Relatório PIBIC

1 Introdução

A Análise de sobrevivência é a área da Estatística aplicada nas áreas da saúde, das ciências sociais e econômicas chegando até às engenharias. Por possuir tantas aplicações, essa área também é chamada de Análise de Confiabilidade. Nesse tipo de análise o objeto de estudo é o tempo até a ocorrência de um evento de interesse. Sua grande diferença com relação às outras áreas da Estatística é a presença de censura no estudo, ou seja, dados cujo a observação de tempo não se concluiu por algum motivo. O tempo juntamente com essa censura, formam a variável resposta. Considerando o tempo como uma variável aleatória, é possível aplicar certos tipos de gráficos e técnicas com o propósito de descobrir uma possível distribuição para esses dados e assim realizar possíveis inferências e tirar as devidas conclusões. Na grande maioria dos trabalhos nessa área mensuram a variável tempo de forma contínua, pela grande quantidade de distribuições de probabilidade as quais os dados podem se ajustar. Quando a variável é medida de forma discreta, é necessário utilizar de alguns artifícios para o estudo. Um desses artifícios é o uso da discretização de uma distribuição de probabilidade. Esse método causa uma adaptação nos métodos tradicionais da análise de sobrevivência. Segundo @colosimo, uma distribuição de probabilidade muito utilizada como alternativa de distribuições mais flexíveis é a Log-Logística, pois é aplicável em muitas situações práticas. O trabalho abordará dois bancos de dados, sendo que o primeiro é o banco de dados utilizado por Barreto et al. (1994). Este banco de dados estuda pouco mais de mil e duzentas crianças e estuda o tempo entre a suplementação de vitamina A ou placebo até o primeiro caso de diarréia. O segundo banco de dados, é um banco cedido pela Universidade Estadual da Paraíba (UEPB). A variável de estudo nesse caso é o tempo que um estudante de química leva para evadir do curso.

2 Revisão de Literatura

2.1 Análise de Sobrevivência

Em determinados tipos de estudo é desejável estudar o tempo até a ocorrência de determi- nado evento de interesse, na Estatística, o nome que se dá a esse tipo de estudo é Análise de

Sobrevivência. O evento de interesse pode ter nomes diferentes para diferentes áreas, em geral o termo usado pela maioria é falha, na medicina esse evento pode ser a morte do paciente, a cura ou a manifestação de uma doença. Na engenharia, em geral, esse termo se refere a falha de um equipamento e nas ciências econômicas esse termo pode se referir a inadimplência de um determinadoo cliente.

O tempo nesse estudo pode ser medido de diversas formas como dias, meses, anos e até intervalos de tempo pré-determinados. Por ser um tipo de estudo que é observado ao longo do tempo, o acompanhamento de determinadas observações pode ser interrompido por diversos motivos. Dentre esses motivos podem estar a desistência de um paciente em participar do estudo por motivos pessoais ou o defeito de um produto por outro motivo que não o desejado. Diferentemente de outras áreas da Estatística, na Análise de Sobrevivência esses dados incompletos também são utilizados e se caracterizam como censuras. Essa censura forma a variável resposta juntamente com as observações completas do tempo. Dentre essas censuras existem três tipos principais, à direita, à esquerda e a intervalar. A censura à direita acontece quando o tempo registrado no estudo é maior que o tempo de início do estudo. A censura à esquerda ocorre quando a evento de interesse ocorre antes mesmo do início do estudo. E a censura intervalar ocorre quando o evento ocorre dentro do intervalo de dois tempos, geralmente é encontrada quando existe um acompanhamento periódico do evento. Dentro da censura à direita, ainda é possível realizar a divisão dessa censura em três tipos. A censura de tipo I ocorre quando existe um tempo limite para a ocorrência desse evento, caso esse tempo seja atingido, todos as observações que não manifestaram o evento são marcadas como censura. Na censura de tipo II, o número de falhas é fixado no começo do estudo e ao atingir esse número de falhas, as outras observações são marcadas como censuras. A censura de tipo III é a censura aleatória e engloba as duas censuras anteriores,

esta censura se caracteriza por possuir censuras que não se sabe o motivo dela ter acontecido. Nesse trabalho será utilizada a censura à direita aleatória. A utilização desse tipo de censura no estudo se mostra importante, pois apesar de não ter apresentado falha, há a informação de que a observação ainda poderia apresentar tal falha caso o estudo tivesse continuado. A ausência dessa censura pode causar um viés às estimativas e ainda não mostrar a verdadeira distribuição dos dados. A função de sobrevivência é definida como a probabilidade de um indivíduo não falhar até um determinado tempo t, ou seja, é a probabilidade de uma observação viver além do tempo t. Dada uma variável aleatória T, contínua, não negativa. Pode-se descrever a função de sobrevivência como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \sum_{j=t+1}^{\infty} P(T = k), \qquad t = 0, 1, 2, 3, ... \end{aligned} \tag{1}$$

A função de risco acumulado é uma função que não possui uma interpretação simples, porém possui importância dentro do campo da análise de sobrevivência. Esta função, denotada como H(t), pode ser definida como o logaritmo da função de sobrevivência multiplicado por menos um, ou seja:

$$H(t) = -\log(S(t)) \tag{2}$$

2.2 Estimação da Função de Sobrevivência por Kaplan Meier

O estimador a ser usado nesse trabalho será o estimador não-paramétrico de Kaplan-Meier. Esse estimador é muito popular em pesquisas que usam análise de sobrevivência. O estimador é escrito da seguinte forma:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j}$$

.

Onde, n_j representa o número de dados em risco de falha, d_j são os dados que falharam no tempo t_j, em que, $0 \leq t_{(1)} \leq \ldots \leq t_{(n)}$, são os tempos distintos de falha. Esta técnica não utiliza covariáveis para a estimação, mas pode usar variável categóricas para verificar se as funções estimadas são diferentes.

A representação gráfica desse método se comporta em uma função da forma de escada, uma vez que a estimação entre o tempo $t_{(j)}$ e $t_{(j+1)}$ é constante.

2.3 Distribuição Log-Logística

Para os casos onde T é uma variável aleatória contínua seguindo uma distribuição Log-Logística, sua função densidade de probabilidade é descrita como

$$f(t) = \frac{\lambda \left(\frac{t}{\mu}\right)^{\lambda-1}}{\mu \left[1 + \left(\frac{t}{\mu}\right)^{\lambda}\right]^2}, \qquad t > 0 \tag{3}$$

$$S(t) = \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\lambda}} \tag{4}$$

E a partir desta função pode-se encontrar a função risco acumulado:

$$H(t) = -\log \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\lambda}} \tag{5}$$

Segundo @damiao, dada uma variável aleatória contínua T, é possível encontrar sua função de probabilidade discretizada a partir de sua função de distribuição de probabilidade e função de Sobrevivência, através de:

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq T < t + 1) \\ &= P(T < t + 1) - P(T \leq t) \\ &= F_T(t + 1) - F_T(t) \\ &= [1 - S_T(t + 1)] - [1 - S_T(t)] \\ &= S_T(t) - S_T(t + 1) \end{aligned} \tag{6}$$

Dado que T é uma variável com distribuição Log-Logística, a função de probabilidade discretizada dessa variável pode ser descrita como:

$$p(t) = \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\lambda}} - \frac{1}{1 + \left(\frac{t+1}{\mu}\right)^{\lambda}} \tag{7}$$

Com relação ao comportamento da função de risco, quando λ é menor que 1, esta é monótona decrescente, enquanto para valores maiores que 1 a função tem um comportamento monótono crescente.

3 Modelo de Regressão Log-Logístico discreto

Uma das técnicas mais utilizadas para tentar explicar a variabilidade de uma variável segundo um conjunto de covariáveis é o modelo de regressão. Na análise de sobrevivência, utiliza-se de covariáveis para tentar explicar o tempo até a ocorrência do evento de interesse, com base na correlação entre a variável tempo e cada uma das covariáveis. Segundo @damiao, dado um vetor de covariáveis $\boldsymbol{x}^T = (1, x_1, ..., x_p)$ utiliza-se uma função de ligação $g(.)$ que conecte a variável resposta ao vetor $\boldsymbol{x}^T$.

Definindo o preditor linear como $\boldsymbol{x}^T\boldsymbol{\beta}$, em que, $\boldsymbol{\beta}$ é o vetor com p+1 coeficientes de regressão. Considerando uma variável aleatória T com distribuição log-logística discreta definida na seção 2.3, pode-se utilizar o parâmetro de escala μ, com $\mu > 0$ como a função de ligação, ou seja, $\mu = g(\eta) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Com isso, é possível descrever o modelo de regressão Log-Logístico discreto como:

$$p(t|x) = \frac{1}{1 + \left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}} - \frac{1}{1 + \left(\frac{t+1}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}}. \tag{8}$$

A partir dessa função, é possível descrever também a função sobrevivência da seguinte forma:

$$S(t|x) = \frac{1}{1 + \left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta}}\right)^{\lambda}} \tag{9}$$

Como a função risco acumulado pode ser descrita como uma função da função de sobrevivência, esta possui a seguinte forma:

$$H(t|x) = -\log \frac{1}{1 + \left(\frac{t}{\boldsymbol{x}^T\boldsymbol{\beta}}\right)^{\lambda}} \tag{10}$$

4 Método de Máxima Verossimilhança para dados discretos

Para a estimação dos parâmetros da função de distribuição, e também para os parâmetros do modelo, existe uma grande variedade de formas para efetuar tal procedimento. Como a característica principal da análise de sobrevivência é a presença de censuras, o procedimento também deve incorporar tal característica. Por esse motivo, é descartado alguns métodos para a estimação.

Um método que consegue incorporar a censura é o método da máxima verossimilhança. Este método tem como objetivo encontrar o valor do parâmetro que maximiza a probabilidade da amostra observada ser encontrada. Este método mostra-se adequado por permitir a incorporação das censuras através da inclusão da função de sobrevivência para os tempos censurados, enquanto os tempos em que ocorreram falha, considera-se a função densidade.

Para os tipos de censura à direita mostrados, a função de máxima verossimilhança a ser maximizada pode ser descrita analíticamente e a menos de constantes, é dada por [**?**]:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[p(t_i;\boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i;\boldsymbol{\theta})\right]^{1-\delta_i}, \tag{11}$$

onde δ_i é a variável indicadora de falha e $\boldsymbol{\theta}$ é o vetor de parâmetros que serão estimados.

Para o modelo Log-Logístico, utilizando as equações 8 e 9, a função de máxima verossimilhança a ser maximizada possui a seguinte forma:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[\frac{1}{1 + \left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}} - \frac{1}{1 + \left(\frac{t+1}{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}\right)^{\lambda}}\right]^{\delta_i} \left[\frac{1}{1 + \left(\frac{t}{\exp(\boldsymbol{x}^T\boldsymbol{\beta}}\right)^{\lambda}}\right]^{1-\delta_i}, \tag{12}$$

A partir da equação **??**, é possível obter os parâmteros do modelo encontrando o ponto de máximo global na função. Isto pode ser feito ao resolver o sistema:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

5 Análise de Dados

A fim de ilustrar o funcionamento do modelo de regressão Log-Logístico, serão utilizados dois bancos de dados com aplicações distintas. O primeiro banco de dados é o banco utilizado por @Barreto. Este banco estuda o tempo desde a suplementação de vitamina A ou placebo, até o primeiro episódio de diarréia em crianças com idades entre 0 e 24 meses. Os dados foram obtidos ao acompanhar pouco mais de 1200 crianças e possui em sua estrutura 3 variáveis explicativas.

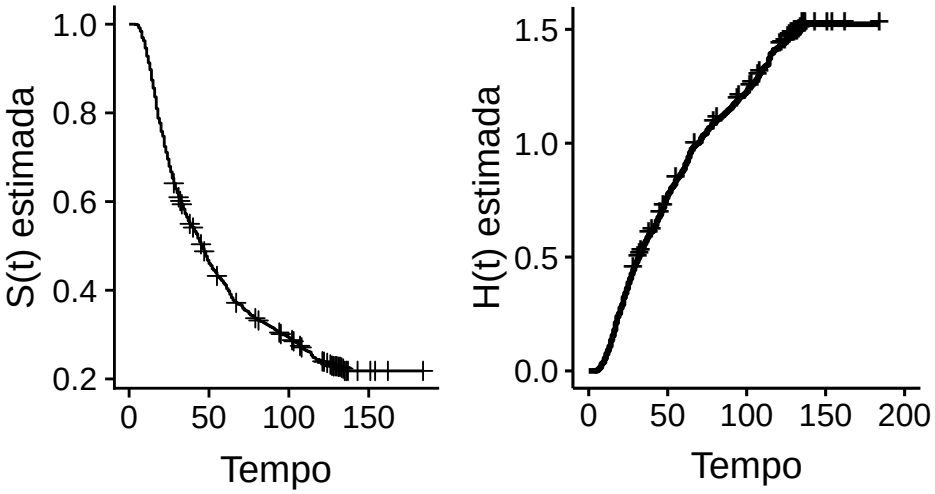
O segundo banco de dados foi cedido pela Universidade Estadual da Paraíba (UEPB). O banco estuda o tempo que um aluno leva para evadir do curso de Química nesta universidade com mais de 600 alunos presentes no conjunto. O banco possui informações sobre os alunos como covariáveis.

5.1 Suplementação de Vitamina em Episódios de Diarréia

No estudo de suplementação de vitamina, o evento de interesse era a ocorrência do primiero episódio de diarréia em crianças com idades entre 0 e 24 meses. Para o estudo, foi definido que um episódio da doença é quando a doença persiste por uma sequência de dias. Além da covariável que indica o uso de vitamina A ou placebo, o banco também conta com a idade da criança e o sexo. Para o estudo, foram coletadas as informações de 1207 crianças.

5.1.1 Análise Descritiva

Para observar inicialmente o comportamento da variável tempo e levar em consideração as censuras ao longo do estudo, utiliza-se o estimador de Kaplan-Meier para a função de sobrevivência e assim poder estimar a curva de sobrevivência. Utiliza-se a função risco acumulado estimada para encontrar uma possível distribuição para realizar a modelagem através de seu gráfico.



5.1.2 Análise Descritiva

6 Referências