0	35	0	35
57	1	3	1	1	0	0	1	22	1	28
56	1	2	*	*	0	0	*	8	1	8
63	1	3	1	0	0	0	*	22	0	22
30	1	3	0	0	0	0	1	47	0	62
34	1	2	1	0	0	0	0	25	0	25
39	1	3	1	1	0	0	1	48	0	58
41	1	3	1	0	1	1	1	49	0	83
58	1	2	0	1	0	0	1	31	0	41
57	2	3	0	0	0	0	0	42	0	42
39	2	3	0	1	0	0	0	30	0	30
65	2	3	0	0	1	1	0	30	0	30
54	2	3	1	0	1	1	0	56	0	56
53	2	3	1	1	0	0	0	32	0	32
49	2	3	0	0	1	1	0	40	0	40
57	2	3	1	1	0	0	1	39	1	44
41	2	3	1	0	1	1	0	37	0	37
62	2	3	0	0	0	0	1	24	0	34
56	2	3	0	1	1	0	0	58	0	58
52	2	2	1	1	1	1	0	29	0	29
49	2	3	1	1	0	0	0	44	0	44
40	2	3	1	1	1	1	0	22	0	22
51	2	2	1	1	1	1	0	31	0	31
48	2	2	0	1	1	1	0	16	0	16

- Idade: Idade da Paciente (0: ≤40 anos; 1:>40 anos)
- Hercep: Uso do medicamento Herceptin® (1: ≥4 ciclos; 2: <4 ciclos)
- Est: Estágio da doença (2 ou 3)
- Cirur: Tipo de cirurgia realizada na Paciente (1:radical; 0:conservadora)
- pCR: Resposta Patológica Completa (1:Sim; 0:Não)
- Estr: Receptor de Estrogênio (1:positivo; 0:negativo)
- Proge: Receptor de Progesterona (1:positivo; 0:negativo)

B. Programas utilizados no Open Bugs

Esse apêndice apresenta os programas computacionais desenvolvidos no software OpenBUGS versão 3.2.3, utilizados nas seções 5 e 7 deste presente trabalho.

B1. Modelo de Weibull sob enfoque bayesiano

```
model {
    for (i in 1:N) {
        t[i] ~ dweib(alpha,theta)I(delta[i],)
    }
    alpha ~ dgamma(0.1,0.1)
    theta ~ dgamma(0.1,0.1)
    b <- pow(theta,1/alpha)
    média.tempo <- (1/b)*exp(loggam(1+1/alpha))
    lambda <- 1/b
}
```

B2. Modelo de Weibull bayesiano na presença de fração de cura

```
model {
    for (i in 1:N) {
        zeros[i] <- 0
        phi[i] <- -log(L[i])
        zeros[i] ~ dpois(phi[i])
        a1[i]<- pow(t[i]/lambda,k-1)
        a2[i]<- pow(t[i]/lambda,k)
        a3[i]<- exp(-a2[i])
        f[i]<- (k/lambda)*a1[i]*a3[i]
        S0[i]<- a3[i]
        L[i]<- exp(delta[i]*log(1-phi1)+ delta[i]*log(f[i])+(1-delta[i])*log(phi1+(1-phi1)*S0[i]))
    }
    k ~ dgamma(1,1)
    phi1~ dbeta(70,30)
    lambda~ dunif(0,300)
}
```

B3. Modelo de Weibull sob enfoque bayesiano com covariáveis

```
model {
    for(i in 1 : N) {
        t[i] ~ dweib(alpha,lambda[i])I(delta[i],)
        lambda[i] <-
exp(beta0+beta1*idade[i]+beta2*herceptin[i]+beta3*estágio[i]+beta4*tipo.cirurgia[i]+beta5*pCR[i]+
beta6* estrogênio [i]+beta7*progesterona[i])
        b[i] <- pow(lambda[i],1/alpha)
        média.tempo[i] <- (1/b[i])*exp(loggam(1+1/alpha))
    }
    alpha ~ dgamma(1,1)
    beta0 ~ dnorm(0,1)
    beta1 ~ dnorm(0,1)
    beta2 ~ dnorm(0,1)
    beta3 ~ dnorm(0,1)
    beta4 ~ dnorm(0,1)
    beta5 ~ dnorm(0,1)
    beta6 ~ dnorm(0,1)
    beta7 ~ dnorm(0,1)
}
```

B4. Modelo de Weibull sob enfoque bayesiano na presença de fração de cura afetando o parâmetro de escala com covariáveis

```
1         model {
2             for (i in 1:N) {
3                 zeros[i] <- 0
4                 phi[i] <- -log(L[i])
5                 zeros[i] ~ dpois(phi[i])
6                 a1[i]<- pow(t[i]/lambda[i],k-1)
7                 a2[i]<- pow(t[i]/lambda[i],k)
8                 a3[i]<- exp(-a2[i])
9                 f[i]<- (k/lambda[i])*a1[i]*a3[i]
10                S0[i]<- a3[i]
11                lambda[i] <-
        exp(beta0+beta1*idade[i]+beta2*herceptin[i]+beta3*estágio[i]+beta4*tipo.cirurgia[i]+beta5*pCR[i]+
        beta6* estrogênio [i]+beta7*progesterona[i])
12                L[i]<- exp(delta[i]*log(1-phi1)+ delta[i]*log(f[i])+(1-delta[i])*log(phi1+(1-phi1)*S0[i]))
13            }
14            k ~ dgamma(1,1)
15            phi1~ dbeta(70,30)
16            beta0 ~ dnorm(0,1)
17            beta1 ~ dnorm(0,1)
18            beta2 ~ dnorm(0,1)
19            beta3 ~ dnorm(0,1)
20            beta4 ~ dnorm(0,1)
21            beta5 ~ dnorm(0,1)
22            beta6 ~ dnorm(0,1)
23            beta7 ~ dnorm(0,1)
24        }
```

B5. Modelo de Weibull sob enfoque bayesiano na presença de fração de cura afetando o parâmetro de escala e a probabilidade de cura com covariáveis

```
1   model {
2       for (i in 1:N) {
3           zeros[i] <- 0
4           phi[i] <- -log(L[i])
5           zeros[i] ~ dpois(phi[i])
6           a1[i]<- pow(t[i]/lambda[i],k-1)
7           a2[i]<- pow(t[i]/lambda[i],k)
8           a3[i]<- exp(-a2[i])
9           f[i]<- (k/lambda[i])*a1[i]*a3[i]
10          S0[i]<- a3[i]
11          lambda[i] <-
    exp(beta0+beta1*idade[i]+beta2*herceptin[i]+beta3*estágio[i]+beta4*tipo.cirurgia[i]+beta5*pCR[i]+beta6*
    estrogênio [i]+beta7*progesterona[i])
12          logit(phi1[i])<-
    alpha0+alpha1*idade[i]+alpha2*herceptin[i]+alpha3*estágio[i]+alpha4*tipo.cirurgia[i]+alpha5*pCR[i]+alp
    ha6*estrogênio [i]+alpha7*progesterona[i]
13          L[i]<- exp(delta[i]*log(1-phi1[i])+ delta[i]*log(f[i])+(1-delta[i])*log(phi1[i]+(1-phi1[i])*S0[i]))
14      }
15      k ~ dgamma(1,1)
16      beta0 ~ dnorm(0,1)
17      beta1 ~ dnorm(0,1)
18      beta2 ~ dnorm(0,1)
19      beta3 ~ dnorm(0,1)
20      beta4 ~ dnorm(0,1)
21      beta5 ~ dnorm(0,1)
```

```
22      beta6 ~ dnorm(0,1)
23      beta7 ~ dnorm(0,1)
24      alpha0 ~ dnorm(0,1)
25      alpha1 ~ dnorm(0,1)
26      alpha2 ~ dnorm(0,1)
27      alpha3 ~ dnorm(0,1)
28      alpha4 ~ dnorm(0,1)
29      alpha5 ~ dnorm(0,1)
30      alpha6 ~ dnorm(0,1)
31      alpha7 ~ dnorm(0,1)
32      }
```

B6. Distribuição exponencial bivariada Block e Basu sem a presença de covariáveis

```
1   model {
2       lambda<- lambda1+lambda2+lambda3
3       lambda12<- lambda1+lambda2
4       lambda13<- lambda1+lambda3
5       lambda23<- lambda2+lambda3
6       a1<- (lambda*lambda1*lambda23)/lambda12
7       a2<- (lambda*lambda2*lambda13)/lambda12
8       mean1<- 1/lambda13+(lambda2*lambda3)/(lambda*lambda12*lambda13)
9       mean2<- 1/lambda23+(lambda1*lambda3)/(lambda*lambda12*lambda23)
10      d1<-lambda2*lambda3*(2*lambda1*lambda+lambda2*lambda3)
11      var1<-1/pow(lambda13,2)+d1/(pow(lambda,2)*pow(lambda12,2)*pow(lambda13,2))
12      sd1<-sqrt(var1)
13      d2<-lambda1*lambda3*(2*lambda2*lambda+lambda1*lambda3)
14      var2<-1/pow(lambda23,2)+d2/(pow(lambda,2)*pow(lambda12,2)*pow(lambda23,2))
15      sd2<-sqrt(var2)
16      b1<-(pow(lambda1,2)+pow(lambda2,2))*lambda3*lambda+lambda1*lambda2*pow(lambda3,2)
17      b2<-            pow(lambda,2)*lambda12*lambda13*lambda23
18      cov12<-b1/b2
19      rho12<-cov12/(sd1*sd2)
20              for (i in 1:N) {
21                      zeros[i] <- 0
22                      phi[i] <- -log(L[i])
23                      zeros[i] ~ dpois(phi[i])
24                      f1[i]<- a1*exp(-lambda1*t1[i]-lambda23*t2[i])
25                      f2[i]<- a2*exp(-lambda13*t1[i]-lambda2*t2[i])
26                      S1[i]<- (lambda/lambda12)*exp(-lambda1*t1[i]-lambda23*t2[i])-
    (lambda3/lambda12)*exp(-lambda*t2[i])
27                      S2[i]<- (lambda/lambda12)*exp(-lambda13*t1[i]-lambda2*t2[i])-
    (lambda3/lambda12)*exp(-lambda*t1[i])
28                      Sstar1t1[i]<- (lambda*lambda1)/(lambda12)*exp(-lambda1*t1[i]-lambda23*t2[i])
29                      Sstar2t1[i]<- (lambda*lambda13)/(lambda12)*exp(-lambda13*t1[i]-lambda2*t2[i])-
    (lambda*lambda3)/(lambda12)*exp(-lambda*t1[i])
30                      Sstar1t2[i]<- (lambda*lambda23)/(lambda12)*exp(-lambda1*t1[i]-lambda23*t2[i])-
    (lambda*lambda3)/(lambda12)*exp(-lambda*t2[i])
31                      Sstar2t2[i]<- (lambda*lambda2)/(lambda12)*exp(-lambda13*t1[i]-lambda2*t2[i])
32                      L[i]<- exp(v[i]*delta1[i]*delta2[i]*log(f1[i])+ (1-v[i])*delta1[i]*delta2[i]*log(f2[i])+
    v[i]*delta1[i]*(1-delta2[i])*log(Sstar1t1[i]) + (1-v[i])*delta1[i]*(1-delta2[i])*log(Sstar2t1[i]) + v[i]*(1-
    delta1[i])*delta2[i]*log(Sstar1t2[i]) + (1-v[i])*(1-delta1[i])*delta2[i]*log(Sstar2t2[i]) + v[i]*(1-
    delta1[i])*(1-delta2[i])*log(S1[i]) + (1-v[i])*(1-delta1[i])*(1-delta2[i])*log(S2[i]))
33              }
34      lambda1~ dgamma(1,100)
35      lambda2~ dgamma(1,100)
36      lambda3~ dgamma(1,100)
37      }
```

B7. Distribuição exponencial bivariada Block e Basu com a presença de covariáveis

```
model {
    for (i in 1:N) {
        lambda1[i]<- alpha1*
exp(beta11*idade[i]+beta12*herceptin[i]+beta13*estágio[i]+beta14*tipo.cirurgia[i]+beta15*pCR[i]+beta16
*estrogênio[i]+beta17*progesterona[i])
        lambda2[i]<- alpha2*
exp(beta21*idade[i]+beta22*herceptin[i]+beta23*estágio[i]+beta24*tipo.cirurgia[i]+beta25*pCR[i]+beta26
*estrogênio[i]+beta27*progesterona[i])
        lambda[i]<- lambda1[i]+lambda2[i]+lambda3
        lambda12[i]<- lambda1[i]+lambda2[i]
        lambda13[i]<- lambda1[i]+lambda3
        lambda23[i]<- lambda2[i]+lambda3
        a1[i]<- (lambda[i]*lambda1[i]*lambda23[i])/lambda12[i]
        a2[i]<- (lambda[i]*lambda2[i]*lambda13[i])/lambda12[i]
        zeros[i] <- 0
        phi[i] <- -log(L[i])
        zeros[i] ~ dpois(phi[i])
        f1[i]<- a1[i]*exp(-lambda1[i]*t1[i]-lambda23[i]*t2[i])
        f2[i]<- a2[i]*exp(-lambda13[i]*t1[i]-lambda2[i]*t2[i])
        S1[i]<- (lambda[i]/lambda12[i])*exp(-lambda1[i]*t1[i]-lambda23[i]*t2[i])-
(lambda3/lambda12[i])*exp(-lambda[i]*t2[i])
        S2[i]<- (lambda[i]/lambda12[i])*exp(-lambda13[i]*t1[i]-lambda2[i]*t2[i])-
(lambda3/lambda12[i])*exp(-lambda[i]*t1[i])
        Sstar1t1[i]<- (lambda[i]*lambda1[i])/(lambda12[i])*exp(-lambda1[i]*t1[i]-lambda23[i]*t2[i])
        Sstar2t1[i]<- (lambda[i]*lambda13[i])/(lambda12[i])*exp(-lambda13[i]*t1[i]-
lambda2[i]*t2[i])- (lambda[i]*lambda3)/(lambda12[i])*exp(-lambda[i]*t1[i])
        Sstar1t2[i]<- (lambda[i]*lambda23[i])/(lambda12[i])*exp(-lambda1[i]*t1[i]-
lambda23[i]*t2[i])- (lambda[i]*lambda3)/(lambda12[i])*exp(-lambda[i]*t2[i])
        Sstar2t2[i]<- (lambda[i]*lambda2[i])/(lambda12[i])*exp(-lambda13[i]*t1[i]-lambda2[i]*t2[i])
        L[i]<- exp(v[i]*delta1[i]*delta2[i]*log(f1[i])+ (1-v[i])*delta1[i]*delta2[i]*log(f2[i])+
v[i]*delta1[i]*(1-delta2[i])*log(Sstar1t1[i]) + (1-v[i])*delta1[i]*(1-delta2[i])*log(Sstar2t1[i]) + v[i]*(1-
delta1[i])*delta2[i]*log(Sstar1t2[i]) + (1-v[i])*(1-delta1[i])*delta2[i]*log(Sstar2t2[i]) + v[i]*(1-
delta1[i])*(1-delta2[i])*log(S1[i]) + (1-v[i])*(1-delta1[i])*(1-delta2[i])*log(S2[i]))
        mean1[i]<- 1/lambda13[i]+(lambda2[i]*lambda3)/(lambda[i]*lambda12[i]*lambda13[i])
        mean2[i]<- 1/lambda23[i]+(lambda1[i]*lambda3)/(lambda[i]*lambda12[i]*lambda23[i])
    }
lambda3~ dgamma(1,100)
alpha1~ dgamma(1,1)
alpha2~ dgamma(1,1)
beta11~ dnorm(0,1)
beta12~ dnorm(0,1)
beta13~ dnorm(0,1)
beta14~ dnorm(0,1)
beta15~ dnorm(0,1)
beta16~ dnorm(0,1)
beta17~ dnorm(0,1)

beta21~ dnorm(0,1)
beta22~ dnorm(0,1)
beta23~ dnorm(0,1)
beta24~ dnorm(0,1)
beta25~ dnorm(0,1)
beta26~ dnorm(0,1)
beta27~ dnorm(0,1)
}
```

B8. Distribuição geométrica bivariada Arnold sem a presença de covariáveis

```
model {
    gamma1 <- 1-theta1-theta2
    gamma2 <- 1-theta1
    gamma3 <- 1-theta2
        for (i in 1:N) {
            zeros[i] <- 0
            phi[i] <- -log(L[i])
            zeros[i] ~ dpois(phi[i])
            a1[i]<- pow(gamma1,t1[i]-1)
            a2[i]<- pow(gamma3,t2[i]-t1[i]-1)
            a3[i]<- pow(gamma1,t2[i]-1)
            a4[i]<- pow(gamma2,t1[i]-t2[i]-1)
            P1[i]<- theta1*theta2*a1[i]*a2[i]
            P2[i]<- theta1*theta2*a3[i]*a4[i]
            a5[i]<- pow(gamma1,t2[i])
            a6[i]<- pow(gamma2,t1[i]-t2[i]-1)
            S1[i]<- theta1*a1[i]*a2[i]
            S2[i]<- theta1*a5[i]*a6[i]
            a7[i]<- pow(gamma1,t1[i])
            a8[i]<- pow(gamma3,t2[i]-t1[i]-1)
            a9[i]<- pow(gamma2,t1[i]-t2[i])
            R1[i]<- theta2*a8[i]*a7[i]
            R2[i]<- theta2*a9[i]*a3[i]
            a10[i]<- pow(gamma3,t2[i]-t1[i])
            U1[i]<- a10[i]*a7[i]
            U2[i]<- a9[i]*a5[i]
            L[i]<- exp(v[i]*delta1[i]*delta2[i]*log(P1[i])+(1-
v[i])*delta1[i]*delta2[i]*log(P2[i])+v[i]*delta1[i]*(1-delta2[i])*log(S1[i])+
            (1-v[i])*delta1[i]*(1-delta2[i])*log(S2[i]) + v[i]*(1-delta1[i])*delta2[i]*log(R1[i]) +
            (1-v[i])*(1-delta1[i])*delta2[i]*log(R2[i]) + v[i]*(1-delta1[i])*(1-delta2[i])*log(U1[i])
+ (1-v[i])*(1-delta1[i])*(1-delta2[i])*log(U2[i]))
            }
    theta1<- p[1]
    theta2<- p[2]
    r<- p[3]
    p[1:3]~ddirich(alpha[])
    mean1<-(1-theta1)/theta1
    mean2<-(1-theta2)/theta2
}
```

B9. Distribuição geométrica bivariada Arnold com a presença de covariáveis

```
model{
    for (i in 1:N) {
        zeros[i] <- 0
        phi[i] <- -log(L[i])
        zeros[i] ~ dpois(phi[i])
        a1[i] <- 1-theta[i]
        p1[i] <- theta[i]*pow(a1[i],t1[i])
        p2[i] <- pow(a1[i],t1[i])
        L[i]<- exp(delta1[i]*log(p1[i])+(1-delta1[i])*log(p2[i]))
        logit(theta[i]) <-beta10+beta11*idade[i]+beta12*herceptin[i]+
    beta13*estágio[i]+beta14*tipo.cirurgia[i]+beta15*pCR[i]+beta16*estrogênio[i]+beta17*progesterona[i]
        mean[i]<-(1-theta[i])/theta[i]
    }
beta10~dnorm(0,1)
beta11~ dnorm(0,1)
beta12~dnorm(0,1)
beta13~dnorm(0,1)
```

```
beta14~dnorm(0,1)
beta15~dnorm(0,1)
beta16~dnorm(0,1)
beta17~ dnorm(0,1)
}
```

B10. Distribuição geométrica bivariada Arnold com a presença de covariáveis e utilizando distribuições a priori informativas

```
model {
    for (i in 1:N) {
    zeros[i] <- 0
    phi[i] <- -log(L[i])
    zeros[i] ~ dpois(phi[i])
    gamma1[i] <- 1-theta1[i]-theta2[i]
    gamma2[i] <- 1-theta1[i]
    gamma3[i] <- 1-theta2[i]
    a1[i]<- pow(gamma1[i],t1[i]-1)
    a2[i]<- pow(gamma3[i],t2[i]-t1[i]-1)
    a3[i]<- pow(gamma1[i],t2[i]-1)
    a4[i]<- pow(gamma2[i],t1[i]-t2[i]-1)
    P1[i]<- theta1[i]*theta2[i]*a1[i]*a2[i]
    P2[i]<- theta1[i]*theta2[i]*a3[i]*a4[i]
    a5[i]<- pow(gamma1[i],t2[i])
    a6[i]<- pow(gamma2[i],t1[i]-t2[i]-1)
    S1[i]<- theta1[i]*a1[i]*a2[i]
    S2[i]<- theta1[i]*a5[i]*a6[i]
    a7[i]<- pow(gamma1[i],t1[i])
    a8[i]<- pow(gamma3[i],t2[i]-t1[i]-1)
    a9[i]<- pow(gamma2[i],t1[i]-t2[i])
    R1[i]<- theta2[i]*a8[i]*a7[i]
    R2[i]<- theta2[i]*a9[i]*a3[i]
    a10[i]<- pow(gamma3[i],t2[i]-t1[i])
    U1[i]<- a10[i]*a7[i]
    U2[i]<- a9[i]*a5[i]
    logit(theta1[i]) <-
beta10+beta11*idade[i]+beta12*herceptin[i]+beta13*estágio[i]+beta14*tipo.cirurgia[i]+beta15*pCR[i]+bet
a16*estrogênio[i]+beta17*progesterona[i]
    logit(theta2[i]) <- beta20+
beta21*idade[i]+beta22*herceptin[i]+beta23*estágio[i]+beta24*tipo.cirurgia[i]+beta25*pCR[i]+beta26*estr
ogênio[i]+beta27*progesterona[i]
    L[i]<- exp(v[i]*delta1[i]*delta2[i]*log(P1[i])+(1-v[i])*delta1[i]*delta2[i]*log(P2[i])+v[i]*delta1[i]*(1-
delta2[i])*log(S1[i])+(1-v[i])*delta1[i]*(1-delta2[i])*log(S2[i])+v[i]*(1-delta1[i])*delta2[i]*log(R1[i])+(1-
v[i])*(1-delta1[i])*delta2[i]*log(R2[i]) + v[i]*(1-delta1[i])*(1-delta2[i])*log(U1[i])+(1-v[i])*(1-
delta1[i])*(1-delta2[i])*log(U2[i]))
    }
beta10~ dnorm(-1.1500,1)
beta11~ dnorm(-0.6106,1)
beta12~ dnorm(-0.7497,1)
beta13~ dnorm(-0.7921,1)
beta14~ dnorm(0.1646,1)
beta15~ dnorm(-0.5464,1)
beta16~ dnorm(-0.3122,1)
beta17~ dnorm(-0.4517,1)
beta20~ dnorm(-1.1310,1)
beta21~ dnorm(0.0109,1)
beta22~ dnorm(-0.8977,1)
beta23~ dnorm(-1.3640,1)
beta24~ dnorm(0.8478,1)
```

```
beta25~ dnorm(-0.7906,1)
beta26~ dnorm(-0.0024,1)
beta27~ dnorm(-0.7040,1)
}
```

B11. Distribuição geométrica bivariada Basu-Dhar sem a presença de covariáveis

```
model {
    for (i in 1:N) {
        zeros[i] <- 0
        phi[i] <- -log(L[i])
        zeros[i] ~ dpois(phi[i])
        z1[i]<-max(t1[i]-1,t2[i])
        z2[i]<-max(t1[i],t2[i])
        z3[i]<-max(t1[i],t2[i]-1)
        log(A1[i])<-(t1[i]-1)*log(p1)+ (t2[i]-1)*(log(p2)+log(p12))+log(1-p1)+log(1-p2*p12)
        log(A2[i])<-(t1[i]-1)*(log(p1)+log(p2)+log(p12))+ log(1-p1*p2-p2*p12+p1*p2*p12)
        log(A3[i])<-(t2[i]-1)*log(p2)+ (t1[i]-1)*(log(p1)+log(p12))+log(1-p2)+log(1-p1*p12)
        log(P11[i])<-delta1[i]*(1-delta2[i])*(1-delta3[i])*log(A1[i])+
        delta3[i]*(1-delta1[i])*(1-delta2[i])*log(A2[i])+ delta2[i]*(1-delta1[i])*(1-
delta3[i])*log(A3[i])
        log(P10[i])<- (t1[i]-1)*log(p1)+ t2[i]*log(p2)+log(pow(p12,z1[i])-p1*pow(p12,z2[i]))
        log(P01[i])<- t1[i]*log(p1)+(t2[i]-1)*log(p2)+log(pow(p12,z3[i])-p2*pow(p12,z2[i]))
        log(P00[i])<- t1[i]*log(p1)+t2[i]*log(p2)+z2[i]*log(p12)
        log(L[i])<- v1[i]*v2[i]*log(P11[i])+v1[i]*(1-v2[i])*log(P10[i])+(1-
v1[i])*v2[i]*log(P01[i])+(1-v1[i])*(1-v2[i])*log(P00[i])
    }
p1~ dunif(0,1)
p2~ dunif(0,1)
p12~ dunif(0,1)
mean1<-1/(1-p1*p12)
mean2<-1/(1-p2*p12)
}
```

B12. Distribuição geométrica bivariada Basu-Dhar com a presença de covariáveis

```
model {
for (i in 1:N) {
zeros[i] <- 0
phi[i] <- -log(L[i])
zeros[i] ~ dpois(phi[i])
z2[i]<-max(t1[i],t2[i])
logit(p1[i]) <-
beta10+beta11*idade[i]+beta12*herceptin[i]+beta13*estágio[i]+beta14*tipo.cirurgia[i]+beta15*pCR[i]+bet
a16*estrogênio[i]+beta17*progesterona[i]
logit(p2[i]) <-
beta20+beta21*idade[i]+beta22*herceptin[i]+beta23*estágio[i]+beta24*tipo.cirurgia[i]+beta25*pCR[i]+bet
a26*estrogênio[i]+beta27*progesterona[i]
logit(p12[i]) <-
beta30+beta31*idade[i]+beta32*herceptin[i]+beta33*estágio[i]+beta34*tipo.cirurgia[i]+beta35*pCR[i]+bet
a36*estrogênio[i]+beta37*progesterona[i]
log(A1[i])<-(t1[i]-1)*log(p1[i])+ (t2[i]-1)*(log(p2[i])+log(p12[i]))+log(1-p1[i])+log(1-p2[i]*p12[i])
log(A2[i])<-(t1[i]-1)*(log(p1[i])+log(p2[i])+log(p12[i]))+ log(1-p1[i]*p2[i]-
p2[i]*p12[i]+p1[i]*p2[i]*p12[i])
log(A3[i])<-(t2[i]-1)*log(p2[i])+ (t1[i]-1)*(log(p1[i])+log(p12[i]))+log(1-p2[i])+log(1-p1[i]*p12[i])
log(P11[i])<-delta1[i]*(1-delta2[i])*(1-delta3[i])*log(A1[i])+delta3[i]*(1-delta1[i])*(1-
delta2[i])*log(A2[i])+ delta2[i]*(1-delta1[i])*(1-delta3[i])*log(A3[i])
log(P10[i])<- ((t1[i]-1)*log(p1[i])+ t2[i]*log(p2[i])+ t2[i]*log(p12[i])+ log(1-p1[i]))* delta1[i]+((t1[i]-
1)*log(p1[i])+ t1[i]*log(p2[i])+ t1[i]*log(p12[i])+ log(1-p1[i]))*delta3[i]+ ((t2[i])*log(p1[i])+(t1[i]-
1)*log(p1[i])+(t1[i]-1)*log(p12[i])+ log(1-p1[i]*p12[i]))*delta2[i]
```

```
log(P01[i])<- ((t1[i])*log(p1[i])+ (t2[i]-1)*log(p2[i])+ (t2[i]-1)*log(p12[i])+ log(1-
p2[i]*p12[i]))*delta1[i]+((t2[i]-1)*log(p2[i])+ t1[i]*log(p1[i])+ t1[i]*log(p12[i])+ log(1-
p2[i]))*delta3[i]+((t2[i]-1)*log(p2[i])+(t1[i])*log(p1[i])+(t1[i])*log(p12[i])+log(1-p2[i]))*delta2[i]
log(P00[i])<- t1[i]*log(p1[i])+t2[i]*log(p2[i])+z2[i]*log(p12[i])
log(L[i])<- v1[i]*v2[i]*log(P11[i])+v1[i]*(1-v2[i])*log(P10[i])+(1-v1[i])*v2[i]*log(P01[i])+(1-v1[i])*(1-
v2[i])*log(P00[i])
mean1[i]<-(1/(1-p1[i]*p12[i]))
mean2[i]<-(1/(1-p2[i]*p12[i]))
}
beta10~ dnorm(0,1)
beta11~ dnorm(0,1)
beta12~ dnorm(0,1)
beta13~ dnorm(0,1)
beta14~ dnorm(0,1)
beta15~ dnorm(0,1)
beta16~ dnorm(0,1)
beta17~ dnorm(0,1)
beta20~ dnorm(0,1)
beta21~ dnorm(0,1)
beta22~ dnorm(0,1)
beta23~ dnorm(0,1)
beta24~ dnorm(0,1)
beta25~ dnorm(0,1)
beta26~ dnorm(0,1)
beta27~ dnorm(0,1)
beta30~ dnorm(0,1)
beta31~ dnorm(0,1)
beta32~ dnorm(0,1)
beta33~ dnorm(0,1)
beta34~ dnorm(0,1)
beta35~ dnorm(0,1)
beta36~ dnorm(0,1)
beta37~ dnorm(0,1)
}
```