INTERVALOS DE CREDIBILIDADE PARA A RAZÃO DE RISCOS DO MODELO DE COX, CONSIDERANDO ESTIMATIVAS PONTUAIS *BOOTSTRAP*

MARCELINO ALVES ROSA DE PASCOA

2008

MARCELINO ALVES ROSA DE PASCOA

INTERVALOS DE CREDIBILIDADE PARA A RAZÃO DE RISCOS DO MODELO DE COX, CONSIDERANDO ESTIMATIVAS PONTUAIS *BOOTSTRAP*

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

Orientador

Prof. Dr. Mário Javier Ferrua Vivanco

LAVRAS

MINAS GERAIS-BRASIL

2008

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Pascoa, Marcelino Alves Rosa de.

Intervalos de credibilidade para a razão de riscos do modelo de Cox, considerando estimativas pontuais *boostrap* / Marcelino Alves Rosa de Pascoa. – Lavras : UFLA, 2008.

71 p. : il.

Dissertação (Mestrado) Universidade Federal de Lavras, 2008
Orientador: Mário Javier Ferrua Vivanco.
Bibliografia.

1. Modelo de Cox. 2. Razão de riscos. 3. Boostrap. 4. Inferência bayesiana. I. Universidade Federal de Lavras. II.Título.

CDD - 519.546

MARCELINO ALVES ROSA DE PASCOA

INTERVALOS DE CREDIBILIDADE PARA A RAZÃO DE RISCOS DO MODELO DE COX, CONSIDERANDO ESTIMATIVAS PONTUAIS *BOOTSTRAP*

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

APROVADA em 15 de fevereiro de 2008.

Prof. Dr. Rogério de Melo Costa Pinto — UFU

Prof. Dr. João Domingos Scalon — UFLA

Prof. Dr. Mário Javier Ferrua Vivanco
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

Aos meus pais, Sebastião e Lucieni,

pela educação que me deram,

carinho, apoio.

AGRADECIMENTOS

Hoje eu sei, tenho muito a agradecer ...

Agradeço a Deus por ser minha força maior neste caminho, que sempre me ajudou e que está sempre comigo.

Ao meu irmão Daniel pelo apoio e amizade, presentes em todos os momentos da minha vida.

À Universidade Federal de Lavras, em especial ao Departamento de Ciências Exatas, pela oportunidade de concretização deste trabalho.

Ao professor Dr. Mário Vivanco, pela orientação, ensinamentos, amizade, conselhos, apoio e confiança durante os últimos dois anos.

A todos os professores que contribuíram para o enriquecimento dos conhecimentos indispensáveis a este trabalho.

Aos funcionários do Departamento de Ciências Exatas, em especial à Selminha, Edila, Josi e Joyce pela eficiência e amizade.

Aos meus amigos de mestrado. Jamais esquecerei de nossa convivência, dos momentos de risos, de choro e de estudos. Agradeço, também, aos amigos das outras turmas de mestrado e doutorado.

Em especial à Graziela e ao Fabricio por sempre estarem prontos a ajudar, pelo apoio durante esta caminhada e pela sincera amizade que demonstraram para comigo.

Ao inesquecível quarteto de probabilidade (quarteto fantástico), pelas horas de estudo, horas de sofrimentos e horas de bate papo entre verdadeiros amigos.

À CAPES e ao CNPq, pelo apoio financeiro.

MUITO OBRIGADO!!!

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

PASCOA, Marcelino Alves Rosa de. **Intervalos de credibilidade para a razão de riscos do modelo de Cox, considerando estimativas pontuais *bootstrap***. 2008. 71 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG.*

O objetivo deste trabalho foi propor uma alternativa de solução ao problema relacionado à interpretação da estimativa intervalar da razão de riscos do modelo de Cox, cuja interpretação fica ambígua quando o intervalo de confiança abrange a unidade, pois valores acima de 1 indicam sobre-risco e valores abaixo de 1 indicam proteção. Para alcançar tal objetivo, foram utilizados o método *bootstrap*, a inferência bayesiana e o algoritmo MCMC para a estimação dos intervalos de credibilidade, para o parâmetro β_j, relacionados a cada covariável selecionada pelo modelo de Cox. Respectivamente, por meio do método *bootstrap*, obtiveram-se novas estimativas do parâmetro dos coeficientes do modelo de regressão de Cox, β_j, $j = 1, ..., p$ relacionadas a cada covariável selecionada pelo modelo de Cox. A inferência bayesiana foi empregada considerando-se as estimativas pontuais *bootstrap*, de modo que a escolha da *priori* adequada se deu em função dessas estimavas. A função de verossimilhança adotada seguiu o modelo normal. A *priori* utilizada foi a normal truncada. Utilizou-se tal *priori* devido à necessidade de que os β_s assumam valores exclusivamente positivos ou estritamente negativos, dependendo da estimativa pontual *bootstrap* obtida. Desse modo, para resolver o problema da ambigüidade dos intervalos de confiança, construíram-se os intervalos de credibilidade, baseando-se no algoritmo Metropolis-Hastings. Concluiu-se que a aplicação da metodologia proposta neste trabalho obteve resultados satisfatórios, de forma que, no intervalo de credibilidade construído, obteveram-se valores que indicavam apenas sobre-risco ou valores que indicavam apenas proteção. Foram utilizados dados reais referentes a dependentes químicos e tempo de aleitamento materno. Concluiu-se também que os comprimentos dos intervalos de confiança e de credibilidade não têm qualquer diferença expressiva, quanto à sua amplitude.

*__Orientador:__ Prof. Dr. Mário Javier Ferrua Vivanco - UFLA.

ABSTRACT

PASCOA, Marcelino Alves Rosa de. **Credibility Intervals for the risk ratio of the Cox model, considering point *bootstrap* estimates**. 2008. 71 p. Dissertation (Master in Statistics and Agricultural Research) - Federal University of Lavras, Lavras, MG.*

The objective of this study was to propose an alternative solution to the problem of interpreting the interval estimation of the risk ratio of the model Cox, whose interpretation is misleading when the confidence interval covers the unit, because values over 1 indicate over-risk and values below 1 indicate protection. To achieve this goal the *bootstrap* method has been used, associated with Bayesian inference and the MCMC algorithm for estimating credibility intervals for the β_j parameter, related to each covariate selected by the Cox model. Respectively, through the *bootstrap* method we computed new estimates of the parameter of the coefficients of the Cox regression model, β_j, $j = 1, ..., p$ related to each covariate selected by the Cox model. Bayesian inference was used with the point *bootstrap* estimates, so that the choice of a suitable *priori* was base on such estimates. The adopted likelihood function followed the normal model. It was used a truncated normal as *priori*. The reason for using such *priori* is due to the need of the β_s to assume exclusively positive or negative values, depending on the point *bootstrap* estimate obtained. Thus, to solve the ambiguity problem of the confidence intervals, credibility intervals were built based on the Metropolis-Hastings algorithm. Results of proposed methodology were satisfactory, yielding credibility intervals indicating either over-risk or protection. Drug addict and breastfeeding time data were used at the practical example. In addition, it was detected that the lengths of the confidence and credibility intervals did not strongly differ.

*Adviser: Prof. Dr. Mário Javier Ferrua Vivanco - UFLA.

1 INTRODUÇÃO

Segundo Paz (2005), a análise de sobrevivência é o conjunto de técnicas e modelos estatísticos usados na análise do comportamento de variáveis positivas, tais como: tempo decorrido entre o início do tratamento até a morte do paciente, período de remissão de uma doença, tempo até o desenvolvimento de uma doença ou simplesmente tempo até a morte. Em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Esses eventos são, na maioria dos casos, indesejáveis e usualmente chamados de falhas. Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falhas, cuja característica importante é a presença de censura, que representa a observação parcial da resposta.

Para modelar os dados de sobrevivência existem três tipos de modelos, paramétricos, semiparamétrico e o não-paramétrico, sendo o modelo semiparamétrico de maior interesse para este trabalho.

O modelo de riscos proporcionais de Cox é aplicado para avaliar o efeito das covariáveis no risco de falha (Cox, 1972). Ele apresenta a vantagem de possuir uma interpretação simples dos resultados. Isto se deve ao fato de que a suposição básica da Razão de Riscos (RR), para dois indivíduos ou grupos de indivíduos homogêneos, não depende do tempo. Dessa forma, a razão de riscos é a mesma durante todo o período de acompanhamento do indivíduo ou grupo.

Diversos autores, como Collet (1994), Carvalho et al. (2005) e Colosimo & Giolo (2006), constroem intervalos de confiança para a Razão de Riscos a partir da estimação intervalar dos coeficientes "β" no modelo de Cox. Esse vetor de parâmetros é estimado pelo método de máxima verossimilhança parcial (Colosimo

& Giolo, 2006). Os intervalos de confiança para os parâmetros β são construídos da seguinte maneira $[\widehat{\beta} - 1.96\ e.p.(\widehat{\beta}); \widehat{\beta} + 1.96\ e.p.(\widehat{\beta})]$, em que $e.p.$ é o erro-padrão. É demonstrável que os estimadores de maxima verossimilhanca parcial têm propriedades desejáveis, como consistência e normalidade assintótica.

Para uma determinada covariável, X_j a RR_j é determinada a partir da seguinte relação: $RR_j = exp(\beta_j)$. Sendo assim, o intervalo de confiança para a RR é dado pela exponencial dos limites superior e inferior do intervalo de confiança obtido para os respectivos coeficientes β. No entanto, a interpretação da estimativa intervalar da RR fica ambígua quando o intervalo de confiança abrange a unidade. Nesses casos, valores acima de 1 indicam sobre-risco, valores abaixo de 1 indicam proteção. Por exemplo, se para uma determinada covariável W, o intervalo de confiança estimado para a razão de riscos é $[0,5; 2,3]$, a interpretação deste intervalo é ambíguo, pois temos valores acima de 1 que indicam sobre-risco e valores abaixo de 1 que indicam proteção.

O objetivo deste trabalho é propor uma alternativa de solução ao problema mencionado acima, de forma que em um mesmo intervalo não existam valores que indiquem proteção e sobre-risco.

2 REFERENCIAL TEÓRICO

Nesta seção, serão apresentados alguns conceitos que serão úteis para atingir o objetivo proposto neste trabalho.

2.1 O Modelo Normal

Esse modelo foi proposto, por volta de 1733, pelo matemático frânçes Abraham De Moivre. Ele, às vezes, é chamado de Gauss ou de De Moivre. Sua definição é dada a seguir:

- Dizemos que a variável aleatória X tem distribuição normal com média μ e variância σ^2, que denotamos por $X \sim N(\mu, \sigma^2)$, se a função de densidade de probabilidade de X é dada por:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \tag{2.1}$$

sendo μ e σ^2 ($\mu \, \epsilon \, \Re$, $\sigma^2 \, \epsilon \, \Re^+$) parâmetros do modelo que representam, respectivamente, a média e a variância.

$$\mu = E(X) \quad ou \quad \sigma^2 = Var(X)$$

A distribuição Normal tem grande importância na Estatística. Ela serve como modelo para quantidades de interesse em inferência estatística e também é usada em aproximações. Sua densidade é simétrica ao redor de μ e vai diminuindo a massa de probabilidade à medida que seus valores se movem para as extremidades (Magalhães, 2004).

2.2 Distribuição Truncada

Segundo Mood et al. (1974) truncamento pode ser definido, em geral, da seguinte maneira: se X é uma variável aleatória com função densidade $f_X(.)$ e função de distribuição $F_X(.)$, então, a densidade de X truncada à esquerda em a e à direita em b é determinada por:

$$f_{Truncada}(x) = \frac{f_X(x) I_{(a,b)}(x)}{F_X(b) - F_X(a)}. \tag{2.2}$$

Neste trabalho, a necessidade de truncar uma distribuição deve-se ao fato de querer representar variáveis aleatórias que assumem valores apenas num trecho da reta real, por distribuições que assumem valores em toda a reta real. Este é o caso dos coeficientes β do modelo de Cox, os quais, em determinado momento, precisamos que assumam valores estritamente positivos e em outros, valores estritamente negativos.

2.3 Censura

A análise de sobrevivência trata com dados obtidos a partir do segmento de indivíduos ou elementos desde um tempo inicial até a ocorrência de um determinado evento. Acontece que, por diversos motivos, não é para todos os indivíduos que o evento ocorre. A informação obtida naqueles elementos ou indivíduos em que o evento não ocorre é apenas parcial e é chamada de “dado censurado", ou, simplesmente, censura.

Por exemplo, se estamos acompanhando um paciente desde o momento em que realizou uma determinada cirurgia até a morte e o tempo de estudo é de dez

anos, se, ao término dos dez anos, o paciente não morreu, diremos que o evento (morte) não aconteceu. Portanto, o dado registrado para esse paciente será um dado censurado.

De acordo com Colossimo & Giolo (2006), mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida de pacientes e a omissão de censuras no cálculo das estatísticas de interesse pode acarretar em conclusões viciadas.

Segundo Lawless (1982), existem três tipos de censuras mais utilizadas. São elas:

- censura à direita: aquela em que o tempo até a ocorrência do evento é maior que um determinado tempo “C";

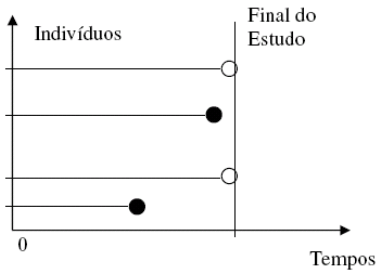

FIGURA 2.1: Representação gráfica de censura à direita, em que • representa a falha e ◦ a censura.

- censura à esquerda: ocorre quando o tempo registrado é maior que o tempo de falha, isto é, o evento de interesse já aconteceu quando o indivíduo foi observado;

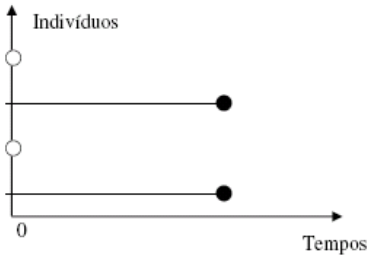

FIGURA 2.2: Representação gráfica de censura à esquerda, em que • representa a falha e ◦ a censura.

- censura aleatória: acontece quando um item é retirado no decorrer do evento, sem que a falha tenha ocorrido.

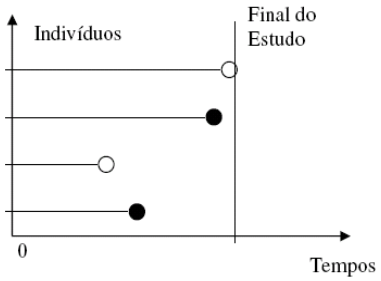

FIGURA 2.3: Representação gráfica de censura aleatória, em que • representa a falha e ◦ a censura.

2.4 Especificação do tempo de falha

Em análise de sobrevivência, o tempo de falha, representado pela variável aleatória não-negativa T, pode ser especificado tanto pela função de sobrevivência quanto pela função de taxa de falha (ou risco).

2.4.1 Função de sobrevivência

Suponha que a variável aleatória T tenha uma distribuição de probabilidade com função de densidade de probabilidade $f(t)$. A função de distribuição de T é então, dada por:

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(x)dx,$$

e representa a probabilidade de que o tempo de sobrevivência seja menor que algum valor t.

A função de sobrevivência é uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência. Ela é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de um indivíduo sobreviver ao tempo t. Em termos probabilísticos, isto é escrito como:

$$S(t) = P(T \geq t) \tag{2.3}$$

Conseqüentemente, a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver ao tempo t, isto é, $F(t) = 1 - S(t)$.

A função de sobrevivência tem as seguintes propriedades:

1. é uma função monótona decrescente;
2. é contínua no tempo;
3. $S(0) = 1$, isto é, a probabilidade de sobreviver ao tempo zero é um;
4. $\lim_{t\to\infty} S(t) = 0$, isto é, a probabilidade de sobreviver no tempo infinito é zero.

2.4.2 Função de taxa de falha ou de risco

A função risco é bastante utilizada na análise de sobrevivência, devido ao fato de que a sua interpretação ajuda a entender a distribuição do tempo de vida de pacientes. Isto é possível pois a mesma descreve a forma como a taxa instantânea de falha muda com o tempo.

A função de risco, $\lambda(t)$, é definida como o risco instantâneo de um indivíduo sofrer o evento entre o tempo t e $t + \Delta t$, dado que ele sobreviveu até o tempo t. Para se obter uma definição formal da função de risco, considere um intervalo de tempo $[t, t + \Delta t)$. A probabilidade da falha ocorrer neste intervalo pode ser expressa, em termos da função de sobrevivência, como:

$$S(t) - S(t + \Delta t).$$

A taxa de falha no intervalo $[t, t + \Delta t)$ é definida como o risco de que a falha ocorra neste intervalo, dado que não ocorreu antes de t, dividida pelo comprimento do intervalo. Dessa forma, a taxa de falha no intervalo $[t, t + \Delta t)$ é expressa por:

$$\frac{S(t) - S(t + \Delta t)}{[(t + \Delta t) - t]S(t)}.$$

De forma geral, $\lambda(t)$ pode ser escrita como:

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Para Δt pequeno, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. Deve-se atentar para o fato de que as taxas de falha são números positivos, sem limite superior.

Matematicamente, a definição de risco pode ser expressa pela fórmula:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} \tag{2.4}$$

A função risco pode ser definida, em termos da função de distribuição $F(t)$ e da função de densidade de probabilidade $f(t)$, da seguinte forma:

$$\lambda(t) = \frac{f(t)}{1 - F((t)} = \frac{f(t)}{S(t)}. \tag{2.5}$$

A função taxa de falha acumulada é dada por:

$$\Lambda(t) = \int_0^t \lambda(x) dx.$$

2.5 Modelo de riscos proporcionais de Cox

2.5.1 Introdução

Os primeiros modelos de regressão para análise de sobrevida foram desenvolvidos na década de 1960 (Harris & Albert, 1991) e eram totalmente paramétricos, ou seja, baseados nas premissas de validade da estatística tradicional. Em 1972, Cox desenvolveu um modelo de regressão semiparamétrico, também conhecido como modelo de riscos proporcionais de Cox, modelo de Cox, ou regressão de Cox (Cox, 1972).

2.5.2 O modelo de Cox

O modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustado por covariáveis (Colosimo & Giolo, 2006). Sob a suposição de riscos proporcionais, Cox propôs, em 1972, o modelo de riscos proporcionais de Cox, cuja expressão geral é dada por:

$$\lambda(t) = \lambda_0(t) g(X^{'}\beta), \tag{2.6}$$

em que g é uma função não negativa que deve ser especificada, tal que $g(0) = 1$, β é o vetor de parâmetros regressores associado com as covariáveis e X é o vetor de covariáveis, $X = (x_1, ..., x_p)^{'}$. Este modelo é composto pelo produto de dois componentes de naturezas distintas, um não-paramétrico e outro paramétrico, o que o torna bastante flexível.

Um exemplo da flexibilidade deste modelo é envolver alguns modelos conhecidos como casos particulares, tal como o modelo de regressão weibull (Kalbfleisch & Prentice, 1980). O componente não-paramétrico $\lambda_0(t)$ não é especificado e é uma função não negativa do tempo. Ele é usualmente chamado de função de base ou basal. A parte paramétrica do modelo, $g(X^{'}\beta)$, assume, geralmente, a forma $exp(X^{'}\beta)$ (Cox, 1972).

O modelo de regressão de Cox é caracterizado pelos coeficientes β_s que medem os efeitos das covariáveis sobre a função taxa de falha. Dessa maneira, é necessário um método de estimação para se fazer inferências no modelo. O método de máxima verossimilhança usual, bastante conhecido e freqüentemente usado (Cox & Hinkley, 1974), não pode ser utilizado aqui, pois a presença do compo-

nente não paramétrico $\lambda_0(t)$ na função de verossimilhança o torna inapropriado. Frente a tal dificuldade, Cox (1975) propôs o método de máxima verossimilhança parcial, que condiciona a verossimilhança à história dos tempos de sobrevivência e censuras anteriores e, dessa forma, elimina a função base desconhecida $\lambda_0(t)$.

2.5.3 Método de máxima verossimilhança parcial

Nos intervalos de tempo em que nenhuma falha ocorre, não existe nenhuma informação sobre o vetor de parâmetros β, pois $\lambda_0(t)$ pode, teoricamente, ser identicamente igual a zero em tais intervalos. Uma vez que é necessário um método de análise válido para todas $\lambda_0(t)$ possíveis, a consideração de uma distribuição condicional é necessária.

Considere que, em uma amostra de n indivíduos, existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 < ... < t_k$. A probabilidade condicional da i-ésima observação vir a falhar no tempo t_i, conhecendo quais observações estão sob risco em t_i, é:

$$\frac{\lambda_i(t|X_i)}{\sum_{j\in R(t_i)}\lambda_j(t|X_j))} = \frac{\lambda_0(t)exp(X_i'\beta)}{\sum_{j\in R(t_i)}\lambda_0(t)exp(X_j'\beta)} = \frac{exp(X_i'\beta)}{\sum_{j\in R(t_i)}exp(X_j'\beta)}, \tag{2.7}$$

em que $R(t_i)$ é o conjunto dos índices dos indivíduos sob risco no tempo t_i. Pode-se verificar que, ao utilizar a probabilidade condicional, o componente não-paramétrico $\lambda_0(t)$ desaparece da equação (2.7).

A função de verossimilhança parcial $L(\beta)$ é obtida fazendo o produto dessas

probabilidades condicionais, associadas aos distintos tempos de falha, ou seja,

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(X_i'\beta)}{\sum_{j \in R(t_i)} exp(X_j'\beta)} = \prod_{i=1}^{k} \left(\frac{exp(X_i'\beta)}{\sum_{j \in R(t_i)} exp(X_j'\beta)} \right)^{\delta_i}, \qquad (2.8)$$

em que $\delta_i = 0$, se o i-ésimo tempo de sobrevivência é censurado e 1, caso contrário.

2.5.4 Estimação de parâmetros

Sendo a função $l(\beta)$ obtida pelo logaritmo da função de verossimilhança parcial, ou seja, $l(\beta) = log(L(\beta))$ e $U(\beta)$ é o vetor escore de derivadas de primeira ordem da função $l(\beta)$. Estimadores para o vetor de parâmetros β podem ser obtidos maximizando-se o logaritmo da função de verossimilhança parcial (2.8), ou seja, resolvendo o sistema de equações definido por $U(\beta) = 0$. Isto é,

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in R(t_i)} x_j exp(X_j'\widehat{\beta})}{\sum_{j \in R(t_i)} exp(X_j'\widehat{\beta})} \right] \qquad (2.9)$$

O procedimento de estimação requer um método interativo que é, geralmente, o método de Newton-Raphson, pois as equações encontradas em (2.9) não apresentam forma fechada.

Cox (1975) mostra informalmente que o método usado para construir esta verossimilhança gera estimadores que são consistentes e assintóticamente normalmente distribuídos, com matriz de covariâncias assintóticamente estimadas consistentemente pelo inverso do negativo da matriz de segundas derivadas parciais do logaritmo da função de verossimilhança. Provas formais dessas propriedades foram apresentadas por Andersen & Gill (1982).

2.5.5 Razão de riscos

Considere o modelo $\lambda(t) = \lambda_0(t)exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$. A razão de riscos é um parâmetro. Ela é medida em apenas uma covariável, na qual é possível separar grupos em função de um ou mais pontos de corte. Por exemplo, considere a covariável X_1. Essa covariável está dicotomizada em 0 e 1. A razão de riscos neste caso é dada por:

$$RR = \frac{Risco\ (grupo\ 0)}{Risco\ (grupo\ 1)}.$$

A expressão nos dá um valor, por exemplo, 2. Neste caso, temos que: Risco (grupo 0) = 2 Risco (grupo 1). Ou seja, o risco para o grupo 0 é duas vez maior que o risco para o grupo 1.

A relação existente entre a razão de riscos e os parâmetros do modelo de Cox se dá por meio da seguinte expressão:

$$RR = exp(\beta),$$

em que RR e β são, respectivamente, a razão de riscos e o coeficiente associado à covariável em estudo. Desse modo, valores de β negativos implicam proteção e valores de β positivos indicam sobre-risco.

2.5.6 Teste da razão de verossimilhanças

O teste da razão de verossimilhanças é baseado na função de verossimilhança. Ele envolve a comparação dos valores do logaritmo da função de verossimilhança

maximizada sem restrição e sob H_0, ou seja, a comparação de $logL(\widehat{\beta})$ e $logL(\widehat{\beta}_0)$. Sua estatística é dada por:

$$TRV = -2log\left[\frac{L(\widehat{\beta}_0)}{L(\widehat{\beta})}\right] = 2\left[logL(\widehat{\beta}) - logL(\widehat{\beta}_0)\right],$$

essa estatística segue, assintoticamente, uma distribuição qui-quadrado (χ^2) com número de graus de liberdade igual ao número de parâmetros que estão sendo testados.

2.6 Bootstrap

O *bootstrap* é uma técnica estatística computacionalmente intensiva que permite a avaliação da variabilidade de estatísticas, com base nos dados de uma única amostra existente. Essa técnica foi introduzida por Efron (1979) e, desde então, tem merecido profundo estudo por parte dos estatísticos, não só na parte teórica, como também na aplicada.

O *bootstrap* pode ser implementado tanto na estatística não-paramétrica quanto na paramétrica, dependendo apenas do conhecimento do problema. No caso não-paramétrico, o método *bootstrap* reamostra os dados com reposição, de acordo com uma 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 de distribuição dos dados, a amostra *bootstrap* é formada realizando-se a amostragem diretamente nessa distribuição com os parâmetros desconhecidos substituídos por estimativa paramétricas. A distribuição da estatística de interesse aplicada aos valores da amostra *bootstrap*, condicional aos dados observados, é definida como a distribuição *bootstrap* dessa estatística

(Lavoranti, 2003).

A técnica consiste em se retirar uma amostra de tamanho n da população e reamostrá-la com reposição, obtendo-se uma nova amostra de tamanho n da amostra original. Cada uma das amostras obtidas pelas reamostragens é uma amostra *bootstrap*. Esse procedimento é executado milhares de vezes, obtendo-se, assim, as estimativas dos parâmetros que serão usadas para gerar a distribuição denominada distribuição *bootstrap*. Esta distribuição é uma estimativa da verdadeira densidade populacional.

O número de reamostras *bootstrap* sugerido na literatura parece crescer com o avanço computacional, seja para hardware e ou software. Efron & Tibshirani (1993) comentam que $B = \infty$ é o número ideal de reamostras *bootstrap*. Naturalmente, na prática, B deve ser um número finito restrito ao poder computacional disponível.

2.7 Inferência bayesiana

A inferência bayesiana é o processo de encontrar um modelo de probabilidade para um conjunto de dados e resumir o resultado por uma distribuição de probabilidades sobre os parâmetros do modelo e sobre quantidades não observadas, tais como predição para novas observações (Gelman et al., 2000).

Devido à grande versatilidade na resolução de problemas, nunca antes solucionados, a metodologia bayesiana é, atualmente, um dos principais assuntos da comunidade científica envolvida com o desenvolvimento e a aplicação de procedimentos estatísticos.

A abordagem bayesiana requer um modelo amostral, a função de verossimilhança, que é representada por $L(\theta|y_1, ..., y_n)$, sendo n o número de obser-

vações e, em adição, uma distribuição a *priori*, dada por $\pi(\theta)$, para os parâmetros (Carlin & Louis, 1996). Parâmetros desconhecidos são considerados aleatórios e todas as conclusões estão baseadas na distribuição condicional dos parâmetros em relação aos dados observados, a distribuição a *posteriori*, expressa por $\pi(\theta|Y)$. Todo processo de inferência bayesiana tem como base o Teorema de Bayes, que é apresentado a seguir:

Para θ contínuo,

$$\pi(\theta|Y) = \frac{L(\theta|Y)\pi(\theta)}{\int_{\forall\theta} L(\theta|Y)\pi(\theta)d\theta}, \tag{2.10}$$

para θ discreto, temos:

$$\pi(\theta|Y) = \frac{L(\theta|Y)\pi(\theta)}{\sum_{\forall\theta} L(\theta|Y)\pi(\theta)}, \tag{2.11}$$

sendo $Y = \{y_1, y_2, ..., y_n\}$. O denominador, tanto para (2.10) quanto para (2.11), não depende de θ logo, temos:

$$\pi(\theta|Y) \propto L(\theta|Y)\pi(\theta), \tag{2.12}$$

em que $\propto$ representa proporcionalidade.

2.7.1 Distribuições a *priori*

A distribuição a *priori* tem um importante papel na análise bayesiana, sendo usada para descrever uma informação sobre os parâmetros desconhecidos antes que se possa avaliar os dados em questão (Box & Tião, 1992). Em outras palavras, ela é uma crença subjetiva e expressa o conhecimento que o pesquisador tem a respeito do problema, sendo muitas vezes baseada simplesmente na expectativa

subjetiva do investigador (ou estatístico), causando assim uma controvérsia pertinente de pesquisador para pesquisador.

Quando, em determinado estudo, o pesquisador tem alguma informação prévia sobre o que se está estudando, ou seja, ele tem conhecimento sobre o parâmetro θ, ele pode usar uma *priori* informativa. Se há pouca ou nenhuma informação para incorporar a *priori*, considera-se uma distribuição não-informativa, por exemplo, a *priori* de Jeffreys (Jeffreys, 1961). Uma família de distribuições a *priori* é conjugada se as distribuições a *posteriori* pertencem à mesma família de distribuições.

2.7.2 Função de verossimilhança

O conceito de função de verossimilhança é enunciado a seguir, segundo Bolfarine & Sandoval (2001):

- sejam $y_1, y_2, ..., y_n$ uma amostra aleatória de tamanho n da variável aleatória Y com função de densidade (ou de probabilidade) $f(y|\theta)$, com $\theta \, \epsilon \, \Theta$, sendo Θ é o espaço paramétrico. A função de verossimilhança de θ correspondente à amostra aleatória observada dada por:

$$L(\theta|Y) = \prod_{i=1}^{n} f(y_i|\theta).$$

A função de verossimilhança tem papel fundamental como veículo portador da informação dada pela amostra. Assim, o princípio da verossimilhança sustenta que toda a informação dada pela amostra ou pela experiência está contida na função de verossimilhança (Paulino et al., 2003).

2.7.3 Distribuição a *posteriori*

A distribuição a *posteriori* $\pi(\theta|Y)$ é a descrição completa do conhecimento corrente de θ obtido da quantificação à informação a *priori* em θ e da informação amostral em $L(\theta|Y)$.

O gráfico desta distribuição é a melhor descrição do processo inferêncial, Estas informações podem ser resumidas por meio de valores numéricos, tais como: média, mediana, moda e intervalos de credibilidade.

2.7.4 Intervalo de credibilidade

Em inferência bayesiana, em geral, temos interesse em determinar uma região R (do espaço paramétrico) para o qual a probabilidade de conter a densidade a *posteriori* para um determinado parâmetro θ é $\gamma = 1 - \alpha$ (Louzada Neto, 1991).

Como discutido por Box & Tiao (1992), tal região é chamada de região a *posteriori* de maior densidade, ou "highest posterior density"(HPD), com probabilidade $100(1 - \alpha)\%$ se,

$$\int_R \pi(\theta|Y)d\theta = \gamma,$$

em que R é a menor região do espaço paramétrico com probabilidade γ; $\pi(\theta|Y)$ é a densidade a *posteriori* e Y são os dados.

2.8 Métodos de Monte Carlo baseado em cadeias de Markov (MCMC)

Os métodos computacionais de Monte Carlo via Cadeias de Markov têm sido largamente usados na estatística bayesiana, possibilitando simular amostras de uma determinada densidade a *posteriori* $\pi(\theta|Y)$, cuja geração direta é difícil ou complicada.

A idéia básica do método é construir uma cadeia de Markov com distribuição de equilíbrio igual à de interesse; nessa cadeia, cada estado pode ser atingido a partir de qualquer outro com um número finito de iterações. Após um número suficientemente grande de iterações, a cadeia converge para uma distribuição de equilíbrio, dando origem à amostra da distribuição de interesse, que pode ser usada para fazer inferências.

Existem vários métodos para a construção da cadeia de Markov. Dentre eles, os mais utilizados são o método de Metropolis-Hastings e o amostrador de Gibbs. Gamerman (1997) apresenta uma descrição detalhada dos métodos de simulação baseados nos métodos de Monte Carlo via Cadeias de Markov.

2.8.1 Metropolis-Hastings

O algoritmo de Metropolis-Hastings é um método de simulação baseado em cadeias de Markov e foi descrito, pela primeira vez, por Hastings (1970), como uma generalização do algoritmo de Metropolis, que foi desenvolvido por Metropolis et al. (1953). Ele está estruturado nos seguintes passos:

1. inicializar o contador de iterações $t = 0$ e especificar valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$;

2. gerar um novo valor θ^c da distribuição proposta $q(.|\theta_1)$;

3. calcular a probabilidade de aceitação $\alpha(\theta_1, \theta^c)$ e gerar $u \sim U(0,1)$;

$$\alpha(\theta_1, \theta^c) = min\left\{1, \frac{\pi(\theta^c|\theta_2, ..., \theta_p)q(\theta_1|\theta^c)}{\pi(\theta_1|\theta_2, ..., \theta_p)q(\theta^c|\theta_1)}\right\}$$

4. se $u \leq \alpha$, então, aceitar o novo valor e fazer $\theta_1^{(t+1)} = \theta^c$. Caso contrário, rejeitar e fazer $\theta_1^{(t+1)} = \theta^t$;

5. incrementar o contador de t para $t+1$ e voltar ao passo 2, até atingir a convergência.

Este algoritmo permite gerar uma amostra da distribuição conjunta a *posteriori* $\pi(\theta_1, \theta_2, ..., \theta_p|y)$, sendo p parâmetros, a partir das distribuições condicionais completas com formas desconhecidas. Eles usam a idéia de que um valor é gerado de uma distribuição auxiliar ou candidata e este é aceito com uma dada probabilidade (Metropolis et al., 1953; Hastings, 1970).

2.8.2 Avaliação da convergência dos métodos de Monte Carlo via cadeias de Markov (MCMC)

Para a avaliação da convergência da cadeia de Markov para o estado de equilíbrio, vários testes foram propostos. No entanto, não há um método que se possa dizer ser o melhor ou mais eficiente e a utilização de métodos diferentes para o mesmo problema pode conduzir a respostas bastante diferentes. Dentre eles, os mais relevantes são as de Geweke (1992), Gelman & Rubin (1992), Heidelberg & Welch (1993) e Raftery-Lewis (1992). Estes testes estão implementados no pacote BOA (*Bayesian Output Analysis*) do software livre R (R Development Core Team, 2007).

O método de Geweke (1992) propõe o diagnóstico de convergência para cadeias de Markov, baseado no teste de igualdade de médias da primeira e última parte da cadeia de Markov, geralmente dos primeiros 10% e dos últimos 50%. Ele se baseia em técnicas de análise espectral.

O critério de Gelman & Rubin (1992) pressupõe que *m* cadeias tenham sido geradas em paralelo. Este analisa as variâncias dentro e entre as cadeias e utiliza esta informação para estimar o fator pelo qual o parâmetro escalar da distribuição marginal a *posteriori* deveria ser reduzido se a cadeia fosse repetida infinitas vezes. Este fator é expresso pelo valor $\sqrt{\widehat{R}}$ (fator de redução de escala potencial ou fator de diagnóstico da convergência) e sugere que valores de $\sqrt{\widehat{R}}$ próximos a 1 indicam que a convergência foi atingida para *n* iterações.

O método de Raftery & Lewis (1992) é baseado na acurácia de estimação do quantil. O método fornece as estimativas do *burn-in*, que é o número de iterações que devem ser descartadas, o número de iterações que devem ser computadas e o k, que é a distância mínima de uma iteração à outra para se obter a subamostra aproximadamente independente (*thin*).

O método de Heidelberger & Welch (1993), por meio de testes estatísticos, testa a hipótese nula de estacionariedade da amostra gerada. Se a hipótese nula é rejeitada para um dado valor, o teste é repetido depois de descartadas os primeiros 10% das iterações. Se a hipótese é novamente rejeitada, outros 10% são descartados após o descarte dos 10% primeiros. Este processo é repetido até se ter uma proporção de 50% dos valores iniciais descartados. Se a hipótese for novamente rejeitada ou o teste não conseguir ser realizado, isso indica falha da estacionariedade, implicando que é necessário um número maior de iterações. Se o teste for satisfatório, o número inicial de iterações descartadas é indicado como o tamanho do *burn-in* (Nogueira, 2004).

A verificação informal da convergência utilizando técnicas gráficas é também um procedimento bastante útil. Segundo Gelman et al. (2000), os gráficos mais freqüentes nesta análise são o gráfico de θ ao longo das iterações e um gráfico da estimativa da distribuição a *posteriori* de θ, por exemplo, um histograma. Problemas podem ser verificados com estas técnicas, pois a utilização de gráficos como única ferramenta pode não ser adequada. A escala dos gráficos pode fornecer uma falsa impressão de igualdade.

Nogueira (2004), após realizar um intenso estudo de avaliação dos critérios de convergência, concluiu que, para uma avaliação precisa da convergência, deve-se seguir o seguinte procedimento:

1. aplicar Raftery & Lewis em uma amostra piloto e determinar o tamanho ideal da seqüência;

2. determinar o tamanho do *burn-in* pelo critério de Heidelberger & Welch;

3. monitorar a convergência das seqüências nas proximidades do tamanho ideal, indicado pelo critério de Raftery & Lewis, por meio dos critérios de Gelman & Rubin e Geweke.

3 MATERIAL E MÉTODOS

3.1 Material

Para a realização deste trabalho foram avaliados dois conjuntos de dados distintos.

O primeiro conjunto de dados foi fornecido pela Associação Mãe Admirável, situada na cidade de Caratinga, MG. Foram avaliados 141 residentes, dependentes químicos, no período de 2000 a 2005. A variável resposta foi o tempo de permanência na comunidade até a desistência do tratamento, considerando que cada residente permanece na Comunidade por um período máximo de 270 dias, sem qualquer contato com as drogas e quem alcança esta meta foi considerado, neste trabalho, como um dado censurado. De cada paciente foram obtidas as informações dispostas na Tabela 3.1.

O segundo conjunto de dados utilizado neste trabalho foi retirado do livro de Colosimo & Giolo (2006), que apresenta um estudo realizado pelos professores Eugênio Goulart e Cláudia Lindgren, do Departamento de Pediatria da Universidade Federal de Minas Gerais, no Centro de saúde São Marcos, em Belo Horizonte, MG. O estudo foi realizado com os objetivos principais de conhecer a prática do aleitamento materno de mães que utilizam este centro, assim como os possíveis fatores de risco ou de proteção para o desmame precoce. Um inquérito epidemiológico composto por questões demográficas e comportamentais foi aplicado a 150 mães de crianças menores de 2 anos de idade. A variável resposta de interesse foi estabelecida como sendo o tempo máximo de aleitamento materno, ou seja, o tempo contado a partir do nascimento até o desmame completo da criança.

TABELA 3.1: Descrição das covariáveis usadas no estudo de dependência química.

Dados	**Legenda**	**Situação**
Tempo de tratamento (em dias)	tempo	
Censura	cens	0 (censurado) ou 1 (falha)
Problemática	P	0 (álcool ou outro TD) ou 1 (álcool e outro TD)
Estado civil do residente	EC	0 (solteiro) ou 1 (casado/separado)
Se tem filhos	F	0 (não) ou 1 (sim)
Estado civil dos pais do residente	EP	0 (casados) ou 1 (separados)
Fonte de renda própria	FR	0 (não) ou 1 (sim)
Nível de escolaridade	ES	0 (nenhuma ou FI) ou 1 (FC a superior)
Idade (em anos)	I	$1 \geq 30$ anos e $0 < 30$ anos
Diversidade de drogas	DD	0 (uma) ou 1 (mais de uma)
Caso de droga/álcool na família	CF	0 (não) ou 1 (sim)
Tratamento anterior	TA	0 (não) ou 1 (sim)
Problemas com a justiça	PJ	0 (não) ou 1 (sim)
Tempo de uso (em anos)	TU	$1 \geq 15$ anos e $0 < 15$ anos

Sendo: TD = tipo de droga; FI = ensino fundamental incompleto; FC = ensino fundamental completo e Superior = ensino superior completo ou não. Os dados correspondentes a este estudo encontram-se no anexo A deste trabalho.

Nesse estudo foram registradas 11 covariáveis e a variável resposta. Algumas crianças não foram acompanhadas até o desmame e, portanto, registra-se a presença de censuras. As variáveis analisadas encontram-se na Tabela 3.2, das quais 11 são covariáveis (fatores), uma é variável resposta, representada pelo tempo de acompanhamento e uma, a variável indicadora de ocorrência do desmame.

TABELA 3.2: Descrição das covariáveis usadas no estudo sobre aleitamento materno em Belo Horizonte (Colosimo & Giolo, 2006).

Dados	**Legenda**	**Situação**
Tempo até o desmame (meses)	tempo	
Censura	cens	0 (censurado) ou 1 (falha)
Experiência anterior amamentação	V1	0 se sim e 1 se não
Número de filhos vivos	V2	0 se $\geqslant 2$ e 1 se < 2
Conceito materno sobre o tempo ideal de amamentação	V3	0 se > 6 meses e 1 se $\leqslant 6$ meses
Dificuldades para amamentar nos primeiros dias pós-parto	V4	0 se não e 1 se sim
Tipo de serviço em que realizou o pré-natal	V5	0 se público e 1 se privado/convênios
Recebeu exclusivamente leite materno na maternidade	V6	0 se sim e 1 se não
A criança teve contato com o pai	V7	0 se sim e 1 se não
Renda per capita (em SM/mês)	V8	0 se $\geqslant 1$ SM e 1 se < 1 SM
Peso ao nascimento	V9	0 se $\geqslant 2,5$Kg e 1 se $< 2,5$Kg
Tempo de separação mãe-filho pós-parto	V10	0 se $\leqslant 6$ horas e 1 se > 6 horas
Permanência no berçário	V11	0 se não e 1 se sim

3.2 Métodos

Inicialmente, foi feito o ajuste com base no modelo semiparamétrico de Cox, apresentado na Seção 2.5, para a seleção de covariáveis dos dois conjuntos de dados descritos anteriormente, de acordo com a proposta de Collet (1994). Este autor defende que o estatístico e o pesquisador devem ter uma postura pró-ativa neste processo. Isto implica, por exemplo, que covariáveis importantes em termos clínicos devem ser incluídas independente de significância estatística, assim como

a importância clínica deve ser considerada em cada passo de inclusão ou exclusão no processo de seleção de covariáveis. Os passos utilizados no processo de seleção são apresentados a seguir.

1. Ajusta-se todos os modelos contendo uma única covariável. Inclui-se todas as covariáveis que forem significativas ao nível de 10%, utilizando o teste da razão de verossimilhanças neste passo.

2. As covariáveis significativas no passo 1 são então ajustadas conjuntamente. Na presença de certas covariáveis, outras podem deixar de ser significativas. Conseqüentemente, ajusta-se modelos reduzidos, excluindo uma única covariável. Verifica-se as covariáveis que provocam um aumento estatisticamente significativo na estatística da razão de verossimilhanças. Somente aquelas que atingiram a significância permanecem no modelo.

3. As eventuais covariáveis significativas no passo 2 são incluídas ao modelo. Neste passo retorna-se com as covariáveis excluídas no passo 1 para confirmar que elas não são estatisticamente significativas.

4. O modelo final fica determinado pelos efeitos principais identificados no passo 3.

Ao utilizar este procedimento de seleção, deve-se incluir as informações clínicas no processo de decisão e evitar ser muito rigoroso ao testar cada nível individual de significância.

Como discutido na Seção $2.5.5$, a suposição de riscos proporcionais deve ser atendida para que o modelo de Cox possa ser considerado adequado aos dados em estudo. Neste caso, para fazer a verificação da proporcionalidade, foi utilizado o gráfico do tempo versus $log\left(\widehat{\Lambda}_0(t)\right)$, em que $\widehat{\Lambda}_0(t)$ é uma estimativa simples para

a função de taxa de falha acumulada $\Lambda_0(t)$, proposta por Breslow (1972), que é expressa por:

$$\widehat{\Lambda}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l\epsilon R_j} exp\left\{Y_l^{'}\widehat{\beta}\right\}},$$

sendo d_j o número de falhas em t_j, $\widehat{\beta}$ é a estimativa do vetor de parâmetros associado às covariáveis e Y o vetor de covariáveis com os componentes $Y = (y_1, ..., y_p)^{'}$.

Após verificar a proporcionalidade dos riscos, ajustou-se o modelo de Cox, dado pela expressão $\lambda(t) = \lambda_0(t)g(X^{'}\beta)$ para os dados em estudo, obtendo-se assim uma estimativa de β_j (j: 1, ..., p-covariáveis), associada a cada covariável selecionada. O método *bootstrap*, não-paramétrico foi então empregado para se obter novas estimativas do parâmetro β_j relacionadas a cada covariável selecionada pelo modelo de Cox. Foram utilizadas neste trabalho $B = 10.000$ reamostragens *bootstrap* para cada conjunto de dados analisado. Ou seja, para cada covariável, obteve-se uma amostra de 10.000 $\widehat{\beta}_s$ estimados pelo método *bootstrap*.

Foi empregada a metodologia bayesiana, para a construção de intervalos de credibilidade para o vetor de parâmetros β_j referente a cada covariável. Para que fosse possível realizar inferências fundamentadas em determinado modelo, foi necessário obter a função de verossimilhança e especificar as distribuições a *priori* para os parâmetros de interesse.

3.2.1 Função de verossimilhança

A função de verossimilhança adotada neste trabalho segue um modelo normal, com parâmetros β_j e σ^2. Utilizou-se o modelo normal pelo fato de a amostra de da-

dos ter como elementos as estimativas de máxima verossimilhança dos coeficientes β. Tais estimativas foram obtidas a partir das amostras *bootstrap* da amostra de dados original. Como demonstram Bolfarine & Sandoval (2001), os estimadores de máxima verossimilhança são assintoticamente normais. Desta forma a função de verossimilhança para o modelo normal, dado em (2.1), é expressa por:

$$L_j(\beta_j, \sigma^2|Y_j) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{ij} - \beta_j)^2\right\}. \qquad (3.1)$$

Sendo Y_j o vetor dos dados associado a cada covariável, em que j: 1, ..., p-covariáveis.

3.2.2 Distribuições a *priori*

Neste estudo, optou-se por utilizar a normal truncada como *priori*, pois a normal é uma distribuição que assume valores no intervalo de $(-\infty, \infty)$ e, devido à relação existente entre a RR_j e o respectivo coeficiente β_j, detalhada na Seção (2.5.5), foi necessário obter intervalos para os $\beta_s^{'}$ acima de zero e para $\beta_s^{'}$ abaixo de zero (não nulos). Por isso, decidiu-se truncar a distribuição normal para $\beta_j > 0$ e para $\beta_j < 0$. Essa densidade foi truncada de acordo com a expressão (2.2).

- *Priori* normal truncada para $\beta_j > 0$.

$$\pi(\beta_j) \propto \frac{exp\left\{-\frac{1}{2}\left(\frac{\beta_j - \widehat{\mu}}{\widehat{\sigma}}\right)^2\right\}}{\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\left(1 - \Phi\left(-\frac{\widehat{\mu}}{\widehat{\sigma}}\right)\right)} I_{(0,\infty)}(\beta_j), \qquad (3.2)$$

- *Priori* normal truncada para $\beta_j < 0$.

$$\pi(\beta_j) \propto \frac{exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\widehat{\mu}}{\widehat{\sigma}}\right)^2\right\}}{\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\left(1-\Phi\left(-\frac{\widehat{\mu}}{\widehat{\sigma}}\right)\right)} I_{(-\infty,0)}(\beta_j), \tag{3.3}$$

em que,

$\widehat{\mu}$ é a média estimada dos $\beta_s^{'}$ pelo método *bootstrap*;

$\widehat{\sigma}^2$ é a variância estimada dos $\beta_s^{'}$ pelo método *bootstrap*;

$\Phi(.)$ é a função de distribuição acumulada da normal padrão .

$\widehat{\mu}$ e $\widehat{\sigma}^2$ foram utilizados diretamente como hiperparâmetros devido ao desconhecimento total deles. Assim, acreditou-se que seria melhor usar esses hiperparâmetros no lugar de buscar uma aproximação aleatória para eles.

3.2.3 Distribuições conjuntas a *posteriori*

De acordo com a teoria bayesiana, para cada distribuição a *priori* especificada deve-se obter uma distribuição a posteriori, considerando, no presente estudo, a mesma função de verossimilhanças. Partindo do pressuposto de que $\pi(\theta|Y) \propto L(\theta|Y)\pi(\theta)$, tem-se:

- Posteriori conjunta para o caso em que $\beta_j > 0$.

$$\pi(\beta_j, \sigma^2|Y_j) \propto \frac{exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\widehat{\mu}}{\widehat{\sigma}}\right)^2\right\}}{\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\left(1-\Phi\left(-\frac{\widehat{\mu}}{\widehat{\sigma}}\right)\right)} I_{(0,\infty)}(\beta_j)(\sigma^2 2\pi)^{-\frac{n}{2}}$$

$$\times exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{ij}-\beta_j)^2\right\}. \tag{3.4}$$

- Posteriori conjunta para o caso em que $\beta_j < 0$.

$$\pi(\beta_j, \sigma^2|Y_j) \propto \frac{exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\widehat{\mu}}{\widehat{\sigma}}\right)^2\right\}}{\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\left(1-\Phi\left(-\frac{\widehat{\mu}}{\widehat{\sigma}}\right)\right)} I_{(-\infty,0)}(\beta_j)(\sigma^2 2\pi)^{-\frac{n}{2}}$$

$$\times exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{ij}-\beta_j)^2\right\}. \tag{3.5}$$

3.2.4 Distribuições condicionais completas a *posteriori*

Para se obter as distribuições marginais a *posteriori*, é necessário realizar integrações na distribuição conjunta a *posteriori*. A inferência será baseada em amostras obtidas por meio das distribuições condicionais completas a *posteriori*, usando algoritmos MCMC. Portanto, é necessário apresentar distribuições condicionais completas a *posteriori*, necessárias à implementação do algoritmo Metropolis-Hastings, verificando-se que estas possuem forma desconhecida. Tais distribuições condicionais completas a posteriori são apresentadas a seguir:

- Posteriori condicional completa para $\beta_j > 0$.

$$\pi(\beta_j|\sigma^2, Y_j) \propto exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\widehat{\mu}}{\widehat{\sigma}}\right)^2\right\} I_{(0,\infty)}(\beta_j)$$

$$\times \left[\sum_{i=1}^{n}(y_{ij}-\beta_j)^2\right]^{\frac{2-n}{2}}. \tag{3.6}$$

- Posteriori condicional completa para $\beta_j < 0$.

$$\pi(\beta_j|\sigma^2, Y_j) \propto exp\left\{-\frac{1}{2}\left(\frac{\beta_j - \widehat{\mu}}{\widehat{\sigma}}\right)^2\right\} I_{(-\infty,0)}(\beta_j)$$

$$\times \left[\sum_{i=1}^{n}(y_{ij} - \beta_j)^2\right]^{\frac{2-n}{2}}. \tag{3.7}$$

- Posteriori condicional completa para σ^2.

$$\pi(\sigma^2|\beta_j, Y_j) \propto \sqrt{\sigma^2}. \tag{3.8}$$

3.2.5 Algoritmo MCMC

A escolha de qual *priori* utilizar, se a distribuição normal truncada à esquerda ou à direita, será baseada na estimativa pontual *bootstrap*, no caso, a média. Para a distribuição condicional do parâmetro β_j, que para as prioris especificadas não apresenta forma definida, deve-se utilizar o algoritmo Metropolis-Hastings, que está implementado no software estatístico R (R Development Core Team, 2007).

O algoritmo Metropolis-Hastings foi implementado, na análise dos dados, considerando uma cadeia de 50.000 iterações, das quais 10% dos valores iniciais foram eliminados (*burn-in*), para evitar o efeito dos valores iniciais adotados. Utilizou-se um *thin* de tamanho 10 para se obter a subamostra aproximadamente independente. Para a obtenção de valores amostrados dos parâmetro β_j e σ^2 utilizaram-se funções condidatas a distribuição normal truncada (com média $(\widehat{\mu})$ e variância $(\widehat{\sigma}^2)$ estimadas pelo procedimento *bootstrap*) para o parâmetro β_j e a Uniforme (0, 1) para σ^2.

A verificação final da convergência foi feita segundo o procedimento de Nogueira (2004), mediante o pacote BOA (*Bayesian Output Analysis*) do software R (R Development Core Team, 2007). Verificada a convergência, construiu-se o intervalo de credibilidade (HPD), para os $\beta_s^{'}$, para cada covariável selecionada no modelo. Após isso, o intervalo de credibilidade para a razão de riscos foi obtido exponenciando-se os limites (limite inferior e superior para os $\beta_s^{'}$), considerando a relação:

$$RR_j = exp(\beta_j),$$

em que,

RR_j é a razão de riscos associada à covariável "j", sendo j: 1, ..., p-covariáveis;

β_j é o coeficiene associado à covariável "j".

4 RESULTADOS E DISCUSSÃO

4.1 Análise dos dados de dependência química.

Utilizou-se o modelo de Cox para modelar o tempo de permanência dos pacientes na comunidade até a desistência do tratamento, em função das covariáveis apresentadas na Tabela 3.1, de acordo com a proposta de Collet (1994). Para decidir se um termo deve ou não ser incluído no modelo, usou-se um nível de significância ($\alpha = 0,10$). Considerando, então, este modelo, os passos da implementação da estratégia de seleção das covariáveis podem ser vistos na Tabela 4.1.

TABELA 4.1: Seleção de covariáveis usando o modelo de regressão de Cox ($\alpha = 0,10$).

Passos	Modelo	$-2logL(\theta)$	Estatística de teste (TRV)	Valor p
Passo 1	Nulo	$885,301$	-	-
	P	$885,299$	$0,002$	$0,964$
	EC	$884,033$	$1,268$	$0,260$
	F	$885,201$	$0,100$	$0,752$
	EP	$885,183$	$0,118$	$0,731$
	FR	$885,299$	$0,002$	$0,964$
	ES	$878,284$	$7,017$	$0,008$
	I	$880,640$	$4,661$	$0,031$
	CF	$885,050$	$0,251$	$0,617$
	TA	$884,494$	$0,807$	$0,369$
	PJ	$884,923$	$0,378$	$0,539$
	TU	$882,081$	$3,220$	$0,073$
Passo 2	ES+I+TU	$866,580$	-	-
	ES+I	$872,020$	$5,440$	$0,020$
	ES+TU	$875,240$	$8,660$	$0,003$
	I+TU	$875,240$	$8,668$	$0,003$

continua...

Passos	Modelo	$-2logL(\theta)$	Estatística de teste (TRV)	Valor p
Passo 3	ES+I+TU	$866,580$	-	-
	ES+I+TU+P	$866,292$	$0,288$	$0,591$
	ES+I+TU+EC	$865,870$	$0,702$	$0,402$
	ES+I+TU+F	$866,468$	$0,112$	$0,738$
	ES+I+TU+EP	$866,500$	$0,080$	$0,777$
	ES+I+TU+FR	$866,087$	$0,493$	$0,483$
	ES+I+TU+CF	$866,391$	$0,189$	$0,664$
	ES+I+TU+TA	$865,858$	$0,722$	$0,395$
	ES+I+TU+PJ	$866,554$	$0,262$	$0,609$
Modelo Final	ES+I+TU+EC	$865,870$		

Após o processo de seleção, o modelo de Cox resultante incluiu as seguintes covariáveis: nível de escolaridade (ES), idade (I), tempo de uso de drogas (TU) e estado civil (EC). Apesar da covariável (EC) não apresentar significância estatística, sua inclusão no modelo foi baseada em evidências clínicas.

Como discutido na Seção 2.5, a suposição de riscos proporcionais deve ser atendida para que o modelo de Cox possa ser considerado adequado aos dados deste estudo. Na Figura 4.1 é apresentado um método gráfico envolvendo o logaritmo da função risco, para essa finalidade.

Na Figura 4.1, encontram-se os gráficos envolvendo o logaritmo da função risco acumulado de base para as covariáveis ES, I, TU e EC. Pode-se observar, nesta Figura, que as curvas não indicam violação da suposição de riscos proporcionais. Embora as mesmas não sejam perfeitamente paralelas ao longo do eixo do tempo, não existem, em termos descritivos, afastamentos marcantes desta característica. A situação extrema de violação é caracterizada por curvas que se cruzam.

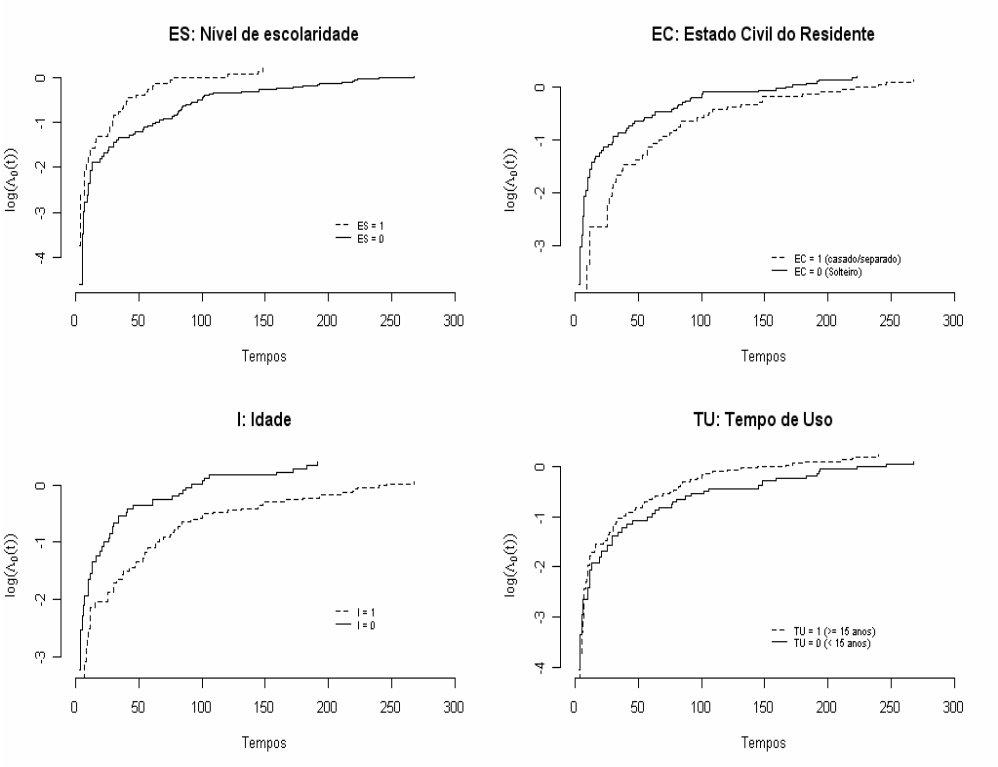

FIGURA 4.1: $Log(\widehat{\Lambda}_{0j}(t))$ *versus* tempo para as covariáveis ES, EC, I e TU.

Após se verificar a proporcionalidade dos riscos, ajusta-se o modelo de Cox, dado pela expressão (2.6) para os dados em estudo. Os resultados obtidos do modelo de Cox com as covariáveis selecionadas (ES, I, TU e EC) encontram-se na Tabela 4.2.

Pelos dados da Tabela 4.2, observa-se que, para as covariáveis TU e EC, os intervalos de confiança obtidos englobam a unidade. Temos assim, num mesmo intervalo, valores que indicam sobre-risco (acima de 1) e valores que indicam proteção (abaixo de 1), desse modo causando uma dupla interpretação dos resultados obtidos. Para contornar esse problema, implementou-se a metodologia exposta na

TABELA 4.2: Estimativas do ajuste do modelo de Cox para os dados de dependência química e correspondentes razões de risco (RR_j).

Covariável	Estimativa	Erro padrão	RR_j	$IC_{95\%}(RR_j)$
ES	$0,581$	$0,213$	$1,788$	$(1,177; 2,720)$
I	$-0,533$	$0,218$	$0,587$	$(0,383; 0,900)$
TU	$0,281$	$0,210$	$1,324$	$(0,877; 2,000)$
EC	$-0,183$	$0,216$	$0,833$	$(0,545; 1,270)$

Seção 3.2.

Em primeiro lugar, aplicou-se o método *bootstrap*, descrito na Seção 2.6, para se obter novas estimativas do parâmetro β_j relacionadas a cada covariável selecionado pelo modelo de Cox. Foram utilizadas $B = 10.000$ reamostragens *bootstrap* para cada covariável selecionada. Dessa forma, cada covariável será composta por uma amostra de 10.000 $\widehat{\beta}_s^{'}$ estimados pelo método *bootstrap*. Na Tabela 4.3 encontram-se a média $(\widehat{\mu})$ e a variância $(\widehat{\sigma}^2)$ das amostragens *bootstrap* relacionadas a cada covariável mensurada no estudo.

TABELA 4.3: Estimativas da média $(\widehat{\mu})$ e variância $(\widehat{\sigma}^2)$ das amostragens *bootstrap* relacionadas a cada covariável mensurada no estudo.

Covariável	$\widehat{\mu}$	$\widehat{\sigma}^2$
ES	$0,5844$	$0,0621$
I	$-0,5347$	$0,0634$
TU	$0,3282$	$0,0477$
EC	$-0,2209$	$0,0535$

Considerou-se o valor de $\widehat{\mu}$ como estimativa pontual dos respectivos coeficientes β_j, correspondentes a cada covariável.

Em segundo lugar, aplicou-se a metodologia bayesiana para a construção de intervalos de credibilidade para o vetor de parâmetros β_j. Como descrito na Seção 3.2.5, a *priori* escolhida foi a normal truncada à direita, dada pela expressão (3.2), para as covariáveis ES e TU, devido ao fato de a estimativa pontual *bootstrap*,

no caso, a média, ser maior do que zero. Para as covariáveis I e EC, a *priori* escolhida foi a normal truncada à esquerda, dada pela expressão (3.3), pelo fato de a estimativa pontual *bootstrap*, no caso a média, ser menor do que zero. Conseqüentemente, as inferências foram baseadas em amostras obtidas por meio das distribuição condicional completa a *posteriori*, dada pelas expressões (3.6) e (3.7), usando algoritmo MCMC.

Para se avaliar a convergência da seqüencia que foi gerada pelo algoritmo MCMC, foi utilizado, na definição do número de iterações necessário à convergência, assim como na determinação dos valores de *burn-in* e *thin*, o procedimento recomendado por Nogueira (2004), já descrito na Seção 2.8.2. De acordo com este procedimento, tem-se que o teste de Raftery & Lewis sugeriu um processo com 50.000 iterações. Para assegurar a independência da amostra, considerou-se um espaçamento, entre os pontos amostrados, de tamanho 10 (*thin*). O teste de Heidelberger & Welch recomendou que fossem descartadas as 5.000 interações iniciais (*burn-in*), ou seja, obteve-se uma amostra final, para cada parâmetro, de tamanho 4.500.

Para a avaliação da convergência das cadeias geradas dos parâmetros, também foi utilizada a inspeção da visualização gráfica do traço. Na Tabela 4.4 são apresentados os resultados dos testes de diagnóstico para os procedimentos de Geweke (1992) e de Gelman & Rubin (1992), para cada covariável selecionada pelo modelo de Cox.

TABELA 4.4: Testes de diagnóstico para o parâmetro β_j, referente às covariáveis selecionadas.

Covariável	Geweke	Gelman & Rubin
ES	0,182	1,00
I	0,281	1,01
TU	-0,097	0,99
EC	0,576	1,02

Os testes de diagnóstico mostraram que, em todos os casos, a cadeia convergiu e apresenta características de uma cadeia estacionária. Sendo assim, pelo critério de Geweke (1992), não houve nenhum valor que não estivesse entre os limites da normal padronizada. E, no caso do Gelman & Rubin (1992), este sempre apresentou valores de $\sqrt{\widehat{R}}$ iguais a 1 ou próximos deste valor.

As cadeias geradas pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente a cada covariável selecionada pelo modelo, encontram-se nas Figuras 4.2 a 4.5.

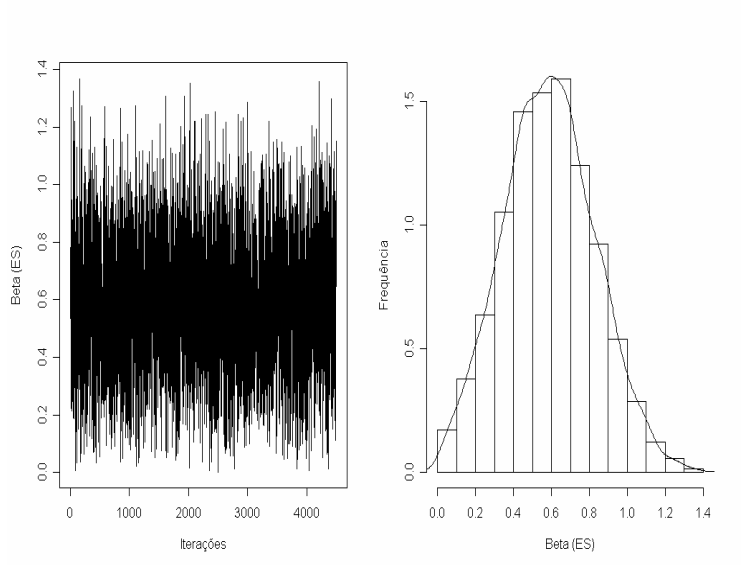

FIGURA 4.2: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável ES, selecionada pelo modelo.

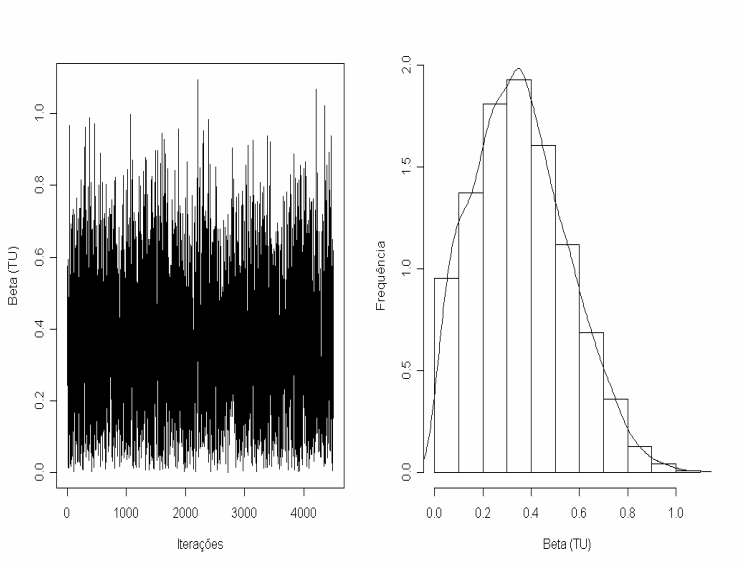

FIGURA 4.3: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável TU, selecionada pelo modelo.

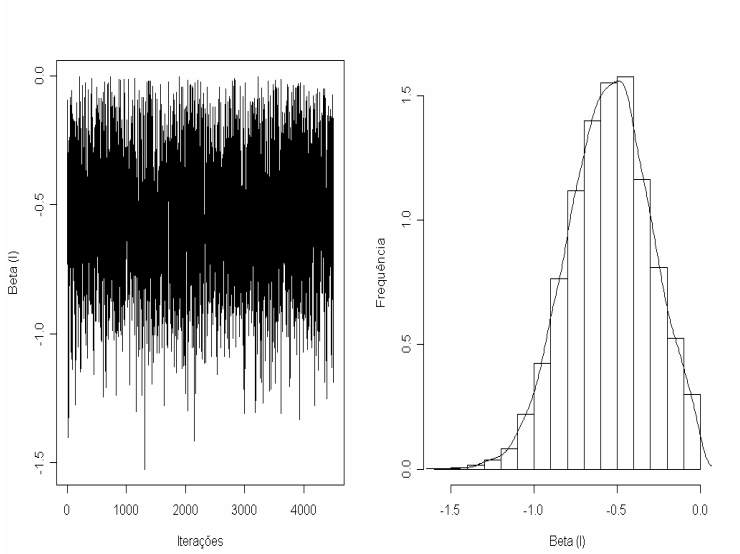

FIGURA 4.4: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável I, selecionada pelo modelo.

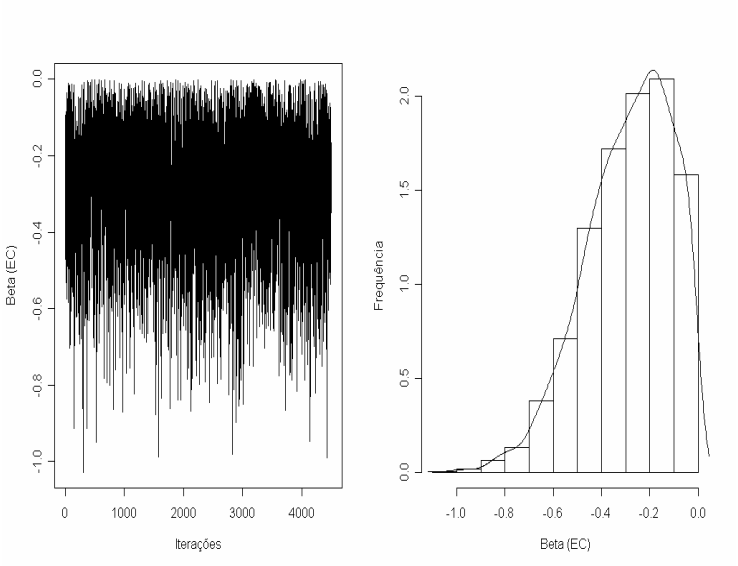

FIGURA 4.5: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável EC, selecionada pelo modelo.

Nas Figuras de 4.2 a 4.5, observa-se claramente uma rápida convergência dos parâmetros β_j estimados para as covariáveis selecionadas pelo modelo de Cox, em torno do valor real (média *bootstrap*). Nota-se que a distribuição a *posteriori* é assimétrica, para todas as covariáveis, o que era esperado, devido ao fato de ter sido utilizada como *priori* a distribuição normal truncada.

Na Tabela 4.5 são mostradas as estimativas pontuais (médias a *posteriori*), o desvio padrão e por intervalo (intervalos de credibilidade (95%)), para o parâmetro β_j, relacionadas a cada covariável em estudo.

TABELA 4.5: Estimativas da média $(\widehat{\beta})$, desvio padrão (DP) e intervalos de credibilidade 95% (HPD), para o parâmetro β_j, referente a cada covariável.

Covariável	$\widehat{\beta}$	DP	2,5%	97,5%
ES	$0,5886$	$0,2402$	$0,1433$	$1,0790$
I	$-0,5469$	$0,2457$	$-0,9736$	$-0,0311$
TU	$0,3561$	$0,1916$	$0,0044$	$0,6970$
EC	$-0,2929$	$0,1836$	$-0,6239$	$-0,0003$

Os limites dos intervalos de credibilidade dos β_j, apresentados na Tabela 4.5, foram utilizados para se obter o correspondente intervalo de credibilidade para as razões de risco, (RR_j). Para isso, bastou utilizar a relação existente entre as RR_j e os coeficientes β_j (Seção 2.5.5), ou seja, $RR_j = exp(\beta_j)$. Esta relação foi aplicada em cada um dos limites do intervalo obtido para os $\beta_s^{'}$, referente a cada covariável. Os intervalos de credibilidade para as RR_j podem ser observados na Tabela 4.6.

TABELA 4.6: Intervalos de credibilidade 95% para as razões de risco (RR_j).

Covariável	$RR_j = exp(\widehat{\mu})$	$2,5\%$	$97,5\%$
ES	$1,79$	$1,1541$	$2,7704$
I	$0,58$	$0,3777$	$0,9694$
TU	$1,39$	$1,0044$	$2,0077$
EC	$0,80$	$0,5358$	$0,9997$

Pode-se observar que o problema encontrado nos intervalos de confiança apresentados na Tabela 4.2, em relação aos limites abaixo e acima de 1 num mesmo intervalo, foi resolvido quando aplicou-se o método exposto neste trabalho. Isto oferece uma maior credibilidade nos resultados. Sendo assim, pode-se interpretar os resultados da Tabela 4.6 da seguinte forma:

1. o risco de abandono da comunidade de pacientes com baixa escolaridade é 1,79 vez o risco de abandono de pacientes com maior tempo de estudo.

Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 1,1541 e 2,7704;

2. o risco de abandono da comunidade de pacientes com idade inferior a 30 anos é 0,58 vezes o risco de abandono de pacientes com idade igual ou superior a 30 anos. Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 0,3777 e 0,9694;

3. o risco de abandono da comunidade de pacientes com tempo de uso de drogas maior ou igual a 15 anos é 1,39 vez o risco de abandono de pacientes com tempo de uso inferior a 15 anos. Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 1,0044 e 2,0077;

4. o risco de abandono da comunidade de pacientes solteiros é 0,8 vezes o risco de abandono de pacientes casados. Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 0,5358 e 0,9997.

4.2 Análise dos dados de aleitamento materno

Colosimo & Giolo (2006) utilizaram o modelo de Cox para modelar o tempo máximo de aleitamento materno, em função das covariáveis apresentadas na Tabela 3.2, de acordo com a proposta de Collet (1994). Os autores consideraram um nível de significância de 0,10.

Após o processo de seleção, o modelo de Cox resultante incluiu as seguintes covariáveis: experiência anterior de amamentação (V1), conceito materno sobre tempo ideal de amamentação (V3), dificuldades de amamentação nos primeiros dias pós-parto (V4) e recebimento exclusivo de leite materno na maternidade (V6).

Após se verificar a proporcionalidade dos riscos, Colosimo & Giolo (2006) obtiveram os resultados apresentados na Tabela 4.7.

TABELA 4.7: Estimativas do ajuste do modelo de Cox para os dados de aleitamento materno e correspondentes razões de risco (RR_j).

Covariável	Estimativa	Erro padrão	RR_j	$IC_{95\%}(RR_j)$
V1	$0,471$	$0,268$	$1,60$	$(0,94; 2,71)$
V3	$0,579$	$0,262$	$1,78$	$(1,07; 2,99)$
V4	$0,716$	$0,264$	$2,05$	$(1,22; 3,43)$
V6	$0,578$	$0,264$	$1,78$	$(1,06; 2,99)$

Pelos dados da Tabela 4.7, pode-se observar que para a covariável V1, o intervalo de confiança obtido engloba a unidade. Têm-se assim, num mesmo intervalo, valores que indicam sobre-risco (acima de 1) e valores que indicam proteção (abaixo de 1), desse modo causando uma dupla interpretação dos resultados obtidos.

Para contornar este problema, implementou-se a metodologia exposta na Seção 3.2.

Em primeiro lugar, aplicou-se o método *bootstrap*, descrito na Seção 2.6, para se obter novas estimativas do parâmetro β_j relacionada a cada covariável selecionado pelo modelo de Cox. Foram utilizadas $B = 10.000$ reamostragens *bootstrap* para cada covariável selecionada. Na Tabela 4.8 encontram-se a média $(\widehat{\mu})$ e a variância $(\widehat{\sigma}^2)$ das amostras *bootstrap* relacionadas a cada covariável mensurada no estudo.

Considerou-se o valor de $\widehat{\mu}$ como estimativa pontual dos respectivos coeficientes β_j, correspondentes a cada covariável.

Em segundo lugar, aplicou-se a metodologia bayesiana para a construção de intervalos de credibilidade para o vetor de parâmetros β_j. Como descrito na Seção 3.2.5, a *priori* escolhida foi a normal truncada à direita, dada pela expressão (3.2),

TABELA 4.8: Estimativas da média ($\widehat{\mu}$) e variância ($\widehat{\sigma}^2$) das amostras *bootstrap* relacionadas a cada covariável mensurada no estudo.

Covariável	$\widehat{\mu}$	$\widehat{\sigma}^2$
V1	$0,4654$	$0,0792$
V3	$0,6185$	$0,0737$
V4	$0,7094$	$0,0844$
V6	$0,6160$	$0,0805$

devido ao fato de a estimativa pontual *bootstrap*, no caso a média, ser maior do que zero para todas as covariáveis analisadas. Conseqüentemente, as inferências foram baseadas em amostras obtidas por meio da distribuição condicional completa a *posteriori*, dada pela expressão (3.6), usando o algoritmo MCMC.

Para se avaliar a convergência da seqüência que foi gerada pelo algoritmo MCMC, foi utilizado, para se definir o número de iterações necessárias à convergência, assim como na determinação dos valores de *burn-in* e *thin*, o procedimento recomendado por Nogueira (2004), já descrito na Seção 2.8.2. De acordo com esse procedimento, tem-se que o teste de Raftery & Lewis sugeriu um processo com 50.000 iterações, sendo que, para assegurar a independência da amostra, considerou-se um espaçamento, entre os pontos amostrados, de tamanho 10 (*thin*). O teste de Heidelberger & Welch recomendou que fossem descartadas as 5.000 interações iniciais (*burn-in*), ou seja, obteve-se uma amostra final, para o parâmetro β_j, de tamanho 4.500.

Para a avaliação da convergência das cadeias geradas dos parâmetros, também foi utilizada a inspeção da visualização gráfica do traço. Na Tabela 4.9 são apresentados os resultados dos testes de diagnóstico para os procedimentos de Geweke (1992) e de Gelman & Rubin (1992), para cada covariável selecionada pelo modelo de Cox.

Os testes de diagnóstico mostraram que, em todos os casos, a cadeia convergiu

TABELA 4.9: Testes de diagnóstico para o parâmetro β_j referente às covariáveis selecionadas.

Covariável	Geweke	Gelman & Rubin
V1	0,011	1,01
V3	-1,380	1,00
V4	-0,717	1,00
V6	0,002	0,98

e apresenta características de uma cadeia estacionária. Sendo assim, pelo critério de Geweke (1992), não houve nenhum valor que não estivesse entre os limites da normal padronizada. E no caso do Gelman & Rubin (1992) este sempre apresentou valores de $\sqrt{\widehat{R}}$ iguais a 1 ou próximos deste valor.

As cadeias geradas pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente a cada covariável selecionada pelo modelo, encontram-se nas Figuras 4.6 a 4.9.

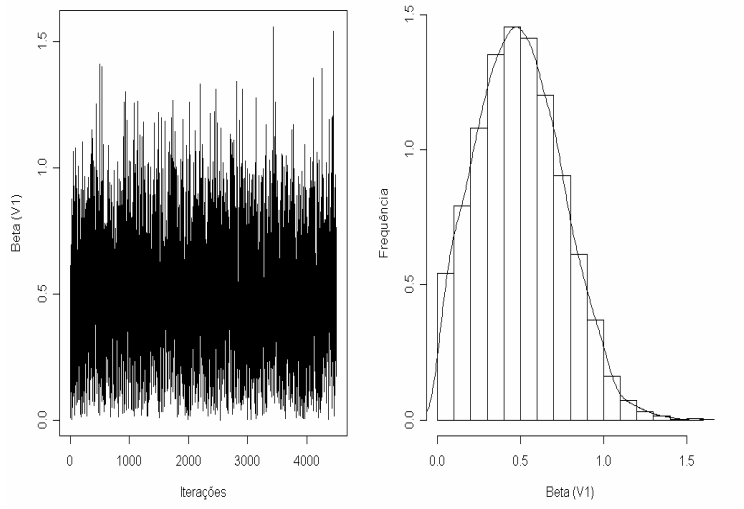

FIGURA 4.6: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável V1, selecionada pelo modelo.

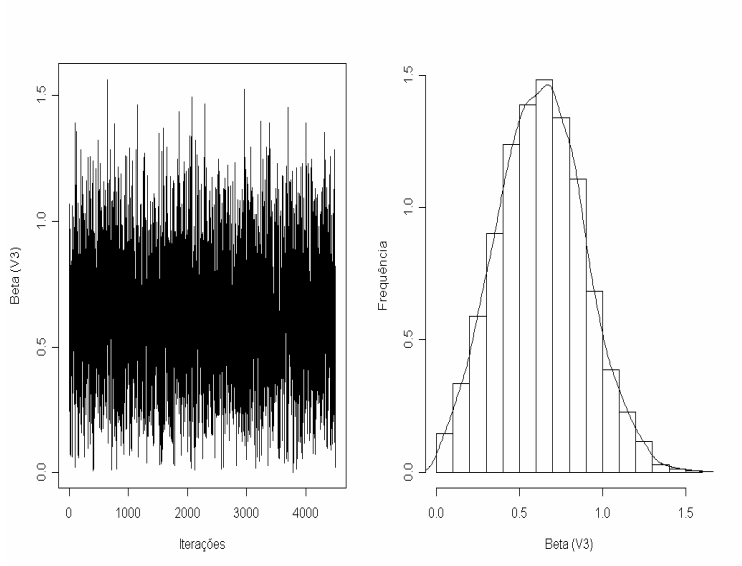

FIGURA 4.7: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável V3, selecionada pelo modelo.

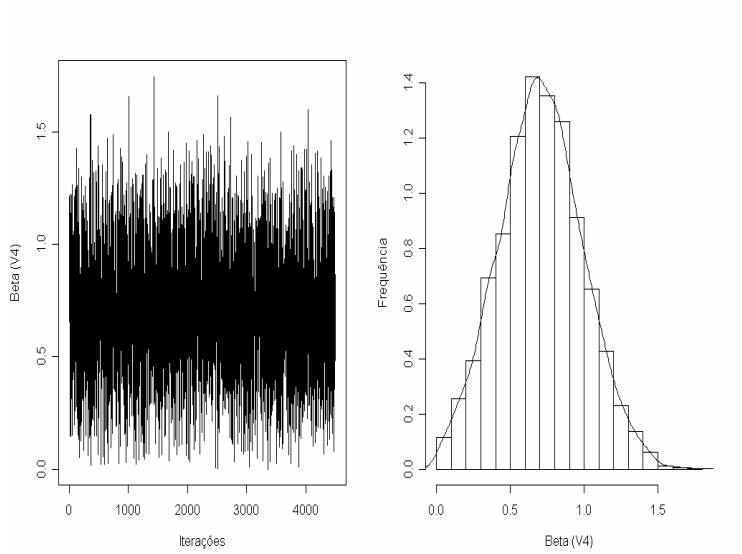

FIGURA 4.8: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável V4, selecionada pelo modelo.

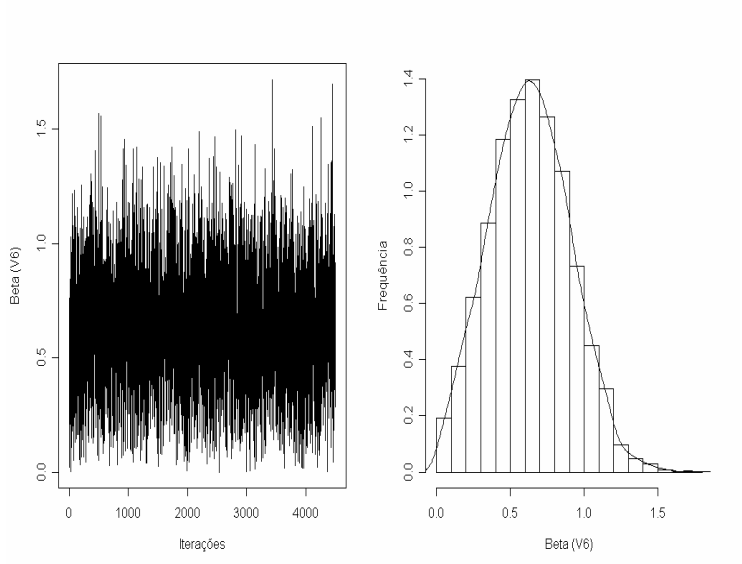

FIGURA 4.9: Cadeia gerada pelo método MCMC e densidade a posteriori para o parâmetro β_j, referente à covariável V6, selecionada pelo modelo.

Nas Figuras 4.6 a 4.9, observa-se claramente uma rápida convergência dos parâmetros β_j estimados para as covariáveis selecionadas pelo modelo de Cox, em torno do valor real (média *bootstrap*). Nota-se que a distribuição a *posteriori* é assimétrica, para a covariável V1 e simétrica para as demais covariáveis em estudo.

Na Tabela 4.10 são mostradas as estimativas pontuais (médias a *posteriori*), o desvio padrão e por intervalo (intervalos de credibilidade (95%)), para o parâmetro β_j, relacionada a cada covariável em estudo.

TABELA 4.10: Estimativas da média $(\widehat{\beta})$, desvio padrão (DP) e intervalos de credibilidade 95% (HPD), para o parâmetro β_j, referente a cada covariável.

Covariável	$\widehat{\beta}$	DP	$2,5\%$	$97,5\%$
V1	$0,4961$	$0,2558$	$0,0267$	$0,9593$
V3	$0,6287$	$0,2593$	$0,1228$	$1,1325$
V4	$0,7081$	$0,2824$	$0,1450$	$1,2556$
V6	$0,6321$	$0,2736$	$0,0901$	$1,1325$

Os limites dos intervalos de credibilidade dos β_j, apresentados na Tabela 4.10, foram utilizados para se obter o correspondente intervalo de credibilidade para as razões de risco, (RR_j). Para isso, bastou utilizar a relação existente entre as RR_j e os coeficientes β_j (Seção 2.5.5), ou seja, $RR_j = exp(\beta_j)$. Esta relação foi aplicada em cada um dos limites do intervalo obtido para os $\beta_s^{'}$ referente a cada covariável. Os intervalos de credibilidade para as RR_j podem ser observados na Tabela 4.11.

TABELA 4.11: Intervalos de credibilidade 95% para as razões de risco (RR_j).

Covariável	$RR_j = exp(\widehat{\mu})$	$2,5\%$	$97,5\%$
V1	$1,59$	$1,0270$	$2,6099$
V3	$1,86$	$1,1307$	$3,1034$
V4	$2,03$	$1,1560$	$3,5099$
V6	$1,85$	$1,0943$	$3,1005$

Pode-se observar que o problema encontrado nos intervalos de confiança apresentados em Colosimo & Giolo (2006), em relação aos limites abaixo e acima de 1 num mesmo intervalo, foi resolvido quando aplicou-se o método exposto neste trabalho. Isto oferece uma maior credibilidade nos resultados. Sendo assim, pode-se interpretar os resultados da Tabela 4.11 da seguinte forma:

1. o risco de desmame precoce em mães que não tiveram experiência anterior de amamentação é 1,59 vez o risco das mães que tiveram essa experiência. Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 1,0270 e 2,6099;

2. o risco de desmame precoce em mães que acreditam que o tempo ideal de amamentação é $\leq$ a 6 meses é 1,86 vez o risco das mães que acreditam que o tempo ideal de amamentação é superior a 6 meses. Além disso, pode-se

dizer, com 95% de probabilidade, que esse risco varia entre 1,1307 e 3,1034;

3. o risco de desmame precoce em mães que apresentaram dificuldades de amamentar nos primeiros dias pós-parto é 2,03 vezes o risco das mães que não apresentaram essas dificuldades. Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 1,156 e 3,5099;

4. o risco de desmame precoce em crianças que não receberam exclusivamente leite materno na maternidade é 1,85 vezes o risco de desmame precoce em crianças que receberam exclusivamente o leite materno. Além disso, pode-se dizer, com 95% de probabilidade, que esse risco varia entre 1,0943 e 3,1005.

5 CONCLUSÃO

Diante do problema relacionado com a estimação dos intervalos de confiança para a razão de riscos, em que, num mesmo intervalo, observaram-se valores acima de 1, que indicam sobre-risco e valores abaixo de 1, que indicam proteção, observou-se também que a aplicação da metodologia proposta neste trabalho proporcionou resultados satisfatórios.

Para a sua realização, foram utilizados o método *bootstrap*, a inferência bayesiana e o algoritmo MCMC para a estimação dos intervalos de credibilidade. Por meio do método *bootstrap*, obtiveram-se novas estimativas do parâmetro dos coeficientes do modelo de regressão de Cox, β_j, $j = 1, ..., p$ relacionadas a cada covariável selecionada pelo modelo de Cox. A inferência bayesiana foi empregada considerando-se as estimativas pontuais *bootstrap*, de modo que a escolha da *priori* adequada se deu em função dessas estimavas. A função de verossimilhança adotada seguiu o modelo normal. A *priori* utilizada foi a normal truncada. O motivo de se utilizar tal *priori* deve-se à necessidade de que os β_s assumam valores exclusivamente positivos ou estritamente negativos, dependendo da estimativa pontual *bootstrap* obtida. Desse modo, para resolver o problema da ambigüidade dos intervalos de confiança, construíram-se os intervalos de credibilidade, baseando-se no algoritmo Metropolis-Hastings.

Ao aplicar este método, foram encontrados resultados aceitáveis, como se pode ver nas aplicações relacionadas com os exemplos descritos na Seção 4.1 e 4.2, em que os intervalos de confiança para a covariável TU ([0,877; 2,000]), EC ([0,545; 1,270]) e V1 ([1,94; 2,71]) agora só fornecem limites acima de um quando a estimativa *bootstrap* é maior que 1 e abaixo de um quando a estimativa *bootstrap* é

menor que 1.

Ao comparar os comprimentos dos intervalos de confiança e de credibilidade, percebe-se que os mesmos não têm qualquer diferença expressiva quanto à amplitude.

6 ESTUDOS FUTUROS

1. Implementar o procedimento no software R, por meio de um pacote. De modo que, o pesquisador possa introduzir os dados no software, este construiria o intervalo de credibilidade para as Razões de Risco, relacionada a covariável de interesse.

2. Aplicar o mesmo procedimento usando outras *prioris*, como por exemplo, a distribuição Gama (α, β) e a distribuição Beta (a, b), para uma análise comparativa dos resultados. Outras distribuições utilizadas como *priori* poderiam também ser comparadas.

3. Avaliar a performance do procedimento proposto via simulação.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSEN, P. K.; GILL, R. Cox's regression model for counting processes: a large sample study. **Annals of Statistics**, Amsterdam, v. 10, p. 1100-1200, Dec 1982.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001. 125 p.

BOX, G. E. P.; TIAO, G. C. **Bayesian inference in statistical analysis**. New York: J. Wiley, 1992. 603 p.

BRESLOW, N. Contribuição à discussão do artigo de D. R. Cox. **Journal of the Royal Statistical Society B**, London: Imperial College, v. 34, p. 216-217, May 1972.

CARLIN, B. P.; LOUIS, T. A. **Bayes and empirical Bayes methods for data analysis**. London: Chapman and Hall, 1996. 399 p.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de sobrevida teoria e aplicações em saúde**. Rio de Janeiro: Ed. Fiocruz, 2005. 396 p.

COLLETT, D. **Modelling survival data in medical research**. New York: Chapman and Hall, 1994. 346 p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência**. São Paulo: Edgard Blücher, 2006. 369 p.

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. London, Chapman & Hall,

1974. 511 p.

COX, D. R. Partial likelihood. **Biometrika**, London: Imperial College, v. 62, n. 8, p. 269-276, 1975.

COX, D. R. Regression models and life tables (with discussion). **Journal Royal Statistical Society B**, London: Imperial College, v. 34, n. 8, p. 187-220, March 1972.

EFRON, B. *Bootstrap* methods: another look at jakknife. **Annals of Statistics**, Amsterdam, v. 7, n.1, p. 1-26, Jan 1979.

EFRON, B.; TIBSHIRANI, R. **An introduction to the bootstrap**. New York: Chapman & Hall, 1993. 436 p.

GAMERMAN, D. **Markov chain Monte Carlo - stochastic simulation for bayesian inference**. London: Chapman & Hall, 1997. 245 p.

GELMAN, A.; CARLIN, J. B.; STER, H. S.; RUBIN, D. B. **Bayesian data analysis**. Boca Raton: Chapman & Hall/CRC, 2000. 526 p.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v. 7, n. 4, p. 457-511, May 1992.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calcualting posterior moments. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian statistics.** New York: Oxford University, 1992. p. 625-631.

HARRIS, E. K.; ALBERT, A. **Survivorship analysis for clinical studies**. New

York: Marcel Dekker Incorporation, 1991. 197 p.

HASTINGS, W. K. Monte carlo sampling methods using Markov Chains and theirs applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, Apr 1970.

HEIDELBERG, P.; WELCH, P. Simulation run lenght control in the presence of an initial transient. **Operations Research**, Landing, v. 31, n. 6, p. 1109-44, Nov./Dec. 1993.

JEFFREYS, H. **Theory of probability**. Oxford: Claredon, 1961. 325 p.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. New York: John Wiley and Sons, 1980. 385 p.

LAVORANTI, O. J. **Estabilidade e adaptabilidade fenotípica através da reamostragem *Bootstrap* no modelo AMMI**. 2003. 184 p. Tese (Doutorado em agronomia) - Escola Superior de Agricultura Luiz de “Queiroz", Piracicaba, SP

LAWLESS, J. F. **Statisticals models and methods for lifetime data**. New York: John Wiley and Sons, 1982. 580 p.

LOUSADA NETO, F. **Testes de sobrevivência acelerados: uma análise bayesiana do modelo de eyring**. 1991. 132 p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemática, Universidade de São Paulo, São Carlos.

MAGALHAES, M. N. **Probabilidade e variáveis aleatórias**. São Paulo: IME-USP, 2004. 414 p.

METROPLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A.

H.; TELLER, E. Equation of state calculations by fast computing machine. **Journal of Chemical Phisics**, Chicago, v. 21, n. 6, p. 1089-1091, June 1953.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3. ed. New York: J. Wiley & Sons, 1974. 564 p.

NOGUEIRA, D. A. **Proposta e avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov:** casos uni e multivariados. 2004. 142 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG.

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

PAZ, L. M. **Modelo de Cox para dados com censura intervalar**. 2005. 77 p. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Belo Horizonte, MG.

RAFTERY, A. L.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J. M. et al. (Ed.). **Bayesian statistics.** Oxford: University, 1992. p. 763-774.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 20 dez. 2007.

ANEXOS

ANEXO A: Dados utilizados no estudo sobre dependência química.

P	EC	F	EP	FR	ES	I	QD	CF	TA	PJ	TU	tempos	cens
0	0	0	0	1	1	1	0	1	0	0	1	7	1
0	1	1	0	0	1	1	0	1	1	1	0	25	1
0	1	1	0	0	1	1	0	1	1	0	1	48	1
0	0	1	0	0	1	1	0	1	0	0	1	55	1
1	0	0	1	0	0	0	1	1	0	0	0	76	1
1	1	1	0	1	0	0	1	1	1	0	0	102	1
0	0	1	0	0	1	0	1	1	0	1	1	40	1
0	0	0	0	1	0	1	0	1	1	0	1	54	1
1	0	1	1	0	0	0	1	1	0	1	0	45	1
1	0	1	0	0	1	0	1	1	1	1	0	183	1
0	0	0	0	0	0	0	0	1	0	0	0	13	1
0	1	1	0	0	1	0	0	0	1	0	0	11	1
0	0	0	0	0	0	0	0	0	0	0	1	3	1
0	0	1	0	1	1	1	0	1	0	0	1	15	1
0	1	1	1	1	1	1	0	0	0	0	0	148	1
0	1	1	0	1	0	1	0	1	0	0	0	80	1
1	0	1	0	0	0	0	1	1	0	0	0	34	1
0	1	1	0	0	0	1	0	1	0	1	0	57	1
0	0	0	0	0	0	1	1	1	0	0	1	81	1
0	0	0	0	0	0	1	0	1	1	0	1	30	1
0	0	1	0	0	0	0	0	1	0	0	1	22	1
0	0	0	0	1	0	1	0	0	0	0	1	42	1
0	0	0	0	1	0	1	0	0	0	1	1	6	1

cont...

P	EC	F	EP	FR	ES	I	QD	CF	TA	PJ	TU	tempos	cens
0	0	1	0	1	1	1	0	1	0	0	1	7	1
0	1	1	0	0	1	1	0	0	1	0	1	144	1
0	1	0	1	0	0	1	0	0	0	0	1	109	1
0	0	1	0	0	0	0	0	0	0	0	1	26	1
0	0	1	0	0	0	0	0	1	1	0	0	87	1
0	0	1	0	1	1	0	0	1	1	0	1	34	1
1	1	1	0	1	0	0	1	1	1	1	0	106	1
0	0	0	1	0	0	0	0	0	0	1	1	172	1
0	1	1	1	0	0	1	0	1	1	1	1	32	1
0	0	1	0	0	0	0	0	1	1	1	0	5	1
0	0	0	0	0	1	0	1	1	1	0	0	3	1
1	1	1	0	0	1	0	1	1	1	1	1	27	1
0	1	1	0	0	0	1	1	1	1	0	1	11	1
0	1	1	0	0	1	1	0	0	0	1	0	246	1
1	0	0	0	0	0	1	1	1	1	0	1	10	1
0	0	0	0	0	0	1	0	1	0	1	0	223	1
0	1	1	1	0	1	1	0	1	0	0	1	57	1
0	0	0	0	0	0	1	0	0	0	0	0	193	1
0	0	1	1	0	0	1	0	1	0	0	1	100	1
0	0	1	1	0	1	0	0	0	0	1	1	10	1
0	0	1	0	1	1	0	0	0	0	1	0	4	1
0	0	0	0	1	1	0	0	0	0	0	1	4	1
1	1	1	0	1	0	1	1	0	1	1	1	25	1
0	0	0	1	0	0	0	0	0	0	1	1	13	1

cont...

P	EC	F	EP	FR	ES	I	QD	CF	TA	PJ	TU	tempos	cens
0	1	0	0	0	1	1	0	1	0	0	0	37	1
0	0	1	0	0	0	0	0	0	1	1	0	92	1
0	0	0	0	0	0	1	0	1	0	0	1	11	1
0	1	1	0	1	0	1	0	1	0	0	1	84	1
0	1	0	0	0	0	1	0	1	0	0	1	9	1
1	1	1	0	0	0	1	1	1	1	0	1	180	1
0	0	0	0	0	1	1	1	1	0	1	0	63	1
0	1	1	1	0	0	1	0	1	0	0	1	240	1
0	0	1	0	0	0	1	0	0	0	0	1	63	1
0	1	1	0	0	0	1	0	1	0	0	1	5	1
1	0	0	1	0	0	1	1	1	0	0	1	219	1
0	1	1	0	1	1	1	0	1	0	1	0	148	1
0	1	0	0	0	1	0	0	1	0	0	0	29	1
0	1	1	0	0	0	1	0	1	0	0	0	66	1
0	1	1	0	0	0	1	0	1	0	1	1	83	1
0	1	1	0	1	1	1	0	1	0	0	1	38	1
0	0	0	0	1	1	1	0	1	1	0	1	8	1
0	0	1	1	1	0	1	0	1	0	1	0	11	1
0	0	1	0	0	0	0	1	0	1	1	0	159	1
0	0	0	0	0	1	1	0	1	1	0	1	167	1
0	0	0	0	0	0	1	0	1	0	1	1	91	1
1	0	0	0	0	1	0	1	0	1	1	1	16	1
1	0	0	0	0	0	0	1	1	1	1	1	61	1
0	1	1	0	0	0	1	0	1	0	0	1	131	1

cont...

P	EC	F	EP	FR	ES	I	QD	CF	TA	PJ	TU	tempos	cens
0	0	0	0	0	0	1	0	0	0	0	1	101	1
0	0	0	0	0	0	0	0	1	0	1	1	7	1
1	0	0	1	0	0	1	1	1	0	0	1	47	1
0	0	0	0	1	0	1	1	1	0	1	0	145	1
0	1	1	0	1	0	1	0	0	0	0	0	268	1
1	0	0	0	0	0	0	1	1	0	1	0	21	1
0	1	1	0	0	0	1	0	0	0	0	1	53	1
0	0	1	0	1	1	0	1	1	0	1	1	30	1
1	0	0	0	0	0	1	1	0	0	0	1	78	1
0	0	0	0	0	0	0	1	1	0	1	1	85	1
0	1	1	0	0	0	1	1	1	0	0	1	221	1
1	1	1	0	0	1	1	1	0	0	0	1	75	1
0	1	1	0	0	1	1	0	1	0	0	1	120	1
0	0	1	0	0	0	0	0	0	0	0	0	19	1
1	0	1	1	0	0	0	1	1	0	1	1	82	1
0	0	1	0	0	1	0	1	0	0	0	0	41	1
0	1	1	0	0	1	1	0	1	1	0	0	194	1
0	0	0	0	0	0	0	0	1	0	1	1	100	1
0	1	1	0	0	1	0	0	0	0	1	0	61	1
1	0	0	0	0	0	0	1	0	0	1	0	191	1
0	0	0	1	0	0	0	0	0	0	0	0	10	1
0	1	0	0	1	0	1	0	0	0	0	1	30	1
0	1	1	0	0	0	1	0	1	0	0	1	211	1
0	1	1	0	0	0	1	0	1	0	0	1	69	1

cont...

P	EC	F	EP	FR	ES	I	QD	CF	TA	PJ	TU	tempos	cens
0	0	1	0	0	0	0	1	1	1	1	0	6	1
0	0	0	0	0	1	0	0	1	0	0	0	29	1
0	1	1	0	0	1	1	0	1	0	0	0	77	1
0	1	1	0	0	0	1	0	1	0	1	1	96	1
0	1	1	0	1	0	1	0	1	0	1	1	270	0
1	0	0	0	1	1	0	1	0	1	1	1	270	0
0	0	0	0	0	0	1	0	1	0	0	0	270	0
0	1	1	0	0	1	1	0	1	0	0	1	270	0
0	0	0	0	0	1	1	0	1	0	1	1	270	0
0	0	1	1	0	0	0	1	0	0	1	0	270	0
0	1	1	0	1	0	0	0	1	0	0	0	270	0
1	0	0	1	0	0	1	1	1	1	0	1	270	0
0	0	0	0	0	0	0	0	0	0	1	0	270	0
0	0	0	0	0	0	1	0	1	1	0	0	270	0
1	1	1	0	0	0	1	1	1	0	0	1	270	0
0	0	0	0	0	1	0	1	1	0	0	1	270	0
0	0	1	0	1	0	1	0	1	0	1	1	270	0
0	0	0	0	0	0	0	1	0	0	1	0	270	0
0	0	0	0	0	0	1	0	1	0	1	0	270	0
0	0	0	0	1	0	1	0	0	0	0	1	270	0
0	0	0	0	1	0	1	0	1	0	0	1	270	0
0	1	1	0	1	0	1	0	1	0	0	1	270	0
0	1	1	0	0	0	1	0	0	0	0	1	270	0
0	1	1	0	0	0	1	0	1	1	0	1	270	0

cont...

P	EC	F	EP	FR	ES	I	QD	CF	TA	PJ	TU	tempos	cens
0	0	0	0	1	0	1	0	1	0	0	0	270	0
0	1	1	0	0	0	1	0	0	0	0	1	270	0
0	0	1	0	0	0	1	0	1	0	1	1	270	0
0	0	1	0	0	0	0	1	0	0	0	0	270	0
0	1	1	0	0	0	1	0	1	0	0	0	270	0
0	1	1	1	0	0	1	0	1	1	0	0	270	0
0	0	0	0	0	0	1	0	1	1	0	0	270	0
0	1	1	0	0	1	1	0	1	0	0	1	270	0
0	1	1	0	1	1	1	0	1	0	0	1	270	0
0	1	1	0	0	0	1	0	0	0	0	1	270	0
1	0	0	0	0	0	0	1	0	0	1	0	270	0
0	0	0	0	0	0	1	0	0	0	0	0	270	0
1	0	1	1	0	0	1	1	1	0	1	0	270	0
0	1	1	0	1	0	0	0	1	0	0	1	270	0
1	1	1	1	0	0	0	1	1	0	0	1	270	0
0	1	1	0	1	0	1	0	0	0	0	0	270	0
1	0	0	1	0	0	1	1	0	1	0	1	270	0
0	0	1	0	0	0	1	0	1	0	0	0	270	0
0	1	0	0	0	0	0	0	0	0	1	1	270	0
0	1	0	1	0	0	1	0	1	1	1	0	270	0
0	0	0	0	0	1	1	0	1	0	0	0	270	0
0	0	0	0	0	0	1	0	0	1	0	1	270	0

ANEXO B: Rotina para a análise dos dados.

```
dados<-read.table(''dados.txt'', head=T)
attach(dados)
# dados
require(survival)
fit<-coxph(Surv(tempos,cens)~EC+ES+I+TU,data=dados,
x=T,method=''breslow'')
summary(fit)
fit$loglik
#-----------------------------
#Obtenção da Figura 4.1
par(mfrow=c(2,2))
fit1<-coxph(Surv(tempos[ES==0],cens[ES==0])~1,data=
dados,x=T,method=''breslow'')
ss<- survfit(fit1)
s0<-round(ss$surv,digits=5)
H0<- -log(s0)
plot(ss$time,log(H0),xlim=range(c(0,300)),xlab=
''Tempos'',ylab=expression(log(Lambda[0]*(t))),
bty=''n'',type=''s'')
fit2<-coxph(Surv(tempos[ES==1],cens[ES==1])~1,data=
dados,x=T,method=''breslow'')
ss<- survfit(fit2)
s0<-round(ss$surv,digits=5)
H0<- -log(s0)
lines(ss$time,log(H0),type=''s'',lty=2)
```

```
legend(200,-3,lty=c(2,1),c(''ES = 1'',''ES =0''),
lwd=1,bty=''n'',cex=0.7)
title(''ES: Nível de escolaridade'')
# Obs: Análogo para as demais covariáveis.
#----------------------------
B<-10000
lin<-nrow(dados)
x<-dados
# matriz que guarda os coeficientes
coefi<-matrix(0,B,4)
coefi[1,]<-fit$coef
set.seed(54321)
for(i in 2:B){
# obter a nova amostra
h=as.integer(runif(lin,1,141))
for(j in 1:lin){
hh=h[j]
x[j,]=dados[hh,]
}
# calcular os coeficientes
fit<-coxph(Surv(tempos,cens)~ES+I+TU+EC,data=x,x=T,
method=''breslow'')
coefi[i,]<-fit$coef
}
# armazenar os valores de ES, I, TU e EC
ES<-coefi[,1]
```

```
I<-coefi[,2]
TU<-coefi[,3]
EC<-coefi[,4]
# médias e variâncias dos dados
mean(ES)
mean(I)
mean(TU)
mean(EC)
var(ES)
var(I)
var(TU)
var(EC)
#------------------------------
# iteração
iter <- 50000
# Variáveis
Beta <- c(0,iter)
Sigma <- c(0,iter)
# valor inicial
Beta[1] <- 0.6
Sigma[1] <- 0.06
# taxa de aceitação
txacBeta <- 0
txacSigma <- 0
Media <-  0.5843961
sigmaaux <- 0.06212787
```

```
# dados
x <- ES
n <- 10000
# truncamento à direita de lim
rnt<- function(n,m,s,lim)
{
res<-qnorm(pnorm(lim,mean=m,sd=s)+runif(n)*(1-pnorm
(lim,mean=m,sd=s)),mean=m,sd=s)
}
# processo iterativo
for (i in 2:iter) {
# Beta
soma <- 0
BetaCand <- rnt(1,Media,sqrt(sigmaaux),0)
cand1 <- log(dnorm(BetaCand,0,10000000000000000))
cand2 <- log(dnorm(Beta[i-1],0,10000000000000000))
for (j in 1:n) {
soma <- soma + (x[j]-Beta[i-1])^2
}
cond1 <- -1/2*((Beta[i-1]-Media)/Sigma[i-1])^2+
((2-n)/2)*log(soma)
for (j in 1:n) {
soma <- soma + (x[j]-BetaCand)^
}
cond2 <- -1/2*((BetaCand-Media)/Sigma[i-1])^2+
((2-n)/2)*log(soma)
```

```
b <- cond1+cand1-cond2-cand2
alfa <- min(exp(b) ,1)
a <- runif(1,0,1)
if (a<=alfa) {
Beta[i] <- BetaCand
txacBeta <- txacBeta + 1
}else {
Beta[i] <- Beta[i-1]
}
# sigma
SigmaCand <- runif(1,0,1)
cand1 <- dunif(SigmaCand,0,1)
cand2 <- dunif(Sigma[i-1],0,1)
cond1 <- sqrt(SigmaCand)
cond2 <- sqrt(Sigma[i-1])
b <- (cond1*cand1)/(cond2*cand2)
alfa <- min(b,1)
# parte da aceitação
a <- runif(1,0,1)
if (a<=alfa) {
Sigma[i] <- SigmaCand
txacSigma <- txacSigma + 1
}else {
Sigma[i] <- Sigma[i-1]
}
}
```

```
# convergência
# carregar biblioteca do BOA
library(boa)
# Cria a matriz de análise
Conv<-matrix(0,iter,2,dimnames=list(c(1:iter),
c(''Beta'',''Sigma'')))
Conv[,1] <- Beta
Conv[,2] <- Sigma
# Raftery e Lewis
boa.randl(Conv, 0.025, 0.005, 0.95, 0.001)
# Geweke
boa.geweke(Conv, 0.1, 0.5)
# Heidelberger e Welch
boa.handw(Conv, 0.1, 0.05)
# Gelman e Rubin
# Executado pelo boa.menu() seguir os passos
dessa função
# Função de autocorrelação
boa.acf(Conv,c(1,5,10,20,30,40,50))
#------------------------------
# Tamanho da amostra Final
#------------------------------
# burn= período de aquecimento da cadeia
# thin = salto para independência
burn <- 5000
thin <- 10
```

```
# tamanho final da amostra
final <- (iter-burn)/thin
aux4 <- burn+1
# Novas variáveis guardam os valores finais
Betaf <- rep(0,final)
Sigmaf <- rep(0,final)
for (i in 1:final) {
Betaf[i] <- Beta[aux4]
Sigmaf[i] <- Sigma[aux4]
aux4 <- aux4+thin
}
# Média e desvio-padrão
cat(mean(Betaf),sd(Betaf),''\n'')
cat(mean(Sigmaf),sd(Sigmaf),''\n'')
library(boa)
# Intervalo de credibilidade HPD - cada um
boa.hpd(Betaf,0.05)
boa.hpd(Sigmaf,0.05)
#-----------------------------
# Obtenção da FIGURA 4.2
par(mfrow=c(1,2))
plot(Betaf, type=''l'',xlab=''Iterações'',ylab=
''Beta (ES)'')
f1.d <- density(Betaf)
hist(Betaf, prob=T,xlab=''Beta (ES)'',
ylab=''Frequência'',main='''')
```

```
lines(f1.d)
par(mfrow=c(1,2))
plot(Betaf, type=''l'',xlab=''Iterações'',
ylab=''Beta (ES)'')
f1.d <- density(Betaf)
hist(Betaf, prob=T,xlab=''Beta (ES)'',
ylab=''Frequência'',main='''')
lines(f1.d)
#------------------------------
# O procedimento é análogo para as covariáveis TU, I
e EC.
#------------------------------
# Caso a priori for truncada à esquerda de lim usa-se:
rnt <- function(n,m,s,lim)
{
res<-qnorm(pnorm(lim,mean=m,sd=s)*runif(n),mean=m,sd=s)
}
#------------------------------
```