

Calisto Manuel Máquina

*Análise de Dados de Sobrevida  
Espacialmente Correlacionados*

Belo Horizonte - MG

Fevereiro/ 2014

Calisto Manuel Máquina

*Análise de Dados de Sobrevivência  
Espacialmente Correlacionados*

Dissertação apresentada ao Programa de Pós-Graduação em Estatística como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Fábio Nogueira Demarqui  
Coorientador: Prof. Dr. Marcos Oliveira Prates

MESTRADO EM ESTATÍSTICA  
DEPARTAMENTO DE ESTATÍSTICA  
INSTITUTO DE CIÊNCIAS EXATAS  
UNIVERSIDADE FEDERAL MINAS GERAIS

Belo Horizonte – MG

Fevereiro / 2014

## *Agradecimentos*

Em primeiro lugar, agradeço a Deus, por tudo que tem me proporcionado e mais esta conquista.

À família Máquina, especialmente ao meu pai (Manuel do Rosário Máquina), minha mãe (Maria de Lurdes Muaviaca) e aos meus irmãos (Cristina), (Florêncio), (Ana Maria), (Angélica) e (Orlando).

Ao Colegiado de Pós-Graduação em Estatística da UFMG, por ter me aceite neste excelente programa de Pós-graduação.

Ao meu orientador Prof. Dr. Fábio Nogueira Demarqui e ao meu co-orientador Prof. Dr. Marcos Oliveira Prates, pelo apoio e pela paciência que proporcionaram, que para mim foi uma honra tê-los como orientadores durante a minha formação neste curso.

Um agradecimento à banca composta pelo Prof. Dr. Marcelo Azevedo Costa e pelo Prof. Dr. Vinícius Diniz Mayrink pelas críticas e comentários que ajudaram a melhorar o meu trabalho.

Aos Professores do Departamento, em especial ao Prof. Dr. Roger William Câmara Silva e a Profa. Dra. Rosângela Helena Loschi, que sem dúvidas ajudaram-me bastante durante estes dois anos.

Às Secretárias Rogéria, Rose e a Kate, por todo apoio que demonstraram em todos os momentos da minha formação e a Maísa pelo café, que sem dúvida, serviu de força do meu dia-a-dia dos meus estudos.

À CNPq por todo apoio financeiro prestado durante à minha formação neste curso de mestrado.

Sem me esquecer aos colegas Felipe, Renan, Raquel e William, que compartilharam os seus saberes com os meus.

*“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito.  
Não sou o que deveria ser, mas Graças à Deus, não sou o que era antes” (Marthin  
Luther King).*

# *Sumário*

<b>Lista de Figuras</b>	p. v
<b>Lista de Tabelas</b>	p. vii
<b>Resumo</b>	p. viii
<b>Abstract</b>	p. ix
<b>1 Introdução</b>	p. 1
<b>2 Conceitos básicos em análise de sobrevivência</b>	p. 4
2.1 Funções de Interesse em Análise de Sobrevivência . . . . .	p. 7
2.2 Modelo Weibull . . . . .	p. 9
2.3 Função de Verossimilhança . . . . .	p. 10
2.4 Modelo de Riscos Proporcionais . . . . .	p. 11
<b>3 Inferência Bayesiana</b>	p. 13
3.1 Estimação Pontual . . . . .	p. 14
3.2 Estimação por Intervalo . . . . .	p. 15
3.3 Amostrador de Gibbs . . . . .	p. 16
3.3.1 Seleção de Modelos . . . . .	p. 17
<b>4 Modelos de Fragilidade</b>	p. 19
4.1 Modelo de Fragilidade Gama . . . . .	p. 22
4.2 Modelo de Fragilidade Log-normal . . . . .	p. 22

4.3	Fragilidade espacial . . . . .	p. 23
<b>5</b>	<b>Aplicações</b>	p. 25
5.1	Aplicação a dados de Câncer . . . . .	p. 26
5.2	Aplicação a dados de crimes . . . . .	p. 32
<b>6</b>	<b>Conclusões e trabalhos futuros</b>	p. 41
	<b>Apêndice</b>	p. 43
	<b>Referências Bibliográficas</b>	p. 52

## *Lista de Figuras*

2.1	Ilustração de mecanismos de censura, em que "●" representa falha e "○" representa a censura. Fonte: Giolo e Colosimo (2006) . . . . .	p.6
2.2	Exemplos de funções risco associadas a distribuição Weibull . . . . .	p.10
5.1	<i>Boxplots</i> da distribuição <i>a posteriori</i> de $\alpha$ para dados de câncer . . . . .	p.28
5.2	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>idade</i> para dados de câncer . . . . .	p.29
5.3	<i>Boxplot</i> das distribuições <i>a posteriori</i> dos $\beta$ 's os para dados de câncer . . . . .	p.30
5.4	<i>Boxplot</i> das distribuições <i>a posteriori</i> dos $\alpha$ 's para dados de câncer . . . . .	p.30
5.6	<i>Boxplots</i> da distribuição <i>a posteriori</i> de $\alpha$ para dados de crimes . . . . .	p.33
5.7	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>sexo</i> para dados de crimes . . . . .	p.35
5.9	<i>Boxplot</i> das distribuições <i>a posteriori</i> dos $\alpha$ 's para dados de crimes . . . . .	p.36
5.5	Mapas das distribuições de fragilidades para dados de câncer . . . . .	p.38
5.8	<i>Boxplot</i> dos $\beta$ 's da Distribuição <i>a posteriori</i> para dados de crimes . . . . .	p.39
5.10	Mapas das distribuições de fragilidades para dados de crimes . . . . .	p.40
6.1	<i>Boxplots</i> da distribuição <i>a posteriori</i> do <i>intercepto</i> para dados de câncer . . . . .	p.44
6.2	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>sexo</i> para dados de câncer . . . . .	p.45
6.3	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>raça</i> para dados de câncer . . . . .	p.46
6.4	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>nº de tumores primários</i> para dados de câncer . . . . .	p.47
6.5	<i>Boxplots</i> da distribuição <i>a posteriori</i> do <i>intercepto</i> para dados de crimes . . . . .	p.48

6.6	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>idade</i> para dados de crimes . . . . .	p. 49
6.7	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>lugar</i> para dados de crimes . . . . .	p. 50
6.8	<i>Boxplots</i> da distribuição <i>a posteriori</i> do coeficiente de regressão associado à covariável <i>consumado</i> para dados de crimes . . . . .	p. 51



## *Lista de Tabelas*

5.1	Análise de sensibilidade das distribuições <i>a priori</i> dos hiperparâmetros $\eta$ , $\tau$ e $\phi$ - Aplicação aos dados de câncer-Valores de DIC. . . . .	p. 27
5.2	Estimativas das distribuições <i>a posteriori</i> dos parâmetros associados ao modelo M4 para dados de câncer . . . . .	p. 31
5.3	Análise de sensibilidade das distribuições <i>a priori</i> dos hiperparâmetros $\eta$ , $\tau$ e $\phi$ - Aplicação aos dados de crimes-Valores de DIC . . . . .	p. 32
5.4	Estimativas das distribuições <i>a posteriori</i> dos parâmetros associados ao modelo M4 para dados de crimes . . . . .	p. 37
6.1	Testes de convergência para as cadeias do parâmetro do Modelo M4 para dados de câncer . . . . .	p. 43
6.2	Testes de convergência para as cadeias do parâmetro do Modelo M4 para dados de crimes . . . . .	p. 43

## *Resumo*

Nesta dissertação são apresentadas as análises de dois conjuntos de dados reais que dispõem de informação espacial. O primeiro banco de dados corresponde a uma aplicação da área médica em que o objetivo é estudar o tempo de vida de pacientes com câncer de nasofaringe no estado de Connecticut, EUA, e o segundo banco de dados é constituído por uma amostra de indivíduos que vivem em municípios pertencentes à região metropolitana de Belo Horizonte, e que foram vítimas de tentativa de homicídio. Considerando uma abordagem Bayesiana e assumindo que os tempos de falha seguem distribuição Weibull, os modelos de fragilidade gama e lognormal, que não levam em conta a estrutura de dependência espacial presente nos dados, e o modelo CAR intrínseco, que acomoda tal estrutura de dependência, foram ajustados para ambos os bancos de dados. A fim de se avaliar o efeito de diferentes especificações a priori sobre os modelos ajustados uma análise de sensibilidade foi conduzida. Os modelos que apresentaram o melhor ajuste foram então comparados, e mapas com as fragilidades estimadas para cada área foram construídos. Com base nos mapas obtidos foi possível observar a existência de um padrão do risco da ocorrência do evento de interesse nas áreas estudadas, indicando a presença de estrutura de dependência espacial nos dois bancos de dados analisados.

**Palavras-chave:** *Dados de sobrevivência, modelo de riscos proporcionais, modelos de fragilidade.*

## *Abstract*

In this work the analyses of two real data sets with spatial information available are presented. The first data set corresponds to an application of the medical field in which the objective is to study the life of patients with nasopharyngeal cancer in Connecticut, USA, and the second data set comprises a sample of individuals living in municipalities belonging to the metropolitan region of Belo Horizonte, who were victims of attempted murder. Considering a Bayesian approach and assuming that the failure times follow Weibull distribution, the gamma and lognormal frailty models, which do not take into account the spatial dependence structure present in the data, and the intrinsic CAR model, which accommodates such dependence structure, were adjusted for both data sets. In order to evaluate the effect of different prior specifications on the fitted models a sensitivity analysis was conducted. The best fitted models were then compared, and maps displaying the estimated frailties for each area were built. Based on such maps it was possible to observe the existence of a risk pattern of the occurrence of the event interest over the areas under study, indicating the presence of spatial dependence structure in both databases analyzed.

**Keywords:** *Survival data, proportional hazard models, frailty models.*

# 1 *Introdução*

A análise de sobrevivência é uma das áreas da estatística que vem crescendo a partir da década de 80 sob forte influência da evolução dos computadores e das técnicas estatísticas para analisar dados de sobrevivência. Um fato importante é que esta técnica é aplicada principalmente em situações na qual os dados são censurados, por exemplo, quando alguns indivíduos envolvidos no estudo não "experimentam" o evento de interesse, ou seja, não "falham" até o fim do estudo (GIOLO; COLOSIMO, 2006). Sem a presença da censura nos dados, continuam válidas outras técnicas estatísticas além da análise de sobrevivência, como análise de regressão e planejamento de experimentos.

Segundo Banerjee *et al.* (2003), o uso de modelos de sobrevivência para a modelagem de dados que envolvem efeitos aleatórios (ou fragilidades, como usualmente são referidos estes efeitos em análise de sobrevivência), estão se tornando cada vez mais comuns. O termo *fragilidade* foi introduzido por Vaupel *et al.* (1979) em um contexto envolvendo dados de mortalidade e Lancaster (1979) incorporou este termo na modelagem do tempo de duração de desemprego.

Apesar de introduzido em 1979, Clayton (1978) já havia aplicado o termo fragilidade (mas sem levar em conta o conceito de fragilidade) em um estudo sobre análise multivariada de dados de sobrevivência. Mais tarde Clayton e Cuzick (1985) incorporou também o termo de fragilidade em contexto de modelagem da correlação entre os grupos. Whitmore e Lee (1991) baseando na definição dos modelos de fragilidade, examinaram a distribuição *gama inversa* sobre o termo de fragilidade. Por sua vez Banerjee *et al.* (2003) utiliza o modelo condicional autorregressivo (CAR), para incorporar em sua modelagem um termo de fragilidade espacial nas suas análises sobre os dados de mortalidade infantil no estado de Minnesota.

Existem diversas razões para incluirmos um ou mais termos de fragilidade na modelagem de dados de sobrevivência. Por exemplo, podemos incluir o termo de fragilidade para capturar a heterogeneidade não observada associada aos indivíduos/elementos que

são objeto de estudo. De uma maneira geral, podemos dizer que há sempre um grande número de covariáveis que, caso pudessem ser medidas, dariam informação suficiente para mostrar as diferenças entre indivíduos. Neste contexto, o termo fragilidade pode ser utilizado para representar a ausência daquelas covariáveis que não foram incluídas no modelo porque não dispomos de informação acerca dos seus valores individuais, ou simplesmente pelo nosso desconhecimento da sua existência. Modelos de fragilidade também são bastante utilizados em situações nas quais os dados de sobrevivência são observados em grupos ou (conglomerados) de indivíduos/elementos. Nestes casos, usualmente termos de fragilidades são adicionados à modelagem para representar os diferentes grupos, que por sua vez são assumidos serem independentes. Tais modelos são conhecidos na literatura de análise de sobrevivência como modelos de fragilidade compartilhada, uma vez que todos os indivíduos/elementos pertencentes ao mesmo grupo compartilham o mesmo termo de fragilidade.

Neste trabalho consideraremos a situação em que dados de sobrevivência são observados em grupos que são espacialmente organizados. Especificamente, nossos grupos representam regiões geográficas, como, por exemplo, cidades ou condados de um determinado estado, etc. Assim, é razoável suspeitarmos que termos de fragilidade associados a áreas ou regiões que são próximas uma das outras tendem a ser similares em magnitude. Em outras palavras, estamos interessados em modelar a possível estrutura (não observada) de dependência espacial presente nos dados de sobrevivência. Modelos de sobrevivência que incluem informação espacial tem se tornado cada vez mais comuns. Em particular, dados de área tem sido bastante utilizados para se modelar de dados de sobrevivência com informação espacial disponível, e a modelagem deste tipo de dados é, tradicionalmente, feita através do modelo CAR proposto por Besag (1974). Diversos estudos aplicam o modelo CAR, como é observado nos trabalhos de Banerjee e Carlin (2004), Cooner *et al.* (2006), Bastos e Gamerman (2006), Diva *et al.* (2007, 2008), Pan *et al.* (2013), entre outros. Neste trabalho de dissertação, consideraremos dados de área que serão modelados pelo modelo CAR.

Nesta dissertação são apresentadas as análises detalhadas de dois conjuntos de dados reais. O primeiro banco de dados corresponde a uma aplicação da área médica em que o objetivo é estudar o tempo de vida de pacientes com câncer de nasofaringe no estado de Connecticut, EUA. Para este banco de dados serão consideradas para a análise algumas variáveis explicativas, tais como *sexo*, *idade*, *raça*, *número de tumores primários* e o *condado* em que vive cada paciente. O segundo banco de dados é constituído por uma amostra de indivíduos que vivem em municípios pertencentes à região metropolitana

de Belo Horizonte (RMBH), e que foram vítimas de tentativa de homicídio. A variável resposta neste caso é o tempo até a ocorrência da segunda tentativa de homicídio, e o nosso interesse será estudar se existe algum tipo de padrão espacial presente nos dados.

Este trabalho apresentará a seguinte estrutura de organização: no Capítulo 2 serão apresentados os conceitos básicos em análise de sobrevivência, algumas funções de interesse, o modelo *Weibull* e de riscos proporcionais. No Capítulo 3 serão apresentados os procedimentos Bayesianos utilizados para a gerar as amostras *a posteriori*. Os diferentes modelos com fragilidade serão descritos no Capítulo 4 e a aplicação com base nos dados reais mencionados anteriormente será feita no Capítulo 5. Capítulo 6, apresenta as conclusões e os trabalhos futuros.

## *2 Conceitos básicos em análise de sobrevivência*

A análise de sobrevivência consiste em um conjunto de técnicas estatísticas para análise de dados referentes a medições de tempo a partir de um instante inicial até a ocorrência de um determinado evento de interesse numa população. No entanto, nem sempre é possível observar o evento de interesse para todos os indivíduos pertencentes ao estudo, devido, em geral, a certas restrições no processo de coleta de dados, situação em que surgem as designadas observações censuradas. De fato, nos estudos envolvendo dados de sobrevivência, o tempo de vida de alguns indivíduos pode não ser conhecido com exatidão, havendo apenas informação incompleta. Segundo Giolo e Colosimo (2006), observações censuradas podem ocorrer por várias razões, como é o caso, por exemplo, da perda de acompanhamento de pacientes no decorrer do estudo, ou devido a não ocorrência do evento de interesse até o término do experimento.

Observações censuradas se caracterizam em dois tipos de mecanismos de censura, nomeadamente: mecanismo de censura não informativo e mecanismo de censura informativo. O mecanismo de censura não informativo é aquele no qual o conhecimento do tempo de censura de um indivíduo não fornece mais informação sobre a sobrevivência futura do indivíduo além da que, teríamos se ele tivesse continuado no estudo. Na realidade, trata-se de uma hipótese de independência entre os mecanismos geradores de falha e de censura, e garante que indivíduos cujo tempo de vida foi censurado num determinado instante, tenham o mesmo risco de sofrer uma falha futura em comparação com aqueles que ainda não "experimentaram" o evento de interesse nesse instante. Em outras palavras, um mecanismo de censura diz-se não informativo se a ocorrência dessa censura é independente do mecanismo que provoca a falha. Caso contrário, dizemos que mecanismo de censura é informativo.

Em análise de sobrevivência as censuras podem ser classificadas em três tipos:

1. Censura à direita: ocorre quando o tempo até a ocorrência do evento de interesse

está à direita do tempo observado;

2. Censura à esquerda: ocorre quando o tempo até a ocorrência do evento de interesse está à esquerda do tempo observado;
3. Censura intervalar: ocorre quando é conhecido apenas o intervalo de tempo no qual o evento de interesse ocorreu, sendo o tempo exato de ocorrência desconhecido.

A censura à direita é a mais frequente na prática. As circunstâncias nas quais pode ocorrer este tipo de censura são várias. Por exemplo, ao considerarmos um ensaio clínico ou qualquer outro estudo experimental cuja finalidade seja a análise de tempos de vida, é natural ainda existirem indivíduos no fim do estudo para os quais não foi observada a morte. Por outro lado, a perda de acompanhamento de indivíduos em risco, a interrupção do tratamento para alguns doentes devido a fortes efeitos secundários ou a ocorrência da morte por outra causa diferente da causa de interesse, são também situações de censura à direita. A censura à esquerda é o tipo de censura que ocorre com menor frequência na prática. Um exemplo bastante elucidativo de censura à esquerda surge quando se pretende estudar a idade em que uma criança consegue realizar determinada tarefa. Assim, como antes do início do estudo, é possível que algumas crianças já saibam realizar essa tarefa, as observações correspondentes são censuradas à esquerda. Finalmente, a censura intervalar ocorre com bastante frequência em problemas da área médica. Este tipo de censura está associado, em geral, a experimentos nos quais pacientes/elementos são acompanhados periodicamente para o monitoramento da ocorrência do evento de interesse.

Neste trabalho de dissertação será assumido que os tempos de falha estão sujeitos a ocorrência de censura à direita, e que o mecanismo de censura é não informativo. A censura à direita pode ser classificada em:

1. Censura do Tipo I: ocorre quando o estudo é terminado após um período pré-estabelecido de tempo.
2. Censura do Tipo II: ocorre quando o estudo é finalizado após a ocorrência do evento de interesse para um número pré-estabelecido de indivíduos.
3. Censura do Tipo III: também conhecida como censura aleatória, e na prática médica é o que mais ocorre. Ela acontece quando um paciente é retirado do estudo sem ter se verificado o evento de interesse, como por exemplo, se um paciente morre por uma causa diferente da causa em estudo.



A Figura (2.1)(a) ilustra o caso em que todos os indivíduos envolvidos no estudo "experimentam" evento de interesse. As Figuras (2.1)(b),(c) e (d) ilustram os três tipos de censura que podem ocorrer.

Censuras do tipo I e II ocorrem com maior frequência na área de engenharia enquanto a censura do tipo III ou censura aleatória é frequentemente observada na área médica. Ao longo deste trabalho de dissertação, assumiremos o mecanismo de censura aleatório.

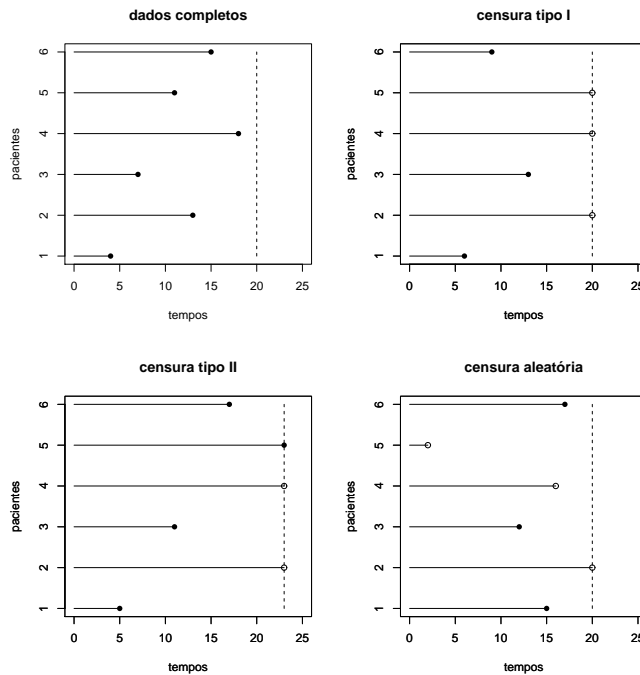


Figura 2.1: Ilustração de mecanismos de censura, em que "●" representa falha e "○" representa a censura. Fonte: Giolo e Colosimo (2006)

De acordo com Giolo e Colosimo (2006) uma representação simples do mecanismo de censura aleatória é feita usando duas variáveis aleatórias  $T$  e  $C$ , em que  $T$  que representa o tempo de falha, isto é, o tempo de ocorrência do evento de interesse, e  $C$  representa o tempo de censura associado ao paciente.

Os dados de sobrevivência são comumente representados pelo par  $(Y_i, \delta_i)$ , em que  $Y_i$  é o tempo de falha ou censura e  $\delta_i$  é uma variável indicadora de falha ou censura, isto é,  $(\delta_i = 1)$  se  $Y_i$  é o tempo de falha e  $(\delta_i = 0)$  se  $Y_i$  é o tempo de censura. Dessa forma,  $Y_i$  é representado por:

$$Y_i = \min\{T_i, C_i\},$$

em que  $T_1, T_2, \dots, T_n$  são variáveis aleatórias independentes e identicamente distribuídas com função de distribuição  $F$  e independentes de  $C_1, C_2, \dots, C_n$  que também são variáveis

aleatórias independentes e identicamente distribuídas com função de distribuição  $G$ .

A variável indicadora  $\delta_i$ ,  $i = 1 \dots, n$ , pode ser representada da seguinte maneira

$$\delta_i = \begin{cases} 1 & \text{se } T_i \leq C_i \\ 0 & \text{se, } T_i > C_i \end{cases},$$

para  $i = 1 \dots, n$ .

## 2.1 Funções de Interesse em Análise de Sobrevivência

De maneira geral, suponha que se pretende dar um acompanhamento à pacientes por um certo período de tempo, após o diagnóstico de uma determinada doença, e que durante esse período registra-se um determinado número de óbitos. Uma forma simples de descrever este cenário é através do cálculo da taxa de mortalidade média. No entanto, em situações em que a taxa de mortalidade não é constante ao longo do tempo, a taxa média perde seu peso para descrever tal cenário, enfatizando a necessidade da análise de sobrevivência. De fato, segundo Carvalho *et al.* (2011), ao empregar técnicas de análise de sobrevivência somos capazes de responder a uma variedade de questões não respondidas pelos métodos descritivos clássicos.

Nesta seção, serão apresentadas funções básicas de análise de sobrevivência. Por meio dessas funções será possível responder à várias situações, nas quais o processo de ocorrência de eventos ao longo do tempo é o foco de interesse.

Seja  $T$  uma variável aleatória contínua não negativa que representa o tempo até ocorrência de um evento de interesse. Em análise de sobrevivência, a distribuição da variável aleatória  $T$ , pode ser especificada através da função de sobrevivência, função da taxa de falha (ou risco), ou da função densidade de probabilidade.

Uma função de grande interesse prático em análise de sobrevivência é a chamada *função de sobrevivência*, denotada por  $S(t)$ , e definida como sendo a probabilidade de um indivíduo sobreviver um tempo superior a  $t$ , ou seja

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t). \end{aligned}$$

A distribuição da variável aleatória  $T$  também pode ser determinada pela *função densidade de probabilidade*, que é definida como sendo o limite da probabilidade do evento de interesse ocorrer em um intervalo de tempo pequeno,  $[t, t + \Delta t)$ , dividida pela amplitude do intervalo,  $\Delta t$ , isto é,

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

A *função risco*,  $h(t)$ , é definida como sendo a probabilidade de falha durante um intervalo de tempo muito pequeno, dado que o evento não ocorreu até o início do intervalo. De maneira simples, assumindo que o indivíduo estava vivo no início do intervalo, qual é a chance dele ter uma falha nesse curto intervalo de tempo. A função risco é dada por

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \quad (2.1)$$

Segundo Lee e Wang (2003), a função risco é também chamada função taxa de falha instantânea. Assim a função risco acumulada é dada por

$$H(t) = \int_0^t h(u) du. \quad (2.2)$$

As três funções apresentadas acima são matematicamente equivalentes, uma vez que sendo conhecida uma delas, as outras duas podem ser obtidas naturalmente por meio de derivação. Por exemplo, a função densidade pode ser obtida derivando-se a função de distribuição sobrevivência, isto é,

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}[1 - S(t)] = -\frac{d}{dt}S(t). \quad (2.3)$$

Substituindo a Equação (2.3) na Equação (2.1) obtemos a função risco:

$$h(t) = \frac{-\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt} \log S(t). \quad (2.4)$$

Integrando a Equação (2.4) no intervalo de  $(0, t]$  e usando o fato de que  $S(0) = 1$ , temos

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u) du\right\}. \quad (2.5)$$

Logo, das Equações (2.2) e (2.5) concluímos :

$$f(t) = h(t) \exp\{-H(t)\}. \quad (2.6)$$

## 2.2 Modelo Weibull

Em análise de sobrevivência, como vimos no começo desta dissertação, a variável resposta é o tempo,  $T$ , até ocorrência de um evento de interesse. Vimos também que esta variável resposta pode ser caracterizada por diversas funções: função de sobrevivência, função de densidade de probabilidade e função risco. Uma questão importante encontrada dentro do campo de análise de sobrevivência é a escolha da distribuição da variável aleatória  $T$ . A informação que se tem, é que esta variável aleatória é contínua e não-negativa, dessa forma a escolha da distribuição normal sobre  $T$  é descartada, uma vez que essa distribuição assume valores negativos.

Carvalho *et al.* (2011) discutem diversas distribuições de probabilidades adotadas sobre a variável aleatória  $T$ . Fazem parte deste grupo as ditribuições *Weibull*, *log-normal* e *gama*. Segundo Carvalho *et al.* (2011) é desejável decidir por distribuições de probabilidades que se adaptam a uma grande variedade de situações, como é o caso da distribuição *Weibull*.

Seja  $T$  uma variável aleatória com função de distribuição *Weibull*, isto é,  $T \sim Weibull(\alpha, \gamma)$ , então as suas funções de densidade e de sobrevivência são respectivamente:

$$f(t|\boldsymbol{\psi}) = \alpha\gamma \exp\{-\gamma t^\alpha\}$$

e

$$S(t|\boldsymbol{\psi}) = \exp\{-\gamma t^\alpha\}$$

em que  $t \geq 0$ ,  $\boldsymbol{\psi} = (\alpha, \gamma)$ ,  $\alpha > 0$  é o parâmetro de forma e  $\gamma > 0$  é parâmetro de escala.

Com base na Equação (2.4), temos que a função risco é definida da seguinte forma:

$$\begin{aligned} h(t) &= -\frac{d}{dt} \log S(t|\boldsymbol{\psi}) \\ &= -\frac{d}{dt} (-\gamma t^\alpha) \\ &= \alpha\gamma t^{\alpha-1}. \end{aligned} \tag{2.7}$$

O valor do parâmetro  $\alpha$  é de grande interesse prático, uma vez que este determina a forma da função risco, como ilustrado na Figura (2.2). A função risco é decrescente se  $\alpha < 1$ , crescente se  $\alpha > 1$  e constante para  $\alpha = 1$ .

Na Figura (2.2) observamos cinco cenários de funções risco associados à distribuição *Weibull*. Notamos que, a cada valor do parâmetro de forma, a curva da função risco se

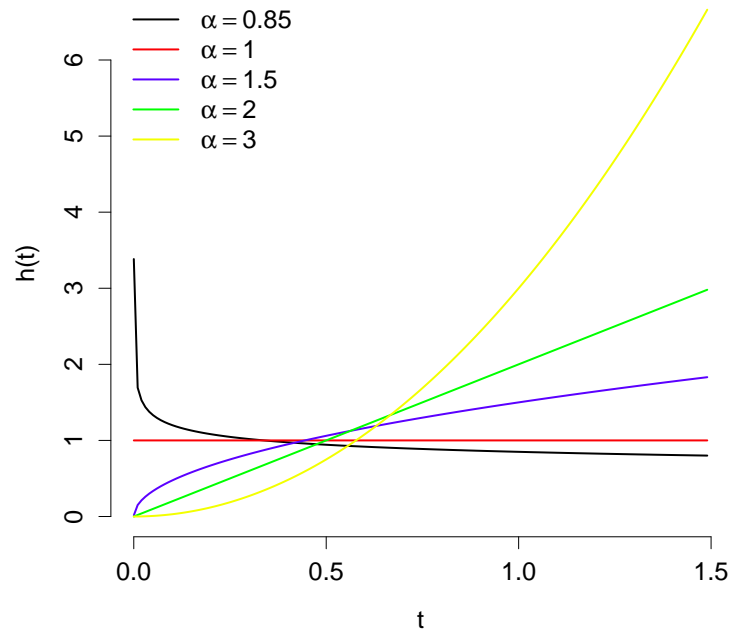


Figura 2.2: Exemplos de funções risco associadas a distribuição Weibull

comporta de maneira diferente. Por exemplo, em situações que o valor do parâmetro de forma é menor que um ( $\alpha = 0.85$ ), o risco é maior no instante inicial e esse risco vai diminuindo à medida que o tempo vai passando. Em situações em que o valor do parâmetro de forma é maior que um, como por exemplo  $\alpha = 3$ , notamos um crescimento rápido da curva da função risco.

## 2.3 Função de Verossimilhança

Vimos no início deste capítulo que a distinção entre falha e censura é geralmente feita com base na variável indicadora e que o tipo de mecanismo de censura à direita mais comum, assumido nesta dissertação, é a censura aleatória. Nesta seção, veremos como construir a função de verossimilhança para dados de censura à direita.

Sejam  $T$  e  $C$  variáveis aleatórias independentes representando, respectivamente, os tempos de falha e censura. Assuma que  $g(c)$  e  $G(c)$  são funções de densidade e de sobrevivência, respectivamente, da variável aleatória  $C$ . Então, para o  $i$ -ésimo indivíduo temos que

- Se for observada uma censura

$$p(Y_i = t, \delta_i = 0) = P(C_i = t, T_i > C_i) = P(C_i = t, T_i > t) = g(t|\boldsymbol{\lambda})S(t; \boldsymbol{\psi});$$

- Se for observada uma falha

$$p(Y_i = t, \delta_i = 1) = P(T_i = t, T_i \leq C_i) = P(T_i = t, C_i \geq t) = f(t|\boldsymbol{\psi})G(t|\boldsymbol{\lambda}).$$

A combinação dos dois cenários resulta a seguinte verossimilhança:

$$L(\boldsymbol{\psi}|D) = \prod_{i=1}^r f(t_i|\boldsymbol{\psi})G(t_i|\boldsymbol{\lambda}) \prod_{i=r+1}^n g(t_i|\boldsymbol{\lambda})S(t_i|\boldsymbol{\psi}), D = \{(t_i, \delta_i) : i = 1, \dots, n\}$$

em que  $\boldsymbol{\psi}$  e  $\boldsymbol{\lambda}$  são vetores de parâmetros associados à  $f()$  e  $G()$ , respectivamente,  $r$  representa o número de observações não-censuradas  $n - r$  corresponde o número de observações censuradas.

Baseando-se na suposição de que o mecanismo de censura é não-informativo, isto é, a função de distribuição  $G$  não envolve o vetor de parâmetros  $\boldsymbol{\psi}$ , isto é, como  $\boldsymbol{\psi} \neq \boldsymbol{\lambda}$ , a função de verossimilhança toma a seguinte forma:

$$L(\boldsymbol{\psi}) \propto \prod_{i=1}^r f(t_i|\boldsymbol{\psi}) \prod_{i=r+1}^n S(t_i|\boldsymbol{\psi}). \quad (2.8)$$

Aplicando a variável indicadora  $\delta_i$ ,  $i = 1, \dots, n$  na Equação (2.8), temos

$$L(\boldsymbol{\psi}) \propto \prod_{i=1}^n [f(t_i|\boldsymbol{\psi})]^{\delta_i} [S(t_i|\boldsymbol{\psi})]^{1-\delta_i}. \quad (2.9)$$

Substituindo na Equação (2.9) a relação (2.6) segue que

$$L(\boldsymbol{\psi}|D) \propto \prod_{i=1}^n [h(t_i|\boldsymbol{\psi})]^{\delta_i} S(t_i|\boldsymbol{\psi}) \quad (2.10)$$

## 2.4 Modelo de Riscos Proporcionais

Em análise de sobrevivência, variáveis explicativas (ou covariáveis) são usualmente introduzidas na modelagem através da função risco. Neste contexto, o modelo de riscos proporcionais, proposto por Cox (1972), surge como o modelo de regressão mais popular encontrado na literatura.

O modelo de Cox é dado por

$$h(t_i|\mathbf{x}) = h_0(t_i) \exp\{\mathbf{x}_i\boldsymbol{\beta}\}, i = 1, \dots, n, \quad (2.11)$$

em que  $h_0(\cdot)$  é chamada função risco base, que é afetada multiplicativamente pelo termo exponencial que envolve as covariáveis, e pode ser modelada paramétrica ou não-parametricamente,  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$  é o vetor de covariáveis e  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  é o vetor dos coeficientes de regressão.

Devido ao fato da função risco base  $h_0(\cdot)$  poder ser modelada paramétrica ou não-parametricamente, o modelo de riscos proporcionais é frequentemente referido na literatura como um modelo semiparamétrico. Além disso, o termo "riscos proporcionais" deve-se ao fato de que a razão entre as funções risco associadas a quaisquer dois elementos é constante com relação ao tempo, isto é, dados  $i$  e  $j$  dois indivíduos diferentes

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} = \frac{h_0(t) \exp\{\mathbf{x}_i\boldsymbol{\beta}\}}{h_0(t) \exp\{\mathbf{x}_j\boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i\boldsymbol{\beta} - \mathbf{x}_j\boldsymbol{\beta}\}.$$

Quando a função risco base é modelada parametricamente, como o caso considerado nesta dissertação, a função de verossimilhança é obtida de maneira similar àquela apresentada na Seção 2.3. Logo, segue que

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}|D) = \prod_{i=1}^n [h_0(t_i|\boldsymbol{\psi}) \exp(\mathbf{x}_i\boldsymbol{\beta})]^{\delta_i} \exp\{-H_0(t_i|\boldsymbol{\psi}) \exp(\mathbf{x}_i\boldsymbol{\beta})\},$$

em que  $D = \{(t_i, \delta_i, \mathbf{x}_i) : i = 1, \dots, n\}$ .

### 3 *Inferência Bayesiana*

Em inferência estatística, é fundamental medir a informação sobre a quantidade de interesse  $\theta$  com base nos dados e conhecimentos *a priori* sobre  $\theta$ . Um dos procedimentos utilizados para a solução deste problema é o uso da abordagem Bayesiana, que consiste em caracterizar a informação dos dados combinada com conhecimento *a priori* sobre a quantidade  $\theta$ . Esta incerteza, pode assumir diferentes graus, que por sua vez, são representados através de modelos probabilísticos sobre  $\theta$  assumindo valores no espaço paramétrico  $\Theta$ .

Seja  $H$  a informação inicial sobre um parâmetro  $\theta$ ,  $\theta \in \Theta$ . Considerando que esta informação pode ser expressa de forma probabilística,  $p(\theta|H)$ . Segundo Migon e Gamerman (1999) normalmente a informação em  $H$  não é suficiente, sendo necessário métodos de experimentação, que consistem na observação das quantidades aleatórias, denotadas por  $\mathbf{Y}$ , que caracterizam a população em estudo. Antes de se observar a quantidade  $\mathbf{Y}$ , é sempre importante avaliar a distribuição amostral dessa quantidade, representada por  $p(\mathbf{Y}|\theta, H)$ , vale ressaltar que estas quantidades aleatórias dependem do parâmetro  $\theta$ .

Seja  $\mathbf{y} = (y_1, \dots, y_n)$ , uma realização do vetor aleatório  $\mathbf{Y}$ . Então a quantidade de informação disponível sobre o parâmetro  $\theta$  é aumentada de  $H$  para  $H^* = H \cap \{\mathbf{Y} = \mathbf{y}\}$ , em que,  $H^* \subset H$ . Observemos agora que toda informação disponível sobre o parâmetro  $\theta$  é descrita por  $p(\theta|\mathbf{y}, H)$ , por meio da passagem de  $p(\theta|H)$  para  $p(\theta|\mathbf{y}, H)$ , através do Teorema de Bayes.

**Teorema 1** *Seja  $\theta$  uma quantidade de interesse e desconhecida. A informação que dispomos sobre  $\theta$ , que é resumida probabilisticamente por meio de  $p(\theta|H)$ , pode ser aumentada observando-se uma quantidade aleatória  $\mathbf{Y}$  associada à  $\theta$ . Seja  $p(\mathbf{y}|\theta, H)$  a distribuição amostral que define esta relação. A distribuição atualizada de  $\theta$  é obtida por*

$$p(\theta|\mathbf{y}, H) = \frac{p(\theta, \mathbf{y}|H)}{p(\mathbf{y}|H)} = \frac{p(\mathbf{y}|\theta, H)p(\theta|H)}{p(\mathbf{y}|H)}$$



em que

$$p(\mathbf{y}|H) = \int_{\Theta} p(\mathbf{y}, \theta|H) d\theta.$$

Este resultado é conhecido como Teorema de Bayes, em homenagem ao seu idealizador, Reverendo Thomas Bayes. Note que o termo  $p(\mathbf{y}|H)$  não depende de  $\theta$ , isto é, o termo  $1/p(\mathbf{y}|H)$  é uma constante normalizadora. Note também que o fator  $H$  é comum a todos os termos, e pode ser omitido para facilitar a notação. Desta forma, o teorema de Bayes pode ser reescrito da seguinte maneira

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$$

em que  $p(\mathbf{y}|\theta)$  é a função que relaciona a observação  $\mathbf{y}$  e o parâmetro desconhecido  $\theta$ . A cada valor fixo de  $\mathbf{y}$ , esta relação fornece a plausibilidade ou verossimilhança, para cada um dos possíveis valores de  $\theta \in \Theta$  e  $p(\theta)$  é chamada distribuição *a priori*.

A distribuição *a posteriori* de  $\theta$ ,  $p(\theta|\mathbf{y})$ , contém toda informação de que dispomos para fazermos inferências sobre  $\theta$ . Nas seções 3.1 e 3.2 será descrito brevemente como é feita a estimação pontual e por intervalo respectivamente, sob o paradigma Bayesiano.

### 3.1 Estimação Pontual

Suponha, sem perda de generalidade, que  $\theta$  é um escalar. Então, dada a distribuição *a posteriori*  $p(\theta|\mathbf{y})$ , uma estimativa pontual para  $\theta$  pode ser obtida a partir de alguma medida de tendência central. Três escolhas são: a média *a posteriori*

$$E[\theta|\mathbf{y}] = \int_{\theta \in \Theta} \theta p(\theta|\mathbf{y}) d\theta,$$

a mediana *a posteriori*

$$\hat{\theta} : \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{y}) d\theta = 0.5,$$

e a moda *a posteriori*

$$\hat{\theta} : p(\hat{\theta}|\mathbf{y}) = \sup_{\theta} p(\theta|\mathbf{y}).$$

É importante ressaltar que, se a distribuição *a posteriori* for unimodal e simétrica, como é o caso da distribuição normal, a média, a mediana e a moda serão iguais. No

entanto, no caso de distribuição multimodal, a moda será a pior medida de tendência central a escolher. Por outro lado, a média *a posteriori* pode sofrer influência dos valores extremos, tornando-se assim uma medida não robusta, conluo que devo optar pela moda. No entanto apesar dos problemas com *outliers*, a média é bastante utilizada na literatura.

Analogamente, como medida de variabilidade podemos tomar a variância *a posteriori*,

$$\begin{aligned} \text{Var}[\theta|\mathbf{y}] &= E[\theta^2|\mathbf{y}] - (E[\theta|\mathbf{y}])^2 \\ &= \int_{\theta \in \Theta} \theta^2 p(\theta|\mathbf{y}) d\theta - \left( \int_{\theta \in \Theta} \theta p(\theta|\mathbf{y}) d\theta \right)^2, \end{aligned}$$

ou ainda o desvio padrão *a posteriori* definido por

$$\text{Sd}[\theta|\mathbf{y}] = \sqrt{\text{Var}[\theta|\mathbf{y}]}.$$

Uma outra quantidade considerada é o percentil  $\xi$  *a posteriori*, que é uma medida de posição e indica valor abaixo do qual cai um grupo de observações e é dado por

$$Q(\xi) : \int_{-\infty}^{\theta'} p(\theta|\mathbf{y}) d\theta = \xi,$$

em que  $\theta' \in \Theta$  e  $\xi \in (0, 1)$ .

## 3.2 Estimação por Intervalo

Na inferência clássica, um dos procedimentos utilizados para verificar se covariáveis são estatisticamente significativas consiste na avaliação dos intervalos de confiança dos parâmetros associados à cada covariável. Este procedimento é também válido no caso Bayesiano. Na realidade, no contexto de inferência Bayesiana, recorre-se aos chamados intervalos de credibilidade ( $C$ ).

Suponha que se deseja encontrar os percentis  $\xi/2$  e  $1 - \xi/2$  da distribuição *a posteriori* de  $\theta$ , isto é, os pontos  $q_L$  e  $q_U$  tais que

$$\int_{-\infty}^{q_L} p(\theta|\mathbf{y}) d\theta = \xi/2$$

e

$$\int_{q_U}^{\infty} p(\theta|\mathbf{y}) d\theta = 1 - \xi/2.$$

É fácil notar que  $P(q_L < \theta < q_U) = 1 - \xi$ , isto é, a probabilidade de  $\theta$  pertencer ao intervalo  $(q_L, q_U)$  é de  $100(1 - \xi)\%$ . O intervalo apresentado acima é chamado intervalo

de credibilidade e possui uma interpretação probabilística direta.

### 3.3 Amostrador de Gibbs

Na maioria das vezes não é possível obter uma expressão fechada a para distribuição *a posteriori* de  $\theta$ , sendo necessário o uso de métodos numéricos para que a análise Bayesiana possa ser realizada. Existem vários métodos numéricos utilizados para solucionar este problema, dentre os quais destacam-se os métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC), e em particular o amostrador de Gibbs, que será utilizado neste trabalho.

Os métodos MCMC são métodos de simulação estocásticos baseados na construção de cadeias de Markov cuja distribuição estacionária é  $p(\theta|\mathbf{y})$ . Em linhas gerais os métodos MCMC requerem apenas o conhecimento do núcleo de  $p(\theta|\mathbf{y})$ , e fornecem uma maneira de se obter amostras de  $p(\theta|\mathbf{y})$ . Inferências sobre as quantidades de interesse são, então, realizadas com base nas amostras geradas da distribuição *a posteriori*.

O amostrador de Gibbs é um método MCMC bastante popular devido a sua eficiência na solução de problemas multidimensionais complexos. Considere um modelo com  $k$  parâmetros, isto é,  $\theta = (\theta_1, \dots, \theta_k)'$ . Para implementar o amostrador de Gibbs, é necessário que seja possível obter cada uma das distribuições condicionais completas *a posteriori*,  $\{p(\theta_i|\theta_{j \neq i}, \mathbf{y}), i = 1, \dots, k\}$ .

Dado um conjunto arbitrário de valores iniciais  $\{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$ , o algoritmo do amostrador de Gibbs funciona da seguinte maneira:

Para ( $s = 1 : S$ ):

**Passo 1:** Seleciona-se  $\theta_1^{(s)}$  a partir de  $p(\theta_1|\theta_2^{s-1}, \theta_3^{s-1}, \dots, \theta_k^{s-1}, \mathbf{y})$

**Passo 2:** Seleciona-se  $\theta_2^{(s)}$  a partir de  $p(\theta_2|\theta_1^s, \theta_3^{s-1}, \dots, \theta_k^{s-1}, \mathbf{y})$

⋮

**Passo k:** Seleciona-se  $\theta_k^{(s)}$  a partir de  $p(\theta_k|\theta_1^s, \theta_2^s, \dots, \theta_{k-1}^s, \mathbf{y})$

Segundo Geman e Geman (1984), sob certas condições de regularidades que são satisfeitas na maioria dos problemas, pode-se mostrar que, após um número suficientemente grande de iterações, a cadeia de Markov formada pelas distribuições condicionais comple-

tas apresentadas acima converge para a distribuição *a posteriori* de  $\theta$ , obtendo portanto, uma amostra da distribuição a posteriori.

A questão que se coloca é de como saber o número de iterações necessárias para a convergência da cadeia para a distribuição estacionária  $p(\theta|y)$ . Vários procedimentos foram propostos para se testar o problema da convergência cadeia. Uma excelente revisão sobre tais testes de convergência pode ser encontrada em (COWLES; CARLIN, 1996). Nesta dissertação serão utilizados os testes propostos Heidelberger e Welch (1983), Geweke (1992), disponíveis no pacote *coda* do *software R*.

### 3.3.1 Seleção de Modelos

Um critério para comparação de modelos Bayesianos bastante utilizado na literatura é o chamado *Critério de Informação do Desvio* (DIC). Grande parte da popularidade do DIC deve-se ao fato desse método estar implementado em *softwares* como *WinBUGS* e *OpenBUGS*. Nesta dissertação será utilizado o *software OpenBUGS* para o ajuste dos modelos em estudo, então a comparação dos modelos ajustados será feita utilizando-se o DIC.

O DIC foi proposto por Spiegelhalter *et al.* (2002) como uma generalização do *Critério de Informação de Akaike* (AIC). Tal generalização é baseada na distribuição *a posteriori* da estatística desvio,

$$D(\theta) = -2 \log f(\mathbf{y}|\theta) + 2 \log m(\mathbf{y}),$$

em que  $f(\mathbf{y}|\theta)$  é a função de verossimilhança e  $m(\mathbf{y})$  é uma função que depende apenas dos dados observados.

A proposta do DIC é resumir o ajuste do modelo através do valor esperado *a posteriori* da estatística *desvio*,  $\bar{D} = E_{\theta|\mathbf{y}}[D]$ , e da *complexibilidade* do modelo determinada pelo número de parâmetros efetivos, denotado por  $p_D$ . Spiegelhalter *et al.* (2002) mostraram que, em situações envolvendo modelos Gaussianos, é possível mostrar-se que o número de parâmetros efetivos é o valor esperado da estatística *desvio* menos o desvio das esperanças calculadas com base nas distribuições *a posteriori*, isto é,

$$\begin{aligned} p_D &= E_{\theta|\mathbf{y}}[D] - D(E_{\theta|\mathbf{y}}[\theta]) \\ &= \bar{D} - D(\bar{\theta}). \end{aligned}$$

Logo, temos o DIC definido como

$$\begin{aligned} DIC &= \bar{D} + p_D \\ &= 2\bar{D} - D(\bar{\theta}). \end{aligned}$$

A escolha do melhor modelo é feita comparando os valores do DIC de todos os modelos ajustados; quanto menor o valor do DIC, melhor será ajuste do modelo.

## 4 *Modelos de Fragilidade*

Segundo Rocha (2011), é prática comum em análise de sobrevivência assumirmos que indivíduos com características (covariáveis) similares sejam homogêneos. Entretanto, em diversas situações práticas essa suposição pode não ser razoável. Como exemplo desta situação, consideremos um estudo clínico no qual são envolvidos indivíduos com características idênticas, frequentemente observa-se que apesar de estes indivíduos fazerem parte de um grupo homogêneo em relação às covariáveis consideradas, diferem entre si na reação quando submetidos ao mesmo tratamento, ou no modo como cada indivíduo absorve outros fatores.

De uma maneira geral, a heterogeneidade entre indivíduos que apresentam covariáveis similares pode ser vista como decorrência da não inclusão de uma ou mais variáveis explicativas na modelagem de dados. A ausência de tal(is) covariável(is), por sua vez, pode ser atribuída ao desconhecimento da(s) mesma(s), ou ainda, devido a impossibilidade de se observar ou medir tal(is) variável(is). Neste contexto, a necessidade de se incluir na modelagem de dados de sobrevivência fontes de heterogeneidade não observadas levou ao surgimento dos chamados modelos de fragilidade. Introduzido por Vaupel *et al.* (1979), o termo fragilidade é usado para designar as covariáveis não observadas que descrevem os fatores de risco desconhecidos ou que não podem ser medidos.

Modelos de fragilidade também têm sido extensivamente utilizados na modelagem de dados de sobrevivência multivariados que compreendem, por exemplo, tempos de sobrevivência associados à: (i) membros de uma mesma família; (ii) indivíduos sujeitos a múltiplos eventos do mesmo tipo (eventos recorrentes); ou ainda, (iii) grupos de indivíduos provenientes de diferentes áreas ou regiões interligadas entre si. Em qualquer um desses casos, a suposição de independência entre os tempos associados a indivíduos pertencentes a um mesmo grupo (ou recorrência de eventos em um mesmo indivíduo) não é razoável.

No contexto multivariado, a fragilidade é um efeito aleatório (não observável) que é compartilhado por elementos pertencentes a um mesmo grupo ou conglomerado e atua de

maneira multiplicativa na função risco. O modelo de fragilidade mais comum é o chamado modelo de fragilidade compartilhada, que corresponde a uma generalização do modelo de riscos proporcionais, através da introdução do efeito aleatório, e é definido do seguinte modo

$$h(t_{il}|\mathbf{x}_{il}, z_l) = h_0(t_{il})z_l \exp(\mathbf{x}_{il}\boldsymbol{\beta}), \quad l = 1, \dots, L \text{ e } i = 1, \dots, n_l, \quad (4.1)$$

em que  $h_0(t_{il})$  é a função risco de base,  $\mathbf{x}_{il}$  é um vetor  $1 \times p$  de covariáveis associado ao  $i$ -ésimo indivíduo do  $l$ -ésimo conglomerado,  $\boldsymbol{\beta}$  é um vetor  $p \times 1$  de coeficientes de regressão e  $z_l$  é a fragilidade (não observada) associada ao  $l$ -ésimo grupo com distribuição de probabilidade conhecida com média 1 e variância desconhecida  $\eta$ . Valores grandes de  $\eta$  indicam grande heterogeneidade entre os indivíduos ou grupos, enquanto pequenos valores indicam ausência ou pequena heterogeneidade entre os indivíduos ou grupos.

A partir do modelo (4.1), pode-se observar que indivíduos com valores de fragilidades superiores a um tendem a ter maior possibilidade de experimentar o evento de interesse enquanto os indivíduos com valores menores que um são menos propensos à ocorrência do evento de interesse.

O modelo com fragilidade apresentado em (4.1) preserva a estrutura do modelo de riscos proporcionais dado em (2.11). A razão das funções risco de dois elementos, entretanto, pode depender tanto da diferença entre as covariáveis quanto da diferença entre os grupos. Temos os seguintes cenários a considerar:

1. Para dois elementos  $i$  e  $j$  com covariáveis diferentes pertencentes a um mesmo grupo  $l$ , temos

$$R(t) = \frac{h_0(t) z_l \exp\{\mathbf{x}_i \boldsymbol{\beta}\}}{h_0(t) z_l \exp\{\mathbf{x}_j \boldsymbol{\beta}\}} = \exp\{(\mathbf{x}_i - \mathbf{x}_j) \boldsymbol{\beta}\},$$

ou seja, a razão das funções risco entre os elementos  $i$  e  $j$  depende apenas da diferença entre suas respectivas covariáveis.

2. Analogamente, para indivíduos pertencentes a grupos diferentes (digamos  $l$  e  $k$ ) que apresentam um mesmo conjunto de covariáveis  $\mathbf{x}$  temos

$$R(t) = \frac{h_0(t) z_l \exp\{\mathbf{x} \boldsymbol{\beta}\}}{h_0(t) z_k \exp\{\mathbf{x} \boldsymbol{\beta}\}} = \frac{z_l}{z_k},$$

isto é, o risco relativo é função da razão entre as fragilidades associadas aos respectivos grupos dos indivíduos.

3. Finalmente, se forem comparados dois indivíduos com covariáveis diferentes,  $\mathbf{x}_i$  e

$\mathbf{x}_j$ , pertencentes em grupos distintos  $l$  e  $k$ , então

$$R(t) = \frac{h_0(t)z_l \exp\{\mathbf{x}_i\boldsymbol{\beta}\}}{h_0(t)z_k \exp\{\mathbf{x}_j\boldsymbol{\beta}\}} = \frac{z_l}{z_k} \exp\{(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\beta}\},$$

ou seja, o risco relativo depende tanto da diferença entre as covariáveis quanto da diferença entre os grupos, tornando impossível a sua interpretação.

Assim,  $h(t_{il}|\mathbf{x}_{il}, z_l)$  é uma função risco condicional ao efeito aleatório  $z_l$ . O modelo marginal (integrando com respeito a  $z$ ), em geral não preserva a estrutura de riscos proporcionais. De fato, a estrutura de riscos proporcionais só é preservada se a fragilidade segue uma distribuição estável positiva (KLEIN; MOESCHBERGER, 1997; HOUGARD, 2000).

A verossimilhança (condicional) envolvendo os dados completos (dados observados e dados não observados) é dada por

$$L(\boldsymbol{\beta}, \boldsymbol{\psi} | D_{comp}) = \prod_{l=1}^L \prod_{i=1}^{n_l} [h_0(t_{il}|\boldsymbol{\psi}) z_l \exp(\mathbf{x}_{il}\boldsymbol{\beta})]^{\delta_{il}} \exp\{-H_0(t_{il}|\boldsymbol{\psi}) z_l \exp(\mathbf{x}_{il}\boldsymbol{\beta})\}, \quad (4.2)$$

em que  $D_{comp} = \{(t_{il}, \delta_{il}, \mathbf{x}_{il}, z_l) : l = 1, \dots, L, i = 1, \dots, n_l\}$  denota os dados completos.

Logo, a verossimilhança marginal é obtida integrando-se (4.2) com respeito a densidade  $g(\mathbf{z}|\boldsymbol{\phi})$ , isto é,

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\phi} | D_{obs}) = \int_{\mathbf{z}} L(\boldsymbol{\beta}, \boldsymbol{\psi} | D_{comp}) g(\mathbf{z}|\boldsymbol{\phi}) d\mathbf{z} \quad (4.3)$$

em que  $D_{obs} = \{(t_{il}, \delta_{il}, \mathbf{x}_{il}) : l = 1, \dots, L, i = 1, \dots, n_l\}$  denota os dados observados.

A função de verossimilhança dada em (4.3) é bastante complicada uma vez que, em geral não possui uma forma analítica fechada, sendo necessário uso de métodos de integração numérica. Entretanto, sob o paradigma Bayesiano, este problema pode ser facilmente contornado se utilizarmos o amostrador de Gibbs para gerarmos uma amostra da distribuição conjunta a *posteriori* de  $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{z})$ .

Uma questão que naturalmente se coloca, ao reconhecermos a necessidade da utilização de um modelo que inclua heterogeneidade não observada, é a da escolha da distribuição da variável aleatória  $Z$  (fragilidade não observada ou efeito aleatório). De fato, precisamos identificar qual é a distribuição de probabilidade adequada sobre  $Z$  através da qual conseguiremos estimar a variância  $\eta$ .

Como vimos no começo deste Capítulo 4, existem dois cenários distintos sobre o efeito aleatório. Um dos cenários está associado aos efeitos aleatórios não correlacionados, em



que se assume independência entre os indivíduos ou grupos, e o outro cenário que tange aos efeitos aleatórios correlacionados, neste caso tomando em consideração a estrutura de dependência entre os grupos. Cada um dos cenários descritos acima, suporta certos tipos de distribuições. Segundo Lawson (2008), as distribuições *gama* e *log-normal* tem sido bastante utilizadas para modelagem do efeito aleatório quando não se assume a suposição de dependência entre os grupos, e os modelos (CAR), especificamente os modelos Condicionais Autoregressivos Impróprios ou Intrínsecos (ICAR) demonstram maiores vantagens na modelagem do efeito aleatório assumindo dependência entre os grupos (BANERJEE *et al.*, 2004).

## 4.1 Modelo de Fragilidade Gama

Segundo Vaupel e Yashin (1983), Hougaard (1984, 2000), a distribuição *gama* tem sido mais utilizada como distribuição da fragilidade, por apresentar vantagens sob o ponto de vista do tratamento matemático, uma vez que a distribuição *gama* corresponde a única distribuição de fragilidade para qual a função de verossimilhança marginal apresenta forma fechada. O modelo de fragilidade *gama* é definido como no modelo (4.1), isto é,

$$h(t_{il}|\mathbf{x}_{il}, z_l) = h_0(t_{il}) z_l \exp\{\mathbf{x}_{il}\boldsymbol{\beta}\},$$

considerando agora que as fragilidades  $z_l$ ,  $l = 1, \dots, L$ , são assumidas serem uma amostra independente de variáveis aleatórias  $Z_l$  com distribuição *gama* de parâmetros  $\eta$  e  $\eta$ , isto é,  $Z_l \sim \Gamma(\eta, \eta)$ .

A suposição assumida no modelo com fragilidade *gama* é que a variável aleatória  $Z$  deve ter média um para garantir a identificabilidade do modelo. Com base nos parâmetros assumidos nesta distribuição, esta suposição não é violada, uma vez que,  $E[z_l] = 1$  e  $Var(z_l) = 1/\eta$ . Teremos heterogeneidade entre os grupos para pequenos valores de  $\eta$ .

## 4.2 Modelo de Fragilidade Log-normal

Outra distribuição bastante utilizada na prática para os efeitos aleatórios é a distribuição *log-normal* (HOUGAARD, 2000). No entanto esta distribuição apresenta uma desvantagem uma vez que a verossimilhança marginal não possui forma fechada, sendo necessário aplicação do método de integração numérica. O método de quadratura gaussiana tem-se destacado bastante para solucionar o problema da verossimilhança marginal

para esta distribuição.

O modelo com fragilidade *log-normal* é um modelo que ainda preserva a estrutura de riscos proporcionais (veja Equação (2.11)) e é definido do seguinte modo

$$\begin{aligned} h(t_{il}|x_{il}, z_l) &= h_0(t_{il}) z_l \exp\{\mathbf{x}_{il}\boldsymbol{\beta}\} \\ &= h_0(t_{il}) \exp\{\mathbf{x}_{il}\boldsymbol{\beta} + w_l\} \end{aligned}$$

em que  $w_l = \log z_l$ ,  $l = 1, \dots, L$ , são variáveis aleatórias independentes e identicamente distribuídas. Banerjee *et al.* (2003) trabalharam em seu artigo, com a distribuição normal de média zero e precisão  $\tau$  desconhecida, isto é,  $w_l \stackrel{i.i.d}{\sim} N(0, \tau)$ . De fato, para que o modelo de fragilidade *log-normal* seja identificável, o termo do efeito aleatório deve ter média zero. Neste caso valores pequenos de  $\tau$  indicam um alto grau de heterogeneidade entre os grupos.

### 4.3 Fragilidade espacial

Os modelos com fragilidade compartilhada *gama* e *lognormal* apresentados nas seções anteriores foram contruídos com base na suposição de independência entre os grupos ou conglomerados. Como já foi mencionado anteriormente, se os grupos representarem áreas, no entanto, a suposição de independência entre os grupos pode deixar de ser razoável.

O modelo CAR impróprio, proposto por Besag *et al.* (1991), fornece uma maneira bastante atrativa para modelarmos efeitos espaciais associados a dados de área. Este modelo depende de uma matriz de proximidade  $V_{L \times L}$ , cujas entradas  $v_{kl}$  assumem valor 1 se as áreas  $k$  e  $l$  possuem fronteiras comuns (evento denotado por  $k \sim l$ ), e zero caso contrário. É possível mostrar que a distribuição condicional completa (a *priori*) associada a cada unidade de área é dada por

$$(\omega_k | \boldsymbol{\omega}_{(-k)}) \sim N \left( \sum_{l \in \varphi_k} w_l / v_k; \phi v_k \right), \quad (4.4)$$

em que  $\boldsymbol{\omega}_{(-k)} = \{\omega_l : l \neq k\}$ ,  $\varphi_k$  denota o conjunto de índices associados às áreas vizinhas à área  $k$ ,  $v_k = \sum_{l \neq k} v_{kl}$  é o número de áreas vizinhas a área  $k$  e  $\phi$  é um parâmetro de precisão que controla a variabilidade comum associada ao conjunto de todas as áreas sob investigação. A fim de garantirmos a identificabilidade do modelo (4.4) devemos impor a restrição  $\sum_{k=1}^L \omega_k = 0$ . Esta restrição pode ser imposta numericamente através da recentralização do vetor  $\boldsymbol{\omega}$  ao redor de sua própria média em cada iteração do amostrador

Gibbs.

## 5 Aplicações

Neste capítulo serão apresentadas as análises de dois bancos de dados reais. No primeiro banco de dados a variável resposta é o tempo até a morte de pacientes diagnosticados com câncer de nasofaringe em 8 condados do estado de Connecticut, EUA. O segundo banco de dados reais é referente aos crimes em 8 municípios pertencentes à RMBH, em que a variável resposta é o tempo entre a primeira e a segunda tentativa de homicídio.

Para a geração das amostras *a posteriori* das quantidades de interesse foram utilizados os softwares *R* e *OpenBUGS* interligados por meio do pacote *rbugs*. Os *scripts* utilizados para o ajuste dos modelos estão disponíveis no Apêndice. Para cada modelo ajustado, foi gerada uma única cadeia com 600 mil iterações, das quais foram descartadas as primeiras 100 mil iterações (período de aquecimento da cadeia). Para diminuir a autocorrelação da amostra gerada foi considerado um salto (*lag*) de 100 iterações, resultando em amostras de tamanho 5000 das distribuições *a posteriori* das quantidades de interesse.

Para ambos os bancos de dados analisados nesta dissertação foram ajustados o modelo de riscos proporcionais de Cox (*M1*), e três modelos de fragilidade compartilhada, a saber: fragilidade *gama* (*M2*), fragilidade *log-normal* (*M3*), e o modelo CAR intrínscico para fragilidade espacial (*M4*).

Nesta dissertação foi assumido que os tempos de resposta em ambos bancos de dados seguem a distribuição *Weibull*, o que implica que função risco base considerada será a Equação (2.7). A fim de garantir identificabilidade do modelo, o parâmetro de escala  $\gamma$  da distribuição *Weibull* será assumido como sendo igual a 1, uma vez que consideraremos o intercepto na modelagem. A função risco considerada na modelagem é, então, dada por

$$h(t_{il}|\mathbf{x}_i, z_l) = \alpha t^{\alpha-1} z_l \exp\{\mathbf{x}_i \boldsymbol{\beta}\}.$$

Para o parâmetro de forma  $\alpha$  da distribuição *Weibull* e os coeficientes da regressão  $\beta$ 's, comuns em todos os modelos ajustados, as seguintes especificações *a priori* foram consideradas:  $\alpha \sim \Gamma(10^{-3}, 10^{-3})$  e  $\beta_k \sim N(0, 10^{-3})$ ,  $k = 1, \dots, p$ . A escolha da dis-

tribuição *a priori* do parâmetro de forma  $\alpha$  da distribuição *Weibull* foi feita de modo a ter média 1 (note que, quando  $\alpha = 1$  a função risco é constante) e variância 1000, isto é,  $\alpha \sim \Gamma(10^{-3}, 10^{-3})$ , e dos coeficientes de regressão de modo a terem média zero (estatisticamente não significativos) e variância 1000, isto é,  $\beta_k \sim N(0, 10^{-3})$ . Para os hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$  associados aos modelos *M2*, *M3* e *M4*, respectivamente, foi conduzida uma análise de sensibilidade, em que foram consideradas as seguintes distribuições *a priori*:  $\Gamma(1, 1)$ ,  $\Gamma(0.1, 0.1)$ ,  $\Gamma(0.01, 0.01)$  e  $\Gamma(0.001, 0.001)$ . Para obtenção de melhores resultados da análise de sensibilidade, foram consideradas *priori* mais informativas e menos informativas, com a mesma média (média=1) e diferentes valores da variância.

A convergência da cadeia foi verificada mediante aplicação das técnicas de Heidelberger e Welch (1983), Geweke (1992). Os resultados dos testes de convergência são mostrados no Apêndice.

## 5.1 Aplicação a dados de Câncer

Os dados de câncer de nasofaringe nos condados do estado de Connecticut, nos EUA, são referentes à um período de acompanhamento de 34 anos (1973 à 2007). Nesse período temos uma amostra de 560 pacientes sendo registradas 390 mortes (69.64%). Foram consideradas as seguintes variáveis explicativas: *sexo* (feminino=1, masculino=0), *raça* (branco=1, outros=0), *idade* (em anos) e *número de tumores primários*.

Na Tabela 5.1 são apresentados os valores do DIC e o número de parâmetros efetivos,  $p_D$ , dos três modelos de fragilidade ajustados considerando-se as diferentes especificações *a priori* para os hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$ . Com base na Tabela 5.1 verificou-se que o melhor modelo com fragilidade *gamma*, (*M2*), é obtido considerando-se  $\eta \sim \Gamma(0.001, 0.001)$ , uma vez que apresenta menor valor do DIC (4125), entre todas as especificações *a priori* do hiperparâmetro  $\eta$ , exceto a especificação  $\Gamma(0.01, 0.01)$ . De fato, apesar das distribuições *a priori*  $\Gamma(0.001, 0.001)$  e  $\Gamma(0.01, 0.01)$  apresentarem o mesmo valor de DIC, considerando  $\Gamma(0.001, 0.001)$  obtém-se um modelo com um número de parâmetros efetivos ( $p_D = 7.968$ ) próximo ao número de parâmetros incluídos no modelo, 6 (5 componentes de  $\beta$  mais 1 do parâmetro da distribuição *Weibull*,  $\alpha$ ).

De modo análogo, o melhor modelo de fragilidade *log-normal* (*M3*), é obtido considerando-se  $\tau \sim \Gamma(0.001, 0.001)$ , uma vez que apesar de ilustrar o mesmo valor de DIC em relação a especificação  $G(0.01, 0.01)$ , apresenta um número de parâmetros efetivos,  $p_D = 7.92$  próximo ao número de parâmetros incluídos no modelo. No caso do modelo de fragilidade

Tabela 5.1: Análise de sensibilidade das distribuições *a priori* dos hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$  - Aplicação aos dados de câncer-Valores de DIC.

Modelo	Hiperp.	Distribuições <i>a priori</i>			
		$\Gamma(1, 1)$	$\Gamma(0.1, 0.1)$	$\Gamma(0.01, 0.01)$	$\Gamma(0.001, 0.001)$
M2	$\eta$	4130.0 (12.09)	4127.0 (10.52)	4125.0 (8.894)	<b>4125.0 (7.968)</b>
M3	$\tau$	4130.0 (12.05)	4127.0 (10.49)	4125.0 (8.46)	<b>4125.0 (7.92)</b>
M4	$\phi$	4127.0 (8.998)	<b>4121.0 (3.951)</b>	4125.0 (7.249)	4123.0 (5.221)

espacial, (*M4*), o melhor modelo é obtido considerando-se  $\phi \sim \Gamma(0.1, 0.1)$ . Estas constatações, nos levam a concluir que nos dois primeiros cenários, quanto menos informativa for a especificação *a priori* do hiperparâmetro que controla a variância do efeito aleatório, melhor é a qualidade de ajuste, e neste último modelo, (*M4*), a distribuição *a priori* do hiperparâmetro  $\phi$ , informativa tende a determinar um bom ajuste do modelo.

Também foi ajustado um modelo sem efeito aleatório (*M1*), isto é, o modelo de Cox, tendo se obtido como valor de DIC=4125, com número de parâmetros efetivos,  $p_D = 6.074$ . Comparando este valor de DIC com o dos outros três modelos, *M2*, *M3* e *M4*, conclui-se que o modelo *M4*, com hiperparâmetro  $\phi$ , que tem como distribuição *a priori*  $\Gamma(0.1, 0.1)$ , possui melhor ajuste para modelar os dados de câncer de nasofaringe, uma vez que apresentou o menor valor de DIC (4121).

A fim de avaliar o impacto da variação das especificações das distribuições *a priori* dos hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$ , sobre o parâmetro de forma  $\alpha$  da *Weibull*, recorreu-se aos *boxplots* como ilustrado na Figura 5.1. Observamos que os modelos *M2* e *M3* apresentam um padrão idêntico na sua distribuição *a posteriori* à medida que variamos de uma distribuição *a priori* para outra. Entretanto, a distribuição *a posteriori* de  $\alpha$  obtida a partir do modelo *M4* mostrou-se um pouco sensível as especificações *a priori* para  $\alpha$ . A medida que a variância aumenta, a distribuição *a posteriori* de  $\phi$  tende a concentrar-se em valores menores.

Do modo análogo ao parâmetro de forma da distribuição *Weibull*, foi feita uma análise de sensibilidade para cada um dos coeficientes de regressão associados às covariáveis consideradas na modelagem, de acordo com as diferentes especificações *a priori* para  $\eta$ ,  $\tau$  e  $\phi$ .

A Figura 5.2 apresenta o *boxplot* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *idade*. Verificamos que em todos os modelos de fragilidade este coeficiente apresenta características *a posteriori* similares, indicando não ser sensível para as diferentes especificações *a priori* dos hiperparâmetros  $\eta$  e  $\tau$ , respectivamente associados

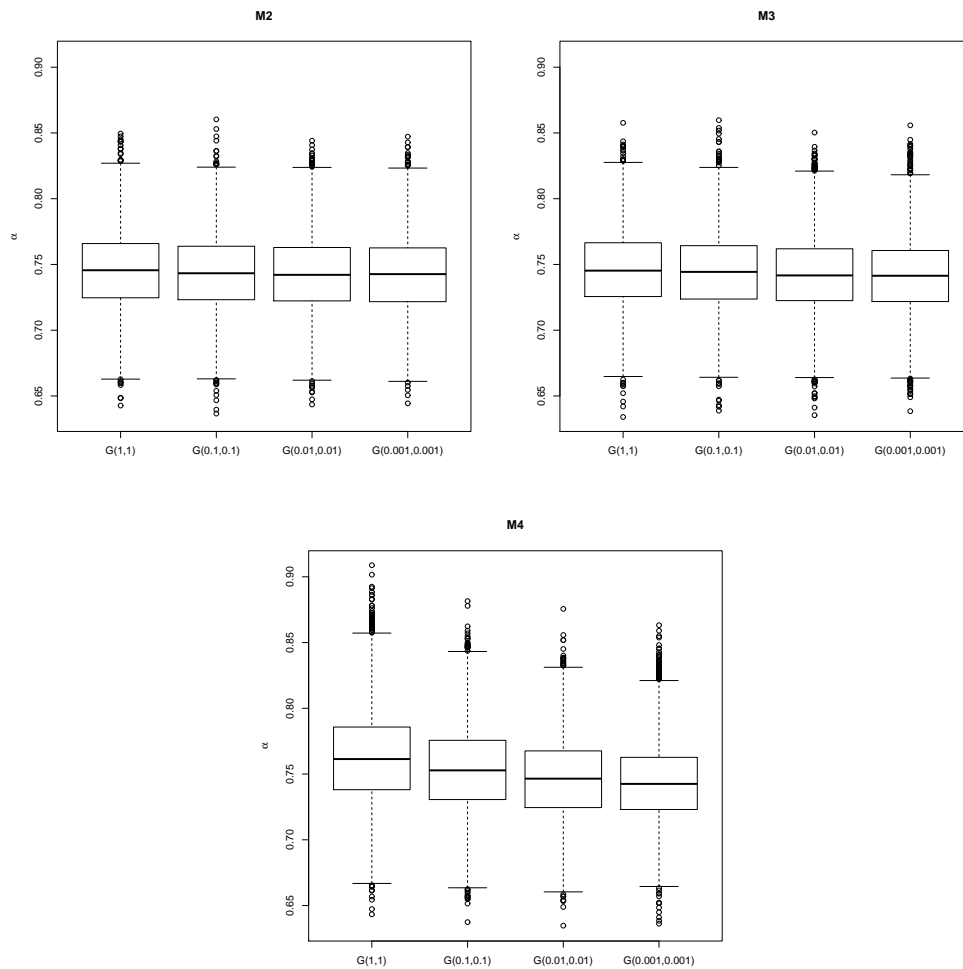


Figura 5.1: *Boxplots* da distribuição *a posteriori* de  $\alpha$  para dados de câncer

aos modelos  $M2$  e  $M3$ ) e com tendências de apresentar os mesmos valores das medianas *a posteriori*. O cenário observado com base na covariável *idade*, foi igualmente verificado nas demais covariáveis incluídas no modelo.

Para o modelo  $M1$  e os modelos  $M2$ ,  $M3$  e  $M4$ , selecionados na análise de sensibilidade exibida na Tabela 5.1, foi feita uma análise dos parâmetros associados às covariáveis. Na Figura 5.3 estão ilustrados os *boxplots* referentes as distribuições *a posteriori* desses parâmetros. Observamos similaridades nas características dos parâmetros em suas distribuições, isto é, a distribuição *a posteriori* de cada um dos parâmetros não se diferencia de um modelo para o outro. De fato, se considerarmos por exemplo o parâmetro associado à covariável *idade* nos modelos  $M1$ ,  $M2$ ,  $M3$  e  $M4$ , notamos um comportamento similar em sua distribuição *a posteriori*.

Como na Figura 5.3, na Figura 5.4 a similaridade da distribuição *a posteriori* é observada apenas nos três primeiros modelos,  $M1$ ,  $M2$ ,  $M3$ , desta vez considerando o pa-

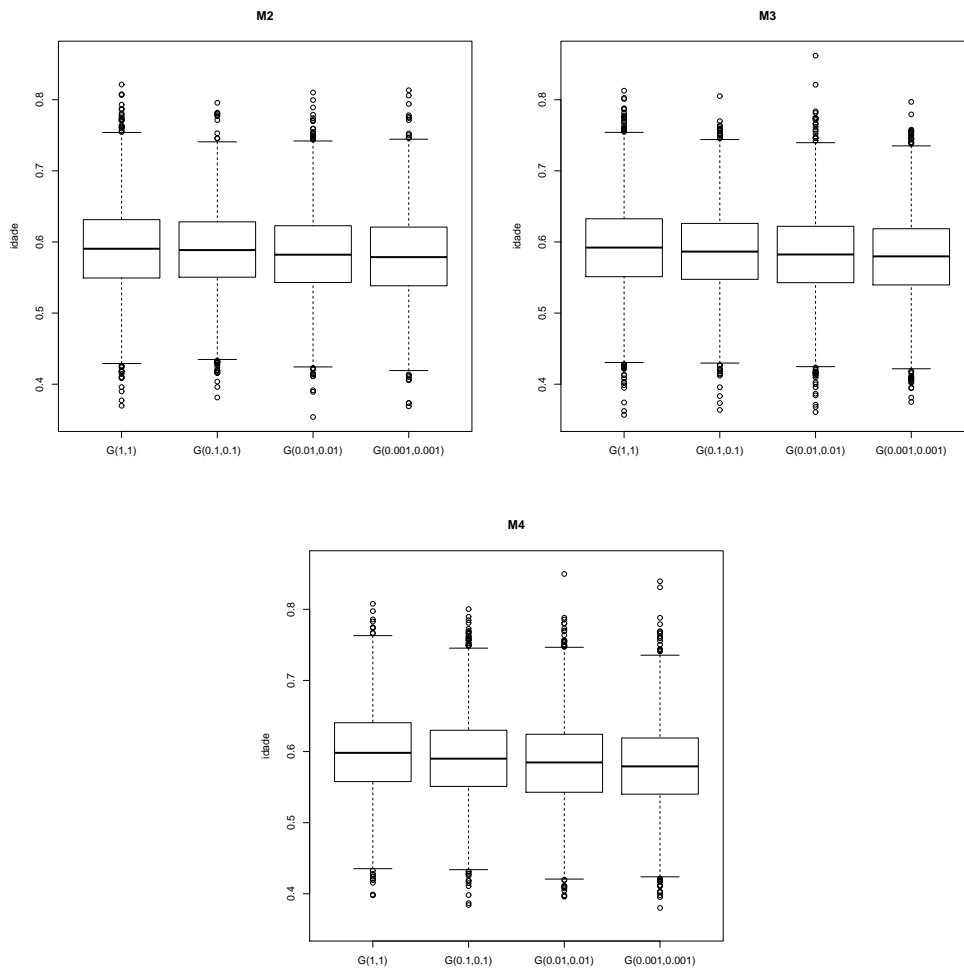


Figura 5.2: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *idade* para dados de câncer

râmetro de forma  $\alpha$  da distribuição *Weibull*. De fato, a distribuição *a posteriori* deste parâmetro apresenta valor mediano superior no modelo *M4*, em relação ao valor apresentado por outros modelos.



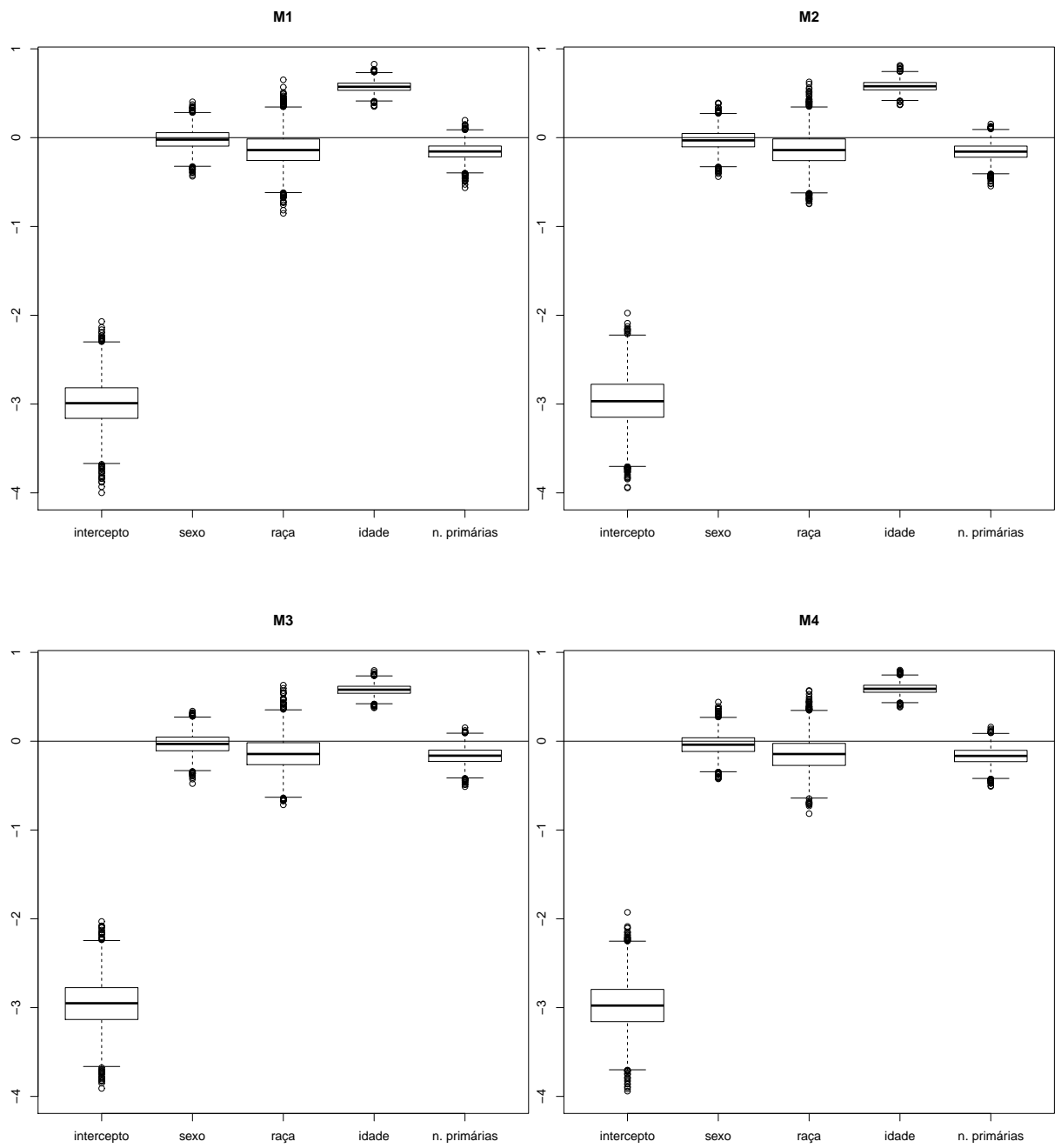


Figura 5.3: *Boxplot* das distribuições *a posteriori* dos  $\beta$ 's os para dados de câncer

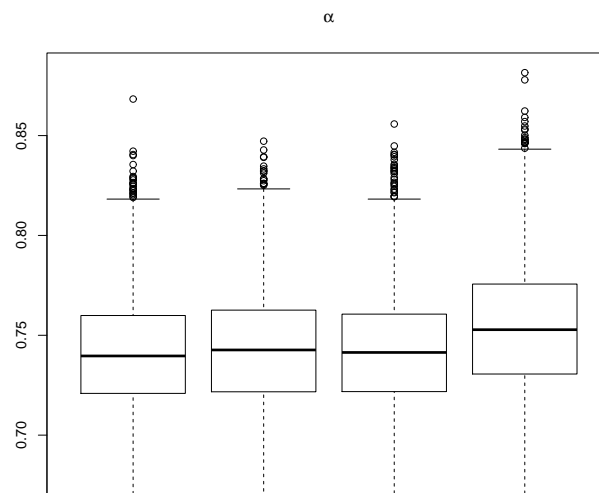


Tabela 5.2: Estimativas das distribuições *a posteriori* dos parâmetros associados ao modelo M4 para dados de câncer

	média	desvio	limite inferior	limite superior
$\alpha$	0.7533	0.0328	0.6925	0.8188
intercepto	-2.9761	0.2729	-3.5070	-2.4520
sexo	-0.0371	0.1142	-0.2633	0.1823
raça	-0.1459	0.1852	-0.4905	0.2378
idade	0.5909	0.0583	0.4825	0.7119
n° de tumores primários	-0.1679	0.0929	-0.3418	0.0176
$\phi$	11.7564	8.6700	0.4279	29.2000

Com o objetivo de avaliar a contribuição de cada uma das covariáveis assumidas na modelagem, recorreu-se ao melhor modelo mais bem ajustado (*M4*) escolhido anteriormente com base no DIC. Os resultados apresentados na Tabela 5.2, mostram a *média*, *desvio padrão* e *intervalos HPD* para o parâmetro de forma  $\alpha$  da distribuição *Weibull*, para o hiperparâmetro que controla a variância da fragilidade espacial e também para cada uma das covariáveis incluídas no modelo. Notamos que apenas os parâmetros  $\beta_1$  (*intercepto*) e  $\beta_4$  (*idade*), são estatisticamente significativos, uma vez que os intervalos HPD de cada um deles não contém zero.

Com base no modelo *M4*, mantendo inalteradas as covariáveis *sexo*, *raça* e *número de tumores primários*, a cada variação unitária na covariável *idade*, condicionado em cada fragilidade, o risco dos pacientes com câncer no estado de Connecticut experimentarem o evento de interesse (morte), aumenta em média  $e^{0.5909} = 1.81$ .

Ainda na Tabela 5.2 observamos que o intervalo HPD está a esquerda de 1, implicando que a função risco é decrescente, conforme mostra a Seção 2.2.

Um outro parâmetro de interesse neste estudo é aquele que controla a precisão da fragilidade, neste caso representado por  $\phi$ . O valor da média *a posteriori* do parâmetro  $\hat{\phi} = 11.7564$ , indicando um baixo nível de heterogeneidade entre as áreas.

Uma maneira de analisar dados de sobrevivência espacialmente arrajandos é pelo mapeamento dos termos do efeito aleatório. Na Figura 5.5 são apresentadas as fragilidades estimadas pelos modelos *M2*, *M3* e *M4* mais bem ajustados na análise de sensibilidade. Com base no mapa fornecido pelo modelo *M4*, que apresentou o menor DIC dentre todos os modelos considerados, nota-se que os condados localizados na região sudoeste (Fairfield e New Haven) têm menor risco de morte em decorrência de câncer de nasofaringe que nos demais condados do estado. Os condados de Litchfield e Middlesex, por outro lado, apresentaram o maior risco dentre todos os condados do estado, enquanto os condados

Tabela 5.3: Análise de sensibilidade das distribuições *a priori* dos hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$  - Aplicação aos dados de crimes-Valores de DIC

Modelo	Hiperp.	Distribuições <i>a priori</i>			
		$\Gamma(1, 1)$	$\Gamma(0.1, 0.1)$	$\Gamma(0.01, 0.01)$	$\Gamma(0.001, 0.001)$
M2	$\eta$	1437 (10.1)	1434 (8.553)	1432 (7.324)	<b>1431 (6.54)</b>
M3	$\tau$	1437 (9.979)	1434.450 (8.485)	1432 (7.271)	<b>1431 (6.692)</b>
M4	$\phi$	1435 (5.536)	1432 (5.672)	<b>1429 (3.665)</b>	1431 (6.002)

de New London, Windham, Tolland e Hartford apresentaram risco intermediário de óbito devido ao câncer de nasofaringe. É importante observar, também, que os mapas fornecidos pelos modelos M2 e M3, que não consideram a estrutura espacial, diferem do mapa fornecido pelo modelo M4, podendo levar a conclusões enganosas.

## 5.2 Aplicação a dados de crimes

Como abordado anteriormente no começo deste Capítulo, a segunda aplicação deste trabalho de dissertação está relacionada aos dados sobre crimes ocorridos na RMBH, registados num período de 4 anos (2008 à 2012) tomando em consideração uma amostra de 118 indivíduos que sofreram duas tentativas de homicídio. O homicídio foi consumado em 26.3% dos casos registados. As análises foram feitas considerando as seguintes covariáveis: *sexo*, (branco=1, outros=0), *idade*, (em anos), *local* (indicadora: 1= as duas tentativas de homicídio ocorreram no mesmo município e 0 caso contrário) e *consumado* (indicadora: 1= o homicídio foi consumado na segunda tentativa do crime e 0 caso contrário). A variável resposta para esta aplicação como foi abordado ainda no começo deste Capítulo é o tempo entre a primeira e a segunda tentativa de homicídio.

Os resultados ilustrados na Tabela 5.3 fornecem os valores do DIC e o número de parâmetros efetivos,  $p_D$  em parêntesis, a partir dos quais foi possível fazer uma análise de sensibilidade dos três modelos de fragilidade ajustados considerando as diferentes especificações *a priori* sobre os hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$ . Com base na Tabela 5.3, verificou-se que os melhores modelos de fragilidade (menor DIC=1431) *gama* (M2) e *log-normal* (M3) foram obtidos considerando-se como especificação *a priori* para  $\eta$  e  $\tau$  a  $\Gamma(0.001, 0.001)$ . De modo análogo, no caso do modelo de fragilidade espacial, (M4), o melhor modelo (menor DIC=1429) é obtido considerando a  $\Gamma(0.01, 0.01)$  como distribuição *a priori* para  $\phi$ . Em geral, especificações *a priori* menos informativas, tendem a determinar melhor ajuste com base nestes dados de crimes.

Foi igualmente ajustado um modelo M1 sem a fragilidade (não observada), tendo se

obtido o valor de  $DIC=1430$ , com número de parâmetros efetivos,  $p_D = 5.975$ . Comparando este DIC com o valor dos menores DIC dos três modelos,  $M2$ ,  $M3$  e  $M4$ , opta-se pelo modelo  $M4$  com distribuição *a priori*  $\phi \sim \Gamma(0.01, 0.01)$  e o menor DIC.

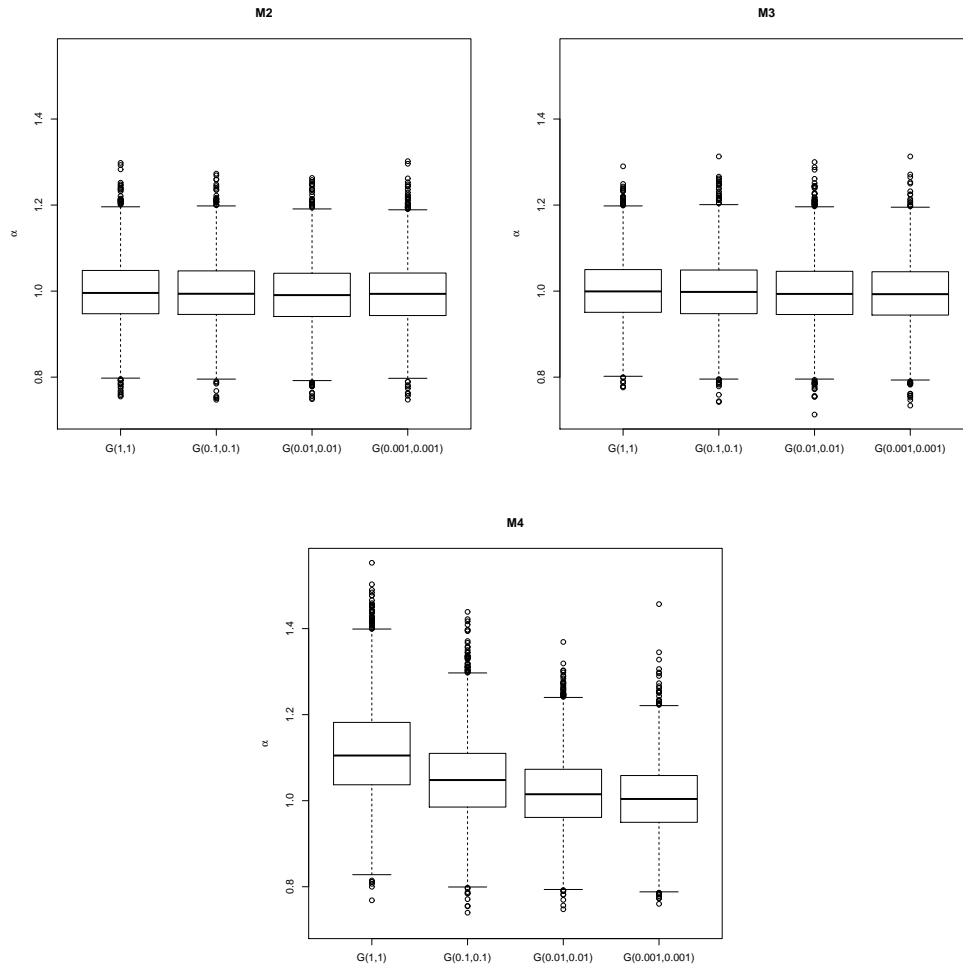


Figura 5.6: *Boxplots* da distribuição *a posteriori* de  $\alpha$  para dados de crimes

Para analisar a sensibilidade do parâmetro de forma  $\alpha$  da distribuição *Weibull*, utilizando as diferentes especificações das distribuições *a priori* dos hiperparâmetros  $\eta$ ,  $\tau$  e  $\phi$  associados às variâncias dos efeitos aleatórios inclusos nos modelos  $M2$ ,  $M3$  e  $M4$  respectivamente, foram construídos os *boxplots*, da Figura 5.6. No modelo  $M2$  notamos que cada amostra *a posteriori* tende a mostrar, características similares às observadas nas outras amostras. Como por exemplo, a variabilidade observada com base na *priori*  $\Gamma(1, 1)$  no modelo  $M2$  é identicamente observada na *priori*  $\Gamma(0.001, 0.001)$ , assim como em outras especificações *a priori* ao nível do modelo  $M2$ .

Estas características da distribuição *a posteriori* de  $\alpha$ , observadas nos *boxplots* obtidos para o modelo  $M2$ , são verificadas também nos *boxplots* obtidos com base no modelo

*M3*. O cenário muda quando consideramos os resultados para o modelo *M4*; note que o parâmetro de forma da distribuição *Weibull* sofre influência das variações de especificações *a priori* do hiperparâmetro  $\phi$  que controla a variância da fragilidade espacial. De fato, considerando distribuições *a priori* menos informativas, observamos que os valores das medianas *a posteriori* tornam-se menores.

Em geral, o parâmetro de forma da distribuição *Weibull* não foi influenciado pelas variações de especificação *a priori* ao nível dos modelos *M2* e *M3* e não sofreu efeitos na mudança do modelo *M2* para o *M3*. Entretanto, notamos evidentes efeitos da variação de *priori* ao nível do modelo *M4*, indicando diminuição gradual da mediana *a posteriori* à medida que consideramos *priori* menos informativas.

Os procedimentos aplicados ao parâmetro de forma da distribuição *Weibull*, no que concerne à análise de sensibilidade, também foram usados para cada uma das covariáveis incluídas nos modelos de fragilidade. A Figura 5.7 apresenta os *boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *sexo*. Verifica-se características similares em todos os modelos de fragilidade, com *priori* menos informativas apresentando maiores valores da mediana *a posteriori*. Para as demais covariáveis incluídas no modelo foi observado que diferentes especificações *a priori* para  $\eta$ ,  $\tau$  e  $\phi$  não influenciam nas distribuições *a posteriori* dos coeficientes de regressão. Os gráficos destas covariáveis são mostrados no Apêndice.

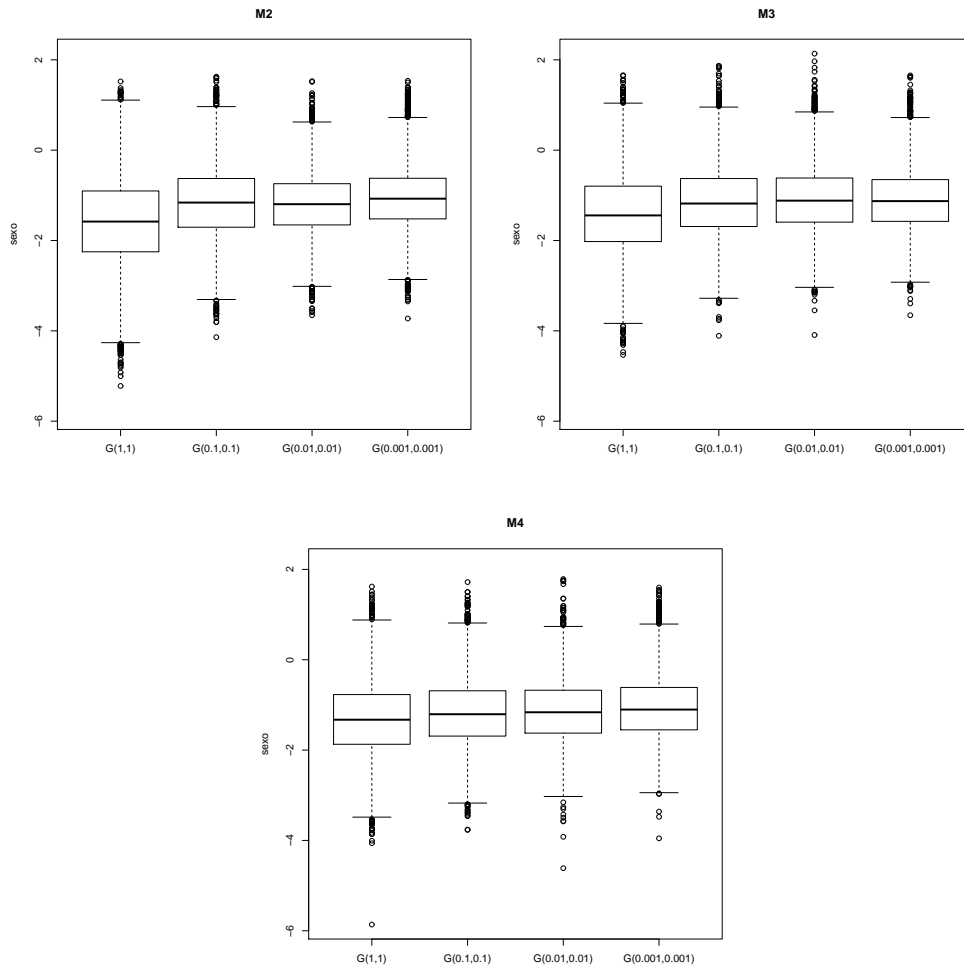


Figura 5.7: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *sexo* para dados de crimes

Os *boxplots* mostrados na Figura 5.8 ilustram o comportamento da distribuição *a posteriori* dos coeficientes associados às covariáveis incluídas no modelo *M1* e nos modelos *M2*, *M3* e *M4* mais bem ajustados de acordo com DIC. Observamos uma similaridade de comportamento dos coeficientes de regressão em todos os modelos, como por exemplo, a presença dos valores extremos (*outliers*) observados em cada coeficiente de regressão do modelo *M4*, tende a se manter nos restantes modelos. Uma outra similaridade nos quatro modelos considerados, é observada para o *intercepto* que apresenta um valor da mediana *a posteriori* inferior aos valores medianos *a posteriori* dos coeficientes das covariáveis consideradas na modelagem, além de indicar maior variabilidade, uma vez que apresentou maior intervalo interquartílico em todos os modelos. Em geral, cada coeficiente de regressão *a posteriori* tende a manter o seu comportamento, à medida que mudamos de um modelo para o outro.

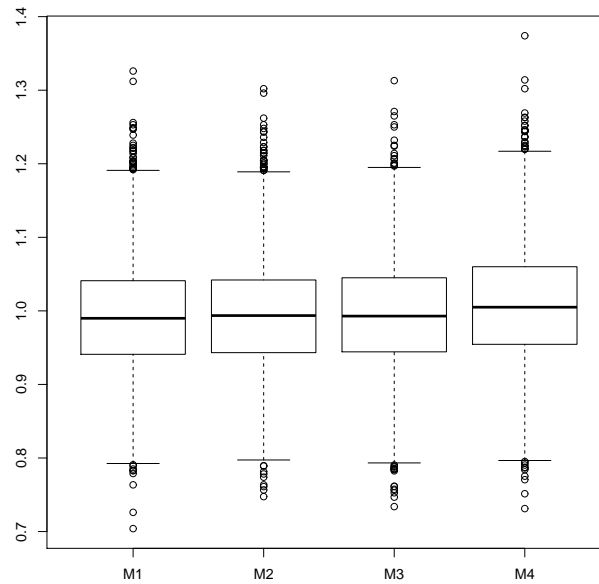


Figura 5.9: *Boxplot* das distribuições *a posteriori* dos  $\alpha$ 's para dados de crimes

Na Figura 5.9 observamos os *boxplots* da distribuição *a posteriori* do parâmetro de forma  $\alpha$  da distribuição *Weibull* para cada um dos quatro modelos considerados. Notamos a presença de características que dão similaridade na sua distribuição nos três primeiros modelos,  $M1$ ,  $M2$  e  $M3$ . Por exemplo, em cada modelo, o parâmetro tende a ilustrar a mesma variabilidade *a posteriori*, uma vez que apresenta a mesma diferença interquartílica, com tendências de apresentar o mesmo valor da mediana *a posteriori*. Entretanto, este parâmetro sofre mudanças da sua distribuição *a posteriori*, quando consideramos o modelo de fragilidade espacial,  $M4$ . De fato, com base neste modelo, o parâmetro possui um valor da mediana *a posteriori* superior, o mesmo se observa para os valores mínimo e máximo. Em geral, o parâmetro de forma  $\alpha$  da distribuição *Weibull* tem comportamento *a posteriori* similar nos modelos sem fragilidade e com fragilidade *gamma* e *log-normal*, sofrendo modificações quando consideramos o modelo de fragilidade espacial.

A fim de analisar a contribuição de cada covariável recorreu-se ao modelo  $M4$  com *priori*  $\phi \sim \Gamma(0.01, 0.01)$  e menor DIC. Na Tabela 5.4, além de observarmos as estimativas da média *a posteriori*, do desvio padrão *a posteriori* e do intervalo HPD de 95% para cada coeficiente de regressão associado a cada covariável, temos também estimativas para as distribuições *a posteriori* do parâmetro de forma  $\alpha$  da distribuição *Weibull* e do hiperparâmetro  $\phi$  que controla a variância da fragilidade espacial. Notamos que nenhuma covariável incluída no modelo é estatisticamente significativa, uma vez que o intervalo

Tabela 5.4: Estimativas das distribuições *a posteriori* dos parâmetros associados ao modelo M4 para dados de crimes

	média	desvio	limite inferior	limite superior
$\alpha$	1.0187	0.0827	0.8576	1.1850
intercepto	-3.8551	0.8389	-5.4970	-2.2580
sexo	-1.1365	0.7021	-2.3920	0.3277
idade	-0.0022	0.0085	-0.0194	0.0141
lugar	-0.1034	0.2810	-0.6483	0.4692
consumado	-0.1306	0.2248	-0.5948	0.2813
$\phi$	43.0138	55.2408	0.0758	152.5000

HPD de cada um dos coeficientes de regressão contém o zero. Estes resultados mostram que, por exemplo, não faz diferença se as tentativas de homicídios ocorrerem no mesmo lugar. De modo análogo verificamos que o parâmetro de forma da distribuição *Weibull* não é estatisticamente significativo, uma vez que não se rejeita a hipótese nula deste parâmetro assumir o valor um, isto é, ( $H_0 : \alpha = 1$ ), que implica numa função risco constante ao passar do tempo.

Na Figura 5.10 são apresentados as fragilidades estimadas *a posteriori* associadas aos modelos *M2*, *M3* e *M4* mais bem ajustados. O mapa obtido com base no modelo *M4* aponta que as áreas localizadas na região oeste do mapa tendem a apresentar maior risco de ocorrência de homicídios que os municípios localizados na região leste do mapa. Tal resultado está de acordo com o que era esperado, uma vez que os municípios localizados na região oeste do mapa são sabidamente mais violentas que aqueles localizados na região leste. É importante observar também que o padrão de distribuição do risco de ocorrência da segunda tentativa de homicídio associado ao modelo *M4* é diferente daquele observado nos mapas obtidos com base nos modelos *M2* e *M3*. Tal resultado pode ser explicado pela estrutura de dependência espacial considerada pelo modelo *M4*, que não é levada em conta nos modelos *M2* e *M3*.



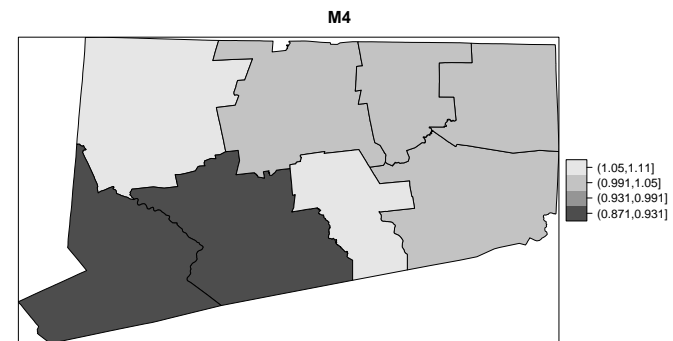
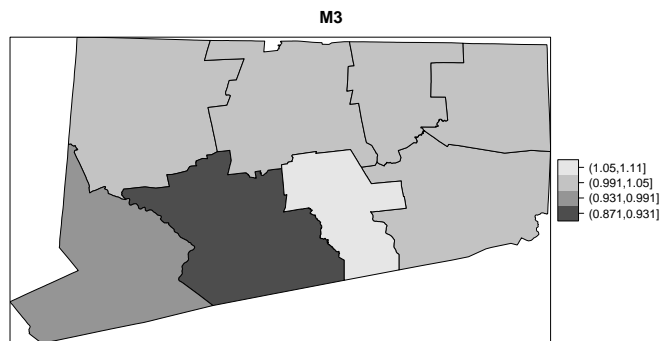
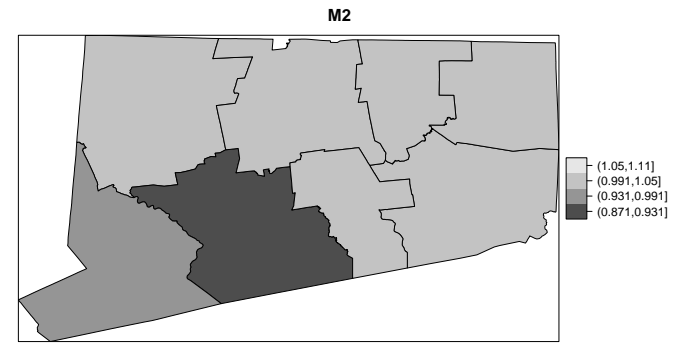
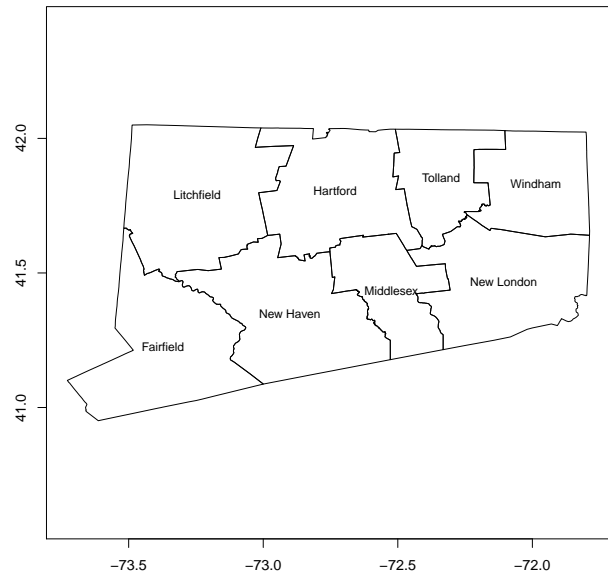


Figura 5.5: Mapas das distribuições de fragilidades para dados de câncer

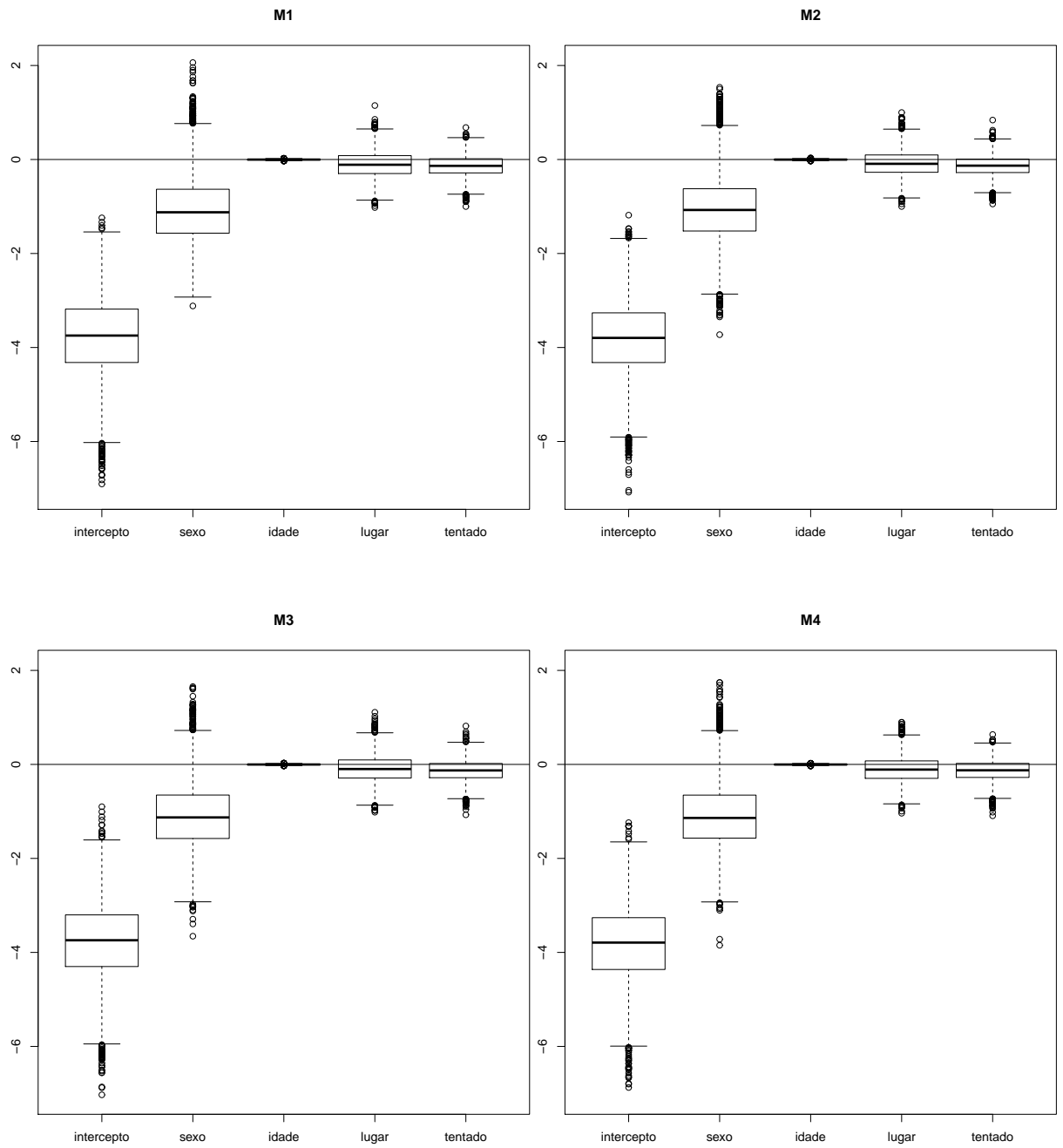


Figura 5.8: *Boxplot* dos  $\beta$ 's da Distribuição *a posteriori* para dados de crimes

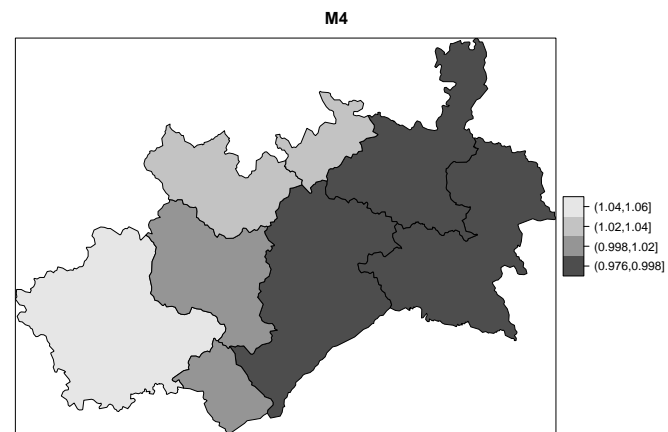
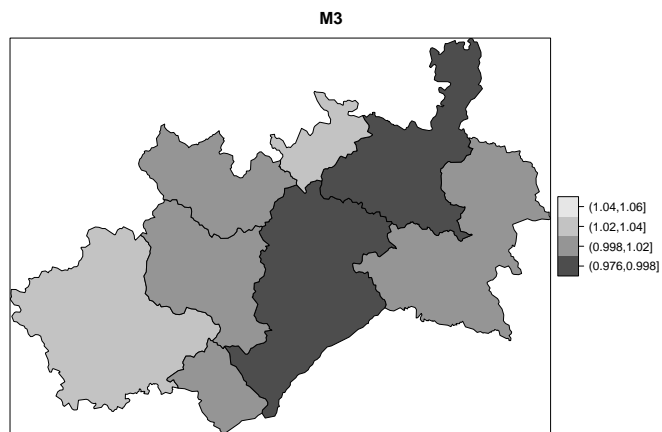
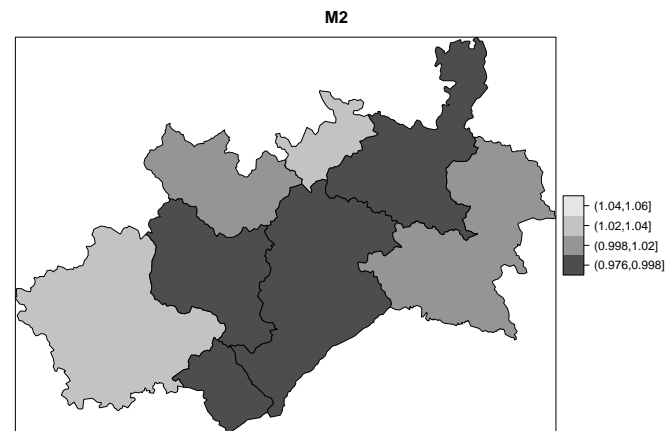
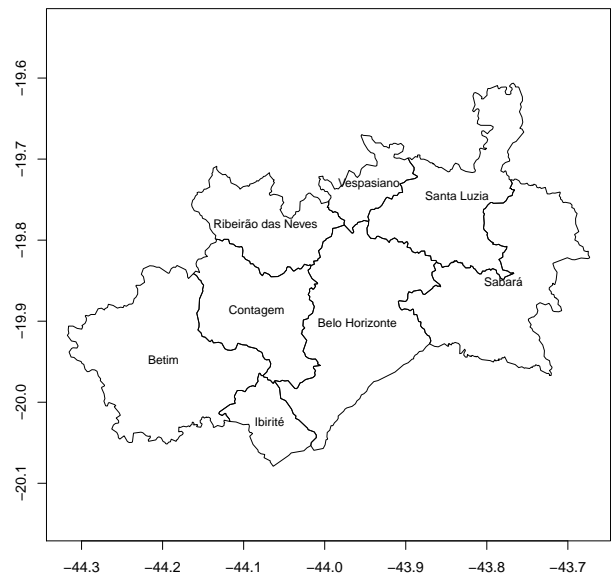


Figura 5.10: Mapas das distribuições de fragilidades para dados de crimes

## 6 *Conclusões e trabalhos futuros*

Neste trabalho foram feitos estudos de modelos de sobrevivência com efeitos aleatórios. Foram considerados o modelo de riscos proporcionais e os modelos de fragilidade, que são uma generalização do modelo de riscos proporcionais. Dentro dos modelos de fragilidade, três distribuições sobre o termo do efeito aleatório foram assumidas, duas delas suportam situações em que não se assume a suposição de dependência espacial, nomeadamente, as distribuições *gama* e *log-normal* e a terceira distribuição assumida, é bastante aplicada na situação em que se assume a suposição de dependência espacial, neste caso o modelo ICAR.

Os modelos estudados foram aplicados a dois bancos de dados reais, referentes aos dados de câncer de nasofaringe no estado de Connecticut, EUA, e dados de crimes ocorridos em municípios pertencentes à região metropolitana de Belo Horizonte, Brasil. Para cada banco de dados foi realizada uma análise de sensibilidade com base nas diferentes especificações *a priori* (com diferentes graus de informação) para os hiperparâmetros associados às variâncias dos efeitos aleatórios de cada modelo de fragilidade considerado. Com base nas análises de sensibilidade realizadas, observou-se que os modelos *M2* e *M3* apresentaram desempenho bastante similar em termos de DIC e  $p_D$  para ambos os bancos de dados analisados. Verificou-se também que os melhores ajustes para os modelos *M2* e *M3* nos dois bancos de dados foram obtidos com base nas especificações *a priori* menos informativas, diferentemente do modelo *M4*, que apresentou melhor desempenho quando distribuições *a priori* mais informativas foram consideradas.

Em geral o parâmetro de forma  $\alpha$  da distribuição *Weibull* não se mostrou sensível às especificações *a priori* dos hiperparâmetros associados aos modelos *M2* e *M3* para ambos os bancos de dados. Entretanto verificou-se que a massa da distribuição *a posteriori* de  $\alpha$  tende a concentrar-se em valores menores a medida que as distribuições *a priori* menos informativas são assumidas para os hiperparâmetros da distribuição da fragilidade. Verificou-se também que as distribuições *a posteriori* dos coeficientes de regressão associados às covariáveis incluídas nos modelos *M2*, *M3* e *M4* são similares para dados de

câncer. Entretanto, para os dados de crimes o *intercepto* e o coeficiente de regressão associado à covariável *sexo* mostraram-se sensíveis às especificações *a priori* para todos os modelos estudados.

As distribuições dos coeficientes de regressão associados aos modelos mais bem ajustados mostraram-se bastante similares, independentemente do banco de dados analisado. No entanto o parâmetro de forma da distribuição *Weibull* associado ao modelo *M4* e aos dados de crimes apresentou uma pequena diferença em relação aos demais modelos estudados.

A função risco de base estimada para os dados de câncer é decrescente, o que implica que o risco de morte em decorrência de câncer de nasofaringe para a população em estudo diminui com o passar do tempo. Foi observado também que apenas a covariável idade contribui significativamente para explicar o risco de ocorrência de óbito devido ao câncer de nasofaringe. A função risco para dados de crimes que é afetada unicamente pelos termos de efeito aleatório, tende a se manter constante ao passar do tempo. Em geral, embora a função risco de se "experimentar" o evento de interesse seja constante, em certos municípios da RMBH ela tende a diminuir e em outros aumenta, igualmente observamos este comportamento nos condados do estado de Connecticut.

Para ambos os bancos de dados analisados, verificou-se que os mapas com as fragilidades estimadas pelos modelos *M2* e *M3*, que não levam em consideração a estrutura de dependência espacial, diferiram dos mapas com as fragilidades estimadas fornecidas pelo modelo *M4*, que foi apontado como o modelo mais bem ajustado. Portanto, os resultados encontrados nessa dissertação ressaltam a importância de se utilizar os modelos que levem em consideração a estrutura de dependência espacial presente nos dados de sobrevivência a fim de se obter resultados mais confiáveis.

Para trabalhos futuros, serão feitas análises dos dados sobre crimes na região metropolitana de Belo Horizonte, com base na modelagem *exponencial*. Seria interessante também desenvolver estudos simulados para avaliar o comportamento do modelo diante de diferentes cenários.

## *Apêndice*

Tabela 6.1: Testes de convergência para as cadeias do parâmetro do Modelo M4 para dados de câncer

	Heidelberg and Welch- $P_{valor}$	Geweke( $Z_{Observados}$ )
$\alpha$	0.9365	-0.1030
int	0.1607	0.7793
sexo	0.2342	0.6018
raça	0.0555	-1.4311
idade	0.0849	1.0375
n° primárias	0.6115	0.9888
$\phi$	0.6811	-1.3597

Tabela 6.2: Testes de convergência para as cadeias do parâmetro do Modelo M4 para dados de crimes

	Heidelberg and Welch- $P_{valor}$	Geweke( $Z_{Observados}$ )
$\alpha$	0.6252	0.6655
intercepto	0.8262	0.4382
sexo	0.9470	-0.2288
idade	0.2573	-1.2427
lugar	0.8759	-0.4924
consumado	0.5964	-1.1074
$\phi$	0.9418	-0.5523

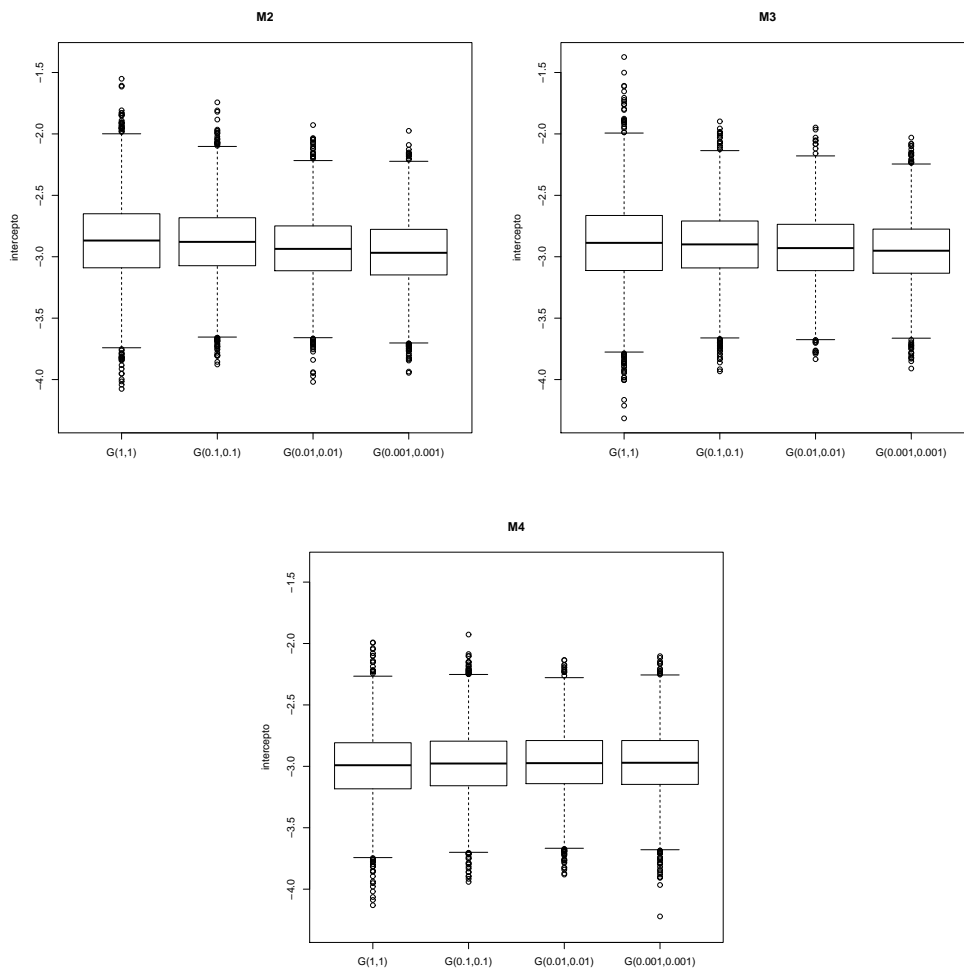


Figura 6.1: *Boxplots* da distribuição *a posteriori* do *intercepto* para dados de câncer

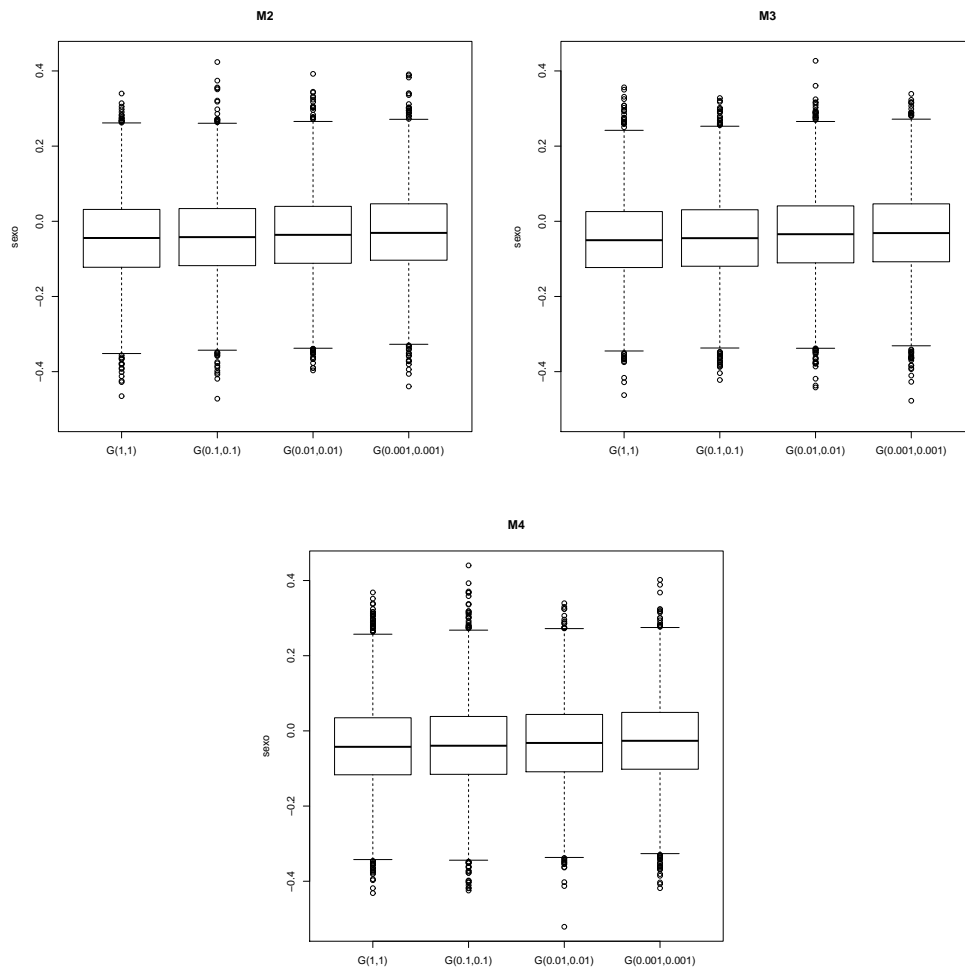


Figura 6.2: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *sexo* para dados de câncer



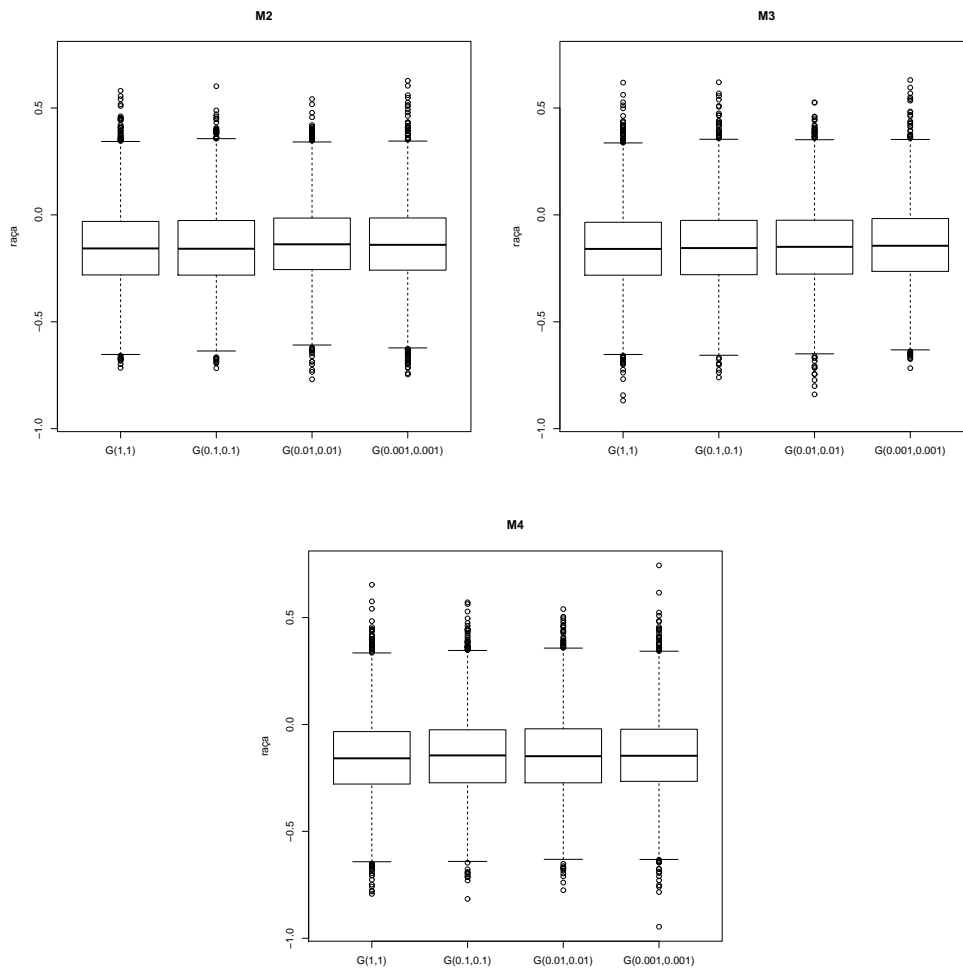


Figura 6.3: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *raça* para dados de câncer

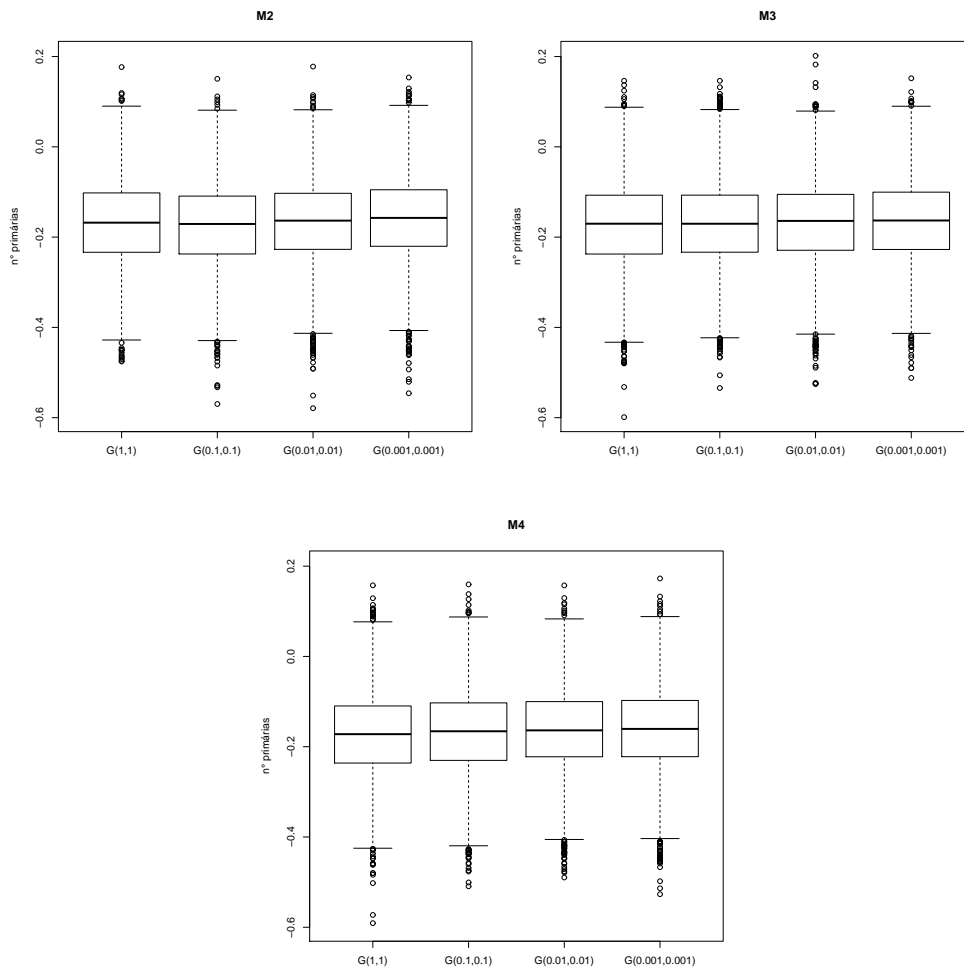


Figura 6.4: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *n*<sup>o</sup> de tumores primários para dados de câncer

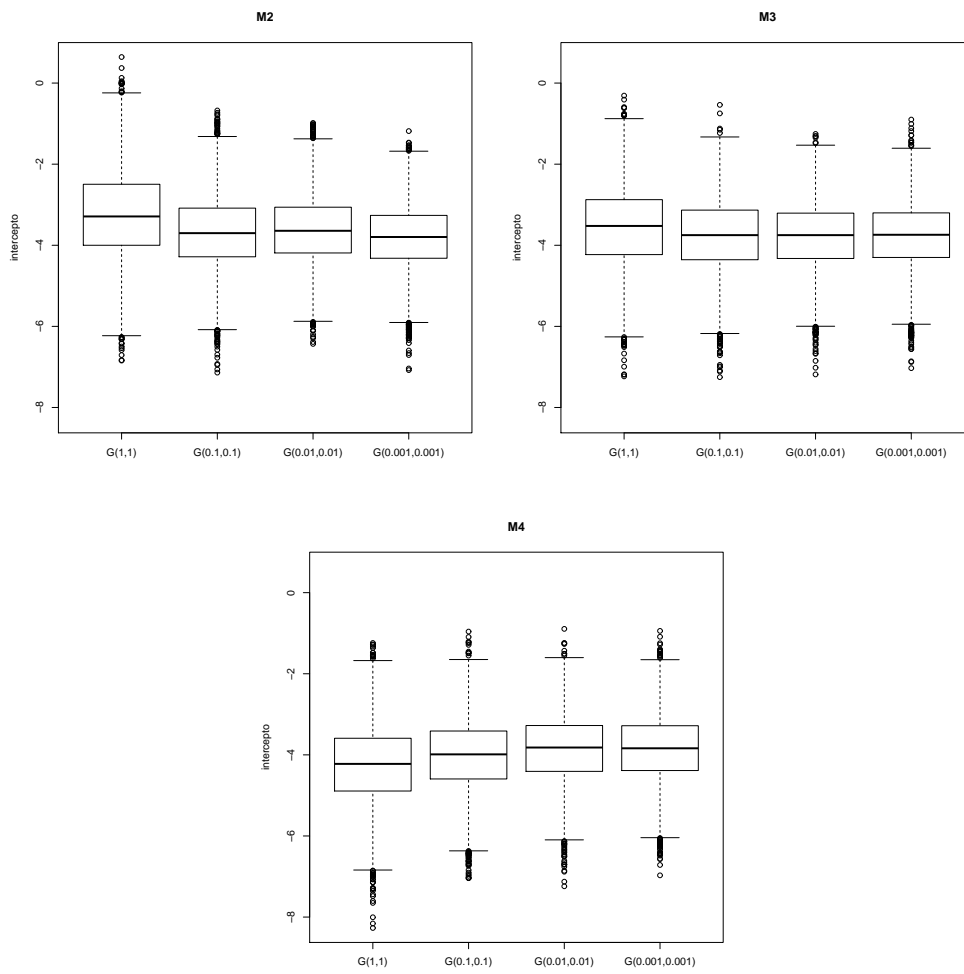


Figura 6.5: *Boxplots* da distribuição *a posteriori* do *intercepto* para dados de crimes

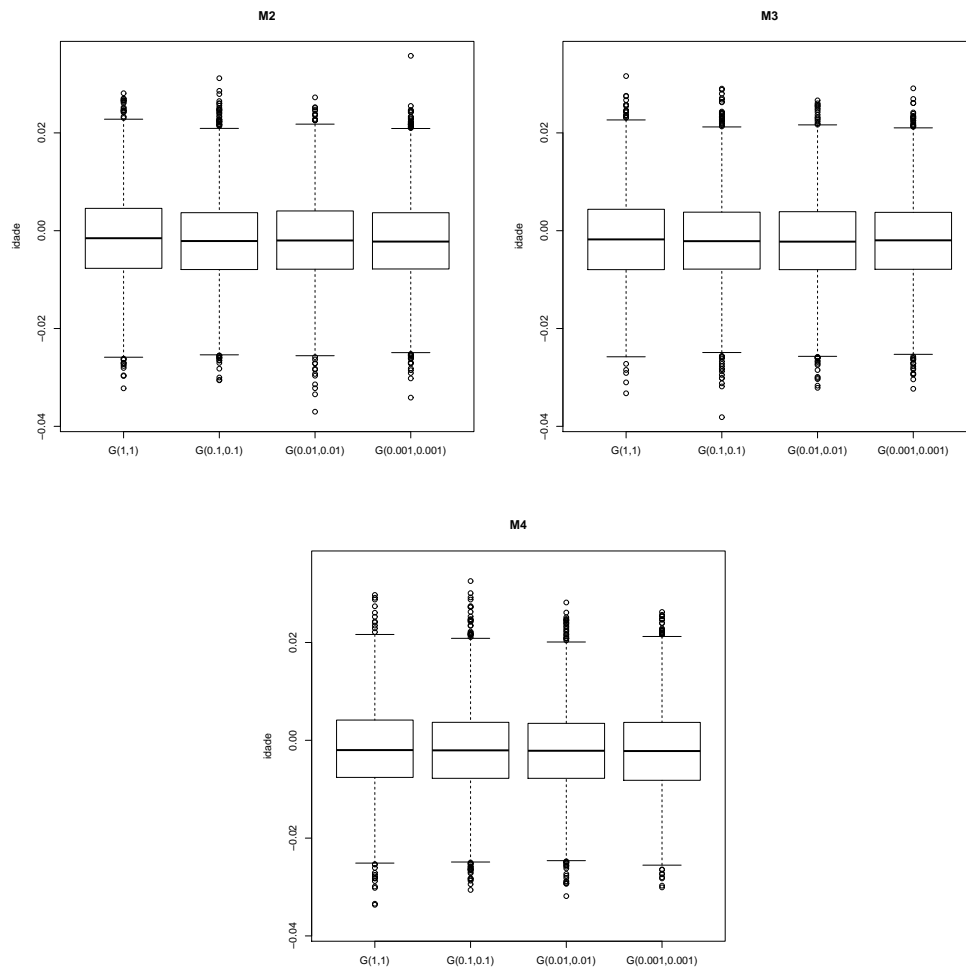


Figura 6.6: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *idade* para dados de crimes

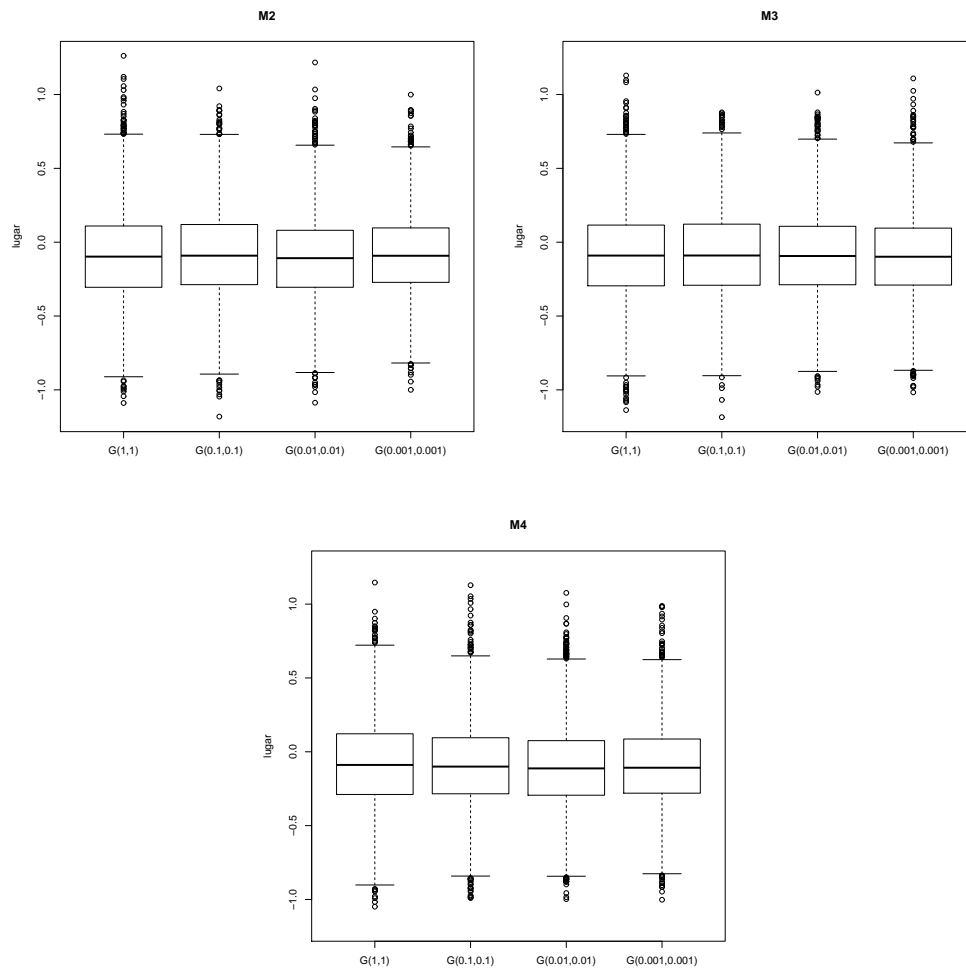


Figura 6.7: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *lugar* para dados de crimes

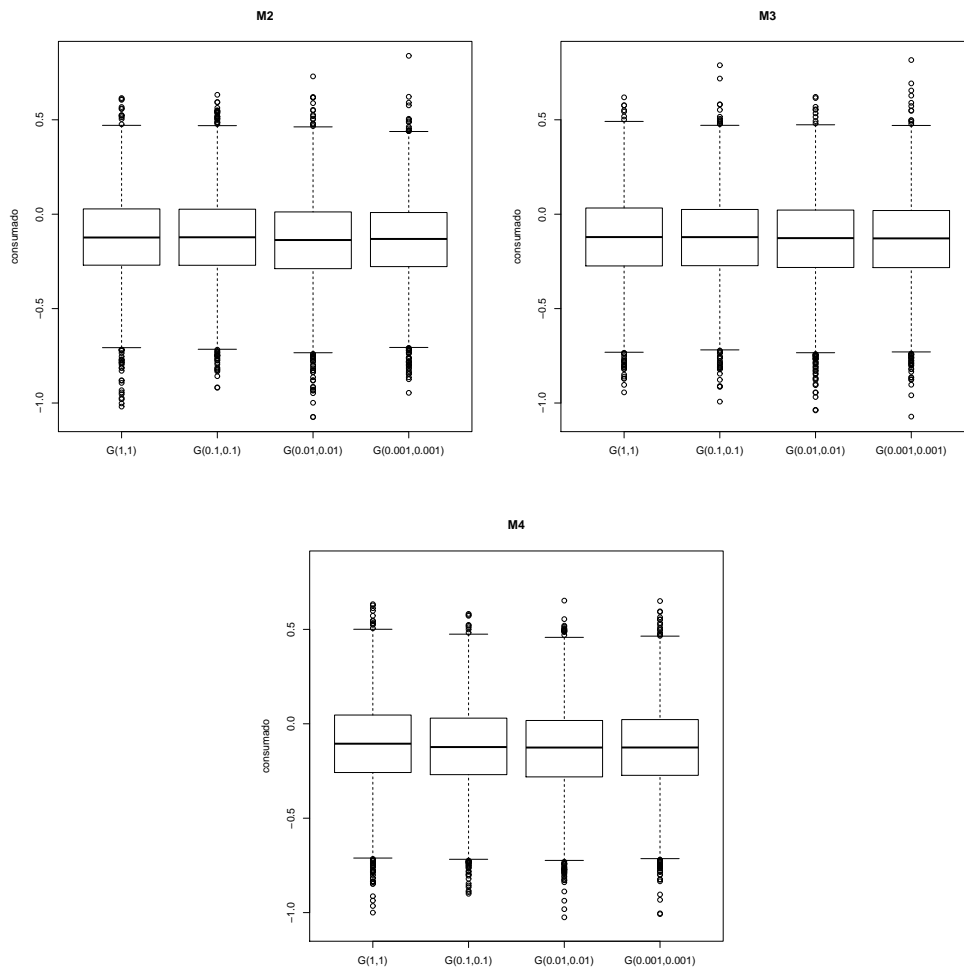


Figura 6.8: *Boxplots* da distribuição *a posteriori* do coeficiente de regressão associado à covariável *consumado* para dados de crimes

## *Referências Bibliográficas*

- BANERJEE, S.; CARLIN, B. P. Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, Blackwell Publishing, v. 60, n. 1, p. 268–275, 2004. ISSN 1541-0420. Disponível em: <<http://dx.doi.org/10.1111/j.0006-341X.2004.00032.x>>.
- BANERJEE, S.; CARLIN, B. P.; GELFAND, A. *Hierarchical Modeling and Analysis For Spatial Data*. [S.l.]: Chapman & Hall/CRC, 2004.
- BANERJEE, S.; WALL, M. M.; CARLIN, B. P. Frailty modeling for correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, v. 4,1, p. 123–142, 2003.
- BASTOS, L.; GAMERMAN, D. Dynamic survival models with spatial frailty. *Lifetime Data Analysis*, Kluwer Academic Publishers-Plenum Publishers, v. 12, n. 4, p. 441–460, 2006. ISSN 1380-7870. Disponível em: <<http://dx.doi.org/10.1007/s10985-006-9020-2>>.
- BESAG, J. Spatial interation and the staistical analysis of lattice systems ((with discussion). *J. Roy. Statist. Soc., Ser. B*, 36, p. 192–236, 1974.
- BESAG, J.; YORK, J.; MOLLIE, A. Bayesian image restoration, with two applications in spatial statistics(with discussion). *Ann. Inst. Statist. Math.*, v. 43, ., p. 1–59, 1991.
- CARVALHO, M. S. *et al. Análise de Sobrevidência. Teoria e aplicações em saúde*. [S.l.]: Fiocruz, 2011.
- CLAYTON, D. G. A model for association in bivariate life tables and its aplications in epidemiology studies for familial tendency in chronic disease incidence. *Biometrika*, v. 64(1), p. 141–151, 1978.
- CLAYTON, D. G.; CUZICK, J. Multivariate generalizations of the proportional hazards model (with discussion). *Journal of Royal Statistical Society, Series, Ser. A*, 148, p. 82–117, 1985.
- COONER, F.; BANERJEE, S.; MCBEAN, A. M. Modelling geographically referenced survival data with a cure fraction. *Statistical Methods in Medical Research*, v. 15, n. 4, p. 307–324, 2006. Disponível em: <<http://smm.sagepub.com/content/15/4/307.abstract>>.
- COWLES, M. K.; CARLIN, B. P. Markov chain monte carlo convergence diagnostics: A comparative review. *American Statistical Association*, v. 91, p. 883–904, 1996.
- COX, D. R. Regression models and life tables (with discussion). *Journal Royal Statistical Society, Series, B*, 34, p. 187–202., 1972.
- DIVA, U.; BANERJEE, S.; DEY, D. K. Modelling spatially correlated survival data for individuals with multiple cancers. *Statistical Modelling*, v. 7, n. 2, p. 191–213, 2007. Disponível em: <<http://smj.sagepub.com/content/7/2/191.abstract>>.

- DIVA, U.; DEY, D. K.; BANERJEE, S. Parametric models for spatially correlated survival data for individuals with multiple cancers. *Statistics in Medicine*, John Wiley & Sons, Ltd., v. 27, n. 12, p. 2127–2144, 2008. ISSN 1097-0258. Disponível em: <<http://dx.doi.org/10.1002/sim.3141>>.
- GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions and bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 6, p. 721–741, 1984.
- GEWEKE, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Oxford University Press, v. 4, p. 169–193, 1992.
- GIOLO, S. R.; COLOSIMO, E. A. *Análise de Sobrevida aplicada*. São Paulo: Edgard Blucher, 2006.
- HEIDELBERGER, P.; WELCH, P. Simulation run length control in the presence of an initial transient. *Operations Research*, v. 31, p. 1109–1144, 1983.
- HOUGAARD, P. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, v. 71, p. 75–83., 1984.
- HOUGAARD, P. *Analysis of multivariate survival data*. [S.l.]: Statistical for biology and health, Library of Congress Cataloging-in-Publication Data, 2000.
- KLEIN, J. P.; MOESCHBERGER, M. L. *Survival Analysis: techniques for censored and truncated data*. [S.l.]: Statistical for biology and health, Library of Congress Cataloging-in-Publication Data, 1997.
- LANCASTER, T. *Econometric methods for the duration of unemployment*. [S.l.]: The Econometric Society, 1979.
- LAWSON, A. B. *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*. [S.l.]: CRC Press, 2008.
- LEE, E.; WANG, J. W. *Statistical Methods for Survival Data Analysis*. New Jersey: John Wiley and Sons, 2003.
- MIGON, H. S.; GAMERMAN, D. *Statistical Inference An Integrated Approach*. [S.l.]: Oxford University Press Inc, 1999.
- PAN, C. *et al.* Bayesian semiparametric model for spatially correlated interval-censored survival data. *Computational Statistics & Data Analysis*, n. 0, p. –, 2013. ISSN 0167-9473. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167947313004544>>.
- ROCHA, C. S. Modelos com fragilidade: aplicação à modelagem da heterogeneidade não observada. *Boletim*, v. 4, p. 26–34, 2011.
- SPIEGELHALTER, D. J. *et al.* Bayesian measures of model complexity fit (pkg: P583-639). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, v. 64, p. 583–616, 2002.



VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on dynamics of mortality. *Demography*, v. 16, p. 439–454, 1979.

VAUPEL, J. W.; YASHIN, A. I. *The Deviance Dynamics of Death in Heterogeneous Populations*. [S.l.]: RR-83-1, Luxemburg, Austria: International Institute for Applied System Analysis, 1983.

WHITMORE, G. A.; LEE, M. L. T. A multivariate survival distribution generated by an inverse gaussian mixture of exponentials. *Technometrics*, v. 33, p. 39–50, 1991.