Universidade de Brasília
IE - Departamento de Estatística
Estágio Supervisionado 1

Análise de Sobrevivência para Dados Grupados

Daniel Lima Viegas

Orientador: Prof.ª Juliana Betini Fachini Gomes

Brasília
Setembro de 2017

Daniel Lima Viegas

Análise de Sobrevivência para Dados Grupados

Orientadora:
Prof[a]. Dr[a]. **Juliana Betini Fachini Gomes**

Monografia apresentada para a obtenção do título de Bacharel em Estatística.

Brasília
2018

Sumário

1 Introdução

A análise de sobrevivência é um tópico importante utilizado em diversas áreas, como biologia, engenharia, medicina, entre outros. O principal objetivo desta análise é explicar ou predizer o tempo até a ocorrência do evento estudado, esse tempo é chamado de tempo de falha. A principal diferença desta técnica de modelagem para as demais é a capacidade de levar em consideração também os tempos em que não foi possível observar o evento de interesse, esse tipo de ocorrência é chamado de censura.

Dentre os tipos de censuras existentes, a mais genérica é a censura intervalar. Esse tipo de censura ocorre quando não é possível determinar o tempo de ocorrência, mas se tem o intervalo de tempo onde ele ocorreu. Por exemplo, no estudo sobre o tempo até uma lâmpada queimar, deixa-se a lâmpada ligada até que ela queime, em um dia, ela está funcionando, o pesquisador sai da área onde está acontecendo o experimento e quando retorna, a lâmpada está queimada. Neste caso, sabe-se que o intervalo de tempo onde a lâmpada queimou é entre o tempo em que o pesquisador saiu e o que ele voltou. O objetivo deste trabalho é estudar o comportamento de dados grupados, que é um caso particular da censura intervalar.

Em diversos estudos de sobrevivência, estuda-se o relacionamento de covariáveis e o tempo, tendo o objetivo de realizar as análises estatísticas e tentando encontrar o melhor uso dessas variáveis para a criação de um modelo de regressão para dados censurados.

O trabalho possui como objetivo, sugerir um modelo para dois bancos de diferentes áreas utilizando a metodologia de dados grupados. Um dos bancos foi utilizado no estudo de (Barreto et al. 1994) na área da saúde e foi cedido pela Universidade Federal da Bahia. O outro banco de dados é um banco de dados na área da educação cedido pela Universidade Estadual da Paraíba.

2 Revisão de Literatura

A fim de dar fundamento teórico ao trabalho na área de análise de sobrevivência, serão apresentados, a seguir, conceitos e notações presentes nas literaturas do tema.

2.1 variável resposta e censuras

Chama-se evento de interesse, aquilo que se deseja encontrar informações sobre a ocorrência. Na análise de sobrevivência, esse evento pode ser a morte de um indivíduo, a cura, um casamento, divórcio ou funcionamento de um dispositivo ou componente de uma máquina. Em análise de sobrevivência, a variável resposta é geralmente o tempo até a ocorrência de um evento de interesse, sendo esse tempo denominado tempo de falha.

A principal característica dos dados de sobrevivência é a presença de censuras, ou seja, observações que por algum motivo não consegue-se determinar o tempo com precisão. Existem três tipos principais de censura, a mais usual é a censura à direita, esta censura acontece quando não se consegue registrar a ocorrência do evento de interesse. Em estudos médicos que analisam o tempo desde a obtenção da doença até a morte do paciente, por exemplo, este tipo de censura acontece quando o paciente é curado, morre por outra razão ou simplesmente não se pode mais observar tal paciente. Este tipo de censura tem três tipos de classificação:

- Censura do Tipo I: acontece quando a pesquisa tem um tempo pré-determinado. Ao final do estudo, as observações que não falharam são consideradas censuras. Nesse tipo de estudo, o percentual de censura é descrito como uma variável aleatória.
- Censura do Tipo II: É encontrada quando se obtém um determinado número de falhas dentro do experimento. Nesse tipo de experimento, o número de falhas deve ser determinado antes de começar o experimento, fazendo com que o número de falhas seja constante. O número de falhas, claramente deve ser menor do que o tamanho da amostra.
- Censura Aleatória: engloba os outros dois tipos de censura. Acontece quando alguns componentes não podem mais ser acompanha-

dos ou quando o motivo da observação falhar é diferente do que interessa. Esta censura ocorre sem intervenção do pesquisador.

Além dessa censura também existem censuras importantes como a censura à esquerda e a censura intervalar. A censura à esquerda ocorre quando o evento ocorre antes do começo do experimento. Por exemplo, deseja-se observar o tempo até uma criança aprender a ler, porém no começo do estudo, algumas crianças podem já ter aprendido a ler sem saber exatamente o tempo em que ela aprendeu, isto é caracterizado como censura à esquerda.

A censura intervalar pode ser dita como um caso genérico das outras censuras. Chama-se censura intervalar quando não se sabe o tempo em que ocorreu o evento de interesse ocorreu, porém sabe-se que ele não ocorreu antes de um determinado tempo. Por exemplo, em um estudo médico é necessário que hajam visitas regulares para a detecção de certas doenças, tal como câncer. Nesse tipo de experimento, sabe-se que a doença apareceu antes do tempo de uma consulta (V), mas também sabe-se que ela apareceu depois de uma consulta (U), ou seja, a doença se manifestou no intervalo [U, V). Quando $V = \infty$ tem-se a censura a direita, e quando a o tempo $U = 0$, essa censura se torna a esquerda. Daí vem o conhecimento de caso genérico da censura.

2.2 Tempo

Como visto anteriormente, a variável resposta do experimento em questão é o tempo até o evento de interesse. Quando esse tempo pode assumir qualquer ponto real não-negativo, descreve-se essa variável como contínua. Esse é o tipo mais comum de variável na análise de sobrevivência, devido a grande diversidade de distribuições contínuas.

Em alguns casos, não faz sentido utilizar uma distribuição contínua para descrever o tempo, mas sim uma discreta. Pode-se ter como exemplo o tempo que um aluno leva para sair da universidade, pode levar 8 semestres, 9 semestres e assim por diante, ou seja, nunca vai levar um tempo real e sim um tempo pertencente aos naturais.

2.3 Funções

2.3.1 Função densidade de probabilidade

Dada uma variável aleatória contínua, não negativa, que represente o tempo de falha de uma observação. Chama-se função densidade, uma função *f*, que descreva a probabilidade de um indivíduo falhar em um intervalo de tempo, quando esse intervalo tende a zero. Sendo assim, essa função descreve a distribuição de probabilidade ao longo do intervalo de zero a infinito.

A partir desta função, é possível obter-se a função de distribuição acumulada, denominada função *F*. No caso contínuo, esta função é obtida a partir do cálculo da integral da função densidade sobre todo seu suporte. Ou seja, dada uma variável aleatória T:

2.3.2 função densidade de probabilidade

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

1. $f(x) \geq 0$ para todo x,
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$,
3. para quaisquer a, b com $-\infty < a < b < +\infty$, teremos $P(a \leq X \leq b) = \int_a^b f(x)dx$.

2.3.3 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:

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

2.3.4 Função de Risco

Segundo , esta funçao especifica a taxa de falha instantânea no tempo t dado que o indivíduo não falhou até esse tempo. Desta forma, é possível definir a função de risco como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, \delta t)$, supondo que esse indivíduo não falhou até o tempo t, dividida pelo comprimento do intervalo infinitesimal δt

2.3.5 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:

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

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.

2.4 Estimadores da função de sobrevivência

2.4.1 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.

2.4.2 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.

2.5 Modelos de Probabilidade

2.5.1 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

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

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:

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

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

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

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.

2.6 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 (Colosimo E e Giolo 2006):

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \left[f(t_i;\boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i;\boldsymbol{\theta})\right]^{1-\delta_i}, \tag{6}$$

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

Para encontrar o estimador de máxima verossimilhança, é necessário encontrar o vetor gradiente da função. Caracteriza-se como um candidato a estimador os valores em que o gradiente da função é igual a 0. O candidato deve atender ao requisito de que a matriz gerada pela segunda derivada da função é positiva definida.

2.7 Dados Grupados

A análise de sobrevivência com censura intervalar, tem como caso particular os dados grupados. Essa particularidade acontece quando acontece um número excessivo de empates, ou seja, os tempos de vida se

repetem diversas vezes. Este tipo de experimento pode ser encontrado em estudos de medidas repetidas longitudinais.

Para esse tipo de estudo, os tempos são separados em intervalos arbitrários e disjuntos isso significa que os intervalos não precisam ter o mesmo tamanho e que não podem haver interseções entre os intervalos. As funções utilizadas na análise, tais como função de sobrevivência e de risco são as mesmas utilizadas quando o tempo é contínuo.

A função de verossimilhança

3 Metodologia

3.1 Material

A fim de propor um modelo de regressão para dados grupados, irá ser utilizado um banco de dados cedido pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia. Esses dados foram obtidos a partir de um estudo conduzido por Barreto et al.(1994). O banco de dados é formado por 1207 crianças com idade entre 6 e 48 meses no início do estudo, que receberam placebo ou vitamina A. Dentre as variáveis se encontra a variável tempo, que é o tempo entre a primeira dose de placebo ou vitamina A e a ocorrência de diarreia na criança, a variável idade, tipo de tratamento e a variável sexo.

3.2 Métodos

Para realizar uma análise preliminar no tempo até a ocorrência de diarréia na criança, é necessário estimar a função de sobrevivência para a análise descritiva da variável. Para essa estimação, será usado o estimador de Kaplan-Meier. Este estimador foi escolhido por ser um estimador de máxima verossimilhança, possuindo assim as propriedades de um estimador deste tipo. A partir da estimação da função de sobrevivência, pela propriedade da invariância, pode-se estimar a taxa de risco acumulado. Por meio desta função, encontra-se possíveis distribuições para a variável resposta.

Neste trabalho, tem-se como um objetivo verificar o efeito de covariáveis, como sexo e idade, e a resposta, tempo. Para tal, será proposto um modelo de regressão, que é uma extensão da distribuição de probabilidade assumida para a variável resposta.

Para a estimação dos parâmetros do modelo, não pode-se usar alguns métodos de estimação, tais como o método de mínimos quadrados e de momentos, pois eles não levam em consideração a censura presente nos dados de sobrevivência. Sendo assim, para incorporar a censura na análise dos dados, será utilizado uma adaptação do método de máxima verossimilhança.

Para a análise dos dados, será utilizado o software estatístico R por meio da IDE Rstudio.

Referências

[1]COLOSIMO, E, A. and Giolo, S. R. [S.l.: s.n.].

[Barreto et al. 1994]BARRETO, M. L. et al. Effect of vitamin a supplementation on diarrhoea acute lower-respiratory-tract infection in young children in brazil. *Lancet 344*, p. 228–231, 1994.

[Colosimo E e Giolo 2006]COLOSIMO E, A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Editora Bucher, 2006.