UNIVERSIDADE ESTADUAL DA PARAÍBA

CAMPUS I

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Kildery Truta Diniz

USO DE REGRESSÃO A DADOS DE LEUCEMIA LINFOBLÁSTICA POR MEIO DE ANÁLISE DE SOBREVIVÊNCIA CLÁSSICA E BAYESIANA

Campina Grande - PB

31/10/2016

Kildery Truta Diniz

USO DE REGRESSÃO A DADOS DE LEUCEMIA LINFOBLÁSTICA POR MEIO DE ANÁLISE DE SOBREVIVÊNCIA CLÁSSICA E BAYESIANA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira

Campina Grande - PB

31/10/2016

D585u Diniz, Kildery Truta.
Uso de regressão a dados de leucemia linfoblástica por meio de análise de sobrevivência Clássica e Bayesiana [manuscrito] / Kildery Truta Diniz. - 2016.
32 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2016.
"Orientação: Prof. Dr. Tiago Almeira de Oliveira, Departamento de Estatística".

1. Análise de sobrevivência. 2. Leucemia linfoblástica. 3. Inferência Bayesiana. 4. Estatística. I. Título.

21. ed. CDD 519.5

Kildery Truta Diniz

USO DE REGRESSÃO A DADOS DE LEUCEMIA LINFOBLÁSTICA POR MEIO DE ANÁLISE DE SOBREVIVÊNCIA CLÁSSICA E BAYESIANA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 31 de outubro de 2016.

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Silvio Fernando Alves Xavier Junior
Universidade Estadual da Paraíba (UEPB)

Eu dedico esse Trabalho de Conclusão de Curso, primeiramente a Deus que me deu toda força necessária para que eu pudesse chegar onde estou hoje. Dedico também a minha família que sempre esteve ao meu lado, me apoiando sempre nas horas de felicidade e de dificuldades que surgiram na minha vida pessoal e no decorrer da minha vida profissional e por fim, aos meus amigos que me apoiaram com palavras de força quando tudo parecia impossível de ser alcançado.

Agradecimentos

Agradeço a Deus por tantas bençãos alcançadas durante minha vida estudantil, principalmente na vida acadêmica, conquista alcançada tanto como pessoa quanto profissional, agradeço a Ele por ter me proporcionado não somente vitórias, mas também situações ao qual foram testadas minha fé, força, perseverança e persistência para fazer o correto sempre.

Agradeço aos meus pais **Roberto** e **Magna**, a minha irmã **Roberta Vitória**, por terem me apoiado sempre, desde o início de mais uma etapa da minha vida, fazendo com que nunca deixasse me abater pelas dificuldades da vida, mostrando que, a vida só vale a pena quando existem dificuldades para que possamos sair cada vez mais forte e experiente de cada situação, mas também desfrutando dos momentos de felicidade, compartilhando cada vitória alcançada, algo que para mim, todas foram de grande representatividade.

Aos meus amigos Rodolfo, Alisson, Leandro, Arthur, que me ouviam e apoiavam (ou não) nos momentos difíceis, debatendo situações e resoluções para sanar qualquer incômodo existente no âmbito pessoal e profissional sempre com muita alegria, parceria e brincadeiras. Não podendo esquecer também dos meus parceiros de classe Cleanderson e Antônio, pessoas aos quais devo muito também, por me aguentarem quando discutíamos para resolver determinado problema de sala, mas também por estarmos sempre unidos, batalhando para vencermos juntos, assim como os Três Mosqueteiros, nunca deixando um ou outro para trás.

À todos os professores do curso de Bacharelado em Estatística, que fizeram esse momento ser possível, capacitando-me para que eu pudesse crescer academicamente e assim qualificar-me mais e mostrar tudo o que meu curso pode proporcionar na academia ou no mercado de trabalho.

"Quando você pensar que é tarde demais tenha cuidado,
não deixe que isso seja uma desculpa para você desistir,
ninguém pode impedi-lo de ter sucesso, exceto você mesmo.
Seja o mais brilhante possível quando for chegada a hora de brilhar."
(Dushun Wang)

Resumo

Esse trabalho objetivou utilizar modelo de regressão com abordagem clássica e bayesiana a dados de pacientes com leucemia. Para atingir tais objetivos, foi realizada uma revisão de Análise de Sobrevivência e Inferência Bayesiana com teorias e aplicabilidades de cada área. Teve-se como finalidade analisar a recuperação dos pacientes com leucemia após o transplante de medula óssea. Dentre os 137 pacientes estudados, 58% eram do sexo masculino e 42% do sexo feminino com idades entre 7 e 52 anos. Para esse estudo foram analisadas as covariáveis, tempo de sobrevivência, idade e sexo do paciente, idade e sexo do doador, tempo até desenvolver doença do hospedeiro aguda, tempo até desenvolver doença do hospedeiro crônica, hospitais que realizaram o procedimento, foi analisado também se o paciente desenvolveu a doença do hospedeiro aguda ou se desenvolveu a doença do hospedeiro crônica, entre outras, utilizando modelo de regressão Weibull com abordagens clássica e bayesiana.

Palavras-chave: Análise de Sobrevivência. Leucemia Linfoblástica. Inferência Bayesiana.

Abstract

This work aims to use classical regression models and a Bayesian approach to leukaemia patients data. There was a revision of theories and applicability in survival analysis and Bayesian inference. 19 specific variables were analysed in 4 American hospitals; Ohio State University Hospital (OSU), Hahnemann University (HU), Hospital St. Vicent (SVH) and Alfred Hospital (AH). Among 137 patients studied, 58% were male, and 42% were women, aged about 7 and 52 years old. Some covariates were analysed, amongst them; patients age, donor age, patient sex, donor sex, time to acute GVHD, acute GVHD, time to chronic GVHD, chronic GVHD, platelet recovery, hospital, FAB and others. Weibull Regression Model was used for the analysis.

Key-words: Survival. Lymphoblastic Leukemia. Bayesian Inference

Sumário

1 Introdução

Análise de Sobrevivência (A.S.) é a expressão utilizada para designar a análise estatística de dados quando a variável em estudo representa o tempo, desde um instante inicial bem definido até à ocorrência de determinado acontecimento de interesse. De acordo com Colosimo e Giolo (2006), a variável aleatória em estudo é não negativa e pode representar o tempo até a ocorrência de um evento de interesse, denominada tempo de falha.

Uma característica importante da análise de sobrevivência é a presença de censura. Em estudos de sobrevivência, para cada indivíduo, as observações são representadas por um vetor que contenha o tempo de falha (t_i) e uma variável indicadora de falha ou censura (δ_i). Se para cada indivíduo do estudo houver vetor de covariáveis (x_i), então o vetor é representado por (t_i, δ_i, x_i) segundo (STRAPASSON, 2007).

Em A.S., existem três tipos de censuras, são elas, censura à direita, à esquerda e intervalar. A censura à direita é utilizada com maior frequência em pesquisas médicas que não se observa o desfecho do tratamento aplicado no paciente. Em contra-ponto, são utilizadas todas as informações obtidas durante o período vigente do tratamento nos pacientes. A censura à esquerda, acontece quando não conhecemos o momento de ocorrência do evento, mas sabe-se que o fenômeno decorreu antes do tempo inicial do estudo. E a censura intervalar é quando não se sabe o tempo exato da ocorrência do evento de interesse, apenas é constatado que sucedeu-se dentro de um intervalo de tempo específico. Toda informação obtida acerca do estudo, seja antes ou depois do estudo, deve-se levar em consideração no momento das análises, com o intuito de evitar subestimar ou superestimar o risco de sofrer o evento (CARVALHO, 2011).

É sabido que, pode-se utilizar a Inferência Bayesiana para diversos estudos, sejam eles Saúde, Economia, Geo-Estatística, entre outros. Conforme citado em Resende (2001), na Inferência Bayesiana, os parâmetros a serem estimados são considerados variáveis aleatórias e são estimadas considerando as incertezas associadas a ela.

Para se realizar um estudo de inferência bayesiana, é levado em consideração a distribuição à priori, definida como o conhecimento prévio que o especialista tem sobre o parâmetro de interesse θ, anterior a observação dos dados, a verossimilhança, que resulta na estimação dos parâmetros de interesse a partir dos dados e a distribuição à posteriori, que pode ser interpretada uma ponderação das duas fontes de informação estimadas.

O trabalho teve como objetivo realizar o estudo em A.S. clássica e bayesiana com o uso de softwares R (TEAM, 2013) e OpenBUGS (THOMAS, 2004) e interpretar estes resultados em um modelo de regressão com aplicação a dados de Leucemia Linfoblástica.

2 Desenvolvimento

2.1 Análise de Sobrevivência

A análise de sobrevivência é o conjunto de técnicas e modelos estatísticos usados na análise do comportamento de variáveis positivas. Segundo Colosimo e Giolo (2006), a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. 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 e são categorizadas em três tipos, censura à direita, à esquerda e intervalar.

A censura à direita é mais utilizada em casos em que não se observa o desfecho do tratamento aplicado no paciente onde todas as informações obtidas são utilizadas para que não haja superestimação ou subestimação a ponto de influenciar o risco de sofrer o evento de interesse. A censura à esquerda acontece, quando o fenômeno decorreu antes do tempo inicial do estudo. E por fim, a censura intervalar, que é classificada quando o evento sucedeu-se dentro de um intervalo de tempo específico desconhecido, dado por algum fator externo.

Segundo Carvalho (2011), truncamento dos dados ocorre quando indivíduos são excluídos do estudo por motivo relacionado à ocorrência do evento. Em outras palavras, acontece quando o estudo inclui somente indivíduos em que o evento ocorreu dentro de uma janela temporal pré-estabelecida denominada truncamento à esquerda T_E ou truncamento à direita T_D).

2.1.1 Tempo

A variável aleatória T é não negativa, representada pelo tempo de falha e é usualmente especificada pela função de sobrevivência ou função de risco, segundo Strapasson (2007). Então, consideremos que T assume distribuição contínua com função de probabilidade $f(t)$, a função de distribuição acumulada é descrita por:

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{2.1}$$

A **função de sobrevivência** é definida como a probabilidade de um indivíduo sobreviver até um certo tempo t, sem a ocorrência do evento em estudo, descrita pela equação abaixo:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du, \tag{2.2}$$

em que a probabilidade de sobrevivência $S(t) = 1$, quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$ (STRAPASSON, 2007).

A **função de risco**, ou taxa de falha, descreve o risco instantâneo de um indivíduo sofrer o evento entre o tempo t e $t + \Delta t$ dado que ele sobreviveu ao tempo t, definida formalmente por Qian (1994), descrita pela equação:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{2.3}$$

Segundo Strapasson (2007), a função de risco pode ser reescrita, em termos da função de densidade de probabilidade $f(t)$ e da função de sobrevivência $S(t)$, da seguinte forma:

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}. \tag{2.4}$$

2.1.2 Distribuição Weibull

Segundo Strapasson (2007), uma variável aleatória T com distribuição Weibull com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\gamma > 0$, isto é, $T \sim$Weibull(γ,λ) tem função de densidade de probabilidade, descrita por:

$$f(t) = \frac{\gamma}{\lambda^\gamma} t^{\gamma-1} exp\left\{-\left(\frac{t}{\lambda}\right)^\gamma\right\}, t \geq 0 \tag{2.5}$$

função de sobrevivência, descrita por:

$$S(t) = exp\left\{-\left(\frac{t}{\lambda}\right)^\gamma\right\}, t \geq 0 \tag{2.6}$$

e função de risco dada por:

$$h(t) = \frac{\gamma}{\lambda^\gamma} t^{\gamma-1}, t \geq 0 \tag{2.7}$$

2.1.3 Estimação da Função de Sobrevivência

O Estimador de Kaplan Meier é conhecido como não paramétrico pois usa os próprios dados para estimar as quantidades necessárias da análise, sem fazer uso de suposições a respeito da forma da distribuição dos tempos de sobrevivência.

A construção desse estimador considera o número de falhas distintas e os limites dos intervalos do tempo de falha na amostra que são ordenados do primeiro ao último, podendo existir mais de uma falha ao mesmo tempo, $t_1 < t_2 < \ldots < t_i$, segundo Colosimo e Giolo (2006), e é expresso por:

$$\hat{S}_{KM}(t_i) = \prod_{i=1}^{i-1}(1 - q_i) \tag{2.8}$$

em que, $q_i = \frac{d_i}{n_i}$, descrito pelo número de falhas dividido por número de indivíduos sob risco em t_i.

2.1.4 Estimação dos parâmetros

Segundo Colosimo e Giolo (2006), os parâmetros são características dos modelos de probabilidade para estudos de tempo de vida, existindo-se alguns métodos de estimação. O método de máxima verossimilhança é uma opção apropriada para dados censurados, incorporando-se as censuras relativamente simples, por possuir certas propriedades para grandes amostras.

2.1.5 Método de Máxima Verossimilhança

Usualmente a função de máxima verossimilhança é dada pela seguinte equação:

$$L(\theta; x) = \prod_{i=1}^{n} f(t_i|\theta) \tag{2.9}$$

No entanto, na equação acima, o tempo de censura é desconsiderado. Portanto, é necessário utilizar uma variável indicadora, descrita por, $\delta_i = 1$, se houve falha no tempo t_i e $\delta_i = 0$, se ocorreu censura em t_i. Considerando todos os mecanismos de censura, sob a suposição de que eles são não informativos (não carregam informação sobre os parâmetros), a função de verossimilhança, é representada por:

$$L(\theta; x) = \prod_{i=1}^{r} [f(t|\theta)] \prod_{i=1}^{n} [S(t_i|\theta)], \tag{2.10}$$

ou, equivalentemente, por:

$$L(\theta; x) = \prod_{i=1}^{n} [f(t|\theta)]^{\delta_i} \prod_{i=1}^{r} [S(t_i|\theta)]^{1-\delta_i} = \prod_{i=1}^{n} [h(t|\theta)]^{\delta_i} [S(t_i|\theta)]. \tag{2.11}$$

O estimador de máxima verossimilhança de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta; x)$. Segundo Bolfarine (2001, pág. 35-36), em situações em que Θ é discreto ou que o máximo de $l(\theta; x)$ ocorre na fronteira de Θ, o estimador de máxima verossimilhança não pode ser obtido a partir de $l'(\theta; x)$. Para isso, obtêm-se o máximo a partir da inspeção da função de verossimilhança.

2.2 Modelo de Regressão Linear

Em um modelo de regressão linear, tem-se como finalidade estimar valores de uma variável dado que sabemos os valores de uma outra variável. Inicialmente no modelo de regressão, pressupomos que, a relação entre X e Y é linear, sendo assim, descrevemos a função que determina a relação entre as variáveis, na equação abaixo:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.12}$$

em que, Y_i é a variável dependente ou (variável resposta), β_0 representa o intercepto da função, β_1 a inclinação em relação a reta de regressão, X_i é a variável independente e ϵ representa o erro associado a cada valor em relação a reta de regressão.

2.3 Modelo de Regressão Weibull

Segundo Strapasson (2007), podemos descrever o modelo de risco proporcionais via Distribuição, supondo que os valores de p covariáveis são registrados para cada um dos n indivíduos, como sendo:

$$h(t|\mathbf{x_i}) = \frac{\gamma}{\lambda^\gamma} t^{\gamma-1} \tag{2.13}$$

$$\lambda^{-1} = exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\}$$

Uma das características importantes dessa distribuição é que ela apresenta uma grande variedade de formas, por exemplo, quando $\gamma = 1$, a função de taxa de falha é constante e os tempos de sobrevivência têm uma distribuição exponencial. Para outros valores de γ, a função de taxa de falha cresce ou decresce monotonicamente (STRAPASSON, 2007).

2.3.1 Linearização no modelo Weibull

Ao estudar a linearização no modelo Weibull, temos que o logaritmo da função de sobrevivência resulta em:

$$-log[S(t)] = \left(\frac{t}{a}\right)^\theta$$

$$log[-log[S(t)]] = -\theta log(a) + \theta log(t),$$

o que mostra que $-log[S(t)]$ é uma função linear de $log(t)$. Sendo assim, ao expor o gráfico da função Weibull linearizada com o estimador de Kaplan Meier, espera-se que o mesmo resulte em um gráfico aproximadamente linear. Quando além de linear, o gráfico assumir valores na origem e inclinação da reta igual a 1, é um indício a favor do modelo exponencial, (COLOSIMO & GIOLO, 2006).

2.4 Critério de Akaike - AIC

Segundo citado em Emiliano (2010), foi mostrado por (Akaike, 1974) que o viés é dado assintoticamente por p, em que p é o número de parâmetros a serem estimados no modelo e $L(\hat{\theta})$ é definido como a função de verossimilhança, descritos na equação à seguir:

$$AIC = -2logL(\hat{\theta}) + 2p$$

Nesse estudo foi tomado que a função de verossimilhança é dada por $L(\theta; x)$ e p o número de covariáveis explicativas consideradas no modelo, logo podemos reescrever a equação como:

$$AIC = -2logL(\theta; x) + 2(p + 1)$$

2.5 Inferência Bayesiana

Teorema de Bayes

De acordo com Ehlers (2011), a informação que dispomos sobre θ, resumida probabilisticamente através de $p(\theta)$, pode ser aumentada observando-se uma quantidade aleatória X relacionada com θ. A distribuição amostral $p(x|\theta)$ define esta relação. Existe a suposição de que após observar $X = x$, a informação obtida sobre θ aumenta. Abaixo está descrito a equação para tal acréscimo de informação:

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d\theta} \tag{2.14}$$

Note que $1/p(x)$, que não depende de θ, funciona como uma constante normalizadora de $p(\theta|x)$. Para um valor fixo de x, a função $L(\theta; x) = p(x|\theta)$ fornece a plausibilidade ou verossimilhança de cada um dos possíveis valores de θ enquanto $p(\theta)$ é chamada distribuição a priori de θ. Estas duas fontes de informação, priori e verossimilhança, são combinadas levando à distribuição a posteriori de θ, $p(\theta|x)$. Assim, a forma usual do teorema de Bayes é $p(\theta|x) \propto L(\theta; x)p(\theta)$, (EHLERS, 2011).

Conforme citado em Resende (2001), na Inferência Bayesiana, os parâmetros a serem estimados são considerados variáveis aleatórias e são estimadas considerando as incertezas associadas a ela.

Segundo Ehlers (2011), a informação à respeito de uma quantidade θ é desconhecida, logo tentamos mensurar o máximo possível esse valor, com o objetivo de quantificar os graus de incerteza representados através de modelos probabilísticos para θ.

Para se realizar um estudo de inferência bayesiana, é levado em consideração a distribuição à priori, definida como o conhecimento prévio que o especialista tem sobre o parâmetro de interesse θ, anterior a observação dos dados.

Os estudos bayesianos podem ser realizados com prioris informativas, o que implica dizer que, o especialista tem conhecimento sobre o fenômeno observado e, no uso de prioris não informativas, o especialista se exime de qualquer influência no estudo ou não tem conhecimento detalhado sobre o fenômeno.

A função de verossimilhança, permite ao pesquisador estimar o valor do parâmetro de interesse a partir do conjunto de dados. E por fim, definimos a distribuição à posteriori, como sendo uma ponderação das duas fontes de informação estimadas, descrita na equação abaixo:

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

Densidades à posteiori - caso discreto e contínuo

Supondo que tenhamos uma variável aleatória $\mathbf{y} = (y_1, \ldots, y_n)$ e θ assuma valores discretos, temos então que a densidade à posteriori para θ_i dado $\mathbf{y}$ é descrito por:

$$p(\theta_i|\mathbf{y}) = \frac{p(\mathbf{y}|\theta_i)p(\theta_i)}{\sum_{j=1}^{n} p(\mathbf{y}|\theta_j)p(\theta_j)}$$

Para o caso contínuo, supomos que θ assuma valores contínuos em um intervalo qualquer, então a equação anterior resulta em:

$$p(\theta_i|\mathbf{y}) = \frac{p(\mathbf{y}|\theta_i)p(\theta_i)}{\int p(\mathbf{y}|\theta_j)p(\theta_j)d\theta}$$

2.6 Leucemia

A leucemia é uma doença maligna dos glóbulos brancos (leucócitos) de origem, na maioria das vezes, não conhecida. Ela tem como principal característica o acúmulo de células jovens (blásticas) anormais na medula óssea, que substituem as células sangüineas normais. Os principais sintomas da leucemia decorrem do acúmulo dessas células na medula óssea, prejudicando ou impedindo a produção dos glóbulos vermelhos (causando anemia), dos glóbulos brancos (causando infecções) e das plaquetas (causando hemorragias).Depois de instalada, a doença progride rapidamente, exigindo com isso que o tratamento seja iniciado logo após o diagnóstico e a classificação da leucemia, de acordo com o Instituto Nacional do Câncer (INCA, 2016).

O transplante é a substituição de células doentes de medula óssea por células saudáveis. A medula óssea é um tecido líquido que ocupa o interior dos ossos, sendo conhecida popularmente por 'tutano'. Na medula óssea são produzidos os componentes do sangue, por isso, a medula óssea é considerada a fábrica do sangue, segundo (SECRETARIA DE SAÚDE DO TOCANTINS, 2016).

Juntamente com o enxerto de medula entram em circulação no doente células do doador (linfócitos) que atacam quaisquer células do doente que tenham à superfície antigénios que são reconhecidos pelas células do doador como estranhos. As células atingidas libertam produtos químicos que vão estimular outras células do dador e deste conjunto de reações resultam diversas manifestações clínicas que constituem a doença enxerto contra hospedeiro, segundo (SERVIÇO DE TRANSPLANTE DE MEDULA ÓSSEA, 2016).

2.7 Origem dos Dados

O banco de dados estudado é composto de 137 indivíduos (pacientes) onde 99 deles compõem os grupos de baixo e alto risco de desenvolvimento da leucemia linfoblástica e os demais preenchem o grupo de pacientes com leucemia linfoblástica aguda.

Os pacientes foram avaliados em diferentes unidades hospitalares, dentre elas, Hospital da Universidade Estadual de Ohio (OSU) avaliou 76 pacientes, Hahnemann University (HU) avaliou 21 pacientes, no Hospital St. Vicent (SVH) foram estudados os casos de 23 pacientes e por fim, 17 no Alfred Hospital (AH).

O estudo dos hospitais teve como alvo transplantes de medula óssea realizados durante os anos de 1984 à 1989, com o intuito de constatar as causas de ocorrência do evento de interesse ou remissão da doença no paciente tendo em vista todas as covariáveis estudadas por cada instituição, dentre elas foram estudadas:

- Tempo até a morte (t_1);
- Variável indicadora de falha ($\delta = 1 \rightarrow$ falha $\mid \delta = 0 \rightarrow$ censura);
- Grupos de risco (Leucemia Linfoblástica Aguda = 1, Leucemia Mieloblástica Baixo Risco = 2 e Mieloblástica Alto Risco = 3);
- Tempo para doença do hospedeiro aguda (ta);
- Tempo para doença do hospedeiro crônica (tc);
- Doença do hospedeiro aguda (a | Não desenvolveu = 0, Desenvolveu = 1);
- Doença do hospedeiro crônica (c | Não desenvolveu = 0, Desenvolveu = 1);
- Idade do paciente (z_1);
- Idade do doador (z_2);
- Sexo do paciente (z_3 | Feminino = 0, Masculino = 1);
- Sexo do doador (z_4 | Feminino = 0, Masculino = 1);
- Plaquetas em níveis normais (p | Não = 0, Sim = 1);
- FAB[1](z_8);
- Hospital - Hospitais pesquisados (z_9);
- MTX - Metotrexato[2] (z_{10} | Não = 0, Sim = 1).

[1] Classificação franco-americana-britânica que requer uma porcentagem de blastos de pelo menos 30% na medula óssea (BM) ou sangue periférico (PB) para o diagnóstico de leucemia mielóide aguda;

[2] É uma droga antimetabólito e antifolato usada no tratamento do câncer e doenças autoimunes.

2.8 Análise Estatística

A análise clássica foram realizadas utilizando o software R (TEAM, 2013) tendo como foco a utilização do pacote survival (THERNEAU, 2016) para análise clássica, com o intuito de obter os gráficos de estimador de Kaplan-Meier, risco instantâneo, função de risco e linearização para a distribuição estudada.

Primeiramente foi realizado o cálculo do estimador de Kaplan-Meier para os grupos de risco de Leucemia Linfoblástica via software R, obtido a partir da equação $\hat{S}_{KM}(t_i) = \prod_{i=1}^{i-1} \left(1 - \frac{d_i}{n_i}\right)$. O passo seguinte foi verificar o EKM para o desenvolvimento da doença de hospedeiro aguda e crônica.

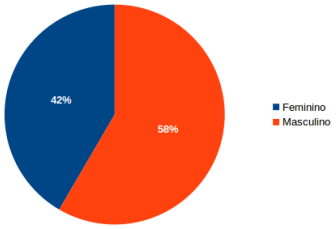
Para seleção do modelo via análise clássica, foram inseridos as covariáveis do estudo para obtenção do modelo de regressão através da ferramenta *stepwise*, que tem como finalidade analisar quais covariáveis são estatisticamente significativas para descrever o modelo selecionado com as respectivas estimativas. Esta ferramenta realiza o procedimento de inserção e remoção das covariáveis provindas do banco de dados implementadas internamente até atingir-se as covariáveis estatisticamente significativas para o modelo.

A análise bayesiana foi realizada com o auxílio do software OpenBUGS (THOMAS, 2004), que utiliza amostragem de Gibbs por meio de Markov Chain Monte Carlo (MCMC) para realização das simulações, possibilitando estimar as densidades à posteriori para cada parâmetro de interesse. A execução final da análise bayesiana foi realizada através do pacote R2OpenBUGS (STURTZ, 2010), com o total de 10000 simulações. Nessa simulação foi considerado uma quantidade de descarte (burn-in) das 1000 primeiras simulações com salto (thin) de 10 em 10 amostras, o que implica numa amostra final de 900 simulações.

3 Resultados

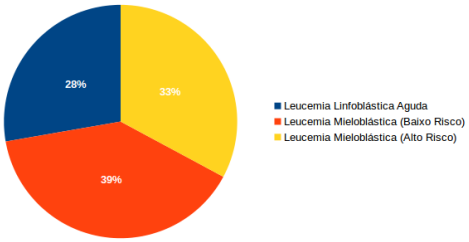
Inicialmente foram expostos alguns gráficos descritivos, tabelas, histogramas, para ter uma explanação mais detalhada sobre algumas das variáveis estudadas do conjunto de dados. Sabendo disso, a figura 1 descreve a porcentagem referente aos sexos dos 137 pacientes submetidos ao transplante e também no estudo, dado que o foco do trabalho é observar o tempo de sobrevivência para cada indivíduo. Logo, observou-se que 58% foram representados por pacientes do sexo masculino e 42% pacientes do sexo feminino.

Figura 1 – Sexo dos pacientes que fazem parte do estudo



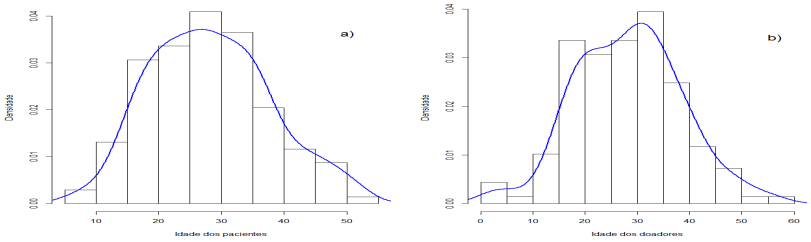
Os pacientes submetidos ao transplante de medula óssea foram classificados antes do início do estudo em três diferentes níveis de risco, sendo eles, Leucemia Linfoblástica Aguda e Leucemia Mieloblástica (Baixo e Alto Risco), descritos na figura 2 com as respectivas porcentagens de cada grupo.

Figura 2 – Pacientes categorizados por grupos de Risco



Na figura 3, estão descritos os histogramas referentes as idades dos pacientes e doadores de medula óssea, descritas em a) e b), respectivamente.

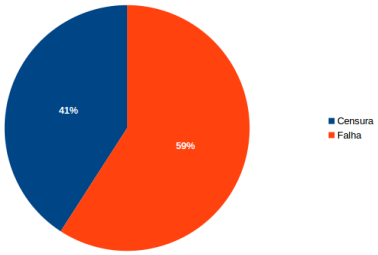
Figura 3 – Histograma da idade dos pacientes e doadores de medula óssea



Em ambos os histogramas observou-se que, a população de pacientes observados para esse estudo tinha idade entre 7 e 52 anos de idade e, a população dos doadores tinha idade observada entre 2 e 56 anos de idade.

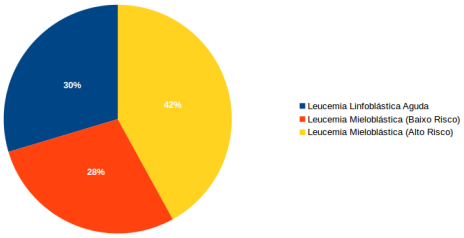
A ocorrência de falha em um paciente, significa dizer, que o mesmo veio à óbito, durante o período vigente do estudo por causas do fenômeno estudado. Na figura 4, constatou-se que, 59% vieram dos pacientes falharam, sem levar em consideração a classificação por grupos de risco.

Figura 4 – Gráfico dos pacientes que falharam durante período vigente do estudo



Do total de 81 pacientes que vieram à óbito, 30% correspondiam ao grupo de Leucemia Linfoblástica Aguda, 42% ao grupo de risco de Leucemia Mieloblástica Alto Risco e 28% representam o grupo de Leucemia Mieloblástica Baixo Risco, como apresentado na figura 5.

Figura 5 – Gráfico dos pacientes classificados por grupo que vieram à óbito



Nas figuras 6 e 7 estão dispostos os gráficos do estimador de Kaplan-Meier com os respectivos tempos de sobrevivência.

Figura 6 – Gráfico do estimador de Kaplan Meier para os grupos de Leucemia

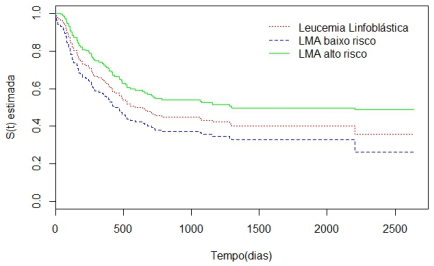
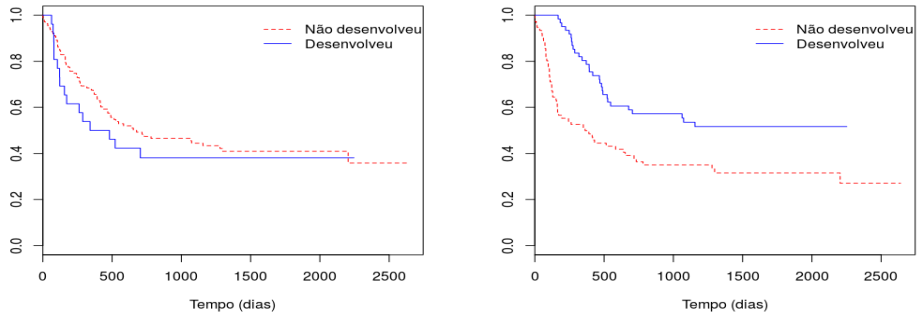


Figura 7 – Gráfico do estimador de Kaplan Meier para a presença ou ausência da doença do hospedeiro aguda e crônica, respectivamente



Observou-se que a curva de sobrevivência estimada não descresce até zero para ambos os gráficos, dando indícios da presença de fração de cura, que descrevem pacientes que não ocorreu o desfecho.

Após a realização do Kaplan Meier, procedeu-se o ajuste do modelo de regressão Weibull. Este modelo foi escolhido em função de um conjunto de fatores estudados previamente, porém não apresentados neste estudo em detalhes, deste modo, serão apenas citados seus resultados. O Teste da razão de verossimilhança entre as distribuições Weibull e Exponencial que levaram ao descarte da distribuição Exponencial, devido o $p - valor$ entre as distribuições ser de $3,55 \times 10^{-7}$ obtidos da distribuição qui-quadrado. O intervalo de confiança para θ na distribuição Weibull foi de [0,50 ; 0,75], não contemplando o valor 1.

Após a escolha da distribuição que modelaria os dados de sobrevivência, ajustou-se um modelo de regressão Weibull com todas as covariáveis e em seguida realizou-se o método de seleção de variáveis de *stepwise* via critério de AIC.

As variáveis selecionadas que descrevem melhor os dados foram, *Tempo para doença do hospedeiro aguda e crônica (ta e tc), doença do hospedeiro aguda e crônica (a e c)*, *Plaquetas em níveis normais (p)* e *FAB* (z_8). Levando em consideração que o modelo foi selecionado a partir do critério de Akaike com menor valor, foi constatado através do software R, que o AIC para distribuição Weibull foi de 1086,715, enquanto o mesmo modelo analisado pela distribuição exponencial, teve o valor de 1123,308. Assim, o modelo ao qual descreve melhor os dados dentre as variáveis estudadas foi:

$$h(t|\mathbf{x_i}) = \frac{\gamma}{\lambda^\gamma} t^{\gamma-1}$$

$$\lambda^{-1} = exp\{\beta_0 + \beta_1 a + \beta_2 c + \beta_3 p + \beta_4 z_8 + \beta_5 ta + \beta_6 tc\}$$

Após a seleção do modelo através do critério AIC, estão descritas as estimativas, erros padrão, estatística z e p-valor para cada parâmetro de interesse, na tabela 1.

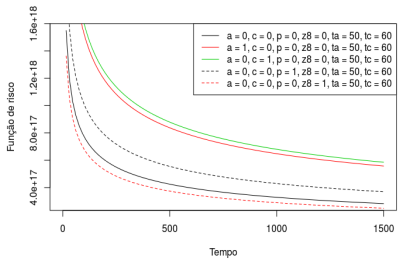
Tabela 1 – Estimativas dos parâmetros do modelo de regressão via análise clássica.

	Estimativa	E.P	z	Valor p
β_0	3,963	0,16	24,97	1,25 x 10^{-137}
β_1	1,077	0,25	4,40	1,08 x 10^{-5}
β_2	1,154	0,20	5,79	6,96 x 10^{--9}
β_3	0,424	0,16	2,67	7,61 x 10^{-3}
β_4	-0,203	0,12	-1,67	9,44 x 10^{-2}
β_5	0,574	0,10	5,80	6,78 x 10^{-9}
β_6	0,723	0,14	5,28	5,31 x 10^{-13}
$\hat{\gamma}$	0,63	0,0617	-	-

A estimativa para β_1, implica dizer que o risco do paciente desenvolver a doença do hospedeiro aguda aumenta em aproximadamente 34,06% com o decorrer do tempo, β_2 aumenta em 31,53% o risco do paciente desenvolver a doença do hospedeiro crônica. β_3 implica no paciente ter nível de plaquetas normais após o transplante que é de 65,44%. β_5 = 56,33% está associado ao tempo para diagnóstico da doença do hospedeiro aguda e β_6 = 48,53% ao tempo para diagnóstico da doença do hospedeiro crônica. Como foi constatado, o risco do paciente ser classificado no nível (4, 5 ou Leucemia Mieloblástica) da covariável FAB aumentou em aproximadamente 22,5%, para estimativa de β_4.

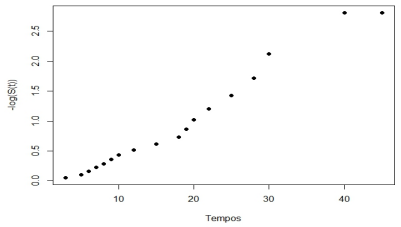
O risco instantâneo é o termo utilizado para analisar o risco ao longo do tempo de ocorrência de falha do indivíduo em função de alguma covariável de interesse. Sabendo disso, na figura 8 está descrito a função de risco para a distribuição Weibull.

Figura 8 – Função de risco via análise clássica



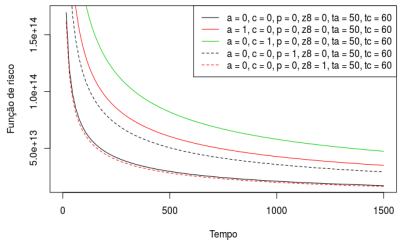
Na figura 9, está disposto o gráfico da função weibull linearizada com o tempo proveniente do conjunto de dados estudado.

Figura 9 – Linearização para a distribuição Weibull



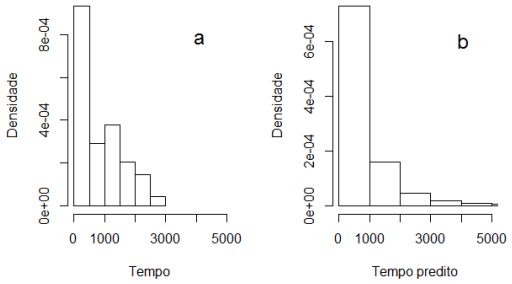
Os resultados descritos a seguir, foram obtidos através das análises realizadas com abordagem bayesiana. No gráfico abaixo, está descrito o risco ao longo do tempo para covariáveis selecionadas para o modelo, com o intuito de mostrar o efeito causado na presença de cada uma delas.

Figura 10 – Função de risco via análise bayesiana



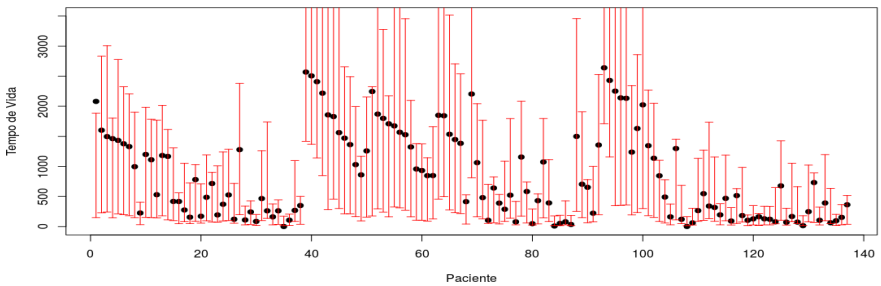
Na figura 11, observa-se o tempo até a morte proveniente dos dados e os tempos preditos, ao qual no histograma[a] (tempo observado) está explicito que o tempo máximo até a morte é de 3000 dias, enquanto que, no histograma[b] (tempo predito), temos que, à partir dos 4000 dias, ambos com uma probabilidade muito perto de zero, de ocorrência.

Figura 11 – Histograma dos tempos até a morte observado e predito, respectivamente.



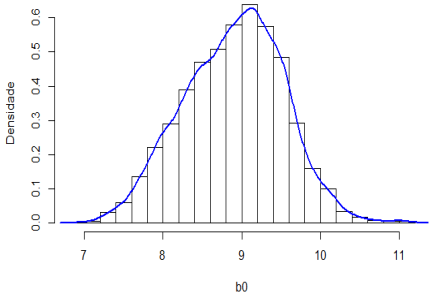
Na figura 12, são mostrados os intervalos de credibilidade com limites inferior e superior, relativo ao tempo de vida de cada paciente do estudo, indicado pelos pontos.

Figura 12 – Gráfico do tempo de vida e intervalos de credibilidade dos pacientes



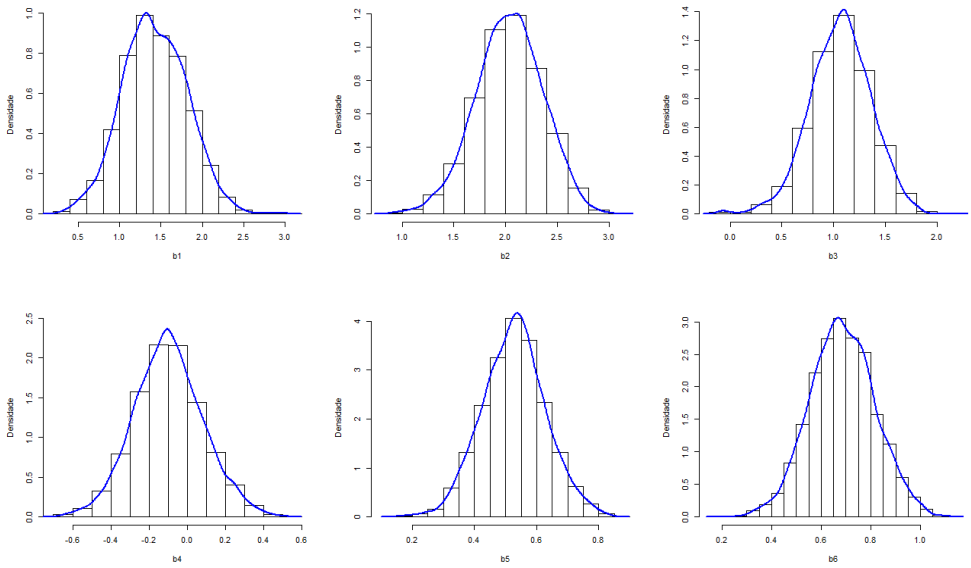
A análise bayesiana foi realizada através de prioris informativas, tomadas a partir das estimativas de cada co-variável no ato da seleção de melhor modelo através da ferramenta *stepwise* via análise clássico. A distribuição à posteriori obtida para representação do intercepto está disposta na figura 13, onde se pode observar que a média de β_0 é de aproximadamente 9, com desvio padrão de $0,63$ e intervalos de credibilidade de [7,7 ; 10,1] para os quantis 2,5% e 97,5%.

Figura 13 – Densidade à posteriori de β_0.



A densidade à posteriori de β_1, teve média de aproximadamente 1,4, com intervalos de credibilidade de [0,66 ; 2,19] para os quantis 2,5% e 97,5%. Os demais β's estimados estão dispostos detalhadamente na tabela 1 e distribuições à posteriori descritos na figura 14, respectivamente.

Figura 14 – Densidades à posteriori para cada $\beta_i \therefore i = 1, \ldots, 6$



As covariáveis selecionadas para representação do modelo são as mesmas tanto na análise clássica quanto na bayesiana. A diferença existente é que, o OpenBUGS segue uma parametrização um pouco diferenciada para implementação das análises, toda via, as estimativas finais de γ são muito próximas dos resultados constatados pela análise clássica, obtendo o modelo descrito abaixo.

$$h(t|\mathbf{x_i}) = \gamma \cdot \lambda t^{\gamma-1}$$

$$\lambda^{-1} = exp\{\beta_0 + \beta_1 a + \beta_2 c + \beta_3 p + \beta_4 z_8 + \beta_5 ta + \beta_6 tc\}$$

Na tabela 2 estão descritas algumas estimativas para os parâmetros de interesse com médias, erros padrão e intervalo de credibilidade, à 2,5% e 97,5%. Para estimativa para β_1, implica dizer que o risco do paciente desenvolver a doença do hospedeiro aguda aumenta em aproximadamente 24% com o decorrer do tempo, β_2 aumenta em 13,07% o risco do paciente desenvolver a doença do hospedeiro crônica. β_3 implica no paciente ter nível de plaquetas normais após o transplante que é de 34,45%. $\beta_5 = 59{,}03\%$ está associado ao tempo para diagnóstico da doença do hospedeiro aguda e $\beta_6 = 50{,}36\%$ ao tempo para diagnóstico da doença do hospedeiro crônica. Como foi constatado, o risco do paciente está no nível de classificação (4, 5 ou Leucemia Mieloblástica) para covariável FAB resultou em um aumento de 10,2% aproximadamente, quantificada pela estimativa de β_4.

Tabela 2 – Estimativas dos parâmetros do modelo de regressão via análise bayesiana.

	Estimativa	E. P.	2,5%	97,5%
β_0	8,8985	0,054	7,658	10,090
β_1	1,4257	0,033	0,6632	2,1890
β_2	2,0349	0,028	1,368	2,6440
β_3	1,0658	0,025	0,4581	1,6290
β_4	-0,0970	0,016	-0.4546	0,2751
β_5	0,5272	0,009	0,3267	0,7350
β_6	0,6859	0,011	0,4301	0,9458
$\hat{\gamma}$	0,5005	0,0316	0,4500	0,5692

Podemos ressaltar que, a disparidade existente entre as estimativas obtidas via análise clássica e bayesiana, são referentes ao fato do estudo bayesiano ter sido realizado através de prioris informativas, o que implica dizer que, existiu mais informação a respeito do banco de dados estudado, levando a intervalos de credibilidade de menor amplitude que os intervalos clássicos.

Ao fim das análises, foi constatado que, a covariável *ta*, tem uma grande influência no resultado do paciente pós transplante, visto que, para a estimativa deβ_5, temos que o tempo influencia diretamente no risco de morte do paciente que é de mais de 50% tanto na análise clássica quanto bayesiana. Para a variável *tc* resultou que, o tempo para desenvolver

a doença do hospedeiro crônica implica num risco de morte maior que 48% em ambas as áreas. Ou seja, à medida que passa o tempo, o risco dos pacientes que foram submetidos ao transplante de medula óssea desenvolverem a doença do hospedeiro aguda ou crônica aumenta.

4 Conclusão

O trabalho desenvolvido teve como objetivo, observar o tempo de sobrevivência dos pacientes submetidos ao transplante de medula óssea.

Foram expostas inicialmente estatísticas descritivas a respeito dos pacientes e doadores envolvidos no estudo. Como método para seleção do melhor modelo para explicação dos dados foi utilizado o critério de Akaike, que resultou nas variáveis a, c, p, ta, tc e z_8, sendo consideradas estatisticamente significativas para compor modelo.

Através da metodologia bayesiana, foram utilizadas prioris informativas obtidas a partir da estimação de cada covariável no intuito de ter uma densidade à posteriori mais informativa. No estudo bayesiano, foram simuladas 10000 observações, com burn-in = 1000 e thin = 10. Com isso, constatou-se que o modelo descrito via análise clássica, obteve DIC igual à 1937 e 8 variáveis no total, dado a quantidade de informação estimada para o modelo.

Ao fim do estudo, evidenciou-se que a Análise de Sobrevivência é uma ferramenta de grande importância na área da saúde, pois adequa-se melhor com os dados da área, principalmente quando vinculada a Inferência Bayesiana, permitindo fazer predições à respeito do fenômeno durante e após o fim do estudo. Não obstante, serão realizados estudos utilizando outras distribuições de probabilidade e fração de cura, com o intuito de observar se existe discrepância entre as variáveis significativas para descrever o modelo de regressão.

5 Referências Bibliográficas

BOLFARINE, H.; SANDOVAL, M.C. **Introdução à inferência estatística**. SBM, 2001.

CARVALHO, M.S. et al. **Análise de Sobrevivência: teoria e aplicações em saúde**. SciELO-Editora FIOCRUZ, 2011.

COLOSIMO, E.A.; GIOLO, S.R. **Análise de sobrevivência aplicada**. In: ABE-Projeto Fisher. Edgard Blücher, 2006.

EHLERS, R.S. **Inferência Bayesiana**. Departamento de Matemática Aplicada e Estatıstica, ICMC-USP, 2011.

EMILIANO, P.C. et al. **Critérios de informação de Akaike versus Bayesiano: análise comparativa**. 19º Simpósio Nacional de Probabilidade e Estatística, 2010.

INSTITUTO NACIONAL DO CÂNCER (INCA). Disponível em: <http://www.inca.gov.br/conteudo_view.asp?id=344>. Acesso em: 4 de nov. 2016.

LEUKEMIA. Disponível em: <http://www.nature.com/leu/journal/v19/n9/full/2403876a.html>. Acesso em: 4 de nov. 2016.

QIAN, J. **A Bayesian Weibull survival model**. 1994. Tese de Doutorado. Duke University.

RESENDE, M.D.V. et al. **Análise de modelos lineares mistos via inferência Bayesiana**. Rev. Mat. Estat, v. 19, p. 41-70, 2001.

SECRETARIA DE SAÚDE DO TOCANTINS. Disponível em: <http://saude.to.gov.br/atencao-a-saude/hemorrede/doacao-de-medula-ssea/>. Acesso em: 4 de nov. 2016.

SERVIÇO DE TRANSPLANTE DE MEDULA ÓSSEA. Disponível em: <http://www.stmo.com.pt/pt/apoio-ao-doente/informacao-ao-doente/reinternamentos/122-doenca-enxerto-contra-hospedeiro.html>. Acesso em: 4 de nov. 2016.

STRAPASSON, E. **Comparação de modelos com censura intervalar em análise de sobrevivência**. 2007. Tese de Doutorado. Escola Superior de Agricultura "Luiz de Queiroz.

STUDIO, R. RStudio: integrated development environment for R. RStudio Inc, Boston, Massachusetts, 2012.

STURTZ, S.; LIGGES, U.; GELMAN, A. R2OpenBUGS: a package for running OpenBUGS from R. URL http://cran. rproject. org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS. pdf, 2010.

TEAM, R. Core et al. R: A language and environment for statistical computing. 2013.

THERNEAU, T.M.; LUMLEY, T. Package 'survival'. 2016.

THOMAS, A.; O'HARA, R.B. OpenBUGS. 2004.

Anexos

Código R - Análise Clássica

```
library(survival)
survival.data = read.table("dados",header = TRUE, sep = ",")
attach(survival.data); ekm <- survfit(Surv(t1,death) ~ 1)
plot(ekm, lty = c(3,2,1), col = c("red", "blue", "green"),
xlab = "Tempo(dias)", ylab = "S(t) estimada")
legend(1500,1,lty = 3:1, c("Leucemia Linfoblástica","LMA baixo risco",
"LMA alto risco"), lwd = 1,bty = "n", col = c("red", "blue", "green"))
st <- ekm$surv
regweib <- survreg(Surv(t1, death) ~ factor(g) + ta + factor(a) + tc +
factor(c) + tp + factor(p) + z1 + z2 + factor(z3) + factor(z4) +
factor(z5) + factor(z6) + z7 + factor(z8) + factor(z9) + factor(z10),
dist = "weibull"); step(regweib)
new_regweib <- survreg(Surv(t1, death) ~ ta + factor(a) + tc +
factor(c) + factor(p) + factor(z8), dist = "weibull"); AIC(new_regweib)
```

Código R - Análise Bayesiana

```
library(R2OpenBUGS)
tempo = c(vetor do tempo observado); Nsim = 10000
model <- function(){
for(i in 1 : 137) {
cens[i] <- 1-falha[i]
tempo[i] ~ dweib(v, lambda[i])%_%C(cens[i],)
lambda[i] <- 1/exp(b0 + b1[a[i]+1]+ b2[c[i]+1] + b3[p[i]+1]+
b4[z8[i] +1] + b5*ta[i]/365+ b6*tc[i]/365)
theta[i] <- 1/lambda[i] #tempo medio de vida
tempopred[i] ~ dweib(v, lambda[i])%_%C(cens[i],)}
b0 ~ dnorm(4, 1.0E-1)
b5 ~ dnorm(0.6, 1.0E-1)
b6 ~ dnorm(0.7, 1.0E-1)
v ~ dunif(0.1, 100)
sig2 <- 1/ v
b1[1] <- 0.0
b1[2] ~ dnorm(1.1, 1.0E+0)
b2[1] <- 0.0
b2[2] ~ dnorm(1.2, 1.0E+0)
b3[1] <- 0.0
b3[2] ~ dnorm(0.42, 1.0E+0)
b4[1] <- 0.0
b4[2] ~ dnorm(-0.20, 1.0E+0)}
data <- list(variáveis selecionadas (tempo, falha, z8, ta, a, tc, c, p))
inits <- function() {list(v = 1)}
out <- bugs(data = data, inits = inits, parameters.to.save = c("tempopred",
"b0", "b1", "b2", "b3", "b4", "b5", "b6"),
model.file = model, n.chains = 1, n.burnin= 1000, n.thin = 10, n.iter = Nsim)
par(mfrow=c(1,2))
hist(out$sims.list$tempopred, xlim=c(0, 5000), breaks = 30, freq = FALSE)
hist(tempo, freq = FALSE, xlim=c(0, 5000))
q1 = q2 = q3 = 0
for(i in 1:137) {
q1[i] = quantile(out$sims.list$tempopred[,i], c(0.025))
q2[i] = quantile(out$sims.list$tempopred[,i], c(0.5))
q3[i] = quantile(out$sims.list$tempopred[,i], c(0.975))}
n = 137
```

```
plot(tempo[1:n], ylab='Tempo de Vida', xlab = 'Paciente', pch=19,
ylim = c(0, 3500))
arrows(1:n, q1, 1:n, q3, length=0.05, angle=90, code=3, col=2)
bi = out$sims.list$bi; acf(bi); hist(bi, main='') .:. (i = 0,...,6)
c(mean(bi), sd(bi), quantile(bi, c(0.025, 0.5, 0.975)))
```