UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓSGRADUAÇÃO EM ESTATÍSTICA

MODELAGEM DE DADOS DE SOBREVIVÊNCIA VIA MODELO DE RISCO LOGÍSTICO GENERALIZADO

Caroline Pires Cremasco

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos como parte dos requisito necessários para obtenção do título de Mestre em Estatística

Orientador: Prof. Dr. Francisco Louzada-Neto

São Carlos
2005

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C915md

Cremasco, Caroline Pires.
Modelagem de dados de sobrevivência via modelo de risco logístico generalizado / Caroline Pires Cremasco. -- São Carlos : UFSCar, 2008.
70 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2005.

1. Estatística matemática. 2. Análise de sobrevivência (Biometria). 3. Modelo de Cox. 4. Função de risco. 5. Regressão logística. I. Título.

CDD: 519.5 (20ª)

Dedico esta monografia à minha família,
em especial aos meus pais, por tudo o que fizeram para que este dia,
que sela meu mestrado, se tornasse
realidade.

Agradecimentos

À minha família por toda a força ao longo de todos esses anos, pela confiança no meu potencial e que sempre dividiram os momentos mais felizes e outros nem tanto.

A Camila Pires Cremasco por toda dedicação e ajuda que me deu no ínicio da minha vida profissional.

Ao meu orientador Francisco Lozada Neto, por dividir seu conhecimento e sabedoria e ao mesmo tempo confiar na minha capacidade de dar este resultado, pelas direções certas ao longo do projeto.

Aos professores ... ... pelas sugestões dadas em meu exame de qualificação.

Aos amigos e colegas, com quem sempre tenho aprendido.sobre todas as coisas.

"Se o Senhor não edificar a casa, em vão trabalham os que a edificam; se o Senhor não guardar a cidade, em vão vigia a sentinela. Inútil vos será levantar de madrugada, repousar tarde, comer o pão que penosamente granjeastes; aos seus amados ele o dá enquanto dormem."

Salmos 127:1,2.

Resumo

A modelagem de dados de sobrevivência com a presença de covariáveis por meio da função de risco tem sido cada vez mais utilizada devido a facilidade de interpretação Um dos exemplos mais importantes de modelos de risco é o modelo de riscos proporcionais proposto por Cox (1972). No entanto, este modelo supõe a proporcionalidade entre as funções de risco para duas ou mais covariáveis. Para acomodar situações em que o modelo de riscos proporcionais não é adequado, vários tipos de modelos não-proporcionais estão sendo desenvolvidos, como o modelo de falha acelerada, proposto por Prentice (1978), o modelo de risco híbrido de Etezadi-Amoli e Ciampi (1987) e os modelos de risco híbrido generalizados de Louzada-Neto (1997 e 1999). Neste trabalho exploramos um uma nova família paramétrica de modelo de risco não-proporcional dependente do tempo (McKenzie, 1999). Este modelo é baseado na generalização da função logística usual e é motivado, em parte, pela necessidade de se considerar o efeito do tempo na modelagem, e, em parte, pela preferência em se considerar uma estrutura paramétrica para a função de risco. Vários procedimentos inferenciais relacionados a esta nova família de modelos são apresentados.

.

Abstract

The modeling of data of survival with the presence of covariáveis by means of the risk function has been each used time more had the easiness of interpretation One of the examples most important of risk models is the model of proportional risks considered by Cox (1972). However, this model assumes the proportionality enters the risk functions in different levels of the covariáveis. To accomodate situations where the model of proportional risks is not adjusted, some types of not-proportional models are being developed, as the model of sped up imperfection, considered for Prentice (1978), the model of hybrid risk of Etezadi-Amoli and Ciampi (1987) and the generalized models of hybrid risk of Louzada-Neto (1997 and 1999). In this work we explore an one new family parametric of model of dependent not-proportional risk of the time (McKenzie, 1999). This model is based on the generalization of the usual logistic function and is motivated, in part, for the necessity of if considering the effect of the time in the modeling, and, in part, for the preference in if considering a parametric structure for the risk function. Some inferenciais procedures related this new family of models are presented.

Sumário

Capítulo 1

Introdução

1.1 Motivação

Vários autores têm preferido modelar dados de sobrevivência na presença de covariáveis por meio da função de risco, fato este, relacionado à sua interpretação. Neste contexto, um dos modelos mais utilizados é o modelo de Cox que tem a limitação de somente acomodar dados com funções de risco proporcionais. Contudo a experiência mostra que vários dados não podem ser acomodados pelo modelo de Cox. Este fato tem sido determinante no desenvolvimento de vários tipos de modelos de riscos não-proporcionais. Entre eles pode-se citar o modelo de falha acelerada (Prentice, 1978), o modelo de risco híbrido (Etezadi-Amoli e Ciampi, 1987) e os modelo de risco híbridos generalizados (Louzada-Neto, 1999).

Neste contexto, Mackenzie (1996) propôs uma nova família paramétrica de modelo

de risco não-proporcional intitulado modelo de risco logístico generalizado dependente do tempo. O modelo é baseado na generalização da função logística usual e é motivado, em parte, por considerar o efeito do tempo em seu ajuste e, em parte, pela necessidade de considerar estrutura paramétrica.

Neste Capítulo, na Seção 1.2, é apresentado uma breve introdução sobre análise de sobrevivência, bem como alguns conceitos introdutórios sobre modelagem de dados via função de risco. O modelo de Cox, um dos mais utilizados na literatura, é descrito na Seção 1.3, com suas propriedades e limitações. Para finalizar, alguns conceitos básicos da função logística usual são descritos resumidamnete na Seção 1.4.

1.2 Conceitos Básicos Análise de Sobrevivência

A Análise de sobrevivência é uma das áreas da estatística na qual são desenvolvidos métodos para analisar dados provenientes de variáveis aleatórias tomando valores positivos. Essas variáveis correspondem ao tempo de ocorrência de um evento de interesse, tais como o tempo até a morte ou cura de um indivíduo, tempo até a falha de um componente eletrônico, dentre outros.

Existem duas características bastante importantes nos dados de sobrevivência. Uma, diz respeito à distribuição dos dados que, geralmente, é assimétrica à direita e a outra é a presença de censuras, que ocorre quando não é possível observar a resposta de algumas unidades que deverão ser incluídas na análise.

O principal interesse na área de sobrevivência está na estimação da probabilidade de

um indivíduo sobreviver até o tempo t, conhecida como função de sobrevivência e a razão instantânea de falha no tempo t, dado que ele sobreviveu até t, chamada de função de risco.

Dentre as técnicas utilizadas para estimar-se a função de sobrevivência destacam-se as paramétricas e as não-paramétricas. Nas técnicas não paramétricas o estimador mais utilizado para a função de sobrevivência é o produto-limite de Kaplan-Meier (Kaplan & Meier, 1958). Um outro estimador não-paramétrico para a função de sobrevivência é a tabela de vida ou o estimador atuarial, que considera os dados agrupados, ou seja, são conhecidos apenas os intervalos em que as falhas ou censura ocorreram.

Estas técnicas não levam em consideração as covariáveis relacionadas com o tempo de sobrevivência. Para considerar estas covariáveis devem-se utilizar os modelos paramétricos ou teste de vida acelerado, em que se supõe uma distribuição de probabilidade conhecida para os tempos.

O tempo de sobrevivência T tem função densidade de probabilidade, $f(t)$, definida como a probabilidade de o indivíduo falhar no intervalo $[t; t + \Delta t]$ por unidade de tempo, que pode ser expressa por,

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}. \tag{1.1}$$

A função de sobrevivência, denotada por $S(t)$, representa a probabilidade do indivíduo não falhar até um certo tempo t, isto é,

$$S(t) = P(T \geq t) = 1 - F(t), \tag{1.2}$$

tal que $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$, e $F(t) = \int_0^t f(u)du$ representa a função de distribuição acumulada. Esta função é geralmente utilizada para determinar o *p-ésimo* percentil do tempo de sobrevivência.

A função de risco é definida como o limite da probabilidade de um indivíduo falecer no intervalo de tempo $[t, t + \Delta t)$, dado que o mesmo tenha sobrevivido até o tempo t, e é expressa por,

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

Esta função também pode ser definida em termos de (1.1) e (1.2) por meio da expressão

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.3}$$

1.2.1 Modelagem Via Função de Risco

Considerando a presença de variáveis explicativas (covariáveis), a função de risco (1.3) pode ser reescrita na forma

$$h\left(t|\mathbf{x}\right) = u\left(t, \boldsymbol{\alpha}'\mathbf{x}\right), \tag{1.4}$$

onde $u\left(\cdot\right)$ é uma função positiva igual a 1 quando seu argumento é zero, $\mathbf{x}^{'} = (x_1, ..., x_p)$ representa o vetor de covariáveis e $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)$ é o vetor de coeficientes do modelo a

serem estimados. Várias formas funcionais podem ser empregadas para $u(\cdot)$. Estas formas permitem a composição de vários tipos de funções de riscos que serão vistas em detalhes a seguir.

Para descrever a heterogeneidade existente numa determinada população pode-se modelar o tempo de vida considerando também as possíveis variáveis explicativas quem podem influenciar na sobrevivência, supondo uma forma para a função de risco básica. Existe uma vasta literatura na modelagem de funções de risco que consideram o efeito de covariáveis. O primeiro modelo de risco proposto e também um dos mais utilizados na literatura é apresentado na próxima Seção.

1.3 Modelo de Cox

O modelo de Cox é sem dúvida um dos mais populares na análise de dados de sobrevivência. Assume-se, nesse modelo, que os tempos $t_i,\ i = 1; ...; p$, são independentes e que o risco é dado por,

$$h\left(t|\mathbf{x}\right) = \exp(\boldsymbol{\alpha}'\mathbf{x})h_0\left(t\right), \tag{1.5}$$

em que h_0 é conhecida como a função de risco base, ou seja, o risco de um indivíduo com covariáveis $\mathbf{x} = \mathbf{0}$, $\boldsymbol{\alpha}$ é o vetor de dimensão p de coeficientes de regressão desconhecidos é $\mathbf{x}$ é o vetor de dimensão p de covariáveis observadas.

Devido a presença do componente não-paramétrico $h_0(t)$ no modelo (1.5), o método da máxima verossimilhança usual não é apropriado para estimar os parâmetros. Desse

modo, Cox(1975) formalizou o método da máxima verossimilhança parcial que consiste em condicionar a função de verossimilhança nos tempos de ocorrência do evento de modo a eliminar a função $h_0(t)$. Além disso, quando acontecem empates não se pode utilizar a função de verossimilhança parcial exata e, nesse caso, existem métodos alternativos tais como os apresentados por Breslow (1974) e Efron (1977).

O modelo dado por (1.5) é denominado modelo de riscos proporcionais pelo fato da razão entre as taxas de falhas para dois indivíduos ser constante no tempo, ou seja, tomando-se a razão de riscos para dois indivíduos i e j e aplicando-se a equação (1.5), obtém-se o resultado abaixo, que não depende do tempo,

$$\frac{h\left(t|\mathbf{x}_i\right)}{h\left(t|\mathbf{x}_j\right)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)h_0\left(t\right)}{\exp(\boldsymbol{\beta}'\mathbf{x}_j)h_0\left(t\right)} = \exp(\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\beta}'\mathbf{x}_j).$$

A suposição básica para o uso do modelo de riscos proporcionais de Cox é, portanto, que as taxas de falha sejam proporcionais. Em outras palavras, isto significa que se um indivíduo apresenta no início do estudo um risco de falha igual a três vezes o risco de um outro indivíduo, então esta taxa será a mesma para qualquer tempo t do período de acompanhamento.

1.4 Conteúdo da Dissertação

O objetivo desta dissertação é discutir o modelo não proporcional apresentado por Mackenzie (1996), apontando suas propriedades e motivações para seu uso. Nesta dissertação apresentamos, no Capítulo 2, os procedimentos clássicos de estimação pontual e

intervalar. Esta análise clássica é baseada em propriedades assintóticas dos estimadores dos parâmetros envolvidos. Sendo assim, também foi realizado alguns estudos numéricos das estimativas clássicas também descritos neste Capítulo.

Uma metodologia de estimação intervalar e a construção de testes de hipóteses para os parâmetros do modelo logístico generalizado dependente do tempo via reamostragem são apresentadas no Capítulo 3. Assim como no Capítulo 2, também foi estudada as propriedades destes estimadores através de um estudo numérico.

A metodologia Bayesiana de estimação para os parâmetros do modelo baseados em técnicas de simulação de Monte Carlo é apresentada no Capítulo 4.

No Capítulo 5 apresenta-se algumas conclusões finais e propostas de continuidade deste trabalho. Um Apêndice foi introduzido com o objetivo de apresentar métodos de diagnósticos de convergência para as cadeias geradas pelo método MCMC (apêndice A). Além disso, os programas desenvolvidos na implementação computacional para os exemplos de aplicação dos modelos propostos podem ser encontrados nos demais apêndices.

Capítulo 2

Modelo Logístico Generalizado Dependente do Tempo

2.1 Introdução

A experiência mostra que alguns dados de sobrevivência não comportam a suposição de riscos proporcionais conduzindo ao desenvolvimento de vários tipos de modelos de riscos não-proporcionais.

Neste contexto, o modelo proposto por MacKenzie (1996) considera uma nova família paramétrica de distribuições de sobrevivência contínuas para a análise de dados com riscos não-porporcionais. O modelo é baseado na generalização da familiar função logística considerando uma forma dependente do tempo, além disso, permite uma interpretação de fragilidade quando se tem uma população heterogênea.

A formulação do modelo e suas principais propriedades são apresentadas na Seção 2.2 e Seção 2.3, respectivamente. Na Seção 2.4, um procedimento de estimação pontual é desenvolvido através da estimação de máxima verossimilhança. Os testes de hipóteses assintóticos são apresentados na Seção 2.5. Para ilustração da metodologia apresentada neste capítulo, um exemplo numérico é desenvolvido na Seção 2.6 e alguns estudos numéricos são descritos na Seção 2.7. A Seção 2.8 apresenta algumas considerações finais, concluindo o Capítulo.

2.2 Formulação do Modelo

Denotando por T uma variável aleatória não negativa representando o tempo de falha, a taxa de falha instantânea, ou a função de risco tem a forma,

$$h(t|\lambda,\alpha,\boldsymbol{\beta}) = \lambda\frac{\exp(t\alpha + x'\beta)}{1+\exp(t\alpha + x'\beta)} \tag{2.1}$$

em que $\lambda > 0$ é um escalar, α é uma medida de efeito do tempo e $\boldsymbol{\beta}' = \left(\beta_1,...,\beta_p\right)'$ é um vetor de p parâmetros desconhecidos medindo a influência das p covariáveis $\mathbf{x}' = (x_1,...,x_p)'$.

Sabendo-se que , em que $S(t|\mathbf{x}) = \exp(-H(t|\mathbf{x}))$,em que $H(t|\mathbf{x})$ é a função de risco acumulada, tem-se que a função de sobrevivência é dada por,

$$S(t|\mathbf{x}) = \left[\frac{1+\exp(t\alpha + x'\beta)}{\exp(x'\beta)}\right]^{-\frac{\lambda}{\alpha}}$$

Então, a equação (2.1) caracteriza uma família tri-paramétrica dada por,

$$\begin{aligned} f(t|\lambda,\alpha,\beta) &= h(t|\lambda,\alpha,\beta)S(t|\lambda,\alpha,\beta) \\ &= \lambda\frac{\exp(t\alpha+x'\beta)}{1+\exp(t\alpha+x'\beta)}\left\{\frac{1+\exp(t\alpha+x'\beta)}{1+\exp(x'\beta)}\right\}^{-\lambda/\alpha} \end{aligned} \tag{2.2}$$

Para $T>0$, tem-se que $S(0|\lambda,\alpha,\beta)=1$ e correspondentemente $S(\infty|\lambda,\alpha,\beta)=0$.

Além disso, podemos notar também que o modelo (2.1) caracteriza uma família tri-paramétrica (λ,α,β) de variáveis aleatórias não negativas com função de probabilidade dada por,

$$f(t|\lambda,\alpha,\beta)=\lambda p(\alpha,\beta)\left\{q(\alpha,\beta)g(\beta)\right\}^{\lambda/\alpha} \tag{2.3}$$

sendo que os componentes individuais são funções simples do modelo logístico múltiplo dependente do tempo,

$$p(\alpha,\beta)=\exp(t\alpha+x'\beta)/\left\{1+\exp(t\alpha+x'\beta)\right\}, \tag{2.4}$$

$$q(\alpha,\beta)=1/\left\{1+\exp(t\alpha+x'\beta)\right\}, \tag{2.5}$$

$$g(\beta)=1+\exp(x'\beta). \tag{2.6}$$

Vários modelos de riscos usuais podem ser obtidos como casos particulares de (2.1). Quando, em (2.5), $q(\alpha,\beta)\approx 1$, o modelo logístico generalizado dependente do tempo (2.1)

se aproxima do modelo de Cox (1.5). Nesta condição as estimativas dos parâmetros de regressão $\boldsymbol{\beta}$ obtidas por meio dos dois modelos são similares. Quando $a = 0$, o modelo (2.1) reduz-se a um modelo de risco logístico múltiplo.

O modelo é incomum pois as covariáveis e o tempo são tratados simetricamente, os parâmetros α e β aparecem conjuntamente numa mesma estrutura. Então, os efeitos das covariáveis são ajustados para o tempo e vice-versa, permitindo que o efeito das covariáveis, medido ao risco básico, decresça com o tempo, então, quando $t \to \infty$, seu efeito é pequeno, um aspecto que geralmente está de acordo a realidade dos estudos médicos.

Além disso, quando $\lambda = 1$, uma nova família logística dependente do tempo é obtida, com função de risco dada por,

$$h^*(t|\alpha, \beta) = \frac{\exp(t\alpha + x'\beta)}{1 + \exp(t\alpha + x'\beta)}. \tag{2.7}$$

Considerando o modelo de risco proporcional, $\exp(x_i'\beta)$, tem uma interpretação de "risco relativo". Contudo, no modelo definido em (2.1), $\exp(x'\beta)$ tem uma interpretação "odds razão", que fica explícita quando $\lambda = 1$, em que surge um modelo logístico dependente do tempo reduzido com,

$$\psi^*(t|x) = \log\left\{\frac{h^*(t|x)}{1 - h^*(t|x)}\right\} = t\alpha + x'\beta.$$

Considerando o risco relativo, de acordo com a equação (2.1), a estimativa de $h(t|x_1)/h(t|x_2)$, para duas determinadas covariadas específicas, podem ser obtidas por,

$$\rho(\alpha, \beta|t, x_1, x_2) = \exp\left[(x_1 - x_2)'\beta\right] \frac{1 + \exp(t\alpha + x_2'\beta)}{1 + \exp(t\alpha + x_1'\beta)}$$

Se x é binária ($x_1 = 1$ e $x_2 = 0$), o principal termo do risco relativo é $\exp\beta$, contudo os risco não são proporcionais, como no modelo de Cox, mas dependentes do tempo.

2.3 Algumas Propriedades

A Figura 1.1 mostra a forma típica da densidade, função de risco e sobrevivência para um conjunto de parâmetros específicos ($\lambda = 1, \alpha = -0.05, \gamma = -1$), sendo que $\gamma = x'\beta$. Neste exemplo, a dependência do tempo é semanal, $\alpha = -0.05$ e a função de risco resultante é aproximadamente linear. Desta figura, nota-se que 50% da população falham antes 2.75 anos e 80% antes de 6.75 anos.

A Figura 1.2 mostra uma série de curvas de risco ilustrando a amplitude das formas que podem ser representadas quando $\lambda = 1$. Em geral, valores positivos de α aumenta a taxa de risco, enquanto que valores negativos a reduzem. Contudo, entre estes intervalos de risco é mais linear, de acordo com a Figura 1.2, quando $\alpha = -0.06$ e $\gamma = -1.6$.

Também, da Figura 1.2, tem-se que uma forte dependência do tempo, como é esperado, para um modelo sigmoidal, e é esta propriedade da função de risco que distingue o modelo logistico dependente do tempo dos demais modelos. Entretanto, sob algumas condições, o modelo logístico generalizado dependente do tempo pode ser aproximado para a distibuição de Gompertz. Por exemplo, na equação (2.1), quando $\frac{1}{1+\exp(t\alpha+\gamma)} \approx 1$, a função de risco é da forma $h_0(t|\alpha, \gamma, \lambda) \approx \lambda \exp(t\alpha + \gamma)$, isto é, segue uma distribuição

de Gompertz com parâmetros $\upsilon = \alpha$ e $\eta = \lambda \exp \gamma$.

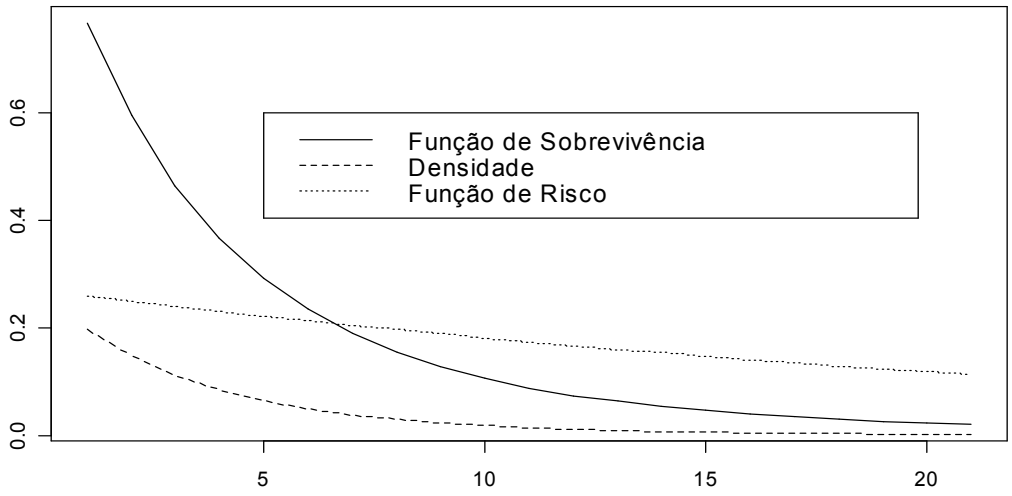

Figura 1.1. Modelo Logístico Generalizado (λ=1.00, α=-0.05,$\gamma = -1.00$)

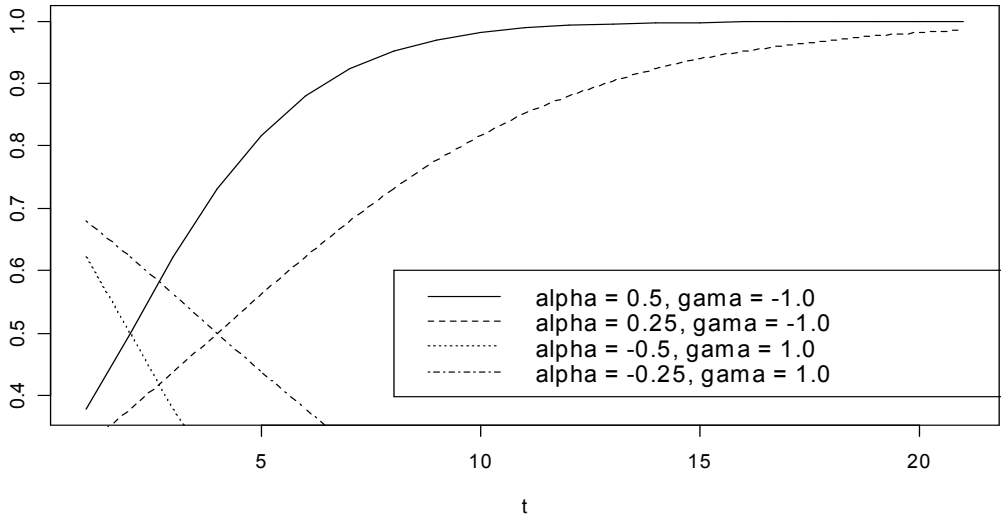

Figura 1.2. Funções de Risco.

Assim, podemos dizer que o modelo logístico generalizado dependente do tempo é uma

família de distribuição de sobrevivência que é diferente da distribuição logística padrão dada em Kalbfleisch e Prentice (1980) e da distribuição generalizada de Verhulst (Ahuja and Nash, 1967). Também não pode ser comparado com com o modelo log-logístico introduzido por Bennett (1983) em que a função de risco é crescente, para alguns valores dos parâmetros.

2.4 Procedimento de Estimação Pontual

Como o modelo logístico generalizado dependente do tempo definido em (2.1) é inteiramente paramétrico, o método de máxima verossimilhança pode ser usado tranquilamente para estimar os parâmetros desconhecidos. Considerando uma amostra aleatória de n indivíduos com as informações disponíveis (t_i, x_i, δ_i), em que $\delta_i = 1$, se o evento de interesse é observado e $\delta_i = 0$, caso contrário, para $i = 1, 2, ..., n$. Então, a contribuição para a verossimilhança dos eventos observados no tempo t_i é da forma,

$$\begin{aligned} L(\lambda, \alpha, \beta) &= \prod_{i=1}^{n} \{\lambda p_i(\alpha, \beta)\}^{\delta_i} \{q_i(\alpha, \beta) g_i(\beta)\}^{\lambda/\alpha} \\ &= \prod_{i=1}^{n} \left\{ \lambda \frac{\exp(t_i\alpha + x_i'\beta)}{1 + \exp(t_i\alpha + x_i'\beta)} \right\}^{\delta_i} \left\{ \frac{1 + \exp(x_i'\beta)}{1 + \exp(t_i\alpha + x_i'\beta)} \right\}^{\lambda/\alpha}. \end{aligned} \tag{2.8}$$

Pois, considerando um esquema de censura não informativo, a contribuição dos tempos censurados é,

$$\begin{aligned} S(t_i|x) &= \{q_i(\alpha,\beta)g_i(\beta)\}^{\lambda/\alpha} \\ &= \left\{\frac{1+\exp(x_i'\beta)}{1+\exp(t_i\alpha+x_i'\beta)}\right\}^{\lambda/\alpha}. \end{aligned}$$

Como $\lambda > 0$ e $\alpha > 0$, é conveniente escrever $\lambda = \exp\varphi$ e $\alpha = \exp\phi$, .para obtermos uma melhor convergência dos dados Considerando o log-verossimilhança, as primeiras derivadas, necessárias para obter os estimadores de máxima verossimilhança são dadas por,

$$\left.\begin{array}{l} U_\varphi = \frac{\partial l}{\partial \varphi} = \sum_{i=1}^n \left[\delta_i + \frac{\exp\varphi}{\exp\phi}\log(q_i(\phi,\boldsymbol{\beta})g_i(\boldsymbol{\beta})\right], \\ U_\phi = \frac{\partial l}{\partial \phi} = \sum_{i=1}^n \left[\delta_i t_i q_i(\phi,\boldsymbol{\beta}) - \frac{\exp\varphi}{\exp\phi}t_i p_i(\phi,\boldsymbol{\beta}) - \frac{\exp\varphi}{\exp(2\phi)}\log(q_i(\phi,\boldsymbol{\beta})g_i(\boldsymbol{\beta})\right], \\ U_{\beta_u} = \frac{\partial l}{\partial \beta_u} = \sum_{i=1}^n \left[\delta_i x_{ui} q_i(\phi,\boldsymbol{\beta}) + \frac{\exp\varphi}{\exp\phi}x_{ui}\left[ri(\boldsymbol{\beta}) - p_i(\phi,\boldsymbol{\beta})\right]\right], \end{array}\right\} \quad (2.9)$$

para $u = 0, ..., p$, e $ri(\boldsymbol{\beta}) = \exp(x_i\acute{}\boldsymbol{\beta})/(1+\exp(x_i\acute{}\boldsymbol{\beta}))$.

Além disso, a matriz de informação observada I, que fornece a consistência dos estimadores da matriz de dispersão assintótica, pode ser obtida através das equações dadas em (2.9). Considerando $\boldsymbol{\theta} = (\varphi, \phi, \boldsymbol{\beta})\acute{}$, então, a matriz de dispersão assintótica é $V(\boldsymbol{\theta}) = I(\boldsymbol{\theta})^{-1}$ e o estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ é assintoticamente $N_p(\boldsymbol{\theta}, V(\hat{\boldsymbol{\theta}}))$.

Como as derivadas (2.9) não apresentam soluções analíticas o método de Newton Raphson foi utilizado para este fim.

Devemos ressaltar também que o modelo (2.8) apresenta dificuldades referentes à estimação de máxima verossimilhança dos parâmetros de interesse. O problema surge porque o modelo é ajustado incluindo o vetor de covariáveis com termo intercepto, resultando em um modelo não identificável, pois λ e β_0 fornecem a mesma informação para os dados.

É importante notar que a constante λ deve estar restrita para valores positivos e o termo β_0 ser retirado, tornando o modelo dado em (2.1) identificável.

2.5 Procedimento de Estimação Intervalar

Nesta Seção estudaremos procedimentos de estimação intervalar bem como testes de hipóteses. Vale ressaltar que as estimativas intervalares e testes de hipóteses para os parâmetros são baseadas na distribuição normal assintótica dos estimadores de máxima verossimilhança (emvs) e na distribuição qui-quadrado da estatística da razão de verossimilhança (RV), respectivamente (Lawless, 1982, apêndice E). A adequação destes procedimentos para amostras pequenas ou moderadas é estudada via simulação mais adiante.

Seja $\hat{\boldsymbol{\theta}} = (\hat{\varphi}, \hat{\phi}, \hat{\beta})$ denotando os estimadores de $\boldsymbol{\theta} = (\varphi, \phi, \beta)$. Assintoticamente, quando $n \to \infty$,

$$\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sqrt{n} \longmapsto N\left(\boldsymbol{\theta}, \mathbf{I}^{-1}(\theta)\right), \tag{2.10}$$

em que N denota os distribuição normal multivariada e $\mathbf{I}$ é a matriz de informação esper-

ada de $\boldsymbol{\theta}$ em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. A matriz de informação observada em, $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, é dada por,

$$\mathbf{I}^{-1}(\hat{\theta}) = \left[-\frac{\partial \log L}{\partial \theta_l \partial \theta_s} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{2.11}$$

é um estimador consistente de $\mathbf{I}^{-1}(\theta)$.

A distribuição assintótica dada em (2.10) pode ser usada para construir regiões de confiança aproximadas para os parâmetros e para funções de interesse.

Para testar a variedade de hipóteses relacionadas com os parâmetros que surgem dos dados, pode-se usar a estatística de Wald, a estatística Score ou a estatística de razão de máxima verossimilhança (RV). Sendo que esta ultima foi considerada em nosso trabalho.

Seja $\boldsymbol{\theta}_0$ e $\boldsymbol{\theta}_1$ denotando os emvs de $\boldsymbol{\theta}$ sob as hipóteses H_0 e H_1. Sob certas condições de regularidade, a RV é dada por,

$$\Lambda = -2 \log \left(\frac{L(\theta_0)}{L(\theta_1)} \right), \tag{2.12}$$

tem uma distribuição assintótica χ^2_d, em que d é a diferença entre o número de parâmetros independentes necessários para especificar H_0 e H_1.

2.6 Exemplo de Aplicação

Kalbfleisch & Prentice (1980) fornecem um conjunto de dados que ilustra dados que consideram várias covariáveis para descrever o prognóstico de câncer pulmonar em 137 pacientes onde 9 são censurados à direita. Após investigações preliminares, optamos por

trabalhar com a variável de performance inicial - (Karnofsky, com escala 0-100, em que valores com scores altos representam pacientes com índice um "bom"índice de performance inicial) para a sobrevivência. Pacientes com performance abaixo de 50 foram classificados como sendo do grupo 1 e os pacientes acima de 50 do grupo 2. A Figura 2.1 mostra a função de risco acumulada e a Figura 2.2 fornece a curva da função de sobrevivência empírica para os dois grupos de performance inicial. Obervando-se as figuras podemos verificar claramente que as curvas se cruzam , concluindo que o modelo de risco proporcional não é apropriado.

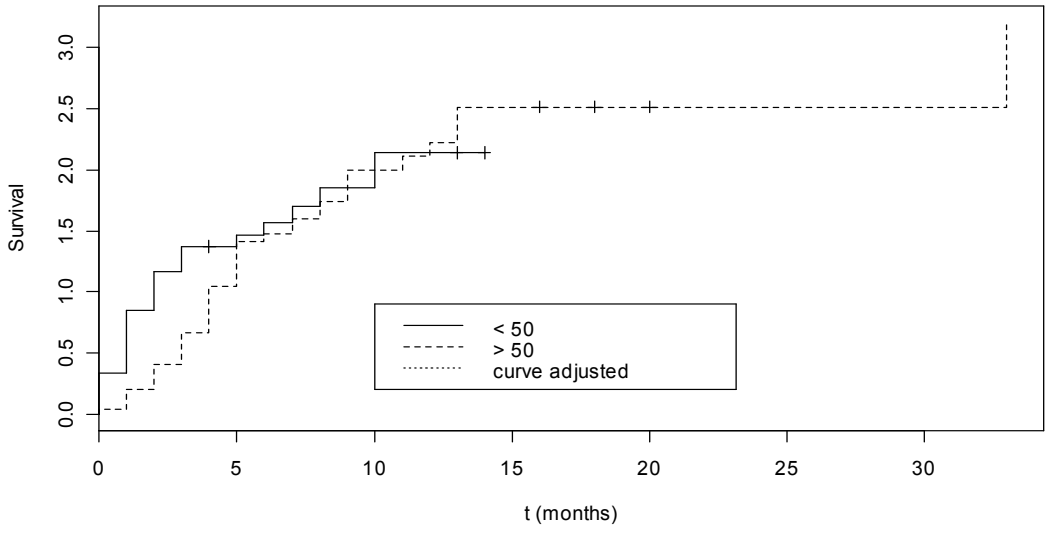

Figura 2.1 Função de Risco Acumulada

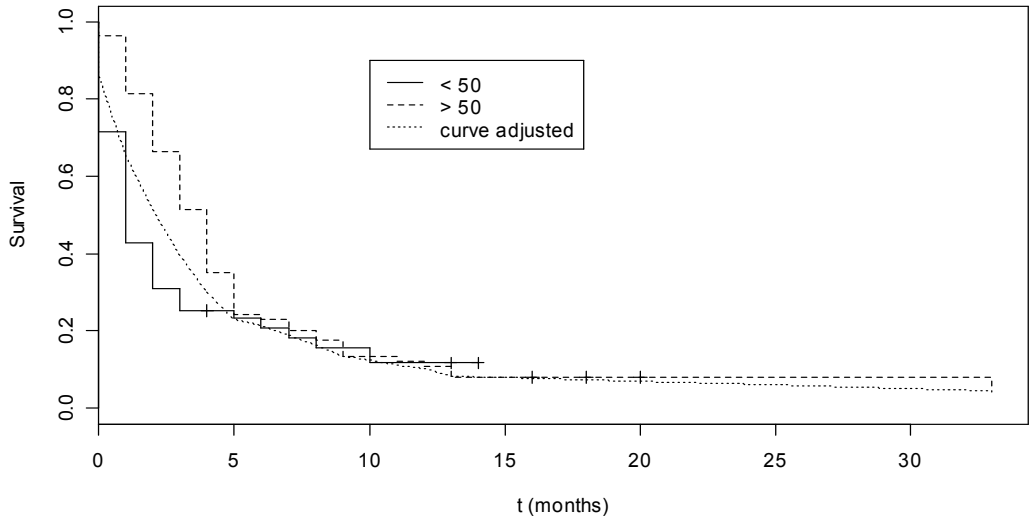

Figura 2.2 Função de Sobrevivência (S(t)), com as estimativas de Kaplan-Meier e o Ajuste do Modelo Logístico Generalizado Dependente do Tempo (—).

As estimativas de máxima verossimilhança foram calculadas considerando o modelo logístico generalizado dependente do tempo (2.8) através do método de Newton-Rapshon. Advertimos que, neste trabalho, consideramos $\lambda > 0$ e $\alpha > 0$ optamos por trabalhar com $\lambda = \exp(\varphi)$ e $\alpha = \exp(\phi)$, para obter uma convergência do método numérico.

Para a implementação do método iterativo de Newton-Raphson foi utilizada a linguagem de programação *R*. Iniciando o método com $\varphi^{(0)} = \phi^{(0)} = \beta^{(0)} = 0$, as estimativas encontradas através das sucessivas aproximações são mostradas na Tabela 2.1. Esta tabela também contém o desvio padrão (DP) e o intervalo de confiança assintótico (IC) das respectivas estimativas.

Tabela 2.1. Estimativas para os Parâmetros do Modelo Logístico Generalizado ajustado para os Dados de Câncer Pulmonar.

Parâmetro	*Estimativa*	DP	$IC(90\%)$	
φ	0.965	0.01	0.95	0.98
ϕ	0.612	0.03	0.55	0.67
β	-0.971	0.09	-1.15	-0.79

A matriz de variância-covariância $[I(\boldsymbol{\theta})]^{-1}$ é dada, neste caso, por

$$[I(\boldsymbol{\theta})]^{-1} = \begin{bmatrix} 0.01 & -0.20 & 0.00 \\ -0.20 & 0.03 & -0.05 \\ 0.00 & -0.05 & 0.09 \end{bmatrix}.$$

Através da Tabela 2.1, podemos concluir que o valor altamente significante de $\hat{\phi}$ indica que o modelo de risco proporcional (1.5) não é apropriado. Portanto, para este conjunto de dados, o modelo Logístico Generalizado Dependente do Tempo, (2.1), fornece um melhor ajuste.

2.7 Alguns Estudos Numéricos

É comum na prática encontrarmos estudos com a possibilidade de amostras pequenas ou moderadas. Para avaliar a aplicabilidade dos resultados assintóticos nestes casos, as propriedades do estimadores devem ser estudadas através de cálculos numéricos baseado na matriz de informação de Fisher. Para isto, y_{is} e os c_{is} foram gerados de uma distribuição exponencial com $n = 30, 50, 100, 300$ e 1000 e taxa de falha igual a 0.4 e 0.5, respectivamente. Os tempos de vida foram definidos como $t_i = \min(y_i, c_i)$ e a censura especificada como $\delta_i = 1$, se $y_i < c_i$ e $\delta_i = 0$, caso contrário. As covariáveis foram geradas como sendo variáveis de Bernoulli com probabilidade de sucesso igual a 0.5.

A adequabilidade foi estudada via simulação Monte Carlo considerando 1000 réplicas dos intervalos de confiança de 90% para os parâmetros do modelo estudado. Além disso, os comprimentos médios dos intervalos de confiança devem decair com o aumento de n, proporcionalmente a $n^{-1/2}$. A verificação deste decaimento assintótico foi feita rea lizando a regressão $log(var(.))$ versus $log(n)$, onde os coeficientes de inclinação de regressão devem ser aproximadamente -1. Os resultados do estudo são apresentados na Tabela 2.2.

As probabilidades de cobertura dos intervalos de confiança de 90% são apresentadas

na Tabela 2.3.

Tabela 2.2. Coeficiente de Inclinação da relação entre o *log* da variância e n.

n	$var(\varphi)$	$var(\phi)$	$var(\beta)$
[15, 30]	-1.401	-0.965	-1.468
[30, 50]	-1.131	-1.096	-1.305
[50, 100]	-0.974	-1.000	-1.160
[100, 300]	-0.946	-0.939	-1.22
[300, 1000]	-0.992	-0.998	-0.938

Tabela 2.3. Probabilidade de Cobertura dos Intervalos de Confiança de 90% para os Parâmetros do Modelo Parametrizado.

n	$\hat{\varphi}$	$\hat{\phi}$	$\hat{\beta}$
15	0.781	0.762	0.833
30	0.887	0.874	0.898
50	0.891	0.889	0.872
100	0.913	0.921	0.893
300	0.899	0.881	0.911
1000	0.933	0.905	0.900

De acordo com a Tabela 2.3, podemos notar que a probabilidade de cobertura dos intervalos de confiança de 90% para os parâmetros φ e ϕ são razoáveis para $n \geq 100$, decrescendo em torno de 0.80 quando $n = 15$. Entretanto, para o parâmetro β podemos

concluir que quando $n \geq 30$ as probabililidades de cobertura já se encontram num nível razoável. Além disso, os comprimentos médios dos intervalos de confiança decaem com o aumento de n na razão de $n^{-1/2}$. A Figura 2.3 apresenta as probabilidades de cobertura versus o tamanho da amostra para os intervalos de confiança dos parâmetros do modelo.

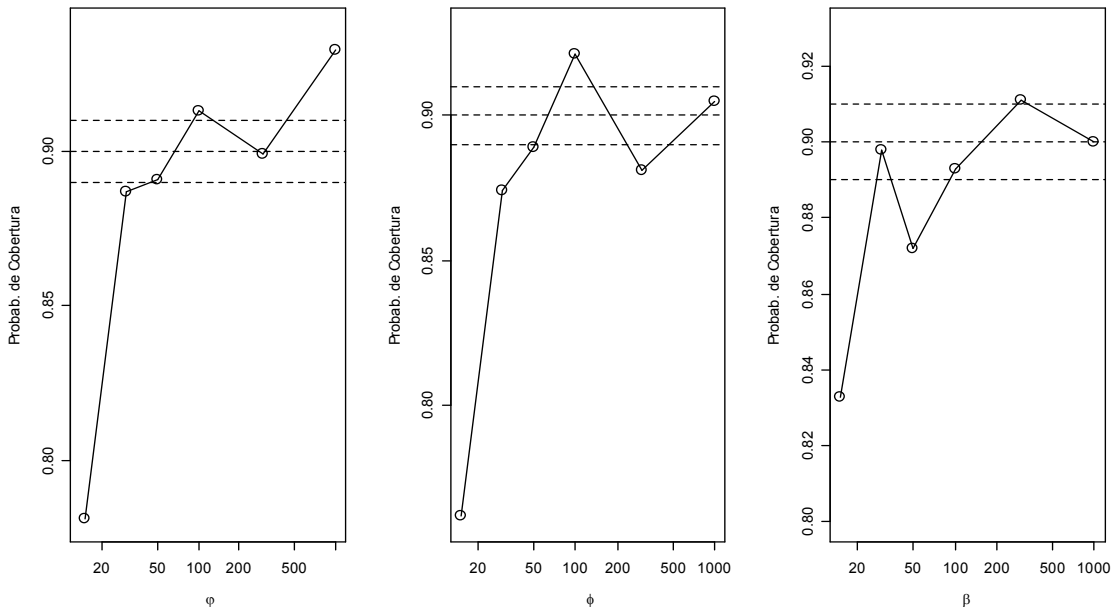

Figura 2.3. Probabilidade de Cobertura dos Intervalos de 90% versus o Tamanho da Amostra, n. Os valores 0.89 e 0.91 correspondem respectivamente aos imites inferiores e superiores do intervalo de 90% da probabilidade de cobertura de 0.9.

2.8 Considerações Finais

Neste Capítulo apresentamos um novo modelo paramétrico para descrever os dados de sobrevivência proposto por Mackenzie em 1996 e intitulado de "Modelo Logístico

Generalizado Dependente do Tempo".

A vantagem desse modelo de risco logístico generalizado dependente do tempo é permitir que o tempo seja considerado no ajuste do modelo. Entretanto, como na análise de sobrevivência é muito comum encontrarmos amostras pequena ou moderada, uma cautela é necessária nos procedimentos de estimação e testes de hipóteses, uma vez que a teoria assintótica usual pode não ser válida.

No estudo de simulação desenvolvido neste Capítulo, observamos que a teoria assintótica não deve ser adotada para pequenas amostras, pois para certos parÂmetros, algumas estimativas podem não estarem adequadas. Além disso, métodos formais de testes de hipóteses e estimação por intervalo via métodos de reamostragem não foram plenamente desenvolvidos.

Capítulo 3

Intervalos de Confiança via Métodos de Reamostragem

3.1 Introdução

O grande objetivo da maioria dos pesquisadores é estimar a incerteza associada aos estimadores para que com isso possamos determinar a região onde se encontra o valor real dos parâmetros. Os estimadores assim como suas incertezas associadas podem ser estimadas através de técnicas estatísticas paramétricas ou não paramétricas.

A maioria das técnicas para estimação dessa incerteza não são apropriadas quando temos uma amostra muito pequena. Neste caso muitas técnicas baseadas em simulação computacional são mais adequadas. As técnicas de simulação atualmente mais conhecidas são as de reamostragem, como por exemplo, o $Bootstrap$ e $Jacknife$. Ambas técnicas de

reamostragem reproduzem a função densidade acumulada do parâmetro que auxiliam na determinação da região de incerteza do parâmetro.

3.2 O Método *Bootstrap*

O *bootstrap* é um método que pode ser facilmente implementado tanto de forma não-paramétrica quanto paramétrica, dependendo do conhecimento do problema. No caso não-paramétrico, o método *bootstrap* reamostra os dados um número **B** de vezes com reposição, em que os valores são escolhidos segundo uma função de distribuição empírica estimada, tendo em vista que, em geral, não se conhece a distribuição subjacente aos dados. No caso paramétrico, quando se tem informação suficiente sobre a forma da distribuição dos dados, a amostra *bootstrap* é formada realizando-se a amostragem diretamente desta distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas.

Considerando-se $Y_1, ..., Y_n$ uma amostra aleatória de tamanho n com distribuição de probabilidade desconhecida F que depende de um parâmetro θ, e sejam $\mathbf{y} = (y_1,\ y_2, ...,\ y_n)$ os valores observados. Como a função F é desconhecida, pode-se estimá-la pela função de distribuição empírica, baseado na amostra de tamanho n que á dada por,

$$\hat{F}(y) = \frac{\#(y_i \leq y)}{n}.$$

A função $\hat{F}$ assume valores $1/n$ para cada valor da amostra y_i, $i = 1, ..., n$.

Supondo-se agora que se queira estimar o parâmetro que depende da distribuição F, isto é, $\theta = t(F)$ baseado em $\mathbf{Y}$. Um estimador de θ é $\hat{\theta} = s(\mathbf{Y})$,resta saber então a

precisão deste estimador. Para isto, utiliza-se o método *bootstrap*, que retira amostras *bootstrap* com reposição $y_1^*, ..., y_n^*$ utilizando a função de distribuição empírica estimada $\hat{F}$. Assim, para a amostra *bootstrap* $\mathbf{y} = (y_1^*, ..., y_n^*)$, calcula-se

$$\hat{\theta}^* = s(\mathbf{y}^*)$$

Ao se obter um grande número de replicações *bootstrap* de $y_1, ..., y_n$, obtém-se a distribuição empírica de $\hat{\theta}^*$, podendo-se calcular medidas de posição, dispersão, intervalos de confiança, dentre outras, de modo a analisar a precisão do estimador obtido inicialmente.

Quando se tem informações suficientes sobre F, um estimador *bootstrap* **paramétrico** $\hat{\theta}^*$ de θ pode ser obtido da mesma forma que o já descrito anteriormente, apenas levando-se em conta a verdadeira distribuição subjacente aos dados.

3.3 Intervalo de Confiança Via *Bootstrap*

Como mencionado, em inferência estatística, tem-se interesse na quantificação do erro cometido ao se estimar um parâmetro de interesse θ através de $\hat{\theta}$. Uma estratégia usual para a busca de medidas de incerteza, que expressem este erro, é a estimação do erro padrão de $\hat{\theta}$. Entretanto, métodos analíticos para a obtenção destas medidas nem sempre estão disponíveis, ou constituem processos altamente complexos. Neste contexto, o método

bootstrap constitui uma eficiente alternativa, fornecendo estimativas do erro padrão de $\hat{\theta}$ livres de complexidades algébricas e possibilitando a obtenção de intervalos de confi-

ança sem necessidade de pressupostos sobre a distribuição do estimador.

Desta forma, o método *bootstrap* é utilizado para a obtenção de estimativas intervalares empíricas para os estimadores dos parâmetros de interesse, através da reamostragem do conjunto de dados originais.

Considerando θ como o parâmetro de interesse, para cada amostra, calcula-se a emv para θ e tem-se no final de B reamostragens, $\hat{\theta}^*_1 < ... < \hat{\theta}^*_B$ valores dos emv ordenados. Utiliza-se, então,

$$\hat{\theta}^*_{1(B+1)\left(\frac{\alpha}{2}\right)} \quad e \quad \hat{\theta}^*_{2(B+1)\left(1-\frac{\alpha}{2}\right)}$$

como sendo os limites inferiores e superiores do intervalo 100(1-α)% de confiança para θ.

3.4 Testes de Hipóteses via Método *Bootstrap*

O uso do método *bootstrap* em teste de hipótese é também de grande interesse em diversas situações. Um dos motivos para isso é que a conexão entre os limites de confiança e testes de significância pode ser explorada para testar certos tipos de hipóteses, como por exemplo, podemos verificar a significância do parâmetro de regressão β, onde se deseja testar a hipótese $H_0 : \beta = 0$. Ou ainda, podemos estar interessados em verificar se efeito do tempo é influente no ajuste do modelo, ou seja, testar $H_0 : \exp(\phi) = 1$ contra $H_1 : \exp(\phi) \neq 1$.

Uma alternativa para testar estas hipóteses é obter a distribuição empírica de Λ, dada em (2.12), através de simulação *bootstrap*,

$$\Lambda = 2(l_{sob\ H_0} - l_{completo}). \tag{3.1}$$

Valores grandes de Λ indicam rejeição da hipótese nula, assim, para a obtenção Λ utiliza-se técnica de reamostragem, isto é, para cada reamostra, calcula-se Λ e no final de **B** reamostras tem-se as estimativas da razão de verossimilhança ordenadas $\Lambda_1^* < ... < \Lambda_B^*$, podendo assim determinar a localização, Λ_0 obtida da amostra original, entre as estimativas da razão de verossimilhança, calculando o *p-valor* empírico, isto é, $\hat{p} = \frac{1}{B}\sum_{r=1}^{B} I\left(\Lambda_B^* \geq \Lambda_0\right)$, onde $I(\cdot)$ é a função indicadora. Rejeita-se a hipótese nula se $\hat{p} \leq \alpha^*$.

3.5 Exemplo de Aplicação

Para ilustrar a metodologia proposta neste capítulo, consideramos o mesmo conjunto de dados utilizado na Seção 2.5. A Tabela 3.1 mostra os intervalos de confiança assintóticos, os intervalos percentis *bootstrap* não paramétrco e paramétrico de 90%, descritos neste capítulo. A diferença entre a teoria assintótica e os resultados obtidos via reamostragem são maiores para φ e ϕ.

Tabela 3.1. EMV´s, intervalos de confiança de 90% assintóticos, *bootstrap* não-paramétrico e paramétrico.

Parâmetros	φ	ϕ	β
EMV	0.965	0.612	-0.971
Assintótico	(0.95, 0.98)	(0.55, 0.67)	(-1.15, -0.79)
bootstrap não- paramétrico	(0.91, 0.99)	(0.32, 0.70)	(-1.27, -0.83)
bootstrap paramétrico	(0.93,0.99)	(0.39, 0.74)	(-1.29, -0.82)

Novamente, percebemos uma certa sensibilidade na estimativa dos parâmetros φ e ϕ.

3.6 Estudo de Simulação

Nesta Seção apresentamos resultados do estudo de simulação realizado para comparar a eficiência das estimativas dos *bootstrap's* paramétrico e não-paramétrico, assim como dos testes de hipóteses para amostras pequenas e moderadas através do cálculo do tamanho.

O estudo de simulação foi baseado em amostras geradas da distribuição exponencial com $n = 15,\ 30,\ 50,\ 100\ e\ 300$ elementos e, assim como na Seção 2.6, os y_{is} e os c_{is} foram gerados de uma distribuição exponencial com taxa de falha igual a 0.4 e 0.5, respectivamente. A amostra foi obtida para os valores gerados $\varphi = \phi = \beta = 0.5$. Os tempos de vida foram definidos como $t_i = \min(y_i, c_i)$ e a censura especificada como $\delta_i = 1$, se $y_i < c_i$ e $\delta_i = 0$, caso contrário. As covariáveis foram geradas como sendo variáveis de Bernoulli com probabilidade de sucesso igual a 0.5.

Considerando $\mathbf{B} = 399$ replicações, a adequabilidade do *bootstrap* foi estudada via simulação Monte Carlo repetindo o procedimento descrito acima 1000 vezes onde a média dos valores obtidos foram considerados como resultados.

3.6.1 Resultados

Probabilidade de Cobertura

A adequabilidade foi estudada via simulação Monte Carlo dos intervalos de confiança de 90% para os parâmetros do modelo estudado apresentados na Tabela 3.2.

Tabela 3.2. Probabilidade de Cobertura dos Intervalos de Confiança de 90% para os parâmetros do Modelo.

	Bootstrap Paramétrico			Bootstrap Não-Paramétrico		
n	φ	ϕ	β	φ	ϕ	β
15	0.865	0.852	0.861	0.871	0.868	0.873
30	0.877	0.870	0.883	0.911	0.881	0.896
50	0.897	0.894	0.896	0.893	0.896	0.898
100	0.910	0.903	0.909	0.914	0.914	0.910
300	0.903	0.892	0.902	0.906	0.897	0.904

De acordo com a Tabela 3.2, podemos notar que a probabilidade de cobertura dos intervalos de confiança de 90% são razoáveis para $n \geq 30$, considerando o método não-paramétrico. Os intervalos de confiança do *bootstrap* paramétrico podemos concluir que

são coerentes para $n \geq 50$. A Figura 3.1 apresenta as probabilidades de cobertura versus o tamanho da amostra para os intervalos de confiança dos parâmetros do modelo.

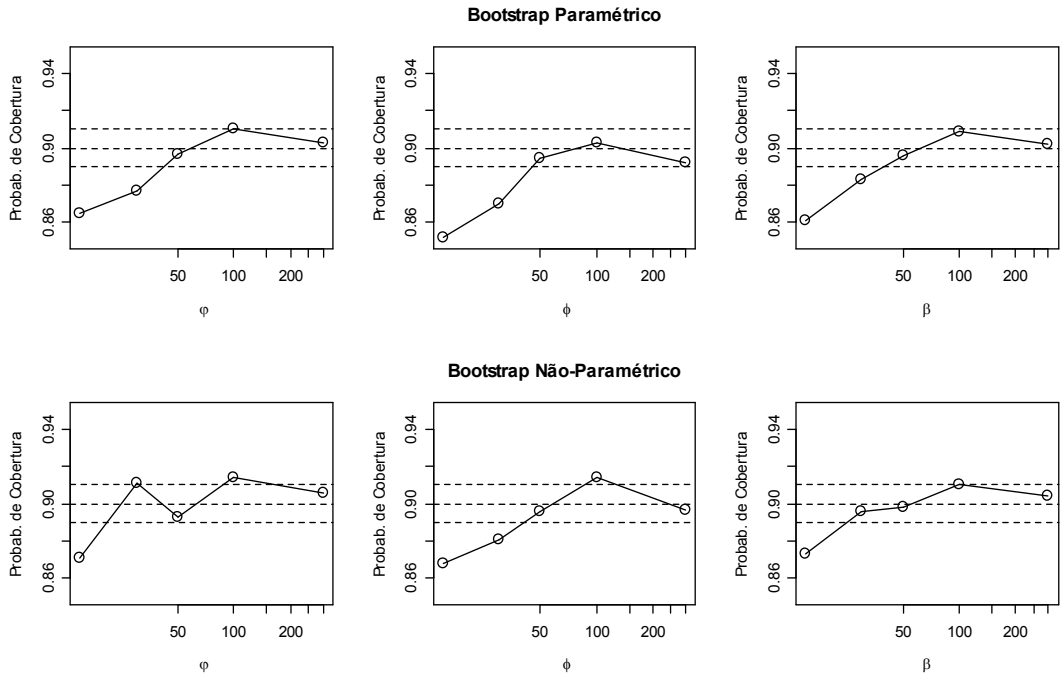

Figura 3.1. Probabilidade de Cobertura dos Intervalos *Bootstrap* de 90% versus o Tamanho da Amostra, n.

Testes de Hipóteses

Para verificar a influência do tempo e da covariável no ajuste do modelo, as hipóteses $H_0 : \beta = 0$ e $H_0 : \exp(\phi) = 1$ podem ser testadas.

Considerando o $H_0 : \beta = 0$ contra $H_1 : \beta \neq 0$, podemos fixar $\beta = 0$ para calcular o tamanho empírico médio.

Para o estudo do tamanho do teste, observou-se quantas vezes a estatística Λ caiu numa região crítica. Foram marcados os percentis 95 da distribuição empírica dos valores

$\Lambda_1^*, ..., \Lambda_{399}^*$ para determinar a região crítica do teste e determinado se Λ_0 do passo inicial estava ou não na região crítica. Este passo foi feito para cada uma das 1000 amostras simuladas. O tamanho do teste foi estimado calculando o número de vezes em que a estatística Λ_0 caiu na respectiva região crítica de 0.05.

A Tabela 3.3 e 3.4 apresenta os resultados obtidos. Notamos que os tamanhos são próximos de 0.05 para os valores grandes de n, nas hipóteses testadas.

Tabela 3.3. Tamanho Empírico Médio do Teste Estatística de Verossimilhança para $H_0 : \exp(\phi) = 1$.

n	15	30	50	100	300
	0.065	0.059	0.058	0.054	0.053

Tabela 3.4. Tamanho Empírico Médio do Teste Estatística de Verossimilhança para $H_0 : \beta = 0$.

n	15	30	50	100	300
	0.066	0.061	0.062	0.056	0.054

3.7 Considerações Finais

Neste capítulo desenvolvemos intervalos de confiança e testes de hipóteses via reamostragem. Um estudo de simulação foi desenvolvido com o objetivo de verificar a adequação dos intervalos de confiança e o tamanho dos testes de hipóteses propostos. Observamos que a teoria de reamostragem é adequada para amostras moderadas.

A seguir apresentaremos um método alternativo de modelagem, a abordagem bayesiana, que também pode ser aplicada para estimar os parâmetros do modelo.

Capítulo 4

Abordagem Bayesiana

Neste Capítulo apresentamos uma abordagem Bayesiana para o modelo logístico generalizado dependente do tempo (2.1). As inferências para os parâmetros do modelo para o mesmo exemplo considerado em estudo são apresentadas na Seção 4.1 e é aplicada, considerando o exemplo estudado, na Seção 4.2. A Seção 4.3 finaliza o capítulo com algumas considerações finais.

4.1 Procedimento Geral

Nesta seção apresentamos alguns conceitos básicos de forma geral para que os modelos possam ser abordados numa perspectiva Bayesiana. A abordagem Bayesiana é adequada quando ocorrem situações onde os modelos avaliados são incompatíveis com modelos simples e outros modelos mais realísticos e, conseqüentemente modelos mais complexos são necessários.

A implementação de técnicas Bayesianas para tais modelos complexos pode resultar em distribuições a *posteriori* de difícil tratamento analítico, como por exemplo, o grande número de parâmetros a serem estimados, a dificuldade na obtenção das densidades marginais de forma analítica, ou também quando as distribuições a *priori* e a *posteriori* não são conjugadas. Nestes casos, os métodos analíticos de aproximação não são indicados, e métodos de aproximação numérica são necessários para estimação dos parâmetros de interesse. Os métodos analíticos de aproximação são eficientes computacionalmente, porém, são baseados em resultados assintóticos e na suposição de normalidade, também são mais difíceis para programar à medida que o número de parâmetros aumenta.

Assim como na inferência clássica, a inferência bayesiana trabalha na presença de observações x cujo valor é inicialmente incerto e descrito através de uma distribuição de probabilidades $f(x|\theta)$. A quantidade θ serve como indexador da família de distribuições das observações representando características de interesse que se deseja conhecer para poder ter uma descrição completa do processo.

Dados os dados observados $\mathbf{x} = (x_1, ..., x_p)$ obtidos sobre um modelo paramétrico $f(x|\theta)$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, o procedimento de inferência bayesiana é baseado na forma familiar do teorema de Bayes,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{l(\boldsymbol{\theta};\mathbf{x})\pi(\boldsymbol{\theta})}{\int l(\boldsymbol{\theta};\mathbf{x})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Se o interesse é centrado em algum componente de $\boldsymbol{\theta}$, digamos θ_i $(i = 1, ..., k)$, a correspondente distribuição marginal pode ser facilmente obtida integrando-se a distribuição

conjunta $p(\boldsymbol{\theta}|\mathbf{x})$,

$$p(\theta_i|\mathbf{x}) = \int l(\boldsymbol{\theta};\mathbf{x})\pi(\boldsymbol{\theta})d\theta_{-i},$$

onde o subscrito $-i$ implica a integração de todos os componentes exceto θ_i. As inferências sumárias na forma de esperanças a *posteriori* são requeridas são obtidas por,

$$E(\psi(\boldsymbol{\theta})|\mathbf{x}) = \int \psi(\boldsymbol{\theta})p(\boldsymbol{\theta};\mathbf{x})d\boldsymbol{\theta},$$

para uma escolha de $\psi(\boldsymbol{\theta})$ de interesse.

Nas últimas décadas, alguns métodos numéricos têm sido desenvolvidos para resolver as questões mencionadas acima. Os métodos numéricos baseados em amostragem englobam os métodos de integração simples de Monte Carlo, os métodos de reamostragem por importância e os métodos de Monte Carlo via Cadeias de Markov (MCMC). Estes são mais simples para implementação e não apresentam restrições quanto ao número de parâmetros a serem estimados.

4.1.1 Cadeias de Markov Monte Carlo (MCMC)

O método de MCMC é uma forma de integração Monte Carlo. A idéia é simular uma cadeia de Markov irredutível aperiódica cuja distribuição estacionária é a distribuição de interesse $\pi(\boldsymbol{\theta})$, ou, para os bayesianos, a densidade a *posteriori* $p(\boldsymbol{\theta}|\mathbf{x})$. Existem dois métodos de gerar cadeia de Markov com distribuição estacionária especificada. Um deles é o algoritmo de Metropolis (Metropolis *et al.*, 1953). Hastings (1970) mostra uma

generalização do algoritmo Metropolis. O outro é o *Gibbs Sampler* (Geman & Geman, 1984). Na subseção seguinte será descrito o algoritmo de Metroplois-Hastings, utilizado neste trabalho para obter as estimativas dos parâmetros de interesse do modelo dado em (2.8).

4.1.2 Metroplois-Hastings

Quando as distribuições condicionais a *posteriori* não possuem forma conhecida, que impossibilita a geração direta a partir destas distribuições usamos o algoritmo Metroplis. Uma generalização do algoritmo dado por Hastings (1970) requer uma densidade de transição $q(\theta, \theta^*)$, que, não necessariamente, tem probabilidade de equilíbrio π, mas que represente uma regra de passagem que defina uma cadeia. Considerando.a probabilidade de aceitação $p(\theta^{(j-1)}, \theta^*)$, o algoritmo segue os seguintes passos,

1. dê um valor inicial $\theta = \theta^{(0)}$ e inicialize o contado de iteração $j = 1$;

2. mova a cadeia para um novo valor de θ^* gerado da densidade $q(\theta^{(j-1)}, \theta^*)$;

3. gere u da distribuição uniforme no intervalo (0,1);

4. aceite o valor gerado, θ^*, se,

$$u \leq \min\left(1, \frac{\pi(\theta^*)q(\theta^*, \theta^{(j-1)})}{\pi(\theta^{j-1})q(\theta^{(j-1)}, \theta^*)}\right).$$

Caso contrário, $\theta^{(j)} = \theta^{(j-1)}$;

5. incremente j e volte para o passo 1.

Para j suficientemente grande, $\theta^{(j)}$ é uma amostra da distribuição a *posteriori* $\pi(\theta)$.

Para o caso vetorial, em que, $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ tem-se a densidade de transição dada por $q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ e uma probabilidade de aceitação dada por $p(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$.

4.1.3 Diagnósticos de Convergência

Apesar dos métodos MCMC serem uma ótima ferramenta para a resolução de muitos problemas práticos, algumas dificuldades surgem com o uso destes métodos. Entre outras podemos citar: o numero de iterações para se obter convergência; a possibilidade das iterações iniciais da amostra serem influenciadas pelos valores iniciais dos parâmetros e, ainda, das seqüências de valores apresentarem correlação entre os parâmetros.

Não existe uma técnica geral para solucionar tais questões, contudo, existem métodos de verificação de convergência que são baseados nas propriedades da cadeia de Markov, para indicar a convergência, ou não, da amostra simulada para a distribuição marginal.

Uma avaliação da convergência pode ser feita preliminarmente analisando-se gráficos ou as medidas descritivas (media, desvio-padrão e os quantis) obtidas a partir dos valores simulados para os parâmetros de interesse. Entre os gráficos mais freqüentes podemos citar o da quantidade de interesse estimada ao longo das iterações e o da estimativa da distribuição marginal a *posteriori* deste parâmetro.

Outra avaliação da convergência é feita utilizando algumas técnicas de diagnostico das cadeias. As técnicas de diagnostico mais populares são descritas por Geweke (1992),

Gelman & Rubin (1992) e Raftery e Lewis (1992). Estas técnicas estão implementadas na biblioteca CODA (*Convergence Diagnosis and Output Analysis Software for Gibbs sampling Output*) (Best et al., 1997). Geralmente, estes métodos são utilizados conjuntamente a fim de obter uma indicação de convergência, uma vez que nenhum deles é considerado melhor do que o outro. Cowles & Carlin (1995) compararam alguns métodos concluindo que, apesar de muitos deles detectarem comumente os problemas na convergência que se propuseram identificar, essas técnicas podem também falhar no seu propósito, não sendo possível afirmar qual delas é mais eficiente. Com isto, é recomendável que o uso destes diagnósticos seja combinado com a analise gráfica e medidas descritivas.

4.1.4 Seleção de Modelos

A comparação Bayesiana é baseada no cálculo das probabilidades a *posteriori* para os modelos competitivos e não apresenta dificuldades na comparação entre modelos com estruturas diferentes. Quando os modelos são, a priori, igualmente prováveis, a chance a *posteriori* de um modelo em relação a outro se reduz ao denominado fator de Bayes. Entretanto, existem dificuldades com esta abordagem quando a informação a priori sobre os parâmetros dos modelos é imprópria. Nesta Seção apresentam-se a definição do fator de Bayes.

4.1.5 Fator de Bayes

Suponha que existam k modelos $M_1, M_2, ..., M_k$ e seja $\pi(\Theta|D, M_i)$ a distribuição a *posteriori* sobre o modelo M_i. Por simplicidade, utiliza-se a mesma notação de Θ, para os parâmetros sobre cada modelo, mas a dimensão de Θ pode ser mudar de modelo para modelo.

Assim, seja $f(D|M_i)$ a verossimilhança marginal para o modelo M_i, em que D denota o conjunto das observações,

$$f(D|M_i) = \int \pi(\Theta|D, M_i) d\Theta. \tag{4.1}$$

Então, o fator de Bayes, que compara os modelos i e r, é definido por,

$$B_{ir} = \frac{f(D|M_i)}{f(D|M_r)}.$$

A probabilidade a *posteriori* do modelo M_i pode ser calculada como,

$$\pi(M_i|D) = \frac{f(D|M_i)\pi(M_i)}{\sum_{j=1}^{k} f(D|M_j)\pi(M_j)},$$

em que $\pi(M_i)$ representa a probabilidade a *priori* do modelo M_i.

Assim, é aproximada a densidade marginal pelas estimativas Monte Carlo, obtidas das S amostras Gibbs geradas, dadas por,

$$\hat{f}(D|M_i) = \sum_{s=1}^{S} f(D|\theta_i^{(s)}, M_j), \tag{4.2}$$

em que $\theta_i^{(s)}$ representa os parâmetros obtidos pelos métodos descritos na Seção 4.1.

O fator de Bayes é um resumo da evidência provida pelos dados a favor de uma teoria específica representada por um modelo estatístico. Jeffreys, em 1961, (Jeffreys, 1961) sugeriu interpretar o fator de Bayes pela escala do logaritmo na base 10. Assim, tem-se,

B_{10}	$\log(B_{10})$	$\log_e(B_{10})$	Evidência contra H_0
1.0 - 3.2	0.0 - 0.5	0.0 - 1.2	insignificante
3.2 - 10.0	0.5 - 1.0	1.2 - 2.3	significativa
10.0 - 100.0	1.0 - 2.0	2.3 - 4.6	forte
> 100.0	> 2.0	> 4.6	decisiva

Esta probabilidade fornece uma melhor escala significativa, mas não são uma calibragem do fator de Bayes.

4.1.6 Estatística CPO (Conditional Predictive Ordinate)

A estatística CPO (Conditional Predictive Ordinate) tambe é outra ferramenta útil na verificação do modelo e vem sendo extensamente usada na literatura. Para uma discussão mais detalhada desta estatística, sugerimos ver Geisser (1993), Gelfand, Dey, and Chang (1992), Dey, Chen and Chang (1997) and Sinhá and Dey (1997).

Considerando a *i-ésima* observação, a estatística CPO é definida por,

$$CPO_i = f\left(t_i|D^{(-i)}\right) = \int f\left(t_i|\Theta, x_i\right)\pi\left(\Theta|D^{(-i)}\right)d\Theta; \tag{4.3}$$

em que t_i representa a variável resposta e x_i é o vetor de covariáveis para o *i-ésimo* caso, $D^{(-i)}$ denota o conjunto de dados após a exclusão do *i-ésimo* caso, e $\pi\left(\Theta|D^{(-i)}\right)$ é a densidade a *posteriori* de Θ baseada nos dados $D^{(-i)}$. Desta forma, temos que a CPO_i é a densidade preditiva marginal a *posteriori* de t_i dado o conjunto de dados $D^{(-i)}$ e pode ser interpretada como a altura desta densidade marginal em t_i. Então, altos valores da CPO_i indicam o melhor ajuste do modelo.

Quando temos mais de um modelo de sobrevivência, uma forma fechada da estatística CPO_i não é possível. Contudo, um estimador Monte Carlo da CPO_i pode ser obtido usando a amostragem MCMC da distribuição a posteriori $\pi(\Theta|D)$, onde D denota o conjunto completo dos dados. Uma implementação mais detalhada para o cálculo da CPO_i pode ser encontrada no Capítulo 10 de Chen, Shao e Ibrahim (2000).

Para comparar dois modelos não-encaixados podemos examinar as $CPO_i\prime s$ sob ambos modelos. A observação com o maior valor da CPO sob um determinado modelo irá mantê-lo. Assim, um gráfico das $CPO_i\prime s$ sob ambos os modelos contra o número de observações nos fornece o melhor modelo, ou seja, o modelo que possui a maioria das $CPO_i\prime s$ acima do outro modelo. Desta forma, para a comparação de vários modelos não encaixados, o valor da CPO_i sob todos os modelos pode ser plotado contra o número de observações num gráfico simples.

Uma alternativa do gráfico da CPO é sumariar a estatística chamada de logaritmo da verossimilhança pseudo-marginal (LPML - logarithm of the Pseudomarginal likelihood) (Geisser e Eddy, 1979), definida como,

$$LPML = \sum_{i=1}^{n} \log\left(CPO_i\right). \tag{4.4}$$

Para compararmos $LPML\prime s$ de diferentes estudos para um determinado modelo podemos usar uma modificação em (4.4) dada pelo calculo da média,

$$ALPML = \frac{LPML}{n}, \tag{4.5}$$

em que n é o tamanho da amostra.

Podemos notar também que a $LPML$ é bem definida quando a densidade de preditiva *a posteriori* é própria. Então, $LPML$ é bem definida sob *prioris* impróprias sendo computacionalmente estáveis. Tendo assim, uma vantagem sobre o fator de Bayes, pois este não é bem definido sob *prioris* impróprias e totalmente sensível quando estas são vagas.

4.2 Exemplo de Aplicação

Para exemplificar a metodologia Bayesiana, nesta subseção será apresentada a ilustração do modelo utilizando o conjunto de dados retirado de Kalbfleisch & Prentice (1980) introduzido na Seção 2.5 do Capítulo 2. Lembrando que o principal objetivo da análise é verificar a diferença entre o Grupo I e o Grupo II.

Foi considerado distribuições a priori independentes para φ, ϕ, β, em que, a distribuição a *priori* conjunta é dada por,

$$\pi(\varphi, \phi, \beta) = \pi(\varphi)\pi(\phi)\pi(\beta), \text{ em que }, \pi(\varphi) = \pi(\phi) = \pi(\beta) = N(0, 5). \qquad (4.6)$$

Estas distribuições foram escolhidas de forma subjetiva porém, uma análise de sensibilidade foi realizada escolhendo outros valores para os hiperparâmetros e constatamos que as outras escolhas não modificaram os resultados apresentados a seguir.

Ressaltamos aqui que não há maiores problemas em centrar a densidade *a priori* numa normal com média zero, pois constatamos que a função de risco empírica dos dados é crescente como pode ser observado na Figura 5.1.

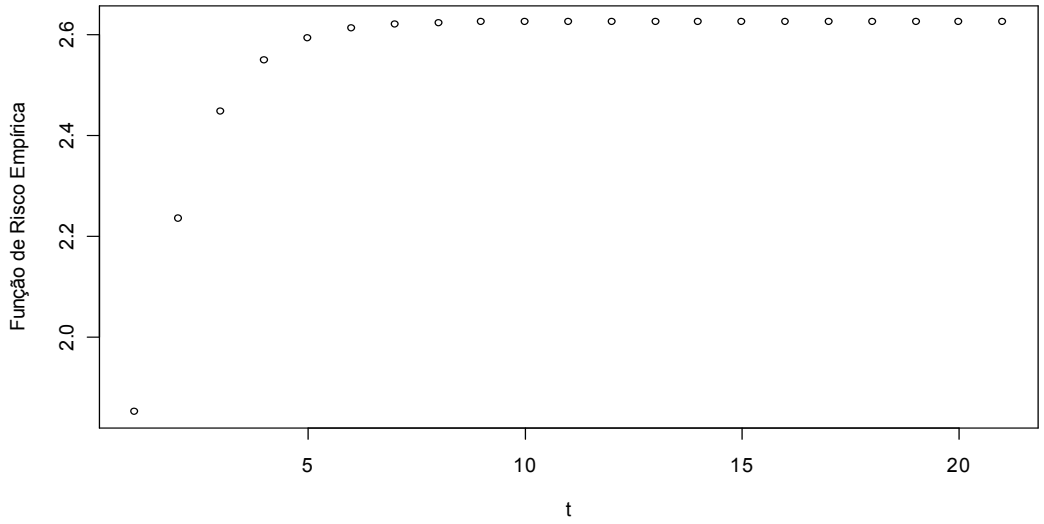

Figura 5.1. Função de Risco Empírica.

As densidades a posteriori marginais da distribuição a posteriori não são facilmente obtidas, isto porque a integração da densidade a *posteriori* conjunta é complicada. Uma alternativa para estas situações é o algoritmo Metropolis-Hastings descrito anteriormente. O software R foi usado para gerar 3 cadeias de 53.000 iterações para os parâmetros. As

primeiras 3.000 iterações foram ignoradas e o restante foi selecionado de 5 em 5 formando uma amostra de tamanho 10.000 para analise.

Para o uso do algoritmo Metropolis-Hastings descrito na Subseção 4.1.2, foram utilizados para gerar valores candidatos as próprias densidades a *priori*. Esse procedimento é artificial, mas é possível aplicá-lo monitorando-se a saída do algoritmo Metropolis-Hastings através da taxa de aceitação (Bessag *et al*, 1995 e Geyes, 1992). Portanto, o uso da *priori* para gerar candidatos e uma taxa de aceitação de 20% a 40% asseguram que o algoritmo Metropolis-Hastings deve levar a convergência para a distribuição de equilíbrio (a *posteriori*). Além disso, foi aplicado o critério de convergência Gelman Rubin (Gelman *et al.*, 1995) com 3 cadeias para assegurar a homogeneidade das seqüências geradas. O numero de iteração foi considerado suficiente para a convergência aproximada quando a redução de escala potencial estimada foi $\hat{R} < 1.1$ (apêndice A), além da verificação gráfica apresentada a seguir.

Assim, para o uso de tal algoritmo consideramos as seguintes distribuições a *posteriori* condicionais completas para os parâmetros em estudo.

Distribuições a *Posteriori* Condicionais

$$\pi(\varphi|\phi,\beta,Dados) \propto \Pi_{i=1}^{n} \left\{ \exp(\delta_i \varphi) \left[\frac{1+\exp(t_i \exp(\phi) + x_i'\beta)}{1+\exp(x_i'\beta)} \right]^{-\frac{\exp\varphi}{\exp\phi}} \right\} \pi(\varphi).$$

$$\pi(\phi|\beta,Dados) \propto \Pi_{i=1}^{n} \left\{ \left[\frac{\exp(t_i \exp(\phi))}{1+\exp(t_i \exp(\phi) + x_i'\beta)} \right]^{\delta_i} \left[\frac{1+\exp(t_i \log(\phi) + x_i'\beta)}{1+\exp(x_i'\beta)} \right]^{-\frac{\exp\varphi}{\exp\phi}} \right\} \pi(\phi).$$

$$\begin{aligned} \pi(\beta_j|Dados,\phi,\beta_{-j}) \quad \propto \quad \Pi_{i=1}^{n} \Bigg\{ & \left[\frac{\exp(x_i'\beta_j)}{1+\exp(t_i \exp(\phi) + x_i'\beta_j)} \right]^{\delta_i} \\ & \left[\frac{1+\exp(t_i \exp(\phi) + x_i'\beta_j)}{1+\exp(x_i'\beta_j)} \right]^{-\frac{\exp\varphi}{\exp(\phi)}} \Bigg\} \pi(\beta_j). \end{aligned}$$

A convergência das cadeias geradas está de acordo com os diagnósticos de convergência implementados no CODA (Best *et al.*, 1997). Os traços das cadeias e a estimação da densidade para cada parâmetro, apresentados na Figura 5.2**,** indicam que não há problemas com a convergência no algoritmo.

As medidas de resumo obtidas são apresentadas na Tabela 4.1.

Tabela 4.1 - Estatísticas Resumo das distribuições a posteriori no Modelo Logístico Generalizado Reduzido, utilizando Distribuição a Priori $N(0,5)$.

Parâmetro	Média	DP	IC (95%)	
φ	0.952	0.042	0.87	1.03
ϕ	0.618	0.097	0.42	0.81
β	-0.960	0.038	-1.03	-0.89

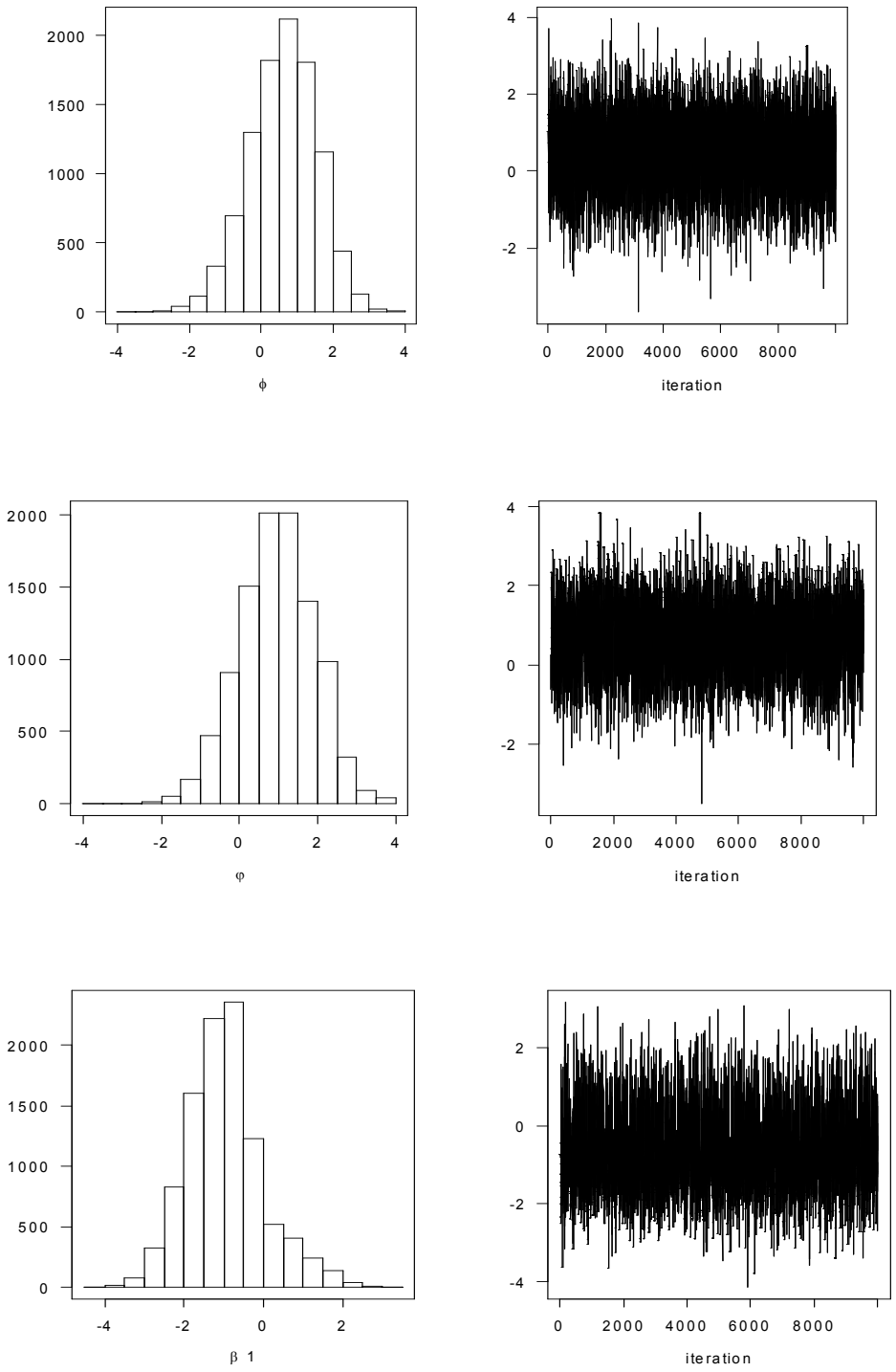

Figura 5.2. Posteriori Marginal e o traço da cadeia gerada para os parâmetros φ, ϕ e β, respectivamente.

É importante notar que há diferença significativa entre os grupos 1 e 2 para o desenvolvimento de tumores indicado pela estimativa de β. Este resultado é confirmado pelo valor do fator de Bayes considerando o modelo completo (M_1) com o modelo sem covariável β (M_0) que é igual a 3.02. Além disso, há um efeito significativo do tempo para o ajuste do modelo, onde o fator de Bayes do modelo completo (M_1) com o modelo sem o termo do efeito do tempo (M_2) é igual a 2.54.

Também apresentamos os resultados utilizando a estatítica CPO para a verificação da significância da variável β no modelo. A Figura 5.3 apresenta o log da razão das CPO´s para o modelo completo M_1 versus o modelo M_0 (sem a covariável β) contra o número de observações. Assim, os pontos maiores que zero favorecem o modelo M_1. Nesta figura podemos notar que aproximadamente 50% das observações suportam o modelo M_1. A diferença nos $LPML's$ entre M_0 e M_1 é aproximadamente 0.92. Desta forma, o modelo M_1 é ligeiramente melhor que o modelo M_0 baseando na estatística CPO.

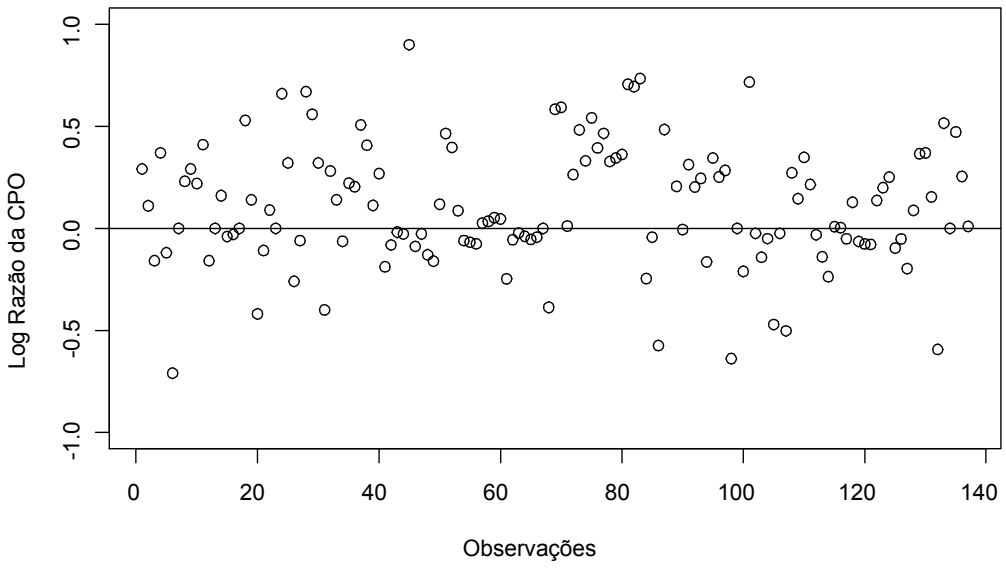

Figura 5.3. Razão das CPO´s para os modelos M_0 e M_1.

Da mesma forma calculamos a estatística CPO para verificação da significância do efeito do tempo no ajuste do modelo. A Figura 5.4 apresenta o log da razão das CPO´s para o modelo completo M_1 versus o modelo M_2 (sem a covariável ϕ) contra o número de observações. A diferença nos $LPMLı s$ entre M_1 e M_2 é aproximadamente 0.97. Desta forma, o modelo M_1 é representa melhor os dados que o modelo M_2.

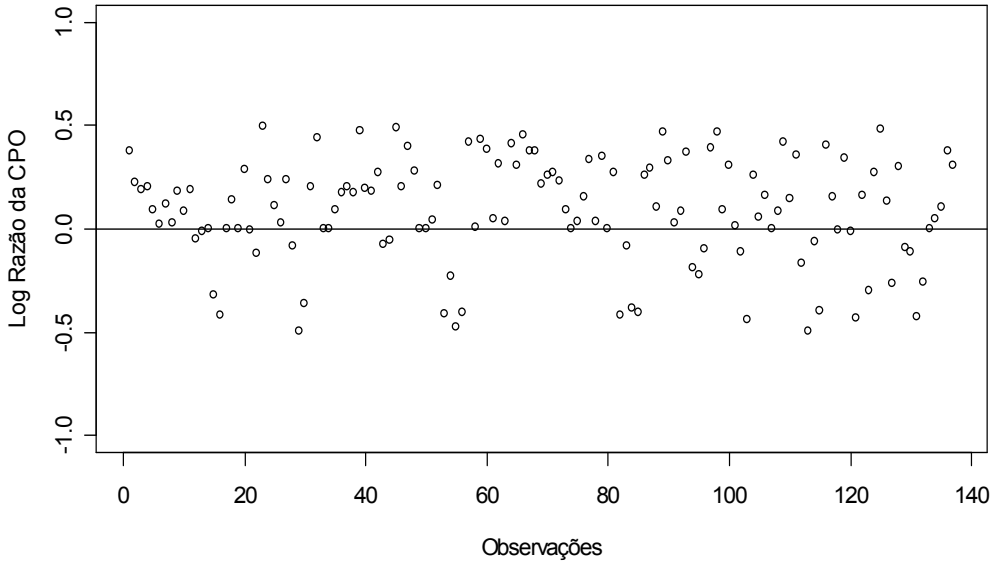

Figura 5.4. Razão das CPO´s para os modelos M_1 e M_2.

4.3 Considerações Finais

A metodologia Bayesiana aplicada neste Capítulo mostrou-se adequada para o estudo em questão, conseguindo estimar de forma simples os parâmetros do modelo, fornecendo resultados consistentes, como pode ser constatado observando-se as técnicas de convergência apresentadas neste trabalho.

Capítulo 5

Conclusões e Perspectivas Futuras

Neste trabalho estudamos sistematicamente o modelo de risco logístico generalizado dependente do tempo bem como suas propriedades, entre elas a de permitir o efeito do tempo no ajuste do modelo. Estudamos este modelo pelos métodos assintóticos, de reamostragem e Bayesianos. As principais contribuições deste trabalho foram as apresentações do estudo detalhado nestes três tipos de metodologia.

Do ponto de vista clássico, verificamos uma certa dificuldade referente ao problema da estimação de máxima verossimilhança. Além disso, os resultados assintóticos encontrados devem ser usados com cautela, uma vez que a teoria clássica usual pode não ser válida, particularmente quando a amostra é pequena.

Alternativamente utilizou-se técnicas de reamostragem na obtenção de intervalos de confiança e das distribuições empíricas das estatísticas de testes. Um estudo de simulação foi realizado com o objetivo de verificar a validade dos intervalos obtidos via *bootstrap*

paramétrico e não-paramétrico bem como verificar o tamanho do teste *boostrap* baseado na estatística de razão de verossimilhança. De acordo com os resultados obtidos, podemos observar que os intervalos de confiança de 90% são adequados para tamanhos de amostras moderadas, assim como o tamanho do teste.

Os resultados obtidos via métodos Bayesianos também são consistentes que aqueles obtidos na análise clássica, sendo que os resultados foram muito similares aos obtidos pelo método de reamostragem.

De forma geral, o modelo mostrou-se útil para resolver problemas que eventulamente podem ocorrer quando os dados de sobrevivência não são proporcionais. Uma análise do modelo logístico generalizado pode conduzir a vários trabalhos futuros, como por exemplo, considerar um modelo de fragilidade alternativo. Além disso, estudar as amplitudes médias dos intervalos de confiança propostos assim como o estudo do poder de testes de hipóteses considerados via simulação Monte Carlo.

Referências Bibliográficas

1. Best, N., Cowles, M. and Vines, K. (1997). CODA - *Convergence Diagonosis and Output Analysis software for Gibbs sampling output: Version 0.4*, MRC Biostatistics Unit, Cambrigde UK.

2. Bickel, P. J., Doksum, K. A. (1981). *In analysis of transformations revisited*, Journal of the American Statistical Association, **76**, 296-311.

3. Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 269-298.

4. Chen, M.H., Dey,D.K., and Sinha,D. (2000). Bayesian analysis of multivariate mortality data with large families. *Applied Statistics*, **49**, 129-144.

5. Chib, S. & Greenberg, E.I. (1995). *Understanding the Metropolis-Hastings Algorithm.* The American Statisticisn, **49**, 327-335.

6. Cox, D. R., 1962, *Further results on tests of separate families*, Journal of the Royal Statistical Society B 24, 406 - 424.

7. Cox, D. R., 1961, *Tests of separate families of Hypoteses*, Proceedings of the 4th Berkeley symposium, Vol 1 (University of California Press, Berkeley. CA) 105 - 123.

8. Cox, D.R. (1972a). *Regression models and life tables (with discussion).* Journal of the Royal Statistical Society B, **34**, 187-220.

9. Cox, D.R., 1975, *Partial likelihood.* Biometrika, **62**.

10. Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Assossiation*, **91**, 883-904.

11. Davison,A. C.; Hinkley, D. V. *Bootstrap methods and their application.* Cambrige: Cambridge University Press, 1997.

12. Dey, D. K., Chen, M. H., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, **53**.

13. Efron, B. (1977). *Censored data and the bootstrap.* Journal of the American Statistical Association, **76**,312-319.

14. Etezadi-Amoli, J. and Ciampi, A. *Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function.* Biometrics, **43**,1987.

15. Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Applied Statistics* **40**, 63-79.

16. Geisser, S. (1993). *Predictive Inference: An Introduction.* London: Chapman & Hall.

17. Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Jounal of the American Statistical Assossiation*, **74**, 153-160.

18. Gelfand, A. E., Dey, D.K.,and Chang,H. (1992). Model determinating using predictive distribuitions with implementation via sampling-based methods (with Discussion). *In Bayesian Statistics*, **4**.

19. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

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

21. Geman, S. and Geman, D. (1984). *Stocastic relation, Gibbs distribution, and the Bayesian restoration of images*. IEEE Transactions on Pattern Analysis and Machine Inteligence, **6**, 721-741.

22. Geweke, J. (1989). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *In Bayesian Statistics* **4**.

23. Hall, P. and Wilson, S. R. (1991). *Two guidelines for bootstrap hipothesis testing*. Biometrics, **47**, 757-762.

24. Hastings W. K. (1970). Monte Carlo sampling methods using Markov chains and their aplications. Biometrika, **57**, 97-109.

25. Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford, U. K. : Oxford University Press.

26. Kalbfleisch, J. F, Prentice, R. L. The Statistical Analisys of Failure Time Data.John Wiley and Sons, New York, 1980.

27. Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley, 580p.

28. Louzada-Neto, F. Mazucheli, J., Achcar, J. Uma Introdução à Análise de Sobrevivência e Confiabilidade. XXVIII Jornads Nacionales de Estadística Chile 2001.

29. MacKenzie, G. (1996). Regression models for survival data: the generalized time-dependent logistic family. *The Statistician*, **45**, 21-34.

30. MacKenzie, G. (2002). A Logistic Model for Survival Data. In: 17th International Workshop in Statistical Modelling , p.431-438.

31. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M., N., Teller, A.H. ande Teller, E. (1953) *Equations of state calculations by fast computing machine*. Journal of Chemical Physics, **21**, 1087-1091.

32. Prentice, R.L., Willians, B.J. and Peterson, A.V. (1978). *On the regression analysis of multivariate failure time data*. Biometrika, **68**, 373-379.

33. Raftery, A.E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? *In Bayesian Statistic*, **4**.

34. Sinha, D. and Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Assossiation* **92**, 1195-1212.

Apêndices

A. Diagnostico de Gelman Rubin

Este método, proposto por Gelman e Rubin (1992) utiliza cadeias múltiplas, inicializadas em pontos distintos, comparando a variância amostral dentro e entre as cadeias,para cada uma das quantidades de interesse. Esta comparação é usada para estimar o fator pelo qual o parâmetro de escala da distribuição a posteriori marginal pode ser reduzido, quando o numero de iterações é grande. Para a obtenção de um resultado melhor no teste, as quantidades de interesse devem ter distribuição aproximadamente normal.

A inspeção visual de similaridade entre a trajetória das varias cadeias após um certo numero de iterações certamente é um indicio forte de convergência. Gelman & Rubin (1992) formalizaram essa idéia, de que as trajetórias das cadeias devem ser a mesma depois de convergirem, através do uso de técnicas de analise de variância. A idéia geral testar se a dispersão intra-cadeias é maior do que a dispersão inter-cadeias. Isto equivale a dizer que o histograma das cadeias como um todo deve ser similar aos histogramas das cadeias tomadas individualmente.

Considerando m cadeias que evoluem em paralelo e uma função real $t(\theta)$, tem-se m trajetórias $\left\{t_i^{(1)}, t_i^{(2)}, ..., t_i^{(n)}\right\}$, $i = 1, ..., m$ para t. Portanto, podem ser obtidas a variância entre as cadeias, B, e a variância, W. As fórmulas correspondentes são dadas por,

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{t}_i - \bar{t}) \qquad \text{e} \qquad W = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(t_i^{(j)} - \bar{t}_i\right)^2,$$

em que $\bar{t}_i$ é a média das observações da cadeia i e $\bar{t}$ é a média dessas médias, $i = 1, ..., m$.

Sob condição de convergência, todos os $m \cdot n$ valores serão gerados da *posteriori e a variância de* t pode ser estimada de forma não viciada,

$$V\left(t\left(\theta\right)\right) = \left(1 - \frac{1}{n}\right)W + \left(\frac{1}{n}\right)B.$$

Se as cadeias ainda não estiverem convergido então essa estimativa é maior que $V\left(t\left(\theta\right)\right)$, pois os valores iniciais ainda estão sendo influenciados pelos valores dos outros parâmetros da cadeia, de forma que a distribuição de equilíbrio ainda não foi atingida, indicando que eles foram escolhidos com dispersão maior que da distribuição de equilíbrio. Por outro lado, W fornece estimativas menores que $V\left(t\left(\theta\right)\right)$, pois a cadeia não terá coberto toda a variabilidade de $t\left(\theta\right)$.

Um indicador de convergência é dado pela chamada redução potencial estimada da escala $R = \sqrt{\hat{V}\left(t\left(\theta\right)\right)/W}$, que é sempre maior que 1. À medida que n cresce, ambos os estimadores acabarão convergindo para $V\left(t\left(\theta\right)\right)$ e R covergirá para 1. Assim, $\hat{R}$ pode ser usado como indicador de convergência pela avaliação de sua proximidade a 1. Gelman (1995) sugere aceitar como garantia de convergência valores de $\hat{R} \leq 1.1$.

B - Programa em R usado o ajuste clássico do modelo Logístico Generalizado Reduzido com dados censurados à direita.

```
psi<-seq(-10000,10000, len=50)

fi<-seq(-10000 ,10000 ,len=50)

beta<-seq(-10000 ,10000 ,len=50)

U<-vector(mode="numeric",length=3)

pi<-vector(mode="numeric",length=1)

qi<-vector(mode="numeric",length=1)

gi<-vector(mode="numeric",length=1)

ri<-vector(mode="numeric",length=1)

si<-vector(mode="numeric",length=1)

U1<-vector(mode="numeric",length=1)

U2<-vector(mode="numeric",length=1)

U3<-vector(mode="numeric",length=1)

J1<-vector(mode="numeric",length=1)

J2<-vector(mode="numeric",length=1)

J3<-vector(mode="numeric",length=1)

J12<-vector(mode="numeric",length=1)

J13<-vector(mode="numeric",length=1)

J23<-vector(mode="numeric",length=1)

J<-matrix(nrow=3,ncol=3)
```

```
raph<-function(psi,fi,beta,x,t,delta){
for (k in 1:n){
pi[k]<-exp(exp(fi)*t[k]+beta*x[k])/(1+exp(exp(fi)*t[k]+beta*x[k]))
qi[k]<-1/(1+exp(exp(fi)*t[k]+beta*x[k]))
gi[k]<-1+exp(beta*x[k])
ri[k]<-exp(beta*x[k])/(1+exp(beta*x[k]))
si[k]<-1-ri[k]
U1[k]<-delta[k]+(exp(psi-fi))*log(qi[k]*gi[k])
U2[k]<-delta[k]*exp(fi)*t[k]*qi[k]-(exp(psi-fi))*log(qi[k]*gi[k])-exp(psi)*t[k]*pi[k]
U3[k]<-delta[k]*x[k]*qi[k]+(exp(psi-fi))*x[k]*(ri[k]-pi[k])
J1[k]<-(exp(psi)/exp(fi))*log(qi[k]*gi[k])
J2[k]<–(t[k]^2)*exp(2*fi)*pi[k]*qi[k]*delta[k]+delta[k]*exp(fi)*t[k]*qi[k]-exp(psi-
fi)*
log(qi[k]*gi[k])+exp(psi)*t[k]*pi[k]-exp(psi+fi)*t[k]*pi[k]*qi[k]
J3[k]<–delta[k]*(x[k]^2)*pi[k]*qi[k]+(exp(psi-fi))*(x[k]^2)*(ri[k]*si[k]-pi[k]*qi[k])
J12[k]<–exp(psi-fi)*log(qi[k]*gi[k])-exp(psi)*pi[k]
J13[k]<–delta[k]*exp(fi)*t[k]*x[k]*pi[k]*qi[k]-exp(psi-fi)*x[k]*(ri[k]-pi[k])-
exp(psi)*x[k]*pi[k]*qi[k]
J23[k]<–x[k]*t[k]*pi[k]*qi[k]*(delta[k]+exp(psi)/exp(fi))-(exp(psi)/exp(fi)^2)*
x[k]*(ri[k]-pi[k])
}
```

```
U[1]<-sum(U1)

U[2]<-sum(U2)

U[3]<-sum(U3)

J[1,1]<-sum(J1)

J[1,2]<-sum(J12)

J[2,1]<-sum(J12)

J[2,2]<-sum(J2)

J[3,3]<-sum(J3)

J[1,3]<-sum(J13)

J[3,1]<-sum(J13)

J[3,2]<-sum(J23)

J[2,3]<-sum(J23)

beta.res<-c(psi,fi,beta) +solve(-J)%*%U

}

solve(J)

# H.bayes<-raph(psi,fi,beta,x,t,delta)# trava o programa

beta.nr<-matrix(nrow=3,ncol=15)

beta.nr[,1]<-c(0.07,0.05,0.09)

for (j in 1:14){

beta.nr[,j+1]<-raph(beta.nr[1,j],beta.nr[2,j],beta.nr[3,j],x,t,delta)

for (k in 1:n){
```

```
pi[k]<-exp(exp(beta.nr[2,j])*t[k]+beta.nr[2,j]*x[k])/(1+exp(exp(beta.nr[2,j])*t[k]+
        beta.nr[2,j]*x[k]))
qi[k]<-1/(1+exp(exp(beta.nr[2,j])*t[k]+beta.nr[2,j]*x[k]))
gi[k]<-1+exp(beta.nr[2,j]*x[k])
ri[k]<-exp(beta.nr[2,j]*x[k])/(1+exp(beta.nr[2,j]*x[k]))
si[k]<-1-ri[k]
U1[k]<-delta[k]+(exp(beta.nr[1,j]-beta.nr[2,j]))*log(qi[k]*gi[k])
  U2[k]<-delta[k]*exp(beta.nr[2,j])*t[k]*qi[k]-(exp(beta.nr[1,j]-beta.nr[2,j]))*log(qi[k]*gi[k])-
exp(beta.nr[1,j])*t[k]*pi[k]
U3[k]<-delta[k]*x[k]*qi[k]+(exp(beta.nr[1,j]-beta.nr[2,j]))*x[k]*(ri[k]-pi[k])
J1[k]<-(exp(beta.nr[1,j])/exp(beta.nr[2,j]))*log(qi[k]*gi[k])
J2[k]<–(t[k]^2)*exp(2*beta.nr[2,j])*pi[k]*qi[k]*delta[k]+delta[k]*exp(beta.nr[2,j])*t[k]*qi[k]-
exp(beta.nr[1,j]-beta.nr[2,j])
    *log(qi[k]*gi[k])+exp(beta.nr[1,j])*t[k]*pi[k]-exp(beta.nr[1,j]+beta.nr[2,j])*t[k]*pi[k]*qi[k]
J3[k]<–delta[k]*(x[k]^2)*pi[k]*qi[k]+(exp(beta.nr[1,j]-beta.nr[2,j]))*(x[k]^2)*(ri[k]*
        si[k]-pi[k]*qi[k])
J12[k]<–exp(beta.nr[1,j]-beta.nr[2,j])*log(qi[k]*gi[k])-exp(beta.nr[1,j])*pi[k]
J13[k]<–delta[k]*exp(beta.nr[2,j])*t[k]*x[k]*pi[k]*qi[k]-exp(beta.nr[1,j]-beta.nr[2,j])*x[k]
          *(ri[k]-pi[k])-exp(beta.nr[1,j])*x[k]*pi[k]*qi[k] J23[k]<–x[k]*t[k]*pi[k]*qi[k]*
           (delta[k]+exp(beta.nr[1,j])/exp(beta.nr[2,j]))-
          (exp(beta.nr[1,j])/exp(beta.nr[2,j])^2)*x[k]*(ri[k]-pi[k])
```

```
}

U[1]<-sum(U1)

U[2]<-sum(U2)

U[3]<-sum(U3)

J[1,1]<-sum(J1)

J[1,2]<-sum(J12)

J[2,1]<-sum(J12)

J[2,2]<-sum(J2)

J[3,3]<-sum(J3)

J[1,3]<-sum(J13)

J[3,1]<-sum(J13)

J[3,2]<-sum(J23)

J[2,3]<-sum(J23)

beta.res<-c(beta.nr[1,j],beta.nr[2,j],beta.nr[2,j]) + solve(-J)%*%U

}

beta.nr

solve(J)

###

## CONVERGÊNCIA

###

## erro_(abs(theta-thetai))/thetai
```

###

C - Programa em WinBUGS usado para o ajuste Bayesiano do modelo Logístico Generalizado Reduzido com dados censurados à direita.

```
model
    {
    for(i in 1 : N) {
p[i] <- exp(alpha*t[i]+beta1*idade[i])/(1+exp(alpha*t[i]+beta1*idade[i]))
q[i] < -1/( 1+exp(alpha*t[i]+beta1*idade[i]))
g[i] < - 1+exp(beta1*idade[i])
S[i] <- pow((1+exp(alpha*t[i]+beta1*idade[i]))/(1+exp(beta1*idade[i])), -lambda/alpha)
L[i] <- pow(lambda*p[i], cen[i])*S[i]
}
            beta1 ~dnorm(0.0, 0.01)
          alpha ~dnorm(0.0, 0.01)
           lambda <- dnorm(0.0, 0.01)
}
```

D - Programa em WinBUGS usado para o ajuste Bayesiano do modelo de Cox com dados censurados à direita.

```
model
    {
    # Set up data
        for(i in 1:N) {
            for(j in 1:T) {
    # risk set = 1 if obs.t >= t
                Y[i,j] <- step(obs.t[i] - t[j] + eps)
    # counting process jump = 1 if obs.t in [ t[j], t[j+1] )
    # i.e. if t[j] <= obs.t < t[j+1]
                dN[i, j] <- Y[i, j] * step(t[j + 1] - obs.t[i] - eps) * fail[i]
            }
        }
    # Model
        for(j in 1:T) {
            for(i in 1:N) {
                dN[i, j] ~dpois(Idt[i, j]) # Likelihood
                Idt[i, j] <- Y[i, j] * exp(beta * idade[i]) * dL0[j]       # Intensity
            }
            dL0[j] ~dgamma(mu[j], c)
```

```
        mu[j] <- dL0.star[j] * c # prior mean hazard
    }
    c <- 0.001
    r <- 0.1
    for (j in 1 : T) { dL0.star[j] <- r * (t[j + 1] - t[j]) }
    beta ~dnorm(0.0,0.000001)
}
```

E - Obtenção da Função de Sobrevivência do Modelo Logístico Generalizado

Da equação,

$$h(t|x) = \lambda \frac{\exp(t\alpha + x'\beta)}{1 + \exp(t\alpha + x'\beta)}$$

pode-se obter a função de risco acumulada, sobre um intervado (t_1, t_2), dada por,

$$H(t_1, t_2|x) = \int_{t_1}^{t_2} h(u|x)du.$$

Fazendo $u = \exp(t\alpha + x'\beta)$ tem-se,

$$H(t_1, t_2|x) = \frac{\lambda}{a} \left\{ \ln \left[\frac{[1 + \exp(t_2\alpha + x'\beta)]}{[1 + \exp(t1\alpha + x'\beta)]} \right] \right\}.$$

Por definição, o evento específico da função de sobrevivência á dada por $\exp\{-H(t_1, t_2|\mathbf{x})\}$ da forma que,

$$S(t_1, t_2|x) = \left\{ \frac{1 + \exp(t_2\alpha + x'\beta)}{1 + \exp(t1\alpha + x'\beta)} \right\}^{-\lambda/\alpha}.$$

Se $t_1 = 0$, então,

$$S(t|x) = \left\{ \frac{1 + \exp(t\alpha + x'\beta)}{1 + \exp(x'\beta)} \right\}^{-\lambda/\alpha}.$$