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 (**??**), 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 (**??**). 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...

2 Revisão de Literatura

2.1 Análise de Sobrevivência

Em determinados tipos de estudo é desejável estudar o tempo até a ocorrência de determinado 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.

2.2 Funções

Pelo tempo estudado na Análise de Sobrevivência ser dado de forma aleatória, sua distribuição pode ser descrita como a aproximação de uma variável aleatória. Considerando-se T = [Y], em que [Y] é a parte inteira dos valores da variável Y, esta variável é descrita como uma variável aleatória discreta. Esse tipo de definição do tempo não é muito utilizado dentro desse tipo de estudo pelo pequeno número de variáveis discretas que representem bem as especificidades do tempo. Como uma forma de realizar esse estudo, é possível discretizar a função distribuição de probabilidade de uma variável. A seguir, serão apresentadas funções utilizadas na Análise de Sobrevivência descritas tanto para variáveis contínuas quanto para discretas.

2.2.1 Função Densidade de Probabilidade

Dada uma variável aleatória, não negativa, que descreva o tempo até a ocorrência de determinado evento. Chama-se função densidade de probabilidade, uma função f, que descreva a probabilidade de um evento, quando o intervalo de ocorrência desse evento tende a zero. Sendo assim, uma função descreve a distribuição de probabilidade ao longo do intervalo de zero a infinito.

Uma função densidade de probabilidade é uma função que satisfaz as seguintes condições:

1. $f(x)\ e\ \geq 0$ para todo x,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ ou $\sum_{i=0}^{\infty} p(x) = 1$, em que f(x) é para uma variável contínua e p(x) para discreta,
3. para quaisquer a e b com $-\infty < a < b < \infty$, $P(a \leq X \leq b) = \int_a^b f(x)dx$ ou $\sum_{i=a}^{b} p(x)$, para variáveis contínuas e discretas respectivamente.

2.2.2 função de sobrevivência

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:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \int_t^{\infty} f(v)dv \end{aligned} \tag{1}$$

Assumindo que T possuam apenas valores inteiros não negativos, ou seja, $T \in \{0, 1, 2, ...\}$, a variável assume caráter discreto e por isso perde as propriedades anteriores. Para esse caso, a função de Sobrevivência é dada por:

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

onde P(T=t) é a função de probabilidade da variável discreta T.

2.2.3 Função de Risco Acumulado

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:

$$H(t) = -\log(S(t)) \tag{3}$$

O gráfico desta função pode assumir algumas diferentes formas. Essas formas são utilizadas para determinar possíveis modelos probabilísticos que melhor se adequam aos dados. Com relação ao comportamento da função, o gráfico pode tomar as seguintes formas:

- Reta diagonal ⇒ Função de risco constante é adequada.
- Curva convexa ou côncava ⇒ Função risco é monotonicamente crescente ou decrescente, respectivamente.
- Curva convexa e depois côncava ⇒ Função risco tem forma de **U**.
- Curva côncava e depois convexa ⇒ Função risco tem comportamento unimodal.

2.3 Estimadores da função de sobrevivência

2.3.1 Estimação simples

A função de sobrevivência pode ser estimada amostralmente, como a proporção dos dados que não falharam até o tempo t. Esse estimador poderia ser escrito da seguinte forma:

$$\hat{S}(t) = \frac{n^o \textit{ de dados com tempo } > \textit{ } t}{n^o \textit{ total de individuos}}, \forall \textit{ } t \textit{ } \in t \geq 0$$

Caso os dados sejam ordenados de forma crescente, pode-se representar a função de sobrevivência da seguinte forma:

$$\hat{S}(t) = \frac{n_j - d_j}{n}$$

Onde n_j é o número de indivíduos que podem falhar, d_j é o número de indivíduos que que falharam no tempo e n é o número total de indivíduos.

2.3.2 Kaplan-Meier

Os estimadores apresentados acima, não podem ser usados nesse tipo de estudo porque não existe nenhuma forma de se incluir censuras.

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.

2.4 Modelos de Probabilidade

2.4.1 Modelo Log-Logístico

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

$$f(t) = \frac{\beta \left(\frac{t}{\mu}\right)^{\beta-1}}{\mu \left[1 + \left(\frac{t}{\mu}\right)^{\beta}\right]^2}, \qquad t > 0 \tag{4}$$

Onde μ e β são constantes, ambas maiores que 0. Com base na função densidade, é possível descrever a função de sobrevivência da variável aleatória T como:

$$S(t) = \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\beta}} \tag{5}$$

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

$$H(t) = -\log \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\beta}} \tag{6}$$

Segundo (**??**), 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:

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq T < t + 1) \\ &= P(T < t + 1) - P(T \leq t) \\ &= F_T(t + 1) - F_T(t) \\ &= [1 - S_T(t + 1)] - [1 - S_T(t)] \\ &= S_T(t) - S_T(t + 1) \end{aligned} \tag{7}$$

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

$$p(t) = \frac{1}{1 + \left(\frac{t}{\mu}\right)^{\beta}} - \frac{1}{1 + \left(\frac{t+1}{\mu}\right)^{\beta}} \tag{8}$$

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

2.5 Método de Máxima Verossimilhança

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 (**??**):

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[f(t_i;\boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i;\boldsymbol{\theta})\right]^{1-\delta_i}, \tag{9}$$

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

3 Metodologia

3.1 Material

Tendo como objetivo a aplicação do modelo de regressão proposto serão utilizados dois bancos de dados com áreas de estudo distintas. Um dos bancos de dados foi cedido pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia, enquanto o segundo banco de dados foi cedido por???????????.

O banco de dados relacionado a Saúde Coletiva foram obtidos a partir de um estudo conduzido por Barreto(**??**), tal estudo tinha como objetivo avaliar avaliar o efeito da suplementação de vitamina A em casos de diarréia no Nordeste do Brasil no período de dezembro de 1990 e dezembro de 1991. O banco acompanha, durante esse período, 1207 crianças com idades entre 6 e 48 meses no início do estudo, que no início do estudo receberam placebo ou vitamina A. A variável referente ao tempo de sobrevivência nesse estudo foi definida como o número de dias entre a primeira dose do suplemento até a primeira sequência de dias onde a criança apresentou diarréia. Essa sequência de dias foi chamada de episódio da doença. O banco compreende um total de 5 variáveis, sendo estas, o tempo de sobrevivência, que possui caráter numérico e discreto, a presença de censura na observação, que possui caráter binário, a idade em meses da criança, com caráter quantitativo discreto, o sexo da criança e o tipo do suplemento, sendo que este pode ser placebo ou vitamina A.

3.2 Métodos

3.2.1 Modelo de Regressão Log-Logístico Discreto

Nos estudos de análise de sobrevivência, pode-se verificar a influência de determinadas covariáveis sobre o tempo de sobrevivência do indivíduo. A partir disto, é comum utilizar estas covariáveis para tentar explicar o tempo de sobrevivência da população em questão por meio de um modelo de regressão.

Utiliza-se uma função de ligação g(.) que liga a variável resposta ao vetor de variáveis explicativas x^T, onde $x^T = (1, x_1, ..., x_p)$. O vetor de parâmetros θ que será estimado a partir das p covariáveis