

Working Paper n° 40

Modelos de Sobrevivência

Cristina Rocha

Working Paper n° 40

ISSN: 0872 - 895X

Depósito Legal n°: 90631/95

Cristina Rocha

Departamento de Estatística e Investigação Operacional, Faculdade de Ciências,
Universidade de Lisboa

Novembro, 1995

Trabalho apresentado no âmbito do Seminário de Estatística e de Econometria -
coordenado pelo Professor Doutor Bento Murteira.

MODELOS DE SOBREVIVÊNCIA

CRISTINA SIMÕES ROCHA

DEIO e CEAUL
Faculdade de Ciências
Universidade de Lisboa

Resumo:

A Análise de Sobrevivência engloba um conjunto de modelos e métodos destinados à análise estatística de dados que representam o tempo de vida para um grupo de indivíduos ou, mais geralmente, o tempo decorrido desde um instante inicial até à ocorrência de um determinado acontecimento. Portanto, tempo de vida ou tempo de sobrevivência pode ser usado simplesmente para designar uma variável aleatória não negativa. No entanto, o que distingue fundamentalmente a Análise de Sobrevivência de outras áreas da Estatística é a possível existência de dados censurados, correspondentes a indivíduos para os quais não foi observado todo o seu tempo de vida.

Pretende-se neste trabalho apresentar uma introdução à análise estatística de dados de sobrevivência fazendo especial referência aos modelos probabilísticos de tempo de vida.

Palavras chave: censura; covariáveis; função de sobrevivência; função hazard; modelo de regressão de Cox; tempo de vida.

Abstract:

Survival Analysis is concerned with statistical models and methods for analysing data representing lifetimes for a group of individuals or, more generally, times from a well-defined starting point until the occurrence of some particular event. Thus, lifetime or survival time can be used merely to designate a non negative random variable. However, what distinguishes Survival Analysis from other fields of statistics is the possible presence of censored observations, corresponding to individuals whose exact lifetimes are not known.

The aim of this working paper is to present an introduction to the statistical analysis of survival time data, with special emphasis on probabilistic lifetime models.

Key words: censoring; covariates; Cox regression model; hazard function; lifetime; survivor function.

1. Introdução

A Análise de Sobrevivência é uma área da Estatística onde se tem verificado, ao longo das duas últimas décadas, um acentuado desenvolvimento. Engloba um conjunto de métodos e modelos destinados à análise estatística de dados de sobrevivência. Este tipo de dados surge quando, para um determinado grupo de indivíduos, é registado o tempo decorrido desde um instante inicial bem definido até à ocorrência de um acontecimento de interesse. Esse acontecimento pode ser:

- morte real de indivíduos ou de animais em estudos clínicos
- uma recaída depois de ter sido atingido um estado de remissão (i.e., ausência de sintomas clínicos) em indivíduos que sofrem de doença incurável
- falha de componentes mecânicas ou electrónicas
- realização de uma determinada tarefa numa experiência de aprendizagem em psicologia
- fim de um período de greve ou de desemprego
- mudança de residência num estudo demográfico

Embora a Análise de Sobrevivência tenha aplicação em áreas tão diversas como economia, física, engenharia, psicologia e demografia, foi a necessidade de obtenção de métodos estatísticos que permitissem a abordagem de problemas na área das ciências biomédicas que motivou, em última análise, o seu desenvolvimento e influenciou de modo determinante a terminologia utilizada. De facto, embora o acontecimento de interesse possa assumir formas tão diversas como as referidas acima, o tempo decorrido desde o instante inicial até à sua realização é habitualmente designado por tempo de vida ou de sobrevivência. Do mesmo modo, *morte* significa ocorrência do acontecimento.

Assim, tempo de vida ou tempo de sobrevivência pode ser usado simplesmente para designar uma variável aleatória não negativa. No entanto, o que distingue a Análise de Sobrevivência de outras áreas da Estatística é a existência de censura, devida a certas restrições na recolha dos dados que impedem que a realização do acontecimento de interesse seja observada para todos os indivíduos em estudo. Frequentemente, são também registados para

cada indivíduo em estudo os valores de certas variáveis, designadas por explanatórias ou covariáveis, que representam factores que se supõe afectarem o tempo de sobrevivência.

Em resumo, a metodologia usada na Análise de Sobrevivência aplica-se a todas as situações em que, para o estudo de uma variável aleatória não negativa, os métodos estatísticos clássicos não se revelam adequados, devido à especificidade de certas questões como, por exemplo, a existência de observações censuradas.

2. Exemplos Práticos

Nesta secção, as características fundamentais dos dados de sobrevivência serão ilustradas através de alguns exemplos práticos, que se encontram referidos em Lawless (1982).

Exemplo 1. Tempo de vida de pacientes com mieloma múltiplo

O mieloma múltiplo é um tumor maligno da medula óssea. Os dados da tabela 1 representam o tempo (em meses) decorrido desde o diagnóstico até à morte devida a mieloma múltiplo para 65 indivíduos e incluem também os valores, registados na altura do diagnóstico, das seguintes variáveis:

- x_1 logaritmo do ácido úrico
- x_2 hemoglobina
- x_3 idade
- x_4 sexo: 0, masculino; 1, feminino
- x_5 cálcio (no soro sanguíneo)

Pretendeu-se identificar, de entre os factores acima mencionados, os que influenciam de forma significativa o tempo de vida dos pacientes. As observações censuradas estão marcadas com * e correspondem ao tempo de vida dos indivíduos que ainda estavam vivos na data em que o estudo terminou.

Tabela 1

t	x_1	x_2	x_3	x_4	x_5	t	x_1	x_2	x_3	x_4	x_5
1	2.218	9.4	67	0	10	26	1.230	11.2	49	1	11
1	1.940	12.0	38	0	18	32	1.322	10.6	46	0	9
2	1.519	9.8	81	0	15	35	1.114	7.0	48	0	10
2	1.748	11.3	75	0	12	37	1.602	11.0	63	0	9
2	1.301	5.1	57	0	9	41	1.000	10.2	69	0	10
3	1.544	6.7	46	1	10	42	1.146	5.0	70	1	9
5	2.236	10.1	50	1	9	51	1.568	7.7	74	0	13
5	1.681	6.5	74	0	9	52	1.000	10.1	60	1	10
6	1.362	9.0	77	0	8	54	1.255	9.0	49	0	10
6	2.114	10.2	70	1	8	58	1.204	12.1	42	1	10
6	1.114	9.7	60	0	10	66	1.447	6.6	59	0	9
6	1.415	10.4	67	1	8	67	1.322	12.8	52	0	10
7	1.978	9.5	48	0	10	88	1.176	10.6	47	1	9
7	1.041	5.1	61	1	10	89	1.322	14.0	63	0	9
7	1.176	11.4	53	1	13	92	1.431	11.0	58	1	11
9	1.724	8.2	55	0	12	4*	1.945	10.2	59	0	10
11	1.114	14.0	61	0	10	4*	1.924	10.0	49	1	13
11	1.230	12.0	43	0	9	7*	1.114	12.4	48	1	10
11	1.301	13.2	65	0	10	7*	1.532	10.2	81	0	11
11	1.508	7.5	70	0	12	8*	1.079	9.9	57	1	8
11	1.079	9.6	51	1	9	12*	1.146	11.6	46	1	7
13	0.778	5.5	60	1	10	11*	1.613	14.0	60	0	9
14	1.398	14.6	66	0	10	12*	1.398	8.8	66	1	9
15	1.602	10.6	70	0	11	13*	1.663	4.9	71	1	9
16	1.342	9.0	48	0	10	16*	1.146	13.0	55	0	9
16	1.322	8.8	62	1	10	19*	1.322	13.0	59	1	10
17	1.230	10.0	53	0	9	19*	1.322	10.8	69	1	10
17	1.591	11.2	68	0	10	28*	1.230	7.3	82	1	9
18	1.447	7.5	65	1	8	41*	1.756	12.8	72	0	9
19	1.079	14.4	51	0	15	53*	1.114	12.0	66	0	11
19	1.255	7.5	60	1	9	57*	1.255	12.5	66	0	11
24	1.301	14.6	56	1	9	77*	1.079	14.0	60	0	12
25	1.000	12.4	67	0	10						

Exemplo 2. Comparação de dois tratamentos para leucemia aguda

Numa experiência clínica, pretendeu-se investigar o efeito do medicamento 6-MP na manutenção do estado de remissão em doentes com leucemia aguda. Para tal, os pacientes foram distribuídos aleatoriamente por dois grupos: a um dos grupos é administrado o medicamento 6-MP, ao outro é dado um placebo (grupo de controlo). Na tabela seguinte estão representados os tempos de remissão (em semanas) para dois grupos de 21 pacientes. Este estudo clínico decorreu durante um ano e as observações censuradas correspondem aos indivíduos que se encontravam ainda em estado de remissão quando o estudo terminou.

6-MP	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Exemplo 3. Estudo de um isolante eléctrico

Amostras de um certo tipo de líquido electricamente isolante foram sujeitas a uma tensão constante e foi registado o tempo decorrido até o líquido perder a sua capacidade isolante. As amostras foram divididas em 3 grupos e testadas a 28 kV, 30 kV e 32 kV. Os resultados obtidos estão indicados na tabela seguinte. Neste caso, não existem observações censuradas, visto que a experiência foi realizada decorreu durante o tempo necessário para observar a *falha* de todas as amostras de líquido.

28 kV	30 kV		32 kV	
68.85	17.05	194.90	0.40	3.91
426.07	22.66	47.30	82.85	0.27
110.29	21.02	7.74	9.88	0.69
108.29	175.88		89.29	100.58
1067.6	139.07		215.10	27.80
	144.12		2.75	13.95
	20.46		0.79	53.24
	43.40		15.93	

3. Notação e Conceitos Básicos

Seja T uma variável aleatória (v.a.) não negativa que representa o tempo de vida de um indivíduo pertencente a uma dada população homogênea. Supomos, deste modo, que os indivíduos não diferem entre si relativamente a factores susceptíveis de influenciar a sua sobrevivência.

Define-se função de sobrevivência (f.s.) como sendo a probabilidade de um indivíduo sobreviver pelo menos até ao instante t e representa-se por

$$S(t) = P(T \geq t), \quad t \geq 0.$$

A f.s.¹ é uma função que goza das seguintes propriedades:

- monótona não crescente
- contínua à esquerda
- $S(0) = 1$ e $S(+\infty) = \lim_{t \rightarrow \infty} S(t) = 0$.

Seja T uma v.a. absolutamente contínua, com função densidade de probabilidade (f.d.p.)

$$\begin{aligned} f(t) &= \lim_{dt \rightarrow 0^+} \frac{P(t \leq T < t + dt)}{dt} \\ &= -S'(t). \end{aligned}$$

A distribuição de T pode também ser caracterizada pela função hazard, que no caso contínuo é definida por

$$h(t) = \lim_{dt \rightarrow 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt}.$$

Esta função, que é também designada por taxa de falha (específica da idade) ou força de mortalidade, é particularmente útil no contexto da Análise de Sobrevivência. Especifica a taxa instantânea de morte no instante t , condicional à sobrevivência do indivíduo até esse instante.

¹Notemos que, ao contrário do habitual, foi adoptada na definição de f.s. uma convenção que leva à continuidade à esquerda da função de distribuição de T , $F(t) = 1 - S(t)$. O objectivo foi o de simplificar algumas fórmulas que envolvem a função hazard. Esta convenção é seguida por Lawless (1982), Cox e Oakes (1984) e Collett (1994), entre outros.

Como consequência das definições anteriores, surgem as seguintes relações entre f.s., f.d.p. e função hazard:

$$\begin{aligned}h(t) &= f(t)/S(t), \\S(t) &= \exp\left(-\int_0^t h(u)du\right), \\f(t) &= h(t)\exp\left(-\int_0^t h(u)du\right).\end{aligned}$$

Notemos que a função hazard verifica as seguintes propriedades:

$$h(t) \geq 0, \quad \int_0^\infty h(t)dt = \infty$$

Define-se ainda a função hazard cumulativa como sendo

$$H(t) = \int_0^t h(u)du, \quad t \geq 0.$$

Logo, $H(t) = -\log S(t)$.

4. Considerações sobre a Função Hazard

A função hazard descreve a evolução ao longo do tempo da probabilidade instantânea de morte de um indivíduo. Representa, portanto, um aspecto da distribuição do tempo de vida que tem significado físico imediato.

Além disso, a informação de natureza qualitativa de que por vezes se dispõe sobre a função hazard pode ajudar-nos na selecção de um modelo para o tempo de vida.

As formas mais comuns que a função hazard apresenta são as seguintes:

- Monótona crescente: é característica de todas as situações em que os indivíduos são observados num período da sua vida durante o qual ocorre um envelhecimento gradual. Neste caso, a proporção de indivíduos que morrem num dado instante, de entre os sobreviventes nesse instante, aumenta com o tempo. Os modelos mais utilizados em Análise de Sobrevivência são os que apresentam este tipo de função hazard.

Ex: tempo decorrido entre a infecção com o vírus HIV e o desenvolvimento da SIDA.

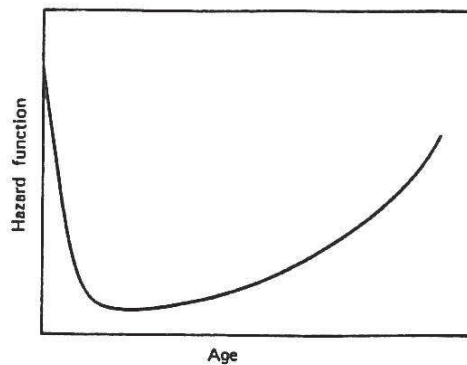
- Monótona decrescente: é menos comum, visto que reflecte uma situação na qual quanto mais tempo o indivíduo sobrevive, menor é a sua probabilidade de morte no instante subsequente.

Ex: tempo de vida de crianças que foram operadas para correcção de um defeito congénito, dado que o principal risco de morte é a própria operação ou complicações imediatas.

- Constante: caracteriza a distribuição do tempo de vida como exponencial.

Ex: tempo de vida de indivíduos cujos únicos riscos de morte são acidentes ou doenças raras; tempo decorrido desde o diagnóstico de cancro do pulmão não operável até à morte.

- Em forma de U (*bathtub-shaped*): ocorre em populações em que os indivíduos são seguidos desde o nascimento até à morte reais, sejam populações de seres vivos ou de objectos manufacturados.



5. Modelos de Sobrevivência Paramétricos

Vamos agora referir algumas das distribuições contínuas univariadas que, por apresentarem certas propriedades, se têm revelado mais úteis na análise de dados de sobrevivência. Notemos, no entanto, que qualquer distribuição com suporte não negativo pode ser usada como modelo para o tempo de vida (ver Lawless (1982), Cox e Oakes (1984), Collett (1994)).

5.1 Distribuição Exponencial

A distribuição é caracterizada por uma função hazard constante

$$h(t) = \lambda \quad \lambda > 0$$

A f.d.p. e a f.s. são dadas, respectivamente, por

$$f(t) = \lambda \exp(-\lambda t), \quad S(t) = \exp(-\lambda t),$$

para $t > 0$. Foi a primeira distribuição amplamente usada em estudos de fiabilidade e estudos clínicos.

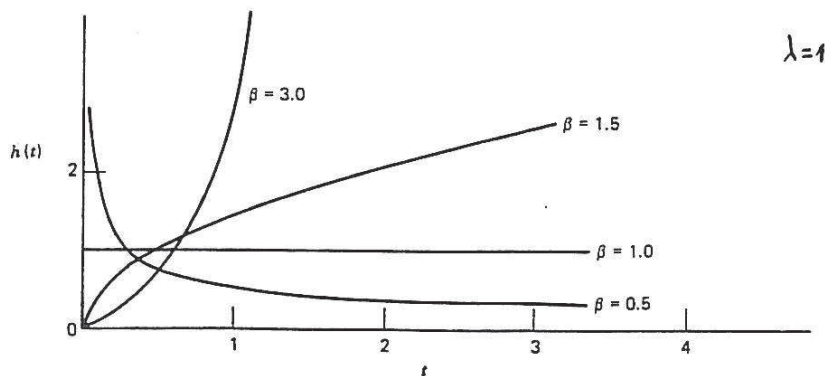
5.2 Distribuição de Weibull

A distribuição de Weibull com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\beta > 0$ tem, para $t > 0$,

$$\begin{aligned} S(t) &= \exp(-(\lambda t)^\beta) \\ h(t) &= \lambda \beta (\lambda t)^{\beta-1} \\ f(t) &= \lambda \beta (\lambda t)^{\beta-1} \exp(-(\lambda t)^\beta). \end{aligned}$$

A simplicidade destas expressões e a razoável flexibilidade do modelo são algumas das razões da sua grande utilização. A função hazard pode ser:

- monótona crescente se $\beta > 1$
- monótona decrescente se $0 < \beta < 1$
- constante se $\beta = 1$



A distribuição de Weibull é provavelmente a distribuição mais usada em Análise de Sobrevida, sendo adequada a um grande número de situações práticas, nomeadamente na área das ciências biomédicas.

5.3 Distribuição Gama

A distribuição gama com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\alpha > 0$ tem, para $t > 0$, f.d.p. da forma

$$f(t) = \frac{\lambda(\lambda t)^{\alpha-1} \exp(-\lambda t)}{\Gamma(\alpha)}.$$

A f.s. é dada por

$$S(t) = 1 - I(\alpha, \lambda t),$$

onde $I(\alpha, x)$ é a função gama incompleta

$$I(\alpha, x) = \frac{1}{\Gamma(\alpha)} \int_0^x u^{\alpha-1} e^{-u} du.$$

A função hazard pode ser:

- monótona crescente se $\alpha > 1$, com $\lim_{t \rightarrow 0+} h(t) = 0$ e $\lim_{t \rightarrow \infty} h(t) = \lambda$
- monótona decrescente se $0 < \alpha < 1$, com $\lim_{t \rightarrow 0+} h(t) = \infty$ e $\lim_{t \rightarrow \infty} h(t) = \lambda$
- constante ($h(t) = \lambda$) se $\alpha = 1$

Embora seja uma das distribuições contínuas positivas mais usadas em Estatística, a sua utilização em Análise de Sobrevida apresenta algumas dificuldades, devidas principalmente à forma da função hazard e da f.s.

5.4 Distribuição Log-Normal

O tempo de vida T tem distribuição log-normal se $\log T$ tem distribuição normal, digamos com valor médio μ e variância σ^2 . A f.d.p. de T é então

$$f(t) = \frac{1}{(2\pi)^{1/2} \sigma t} \exp \left[-\frac{1}{2} \left(\frac{\log t - \mu}{\sigma} \right)^2 \right] \quad t > 0.$$

A f.s. é dada por

$$S(t) = 1 - \Phi \left(\frac{\log t - \mu}{\sigma} \right)$$

onde $\Phi(\cdot)$ é a f.d. da $N(0, 1)$.

A função hazard $h(t)$ é tal que $\lim_{t \rightarrow 0^+} h(t) = 0$, é crescente até um valor máximo e depois decresce, com $\lim_{t \rightarrow \infty} h(t) = 0$.

O facto de $h(t)$ ser decrescente para grandes valores de t é o maior inconveniente da distribuição. No entanto, o modelo é frequentemente apropriado, particularmente quando tais valores não têm interesse (e.g., tempo de recuperação de operações cirúrgicas).

6. Covariáveis

O objectivo fundamental da Análise de Sobrevivência é a análise do tempo de vida de cada indivíduo. No entanto, é plausível admitir que a sua sobrevivência possa ser afectada por diversos factores, designados por factores de risco ou factores de prognóstico. Assim, a partir da década de 70, foi dada uma ênfase particular ao estudo da associação entre o tempo de vida e variáveis designadas por explanatórias ou covariáveis, representando potenciais factores de risco. Sempre que tal for possível, os valores individuais destas variáveis devem ser registados, visto que fornecem informação acerca da heterogeneidade existente na população.

Uma covariável diz-se constante se o seu valor permanece inalterado durante todo o período em que o indivíduo se encontra em observação. Como exemplo, podemos referir:

- uma variável indicatriz do grupo de tratamento a que o indivíduo pertence
- variáveis demográficas (e.g., sexo, idade na altura do diagnóstico)
- variáveis clínicas cujos valores sejam medidos uma única vez, no início do estudo

Uma covariável diz-se dependente do tempo quando o seu valor varia ao longo do período de observação. Como exemplo, podemos referir:

- um factor sob controlo do experimentador, que o faz variar ao longo do estudo de forma pré-determinada (e.g., dosagem de um medicamento)
- variáveis clínicas cujos valores sejam obtidos a intervalos regulares (e.g., pressão arterial)
- o desgaste sofrido por uma componente mecânica durante o seu funcionamento

7. Censura

Conforme já referimos, um aspecto importante a considerar na análise de dados de sobrevivência é a possibilidade de existência de dados censurados, que ocorrem quando, para alguns indivíduos em estudo, não é observada a realização do acontecimento de interesse durante o período em que esses indivíduos estão em observação.

Dispomos apenas de informação parcial sobre o seu tempo de vida, sabendo-se apenas que excede (*censura à direita*) ou é inferior (*censura à esquerda*) a determinado valor.

Censura à esquerda

Ex: numa experiência clínica em que se pretende analisar o tempo até à ocorrência de um tumor, as observações correspondentes aos indivíduos que apresentam metástases são censuradas à esquerda.

Censura à direita

Ex: em estudos de fiabilidade industrial, em que não é economicamente rentável prolongar o estudo até que seja observada a falha de todas as unidades em teste, a observação cessa geralmente quando se verifica um número pré-determinado de falhas. As unidades que ainda se encontram em funcionamento nesse instante dão origem a dados censurados à direita.

Ex: num estudo clínico, com data final pré-determinada, em que o acontecimento de interesse é a morte provocada por determinada doença, os indivíduos que

- sobrevivem até à data final ou
- morreram entretanto devido a alguma outra causa ou
- foram perdidos para o *follow-up*, i.e., estão indisponíveis para continuar em observação por razões alheias ao estudo

dão origem a dados censurados à direita.

A partir de agora, usaremos o termo *censura* para designar censura à direita, visto que corresponde à situação que se encontra com maior frequência na análise de dados de sobrevivência.

Dado que existem diversos motivos para o aparecimento de dados censurados, consideram-se vários tipos de censura, nomeadamente,

Censura de tipo I Obtem-se uma amostra censurada de tipo I quando os períodos de observação $c_1, \dots, c_n$ correspondentes aos n indivíduos em estudo são fixados previamente pelo investigador. O número de mortes observadas é uma v.a. Este tipo de censura surge, por exemplo, em estudos clínicos com data final pré-determinada.

Censura de tipo II São colocados em teste n indivíduos mas a observação cessa no instante de ocorrência da r -ésima morte, sendo r um número pré-determinado ($1 \leq r \leq n$). O tempo de duração do estudo é então uma v.a.

Por vezes, o mecanismo de censura a que estão sujeitos os indivíduos em estudo não é especificado. No entanto, para que os métodos estatísticos usados em Análise de Sobrevivência sejam válidos, é necessário exigir que os indivíduos que são censurados no instante t sejam representativos de todos os indivíduos, com os mesmos valores das covariáveis, que sobreviveram até t . Isto significa que um indivíduo não pode ser retirado do estudo porque parece apresentar um risco de morte invulgarmente elevado ou baixo.

Dizemos então que os indivíduos estão sujeitos a um mecanismo de **censura independente** ou que a censura é não informativa. Este tipo de censura engloba os casos de censura de tipo I e II e abrange a maior parte das situações com que deparamos na prática.

8. A Função de Verosimilhança num Modelo Paramétrico

Os métodos de inferência estatística utilizados em Análise de Sobrevivência são, de um modo geral, baseados na teoria assintótica da máxima verosimilhança, visto que, na maior parte dos casos, a existência de censura torna extremamente difícil a obtenção de distribuições de amostragem exactas. Vamos indicar qual a forma geral da função de verosimilhança num modelo paramétrico, estando os indivíduos sujeitos a um mecanismo de censura independente. Lawless (1982) descreve detalhadamente a sua construção.

Seja $(t_1, \dots, t_n)$ uma amostra aleatória de tempos de vida, possivelmente censurados, correspondentes aos n indivíduos em estudo. Suponhamos que a distribuição do tempo de vida T é conhecida a menos de um vector de parâmetros θ , sobre o qual desejamos realizar inferência, sendo T uma v.a. contínua com f.s. $S(t; \theta)$ e f.d.p. $f(t; \theta)$.

Se a amostra $(t_1, \dots, t_n)$ for completa, i.e., se não existirem observações censuradas, a função de verosimilhança é dada por

$$L(\theta) = \prod_{i=1}^n f(t_i; \theta).$$

Consideremos agora a situação mais usual em que a amostra $(t_1, \dots, t_n)$ contém observações censuradas. Seja então D o conjunto de indivíduos cujos tempos de vida foram observados e C o conjunto de indivíduos que deram origem a dados censurados. Neste caso, a função de verosimilhança é dada por

$$L(\theta) = \prod_{i \in D} f(t_i; \theta) \prod_{i \in C} S(t_i; \theta)$$

Notemos que a contribuição para a verosimilhança de um indivíduo censurado no instante t_i é apenas $S(t_i; \theta)$.

Embora haja ainda algumas questões por resolver nesta área, os resultados assintóticos usuais da teoria da máxima verosimilhança continuam válidos, sob condições de regularidade bastante gerais nos processos de morte e censura.

9. Formulação de um Modelo de Regressão

Para modelar o tempo de vida numa população homogénea, podemos utilizar alguma das distribuições contínuas univariadas que referimos anteriormente.

No entanto, a existência de heterogeneidade entre os indivíduos no que diz respeito a possíveis factores de risco é a situação mais frequente. Uma forma adequada de incorporar, na análise estatística, esses factores que supomos afectarem o tempo de vida, consiste na utilização de um modelo de regressão, em que o tempo de vida é a variável dependente e as covariáveis actuam como variáveis independentes.

É necessário então especificar um modelo para a distribuição do tempo de vida T dado o vector $\mathbf{z} = (z_1, \dots, z_p)'$ de covariáveis associado a determinado indivíduo. Os modelos de regressão paramétricos mais importantes são baseados nas distribuições referidas na secção 5, considerando um dos parâmetros da distribuição como função do vector de covariáveis.

Numa perspectiva que podemos designar por semiparamétrica, Cox (1972) propôs um modelo que rapidamente se tornou no modelo de regressão mais utilizado na análise de tempos de vida, devido à sua flexibilidade e versatilidade. Este modelo será referido com maior detalhe na secção seguinte.

Uma classe de modelos de regressão que engloba grande parte dos modelos mais utilizados em Análise de Sobrevida é a classe dos

Modelos com funções hazard proporcionais

Este tipo de modelos de regressão é caracterizado pela proporcionalidade entre as funções hazard correspondentes a indivíduos com diferentes valores das covariáveis.

Logo, para dois indivíduos com vectores de covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, a razão das funções hazard $h(t; \mathbf{z}_1)/h(t; \mathbf{z}_2)$ não depende de t .

10. Modelo de Cox

Um dos aspectos inovadores do modelo de regressão de Cox (1972) reside no facto de ser formulado com base na relação entre a função hazard e as covariáveis. Com efeito, seja T uma v.a. contínua que representa o tempo de vida de um indivíduo e $\mathbf{z} = (z_1, \dots, z_p)'$ o vector de covariáveis associado a esse indivíduo.

Cox (1972) propôs um modelo em que a função hazard no instante t é da forma

$$h(t; \mathbf{z}) = h_0(t) \exp(\beta_1 z_1 + \dots + \beta_p z_p)$$

em que $\beta_1, \dots, \beta_p$ são os coeficientes de regressão (desconhecidos) que representam o efeito das covariáveis na sobrevivência e $h_0(t)$ é uma função arbitrária não negativa, que representa a função hazard para um indivíduo a que está associado o vector $\mathbf{z} = \mathbf{0}$.

Trata-se de um modelo de hazards proporcionais:

- as funções hazard correspondentes a dois indivíduos com covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$ são proporcionais. De facto, sendo $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$,

$$\frac{h(t; \mathbf{z}_1)}{h(t; \mathbf{z}_2)} = \exp\{\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)\}$$

não depende de t .

- as covariáveis têm um efeito multiplicativo na função hazard, de acordo com o factor $\exp(\boldsymbol{\beta}'\mathbf{z})$, que é designado por risco relativo.

Notemos que, embora o modelo seja baseado na hipótese de hazards proporcionais, não se admite nenhuma forma particular para a distribuição dos tempos de vida. Por isso, o modelo de Cox é considerado um modelo de regressão semiparamétrico.

Um modelo para comparação de dois grupos

Numa experiência clínica, em que os indivíduos em estudo foram distribuídos aleatoriamente por dois grupos de tratamento, pretende-se comparar as respectivas curvas de sobrevivência, utilizando para tal o modelo de Cox.

Seja então z uma covariável binária definida por

$$z = \begin{cases} 0 & \text{se o indivíduo pertence ao grupo 1} \\ 1 & \text{se o indivíduo pertence ao grupo 2} \end{cases}$$

A função hazard para um indivíduo a que está associado a covariável z é dada por

$$h(t; z) = h_0(t) \exp(\beta z)$$

Logo,

- se o indivíduo pertence ao grupo 1

$$h(t; z = 0) = h_0(t)$$

- se o indivíduo pertence ao grupo 2

$$h(t; z = 1) = h_0(t)e^\beta$$

Conclui-se então que:

Se $\beta < 0$, os pacientes do grupo 2 têm melhor prognóstico do que os do grupo 1.

Se $\beta > 0$, os pacientes do grupo 1 têm melhor prognóstico do que os do grupo 2.

BIBLIOGRAFIA

- Collett, D. (1994). *Modelling Survival Data in Medical Research*. London: Chapman and Hall.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B* **34**, 187-220.
- Cox, D.R. e Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.
- Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.