Notas de Aula Análise de Sobrevivência

Alisson Rosa

Índice

Prefácio

Notas de Aula do curso de análise de sobrevivência, as duas referências principais vão ser

- Colosimo (2006)

Em termos práticos, os códigos vão ser desenvolvidos usando as linguagens Python e R, portanto assume-se conhecimentos básicos de ao menos umas dessas linguagens para um bom aproveitamento.

Encontrou algum erro? Pode encaminhar uma issue no repósitório ou se preferir pode fazer um pull request, **qualquer** contribuição construtiva é bem vinda.

1 Introdução

Análise de sobrevivência é a área que estuda o **tempo** até acontecer um evento de interesse acontecer. Como por exemplo:

- Tempo até os modelos convergirem
- Tempo até o equipamento falhar duas vezes
- Tempo até o customer não frequentar mais o local
- Tempo até um indivíduo casar-se por ano
- Tempo para o desenvolvimento da Covid-19

Note portanto, que o evento pode encapsular mais de um fato como **falhar duas vezes**

1.1 Breve História

Originalmente, análise de sobrevivência era usada exclusivamente para para estudos de mortalidade em registros estatísticos. Sabe-se que as primeiras análises estatísticas de processos de sobrevivência foram desenvolvidas pelo estatístico John Graunt, por um longo a análise de sobrevivência foi considerada um instrumento analitico para estudos de biomedicina e estudos demográficos, mas assim como estatística em geral alterou-se fortemente nas últimas decadas junto (causa?) com avanços computacionais, a análise de sobrevivência não foi diferente, pois atualmente possuímos um grande poder computacional para desenvolvimento de métodos estatísticos antes inviaveis

1.2 Porque estudar análise de sobrevivência?

Ok, porque não usar usar modelos de regressão em geral para modelar o Tempo (T)?. Um ponto fundamental aqui, é que nem sempre o evento de interesse acontece, assim gerando o que é conhecido como censura

1.3 Censura

Quando estamos fazendo uma análise de sobrevivência podemos ter indíviduos que por algum motivo tiveram que sair do estudo e não apresentaram o evento de interesse, portanto os dados desses indivíduos são chamados de **censurados**

1.3.1 Tipos de censura

2 Definições básicas

Seja $f(t)$ a função densidade de probabilidade da variável aleatória T (tempo até o evento ocorrer). Definimos a função de distribuição acumulada da variável aleatória T como sendo:

$$F(t) = P(T < t) = \int_0^t f(u)du$$

2.1 Função de Sobrevivência

A função de sobrevivência é o complemento da função de distribuição acumulada, isso é:
A probabilidade de uma observação não falhar até um tempo t:

$$S(t) = P(T > t) = 1 - F(t)$$

Pela Figura 2.1 podemos ter um vislumbre que $S(0) = 1$ e $S(\infty) = 0$.
A demonstração de tais fatos fica a cargo do leitor, basta notar que $S(t) = 1 - F(t)$ e usar as propriedades da F.

Por consequência a probabilidade de não sobreviver até um certo tempo t é:

$$1 - S(t)$$

Assim a probailidade de não sobreviver em um intervalo (t_1, t_2) é dado por:

$$1 - S(t_2) - (1 - S(t_1)) = S(t_1) - S(t_2)$$

2.2 Função taxa de falha

Fornece o potencial instântaneo do evento **ocorrer**, dado que o indivíduo sobreviveu até o tempo t:

$$\lambda(t) = \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

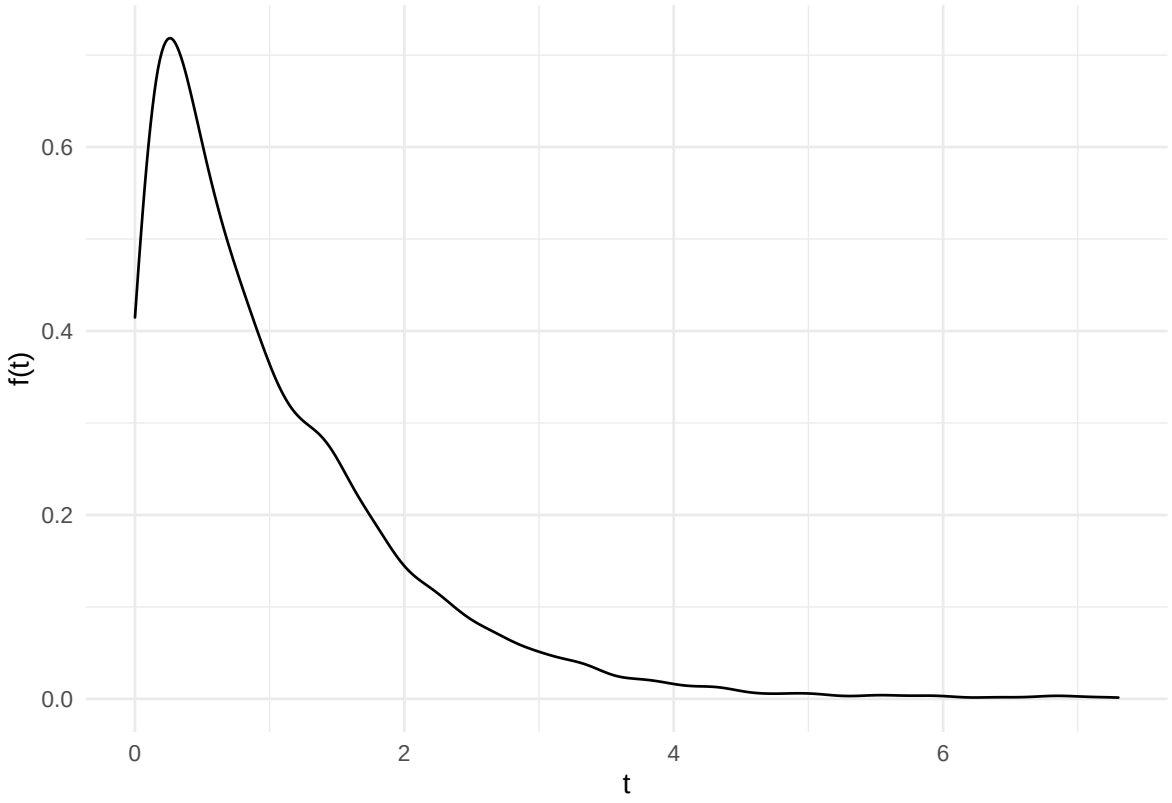

Figura 2.1: Exemplo de uma função de sobrevivência

Note que estamos interessados em que o evento ocorra, ou seja, em termos de interpretação é o **oposto** da função de sobrevivência.

2.2.1 Consequências

- $\lambda(t) \geq 0$
  Demonstração
  É trivial pois por definição medidas de probabilidade $\in [0,1]$ e $\Delta_t \geq 0$ portanto um produto de números positivos
- $\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\Big(log(S(t))\Big)$

3 Programando

Nessa seção vamos aplicar o conceitos em três linguagens de programação, a saber: Python e R

3.1 Layout Básico

Os dados de sobrevivência possuem um layout estabelecido, que é dado a seguir:

Id	T	s_i	X_i	...	X_p
1	t_1	s_1	x_{1i}	...	x_{1p}
2	t_2	s_2	x_{2i}	...	x_{2p}
.	.	.	.	.	.
.	.	.	.	.	.
n	t_n	s_n	x_{ni}	...	x_{np}

R

```
tempo <- c(1, 2, 3, 3, 3, 5, 5, 16, 16, 16, 16, 16, 16,16, 16, 1, 1, 1, 1, 4, 5, 7, 8, 10,
censura <- c(0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0,
grupo <- c(rep(1, 15), rep(0, 14))
dados <- tempo  |>
            cbind(censura)  |>
            cbind(grupo)
dados  |>
        head()
```

```
     tempo censura grupo
[1,]     1       0     1
[2,]     2       0     1
[3,]     3       1     1
[4,]     3       1     1
[5,]     3       0     1
[6,]     5       0     1
```

Python

```
import pandas as pd
import numpy as np

tempo = [1, 2, 3, 3, 3, 5, 5, 16, 16, 16, 16, 16, 16,16, 16, 1, 1, 1, 1, 4, 5, 7, 8, 10, 1
censura = [0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0
grupo = np.concatenate((np.repeat(1, 15), np.repeat(0, 14)))

dados = pd.DataFrame({'tempo':tempo,'censura':censura,'grupo':grupo})
dados.head()
```

```
   tempo  censura  grupo
0      1        0      1
1      2        0      1
2      3        1      1
3      3        1      1
4      3        0      1
```

4 Análise Descritiva

Em Estatística é bastante usual fazer análise descritiva dos dados, como medidas resumo, gráficos e tabelas. Aqui também iremos elaborar análise descritivas, porém com foco nas medidas de sobrevivência e risco.

> **Nota**
>
> O símbolo # aqui é utilizado para indicar a cardinalidade (quantidade de elementos) de um conjunto.
> Por exemplo o conjunto $A = \{a, b, d\}$, possui cardinalidade 3, isso é $\#\text{A} = 3$

4.1 Estimando a função de sobrevivência

Vamos nessa subseção estudar algumas maneiras de estimar a função de sobrevivência $S(t)$

4.1.1 Sem dados de censura

Uma maneira bastante intuitiva para estimarmos $S(t)$ é tomarmos a quantidade de indíviduos que não falharam até o tempo t dividindo pelo total de indíviduos no estudo

$$\hat{S}(t) = \frac{\#\text{Observações que não falharam até t}}{\#\text{Observações}}$$

Exemplo:

5 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier também denominado limite-produto é uma adaptação da idéia 'ingênua' que utilizamos na seção anterior. Ele fornece uma maneira simples, mas eficiente de estimar a função de sobrevivência. De forma intuitiva, dividimos o tempo t em uma série de intervalos de acordo com os eventos observados ou dados censurados, após isso calculamos uma sequência de produto de probabilidade condicionais

5.1 Outra maneira de layout para os dados

t ordenados	int	# de falhas	n_j	m_j
0	$[0, t_1)$	0	k	x_{1p}
t_1	$[t_1, t_2)$	d_2	$k - d_1$	x_{2p}
t_2	$[t_2, t_3)$.	.	.	.
.	.	.	.	.
.	.	.	.	.
k	$[t_k, t_{k+\epsilon})$	d_k	$k - \sum_i d_i$	x_{np}

- **int**: São os intervalos
- **# de falhas**: É o número de falhas naquele intervalo
- n_j : É a quantidade de observações que ainda não falharam naquele intervalo (as vezes chamado de indíviduos sob risco)

Assim fica fácil ver que no tempo 0 temos 0 falhas, porque como vamos ver a seguir a construção dos intervalos começa a partir do primeiro tempo que aconteceo evento, e como temos 0 falhas temos portanto todos os indíviduos sem o evento de interesse

5.2 Formalizando

Sabemos que $S(t) = P(T > t)$, vamos supor que já construímos a tabela e possuímos o tempo 3 e 1, assim queremos calcular, por exemplo:

$$S(3) = P(T > 3)$$

Podemos fazer a seguinte manipulação:

$$S(3) = P(T > 3) = P(T > 1, T > 3) = P(T > 1)P(T > 3|T > 1)$$

> **Nota**
>
> Lembre que $f(X|Y) = \dfrac{f(X,Y)}{f(Y)}$
> E que se A é subconjunto de B, então A $\cap$ B = A (relacione aos intervalos $(1, \infty)$ e $(3, \infty)$

Vamos utilizar o seguinte exemplo para ilustrar a idéia anterior:

t ordenados	int	# de falhas (d_j)	n_j	$\hat{S}(.)$
0	$[0, 1)$	0	14	1
1	$[1, 5)$	3	14	0.78
5	$[5, 7)$	1	9	.
7	$[7, 8)$	1	8	.
8	$[8, 10)$	1	7	.
10	$[10, 16)$	1	6	x_{np}

Para calcular $\hat{S}(1)$, fazemos então:

$$P(T > 1) = P(T > 0, T > 1) = P(T > 0)P(T > 1|T > 0)$$

Sabemos que $P(T > 0) = 1$ como comentado anteriormente. Temos 3 acontecimentos do evento de interesse no tempo $t = 1$ e 14 observações restantes, assim a probabilidade de 'falhar' nesse intervalo é $\dfrac{3}{14}$, portanto a probabilidade de sobreviver é $1 - \dfrac{3}{14}$, substituindo as informações tem-se portanto que $\hat{S}(1) = 0.786$

Para $\hat{S}(5)$, fazemos a mesma decomposição de probabilidade e chegamos em:

$$\hat{S}(5) = P(T > 1)P(T > 5|T > 1)$$

Como calculado anteriormente $P(T > 1) = 1 - \dfrac{3}{14}$ e $P(T > 5|T > 1) = 1 - \dfrac{1}{9}$.

Fica claro então a relação de recursão, pois para o cálculo da estimativa utiliza-se todas as calculadas anteriormente, generalizando temos:

$$S(t_j) = (1 - q_1)(1 - q_2)...(1 - q_j)$$

Onde q_j é a probabilidade de uma observação ter o evento de interesse no intervalo $[t_{j-1}, t_j)$ sabendo-se que não teve em t_{j-1}, formalizando tem-se:

$$q_j = P(T \in [t_{j-1,t_j)}]|T > t_{j-1})$$

Assim o estimador reduz-se a estimar os q_j, reescrevendo usando alguns termos já citados temos:

$$\hat{q}_j = \frac{\#\quad \text{de falhas em} \quad t_j}{\#\quad \text{Observações sobre risco}}$$

6 Summary

In summary, this book has no content whatsoever.

Referências

Colosimo, Suely Ruiz, Enrico Antonio e Giolo. 2006. *Análise de sobrevivência aplicada*. Editora Blucher.