Introdução à análise de sobrevivência — Kaplan Meier Plot e teste de log-rank com Python e R

Como o médico sabe quanto tempo de vida você tem depois de um diagnóstico?

Luísa Mendes Heise · Follow
Published in Turing Talks
8 min read · Jan 31, 2021

Bem-vinde a mais um Turing-Talks. Nesta semana iremos abordar um tipo de análise estatística muito utilizada na área da saúde (mas não só): a análise de sobrevivência (do inglês *survival analysis* ou, mais genericamente, *time-to-event analysis*).

Aqui no Turing-Talks você já deve ter lido sobre regressão logística e regressão linear. De modo simplificado, podemos dizer que a regressão logística nos ajuda a modelar a ocorrência (ou não-ocorrência) de um evento (0 ou 1). Ou seja, se trata de uma modelagem de uma variável binária e sua relação com variáveis contínuas. Por outro lado, a regressão linear nos ajuda a modelar uma variável contínua, como altura ou tempo, e seu relacionamento com outras variáveis contínuas.

Agora, vamos dizer que eu quero fazer algo que "envolve um pouco dos dois mundos". Mais especificamente, quero avaliar o efeito de variáveis categóricas no tempo até um evento.

Vamos pensar num exemplo: você quer verificar se a presença de um gene influencia no tempo que demora para uma criança começar a falar após o nascimento. Você conduz um estudo por 6 anos, acompanhando um grupo de crianças. Quando o estudo acabar, algumas dessas crianças terão falado, outras não. Algumas delas podem ter demorado 3 anos, enquanto outras 5. Além disso, você pode perder o acompanhamento de algumas (que mudaram de país, ou que os pais só não queriam mais que fossem parte do estudo). Como você poderia lidar com todos esses fatores e, por fim, responder a sua questão?

Bom, é aqui que entra a análise de sobrevivência.

O Jargão

Antes de entrarmos em maiores detalhes, é importante definir alguns termos que são parte do 'jargão' da análise de sobrevivência. Vamos fazer isso nos atendo ao primeiro exemplo:

- **Evento (ou falha)**: o evento é a ocorrência de interesse da análise (e essa ocorrência pode ser tanto positiva, como uma criança começar a falar, quanto negativa, como morte). No nosso exemplo: o início da fala da criança.
- **Tempo**: o intervalo desde o início do período de observação até a ocorrência do evento (ou censura do dado… mas vamos falar disso a seguir). No exemplo, o intervalo de tempo desde o nascimento até a fala (ou saída/fim do estudo).
- **Censura/ Dados censurados**: O termo técnico para esses casos em que se perde o acompanhamento ou o evento não ocorre até o final do estudo é de *dados censurados*. Existem diferentes formas para que isso aconteça, mas esses dois tipos são os mais comuns. No nosso exemplo, seria a criança que saiu do país,

que os pais só não queriam mais que fossem parte do estudo ou que não falou até a data do fim do estudo.

- **Função de sobrevivência:** denotada por S(t), essa é a função que nos diz a probabilidade de ocorrência do evento depois de o tempo t. No nosso caso S(t=2 anos) nos daria a probabilidade de uma criança começar a falar após dois anos de tempo de observação.

Entendendo o Plot de Kaplan-Meier e Life-tables

Vamos supor que você foi diagnosticado com uma doença terminal. Provavelmente uma das primeiras perguntas que virá a sua mente é: "quanto tempo eu provavelmente vou viver?" ou então "qual é a chance de eu viver mais de 5 anos". O gráfico de kaplan-meier permite que ambas perguntas sejam respondidas. Ele faz isso dando a **estimativa da função de sobrevivência.**

O cálculo da função de sobrevivência é feito com a ajuda de *life-tables.* Tudo começa com a preparação dos dados para que tenhamos uma tabela com a quantidade de indivíduos "vivos" (ou de casos em que o evento de interesse não foi observado) e indivíduos que "morreram" em instantes de tempo t. Partimos do pressuposto de que, no instante t=0, o evento de interesse (nesse caso morte) não foi observado o evento em nenhum indivíduo. Além disso, só é necessário acrescentar uma linha (para um tempo t) se naquele instante "alguém morreu".

tempo t	vivos no tempo t	mortos no tempo t
0	30	0
1	30	2
3	28	3
5	25	2
7	22	2
10	20	2
15	18	3
20	12	4
22	8	3

O segundo passo é calcular a proporção de "sobreviventes" para cada instante de tempo t.

tempo t	vivos no tempo t	mortos no tempo t	proporção de pacientes sobreviventes no instante t
0	30	0	1.000000
1	30	2	0.933333
3	28	3	0.892857
5	25	2	0.920000
7	22	2	0.909091
10	20	2	0.900000
15	18	3	0.833333
20	12	4	0.666667
22	8	3	0.625000

Nesse caso, denotando a proporção de pacientes sobreviventes no instante t como prop(t), temos:

$$prop(0) = \frac{30}{30} = 100\%$$

E, então:

$$prop(1) = \frac{(30-2)}{30} = 93.9\%$$

Como não temos o instante 2 na tabela, é assumido:

$$prop(1) = prop(2)$$

Ou, mais genericamente, para um instante de tempo t entre t' e t" **que não está na tabela**, temos:

$$prop(t) = prop(t'), \text{ se } t \notin \text{tabela} \wedge t \leq t'' \wedge t \geq t', \text{ com } t' < t''$$

Continuando:

$$prop(3) = \frac{(28-3)}{28} = 89.3\%$$

E, de maneira genérica:

$$prop(t) = \frac{(n_{\text{vivos},t} - n_{\text{mortos},t})}{n_{\text{vivos},t}}$$

Agora, vamos efetivamente calcular a função de sobrevivência: esse cálculo é feito de maneira recorrente, de modo que para calcular S(t+1) precisamos de S(t). Bom, por definição S(t=0)=100%. Daí, o cálculo de S(1), temos:

$$S(1) = S(0) \cdot prop(1)$$

e:

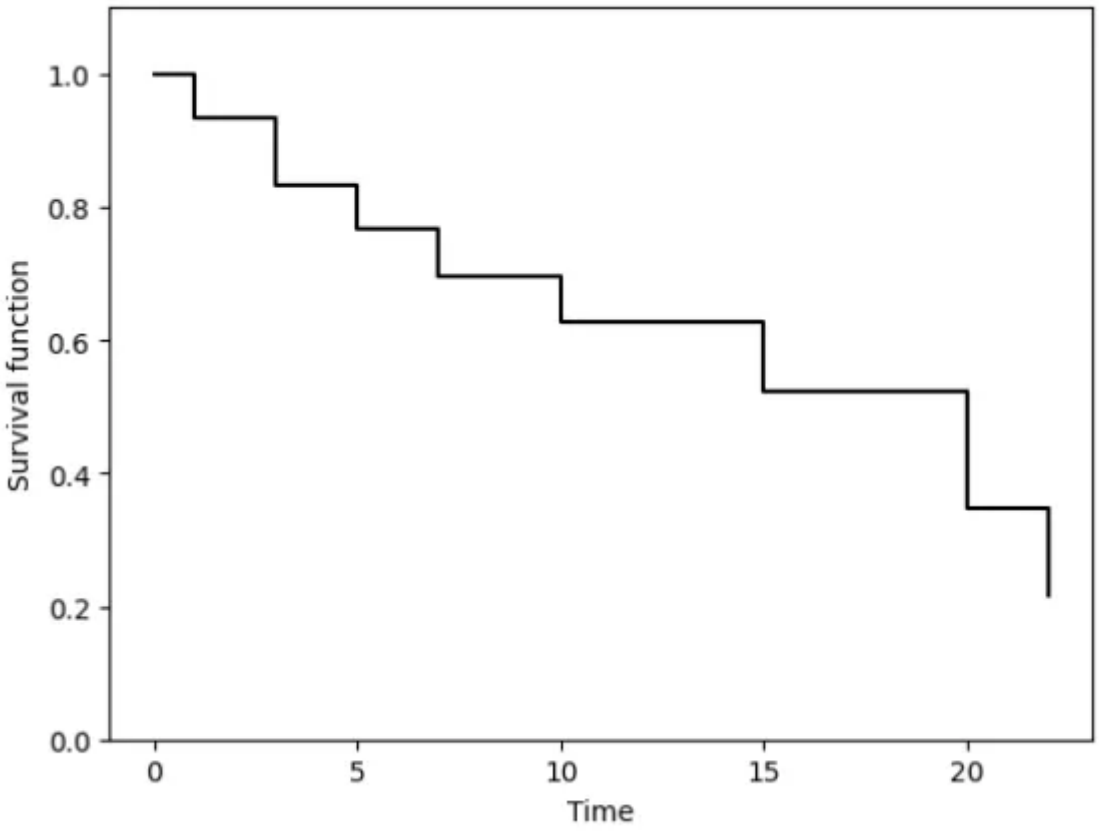
$$S(3) = S(2) \cdot prop(3) = S(1) \cdot prop(3)$$

De maneira genérica, temos:

$$S(t) = S(t-1) \cdot prop(t)$$

tempo t	vivos no tempo t	mortos no tempo t	proporção de pacientes sobreviventes no instante t	função de sobrevivência
0	30	0	1.000000	1.000000
1	30	2	0.933333	0.933333
3	28	3	0.892857	0.833333
5	25	2	0.920000	0.766667
7	22	2	0.909091	0.696970
10	20	2	0.900000	0.627273
15	18	3	0.833333	0.522727
20	12	4	0.666667	0.348485
22	8	3	0.625000	0.217803

O plot de Kaplan-Meier é apenas a função de sobrevivência VS tempo:

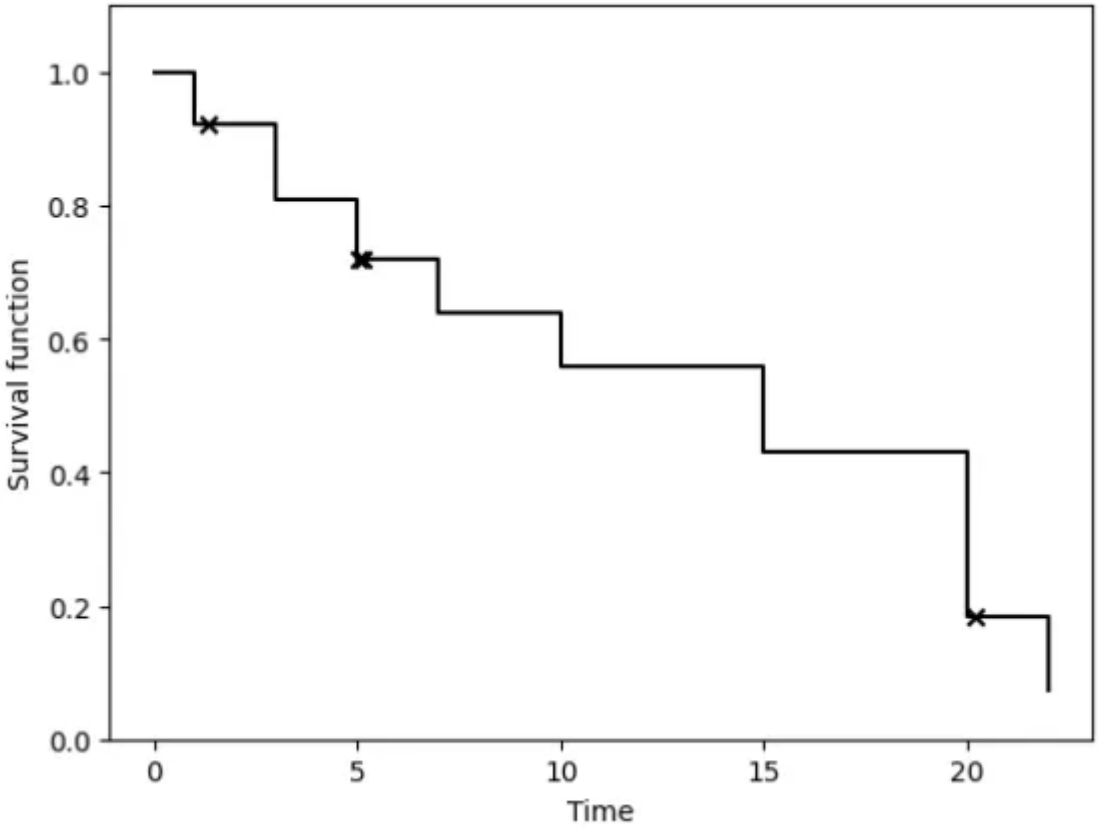


Bom, quase tudo claro... Mas agora precisamos entender onde entram os tais dos *dados censurados*. Quando um paciente é censurado no momento t, sabemos que o paciente estava vivo no momento t, mas não sabemos se o paciente morreu ou sobreviveu. Por esse motivo, os pacientes censurados não são classificados como 'vivos' nem como 'mortos' no momento t. Nós simplesmente os deduzimos do número de pacientes vivos. Daí a única diferença em relação ao caso anterior é que a fórmula de prop(t) fica um pouco diferente:

$$prop(t) = \frac{n_{vivos,t} - n_{mortos,t} - n_{censurados,t}}{n_{vivos,t} - n_{censurados,t}}$$

tempo t	censurados no tempo t	mortos no tempo t	vivos no tempo t	proporção de pacientes sobreviventes no instante t	função de sobrevivência
0	0	0	30	1.000000	1.000000
1	1	2	27	0.923077	0.923077
3	0	3	24	0.875000	0.807692
5	2	2	20	0.888889	0.717949
7	0	2	18	0.888889	0.638177
10	0	2	16	0.875000	0.558405
15	0	3	13	0.769231	0.429542
20	1	4	8	0.428571	0.184089
22	0	3	5	0.400000	0.073636

E no plot de KM, nós identificamos os dados censurados com uma cruz +:



Função de Hazard

Outro conceito muito importante na análise de sobrevivência é o de Hazard. Bem direto ao ponto, podemos dizer que o **Hazard é o risco de morte (ou de haver o evento de interesse) em um dado tempo.** A partir desse conceito, podemos definir **a função de Hazard h(t) (ou de risco), que descreve a mudança desse risco em função do tempo.** Por exemplo, se estamos observando o risco de morte de um paciente após contrair COVID-19, é evidente que esse risco (Hazard) muda com o passar do tempo: há um pico após alguns dias do início do quadro, mas o risco de morte decai

conforme o passar do tempo, com o corpo produzindo anticorpos e vencendo a infecção.

Um tipo de análise muito comum (e útil) em análise de sobrevivência é a de comparar as funções de Hazard de grupos diferentes (pacientes diabéticos e não diabéticos, por exemplo). Ao dividir uma função de Hazard pela outra, obtemos a **razão de hazard (ou de risco)** que pode ou não ser constante ao longo do tempo. Por exemplo: o risco de morte logo após uma dada cirurgia é 2 vezes maior para pacientes diabéticos em relação aos não diabéticos, entretanto, depois de alguns meses, essa razão muda para 1.5. Nesse caso, dizemos que os hazards **não** são proporcionais. Caso eles fossem, independentemente das formas das duas curvas de Hazard/risco, uma seria apenas um múltiplo da outra.

Esse conceito de Hazards é importante para um tipo de análise de sobrevivência chamado modelo de riscos proporcionais de Cox ou regressão de Cox ou, apenas, modelo de Cox, que deve ficar para um Turing Talks futuro.

Teste de log-rank

O teste de log-rank é um teste de hipótese para comparar as distribuições de sobrevivência de duas amostras. Se você não sabe o que é *um teste de hipótese/p-valores*, sugiro que dê uma olhada nos materiais do workshop de estatística com R do Grupo Turing. Esse teste é utilizado para comparar as curvas de Kaplan-Meier de dois grupos e utilizamos seu p-valor para determinar se essas curvas podem ou não ser consideradas diferentes com significância estatística.

A estatística de teste de log-rank compara as estimativas das funções de Hazard (ou de risco) dos dois grupos em cada tempo de evento observado. De modo que a hipótese nula do teste é de que as funções de hazard dos dois grupos comparados é igual:

$$H_0 : h_1(t) = h_2(t)$$
$$H_1 : h_1(t) \neq h_2(t)$$

No caso do estimador de Kaplan-Meier, estar "sob risco" significa que aquele indivíduo observado ainda não morreu nem sofreu censura.

Para computar o teste, começamos calculando uma tabela de contingência para cada instante de tempo t_j. Essa tabela se constitui de três colunas (grupo 1, grupo 2

e total) computando o número de "pacientes sob risco" e pacientes "mortos" para cada uma das colunas.

	Grupo 1	Grupo 2	Total
Evento ocorre	$n_{G_1,falha,t_j}$	$n_{G_2,falha,t_j}$	n_{falha,t_j}
Evento não ocorre	$n_{G_1,risco,t_j}$	$n_{G_2,risco,t_j}$	n_{risco,t_j}

Daí, para esse instante j, calculamos um chamado w_2j para o grupo 2 da seguinte maneira:

$$w_{2j} = \frac{n_{G_2,\text{risco},j} \cdot n_{G_2,\text{falha},j}}{n_{\text{risco},j}}$$

Também calculamos a variância do número de falhas do grupo 2 para cada instante de tempo j, com a fórmula:

$$(V_j)_2 = \frac{n_{G_2,\text{risco},j} \cdot (n_{\text{risco},j} - n_{G_2,\text{risco},j}) \cdot n_{G_2,\text{falha},j} \cdot (n_{\text{risco},j} - n_{G_2,\text{falha},j})}{n_{\text{risco},j}^2 \cdot (n_{\text{risco},j} - 1)}$$

Calculamos, então, uma estatística T.

$$T = \frac{[\sum_{j=1}^{k} (n_{G_2,\text{falha},j} - w_{2j})]^2}{\sum_{j=1}^{k} (V_j)_2}$$

O p-valor do teste pode ser encontrado ao verificar o valor de T na distribuição chi-quadrado com 1 grau de liberdade.

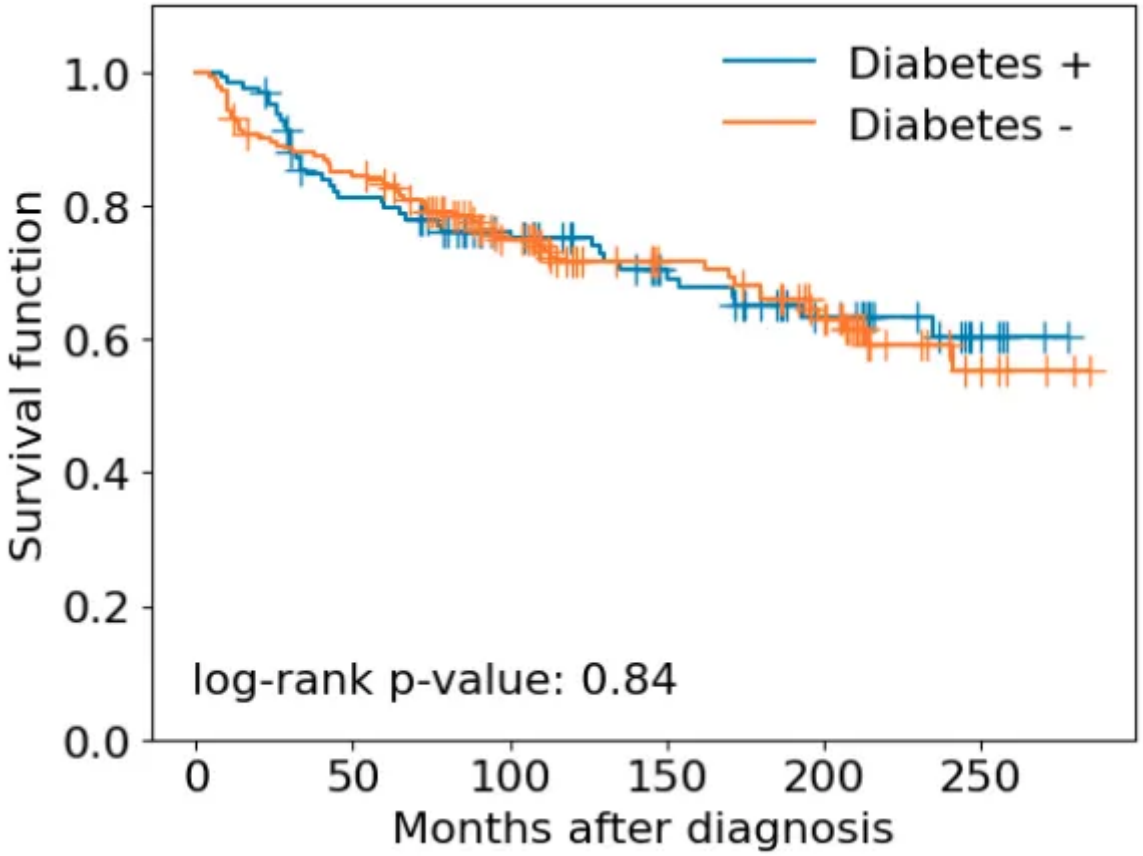
Fazendo uma aplicação em R e no Python

Para a nossa aplicação, vamos utilizar um dataset bem famoso do Kaggle: o Heart Failure Prediction.

Esse dataset contém dados acerca do tempo de sobrevivência de indivíduos com insuficiência cardíaca associado a outras 12 features.

Nesse caso, vamos estudar apenas a presença ou não de diabetes.

Python



Vemos aqui que não há diferença com significância entre a sobrevida depois de diagnóstico de insuficiência cardíaca de pacientes com e sem diabetes.

R