FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

XVII Congresso da Sociedade Portuguesa de Estatística

30 Setembro - 3 Outubro 2009
Sesimbra

Análise de Sobrevivência

Cristina Rocha
Ana Luísa Papoila

XVII Congresso
da Sociedade Portuguesa
de Estatística

30 Setembro - 3 Outubro 2009
Sesimbra

Análise de Sobrevivência

Cristina Rocha Ana Luísa Papoila

Ficha Técnica:

Título: Análise de Sobrevivência
Autores: Cristina Rocha e Ana Luísa Papoila
Editora: Sociedade Portuguesa de Estatística
Capa: Sónia Mariano, Gabinete de Design - FCT/UNL
Impressão: Instituto Nacional de Estatística
Tiragem: 500 exemplares
ISBN: 978-972-8890-22-3
Depósito Legal: nº 297954/09

Prefácio

A Análise de Sobrevivência engloba um conjunto de métodos e modelos destinados à análise estatística de dados de sobrevivência. Embora este tipo de dados possa resultar da observação de tempos de vida no sentido literal do termo, a Análise de Sobrevivência tem actualmente o significado muito mais vasto de análise do tempo decorrido desde um instante inicial até à ocorrência de um acontecimento de interesse, o qual pode assumir formas tão diversas como recaída de uma doença, divórcio, conclusão de uma licenciatura ou compra de um produto após exposição a um anúncio. Um aspecto importante a considerar é a existência de dados censurados, que surgem 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.

O presente texto foi preparado para servir de base ao mini-curso integrado no XVII Congresso Anual da Sociedade Portuguesa de Estatística. Começamos por apresentar os conceitos fundamentais da Análise de Sobrevivência, bem como diversos métodos não paramétricos de inferência que constituem a base de uma correcta análise estatística. Uma especial atenção é dedicada ao modelo de regressão de Cox, dada a grande popularidade deste modelo que revolucionou a análise dos dados de sobrevivência. Abordamos também os mode-

los de regressão paramétricos que, embora menos utilizados, possuem interessantes potencialidades. Os capítulos 6 e 7 constituem apenas uma introdução a dois assuntos sobre os quais existe uma vasta literatura, dadas as limitações de tempo de um curso desta natureza. De facto, não temos a pretensão de, neste texto, descrever exaustivamente os métodos e modelos que têm vindo a ser desenvolvidos para dar resposta aos inúmeros problemas que surgem nos vários campos de aplicação da Análise de Sobrevivência.

Agradecemos à Comissão Organizadora do XVII Congresso Anual da Sociedade Portuguesa de Estatística o convite para a realização deste mini-curso e à Sociedade Portuguesa de Estatística a possibilidade de publicar este livro, com o qual esperamos contribuir para a divulgação desta interessantíssima (em nossa opinião!) área da Estatística.

Lisboa, Julho de 2009

Cristina Rocha e Ana Luisa Papoila

Índice

Capítulo 1

Conceitos fundamentais

1.1 Introdução

A Análise de Sobrevivência é uma área da Estatística onde se tem verificado, ao longo dos últimos 50 anos, 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.

Embora a Análise de Sobrevivência tenha aplicação em áreas tão diversas como economia, física, engenharia, sociologia, 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 morte, mudança de residência ou fim de um período de desemprego, 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.

Um aspecto importante a considerar na análise de dados de sobre-

vivência 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. Podemos então dizer que se dispõe apenas de informação parcial sobre o tempo de vida desses indivíduos, mas o seu período de tempo em observação pode e deve ser registado, sem o que haverá perda de informação. O que os métodos de Análise de Sobrevivência permitem é que esses períodos em observação sejam incluídos na amostra para análise estatística, juntamente com os tempos que são, de facto, valores observados do tempo de vida em estudo. Os dados podem também ser truncados, quando existem indivíduos que são excluídos devido a um processo de selecção inerente ao planeamento do estudo. A presença de censura e/ou truncatura exige, portanto, a utilização de métodos específicos, adequados a tais situações.

Frequentemente, são também registados para cada indivíduo os valores de certas variáveis, designadas por explanatórias ou covariáveis, que representam factores que se supõe afectarem o tempo de sobrevivência. Assim sendo, a análise de regressão é também nesta área uma ferramenta estatística extremamente útil, pelo que foram desenvolvidos modelos de regressão adequados às especificidades dos dados 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 que representa o tempo decorrido até à realização de um determinado acontecimento, os métodos estatísticos clássicos não se revelam adequados, por diversos motivos como, por exemplo, a existência de observações censuradas.

1.2 Alguns marcos históricos

A Análise de Sobrevivência é uma das áreas mais antigas da Estatística, remontando a sua origem ao início do desenvolvimento das Ciências Actuariais e da Demografia no século XVII. A primeira tabela de mortalidade foi apresentada por John Graunt em 1662, no seu estudo sobre a mortalidade em Londres. Motivado pela controvérsia sobre a vacina da varíola, Daniel Bernoulli publicou em 1766 um ensaio onde são lançadas as bases da teoria dos riscos competitivos. Tendo concluído que a força de mortalidade aumenta em progressão geométrica com a idade, Gompertz propôs em 1825 uma "lei da mortalidade humana" da forma bc^x. Esta lei simples provou ser um excelente modelo para diferentes populações e diferentes épocas. Nas primeiras décadas do século XX, foram desenvolvidos métodos de inferência estatística para alguns modelos paramétricos simples, nomeadamente para a distribuição exponencial. Durante alguns anos após o final da 2ª Guerra Mundial, a Análise de Sobrevivência continuou a ser dominada por abordagens clássicas, de tipo paramétrico, das quais resultaram importantes contribuições, como o modelo de cura proposto por Boag em 1949, que utilizou o método da máxima verosimilhança para estimar os parâmetros do modelo. A partir da década de 50, com o desenvolvimento da investigação médica e consequente necessidade de realização de ensaios clínicos, surgiu um tipo de dados de sobrevivência que levou ao desenvolvimento de novos métodos, dos quais o estimador não paramétrico proposto por Kaplan e Meier em 1958 se pode considerar pioneiro. Um avanço decisivo surgiu em 1972 com o trabalho em que Sir David Cox propôs um modelo de regressão que iria revolucionar a análise dos dados de sobrevivência. Desde então, têm sido desenvolvidas novas metodologias com o objectivo de dar resposta a uma gama vastíssima de problemas que surgem nos vários campos onde a Análise de Sobrevivência tem vindo a encontrar apli-

cação, o que faz com que esta seja uma área da Estatística em franco desenvolvimento.

1.3 Notação e conceitos básicos

Seja T uma variável aleatória (v.a.) não negativa, absolutamente contínua, que representa o tempo de vida de um indivíduo pertencente a uma dada população homogénea. Supomos, deste modo, que os indivíduos não diferem entre si relativamente a factores susceptíveis de influenciar a sua sobrevivência. Sendo assim, não iremos, por agora, introduzir quaisquer variáveis explanatórias nas definições que se seguem.

Define-se função de sobrevivência como sendo a probabilidade de um indivíduo sobreviver para além do instante t e representa-se por

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

A função de sobrevivência é uma função que goza das seguintes propriedades:

- monótona decrescente e contínua
- $S(0) = 1$ e $S(+\infty) = \lim_{t\to\infty} S(t) = 0$.

A função densidade de probabilidade é dada por

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

A distribuição de T pode também ser caracterizada pela função de risco (*hazard function*), que no caso contínuo é definida por

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

Esta função, que é também designada por função intensidade, taxa de falha (específica da idade) ou força de mortalidade, é particularmente útil no contexto da Análise de Sobrevivência, visto que 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 directo. Frequentemente, dispomos de informação de natureza qualitativa sobre a função de risco, o que pode ajudar na selecção de uma família de modelos para T. Por exemplo, se os indivíduos são observados num período da sua vida durante o qual ocorre um envelhecimento gradual, serão adequados modelos que apresentem função de risco monótona crescente.

Notemos que a função de risco verifica as seguintes propriedades:

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

Como consequência das definições anteriores, surgem as seguintes relações entre função de sobrevivência, função densidade de probabilidade e função de risco:

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

Define-se ainda a função de risco cumulativa como sendo

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

Logo, $H(t) = -\log S(t) \Leftrightarrow S(t) = \exp[-H(t)]$. $H(t)$ é uma função não negativa e é monótona crescente.

1.4 Considerações sobre a função de risco

A função de sobrevivência é sempre uma função decrescente, sendo por esse motivo muito pouco informativa sobre a evolução do risco de morte ao longo do tempo. Por outro lado, a função de risco é, de facto, uma taxa instantânea de morte (ou falha) e é extremamente útil visto que uma determinada forma desta função irá corresponder ao modo como o risco de morte se altera ao longo do tempo.

As formas mais comuns que a função de risco 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 de risco. Como exemplo, podemos referir o tempo decorrido entre a infecção pelo VIH (Vírus da Imunodeficiência Humana) e o diagnóstico de SIDA (Síndrome da Imunodeficiência Adquirida);
- monótona decrescente: é menos comum, visto que reflecte uma situação na qual quanto mais tempo o indivíduo sobrevive, menor é a probabilidade de morte no instante subsequente; haverá um risco decrescente, por exemplo, no caso de crianças submetidas a uma intervenção cirúrgica para correcção de deficiências congénitas;
- constante: caracteriza univocamente a distribuição do tempo de vida como sendo exponencial. Esta situação surge, por exemplo, quando o tempo de vida representa o tempo até à ocorrência de acidentes ou doenças raras ou quando o estudo é realizado

durante um período de tempo suficientemente curto para se poder considerar que o risco de morte não se altera;

- *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. A função de risco é decrescente no início, constante durante um largo período de tempo e crescente no final da vida, devido ao envelhecimento da população;
- *hump-shaped* ou unimodal: neste caso o risco de morte é inicialmente crescente e passa a ser decrescente ao fim de algum tempo. Surge, por exemplo, em estudos de sobrevivência que envolvam doentes sujeitos a cirurgia, com sucesso, onde há um aumento inicial do risco de morte devido a complicações nas primeiras horas ou dias após a intervenção, seguido de uma diminuição do risco à medida que o paciente recupera.

Portanto, a função de risco pode ser monótona (crescente, decrescente ou constante) ou ser uma função não monótona e a todas estas formas corresponde sempre uma função de sobrevivência decrescente.

1.5 Variáveis explanatórias

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 de prognóstico, tais como tratamentos, propriedades intrínsecas do indivíduo ou variáveis exógenas. Assim, a partir da década de 70 do século XX, foi dada uma ênfase particular ao estudo da associação entre o tempo de vida e variáveis designadas por explanatórias ou covariáveis, representando os factores acima mencionados. Sempre que tal for possível, os valores

individuais destas variáveis devem ser registados, visto que fornecem informação acerca da heterogeneidade existente na população.

De uma maneira geral, as covariáveis podem ser classificadas como dependentes do tempo ou constantes.

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 como o sexo e o país de nascimento
- variáveis clínicas cujos valores sejam medidos uma única vez, geralmente 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 medidos a intervalos regulares (e.g., pressão arterial) ao longo do período de observação
- o desgaste sofrido por uma componente mecânica durante o seu funcionamento

Existe, no entanto, uma outra classificação mais detalhada no que diz respeito à dependência do tempo. Assim, as covariáveis podem ser classificadas em duas categorias principais: covariáveis externas ou internas.

- Uma covariável diz-se **externa** se não está directamente relacionada com o mecanismo que regula a morte dos indivíduos. Pode ainda ser classificada como:

  fixa quando o valor da covariável é medido no início do estudo e permanece constante durante todo o período em que o indivíduo se encontra em observação.

  definida se a sua trajectória, embora não sendo constante, é determinada *a priori* para cada indivíduo. É o caso de um factor sob controlo do experimentador, que o faz variar ao longo do estudo de forma pré-determinada.

  ancilária quando é o resultado de um processo que é exterior ao indivíduo. Por exemplo, uma covariável que represente o nível de poluição atmosférica num estudo da ocorrência de ataques de asma.

- Uma covariável diz-se **interna** se resulta de uma medição feita sobre um indivíduo ao longo do tempo. Logo, é observada apenas enquanto o indivíduo está vivo e não censurado, fornecendo deste modo informação sobre o seu tempo de vida. Numa experiência clínica, podemos referir como exemplo deste tipo de covariável uma determinada medida da condição geral do paciente, feita a intervalos regulares.

1.6 Formulação de um modelo de regressão

Quando pretendemos modelar o tempo de vida numa população homogénea, é habitual utilizar determinadas distribuições contínuas univariadas que, por apresentarem certas propriedades, são particularmente adequadas num grande número de situações práticas. É o caso das distribuições exponencial, gama, Weibull, Gompertz, log-normal e log-logística, entre outras, que iremos referir com algum

detalhe no capítulo 5.

No entanto, a existência de heterogeneidade entre os indivíduos no que diz respeito a possíveis factores de risco/prognóstico 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 (ou variável resposta) e as covariáveis actuam como variáveis independentes. É necessário então especificar um modelo para a distribuição do tempo de vida T dado o vector $\mathbf{z} = (z_1, \ldots, z_p)'$ de variáveis explanatórias associado a determinado indivíduo, o que pode ser feito usando alguma família paramétrica de distribuições ou recorrendo a uma abordagem semi-paramétrica.

De um modo geral, é conveniente definir o vector de covariáveis de modo que $\mathbf{z} = \mathbf{0}$ corresponda a algum conjunto de condições padrão. No que se segue, T é uma v.a. contínua e $\varphi(\mathbf{z})$ e $\psi(\mathbf{z})$ designam funções que relacionam o vector $\mathbf{z}$ com a sobrevivência do indivíduo do seguinte modo: uma função crescente corresponde sempre a um risco de morte crescente, ou seja, a tempo de vida decrescente.

A maior parte dos modelos de regressão utilizados em Análise de Sobrevivência pertencem às seguintes classes que passamos a caracterizar:

Modelos com funções de risco proporcionais

Este tipo de modelos de regressão é caracterizado pela proporcionalidade entre as funções de risco 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 de risco $h(t; \mathbf{z}_1)/h(t; \mathbf{z}_2)$ não depende de t. Então,

a função de risco de T, dado $\mathbf{z}$, pode ser escrita na forma

$$h(t;\mathbf{z}) = h_0(t)\varphi(\mathbf{z}),$$

em que $h_0(t)$ representa a função de risco para um indivíduo a que está associado o vector $\mathbf{z} = \mathbf{0}$, exigindo-se que $\varphi(\mathbf{0}) = 1$. A função $\varphi(\mathbf{z})$ pode ainda ser parametrizada como $\varphi(\mathbf{z};\boldsymbol{\beta})$, e.g., $\varphi(\mathbf{z};\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{z})$.

O factor de proporcionalidade $\varphi(\mathbf{z})$ é designado por risco relativo, e representa a razão entre o risco de morte de um indivíduo a que esteja associado o vector de covariáveis $\mathbf{z}$ e o risco de morte correspondente a um indivíduo para o qual $\mathbf{z} = \mathbf{0}$.

A função de sobrevivência de T dado $\mathbf{z}$ é da forma

$$S(t;\mathbf{z}) = [S_0(t)]^{\varphi(\mathbf{z})},$$

onde $S_0(t) = \exp\left(-\int_0^t h_0(u)du\right)$. A classe de modelos assim obtida é por vezes designada por família de Lehmann gerada por S.

Neste modelo, as covariáveis têm um efeito multiplicativo na função de risco. Admitindo uma forma particular para $h_0(t)$ obtemos modelos de regressão paramétricos de riscos proporcionais. Numa abordagem semi-paramétrica proposta por Cox (1972), a função $h_0(t)$ não é especificada.

Modelos de tempo de vida acelerado

São também designados por modelos de localização-escala para $\log T$ ou modelos log-lineares para T. Sendo $h_0(t)$ e $S_0(t)$ as funções de risco e de sobrevivência subjacentes, i.e., para um indivíduo a que está associado o vector $\mathbf{z} = \mathbf{0}$, a representação do modelo de tempo de vida acelerado em termos de variáveis aleatórias é dada por

$$T = \frac{T_0}{\psi(\mathbf{z})}$$

onde T_0 tem função de sobrevivência S e $\psi(\mathbf{z})$ é tal que $\psi(\mathbf{0}) = 1$. A função de sobrevivência e a função de risco para um indivíduo com covariáveis $\mathbf{z}$ são, respectivamente,

$$S(t;\mathbf{z}) = S_0(t\psi(\mathbf{z}))$$
$$h(t;\mathbf{z}) = h_0(t\psi(\mathbf{z}))\psi(\mathbf{z}).$$

Considerando a função ψ parametrizada por $\boldsymbol{\beta}$, o modelo mais usual é aquele em que $\psi(\mathbf{z};\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{z})$. As covariáveis têm um efeito multiplicativo em t, ou seja, a sua função é acelerar (ou travar) o tempo até à morte. Também nesta classe de modelos podemos considerar uma abordagem paramétrica ou semi-paramétrica, consoante a função $h_0(t)$ é ou não especificada.

Consideremos agora o modelo de tempo de vida acelerado na sua representação log-linear. Então o logaritmo do tempo de vida pode ser escrito como

$$\log T = \mu + \boldsymbol{\alpha}'\mathbf{z} + \sigma\varepsilon,$$

onde μ é o termo independente, $\boldsymbol{\alpha}$ é um vector de parâmetros de regressão, σ é um parâmetro de escala e ε é uma v.a. que representa o erro e cuja distribuição não depende de $\mathbf{z}$. Vejamos como justificar a designação deste modelo como modelo de tempo de vida acelerado. Consideremos a v.a. $T_0 = \exp(\mu + \sigma\varepsilon)$ cuja função de sobrevivência é $S_0(t) = P[\exp(\mu + \sigma\varepsilon) > t]$. Então

$$\begin{aligned} S(t;\mathbf{z}) &= P(T > t|\mathbf{z}) \\ &= P[\exp(\mu + \boldsymbol{\alpha}'\mathbf{z} + \sigma\varepsilon) > t] \\ &= P[\exp(\mu + \sigma\varepsilon) > t\exp(-\boldsymbol{\alpha}'\mathbf{z})] \\ &= S_0(t\exp(-\boldsymbol{\alpha}'\mathbf{z})) = S_0(t/\exp(\boldsymbol{\alpha}'\mathbf{z})). \end{aligned}$$

Portanto, o efeito das covariáveis consiste numa modificação da escala do tempo através do factor $\exp(-\boldsymbol{\alpha}'\mathbf{z})$, que é habitualmente

designado por factor de aceleração. De facto, o tempo de sobrevivência de um indivíduo a que está associado o vector de covariáveis $\mathbf{z}$ é $T = T_0/\exp(-\boldsymbol{\alpha}'\mathbf{z})$. Ora, notemos que:

- se $\exp(-\boldsymbol{\alpha}'\mathbf{z}) > 1 \Leftrightarrow \boldsymbol{\alpha}'\mathbf{z} < 0$, o tempo até à ocorrência do acontecimento de interesse é acelerado por efeito das covariáveis

- se $\exp(-\boldsymbol{\alpha}'\mathbf{z}) < 1 \Leftrightarrow \boldsymbol{\alpha}'\mathbf{z} > 0$, o tempo até à ocorrência do acontecimento de interesse é travado por efeito das covariáveis.

Como consequência de $S(t;\mathbf{z}) = S_0(t\exp(-\boldsymbol{\alpha}'\mathbf{z}))$, constatamos que a mediana do tempo de sobrevivência de um indivíduo com vector de covariáveis $\mathbf{z}$ é igual à mediana do tempo de sobrevivência do indivíduo padrão multiplicada pelo inverso do factor de aceleração. Trata-se de uma característica importante dos modelos de tempo de vida acelerado.

Modelos de possibilidades proporcionais

Define-se possibilidade (*odds*) de sobrevivência para além do instante t como sendo a razão

$$\frac{S(t)}{1-S(t)}.$$

Neste tipo de modelos, a possibilidade de um indivíduo com vector de covariáveis $\mathbf{z}$ sobreviver para além de um determinado instante t é dada pela expressão

$$\frac{S(t;\mathbf{z})}{1-S(t;\mathbf{z})} = e^{\eta}\frac{S_0(t)}{1-S_0(t)}, \tag{1.1}$$

com $\eta = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p$ em que z_j representa o valor da j-ésima covariável, $j = 1, ..., p$, e $S_0(t)$ é a função de sobrevivência subjacente. Neste modelo, as covariáveis têm um efeito multiplicativo

na possibilidade de um indivíduo sobreviver para além de um instante t. Se calcularmos o logaritmo de ambos os termos da expressão (1.1), constatamos que η é o logaritmo da razão entre a possibilidade de sobrevivência para além do instante t de um indivíduo com vector de covariáveis $\mathbf{z}$ e a correspondente possibilidade de um indivíduo padrão. Estamos pois perante um modelo linear para o logaritmo da razão de possibilidades.

Tal como no modelo de riscos proporcionais, a função de risco subjacente pode ser estimada não parametricamente. Para proceder ao ajustamento do modelo aos dados, é então necessário estimar o vector de parâmetros $\boldsymbol{\beta}$ e a função de sobrevivência subjacente (Bennett, 1983a). Se considerarmos uma distribuição específica para o tempo de vida, obtém-se uma versão paramétrica do modelo de possibilidades proporcionais. Para terminarmos esta breve descrição deste tipo de modelos, há que referir uma propriedade que os caracteriza. Assim, pelo facto de

$$\frac{h(t;\mathbf{z})}{h_0(t)} = \{1 + (e^{\eta} - 1)S_0(t)\}^{-1},$$

constata-se que, quando $t = 0$ tem-se que $\frac{h(t;\mathbf{z})}{h_0(t)} = e^{-\eta}$ e quando $t \to \infty$ a razão das funções de risco converge para a unidade. Esta característica do modelo (funções de risco convergentes) pode ser útil, por exemplo, em determinados estudos experimentais prospectivos com os quais se pretende comparar dois grupos, um de controlo (livre da doença) e outro sujeito a um tratamento activo e em que, sendo a possibilidade de cura da doença uma realidade, com o decorrer do tempo os padrões de sobrevivência dos dois grupos tornar-se-iam semelhantes. De um modo geral, sempre que o efeito de uma covariável na sobrevivência se dissipa com o tempo, este modelo é uma boa opção.

Collett (2003) refere que este modelo tem tido pouca utilização prática, visto que, frequentemente, os resultados obtidos ao ajustar este modelo são semelhantes aos obtidos utilizando um modelo de regressão de Cox que inclua uma covariável dependente do tempo, de modo a que as funções de risco sejam não proporcionais.

A distribuição log-logística (Bennett, 1983b) ocupa aqui um lugar de importância, já que o modelo de regressão baseado nesta distribuição é o único que pertence simultaneamente à classe de modelos de tempo de vida acelerado e à classe de modelos de possibilidades proporcionais. Dá-nos, assim, a possibilidade de interpretar os resultados da análise de regressão tanto em função do factor de aceleração como da razão das possibilidades de sobrevivência para além de um instante t, conforme for mais conveniente.

1.7 Censura

Como já referimos, os dados relativos a tempos de vida apresentam-se frequentemente censurados, i.e., existem indivíduos para os quais não foi possível observar o seu tempo de vida com exactidão, havendo apenas uma informação incompleta sobre esse tempo. Notemos que tal se deve a certas restrições na recolha dos dados que impedem que a realização do acontecimento de interesse seja observada para todos os indivíduos em estudo.

Censura à direita

Estamos perante censura à direita quando apenas se sabe que o tempo de vida excede um determinado valor, visto que a observação do indivíduo termina antes da ocorrência do acontecimento de interesse. Consideremos um estudo clínico em que o acontecimento de interesse

é a morte provocada por determinada doença. Os doentes que sobrevivem até à data final pré-determinada, aqueles que são perdidos para o *follow-up* e aqueles que morreram entretanto devido a alguma outra causa dão origem a dados censurados à direita.

Vamos agora descrever alguns tipos de censura à direita, designados por censura de tipo I, censura de tipo II e censura aleatória.

Consideremos uma amostra de n indivíduos e seja T_i a v.a. que representa o tempo de vida do i-ésimo indivíduo. Suponhamos que a cada indivíduo corresponde um período de observação c_i, que podemos designar por tempo de censura potencial, tal que a morte desse indivíduo só será observada se ocorrer durante esse período. Deste modo, as observações são da forma (t_i, δ_i), $i = 1, \ldots, n$, em que

$$t_i = \min(T_i, c_i)$$

e

$$\delta_i = \begin{cases} 1 & \text{se } T_i \leq c_i \\ 0 & \text{se } T_i > c_i. \end{cases}$$

Diz-se que existe censura de tipo I quando os períodos de observação $c_1, \ldots, c_n$ são fixados previamente pelo investigador. O número de mortes observadas é aleatório.

A censura de tipo II surge quando o estudo termina no instante em que é observada a r-ésima morte, sendo r um número pré-determinado $(1 \leq r \leq n)$. O tempo de duração do estudo é então uma variável aleatória.

Um tipo de censura mais geral é a chamada censura aleatória (simples). Em ensaios clínicos, acontece frequentemente que os indivíduos entram de facto de forma aleatória, de acordo com a data de diagnóstico. Se o estudo terminar numa data pré-fixada, então o tempo

decorrido desde que um indivíduo entra em estudo até ao final deste é aleatório.

Neste caso, a cada indivíduo está associado um tempo de vida T_i e um tempo de censura (potencial) C_i, sendo os tempos de censura e os tempos de vida variáveis aleatórias mutuamente independentes e em que T_i e C_i são também variáveis aleatórias independentes. Assim sendo, as observações consistirão nos pares de variáveis aleatórias (Y_i, δ_i), $i = 1, ..., n$, em que $Y_i = \min(T_i, C_i)$ e

$$\delta_i = \begin{cases} 1 & \text{se} \quad T_i \leq C_i \\ 0 & \text{se} \quad T_i > C_i \end{cases}$$

Sendo a censura à direita o tipo de censura mais comum, todos os métodos descritos nos capítulos seguintes destinam-se à análise de dados censurados à direita.

Censura à esquerda

O tempo de vida associado a um indivíduo é considerado censurado à esquerda se apenas se sabe que é menor do que um tempo C_e, que foi registado. Neste caso, pode acontecer que o acontecimento de interesse tenha ocorrido antes da pessoa entrar em observação. Um exemplo bastante elucidativo surge quando o tempo que se pretende estudar é a idade em que uma criança consegue realizar determinada tarefa. Como é possível que algumas crianças já tenham aprendido a realizar essa tarefa antes de entrarem no estudo, as observações correspondentes são censuradas à esquerda, visto que o valor registado será a idade da criança no início do estudo.

Numa situação em que se pretende medir o tempo até à ocorrência de um tumor em indivíduos expostos a uma substância alegadamente cancerígena, as observações correspondentes aos indivíduos que já

apresentam metástases quando são observados são censuradas à esquerda.

Nesta situação, para uma amostra de n indivíduos, as observações consistirão nos pares de variáveis aleatórias (U_i, ϵ_i), $i = 1, ..., n$, em que $U_i = \max(T_i, C_e)$ e

$$\epsilon_i = \begin{cases} 1 & \text{se} \quad U_i \geq C_e \\ 0 & \text{se} \quad U_i < C_e \end{cases}$$

Este tipo de censura é menos comum.

Censura intervalar

Quando não é possível observar o instante exacto em que ocorre o acontecimento de interesse mas, apenas sabemos ter ocorrido num certo intervalo aleatório de tempo, dizemos que estamos perante uma observação censurada num intervalo.

Existem dois tipos de censura intervalar: caso I e caso II. Vejamos quais as suas características.

Assim, se a única informação de que dispomos é saber se, em determinado instante de monitorização, o acontecimento de interesse já ocorreu ou ainda não, dizemos que se trata de censura intervalar-caso I e os dados designam-se por dados do estado actual (*current status data*). Este tipo de dados surge em várias áreas de investigação, tendo sido, no entanto, em demografia que apareceram os primeiros estudos envolvendo dados do estado actual (Diamond *et al.*, 1986; Diamond e McDonald, 1991). De facto, em muitas aplicações demográficas, a principal variável de interesse é a idade em que determinado acontecimento de interesse ocorre, como por exemplo, a idade de desmame, a idade da menarca, a idade do primeiro casamento ou até a idade da morte. Acontece que estes dados provêm habitualmente

de estudos retrospectivos e, inevitavelmente, a sua recolha está sujeita a enviesamentos, dos quais salientamos os provocados por erros de memória. Assim sendo, vários autores (Bergsten-Brucefors, 1976; Quandt, 1987) tentaram perceber quais as consequências destes erros e chegaram à conclusão que os dados que envolvem datas fornecidas pelos entrevistados não são fiáveis, sendo por isso aconselhável a utilização de dados do estado actual. Outra forma de recolha alternativa, mais precisa, seria através de estudos prospectivos; no entanto, os elevados custos que habitualmente acompanham a sua implementação torna-os impeditivos na maioria dos casos. Juntando a estes factos a real inexistência de dados completos nalgumas situações, fica assim justificado o desenvolvimento das metodologias que utilizam dados do estado actual.

Também em alguns estudos de laboratório, em que se pretenda estudar o tempo que decorre entre a exposição de animais a determinado agente carcinogéneo e a ocorrência de tumor (não letal), surgem dados do estado actual. De facto, temos um tempo de monitorização que representa o tempo ao fim do qual os animais são sacrificados ou o tempo até à morte natural, e temos também a informação sobre a existência ou não do tumor, obtida após a morte do animal.

Existem outras áreas em que podemos encontrar dados do estado actual como, por exemplo, a epidemiologia. De facto, estes dados, revelam-se importantes no estudo de algumas características de doenças infecto-contagiosas, sobretudo quando não se consegue observar com exactidão o instante em que ocorre a infecção. Dos dados que se identificam com esta situação, salientamos os provenientes de estudos que envolvem doentes infectados com o VIH ou em risco de o estarem. Jewell e Shiboski (1990) e Shiboski (1998) abordam o grave problema de contágio pelo VIH entre parceiros sexuais. Neste tipo de estudos, são recrutados casais em que um dos parceiros está infectado

com o VIH (por uma via alheia ao outro parceiro) e pretende-se estudar o tempo que decorre até que o outro também se infecte, admitindo que o único meio de este contrair o vírus é através do contacto com o parceiro infectado. Estes estudos representam um marco na história dos dados do estado actual e não podem deixar de ser referenciados sempre que se aborda o tema.

Ainda em epidemiologia, não poderíamos deixar de referir o trabalho desenvolvido por Keiding (1991) e Keiding *et al.* (1996). De facto, é importante conhecer a distribuição da idade em que se contrai determinada doença para a qual existem testes de diagnóstico que nos permitem detectá-la. Assim, ao submetermos determinada população a um teste de rastreio, a presença/ausência de doença em indivíduos com determinada idade origina dados do estado actual sobre a idade em que é contraída a doença. Neste contexto, Keiding (1991) descreve um estimador de máxima verosimilhança para a distribuição da idade de ocorrência de infecção por hepatite A. Keiding *et al.* (1996) estudam ainda a distribuição da idade em que se contrai a rubéola, com base numa amostra de indivíduos do sexo masculino (população não vacinada).

Quando se conhece o intervalo durante o qual se realizou o acontecimento de interesse, dizemos que se trata de censura intervalar-caso II. Neste caso, o acontecimento de interesse ocorreu entre dois instantes observados, ou seja, $T \in [T_e, T_d]$. Dado que estes intervalos são aleatórios e há frequentemente sobreposições, não é possível usar a metodologia usual para dados agrupados.

Este tipo de censura surge com frequência em estudos longitudinais em que há *follow-up* periódico como, por exemplo, se desejarmos estudar a distribuição do tempo que decorre entre o fim do tratamento a um determinado carcinoma em doentes oncológicos e o instante em

que ocorre uma recidiva da doença. As observações resultantes desde estudo serão censuradas num intervalo, uma vez que a recidiva só será detectada numa visita programada de *follow-up* ou numa visita antecipada, dado a existência de queixas. Em ambos os casos, o acontecimento de interesse ocorreu entre duas visitas consecutivas.

Censura dupla

A designação "censura dupla" é utilizada por diferentes autores em duas situações distintas.

Uma delas tem a ver com estudos de sobrevivência em que existem dados censurados à direita, outros à esquerda e as restantes observações são exactas. Este tipo de censura aparece, por exemplo, em estudos sobre o VIH, cujo objectivo é estimar a distribuição do tempo até à infecção pelo vírus. Assim, num grupo de indivíduos em risco, podemos encontrar aqueles que já estão infectados (observação censurada à esquerda), aqueles que chegaram ao fim do estudo sem se terem infectado (observação censurada à direita) e alguns para os quais se conhece o instante em que ocorreu a infecção (observações exactas). Um exemplo desta última situação é a infecção por transfusão de sangue contaminado.

No entanto, esta designação também se aplica quando tanto a origem como o tempo que decorre até à ocorrência do acontecimento de interesse são censurados. Mais uma vez encontramos um bom exemplo para ilustrar este tipo de censura em dados obtidos em estudos sobre o VIH e sobre SIDA. Assim, se pretendemos estimar a distribuição do tempo de latência (tempo que decorre entre a infecção por VIH e o diagnóstico de SIDA), é usual não conhecer, com exactidão, o instante em que ocorreu a infecção mas apenas o intervalo de tempo em que esta ocorreu. Por outro lado, é natural que, no instante em que o

estudo termina, alguns dos indivíduos em estudo ainda não tenham desenvolvido as patologias que os permitem classificar como doentes com SIDA.

Censura independente e não informativa

O mecanismo de censura a que estão sujeitos os indivíduos em estudo poderá não ser especificado. No entanto, designando por p_j a probabilidade de um indivíduo em risco (i.e., vivo e não censurado) imediatamente antes de t_j morrer nesse instante, exige-se que as probabilidades p_j sejam relevantes para toda a população em estudo; assim, se um indivíduo censurado num instante anterior sobrevivesse até t_j, a sua probabilidade de morte seria dada por p_j.

Neste sentido, vamos enunciar duas hipóteses que estão subjacentes aos métodos estatísticos habitualmente utilizados em Análise de Sobrevivência:

1. Dada a história completa do estudo até ao instante t, ou seja, toda a informação sobre mortes e censuras ocorridas até esse instante, assim como informação completa sobre os valores de todas as covariáveis, os mecanismos de morte que regem indivíduos diferentes actuam de modo independente no intervalo $[t, t + dt)$.
2. Para um indivíduo em risco imediatamente antes do instante t, a probabilidade condicional de morte em $[t, t + dt)$ dada a história completa do estudo até ao instante t coincide com a probabilidade condicional de morte em $[t, t + dt)$ dada a sua sobrevivência até ao instante t.

Supõe-se que, no intervalo $[t, t + dt)$, os instantes de morte precedem os instantes de censura.

A hipótese 2 pode ser interpretada como uma hipótese de independência condicional entre os mecanismos de morte e de censura. Um mecanismo de censura que satisfaça esta hipótese diz-se independente. A partir de agora, sempre que existam dados censurados, admitiremos que se encontram sujeitos a um mecanismo de censura independente, visto que tal hipótese é necessária para a validade dos métodos de análise de dados de sobrevivência que iremos descrever.

Em resumo, exige-se que, condicional aos valores das covariáveis, os indivíduos que são censurados no instante t sejam representativos de todos os indivíduos que sobreviveram até t. Isto significa que, em qualquer instante, os indivíduos não podem ser censurados selectivamente por apresentarem um risco de morte invulgarmente elevado ou baixo.

Portanto, para cada indivíduo a censura não deve ser preditiva de morte (não observada) futura. No caso da chamada censura administrativa, ou seja, da censura que ocorre na data em que o estudo termina para os indivíduos ainda vivos, esta hipótese é geralmente verdadeira; no entanto, quando os indivíduos são perdidos para o *follow-up*, podem surgir dúvidas sobre se a censura não estará relacionada com factores associados ao tempo de vida. Por este motivo, em qualquer estudo é fundamental empregar os meios de acompanhamento dos participantes que sejam necessários para evitar, ou pelo menos minimizar, as perdas para o *follow-up*.

É de salientar que a designação "censura não informativa" é usada frequentemente por diversos autores (e.g. Collett, 2003; Klein e Moeschberger, 1997) com o significado acima referido para "censura independente". No entanto, Andersen (2005) clarifica esta questão, salientando que o conceito de "censura independente" é probabilístico, enquanto que "censura não informativa" é um conceito estatístico que

significa que a distribuição do tempo de censura não depende do parâmetro θ que indexa a distribuição do tempo de vida e que é o parâmetro de interesse. Portanto, se a censura for informativa no sentido de Andersen (2005), os tempos de censura podem conter informação sobre θ, o que pode eventualmente levar a alguma perda de eficiência na realização de inferência.

1.8 Truncatura

Censura e truncatura são mecanismos que levam à existência de dados incompletos, como o são frequentemente os dados de sobrevivência. No entanto, são mecanismos de natureza diferente, embora haja por vezes alguma confusão entre os dois conceitos.

A truncatura surge quando, devido a um processo de selecção inerente ao planeamento do estudo, apenas são estudados os indivíduos a quem ocorreu determinado acontecimento. Dito de outra forma, a truncatura consiste numa condição que "oculta" certos indivíduos de modo que o investigador não se apercebe da sua existência.

Há duas formas de truncatura: truncatura à esquerda e truncatura à direita.

Truncatura à esquerda

Apenas são incluídos no estudo os indivíduos que satisfazem determinada condição que deve ser verificada antes da ocorrência do acontecimento de interesse (e.g. exposição a uma doença, ocorrência de um acontecimento intermédio tal como a recaída de leucemia antes da morte pela doença). Se Y é o instante de ocorrência de tal acontecimento e T é o tempo de vida, apenas os indivíduos para os quais $T \geq Y$ são observados, ou seja, o indivíduo é observado apenas se o

seu tempo de sobrevivência excede um determinado valor conhecido.

A situação mais usual em que ocorre truncatura à esquerda é quando o tempo em estudo é a idade de ocorrência de um determinado acontecimento, visto que geralmente os indivíduos não são observados desde o nascimento mas sim desde um instante posterior y_i, correspondente à sua entrada em estudo. Este caso insere-se no âmbito daquilo que, mais geralmente, se pode designar por entrada tardia ou adiada (*delayed entry*): os indivíduos não são seguidos a partir do instante inicial "natural" para o fenómeno em estudo mas sim a partir de um instante posterior y_i (conhecido), desde que tenham sobrevivido até y_i. Esta situação, em que o instante inicial não é o instante em que o indivíduo entra em estudo, ocorre frequentemente em estudos epidemiológicos.

Vejamos um exemplo em que se pretende comparar dois programas de tratamento para indivíduos toxicodependentes. Embora o período livre de drogas (tempo de sobrevivência) tenha sido definido como tendo início no instante em que o indivíduo entrou em tratamento, i.e., quando de forma aleatória lhe foi atribuído um dos programas e deixou de consumir drogas, os indivíduos foram seguidos apenas a partir do momento em que completaram o programa de tratamento. Neste caso há, portanto, um processo de selecção porque só os indivíduos que completaram com sucesso o seu programa são passíveis de ser incluídos na análise estatística. O tempo de sobrevivência será, no mínimo, a duração do tratamento, que não é a mesma para todos os indivíduos. O acontecimento intermédio é aqui completar o programa de tratamento.

Em estudos epidemiológicos é frequente a utilização da chamada coorte prevalente, constituída por indivíduos que já têm uma doença ou outra condição de saúde no momento em que são recrutados para o estudo, sendo então seguidos ao longo do tempo com o objectivo

de observar um determinado acontecimento de interesse (e.g. progressão da doença, recaída ou morte). Quando se conhece a duração do período de tempo em que o indivíduo esteve doente antes do recrutamento, i.e., antes do início do *follow-up*, é possível utilizar os métodos para análise de dados truncados à esquerda, visto que também aqui se pode considerar que se trata de entrada tardia.

Considere-se um estudo cujo objectivo é o de identificar os factores de prognóstico para a sobrevivência de pacientes com uma determinada doença. Suponha-se então que apenas são eligíveis para o estudo indivíduos com a doença que estejam vivos na data C. Então, apenas os indivíduos com tempo de sobrevivência $t_i \geq C - u_i$, onde u_i é a data de diagnóstico, têm a oportunidade de serem incluídos na coorte. De todos os indivíduos a quem foi diagnosticada a doença em u_i, apenas aqueles com sobrevivência mais longa são incluídos no estudo e esta selecção resulta numa sobrestimação das probabilidades de sobrevivência se não forem feitos os ajustamentos necessários. Teríamos portanto um viés de selecção, pois os indivíduos com tempos de vida mais curtos são excluídos selectivamente, fazendo com que o risco de ocorrência do acontecimento de interesse seja subestimado.

Truncatura à direita

O indivíduo é observado apenas se o acontecimento de interesse tiver ocorrido antes de uma data especificada, ou seja, se o seu tempo de sobrevivência for inferior a um determinado valor (conhecido). Tal acontece em estudos em que o critério para recrutamento dos indivíduos se baseia na ocorrência do acontecimento de interesse antes do final do que se considera o período de observação, sendo depois determinado retrospectivamente o respectivo tempo de sobrevivência. Ocorre tipicamente quando os dados provêm de um registo com informação sobre casos confirmados de uma doença. Este tipo de dados

é particularmente relevante em estudos sobre SIDA, bem como sobre outras doenças infecciosas.

Por exemplo, o primeiro estudo sobre o período de incubação ou tempo de latência da SIDA, publicado em 1986, envolveu apenas uma amostra de casos de SIDA, registados na base de dados do CDC (Center for Disease Control and Prevention in the United States), que tinham sido infectados pelo VIH devido a transfusão de sangue contaminado. Estes pacientes foram identificados e seleccionados para o estudo porque já lhes tinha sido diagnosticada SIDA. Outro exemplo é o de um estudo retrospectivo que envolveu a selecção de pacientes pediátricos com SIDA cuja única forma de contaminação conhecida era a transmissão do VIH por via materna (mãe-filho); assumiu-se portanto que a data de infecção era a data de nascimento. Neste caso, o perído de incubação é o tempo decorrido desde o nascimento até ao diagnóstico de SIDA. Estes são exemplos de estudos retrospectivos com dados truncados à direita, nos quais os indivíduos com tempos de sobrevivência mais longos são excluídos selectivamente, porque o acontecimento de interesse pode ainda não ter ocorrido na altura em que os indivíduos são seleccionados para inclusão no estudo.

1.9 Construção da função de verosimilhança

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.

Suponhamos então que a distribuição do tempo de vida T segue um determinado modelo paramétrico, indexado por um vector de parâmetros $\boldsymbol{\theta}$, sobre o qual pretendemos realizar inferência. Ao cons-

truir a função de verosimilhança teremos que considerar, como é habitual, a contribuição dada por cada observação, que será diferente consoante o tipo de censura, caso não se trate de uma observação exacta. Suponhamos então que desejamos construir a função de verosimilhança quando temos observações não censuradas e censuradas à direita. Seja T a v.a. que representa o tempo de vida e C a v.a. que representa o tempo de censura. Então o tempo observado t para um indivíduo é uma observação da v.a. $Y = \min\{T, C\}$. Consideremos agora a distribuição conjunta do par (Y, δ) onde, por simplicidade, consideramos que, por exemplo, $P\{T = t\}$ representa a função densidade de probabilidade de T. A contribuição de um indivíduo cuja morte foi observada ($\delta = 1$) é então dada por

$$\begin{aligned} P\{Y = t, \delta = 1\} &= P\{Y = t, T \leq C\} = P\{T = t, T \leq C\} \\ &= P\{T = t, t \leq C\} = P\{T = t\}P\{C \geq t\}, \end{aligned}$$

Seguindo o mesmo raciocínio, a contribuição de um indivíduo cujo tempo de vida é censurado ($\delta = 0$) será

$$\begin{aligned} P\{Y = t, \delta = 0\} &= P\{Y = t, T > C\} = P\{C = t, T > C\} \\ &= P\{C = t, T > t\} = P\{C = t\}P\{T > t\}. \end{aligned}$$

Ao observarmos as expressões anteriores, constatamos que elas são válidas se assumirmos a independência entre T e C, já referida anteriormente. Sendo assim, a contribuição de um indivíduo para o qual se tenha observado o par (t, δ), virá dada por

$$\left(P\{T = t\}P\{C \geq t\}\right)^{\delta} \left(P\{C = t\}P\{T > t\}\right)^{1-\delta}.$$

Se agora suposermos que T e C são variáveis aleatórias contínuas com funções densidade f e g e com funções de sobrevivência S e $1 - G$, respectivamente, poderemos reescrever a expressão anterior na

seguinte forma

$$\{f(t)[1-G(t)]\}^{\delta}\,[g(t)S(t)]^{1-\delta}.$$

Sendo assim, dada uma amostra de dimensão n, $(t_1,\delta_1),...,(t_n,\delta_n)$, a função de verosimilhança é igual a

$$\prod_{i=1}^{n}\left\{[f(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i}\right\}\prod_{i=1}^{n}\left\{[1-G(t_i)]^{\delta_i}[g(t_i)]^{1-\delta_i}\right\}. \tag{1.2}$$

Admitindo que a censura é não informativa, ou seja, que a distribuição do tempo de censura não depende do vector de parâmetros de interesse $\boldsymbol{\theta}$, podemos basear toda a inferência sobre este vector na verosimilhança, dada por

$$L=\prod_{i=1}^{n}[f(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i}, \tag{1.3}$$

o que é equivalente a

$$L=\prod_{i=1}^{n}[h(t_i)]^{\delta_i}S(t_i).$$

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. Assim, o estimador de máxima verosimilhança $\hat{\boldsymbol{\theta}}$ tem distribuição assintótica normal multivariada com valor médio $\boldsymbol{\theta}$ e matriz de covariância $I(\boldsymbol{\theta})^{-1}$, sendo $I(\boldsymbol{\theta})$ a matriz de informação de Fisher.

Caso a censura seja informativa, será necessário recorrer a métodos baseados em (1.2). No entanto, segundo Andersen (2005), a inferência baseada na verosimilhança (1.3), embora não tão eficiente quanto a inferência baseada em (1.2), ainda é válida e é até aconselhável a

sua utilização, uma vez que, quanto menor for o número de hipóteses sobre a distribuição do tempo de censura, tanto melhor.

Como referimos anteriormente, a contribuição para a função de verosimilhança dada por cada indivíduo é diferente consoante o mecanismo de censura a que está sujeito:

$f(t_i)$	para tempos de vida exactos
$S(t_i)$	para observações censuradas à direita
$1 - S(t_i)$	para observações censuradas à esquerda
$S(l_i) - S(r_i)$	para observações sujeitas a censura intervalar-caso II

Vamos agora indicar qual a forma geral da função de verosimilhança num modelo de regressão paramétrico. Quando os dados contêm apenas tempos de vida exactos ou tempos censurados, a função de verosimilhança é então dada pelo produto dos factores acima referidos:

$$L = \prod_{i \in D} f(t_i) \prod_{i \in C} S(t_i) \prod_{i \in L} (1 - S(t_i)) \prod_{i \in I} [S(l_i) - S(r_i)]$$

onde D é o conjunto de índices associados aos indivíduos com tempos de vida exactos, C é o conjunto de índices associados aos indivíduos com tempos censurados à direita, L é o conjunto de índices associados aos indivíduos com tempos censurados à esquerda e I é o conjunto de índices associados aos indivíduos sujeitos a censura intervalar.

Notemos que a função de verosimilhança para dados sujeitos a truncatura à esquerda no instante Y_i e censura à direita, havendo independência entre os tempos de vida e de truncatura, é dada por

$$L = \prod_{i=1}^{n} \left[\frac{f(t_i)}{S(Y_i)} \right]^{\delta_i} \left[\frac{S(t_i)}{S(Y_i)} \right]^{1-\delta_i}.$$

Capítulo 2

Estimação não paramétrica

2.1 Estimação não paramétrica da função de sobrevivência

2.1.1 Tabelas de mortalidade

A tabela de mortalidade é um dos métodos mais antigos de representação de dados relativos à sobrevivência de um grupo de indivíduos seguidos ao longo do tempo, designado por coorte.

Supondo que os indivíduos constituem uma amostra aleatória proveniente de uma dada população, a tabela de mortalidade permite estimar:

- a probabilidade condicional de morte num intervalo de tempo, dada a sobrevivência no início desse intervalo
- a probabilidade de sobrevivência para além de um dado intervalo.

Consideremos então uma coorte de n indivíduos provenientes da população em estudo. O intervalo $[0, \infty)$ é dividido em $k+1$ intervalos adjacentes e de amplitude fixa

$$I_j = [a_{j-1}, a_j), \qquad j = 1, \ldots, k+1,$$

com $a_0 = 0, a_k = L$ e $a_{k+1} = \infty$, onde L é um limite superior de observação. Os dados consistem no número de indivíduos vivos no início de cada intervalo e no número de indivíduos que morrem ou são censurados em cada intervalo. Seja então

n_j : número de indivíduos em risco no instante a_{j-1}
d_j : número de mortes observadas em I_j
m_j : número de observações censuradas em I_j.

Notemos que

$$\begin{aligned} n_1 &= n \\ n_j &= n_{j-1} - d_{j-1} - m_{j-1} \qquad j = 2, \ldots, k+1. \end{aligned}$$

Seja $S(t)$ a função de sobrevivência populacional. Então, para $j = 1, \ldots, k+1$,

$$\begin{aligned} P_j = &\quad P(\text{um indivíduo sobreviver para além de } I_j) = S(a_j) \\ q_j = &\quad P(\text{um indivíduo morrer em } I_j | \text{ sobreviveu para além de } I_{j-1}) \\ p_j = &\quad 1 - q_j = P_j / P_{j-1}, \end{aligned}$$

em que $P_0 = 1, \quad P_{k+1} = 0, \quad q_{k+1} = 1.$

Então,

$$P_j = p_1 p_2 \cdots p_j, \qquad j = 1, \ldots, k+1.$$

Para obter uma estimativa de P_j comecemos por estimar q_j, usando o estimador actuarial dado por

$$\hat{q}_j = \begin{cases} 1 & n_j = 0 \\ \frac{d_j}{n_j - m_j/2} & n_j > 0. \end{cases}$$

Notemos que $n'_j = n_j - m_j/2$ é o número ajustado de indivíduos em risco no intervalo I_j, assumindo que os instantes de censura se distribuem uniformemente nesse intervalo.

Calculando $\hat{p}_j = 1 - \hat{q}_j$, o estimador de P_j é dado por

$$\hat{P}_j = \hat{p}_1 \cdots \hat{p}_j, \qquad j = 1, \ldots, k+1,$$

sendo a estimativa da variância de $\hat{P}_j$ dada por

$$\widehat{\text{var}}(\hat{P}_j) = \hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{n_i' \hat{p}_i}.$$

A tabela de mortalidade é uma tabela onde, para cada intervalo I_j, são representados todos os valores de n_j, d_j e m_j e as estimativas $\hat{q}_j$ e $\hat{P}_j$. Notemos que $\hat{P}_j$ pode ser obtido recursivamente através da relação $\hat{P}_j = \hat{p}_j \hat{P}_{j-1}$.

Para mais detalhes sobre a estrutura de uma tabela de mortalidade, consultar Marubini e Valsecchi (1995) e Lawless (2002).

2.1.2 Estimador de Kaplan-Meier

Quando não existe censura, a função de sobrevivência num dado instante t é estimada pela proporção de indivíduos que sobreviveram para além do instante t, ou seja, a proporção de tempos de vida observados de valor superior a t. Esta função designa-se por função de sobrevivência empírica e, com base numa amostra de dimensão n, define-se do seguinte modo:

$$\hat{S}(t) = \frac{\text{número de observações} > t}{n} \qquad t \geq 0.$$

Kaplan e Meier (1958) propuseram um estimador não paramétrico da função de sobrevivência, quando existem observações censuradas. Este estimador é designado por estimador de Kaplan-Meier ou estimador "produto-limite".

Sejam $t_{(1)}, \ldots, t_{(r)}$ os instantes de morte distintos numa amostra de

dimensão n ($r \leq n$), d_i o número de mortes ocorridas em $t_{(i)}$ e n_i o número de indivíduos em risco em $t_{(i)}$. O estimador de Kaplan-Meier da função de sobrevivência é então dado por

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.1}$$

sendo $\hat{S}(t) = 1$ para $0 \leq t < t_{(1)}$. Quando um instante de morte e um instante de censura são registados com o mesmo valor, considera-se que o instante de morte precede o instante de censura.

A estimativa $\hat{S}(t)$ pode ser calculada recursivamente:

$$\begin{aligned} \hat{S}(t_{(1)}) &= 1 - \frac{d_1}{n_1} = \hat{S}(t_{(1)}^+) \\ \hat{S}(t_{(i)}) &= \hat{S}(t_{(i-1)}) \left(1 - \frac{d_i}{n_i}\right) \qquad i = 1, \ldots, r. \end{aligned}$$

A estimativa da variância de $\hat{S}(t)$ é dada pela seguinte expressão, conhecida por fórmula de Greenwood:

$$\widehat{\text{var}}\{\hat{S}(t)\} = [\hat{S}(t)]^2 \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

Notemos que:

- quando não há censura, o estimador de Kaplan-Meier coincide com a função de sobrevivência empírica;
- $\hat{S}(t) = 0$ para $t \geq t_{(r)}$, se $t_{(r)}$ for a maior observação registada, i.e., se a maior observação for não censurada;
- se a maior observação registada t^* for censurada, então $\hat{S}(t)$ nunca toma o valor zero e considera-se que a estimativa está definida apenas até esse instante, sendo $\hat{S}(t) = \hat{S}(t_{(r)})$ para $t_{(r)} \leq t \leq t^*$;

- $\hat{S}(t)$ é uma função em escada, com saltos nos instantes de morte observados;
- $\hat{S}(t)$ é um estimador consistente de $S(t)$ e, sob certas condições de regularidade, pode ser considerado como um estimador de máxima verosimilhança não paramétrico de $S(t)$;
- O estimador de Kaplan-Meier é autoconsistente, de acordo com a definição de Efron (1967).

Vejamos com maior detalhe o significado da propriedade da autoconsistência. Considerando uma amostra $(t_1, \ldots, t_n)$ e $\delta_i, i = 1, \ldots, n$ a variável indicatriz usual, seja

$$\psi(y) = \begin{cases} 1 & \text{se } y > t \\ 0 & \text{se } y \leq t. \end{cases}$$

Diz-se que $\hat{S}(t)$ é um estimador autoconsistente de $S(t)$ se

$$\hat{S}(t) = \frac{1}{n}\left[\sum_{i=1}^{n} \psi(t_i) + \sum_{i:t_i \leq t} (1 - \delta_i)\frac{\hat{S}(t)}{\hat{S}(t_i)}\right].$$

A consistência representada por esta equação significa que o número esperado de sobreviventes no instante t é igual ao número de indivíduos que morrem ou são censurados depois de t mais o número de indivíduos censurados até t que se espera que sobrevivam para além de t. De facto, $\hat{S}(t)/\hat{S}(t_i)$ é uma estimativa da probabilidade de que o tempo de vida de um indivíduo, a que corresponde uma observação censurada t_i, seja maior que t. Resolvendo desta forma o problema colocado pelas observações censuradas inferiores ou iguais a t, pretende-se que, também quando existe censura, o estimador represente a proporção de indivíduos que sobrevivem para além de t.

Intervalo de confiança para a função de sobrevivência

Podemos agora construir um intervalo de confiança para o verdadeiro valor da função de sobrevivência num dado instante t_0. Calculados para vários instantes, os intervalos de confiança darão uma indicação de quão preciso é o estimador $\hat{S}(t)$ em cada um desses pontos.

Tendo $\hat{S}(t)$ uma distribuição assintótica normal de valor médio $S(t)$ e variância estimada dada pela fórmula de Greenwood, um intervalo de $100(1-\alpha)\%$ de confiança para $S(t_0)$, ou seja, para a função de sobrevivência no instante t_0 é dado por

$$\left(\hat{S}(t_0) - z_{1-\alpha/2}\sqrt{\widehat{\text{var}}\hat{S}(t_0)}, \hat{S}(t_0) + z_{1-\alpha/2}\sqrt{\widehat{\text{var}}\hat{S}(t_0)}\right).$$

Embora este seja o intervalo mais utilizado e aquele que habitualmente é construído pelo *software* estatístico, não é isento de problemas. De facto, sendo um intervalo simétrico, pode acontecer que os seus limites estejam fora do intervalo (0,1) quando a estimativa $\hat{S}(t_0)$ estiver próxima de zero ou de um. Uma solução possível é substituir o limite superior à unidade por 1.0 ou o limite inferior a zero por 0.0. Um procedimento alternativo consiste em começar por obter um intervalo de confiança para uma transformação de $\hat{S}(t_0)$, como por exemplo $\log[-\log \hat{S}(t_0)]$, e em seguida calcular o intervalo de confiança para $\hat{S}(t_0)$. Para mais detalhes consultar Collett (2003).

Os intervalos obtidos deste modo designam-se por intervalos de confiança ponto-a-ponto (*pointwise*), visto que dizem respeito a um instante específico. Para a obtenção de bandas de confiança, no sentido de que a função de sobrevivência se encontre, com um certo grau de confiança, dentro dessas bandas para todos os valores de t, são necessários métodos diferentes como, por exemplo, o proposto por Hall e Wellner (1980).

Estimativa não paramétrica de quantis do tempo de vida

Dado que, geralmente, a distribuição do tempo de vida é assimétrica positiva, é preferível utilizar a mediana para caracterizar a localização do centro da distribuição. Então, sendo $\hat{S}(t)$ a estimativa de Kaplan-Meier da função de sobrevivência, a estimativa da mediana do tempo de vida é definida como

$$m = \min\{t_{(i)} : \hat{S}(t_{(i)}) \leq 0.5\},$$

onde $t_{(i)}$ é o i-ésimo instante de morte, $i = 1, \ldots, r$.

Em doenças com prognóstico favorável, acontece por vezes que a estimativa da função de sobrevivência é superior a 0.5 para todos os valores de t. Nesse caso, não é possível obter uma estimativa não paramétrica da mediana do tempo de vida. Será então aconselhável obter a estimativa de outro quantil conveniente da distribuição, sendo que a estimativa do quantil de probabilidade p é dada por

$$\hat{\chi}_p = \min\{t_{(i)} : \hat{S}(t_{(i)}) \leq 1 - p\}.$$

Tal como acima foi definido, o estimador de Kaplan-Meier contempla apenas a existência de observações censuradas à direita. No entanto, modificando o modo de obter o conjunto de risco em cada instante de morte, o estimador passa a ser adequado para as situações em que existe também truncatura à esquerda. Vejamos então qual o significado de n_i, o número de indivíduos em risco de ocorrência do acontecimento de interesse em $t_{(i)}$. Para dados censurados à direita, esta quantidade representa o número de indivíduos presentes no estudo no instante $t_{(0)} = 0$ que tinham um tempo em observação não inferior a $t_{(i)}$.

Para dados truncados à esquerda, redefinimos n_i como sendo o número de indivíduos que entraram no estudo antes do instante $t_{(i)}$ e que têm

um tempo em observação não inferior a $t_{(i)}$. Devemos também salientar que o estimador de Kaplan-Meier da função de sobrevivência no instante t é agora um estimador da probabilidade de sobrevivência para além de t, condicional à sobrevivência ao menor dos instantes de entrada em estudo L. Portanto, trata-se de um estimador da função de sobrevivência condicional $P(T > t|T \geq L) = S(t)/S(L)$.

2.2 Estimação não paramétrica da função de risco cumulativa

Um estimador natural de $H(t)$ é

$$\hat{H}(t) = -\log \hat{S}(t),$$

onde $\hat{S}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência. Um estimador alternativo e mais usual, com melhor comportamento para pequenas amostras, é o estimador de Nelson-Aalen definido por

$$\tilde{H}(t) = \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i}.$$

É também designado por função de risco cumulativa empírica. A estimativa da variância de $\tilde{H}(t)$ é dada por

$$\widehat{\text{var}}\{\tilde{H}(t)\} = \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i^2}.$$

A função de risco cumulativa é importante na identificação de modelos para o tempo de vida com base no comportamento da função de risco, visto que $H(t)$ é uma função não decrescente que será

- linear se $h(t)$ for constante

- convexa se $h(t)$ for crescente
- côncava se $h(t)$ for decrescente.

Além disso, como o declive da função de risco cumulativa fornece informação acerca da forma da função de risco, é possível obter uma estimativa "grosseira" desta função a partir do declive do gráfico do estimador de Nelson-Aalen. Dado que a estimativa assim obtida é frequentemente difícil de interpretar, é então necessário recorrer a métodos de suavização.

Notemos que

$$\begin{aligned}\hat{H}(t) &= -\sum_{i:t_{(i)}\leq t} \log\left(1-\frac{d_i}{n_i}\right)\\ &= \sum_{i:t_{(i)}\leq t}\left(\frac{d_i}{n_i}+\frac{d_i^2}{2n_i^2}+\cdots\right).\end{aligned}$$

Constatamos então que o estimador de Nelson-Aalen pode ser considerado como uma aproximação de 1ª ordem do estimador de Kaplan-Meier da função de risco cumulativa. Para modelos contínuos os dois estimadores são assintoticamente equivalentes e dão resultados que não diferem muito, excepto quando há poucos indivíduos em risco.

Sendo o estimador de Nelson-Aalen um estimador da função de risco cumulativa, é óbvio que a partir dele se pode obter um estimador da função de sobrevivência, também designado por estimador de Breslow, dado por

$$\tilde{S}(t) = \exp(-\tilde{H}(t)) = \prod_{i:t_{(i)}\leq t}\exp\left(-\frac{d_i}{n_i}\right).$$

Visto que $e^{-x}\approx 1-x$ quando x é pequeno, tem-se aqui que $\exp(-d_i/n_i)\approx 1-d_i/n_i$ enquanto n_i for grande comparativamente a d_i, i.e., d_i/n_i pequeno. Como regra prática refere-se, por vezes, $n_i \geq 10d_i$.

Em qualquer instante t, a estimativa de Nelson-Aalen da função de sobrevivência é sempre superior ou igual à estimativa de Kaplan-Meier, visto que $\exp(-x) \geq 1 - x$; logo $\hat{S}(t) \leq \tilde{S}(t), \forall t > 0$. É de salientar que, se a maior observação registada t_{max} for não censurada, para $t \geq t_{max}$, $\hat{S}(t) = 0$ enquanto que $\tilde{S}(t) > 0$. Embora, para pequenas amostras, o estimador de Nelson-Aalen apresente um melhor comportamento do que o estimador de Kaplan-Meier, em muitas situações as estimativas serão muito semelhantes, como já referimos.

Capítulo 3

Testes não paramétricos

3.1 Introdução

Quando se pretende comparar a distribuição do tempo de vida para vários grupos de indivíduos, uma primeira abordagem consiste na obtenção da estimativa de Kaplan-Meier da função de sobrevivência para cada um dos grupos e sua representação gráfica num sistema comum de eixos coordenados, o que permite ter uma ideia do comportamento das curvas de sobrevivência e avaliar, de um modo informal, se existem diferenças entre os vários grupos, relativamente ao seu padrão de sobrevivência. Também é aconselhável representar graficamente as estimativas de Nelson-Aalen da função de risco cumulativa para os vários grupos, pela informação que podemos obter sobre a evolução do risco. No entanto, para uma avaliação rigorosa da existência de diferenças significativas entre as várias curvas é necessário recorrer a testes de hipóteses.

Existem vários testes não paramétricos para comparação das curvas de sobrevivência correspondentes a diferentes grupos de indivíduos, mas nenhum deles se pode considerar adequado a todas as situações. De facto, a escolha do teste mais apropriado depende de diversos factores como os padrões de morte e de censura nos vários grupos e a relação entre as correspondentes funções de risco, bem como a

hipótese alternativa de interesse. A natureza da diferença que o investigador espera detectar determina o teste mais potente para testar a hipótese de igualdade das curvas de sobrevivência.

Se existirem diferenças entre as curvas de sobrevivência, estas diferenças podem ocorrer de muitas maneiras. Por exemplo, as diferenças podem existir no período inicial ou no período final do *follow-up*. As funções de sobrevivência podem cruzar-se, sugerindo possíveis diferenças tanto no período inicial como no final mas não no mesmo sentido. As funções de risco podem ser proporcionais ou não proporcionais e, neste último caso, podem ser divergentes, convergentes ou cruzarem-se.

Comecemos por considerar dois grupos de indivíduos, em que $S_i(t)$ representa a função de sobrevivência de um indivíduo no i-ésimo grupo, sendo então as hipóteses a testar

$$H_0 : S_1(t) = S_2(t) \text{ vs } H_1 : S_1(t) \neq S_2(t).$$

3.2 Teste log-rank

Representemos por $t_1 < \ldots < t_k$ os instantes de morte distintos relativos aos $m + n$ pacientes. Seja d_j o número de mortes ocorridas em $t_j, j = 1, \ldots, k$, d_{ij} o número de mortes ocorridas em t_j no grupo $i, i = 1, 2$, n_j o número de indivíduos em risco imediatamente antes de $t_j, j = 1, \ldots, k$ e n_{ij} o número de indivíduos em risco imediatamente antes de t_j no grupo $i, i = 1, 2$.

A informação relevante em cada instante t_j pode ser resumida numa tabela de contingência 2×2:

Grupo	No. de mortes em t_j	No. de sobrev. para além de t_j	No. de ind. em risco em t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
	d_j	$n_j - d_j$	n_j

Mantel e Haenszel (1959) sugeriram considerar a distribuição condicional das frequências observadas em cada célula, dados os totais marginais, sob a validade da hipótese nula. Isto implica considerar a distribuição da frequência de apenas uma célula, digamos d_{1j}, visto que as outras frequências ficam implicitamente determinadas pelos totais marginais fixos. Então, supondo H_0 verdadeira, a distribuição de d_{1j}, condicional aos valores marginais, é hipergeométrica

$$p(d_{1j}|d_j, n_j) = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}.$$

Sob H_0, o valor médio e a variância condicionais de d_{1j} são, respectivamente,

$$e_{1j} = n_{1j} d_j / n_j$$

e

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Notemos que e_{1j} é o número esperado de mortes no instante t_j no grupo 1, sob H_0. De facto, d_j/n_j é a probabilidade de morte em t_j no grupo $i, i = 1, 2$, condicional à sobrevivência até esse instante, visto que, sob H_0, esta probabilidade é igual nos dois grupos.

Para combinar a informação contida nas k tabelas de contingência, de modo a obter uma medida global do desvio dos valores observados

de d_{1j} em relação aos valores esperados, consideremos a estatística

$$U = \sum_{j=1}^{k} (d_{1j} - e_{1j}).$$

Então, $E(U) = 0$ e

$$\text{var}(U) = \sum_{j=1}^{k} v_{1j}.$$

A estatística de teste proposta por Mantel e Haenszel (1959) é

$$Q = U^2/\text{var}(U)$$

que tem distribuição assintótica χ^2_1, sob H_0.

Considerações sobre o teste log-rank e a proporcionalidade das funções de risco

O teste log-rank é o mais potente na detecção de afastamentos da hipótese de igualdade das distribuições que sejam do tipo riscos proporcionais. É ainda bastante potente para alternativas em que as funções de risco sejam não proporcionais mas não se cruzem. Quando as funções de risco se cruzam, o teste log-rank pode não conseguir detectar diferenças significativas entre as curvas de sobrevivência.

Assim sendo, é importante poder avaliar, num caso concreto, a validade da hipótese de riscos proporcionais. Para tal, podemos fazer uso do seguinte resultado: se as funções de risco são proporcionais, então as respectivas funções de sobrevivência não se cruzam. Obviamente, conclui-se então que o cruzamento das funções de sobrevivência invalida a hipótese de riscos proporcionais.

Seja então $h_1(t)$ a função de risco no instante t para um indivíduo do grupo 1 e $h_2(t)$ a função de risco no instante t para um indivíduo do

grupo 2. Se as funções de risco forem proporcionais, então $h_1(t) = \varphi h_2(t)$, em que $\varphi > 0$ é uma constante que não depende de t. Então

$$\int_0^t h_1(u)du = \int_0^t \varphi h_2(u)du$$
$$\exp\left\{-\int_0^t h_1(u)du\right\} = \exp\left\{-\varphi\int_0^t h_2(u)du\right\}$$
$$S_1(t) = [S_2(t)]^{\varphi}.$$

Como a função de sobrevivência toma valores entre zero e um, tem-se que

se $\varphi < 1 \Rightarrow S_1(t) > S_2(t)\ \ \forall t$ e se $\varphi > 1 \Rightarrow S_1(t) < S_2(t)\ \ \forall t$.

Portanto, as funções de sobrevivência não se cruzam. Esta é uma condição necessária, mas não suficiente, para a proporcionalidade das funções de risco. Assim sendo, uma avaliação informal da validade da hipótese de riscos proporcionais pode ser feita através da representação gráfica das estimativas das funções de sobrevivência. Se as estimativas das funções de sobrevivência não se cruzarem, a hipótese de riscos proporcionais pode ser justificada e o teste log-rank é apropriado. Como é óbvio, as funções de sobrevivência estimadas podem-se cruzar, embora as verdadeiras funções de risco sejam proporcionais.

Outro método gráfico, que permite avaliar de forma mais satisfatória a hipótese de riscos proporcionais, baseia-se no seguinte resultado: se $h_1(t)$ e $h_2(t)$ são funções de risco proporcionais, digamos $h_1(t) = e^{\beta}h_2(t)$, então

$$S_1(t) = [S_2(t)]^{\exp(\beta)} \Leftrightarrow \log[-\log S_1(t)] = \beta + \log[-\log S_2(t)],$$

donde se conclui que os gráficos do logaritmo das funções de risco cumulativas correspondentes aos dois grupos são equidistantes ao longo do tempo. Então, sendo $\hat{S}_1(t)$ e $\hat{S}_2(t)$ estimativas de $S_1(t)$ e $S_2(t)$ não baseadas na hipótese de riscos proporcionais (e.g. utilizando o estimador de Kaplan-Meier), o gráfico de $\log[-\log \hat{S}_1(t)]$ *versus* t tenderá

a ser paralelo ao gráfico de $\log[-\log \hat{S}_2(t)]$ *versus* t, quando $h_1(t)$ e $h_2(t)$ são proporcionais. Dever-se-á avaliar, portanto, se a distância entre os gráficos se mantém razoavelmente constante, ao longo do tempo.

3.3 Teste de Gehan

Gehan (1965) propôs uma generalização do teste de Mann-Whitney-Wilcoxon para dados censurados. Este teste é também designado por teste de Wilcoxon generalizado. Sejam m e n as dimensões das amostras correspondentes aos grupos 1 e 2, respectivamente. Formemos então a amostra conjunta de tempos em observação de dimensão $m+n$ e seja δ_i a variável indicatriz usual.

Vamos agora descrever a forma computacional do teste de Gehan sugerida por Mantel (1967). Seja $U_{kj} = U(t_k, t_j)$ a pontuação atribuída ao comparar um tempo t_k, que foi fixado, com cada um dos restantes tempos observados. Então

$$U_{kj} = \begin{cases} +1 & \text{se } (t_k > t_j, \delta_j = 1) \text{ ou } (t_k = t_j, \delta_k = 0, \delta_j = 1) \\ -1 & \text{se } (t_k < t_j, \delta_k = 1) \text{ ou } (t_k = t_j, \delta_k = 1, \delta_j = 0) \\ 0 & \text{caso contrário.} \end{cases}$$

Seja $U'_k = \sum_{j=1}^{m+n} U_{kj}$ para $k = 1, \ldots, m+n$ $(j \neq k)$. Notemos que U'_k é igual à diferença entre o número das restantes $m+n-1$ observações que correspondem a tempos de vida de certeza menores que t_k e o número de observações que correspondem a tempos de vida que são de certeza maiores que t_k. A estatística de teste é dada por

$$U = \sum_{k=1}^{m+n} U'_k \qquad \text{para } k : t_k \in \text{amostra 1.}$$

Sob a validade de H_0, $E(U) = 0$ e

$$\text{var}(U) = \frac{mn}{(m+n)(m+n-1)} \sum_{k=1}^{m+n} (U'_k)^2.$$

Podemos então considerar a estatística $Z = U/\sqrt{\text{var}(U)}$ que, sob H_0, tem distribuição assintótica $N(0,1)$. Ao nível de significância α rejeitamos H_0 se $|z| \geq z_{1-\alpha/2}$.

Notemos que a estatística de teste se pode escrever de uma forma diferente. Seja

$$U_G = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j})$$

onde $e_{1j} = n_{1j}d_j/n_j$. A variância da estatística U_G é dada por $V_G = \sum_{j=1}^{r} n_j^2 v_{1j}$ e a estatística de teste de Gehan é então $W = U_G^2/V_G$. Sob a validade de H_0, W tem distribuição assintótica de Qui-quadrado com 1 grau de liberdade.

Constatamos, então, que cada diferença $(d_{1j} - e_{1j})$ é ponderada por n_j, o número de indivíduos em risco no instante t_j. Assim sendo, é atribuído maior peso às diferenças $(d_{1j} - e_{1j})$ correspondentes aos instantes onde o número total de indivíduos em risco é elevado, ou seja, aos instantes na parte inicial do estudo. Por isso, este teste é menos sensível que o teste log-rank a diferenças entre o número observado e o número esperado de mortes que se verifiquem na cauda direita da distribuição do tempo de vida.

3.4 Outros testes não paramétricos

O teste log-rank e o teste de Gehan pertencem a uma vasta classe que engloba outros testes não paramétricos para os quais a estatística de

teste tem uma forma comum dada por

$$\frac{\left[\sum_{j=1}^{r} w_j(d_{1j} - e_{1j})\right]^2}{\sum_{j=1}^{r} w_j^2 v_{1j}},$$

onde w_j são constantes conhecidas. Se H_0 for verdadeira, esta estatística tem distribuição assintótica Qui-quadrado com 1 grau de liberdade. Notemos que, de acordo com os valores atribuídos aos pesos w_j, obtemos diferentes testes, nomeadamente

$w_j = 1$	teste log-rank
$w_j = n_j$	teste de Gehan
$w_j = \sqrt{n_j}$	teste de Tarone-Ware
$w_j = \prod_{i:t_{(i)} \leq t_{(j)}} \left(1 - \frac{d_i}{n_i+1}\right)$	teste de Peto-Peto.

A escolha do teste adequado é uma questão fundamental na análise estatística. O teste log-rank tem potência óptima para detectar diferenças em que as funções de risco são proporcionais. A estatística de Gehan põe mais peso nas observações mais pequenas e devido a isto é mais potente para detectar os efeitos a curto prazo. Neste caso, os pesos n_j dependem fortemente dos tempos em que ocorrem mortes e da distribuição dos tempos de censura; por este motivo, este teste pode conduzir a conclusões incorrectas quando os padrões de censura em cada amostra são muito diferentes.

O teste proposto por Tarone e Ware (1977) é um compromisso entre os testes de Gehan e log-rank, visto que também atribui maior peso às diferenças na fase inicial, embora menor que o teste de Gehan. Os testes de Gehan e Tarone-Ware dão mais peso a diferenças entre as funções de sobrevivência nos instantes em que está presente a

maior parte dos indivíduos em estudo (geralmente na parte inicial do estudo).

Peto e Peto (1972) e Prentice (1978) propuseram uma versão alternativa do teste de Mann-Whitney-Wilcoxon para dados censurados. Notemos que o peso correspondente ao teste de Peto-Peto (também designado por teste de Peto-Prentice) é um estimador da função de sobrevivência comum para os dois grupos, de valor próximo do estimador de Kaplan-Meier obtido para a amostra conjunta. Este peso depende da sobrevivência observada nas duas amostras combinadas, mas não é afectado por possíveis padrões de censura distintos.

Fleming e Harrington (1981) propuseram uma classe bastante geral de testes que inclui, como casos particulares, o teste log-rank e uma versão do teste de Mann-Whitney-Wilcoxon, muito semelhante à sugerida por Peto e Peto (1972). Seja $\hat{S}(t)$ o estimador de Kaplan-Meier baseado na amostra conjunta. A função peso dos testes de Fleming-Harrington é dada por

$$W_{p,q}(t_i) = \hat{S}(t_{i-1})^p[1 - \hat{S}(t_{i-1})]^q, \qquad p \geq 0, q \geq 0.$$

Notemos que $\hat{S}(t_0) = 1$ e convenciona-se que $0^0 = 1$. Aqui, a estimativa da função de sobrevivência no instante de morte anterior é usada como peso, para assegurar que os valores dos pesos são conhecidos imediatamente antes do instante em que será feita a comparação. Esta propriedade é designada por "passível de ser predito" na terminologia dos processos de contagem.

Quando $p = q = 0$ obtem-se o teste log-rank. Quando $p = 1, q = 0$ obtem-se a versão acima referida do teste de Mann-Whitney-Wilcoxon. Importa referir que, quando $p > 0$ e $q = 0$, é dado maior peso às diferenças iniciais, enquanto que, quando $p = 0$ e $q > 0$, estes testes dão maior peso a diferenças tardias.

Assim sendo, através de uma escolha apropriada dos valores de p e q, podem ser construídos testes que dão maior ênfase a diferenças entre as funções de sobrevivência que ocorram na região do tempo pretendida e deste modo serem mais potentes contra alternativas em que as funções de risco diferem nessa região.

Vejamos um exemplo. Ao comparar o tempo de sobrevivência livre de doença, i.e., o tempo em remissão correspondente a regimes diferentes em transplantes de medula óssea no tratamento de leucemia, é frequentemente usada uma função peso que dá maior peso às diferenças entre as funções de risco que ocorrem a longo prazo. Tal função dá menor peso a diferenças que ocorram no início, que se devem frequentemente à toxicidade dos regimes preparatórios e dá maior peso às diferenças que ocorrem mais tarde e que serão realmente devidas a diferenças na cura da leucemia. Realizou-se um estudo para comparação da eficácia do transplante autólogo (ou autotransplante) com a do transplante alogénico para tratamento da leucemia mieloblástica aguda. Para tal, pretendeu-se comparar o tempo livre de doença após o transplante, ou seja, o tempo até à ocorrência de recaída ou morte, aquilo que acontecer primeiro. Sabe-se que pacientes a quem foi feito um transplante alogénico tendem a ter mais complicações na fase inicial, sendo a mais grave a doença do enxerto contra o hospedeiro, que ocorre nos primeiros 100 dias após o transplante e é frequentemente mortal. Os pacientes a quem foi feito o outro tipo de transplante não estão obviamente em risco para esta doença e tendem a ter uma taxa de sobrevivência mais alta neste período. Os investigadores estão então interessados em comparar a eficácia dos tratamentos entre os sobreviventes a longo prazo.

Klein e Moeschberger (1997) analisaram os dados referentes a 101 pacientes com leucemia mieloblástica aguda, em que 51 pacientes receberam um autotransplante e os restantes um transplante alogénico,

tendo usado o teste de Fleming-Harrington com $p = 0$ e $q = 1$, i.e., com peso $W(t_i) = 1-\hat{S}(t_{i-1})$. Obtiveram um valor-$p = 0.0404$, donde se conclui existir diferença significativa entre os dois tratamentos no que diz respeito ao tempo livre de doença. Por outro lado, os valor-p correspondentes aos testes log-rank e de Gehan foram, respectivamente, 0.5368 e 0.7556. De facto, esta é uma situação em que o teste log-rank e o teste de Gehan não devem ser usados, porque se verifica um cruzamento das funções de risco (cerca dos 12 meses) e estes dois testes não conseguem, por isso, detectar a vantagem a longo prazo dos transplantes alogénicos.

Relativamente a todos os testes acima descritos, tendo a estatística de teste uma distribuição que é assintótica, deve haver cuidado na interpretação dos resultados quando as amostras são de pequena dimensão ou quando há poucas observações não censuradas. Também é assumido que existe independência entre os tempos de vida e de censura.

Os testes referidos anteriormente podem ser generalizados de modo a permitir a comparação de r grupos, para $r \geq 3$. Sob a validade de $H_0 : S_1(t) = \ldots = S_r(t)$, a estatística de teste tem distribuição assintótica Qui-quadrado com $r - 1$ graus de liberdade. Para mais detalhes consultar Klein e Moeschberger (1997) ou Collett (2003).

Capítulo 4

Modelo de regressão de Cox

4.1 Introdução

Numa perspectiva que podemos designar por semi-paramétrica, 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, que tornam adequada a sua utilização num grande número de situações práticas, em áreas tais como medicina e engenharia.

A sua contribuição para o desenvolvimento da Análise de Sobrevivência foi extremamente importante e disso são testemunho inúmeros artigos publicados, quer relatando aplicações do modelo e diferentes abordagens aos problemas de inferência, quer tomando-o como referência para desenvolvimento de novas ideias. De facto, muitos dos trabalhos de investigação que se seguiram dizem respeito a extensões e generalizações do modelo de Cox como, por exemplo, modelos que incluem covariáveis dependentes do tempo e modelos de riscos competitivos.

Um dos aspectos inovadores do modelo de Cox reside no facto de ser formulado com base na relação entre a função de risco e as covariáveis. Com efeito, seja T uma v.a. contínua que representa o tempo de vida. Cox (1972) propôs um modelo em que, no instante t e para um indi-

víduo a que esteja associado o vector de covariáveis $\mathbf{z} = (z_1, \ldots, z_p)'$, a função de risco é da forma

$$\begin{aligned} h(t; \mathbf{z}) &= h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}) \\ &= h_0(t) \exp(\beta_1 z_1 + \cdots + \beta_p z_p), \end{aligned} \tag{4.1}$$

em que $\beta_1, \ldots, \beta_p$ 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 de risco para um indivíduo a que está associado o vector $\mathbf{z} = \mathbf{0}$. Esta função é designada por função de risco subjacente.

É um modelo de regressão semi-paramétrico visto que, embora o efeito das covariáveis seja modelado parametricamente, a função de risco subjacente $h_0(t)$, que descreve a forma comum das distribuições do tempo de vida para os indivíduos em estudo, não é especificada. Tal facto contribui para a flexibilidade do modelo.

Trata-se de um modelo de riscos proporcionais, visto que as funções de risco correspondentes a dois indivíduos com covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$ são proporcionais. De facto,

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

não depende de t.

As covariáveis têm um efeito multiplicativo na função de risco, de acordo com o factor $\exp(\boldsymbol{\beta}' \mathbf{z})$, que é designado por risco relativo. Portanto, o modelo de Cox (1972) pressupõe que a influência das covariáveis na função de risco não sofre qualquer alteração durante o período em que os indivíduos se encontram em observação.

Notemos que

$$\log\left(\frac{h(t; \mathbf{z})}{h_0(t)}\right) = \beta_1 z_1 + \cdots + \beta_p z_p,$$

no espírito da formulação habitual na análise de regressão que consiste em modelar linearmente o efeito das covariáveis, sendo $\boldsymbol{\beta}'\mathbf{z} = \sum_{j=1}^{p} \beta_j z_j$ a componente linear do modelo. Portanto, o modelo de Cox também pode ser encarado como um modelo linear para o logaritmo do risco relativo. A quantidade $\boldsymbol{\beta}'\mathbf{z}_i = \sum_{j=1}^{p} \beta_j z_{ij}$ é designada por *risk score* ou índice de prognóstico para o i-ésimo indivíduo.

É de salientar que a função de risco subjacente representa geralmente a função de risco correspondente a algum tipo de condições padrão e não é necessariamente a função de risco correspondente a um indivíduo com vector de covariáveis nulo. De facto, isto pode não ser realista em muitas situações como, por exemplo, quando uma covariável contínua, tal como a idade, é incluída no modelo. Frequentemente, considera-se que o valor zero de cada covariável corresponde à média dos valores dessa covariável para todos os indivíduos em estudo. Então, ao valor z_{ij} da covariável z_j para o indivíduo i será subtraída a média $\overline{z}_j$ e a função de risco é escrita na forma

$$h(t; \mathbf{z}_i) = h_0(t) \exp(\beta_1(z_{i1} - \overline{z}_1) + \cdots + \beta_p(z_{ip} - \overline{z}_p)).$$

No entanto, como esta redefinição das covariáveis não afecta a inferência sobre a sua influência no risco de morte, trabalharemos com as covariáveis não transformadas e com a função de risco escrita na forma habitual.

4.2 Interpretação dos coeficientes

A interpretação dos coeficientes de regressão é obviamente importante para a compreensão da relação estabelecida entre as variáveis explanatórias e a variável resposta no modelo de Cox. Habitualmente, esta interpretação não é feita em termos de β_j mas sim de $\exp(\beta_j)$, por esta quantidade ter um significado mais directo no que

diz respeito ao risco de morte.

Consideremos dois indivíduos a que estão associados os vectores de covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, que diferem apenas nos valores da covariável z_j. Dada a forma da função de risco, tem-se então que

$$\begin{aligned}\frac{h(t;\mathbf{z}_1)}{h(t;\mathbf{z}_2)} &= \frac{h_0(t)\exp(\beta_1 z_{11}+\ldots+\beta_j z_{1j}+\ldots+\beta_p z_{1p})}{h_0(t)\exp(\beta_1 z_{21}+\ldots+\beta_j z_{2j}+\ldots+\beta_p z_{2p})}\\ &= \exp(\beta_j(z_{1j}-z_{2j})).\end{aligned}$$

Assim sendo, $\exp(\beta_j)$ representa o risco relativo de ocorrência do acontecimento para dois indivíduos que diferem de uma unidade nos valores da covariável z_j, sendo iguais os respectivos valores das restantes covariáveis.

Exemplo 1: num estudo sobre o tempo desde o diagnóstico até à morte por determinada doença, seja z uma covariável binária definida por

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

Se o indivíduo pertence ao grupo 1, então $h(t;z=0)=h_0(t)$; se o indivíduo pertence ao grupo 2, $h(t;z=1)=h_0(t)e^{\beta}$. Portanto, se $\beta<0 \Leftrightarrow e^{\beta}<1$, os pacientes do grupo 2 têm melhor prognóstico do que os do grupo 1; se $\beta>0 \Leftrightarrow e^{\beta}>1$, os pacientes do grupo 1 têm melhor prognóstico do que os do grupo 2.

Exemplo 2: realizou-se um ensaio clínico para comparar os efeitos de um novo medicamento e de um medicamento tradicional, ambos destinados a manter um ritmo cardíaco normal em pacientes sofrendo de fibrilhação auricular. Os doentes foram distribuídos, de forma aleatória, pelos dois grupos de tratamento e foi registado o tempo até à ocorrência de recaída, i.e., o tempo durante o qual o medicamento manteve o ritmo cardíaco normal. Foram considerados também dois

potenciais factores de risco, o que levou à definição das seguintes covariáveis:

z_1 : tratamento (0 = tradicional, 1 = novo)
z_2 : idade no início do estudo (em anos)
z_3 : volume do coração (mm^3)

Então

$$e^{\beta_1} = \frac{h(t; z_1 = 1, z_2 = j, z_3 = k)}{h(t; z_1 = 0, z_2 = j, z_3 = k)}$$

representa o risco de recaída de um indivíduo tratado com o novo medicamento relativamente a um indivíduo que recebe o medicamento tradicional, para indivíduos com a mesma idade no início do estudo e com o mesmo volume do coração,

$$e^{\beta_2} = \frac{h(t; z_1 = i, z_2 = j + 1, z_3 = k)}{h(t; z_1 = i, z_2 = j, z_3 = k)}$$

representa o risco de recaída de um indivíduo com determinada idade no início do estudo relativamente a outro indivíduo um ano mais novo, para indivíduos no mesmo grupo de tratamento e com o mesmo volume do coração e

$$e^{\beta_3} = \frac{h(t; z_1 = i, z_2 = j, z_3 = k + 1)}{h(t; z_1 = i, z_2 = j, z_3 = k)}$$

representa o risco de recaída de um indivíduo com determinado volume do coração relativamente a outro indivíduo com menos 1 mm^3 de volume, para indivíduos no mesmo grupo de tratamento e com a mesma idade no início do estudo.

4.3 Função de verosimilhança

Suponhamos que se encontram em estudo n indivíduos e que foram observados k tempos de vida distintos $t_{(1)} < \ldots < t_{(k)}$, $k < n$. Seja

$$R_i = R(t_{(i)}) = \{j : t_j \geq t_{(i)}\}$$

o conjunto de risco no instante $t_{(i)}$, i.e., o conjunto de índices associados aos indivíduos em observação imediatamente antes do instante $t_{(i)}$. Seja $\mathbf{z}_{(i)}$ o vector de covariáveis associado ao indivíduo que morre em $t_{(i)}$.

Cox (1972) baseou a inferência sobre $\boldsymbol{\beta}$ na seguinte função, utilizada como função de verosimilhança:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_{(i)})}{\sum_{l \epsilon R_i} \exp(\boldsymbol{\beta}'\mathbf{z}_l)}. \tag{4.2}$$

Esta função não depende de $h_0(t)$ e permite assim a realização de inferência sobre o vector de parâmetros $\boldsymbol{\beta}$, sem que seja necessário fazer qualquer restrição à forma de $h_0(\cdot)$.

Notemos que a função (4.2) não é uma verosimilhança no sentido usual, visto não representar a probabilidade de realização de um acontecimento observável. De facto, sob as condições mencionadas na secção 1.9, a função de verosimilhança (1.3) correspondente ao modelo de Cox é da forma

$$\begin{aligned} &L[\boldsymbol{\beta}, h_0(t)] \\ &= \prod_{i=1}^{n} [h_0(t_i) \exp(\boldsymbol{\beta}'\mathbf{z}_i) S_0(t_i)^{\exp(\boldsymbol{\beta}'\mathbf{z}_i)}]^{\delta_i} [S_0(t_i)^{\exp(\boldsymbol{\beta}'\mathbf{z}_i)}]^{1-\delta_i} \\ &= \prod_{i \epsilon D} \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_i)}{\sum_{l \epsilon R_i} \exp(\boldsymbol{\beta}'\mathbf{z}_l)} \prod_{i \epsilon D} \left(h_0(t_i) \sum_{l \epsilon R_i} \exp(\boldsymbol{\beta}'\mathbf{z}_l) \right) \prod_{i=1}^{n} S_0(t_i)^{\exp(\boldsymbol{\beta}'\mathbf{z}_i)}, \end{aligned}$$

onde D designa o conjunto de indivíduos cuja morte foi observada. No entanto, a utilização desta função para a estimação de $\boldsymbol{\beta}$ exigiria a consideração simultânea de $\boldsymbol{\beta}$ e $h_0(t)$, o que não é muito conveniente. Note-se que a verosimilhança (4.2) coincide com o primeiro termo do produto anterior.

Cox (1975) argumentou que (4.2) pode ser interpretada como uma verosimilhança parcial (destinada a permitir a realização de inferência

na presença de parâmetros perturbadores), sendo aqui $h_0(t)$ entendida como uma função perturbadora.

Vários autores (e.g., Andersen e Gill, 1982) consideraram o modelo de Cox no contexto dos processos de contagem e deste modo provaram que, sob condições de regularidade bastante gerais, o estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$ é consistente, assintoticamente normal com valor médio $\boldsymbol{\beta}$ e matriz de covariância $I(\boldsymbol{\beta})^{-1}$, onde

$$I_{jk}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 \log L}{\partial\beta_j \partial\beta_k}\right).$$

4.4 Existência de observações empatadas

Para a obtenção de (4.2) foram consideradas apenas observações distintas, visto que a ocorrência de observações empatadas tem probabilidade nula sob um modelo contínuo. No entanto, a falta de precisão no registo dos dados pode dar origem à existência de valores iguais. De facto, os tempos de vida são geralmente registados arredondando ao dia, mês ou ano mais próximo. Vejamos então como deve ser modificada a função de verosimilhança no caso de existirem observações empatadas.

Para os n indivíduos em estudo, foram observados os tempos de vida distintos $t_{(1)} < \ldots < t_{(k)}$. Seja $\mathbf{z}_{ij}$ o vector de covariáveis associado ao indivíduo j, $j = 1, \ldots, d_i$, que morre em $t_{(i)}$.

Se o número d_i de indivíduos que morrem em $t_{(i)}$ é pequeno comparado com o número de indivíduos pertencentes a R_i, podemos utilizar a seguinte aproximação da função de verosimilhança, proposta por Peto (1972) e Breslow (1974):

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_i)}{\left[\sum_{l\in R_i} \exp(\boldsymbol{\beta}'\mathbf{z}_l)\right]^{d_i}} \tag{4.3}$$

onde $\mathbf{s}_i = \sum_{j=1}^{d_i} \mathbf{z}_{ij}$, para $i = 1, \ldots, k$. Esta é a verosimilhança geralmente usada no *software* estatístico. Quando não há observações empatadas, i.e., quando $d_i = 1$ para $i = 1, \ldots, k$, a função (4.3) reduz-se à verosimilhança parcial (4.2).

Para obter (4.3), Breslow (1974) aproximou a função de risco subjacente por uma função constante entre instantes de morte sucessivos, ou seja, considerou

$$h_0(t) = h_i \qquad t_{(i-1)} < t \leq t_{(i)}, \quad i = 1, \ldots, k$$

sendo $t_{(0)} = 0$. Após alguns cálculos, obtemos a verosimilhança

$$L(h_1, \ldots, h_k, \boldsymbol{\beta}) = \prod_{i=1}^{k} \left\{ h_i^{d_i} e^{\boldsymbol{\beta}' \mathbf{s}_i} \exp\left[-h_i (t_{(i)} - t_{(i-1)}) \sum_{l \in R_i} \exp(\boldsymbol{\beta}' \mathbf{z}_l) \right] \right\} \tag{4.4}$$

Para $\boldsymbol{\beta}$ fixo, maximizemos (4.4) com respeito a $h_i, i = 1, \ldots, k$, resolvendo o sistema de equações $\partial \log L / \partial h_i = 0$, $i = 1, \ldots, k$ que tem como solução, para $i = 1, \ldots, k$

$$\hat{h}_i = \frac{d_i}{(t_{(i)} - t_{(i-1)}) \sum_{l \in R_i} \exp(\boldsymbol{\beta}' \mathbf{z}_l)}. \tag{4.5}$$

Substituindo h_i por $\hat{h}_i$ em (4.4), obtém-se

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left[e^{-1} \frac{d_i}{t_{(i)} - t_{(i-1)}} \right]^{d_i} \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}' \mathbf{s}_i)}{\left[\sum_{l \in R_i} \exp(\boldsymbol{\beta}' \mathbf{z}_l) \right]^{d_i}}$$

que é proporcional a (4.3).

4.5 Estimação da função de sobrevivência

Como

$$S(t; \mathbf{z}) = [S_0(t)]^{\exp(\boldsymbol{\beta}' \mathbf{z})},$$

interessa-nos estimar $S_0(t)$, o que permitirá obter estimativas de $S(t;\mathbf{z})$ para qualquer $\mathbf{z}$. Tendo obtido $\hat{\boldsymbol{\beta}}$ a partir da verosimilhança parcial, Kalbfleisch e Prentice (1973) determinaram um estimador de máxima verosimilhança não paramétrico de $S_0(t)$.

Suponhamos que se encontram em estudo n indivíduos e que foram observados k tempos de vida distintos $t_{(1)} < \ldots < t_{(k)}, k < n$. Seja R_i o conjunto de risco no instante $t_{(i)}$ e D_i o conjunto de índices associados aos d_i indivíduos que morreram em $t_{(i)}$.

Consideremos então um modelo discreto em que a função de risco em $t_{(i)}, i = 1, \ldots, k$ é $h_i = 1 - \alpha_i$ com $\alpha_i = S_0(t_{(i+1)})/S_0(t_{(i)})$.

Supondo $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ e maximizando a função de verosimilhança com respeito a $\alpha_1, \ldots, \alpha_k$, obtemos as seguintes equações de máxima verosimilhança

$$\sum_{l \in D_i} \frac{\exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)}{1 - \alpha_i^{\exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)}} = \sum_{l \in R_i} \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)$$

para $i = 1, \ldots, k$.

Quando $d_i = 1, i = 1, \ldots, k$, a solução desta equação é

$$\hat{\alpha}_i = \left(1 - \frac{\exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_{(i)})}{\sum_{l \in R_i} \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)}\right)^{\exp(-\hat{\boldsymbol{\beta}}' \mathbf{z}_{(i)})}.$$

Caso contrário, é necessário resolver a equação recorrendo a um método iterativo.

O estimador de máxima verosimilhança de $S_0(t)$ é dado por

$$\hat{S}_0(t) = \prod_{i: t_{(i)} \leq t} \hat{\alpha}_i$$

que é uma função em escada, com descontinuidades em cada instante de morte observado $t_{(i)}$.

A função de sobrevivência estimada para um indivíduo a que esteja associado o vector de covariáveis $\mathbf{z}$ é então

$$\hat{S}(t;\mathbf{z}) = \prod_{i:t_{(i)}\leq t} \hat{\alpha}_i^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{z})}.$$

Têm sido propostos muitos outros estimadores para a função de sobrevivência no modelo de Cox. A aproximação de $h_0(t)$ considerada por Breslow (1974) leva a um estimador que não requer a utilização de métodos iterativos quando $d_i > 1$ para algum i:

$$\begin{aligned}\tilde{H}_0(t) &= -\log \tilde{S}_0(t) \\ &= \sum_{i:t_{(i)}\leq t} \frac{d_i}{\sum_{l\in R_i} \exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_l)}.\end{aligned}$$

Notemos que $\tilde{H}_0(t)$ corresponde ao estimador de Nelson-Aalen e $\hat{S}_0(t)$ corresponde ao estimador de Kaplan-Meier. $\tilde{S}_0(t)$ e $\hat{S}_0(t)$ não diferem muito quando $d_i = 1, i = 1, \ldots, k$ ou quando existem poucas observações empatadas, excepto na cauda direita da distribuição.

Sob hipóteses adequadas no mecanismo de censura, prova-se a consistência e a normalidade assintótica de $\hat{S}(t;\mathbf{z})$.

4.6 Comparação de distribuições do tempo de vida

O modelo de Cox pode ser usado para testar a hipótese de igualdade das distribuições do tempo de vida para dois grupos de indivíduos contra a hipótese alternativa de que as distribuições são diferentes (sendo as funções de risco proporcionais).

Seja então z uma covariável indicatriz do grupo, definida por

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

Então, as funções de sobrevivência S_1 e S_2 correspondentes aos dois grupos estão relacionadas por

$$S_2(t) = S_1(t)^{\exp(\beta)}$$

e testar $H_0 : S_1(t) = S_2(t)$ é equivalente a testar $H_0 : \beta = 0$.

Sejam $t_{(1)} < \ldots < t_{(k)}$ os instantes de morte distintos relativos aos $m + n$ pacientes e seja
d_j : número de mortes ocorridas em $t_{(j)}, j = 1, \ldots, k$
d_{ij} : número de mortes ocorridas em $t_{(j)}$ no grupo $i, i = 1, 2$
n_j : número de indivíduos em risco em $t_{(j)}, j = 1, \ldots, k$
n_{ij} : número de indivíduos em risco em $t_{(j)}$ no grupo $i, i = 1, 2$

Sob o modelo de Cox e supondo que existem poucas observações empatadas, tem-se que

$$\log L(\beta) = r_2\beta - \sum_{j=1}^{k} d_j \log(n_{1j} + n_{2j}e^{\beta}),$$

onde $r_2 = \sum_{j=1}^{k} d_{2j}$. Então,

$$U(\beta) = \frac{\partial \log L}{\partial \beta} = r_2 - \sum_{j=1}^{k} \frac{d_j n_{2j} e^{\beta}}{n_{1j} + n_{2j}e^{\beta}}$$

$$I(\beta) = -\frac{\partial^2 \log L}{\partial \beta^2} = \sum_{j=1}^{k} \frac{d_j n_{1j} n_{2j} e^{\beta}}{(n_{1j} + n_{2j}e^{\beta})^2}.$$

Um teste bastante simples para testar $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ que não requer o cálculo de $\hat{\beta}$ é o teste *score*. Tem-se então que, sob H_0, a estatística $Z = \frac{U(0)}{\sqrt{I(0)}}$ tem distribuição assintótica $N(0, 1)$.

Notemos que

$$U(0) = \sum_{j=1}^{k} \left(d_{2j} - \frac{d_j n_{2j}}{n_j} \right)$$

e

$$I(0) = \sum_{j=1}^{k} \frac{d_j n_{1j} n_{2j}}{n_j^2}.$$

Quando há um número substancial de observações empatadas, deve ser usado um teste que entre em linha de conta com a natureza discreta dos dados. Esse teste é ainda baseado na estatística Z, com $U(0)$ dado como anteriormente mas em que

$$I(0) = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Como, sob H_0, Z^2 tem distribuição assintótica χ_1^2, este teste é equivalente ao teste log-rank e, por vezes, é designado por teste de Cox-Mantel.

4.7 Métodos de selecção de variáveis

Numa análise de regressão é habitual pretender identificar quais as covariáveis que têm influência significativa na sobrevivência dos indivíduos, de entre todas as que foram registadas, por exemplo, num estudo clínico. De facto, o modelo de regressão final deverá ser parcimonioso, embora possam ser também incluídas variáveis com relevância clínica que não se tenham revelado estatisticamente significativas. Dado que o coeficiente β_j representa o efeito da covariável z_j na sobrevivência do indivíduo, para avaliar se existe evidência de que essa covariável influencia significativamente o tempo de vida, podemos testar

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

utilizando o teste de Wald, em que a estatística de teste $\hat{\beta}_j^2/\text{var}(\hat{\beta}_j)$ tem, sob H_0, distribuição assintótica χ_1^2.

Notemos que se está a testar a hipótese de que a covariável z_j não tem influência significativa na sobrevivência, na presença das restantes covariáveis. Ora, em geral, as estimativas $\hat{\beta}_j$ não são independentes umas das outras, o que dificulta a interpretação dos resultados de testes sobre os coeficientes associados a covariáveis incluídas num modelo. É então preferível recorrer a métodos que permitam a comparação de modelos alternativos.

Consideremos um modelo de Cox com p covariáveis (modelo 1) e um modelo de Cox em que estão incluídas q covariáveis adicionais (modelo 2):

- Modelo 1
  $h_0(t)\exp(\beta_1 z_1 + \cdots + \beta_p z_p)$
- Modelo 2
  $h_0(t)\exp(\beta_1 z_1 + \cdots + \beta_p z_p + \beta_{p+1} z_{p+1} + \cdots + \beta_{p+q} z_{p+q})$

A questão que aqui se coloca é saber se os q termos adicionais incluídos no modelo 2 melhoram significativamente o poder explanatório deste modelo, relativamente ao modelo 1. Se tal não acontecer, os q termos podem ser omitidos e o modelo 1 é considerado adequado.

A função de verosimilhança resume a informação contida nos dados acerca dos parâmetros desconhecidos num dado modelo. Então, uma estatística adequada (que mede quão bem um modelo se ajusta aos dados) é o valor da função de verosimilhança quando os parâmetros são substituídos pelas suas estimativas de máxima verosimilhança. No entanto, a estatística $-2\log\hat{L}$ não pode ser usada por si só como medida da adequabilidade do modelo, pois o valor de $\hat{L}$ depende da dimensão da amostra. Logo, $-2\log\hat{L}$ só é útil ao compararmos

modelos ajustados aos mesmos dados.

Assim sendo, os modelos podem ser comparados com base na diferença entre os valores da estatística $-2\log\hat{L}$ para cada modelo. Fazemos então um teste de razão de verosimilhanças para testar

$$H_0 : \beta_{p+1} = \ldots = \beta_{p+q} = 0 \quad \text{vs} \quad H_1 : \exists i : \beta_i \neq 0,\ i = p+1, \ldots, p+q$$

Sob H_0, a estatística $-2\log(\hat{L}_1/\hat{L}_2)$ tem distribuição assintótica χ^2_q.

Tendo alertado para os inconvenientes dos métodos de selecção automática de covariáveis, Collett (2003) propôs a seguinte estratégia para selecção do modelo que melhor se ajusta aos dados:

1. Ajustamos os modelos contendo apenas uma covariável, um de cada vez. Calculamos os valores da estatística $-2\log\hat{L}$ para cada um dos modelos e comparamos com o valor da estatística para o modelo nulo (sem covariáveis). Assim, determinamos quais as covariáveis que, por si só, levam a uma redução significativa do valor da estatística e que são potencialmente importantes.
2. Incluimos as covariáveis potencialmente importantes (passo 1) num só modelo e calculamos o valor da estatística $-2\log\hat{L}$. Omitimos então uma covariável de cada vez e retemos no modelo apenas aquelas que levam a um aumento significativo do valor da estatística.
3. Variáveis que, quando consideradas isoladamente, não eram importantes e que, portanto, não foram consideradas no passo 2, podem revelar-se importantes na presença de outras. Estas covariáveis são então incluídas no modelo obtido no passo 2, uma de cada vez, e retemos alguma que leve a uma redução significativa do valor de $-2\log\hat{L}$.
4. É feita uma verificação final, para assegurar que nenhuma covariável pode ser omitida sem levar a um aumento significativo

do valor de $-2 \log \hat{L}$ e que nenhuma covariável não incluída leva a uma redução significativa do valor da estatística.

O nível de significância considerado para a inclusão ou omissão de covariáveis não deve ser muito pequeno; Collett (2003) recomenda que se utilize $\alpha \simeq 0.1$.

4.8 Análise de resíduos

Uma definição apropriada de resíduo é fundamental para se poder avaliar a adequabilidade de um qualquer modelo de regressão. Uma definição natural de resíduo é a diferença entre o valor observado da variável resposta e o valor predito pelo modelo, como acontece na regressão linear. A existência de observações censuradas e a própria forma do modelo de Cox levam a que não se possa fazer uma definição análoga para este modelo. De facto, a definição de resíduo é mais difícil e menos directa na modelação do tempo de vida do que no contexto de outros modelos de regressão.

A ausência de uma definição óbvia levou a que fossem propostos diversos resíduos para o modelo de Cox, cada um dos quais com um papel importante na análise de diferentes aspectos do ajustamento do modelo.

Resíduos de Cox-Snell

Cox e Snell (1968) propuseram um tipo de resíduos baseado na ideia de que, se o modelo é correcto, os resíduos devem comportar-se como uma amostra proveniente de uma determinada distribuição conhecida. Foram os primeiros resíduos propostos para o modelo de Cox e são úteis para avaliar o ajustamento global do modelo final.

Seja T uma v.a. contínua com função de distribuição F e função de sobrevivência S. Então, $F(T) \sim U(0,1)$ e também $S(T) \sim U(0,1)$. Como $H(T) = -\log S(T)$, vem que $H(T)$ tem distribuição exponencial de valor médio 1.

Então, como no modelo de Cox se tem que

$$\begin{aligned} H(t;\mathbf{z}) &= \int_0^t h_0(u)\exp(\boldsymbol{\beta}'\mathbf{z})du \\ &= \exp(\boldsymbol{\beta}'\mathbf{z})\ H_0(t), \end{aligned}$$

o resíduo para o i-ésimo indivíduo, $i = 1, \ldots, n$, é definido como

$$r_i = \hat{H}(t_i) = \exp(\hat{\beta}'\mathbf{z}_i)\hat{H}_0(t_i)$$

em que $\hat{\beta}$ e $\hat{H}_0(t)$ são as estimativas de máxima verosimilhança parcial. Se o modelo que foi ajustado aos dados é satisfatório, então os valores estimados $\hat{H}(t_i)$ terão propriedades semelhantes aos verdadeiros valores $H(t_i)$. Portanto, os resíduos r_i devem comportar-se, aproximadamente, como uma amostra aleatória proveniente de uma população com distribuição Exp(1).

Quando existem dados censurados é necessário ter isso em conta, visto que se determinada observação é censurada também o é o resíduo correspondente. Vejamos então como se obtêm os resíduos de Cox-Snell modificados de forma a explicitamente levarem em conta a censura.

Seja t_i^* uma observação censurada (à direita) e seja t_i o verdadeiro, mas desconhecido, tempo de vida desse indivíduo. O valor correcto do resíduo para esse indivíduo seria $\hat{H}(t_i)$ mas apenas podemos calcular $\hat{H}(t_i^*)$. Como $t_i > t_i^*$, então $\hat{H}(t_i) > \hat{H}(t_i^*)$. Os resíduos de Cox-Snell são então modificados pela adição de uma constante positiva, designada por excesso residual:

$$r_i' = \begin{cases} r_i & \text{se } t_i \text{ é um tempo de vida observado} \\ r_i + \Delta & \text{se } t_i \text{ é uma observação censurada} \end{cases}$$

Qual o valor a atribuir a Δ? Notemos que, se o tempo de vida $T \sim \text{Exp}(1)$, então o tempo de vida residual médio é

$$E(T - x|T \geq x) = \int_x^\infty \frac{S(u)}{S(x)} du = 1$$

donde $E(T|T \geq x) = 1 + x$. Este facto justifica considerar $\Delta = 1$ e tem-se então que os resíduos de Cox-Snell modificados são dados por

$$r_i' = \begin{cases} r_i & \text{se } t_i \text{ é uma observação não censurada} \\ r_i + 1 & \text{se } t_i \text{ é uma observação censurada} \end{cases}$$

Os resíduos de Cox-Snell têm propriedades bastante diferentes dos resíduos usados em regressão linear porque, por exemplo, não se distribuem de forma simétrica em torno de zero e, de facto, nem podem tomar valores negativos. Podem assumir qualquer valor no intervalo $(0, \infty)$, sendo que $r_i' > 1$ se t_i for uma observação censurada. Além disso, como os resíduos de Cox-Snell seguem uma distribuição exponencial quando o modelo que foi ajustado aos dados é apropriado, terão uma distribuição bastante assimétrica.

Para examinar a adequabilidade do modelo, devemos então verificar se a amostra dos resíduos se pode considerar como proveniente de uma população exponencial de valor médio um. Para tal, podemos fazer a representação gráfica dos pontos $\left(r_i', \tilde{H}(r_i')\right)$, onde $\tilde{H}(r_i')$ é a estimativa de Nelson-Aalen da função de risco cumulativa dos resíduos. Se a representação obtida for aproximadamente uma recta de declive um e ordenada na origem nula, conclui-se que o modelo é adequado.

Uma desvantagem destes resíduos é que não indicam o tipo de afastamento do modelo que é detectado quando o gráfico acima referido não é linear. Devem portanto ser usados com alguma prudência, principalmente no caso de pequenas amostras, visto que o afastamento

da distribuição dos resíduos da distribuição exponencial pode ser devido, em parte, à substituição de $\boldsymbol{\beta}$ e $H_0(t)$ pelas suas estimativas de máxima verosimilhança parcial.

Resíduos de Schoenfeld

Estes resíduos foram propostos por Schoenfeld (1982) e diferem dos resíduos de Cox-Snell em dois aspectos:

- não é necessário obter uma estimativa da função de risco cumulativa
- a cada indivíduo corresponde um conjunto de valores, um por cada covariável que foi incluída no modelo de regressão de Cox

Para o i-ésimo indivíduo em estudo, o resíduo de Schoenfeld correspondente à covariável $z_j, j = 1, \ldots, p$ é dado por

$$r_{ji} = \delta_i\{z_{ji} - a_{ji}\}$$

onde

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é uma observação não censurada} \\ 0 & \text{se } t_i \text{ é uma observação censurada} \end{cases}$$

e

$$a_{ji} = \frac{\sum_{l \in R_i} z_{jl} \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)}{\sum_{l \in R_i} \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)}.$$

Para um indivíduo cujo tempo de vida não foi observado, estes resíduos são sempre nulos. Então, para distinguir resíduos que são verdadeiramente iguais a zero daqueles que correspondem a observações censuradas, estes últimos são geralmente indicados como valores omissos.

Para um indivíduo cuja morte foi observada em t_i, o resíduo é a diferença entre o valor da covariável z_j correspondente a esse indivíduo

e uma média ponderada dos valores dessa variável para todos os indivíduos em risco em t_i. O peso associado a um indivíduo $l \in R_i$ é $\exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_l)$. Notemos que

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i \left\{ z_{ji} - \frac{\sum_{l \in R_i} z_{jl} \exp(\boldsymbol{\beta}' \mathbf{z}_l)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}' \mathbf{z}_l)} \right\}$$

onde L é a função de verosimilhança parcial. A i-ésima parcela desta soma, calculada em $\hat{\boldsymbol{\beta}}$, é então o resíduo de Schoenfeld correspondente à covariável z_j para o i-ésimo indivíduo.

Como as estimativas $\hat{\beta}_j$ são tais que

$$\frac{\partial \log L}{\partial \beta_j} \Big|_{\hat{\boldsymbol{\beta}}} = 0,$$

conclui-se que a soma, para todos os indivíduos em estudo, dos resíduos de Schoenfeld correspondentes a cada covariável, é igual a zero. Para grandes amostras, o valor esperado de r_{ji} é zero e os resíduos são não correlacionados.

Se o modelo que foi ajustado aos dados é adequado, gráficos dos resíduos de Schoenfeld *versus* os tempos de vida ou *versus* as ordens dos tempos de vida devem ter o aspecto de uma nuvem aleatória de pontos, centrada em zero.

Os resíduos de Schoenfeld são particularmente úteis na avaliação da hipótese de riscos proporcionais, após o ajustamento aos dados de um modelo de Cox. Grambsch e Therneau (1994) propuseram uma versão destes resíduos que afirmam ser mais eficaz na detecção de afastamentos do modelo assumido. Designam-se por resíduos de Schoenfeld padronizados ou ponderados. Seja $\mathbf{r}_i = (r_{1i}, r_{2i}, \ldots, r_{pi})'$ o vector dos resíduos de Schoenfeld associado ao i-ésimo indivíduo. Os resíduos de Schoenfeld padronizados r^*_{ji} definem-se então como sendo as componentes do vector

$$\mathbf{r}^*_i = k \times \mathrm{var}(\hat{\boldsymbol{\beta}}) \mathbf{r}_i$$

onde k é o número de mortes observadas entre os n indivíduos e $\text{var}(\hat{\boldsymbol{\beta}})$ é a matriz de covariância dos estimadores dos parâmetros β_j no modelo de Cox que foi ajustado aos dados.

Resíduos martingala

Os resíduos martingala constituem uma modificação dos resíduos de Cox-Snell e devem o seu nome ao facto de poderem ser obtidos a partir de resultados sobre martingalas, no âmbito dos processos de contagem (ver Apêndice). São muito úteis na determinação da forma funcional que deve ser usada para uma dada covariável, de modo a explicar o melhor possível o seu efeito na sobrevivência, bem como na detecção de *outliers*.

Quando todas as covariáveis são fixas, no início do estudo, o resíduo martingala associado ao i-ésimo indivíduo, $i = 1, \ldots, n$, é dado por

$$\hat{M}_i = \delta_i - \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_i)\hat{H}_0(t_i) = \delta_i - r_i$$

em que δ_i é a variável indicatriz usual.

Os resíduos martingala exibem grande assimetria e tomam valores no intervalo $(-\infty, 1)$, sendo negativos os resíduos correspondentes a observações censuradas. Para grandes amostras, são não correlacionados e têm valor esperado igual a zero, quando calculado para o verdadeiro (desconhecido) vector de parâmetros $\boldsymbol{\beta}$. Mostra-se que $\sum_{i=1}^{n} \hat{M}_i = 0$.

Notemos que $\hat{M}_i$ representa a diferença entre o número observado de acontecimentos para o i-ésimo indivíduo no intervalo $(0, t_i)$ e o correspondente número esperado, estimado com base no modelo ajustado. De facto, o número observado de "mortes" é um se o tempo t_i é não censurado e zero se t_i é censurado ou seja, é igual a δ_i. Por outro lado, r_i é uma estimativa de $H(t_i)$, o que pode ser interpretado como

o número esperado de "mortes" em $(0, t_i)$ visto estarmos a considerar apenas um indivíduo.

Assim sendo, a análise dos resíduos martingala irá revelar indivíduos fracamente ajustados pelo modelo, i.e., indivíduos que viveram demasiado tempo ou morreram demasiado cedo quando comparados com os outros indivíduos que têm características semelhantes. São estes os indivíduos que designamos por *outliers*. Para detectar a sua existência, faz-se a representação gráfica dos resíduos *versus* o índice de cada indivíduo.

A representação gráfica dos resíduos *versus* uma determinada covariável indica se esta deve ser incluída no modelo tal como foi registada ou se é necessário proceder a uma modificação da sua forma funcional. Conforme referido em Therneau e Grambsch (2000), a abordagem mais simples consiste em representar num gráfico os resíduos martingala resultantes do ajustamento do modelo nulo, i.e., sem covariáveis, *versus* os valores de cada uma das covariáveis incluídas no modelo. Para facilitar a interpretação do gráfico obtido, é aconselhável representar também uma curva de suavização como, por exemplo, a curva obtida pelo LOWESS (Locally Weighted Scatterplot Smoother), proposto por Cleveland (1979). Therneau e Grambsch (2000) mostraram que se o modelo correcto para a covariável z_j é $\exp(f(z_j)\beta_j)$ para alguma função suave f, então a curva de suavização para a covariável z_j mostrará a forma de f, sob certas hipóteses. Se a curva for linear, não é necessário transformar a covariável em causa; se não for linear, pode apresentar diversos aspectos e deve-se proceder à transformação correspondente.

Um caso bastante interessante é aquele em que a covariável é contínua e se coloca a questão de a transformar numa covariável binária, com o objectivo de tornar mais simples a interpretação do modelo. Se a curva de suavização apresentar um "patamar", será então aconselhável

discretizar a covariável, embora esta não seja uma opção consensual. Royston *et al.* (2006) apresentam uma interessante discussão deste assunto.

Desvios residuais

Dado que os resíduos martingala não se distribuem de forma simétrica em torno de zero, mesmo quando o modelo é adequado, os gráficos neles baseados são de difícil interpretação. Os desvios residuais (*deviance residuals*) foram introduzidos por Therneau *et al.* (1990) com o objectivo de colmatar essa falta de simetria. São definidos por

$$r_{Di} = \text{sgn}(\hat{M}_i)\{-2[\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)]\}^{\frac{1}{2}}$$

onde $\hat{M}_i$ é o resíduo martingala para o i-ésimo indivíduo e a função sgn(.) é função sinal. A motivação original para estes resíduos deve-se ao facto de estes serem componentes da estatística *deviance*, que é dada por $D = -2(\log \hat{L}_c - \log \hat{L}_s)$, onde $\hat{L}_c$ e $\hat{L}_s$ são as verosimilhanças parciais maximizadas sob o modelo corrente e sob o modelo saturado, respectivamente. Quanto menor for o valor de D, melhor é o modelo.

Os desvios residuais são tais que $D = \sum_{i=1}^{n} r_{Di}^2$, pelo que as observações correspondentes a resíduos bastante grandes em valor absoluto são as que não estão bem ajustadas pelo modelo. Um gráfico dos desvios residuais *versus* os valores do índice de prognóstico estimado para cada indivíduo constitui um diagnóstico muito útil, neste contexto.

Modelos com interacção

Pode ser apropriado incluir no modelo um termo que corresponda aos efeitos individuais para cada combinação dos vários níveis de dois factores de risco ou prognóstico. Tais efeitos denominam-se in-

teracções. De um modo geral, se A e B são dois factores e o risco de morte depende da combinação dos níveis de A e B, diz-se que há interacção entre A e B. Frequentemente, trata-se de avaliar de que forma variam os efeitos de dois tratamentos para diferentes grupos de pacientes. Uma interacção quantitativa envolve variação na magnitude, mas não no sentido, das diferenças do efeito do tratamento entre subgrupos de pacientes. Caso contrário, se houver alteração no sentido do efeito do tratamento entre subgrupos, a interacção diz-se qualitativa.

Considerando um modelo com duas covariáveis z_1 e z_2, a introdução de um termo de interacção é feita adicionando uma nova covariável $z_3 = z_1 z_2$ ao modelo, pelo que a função de risco será

$$h(t; \mathbf{z}) = h_0(t) \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3)$$

Exemplo: realizou-se um ensaio clínico para avaliar o efeito de um novo tratamento de quimioterapia no tempo até à ocorrência de recaída local em pacientes com cancro da mama, submetidas a cirurgia conservadora. As doentes, que se encontravam em diferentes estádios da doença, foram distribuídas de forma aleatória, por dois grupos de tratamento, após a realização da cirurgia. Foram definidas as seguintes covariáveis:

z_1 : tratamento (0 = tratamento tradicional, 1 = novo tratamento)

z_2 : estádio da doença (0 = estádio I, 1 = estádio II)

Consideremos um modelo de Cox incluindo as covariáveis z_1, z_2 e $z_3 = z_1 z_2$. A função de risco correspondente a cada um dos 4 grupos de pacientes é então:

$$\begin{aligned}
(z_1, z_2, z_3) &= (0, 0, 0) \qquad h_0(t) \\
(z_1, z_2, z_3) &= (1, 0, 0) \qquad h_0(t) e^{\beta_1} \\
(z_1, z_2, z_3) &= (0, 1, 0) \qquad h_0(t) e^{\beta_2} \\
(z_1, z_2, z_3) &= (1, 1, 1) \qquad h_0(t) e^{\beta_1 + \beta_2 + \beta_3}
\end{aligned}$$

Portanto, $\exp(\beta_1)$ representa o risco de recaída de uma doente que recebeu o novo tratamento, relativamente a uma doente que recebeu o tratamento tradicional, para doentes no estádio I. Para as doentes no estádio II, este mesmo risco relativo é dado por $\exp(\beta_1+\beta_3)$. Notemos que, num modelo sem interacção, $\exp(\beta_1)$ representaria o risco de recaída de uma doente que recebeu o novo tratamento, relativamente a uma doente que recebeu o tratamento tradicional, para doentes no mesmo estádio.

4.9 Extensões do modelo de Cox

Partindo do modelo de Cox básico (4.1), vamos aqui referir duas extensões do modelo em que deixa de se verificar a proporcionalidade das funções de risco.

4.9.1 Modelo de Cox estratificado

Por vezes, a hipótese de riscos proporcionais não é válida para alguma covariável. Suponhamos então que às várias categorias de uma dada variável qualitativa correspondem funções de risco que são claramente não proporcionais. No entanto, pode acontecer que as funções de risco sejam proporcionais para subgrupos de indivíduos em cada categoria. Uma extensão do modelo de Cox permite acomodar esta situação, considerando que em cada estrato, correspondente a uma categoria ou nível da covariável, é válido o modelo

$$h_j(t;\mathbf{z}) = h_{0j}(t)\exp(\boldsymbol{\beta}'\mathbf{z})$$

para $j = 1,\ldots,m$, onde m é o número de estratos e $\mathbf{z}$ é o vector constituído pelas restantes covariáveis. Este modelo assume que para indivíduos no estrato j a que estão associados vectores de covariáveis

$\mathbf{z}_1$ e $\mathbf{z}_2$, as funções de risco são proporcionais:

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

No entanto, indivíduos em estratos diferentes podem ter funções de risco não proporcionais, visto que as funções $h_{01}(t), \ldots, h_{0m}(t)$ são arbitrárias e não relacionadas. Os coeficientes de regressão $\boldsymbol{\beta}$ não dependem do estrato, por isso admite-se que o efeito das covariáveis é o mesmo em todos os estratos. Os parâmetros $\boldsymbol{\beta}$ são estimados através da maximização de uma verosimilhança parcial

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} L_j(\boldsymbol{\beta})$$

onde $L_j(\boldsymbol{\beta})$ é a verosimilhança parcial para o estrato j obtida de acordo com (4.3), com as contribuições calculadas em cada instante de morte observado nesse estrato.

É de salientar que a estratificação relativamente a uma certa covariável impede obviamente a estimação do efeito dessa covariável no tempo de vida. Por outro lado, quando a covariável é contínua é necessário categorizá-la, o que também traz problemas.

4.9.2 Modelo de Cox com covariáveis dependentes do tempo

Conforme anteriormente referido, uma covariável dependente do tempo é uma variável cujo valor se pode alterar ao longo do tempo e por isso toma o valor $z_i(t)$ para o i-ésimo indivíduo no instante t. De facto, em muitos estudos em que são gerados dados de sobrevivência, os indivíduos são monitorizados ao longo do período de observação e os valores de certas covariáveis podem ser registados regularmente. É natural que um modelo de regressão que leve em conta esses diferentes

valores seja mais satisfatório que um modelo que inclua apenas os valores dessas variáveis registados no início do estudo. O modelo de Cox pode ser generalizado por inclusão deste tipo de covariáveis, passando a função de risco a depender do valor das covariáveis em cada instante:

$$h(t; \mathbf{z}(t)) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}(t)).$$

Notemos que a notação $\mathbf{z}(t)$ indica que o vector pode conter uma ou mais covariáveis dependentes do tempo, sendo as restantes fixas. Como consequência, as funções de risco correspondentes a dois indivíduos com vectores de covariáveis $\mathbf{z}_1(t)$ e $\mathbf{z}_2(t)$ não são proporcionais. De facto, neste caso

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

já depende de t e portanto o modelo já não é um modelo de riscos proporcionais. Para realizar inferência com este tipo de modelo, a verosimilhança parcial pode ainda ser utilizada. Assim sendo, o logaritmo da função de verosimilhança parcial seria dado por

$$\sum_{i=1}^{n} \delta_i \left\{ \sum_{j=1}^{p} \beta_j z_{ji}(t_i) - \log \sum_{l \in R(t_i)} \exp \left(\sum_{j=1}^{p} \beta_j z_{jl}(t_i) \right) \right\}$$

em que $R(t_i)$ é o conjunto de risco no instante t_i (instante de morte do i-ésimo indivíduo), $i = 1, \ldots, n$. Notemos que é então necessário dispor dos valores de cada uma das variáveis no modelo em cada instante de morte t_i para todos os indivíduos em $R(t_i)$, pelo que por vezes são usados valores aproximados (e.g. o último valor da covariável registado antes do instante em questão).

4.10 Testar a hipótese de riscos proporcionais

Sendo a proporcionalidade das funções de risco uma hipótese crucial do modelo de Cox, é compreensível a existência de um grande número de métodos para a verificação da validade deste pressuposto.

Métodos gráficos

Começamos por referir um método gráfico que pode ser usado na fase inicial da análise estatística, antes do ajustamento do modelo. Como para o modelo de Cox se tem que

$$\log[-\log S(t;\mathbf{z})] = \boldsymbol{\beta}'\mathbf{z} + \log[-\log S_0(t)]$$

então, para dois vectores de covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, a distância entre $\log[-\log S(t;\mathbf{z}_1)]$ e $\log[-\log S(t;\mathbf{z}_2)]$ será constante.

Para p covariáveis fixas que tomem um pequeno número de valores, consideremos os M subgrupos definidos por todas as combinações dos valores que as covariáveis assumem. Então, uma avaliação gráfica da hipótese de riscos proporcionais consiste em obter a estimativa de Kaplan-Meier da função de sobrevivência em cada um desses grupos homogéneos de indivíduos e fazer a representação de $\log[-\log \hat{S}_m(t)]$ *versus* t num mesmo gráfico, para $m = 1, \ldots, M$. Se os gráficos forem razoavelmente paralelos, então um modelo de Cox com as p covariáveis pode ser ajustado aos dados. Frequentemente, este gráfico é feito marcando $\log t$ no eixo das abcissas.

Quanto maior for o número de subgrupos mais difícil se torna a interpretação destes gráficos, além de que podem surgir problemas devido ao pequeno número de indivíduos em cada subgrupo. Por este motivo, é frequente que a validade da hipótese de riscos proporcionais

seja explorada para cada covariável separadamente, apesar dos inconvenientes desta abordagem.

Modelo de Cox estratificado

O modelo de Cox estratificado pode ser usado para testar a validade da hipótese de riscos proporcionais. De facto, o método gráfico descrito anteriormente pode não ser praticável quando o número de subgrupos definidos pelos valores das covariáveis é elevado. Por outro lado, considerar cada covariável isoladamente pode esconder alguma relação existente entre as funções de risco de grupos diferentes. Como compromisso entre estas duas opções, a utilização do modelo de Cox estratificado permite-nos investigar a hipótese de proporcionalidade dos riscos para cada uma das covariáveis consideradas, levando em conta as restantes.

Consideremos o vector de covariáveis $\mathbf{z} = (z_1, \ldots, z_p)'$ para as quais pretendemos investigar a hipótese de riscos proporcionais. Seja então $\mathbf{z} = (z_j, \mathbf{z}^-)$ onde $\mathbf{z}^- = (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_p)'$ é o vector das restantes $p-1$ covariáveis. Se z_j for dicotómica, assumindo os valores 0 e 1, teremos então dois estratos e a função de risco será dada por

$$h(t; \mathbf{z}) = \begin{cases} h_{01}(t) \exp(\boldsymbol{\beta}'^{-} \mathbf{z}^-) & \text{se } z_j = 0 \\ h_{02}(t) \exp(\boldsymbol{\beta}'^{-} \mathbf{z}^-) & \text{se } z_j = 1 \end{cases}$$

Supõe-se que as covariáveis incluídas em $\mathbf{z}^-$ satisfazem a hipótese de riscos proporcionais e a estimativa de $\hat{\boldsymbol{\beta}}^-$ é obtida maximizando a verosimilhança parcial sobre toda a amostra. As funções de sobrevivência subjacentes são então estimadas separadamente em cada estrato, usando alguma das estimativas referidas na secção 4.5 e faz-se a representação gráfica de $\log[-\log \hat{S}_{01}(t)]$ e $\log[-\log \hat{S}_{02}(t)]$ *versus* t. Se os gráficos forem razoavelmente paralelos, não haverá motivos

para duvidar da proporcionalidade dos riscos e a covariável z_j pode ser incluída no modelo de Cox.

Teste baseado em covariáveis dependentes do tempo

Cox (1972) propôs o seguinte procedimento para testar a hipótese de riscos proporcionais para a covariável fixa z_j, na presença das restantes $p-1$ covariáveis:

defina-se uma transformada de z_j, dependente do tempo, da forma $z_j(t) = z_j g(t)$ e inclua-se no modelo de Cox que contém as p covariáveis

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta_j z_j + \gamma z_j(t) + \boldsymbol{\beta}^{-}\mathbf{z}^{-})$$

onde $\mathbf{z}^{-} = (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_p)'$. Para indivíduos com os mesmos valores das covariáveis $\mathbf{z}^{-}$, a razão das funções de risco de dois indivíduos com valores $z_j \neq 0$ e $z_j = 0$ é dada por $\exp(\beta_j z_j + \gamma z_j(t))$. Em geral, $\gamma \neq 0$ significa que existe uma variação no tempo na razão das funções de risco para dois indivíduos com valores diferentes de z_j. A escolha de $g(t)$ é habitualmente restrita a algumas funções monótonas simples como $g(t) = t$ e $g(t) = \log t$. Então,

- $\gamma > 0$ indica que a razão das funções de risco cresce linearmente com o tempo ou o com o logaritmo do tempo

- $\gamma < 0$ indica que a razão das funções de risco decresce linearmente com o tempo ou o com o logaritmo do tempo

Para testar $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$ podemos utilizar o teste de Wald ou o teste de razão de verosimilhanças. Esta é, portanto, uma utilização do modelo de Cox com covariáveis dependentes do tempo no contexto particular da avaliação da validade da hipótese de riscos proporcionais.

Uso dos resíduos de Schoenfeld

Os resíduos de Schoenfeld padronizados são particularmente úteis para avaliar o pressuposto de riscos proporcionais após o ajustamento de um modelo de Cox. Grambsch e Therneau (1994) mostraram que o valor médio do resíduo padronizado de Schoenfeld no instante t_i, para a covariável z_j, é dado por

$$E(r^*_{ji}) \approx \beta_j(t_i) - \hat{\beta}_j$$

onde $\beta_j(t_i)$ é o coeficiente de z_j, que varia com o tempo, calculado no instante de morte t_i e $\hat{\beta}_j$ é a estimativa de β_j no modelo de Cox ajustado aos dados. Então, um gráfico dos valores de $r^*_{ji} + \hat{\beta}_j$ *versus* o tempo, ou alguma função do tempo $g(t)$, dará informação acerca da forma do coeficiente $\beta_j(t)$ e, portanto, da forma como o efeito da covariável poderá depender do tempo. Em particular, uma linha horizontal sugere que o coeficiente de z_j é constante e que a hipótese de riscos proporcionais é válida. Para facilitar a interpretação, é habitual incluir no gráfico uma curva de suavização LOWESS com o respectivo intervalo de confiança. A marcação de linhas horizontais de referência em zero e em $\hat{\beta}_j$ é também aconselhável, visto permitirem uma análise mais completa do gráfico.

É de salientar que a escolha da função $g(t)$, ou seja, da escala do tempo, poderá ter bastante influência no aspecto do gráfico e consequentemente nas conclusões que dele se retiram. As funções mais usadas são $g(t) = t$, $g(t) = \log t$, a ordem dos tempos de vida ou uma transformação baseada na estimativa de Kaplan-Meier da função de sobrevivência. Esta última opção tende a espalhar os resíduos de forma bastante regular ao longo do gráfico, da esquerda para a direita, evitando potenciais problemas com *outliers*. Se os instantes em que ocorrem mortes estão muito concentrados no início do período de observação, não deve ser usada a escala do próprio

tempo ($g(t) = t$), visto que dificultará a análise dos resíduos. A transformação $g(t) = \log t$ é frequentemente uma boa escolha quando a distribuição do tempo tem uma cauda longa.

A observação do gráfico deve ser complementada pela realização de um teste formal proposto por Grambsch e Therneau (1994). O teste é baseado nos resíduos de Schoenfeld padronizados e tem uma versão global e uma versão específica para cada covariável. O teste para cada covariável tem a seguinte motivação: escreva-se $\beta(t)$ como uma regressão em $g(t)$ da seguinte forma,

$$\beta_j(t) = \beta_j + \theta_j(g_j(t) - \overline{g}_j), j = 1, \ldots, p$$

onde $\overline{g}_j$ é a média dos $g_j(t_i)$s. A hipótese nula de riscos proporcionais corresponde a $\theta_j = 0, j = 1, \ldots, p$. A estatística de teste (Grambsch e Therneau, 1994) pode ser interpretada como uma medida da correlação entre os resíduos associados a cada covariável e os tempos de vida e tem, sob H_0, uma distribuição χ^2_1. Caso a hipótese nula seja rejeitada, existe evidência de que a correlação é não nula e portanto existe evidência de não proporcionalidade.

Notemos que se $g(t)$ for uma função específica do tempo, o teste acima referido coincide com o teste para a proporcionalidade das funções de risco proposto por Cox (1972). Por exemplo, se $g(t) = \log t$, então $\beta_j(t) \simeq \beta_j + \theta_j \log t$, o que corresponderia a incluir no preditor linear $\beta_j z_j + \theta_j z_j \log t$. Testar a hipótese $\theta_j = 0$ é, portanto, testar a hipótese de riscos proporcionais.

Estratégias para lidar com a não proporcionalidade

O que se deve fazer quando o diagnóstico utilizado dá forte evidência de não proporcionalidade para uma ou mais covariáveis? Antes de mais, devemos averiguar se é realmente importante. Haverá situações

nas quais uma não proporcionalidade "significativa" pode não fazer diferença para a interpretação de um conjunto de dados, como, por exemplo, quando a amostra é de grande dimensão. De facto, neste caso, a variação em $\hat{\beta}_j(t)$ pode ser pequena em relação à estimativa $\hat{\beta}_j$ que representa o melhor efeito "global" da covariável z_j. Também pode ocorrer que essa não proporcionalidade esteja a ser influenciada pela existência de alguns poucos *outliers*.

Se a não proporcionalidade existir e for importante, há diversas estratégias que podem ser seguidas, algumas no contexto do modelo de Cox:

1. Estratificar

   Considera-se um modelo de Cox estratificado pelas covariáveis para as quais há evidência de não proporcionalidade dos riscos. Esta opção tem os inconvenientes já referidos, além de que as análises estratificadas são menos eficientes.

2. Particionar o eixo do tempo

   Se a hipótese de riscos proporcionais for válida em intervalos de tempo consecutivos, podemos considerar essa partição do eixo do tempo e ajustar, por exemplo, um modelo de Cox *piecewise* (Collett, 2003).

3. Incluir covariáveis dependentes do tempo

   Podemos modelar a não proporcionalidade relativa à covariável z criando uma covariável dependente do tempo $z^*(t)$ tal que $\beta(t)z = \beta z^*(t)$.

4. Usar outro tipo de modelo

   Um modelo de tempo de vida acelerado ou de riscos aditivos pode ser mais apropriado aos dados.

Capítulo 5

Modelos de sobrevivência paramétricos

Em consequência da grande flexibilidade do modelo e da existência de *software* estatístico de fácil acesso e utilização, o modelo de regressão de Cox tem, de facto, dominado a análise de dados de sobrevivência. No entanto, se for razoável admitir um determinado modelo paramétrico para o tempo de vida, teremos a vantagem de dispor de métodos de inferência de aplicação directa. Além disso, Efron (1977) mostrou que, sob certas circunstâncias, os modelos paramétricos levam à obtenção de estimadores dos parâmetros de regressão mais eficientes do que os obtidos com base no modelo de Cox. Vamos em seguida apresentar as distribuições contínuas univariadas mais utilizadas na Análise de Sobrevivência e alguns modelos de regressão paramétricos.

5.1 Algumas distribuições contínuas

Distribuição exponencial

A distribuição é caracterizada por apresentar uma função de risco constante, portanto o risco de morte é o mesmo em qualquer instante, seja qual for o tempo decorrido desde o instante inicial do estudo. Seja T uma v.a. com distribuição exponencial de parâmetro $\lambda > 0$, com

função densidade de probabilidade dada por

$$f(t) = \lambda \exp(-\lambda t), \qquad t \geq 0.$$

Então

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

A distribuição exponencial ocupa um lugar de referência na análise de dados de sobrevivência devido ao seu significado histórico, simplicidade matemática e importantes propriedades. O facto de a função de risco ser constante ao longo do tempo é consequência da "ausência de memória" ou "ausência de envelhecimento" característica da distribuição, o que obviamente restringe a utilização desta distribuição em muitas aplicações industriais e na área da saúde. Actualmente, os meios computacionais disponíveis permitem que, com facilidade, se opte por modelos mais complexos mas mais adequados à realidade.

Distribuição de Weibull

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

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

A função de risco é

- monótona crescente se $\gamma > 1$
- monótona decrescente se $0 < \gamma < 1$
- constante se $\gamma = 1$ (distribuição exponencial).

A mediana da distribuição é dada por

$$\chi_{1/2} = \left(\frac{\log 2}{\lambda}\right)^{1/\gamma}.$$

A distribuição de Weibull é, provavelmente, o modelo paramétrico mais usado em Análise de Sobrevivência, nomeadamente na área das ciências biomédicas. Para tal contribui a razoável flexibilidade da função de risco e também a possibilidade de representar a função de risco e a função de sobrevivência através de expressões analíticas simples. A distribuição de Weibull aqui considerada surge como distribuição limite do mínimo de variáveis aleatórias não negativas e independentes, convenientemente normalizadas.

Por ser conveniente, por vezes, trabalhar com o logaritmo do tempo de vida, vamos aqui referir outra distribuição limite de valores extremos, a distribuição de Gumbel (de mínimos).
Seja $Y = \log T$, onde T segue uma distribuição de Weibull de parâmetros λ e α. Então Y tem distribuição de Gumbel de parâmetros $a = -\log\lambda$ e $b = 1/\alpha$. A função densidade de probabilidade e a função de sobrevivência são definidas, para $-\infty < y < \infty$, por

$$f(y) = \frac{1}{b}\exp\left[\frac{y-a}{b} - \exp\left(\frac{y-a}{b}\right)\right]$$
$$S(y) = \exp\left[-\exp\left(\frac{y-a}{b}\right)\right].$$

Distribuição gama

A distribuição gama com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\alpha > 0$ tem função densidade de probabilidade dada por

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

A função de sobrevivência é

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

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

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

A função de risco pode ser

- monótona crescente se $\alpha > 1$,

  com $h(0) = 0$ e $\lim_{t\to\infty} h(t) = \lambda$
- monótona decrescente se $0 < \alpha < 1$,

  com $\lim_{t\to 0^+} h(t) = \infty$ e $\lim_{t\to\infty} h(t) = \lambda$
- constante se $\alpha = 1$ (distribuição exponencial).

Apesar da flexibilidade da função de risco, esta distribuição é menos usada do que a distribuição de Weibull como modelo de tempo de vida, dada a inexistência de formas funcionais fechadas para a função de sobrevivência e para a função de risco.

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 função densidade de probabilidade de T é então

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

A função de sobrevivência é dada por

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

onde $\Phi(\cdot)$ é a função de distribuição da $N(0,1)$.

Este é um modelo com função de risco unimodal. A função de risco $h(t)$ é tal que $h(0) = 0$, é crescente até um valor máximo e depois decresce, com $\lim_{t\to\infty} h(t) = 0$. O instante em que ocorre o máximo depende do valor de σ. A mediana da distribuição é dada por $\chi_{1/2} = \exp(\mu)$.

Distribuição log-logística

Nas situações em que é necessário considerar um modelo com função de risco unimodal, esta distribuição constitui uma alternativa particularmente interessante à distribuição de Weibull, dada a facilidade de representação analítica da função de risco e da função de sobrevivência. Para uma distribuição log-logística com parâmetro de escala $\lambda > 0$ e parâmetro de forma $\alpha > 0$, com função densidade de probabilidade dada por

$$f(t) = \frac{\alpha\lambda t^{\alpha-1}}{(1+\lambda t^{\alpha})^2},$$

tem-se que

$$S(t) = \frac{1}{1+\lambda t^{\alpha}} \quad \text{e} \quad h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}$$

representam as funções de sobrevivência e de risco, respectivamente.

A função de risco é monótona decrescente se $0 < \alpha \leq 1$. Para $\alpha > 1$ a função é unimodal: é crescente desde zero até um valor máximo, atingido para $t = (\frac{\alpha-1}{\lambda})^{1/\alpha}$ e depois decresce, com $\lim_{t\to\infty} h(t) = 0$. A mediana da distribuição é $\chi_{1/2} = \lambda^{-1/\alpha}$.

Distribuição de Gompertz

Esta distribuição foi introduzida por Gompertz em 1825 como modelo para a mortalidade humana e, desde então, tem sido usada com

sucesso nas ciências actuariais, demografia e biologia. Uma propriedade importante é o facto do logaritmo da função de risco ser uma função linear do tempo. Para $t \geq 0$ e $\theta > 0$, a função de risco e a função de sobrevivência são dadas por

$$\begin{aligned} h(t) &= \theta \exp(\alpha t) \\ S(t) &= \exp\left\{\frac{\theta}{\alpha}[1 - \exp(\alpha t)]\right\} \end{aligned}$$

e a correspondente função densidade de probabilidade é

$$f(t) = \theta \exp(\alpha t) \exp\left\{\frac{\theta}{\alpha}[1 - \exp(\alpha t)]\right\}.$$

O parâmetro α determina a forma da função de risco, que é monótona crescente quando $\alpha > 0$. Notemos que, quando $\alpha < 0$, a função de risco é monótona decrescente e a correspondente função de distribuição é imprópria, visto que $F(\infty) < 1$. Trata-se de uma distribuição de Gompertz que é adequada a situações em que, na população, existem indivíduos para os quais nunca se realiza o acontecimento de interesse, que são designados por indivíduos imunes ou não susceptíveis.

5.2 Avaliação da adequabilidade de um modelo paramétrico

Quando a população em estudo é homogénea e pretendemos averiguar se uma determinada distribuição para o tempo de vida é plausível, uma maneira informal de o fazer consiste em representar graficamente uma estimativa não paramétrica de alguma transformada da função de sobrevivência da distribuição. Para facilitar a interpretação deste tipo de gráfico, a transformação da função de sobrevivência deve ser uma função linear e, como tal, dar origem a uma linha recta se o

modelo considerado for apropriado. Iremos referir, como exemplo, as distribuições de Weibull e log-logística.

1. **Distribuição de Weibull**

   Neste caso, é conveniente considerar a seguinte transformação:

   $$\log[-\log S(t)] = \log\lambda + \gamma \log t.$$

   Assim sendo, se o modelo Weibull for adequado, o gráfico de $\log[-\log \hat{S}(t)]$ *versus* $\log t$ será razoavelmente linear, onde $\hat{S}(t)$ é a estimativa de Kaplan-Meier da função de sobrevivência. O declive e a ordenada na origem da "recta" podem então ser usados como estimativas grosseiras de γ e $\log\lambda$, respectivamente. Se o declive for bastante próximo da unidade, o tempo de vida será bem modelado por uma distribuição exponencial.

2. **Distribuição log-logística**

   A adequabilidade do modelo log-logístico para a análise de dados de sobrevivência pode ser verificada empiricamente fazendo uso da seguinte relação linear:

   $$\log\left\{\frac{S(t)}{1-S(t)}\right\} = -\log\lambda - \alpha \log t.$$

   Deste modo, usando o estimador de Kaplan-Meier da função de sobrevivência para obter a estimativa do logaritmo da possibilidade de sobrevivência para além do instante t, se o gráfico de $\log\left\{\hat{S}(t)/(1-\hat{S}(t))\right\}$ *versus* $\log t$ for razoavelmente linear, podemos considerar que o modelo log-logístico é adequado.

5.3 Modelos de regressão paramétricos

Quando no capítulo 1 descrevemos algumas classes de modelos de regressão, referimos que todos admitiam uma versão paramétrica e uma versão semi-paramétrica, consoante a função de risco subjacente é ou não especificada. Vamos em seguida apresentar dois dos modelos de regressão paramétricos mais utilizados: o modelo Weibull e o modelo log-logístico.

Modelo de regressão Weibull

O modelo Weibull pode ser expresso como modelo de riscos proporcionais ou como modelo de tempo de vida acelerado, sendo, como já referimos, o único modelo de regressão que pertence simultaneamente a estas duas classes.

Comecemos por considerar o modelo Weibull como modelo de riscos proporcionais. A função de risco de um indivíduo com vector de covariáveis $\mathbf{z}$ pode então ser escrita como

$$h(t;\mathbf{z}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}) = \lambda\gamma t^{\gamma-1}\exp(\boldsymbol{\beta}'\mathbf{z}),$$

donde podemos concluir que o tempo de vida desse indivíduo tem distribuição de Weibull com parâmetro de escala $\lambda\exp(\boldsymbol{\beta}'\mathbf{z})$ e parâmetro de forma γ. Portanto, o efeito das covariáveis é modificar o parâmetro de escala da distribuição, enquanto que o parâmetro de forma permanece inalterado.

A função de sobrevivência é então

$$S(t;\mathbf{z}) = \exp(-\lambda t^{\gamma}\exp(\boldsymbol{\beta}'\mathbf{z})). \tag{5.1}$$

Sendo um modelo de tempo de vida acelerado, o modelo Weibull permite também uma representação como modelo log-linear. Vejamos

então quais as relações que se podem estabelecer entre os parâmetros do modelo quando é expresso nestas duas formas. Estas relações são importantes porque a maior parte dos *packages* estatísticos (e.g., R) apenas fornece as estimativas dos parâmetros para o modelo log-linear.

Como vimos anteriormente, o logaritmo do tempo de vida pode ser escrito como

$$\log T = \mu + \boldsymbol{\alpha}'\mathbf{z} + \sigma\varepsilon,$$

em que, para o modelo Weibull, ε segue uma distribuição de Gumbel com função densidade de probabilidade dada por

$$f(\varepsilon) = \exp(\varepsilon - \exp(\varepsilon)), \qquad -\infty < \varepsilon < \infty. \tag{5.2}$$

Para o que se segue é necessário ter em conta que a v.a. $\exp(\varepsilon)$ tem distribuição exponencial de valor médio 1. A função de sobrevivência de T, dado o vector de covariáveis $\mathbf{z}$ é então

$$\begin{aligned} S(t;\mathbf{z}) &= P(T > t) = P(\log T > \log t) \\ &= P(\mu + \boldsymbol{\alpha}'\mathbf{z} + \sigma\varepsilon > \log t) \\ &= P(\varepsilon > (\log t - \mu - \boldsymbol{\alpha}'\mathbf{z})/\sigma) \\ &= P\left[\exp(\varepsilon) > \exp\left(\frac{\log t - \mu - \boldsymbol{\alpha}'\mathbf{z}}{\sigma}\right)\right] \\ &= \exp\left[-\exp\left(\frac{\log t - \mu - \boldsymbol{\alpha}'\mathbf{z}}{\sigma}\right)\right]. \end{aligned}$$

Ora, esta última expressão pode ser reescrita como

$$S(t;\mathbf{z}) = \exp[-\exp(-\mu/\sigma)t^{1/\sigma}\exp((-\boldsymbol{\alpha}/\sigma)'\mathbf{z})]. \tag{5.3}$$

Comparando (5.1) com (5.3), facilmente concluimos que a relação entre os parâmetros do modelo Weibull, representado na forma de modelo de riscos proporcionais e como modelo log-linear, é a seguinte:

$$\lambda = \exp(-\mu/\sigma), \qquad \gamma = 1/\sigma \qquad e \qquad \beta_j = -\alpha_j/\sigma.$$

Modelo de regressão log-logístico

Há obviamente situações em que o modelo Weibull não é adequado para modelar o tempo de vida e onde o modelo log-logístico pode ser uma boa alternativa. Conforme referimos, o modelo log-logístico é o único modelo que pode ser representado como modelo de possibilidades proporcionais e que também admite uma representação na forma log-linear como modelo de tempo de vida acelerado.

Comecemos por considerar o modelo log-logístico como modelo de possibilidades proporcionais. A função de sobrevivência de um indivíduo com vector de covariáveis $\mathbf{z}$ é dada por

$$S(t;\mathbf{z}) = \frac{1}{1+\lambda\exp(\boldsymbol{\beta}'\mathbf{z})t^{\kappa}}, \tag{5.4}$$

ou seja, o tempo de vida de um indivíduo com vector de covariáveis $\mathbf{z}$ segue uma distribuição log-logística com parâmetro de escala $\lambda\exp(\boldsymbol{\beta}'\mathbf{z})$ e parâmetro de forma κ. Então

$$\frac{S(t;\mathbf{z})}{1-S(t;\mathbf{z})} = \frac{1}{\lambda\exp(\boldsymbol{\beta}'\mathbf{z})t^{\kappa}},$$

ou seja,

$$\frac{S(t;\mathbf{z})}{1-S(t;\mathbf{z})} = \exp(-\boldsymbol{\beta}'\mathbf{z})\frac{S_0(t)}{1-S_0(t)},$$

onde $S_0(t) = (1+\lambda t^{\kappa})^{-1}$ é a função de sobrevivência da distribuição log-logística de parâmetros λ e κ.

Escrevendo agora o modelo log-logístico na forma log-linear vem que

$$\log T = \mu + \boldsymbol{\alpha}'\mathbf{z} + \sigma\varepsilon,$$

onde ε segue uma distribuição logística com função densidade de pro-

babilidade dada por

$$f(\varepsilon) = \frac{\exp(\varepsilon)}{(1+\exp(\varepsilon))^2}, \qquad -\infty < \varepsilon < \infty.$$

Ora, como vimos anteriormente, para qualquer modelo log-linear a função de sobrevivência de T, dado o vector de covariáveis $\mathbf{z}$ é

$$S(t;\mathbf{z}) = P\left[\exp(\varepsilon) > \exp\left(\frac{\log t - \mu - \boldsymbol{\alpha}'\mathbf{z}}{\sigma}\right)\right]. \tag{5.5}$$

Seja $X = \exp(\varepsilon)$. A função densidade de probabilidade de X é dada por $f(x) = (1+x)^{-2}$, $x > 0$, ou seja, X tem uma distribuição log-logística de valor médio 0 e parâmetro de escala 1. Então, por (5.5), a função de sobrevivência de T, dado $\mathbf{z}$, pode escrever-se como

$$S(t;\mathbf{z}) = \left[1 + \exp\left(\frac{\log t - \mu - \boldsymbol{\alpha}'\mathbf{z}}{\sigma}\right)\right]^{-1}. \tag{5.6}$$

Comparando (5.4) com (5.6), concluimos que a relação entre os parâmetros do modelo log-logístico, representado na forma de modelo de possibilidades proporcionais e como modelo log-linear, é a seguinte:

$$\lambda = \exp(-\mu/\sigma), \qquad \kappa = 1/\sigma \qquad e \qquad \beta_j = -\alpha_j/\sigma.$$

Construção da função de verosimilhança

Vamos indicar qual a forma geral da função de verosimilhança num modelo de regressão paramétrico, estando os indivíduos sujeitos a um mecanismo de censura independente.

Consideremos que se encontram em estudo n indivíduos e que os dados correspondentes ao i-ésimo indivíduo são da forma $(t_i, \delta_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, onde t_i é tempo de vida ($\delta_i = 1$) ou tempo de censura ($\delta_i = 0$) e $\mathbf{z}_i$ é um vector de covariáveis fixas. Suponhamos que a distribuição do tempo de vida T dado $\mathbf{z}$ é conhecida a menos de um

vector de parâmetros $\boldsymbol{\theta}$, sobre o qual desejamos realizar inferência e que a função de sobrevivência para o i-ésimo indivíduo é $S(t_i; \mathbf{z}_i, \boldsymbol{\theta})$, com a correspondente função densidade de probabilidade $f(t_i; \mathbf{z}_i, \boldsymbol{\theta})$. Estando os indivíduos sujeitos a um mecanismo de censura independente e não informativa, a função de verosimilhança é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \mathbf{z}_i, \boldsymbol{\theta})^{\delta_i} S(t_i; \mathbf{z}_i, \boldsymbol{\theta})^{1-\delta_i}. \tag{5.7}$$

Sob condições de regularidade bastante gerais nos processos de morte e censura, o estimador de máxima verosimilhança $\hat{\boldsymbol{\theta}}$ tem distribuição assintótica normal multivariada com valor médio $\boldsymbol{\theta}$ e matriz de covariância $I(\boldsymbol{\theta})^{-1}$, sendo $I(\boldsymbol{\theta})$ a matriz de informação de Fisher.

Notemos que no caso de um modelo log-linear, a função densidade de probabilidade e a função de sobrevivência de $Y = \log T$ podem ser escritas como

$$f(y; \mathbf{z}) = \frac{1}{\sigma} f_\varepsilon \left(\frac{y - \mu - \boldsymbol{\alpha}'\mathbf{z}}{\sigma} \right)$$

e

$$S(y; \mathbf{z}) = S_\varepsilon \left(\frac{y - \mu - \boldsymbol{\alpha}'\mathbf{z}}{\sigma} \right),$$

onde f_ε e S_ε designam a função densidade de probabilidade e a função de sobrevivência da v.a. ε, respectivamente. Então, a função de verosimilhança (5.7) é dada por

$$L = \prod_{i=1}^{n} \left[\frac{1}{\sigma} f_\varepsilon \left(\frac{y_i - \mu - \boldsymbol{\alpha}'\mathbf{z}_i}{\sigma} \right) \right]^{\delta_i} \left[S_\varepsilon \left(\frac{y_i - \mu - \boldsymbol{\alpha}'\mathbf{z}_i}{\sigma} \right) \right]^{1-\delta_i}.$$

Para a maior parte dos modelos paramétricos e, em particular, para os modelos Weibull e log-logístico, é necessário recorrer a métodos numéricos para obtenção das estimativas de máxima verosimilhança dos parâmetros, o que actualmente é feito utilizando *software* estatístico.

Critério de Informação de Akaike

A escolha do modelo paramétrico mais apropriado, de entre vários modelos possíveis não necessariamente aninhados, pode ser feita com base no Critério de Informação de Akaike dado por

$$AIC = -2\log\hat{L} + 2(p+1+k),$$

onde $\hat{L}$ representa a verosimilhança maximizada e p é o número de parâmetros de regressão do modelo ajustado, $k = 0$ para o modelo exponencial e $k = 1$ para os modelos Weibull, log-logístico e log-normal. Quanto menor for o valor da estatística AIC, melhor é o modelo.

Análise de resíduos

Também no caso dos modelos paramétricos a análise de resíduos constitui um bom método de diagnóstico da adequabilidade do modelo. Os resíduos de Cox-Snell, que permitem uma verificação do ajustamento global do modelo, são definidos por $r_i = \hat{H}(t_i; \mathbf{z}_i)$ para o i-ésimo indivíduo a que corresponde o tempo t_i e o vector de covariáveis $\mathbf{z}_i$, sendo $\hat{H}$ a função de risco cumulativa estimada para o modelo ajustado. Tem-se então que

Modelo Weibull:	$r_i = \hat{\lambda} t_i^{\hat{\gamma}} \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_i)$
Modelo log-logístico	$r_i = \log[1 + \hat{\lambda} t_i^{\hat{\kappa}} \exp(\hat{\boldsymbol{\beta}}' \mathbf{z}_i)]$

Uma abordagem alternativa, mas equivalente, consiste na obtenção dos resíduos padronizados baseados na representação log-linear do modelo. Definem-se então, por analogia com os resíduos utilizados na regressão linear, como

$$s_i = \frac{\log t_i - \hat{\mu} - \hat{\boldsymbol{\alpha}}' \mathbf{z}_i}{\hat{\sigma}}.$$

Se o modelo Weibull for adequado, os resíduos padronizados devem comportar-se, aproximadamente, como uma amostra proveniente da distribuição de Gumbel (5.2); para o modelo log-logístico, os resíduos devem constituir uma amostra proveniente da distribuição logística padrão.

Capítulo 6

Riscos competitivos

6.1 Introdução

Nos capítulos anteriores, considerámos a existência de um único acontecimento de interesse. No entanto, existem situações nas quais vários acontecimentos se podem considerar igualmente relevantes para o problema em estudo.

Consideremos, por exemplo, um estudo envolvendo doentes com cancro da mama, em fase inicial, submetidas a radioterapia. Neste caso, a ocorrência de recaída local, o aparecimento de metástases à distância e a morte por outra causa são todos acontecimentos de interesse clínico. Para a análise dos dados obtidos em situações semelhantes à descrita, são necessários métodos que entrem em consideração com a existência de causas competitivas de "morte". O termo riscos competitivos aplica-se, então, a qualquer situação na qual um indivíduo esteja sujeito (exposto) a duas ou mais causas de morte distintas.

Este problema pode ser formulado de duas maneiras diferentes: uma delas será em termos dos tempos de vida potenciais ou latentes associados a cada causa de morte (conceito que foi introduzido apenas por conveniência matemática), a outra consiste em descrever o problema em termos das funções de risco específicas da causa. É esta a abordagem habitualmente utilizada.

6.2 Funções específicas da causa e seus estimadores

Suponhamos então que uma população está sujeita a m causas de morte. Quando ocorre uma morte, observamos o tempo de vida T e a causa da morte $J, J \in \{1, 2, \ldots, m\}$. A abordagem proposta por Prentice *et al.* (1978) consiste em descrever o problema em termos das funções de risco específicas da causa.

A função de risco específica da causa j ($j = 1, \ldots, m$) é definida por

$$\lambda_j(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt, J = j | T \geq t)}{dt}$$

e descreve a probabilidade instantânea de morte devida à causa j no instante t, na presença das outras causas de morte. Obviamente que a função de risco global

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt},$$

satisfaz a relação

$$\lambda(t) = \sum_{j=1}^{m} \lambda_j(t).$$

Logo, a função de sobrevivência do tempo de vida T pode ser representada por

$$S(t) = \exp\left(- \sum_{j=1}^{m} \int_0^t \lambda_j(u) du \right).$$

A função de sobrevivência específica da causa j é dada por

$$P(T > t, J = j) = \int_t^{\infty} \lambda_j(u) S(u) du.$$

Podemos ainda definir a função de incidência cumulativa da causa j como sendo

$$I_j(t) = P(T \leq t, J = j) = \int_0^t \lambda_j(u)S(u)du.$$

Seja

$$S_j(t) = \exp\left(-\int_0^t \lambda_j(u)du\right) \qquad j = 1, \ldots, m.$$

Notemos que estas funções não podem, geralmente, ser interpretadas como funções de sobrevivência para $m > 1$.

A utilização do estimador de Kaplan-Meier, considerando como censuradas as observações correspondentes aos indivíduos a quem ocorreu outro que não o "acontecimento de interesse", pode levar a uma estimativa da incidência cumulativa que sobrestima substancialmente a incidência do referido acontecimento.

Como exemplo, consideremos um estudo envolvendo pacientes infectados com o VIH com o fim de analisar o tempo de latência da SIDA. Ora, considerar como censurados aqueles que morrem sem desenvolver uma das doenças indicadoras de SIDA, conduz a uma sobrestimação da incidência cumulativa da doença. A abordagem correcta consiste em considerar a morte sem o desenvolvimento de doença como uma causa competitiva. De facto, os doentes que morrem antes de desenvolver a doença, não estão sujeitos a um mecanismo de censura independente porque sabemos que nunca desenvolverão a doença.

Assim sendo, o estimador de Kaplan-Meier foi generalizado de forma a ser utilizado num problema de riscos competitivos. Consideremos então os instantes de morte distintos $t_{j1} < t_{j2} < \ldots < t_{jk_j}$ onde k_j é o número de instantes de morte de tipo j para $j = 1, \ldots, m$. Suponhamos ainda que a falha de tipo j ou morte devida à causa j

ocorre d_{ji} vezes no instante $t_{ji}, i = 1, \ldots, k_j$. O estimador de máxima verosimilhança não paramétrico de S_j é

$$\hat{S}_j(t) = \prod_{i:t_{ji} \leq t} \left(1 - \frac{d_{ji}}{n_{ji}}\right),$$

em que n_{ji} é o número de indivíduos em risco no instante t_{ji}. Ignorando as causas de morte, o estimador de Kaplan-Meier da função de sobrevivência global é $\hat{S}(t) = \prod_{j=1}^{m} \hat{S}_j(t)$, desde que não haja observações empatadas entre instantes de morte de tipos diferentes. O estimador da função de incidência cumulativa da causa j para $j = 1, \ldots, m$ é dado por

$$\hat{I}_j(t) = \sum_{i:t_{ji} \leq t} \frac{d_{ji}}{n_{ji}} \hat{S}(t_{ji}). \tag{6.1}$$

Suponhamos agora que a cada indivíduo está associado um vector de covariáveis $\mathbf{z}$. O modelo de Cox pode ser generalizado de modo a permitir a sua aplicação em problemas de riscos competitivos, sendo a função de risco específica da causa j dada por

$$h_j(t; \mathbf{z}) = h_{0j}(t) \exp(\boldsymbol{\beta_j}' \mathbf{z}).$$

Kalbfleisch e Prentice (1980) obtiveram uma função de verosimilhança parcial para a estimação dos vectores $\boldsymbol{\beta_j}, j = 1, \ldots, m$. Os parâmetros $\boldsymbol{\beta_j}$ são estimados separadamente para cada causa de morte, considerando-se que as mortes devidas às restantes causas dão origem a observações censuradas.

6.3 Tempos de vida latentes

Suponhamos agora que o problema de riscos competitivos é formulado em termos dos tempos de vida latentes $T_1, \ldots, T_m$ correspondentes às causas $1, \ldots, m$. Vários autores definem T_j como sendo o

tempo de vida, correspondente à causa j, que seria observado se as restantes causas de morte fossem eliminadas. Obviamente, o vector $(T_1, \ldots, T_m)$ não é observável. Para cada indivíduo apenas são observados $T = \min(T_1, \ldots, T_m)$ e a causa j tal que $T_j = T$. É necessário admitir então uma função de sobrevivência conjunta

$$S(t_1, \ldots, t_m) = P(T_1 \geq t_1, \ldots, T_m \geq t_m).$$

Deste modo, hipóteses diferentes acerca da distribuição conjunta de $(T_1, \ldots, T_m)$ produzem modelos de riscos competitivos diferentes.

Admitimos que o efeito de eliminar causas de morte é conhecido e pode ser expresso em termos de $S(t_1, \ldots, t_m)$. A hipótese mais usual é a de que o efeito de eliminar a causa j pode ser representado pela anulação do correspondente argumento t_j em $S(t_1, \ldots, t_m)$. Portanto, a função de sobrevivência marginal de T_j é $S_j(t) = P(T_j \geq t), j = 1, \ldots, m$, onde $S_j(t) = S(0, \ldots, t_j, \ldots, 0)$. A correspondente função de risco marginal é dada por $g_j(t) = -d \log S_j(t)/dt$ e representa a função de risco associada à causa j na ausência das outras causas de morte.

A função de sobrevivência de T é $S(t) = P(T \geq t) = S(t, \ldots, t)$, sendo $\lambda(t) = -d \log S(t)/dt$ a função de risco correspondente.

A função de risco específica da causa j $(j = 1, \ldots, m)$ é dada por

$$\lambda_j(t) = -\frac{\partial \log S(t_1, \ldots, t_m)}{\partial t_j}\bigg|_{t_1 = \ldots = t_m = t}.$$

Dado que $\lambda(t) = \sum_{j=1}^{m} \lambda_j(t)$, $S(t)$ pode ser representada como

$$\begin{aligned} S(t) &= \exp\left(-\sum_{j=1}^{m} \int_0^t \lambda_j(u)du\right) \\ &= \prod_{j=1}^{m} \exp\left(-\int_0^t \lambda_j(u)du\right). \end{aligned}$$

Notemos que, sem hipóteses adicionais, as funções $S_j(t)$ e $g_j(t)$ não podem ser expressas em termos da função de risco específica da causa $\lambda_j(t)$, pelo que as distribuições marginais são não identificáveis. Este problema da não identificabilidade, como foi designado por Tsiatis (1975), surge dado que o par observável (T, J) não determina de modo únivoco a distribuição conjunta de $(T_1, \ldots, T_m)$.

A hipótese de independência das causas de morte no modelo de riscos competitivos corresponde a supôr que $T_1, \ldots, T_m$ são v.a. independentes. Então, neste caso,

$$S(t) = \prod_{j=1}^{m} S_j(t),$$

donde se conclui que $g_j(t) = \lambda_j(t)$. Logo, sob a hipótese de independência, as funções de risco associadas à causa j na presença e na ausência das restantes causas de morte são iguais.

É de salientar que Prentice *et al.* (1978), entre outros, criticam esta abordagem que consideram irrealista por supor que podem ser desenvolvidos métodos estatísticos que levem em conta todos os mecanismos complexos envolvidos na eliminação de uma causa de morte.

Capítulo 7

Modelos com fragilidade

7.1 Introdução

Vamos começar por tecer algumas considerações sobre as hipóteses de independência e homogeneidade que, de forma mais ou menos implícita, estão subjacentes à maior parte dos métodos utilizados na análise estatística de tempos de vida, em particular aos que foram referidos anteriormente.

De um modo geral, supomos que, condicional aos valores das covariáveis, os tempos de vida dos indivíduos em estudo são independentes. Esta hipótese de independência aplica-se habitualmente aos tempos até à ocorrência do mesmo acontecimento para indivíduos diferentes, o que é incorrecto em determinadas situações. De facto, é natural supôr que exista associação entre os tempos de vida de indivíduos que partilham factores genéticos não observáveis (e.g., irmãos) ou condições ambientais não quantificáveis, tais como dieta, nível de poluição doméstica, situações de *stress*, etc.(e.g., cônjuges).

Quando se trata dos tempos até à ocorrência de diferentes acontecimentos para o mesmo indivíduo, a hipótese de independência também é questionável. Como exemplos, podemos referir a primeira ocorrência de doença coronária e de acidente vascular cerebral ou a primeira

evidência de hipertensão e de intolerância à glucose. Em ambos os casos, é razoável admitir que exista associação entre os tempos de vida correspondentes ao mesmo indivíduo.

Em análise de dados de sobrevivência, é também usual admitir a existência de homogeneidade entre os indivíduos que apresentam valores comuns das covariáveis observadas. No entanto, esta hipótese é, com frequência, pouco realista, dada a impossibilidade prática de registar todos os factores de risco relevantes. Com efeito, no decorrer de um estudo clínico, é frequente constatar que os indivíduos diferem entre si na evolução natural de determinada doença, na sua reacção a um mesmo tratamento ou no modo como são influenciados por vários factores de risco, apesar de constituirem um grupo homogéneo relativamente às covariáveis consideradas. Esta situação reflecte a existência de uma heterogeneidade individual, não observável, mas que é extremamente importante e deve ser tomada em consideração na interpretação dos resultados obtidos. Com efeito, a heterogeneidade pode explicar alguns resultados inesperados ou fornecer uma explicação alternativa em algumas situações como, por exemplo, quando se observam funções de risco não proporcionais ou decrescentes.

Uma explicação possível para as situações que referimos é a existência de covariáveis que não foram incluídas no modelo porque não dispomos de informação acerca dos seus valores individuais ou mesmo porque desconhecemos a sua existência. De facto, podemos argumentar que há sempre um grande número de variáveis que, caso pudessem ser medidas, dariam informação suficiente para explicar as diferenças individuais.

A necessidade de desenvolver modelos de sobrevivência adequados às situações acima descritas levou ao aparecimento de modelos de efeitos aleatórios, denominados modelos com fragilidade.

Os modelos que iremos descrever, embora sejam representações bastante simplificadas da forma como a heterogeneidade pode actuar, contribuem de forma significativa para uma compreensão sistemática deste problema.

Pretendemos, neste capítulo, apresentar o modelo multiplicativo com fragilidade, bem como as motivações que levaram ao seu desenvolvimento. O modelo será descrito apenas no contexto univariado, ou seja, para modelação da heterogeneidade não observada. Referimos também diversos exemplos práticos de aplicação deste modelo.

7.1.1 O modelo multiplicativo

O termo fragilidade foi introduzido por Vaupel *et al.* (1979) para designar uma variável não observada que descreve factores de risco, desconhecidos ou que não se podem medir, não incluídos no modelo. Esta designação pode ser justificada pelo facto de que valores elevados desta variável correspondem a uma diminuição do tempo de vida do indivíduo, ou seja, a um aumento da função de risco em todo o intervalo de tempo considerado.

Com o objectivo de desenvolver métodos de tabelas de mortalidade para populações cujos membros diferiam na sua susceptibilidade geral às causas de morte, Vaupel *et al.* (1979) propuseram um modelo multiplicativo em que a função de risco no instante t para um indivíduo com fragilidade $W = w$ é

$$\mu(t|w) = w\lambda(t) \tag{7.1}$$

onde W é uma v.a. não negativa e $\lambda(t)$ é uma função do tempo comum a todos os indivíduos e independente de W.

Para modelar a heterogeneidade existente na população em estudo

admite-se que a variável fragilidade segue uma determinada distribuição. É habitual supôr que $E(W) = 1$, o que pode ser justificado pelas razões seguintes: Vaupel *et al.* (1979) e Manton *et al.* (1986) designam $\lambda(t)$ por força de mortalidade padrão, i.e., correspondente a um indivíduo com fragilidade $w = 1$. Por outro lado, Aalen (1988) refere que, considerando que $E(W) = 1$, $\lambda(t)$ pode ser encarada como uma função de risco individual "média", visto que, por (7.1), $\lambda(t)$ é nesse caso o valor esperado da função de risco individual.

O modelo desenvolvido por Vaupel *et al.* (1979) destina-se a uma situação típica de tabelas de mortalidade, em que os únicos factores de risco conhecidos são o sexo e a idade de cada indivíduo. Quando dispomos de informação sobre outros factores e queremos modelar o efeito de uma covariável não observável podemos considerar uma extensão de um modelo de regressão como, por exemplo, a seguinte extensão do modelo de Cox (1972) para inclusão de heterogeneidade não observada e não explicada pelas covariáveis observadas:

$$h(t|\mathbf{z}, w) = w\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}). \tag{7.2}$$

Uma outra aplicação importante dos modelos com fragilidade surge na análise de dados de sobrevivência multivariados. Como referimos anteriormente, pode haver dependência entre os tempos de vida de indivíduos pertencentes a uma mesma família ou ninhada. Para modelar esta dependência admite-se que os indivíduos que estão relacionados partilham uma mesma fragilidade, que se entende como um efeito aleatório comum, ambiental e/ou genético.

Em tais situações, podemos considerar uma generalização do modelo (7.2) em que, condicional ao valor da fragilidade comum w_i, a função de risco para um indivíduo j pertencente ao i-ésimo grupo, com vector de covariáveis observadas $\mathbf{z}_{ij}$ é

$$h(t_{ij}|\mathbf{z}_{ij}, w_i) = w_i h_0(t_{ij}) \exp(\boldsymbol{\beta}'\mathbf{z}_{ij}).$$

7.1.2 Resultados básicos

Consideremos o modelo em que a função de risco individual é dada por (7.1). Então, a função de sobrevivência condicional de T dado $W = w$ é

$$S(t|w) = \exp[-w\Lambda(t)] = \{S_0(t)\}^w,$$

onde $\Lambda(t) = \int_0^t \lambda(u)du$ é a função de risco cumulativa para um indivíduo com fragilidade $w = 1$, que designaremos por função de risco cumulativa subjacente. Portanto, $S_0(t) = \exp[-\Lambda(t)]$ é a correspondente função de sobrevivência subjacente, associada a esse mesmo indivíduo padrão.

As funções $\mu(t|w)$ e $S(t|w)$ correspondem a um modelo individual que não é observável, como refere Aalen (1988). A função de sobrevivência e a função de risco populacionais, correspondentes aos dados de que efectivamente dispomos, são respectivamente, (Hougaard, 1984)

$$\begin{aligned} S(t) &= \int_0^\infty S(t|w) f(w) dw \\ &= \int_0^\infty \exp[-w\Lambda(t)] f(w) dw \\ &= L\{\Lambda(t)\} \end{aligned} \tag{7.3}$$

e

$$h(t) = -\frac{L'\{\Lambda(t)\}}{L\{\Lambda(t)\}}\lambda(t), \tag{7.4}$$

onde $L(s)$ é a transformada de Laplace da distribuição da fragilidade W, $F(w)$ é a correspondente função de distribuição e $L'(s) = dL(s)/ds$.

De (7.3) obtem-se a relação $\Lambda(t) = L^{-1}(S(t))$, cuja resolução pode requerer a utilização de métodos numéricos.

Suponhamos que a fragilidade W é uma v.a. com função densidade de probabilidade $f(w)$. Vamos obter agora a distribuição condicional de W dado $T \geq t$, i.e., a distribuição da fragilidade entre os sobreviventes no instante t ou à idade t. A função densidade de probabilidade correspondente será dada, para $t \geq 0$, por (Hougaard, 1984)

$$\begin{aligned} g(w|T \geq t) &= \frac{\int_t^\infty f(u|w)f(w)du}{P(T \geq t)} \\ &= \frac{\exp[-w\Lambda(t)]f(w)}{L\{\Lambda(t)\}}. \end{aligned}$$

Notemos que a definição de fragilidade que utilizamos pressupõe que cada indivíduo nasce com uma certa fragilidade, a qual não se altera durante toda a sua vida (Vaupel *et al.*, 1979). Portanto, a distribuição da fragilidade coincide com a distribuição da fragilidade entre os sobreviventes no instante $t = 0$, ou seja, com a distribuição da fragilidade à nascença.

A distribuição condicional de W dado $T \geq t$ pode igualmente ser caracterizada pela correspondente transformada de Laplace, que é dada por

$$L_1(s) = E(e^{-sW}|T \geq t) = \frac{L\{s + \Lambda(t)\}}{L\{\Lambda(t)\}}. \tag{7.5}$$

A função densidade de probabilidade de W dado $T = t$, i.e., da distribuição da fragilidade entre os indivíduos que morrem no instante t é (Hougaard, 1984)

$$\begin{aligned} g(w|t) &= \frac{w\lambda(t)\exp[-w\Lambda(t)]f(w)}{-L'\{\Lambda(t)\}\lambda(t)} \\ &= \frac{w\exp[-w\Lambda(t)]f(w)}{-L'\{\Lambda(t)\}}. \end{aligned}$$

Quanto ao valor médio de W dado $T \geq t$, ou seja, a fragilidade média entre os sobreviventes no instante t, é dado por

$$E(W|T \geq t) = -\frac{L'\{\Lambda(t)\}}{L\{\Lambda(t)\}}. \tag{7.6}$$

Então, a função de risco populacional (7.4) pode ser escrita na seguinte forma

$$h(t) = E(W|T \geq t)\lambda(t),$$

a qual evidencia que a função de risco observável $h(t)$ é o valor esperado, entre os sobreviventes no instante t, da verdadeira função de risco $\mu(t|w)$.

A fragilidade média na população sobrevivente decresce com o tempo, visto que os indivíduos que possuem maior fragilidade irão morrer mais cedo do que os outros. De facto, por (7.6) tem-se que

$$\begin{aligned}\frac{\partial E(W|T \geq t)}{\partial t} &= \lambda(t)\frac{\{L'[\Lambda(t)]\}^2 - L''[\Lambda(t)]L[\Lambda(t)]}{\{L[\Lambda(t)]\}^2} \\ &= -\lambda(t)E(W^2|T \geq t) + \lambda(t)\{E(W|T \geq t)\}^2 \\ &= -\lambda(t)var(W|T \geq t) < 0.\end{aligned}$$

Comparando então

$$\mu(t|w) = w\lambda(t)$$

e

$$h(t) = E(W|T \geq t)\lambda(t),$$

podemos concluir que se $\lambda(t)$ for crescente, a função de risco individual cresce mais rapidamente do que a função de risco populacional. Por outro lado, se $\lambda(t)$ for decrescente, então a função de risco populacional (que é a função de risco observada) decresce mais rapidamente

do que a função de risco individual. Portanto, a existência de heterogeneidade não observada é um factor que pode enviesar as conclusões acerca da evolução da mortalidade na população.

É frequente o uso do coeficiente de variação como medida da dispersão de uma distribuição. Nos modelos com fragilidade, o coeficiente de variação é utilizado como medida do grau de heterogeneidade da população. Vejamos porquê. Neste tipo de modelos, interessa-nos estudar a dispersão da distribuição da fragilidade entre os sobreviventes, visto que nos permite tirar conclusões sobre a heterogeneidade existente nessa população num dado momento. Com efeito, uma menor dispersão corresponde a uma distribuição da fragilidade mais concentrada e portanto a uma população em que os indivíduos apresentam valores da fragilidade muito semelhantes. Portanto, uma menor dispersão é indicador de uma população mais homogénea.

Assim, o estudo do coeficiente de variação da distribuição da fragilidade entre os sobreviventes como função do tempo, quanto à sua monotonia, permite avaliar a evolução ao longo do tempo da heterogeneidade na população sobrevivente.

7.2 Escolha da distribuição da fragilidade

Uma questão que naturalmente se coloca, ao reconhecermos a necessidade de utilização de um modelo que inclua heterogeneidade não observada, é a da escolha da distribuição da fragilidade, de entre as que possuem suporte em $[0, \infty)$. Para tal, devemos ter em conta o seguinte:

- Dado que é razoável admitir que a heterogeneidade não observada é devida a um grande número de factores, é preferível considerar que a fragilidade é uma v.a. contínua, excepto num caso

particular. De facto, quando a heterogeneidade é provocada pela existência na população de um grupo de indivíduos que apresenta susceptibilidade nula a determinado acontecimento, o modelo adequado para a fragilidade será uma distribuição de tipo misto, com massa de probabilidade não nula em zero e que varie de modo contínuo em $(0, \infty)$.

- É vantajoso que a família de distribuições seja fechada para a selecção induzida pela mortalidade, ou seja, que a distribuição condicional da variável fragilidade, dada a sobrevivência até determinado instante, pertença ainda à família de distribuições da fragilidade.

Em seguida, vamos indicar algumas distribuições que têm sido estudadas por vários autores como modelos para a fragilidade, referindo também as propriedades que mostram a sua relevância na aplicação a populações em que existe heterogeneidade não observada:

- distribuição gama (Vaupel *et al.*, 1979; Lancaster e Nickell, 1980; Manton *et al.*, 1981, 1986; Hougaard, 1984, 1991; Aalen, 1987, 1988; Klein, 1992)
- distribuição Gaussiana inversa (Hougaard, 1984, 1991; Manton *et al.*, 1986; Aalen, 1988)
- distribuição de Poisson composta (Aalen, 1988, 1992)

7.2.1 Distribuição gama

A distribuição gama tem sido a mais utilizada como distribuição da fragilidade, em parte devido às vantagens que apresenta do ponto de vista do tratamento matemático. Suponhamos então que a fragilidade W segue uma distribuição gama com função densidade de

probabilidade

$$f(w) = \theta^{\delta} w^{\delta-1} \exp(-\theta w)/\Gamma(\delta), \qquad w > 0$$

onde $\theta > 0, \delta > 0$. A correspondente transformada de Laplace é dada por

$$L(s) = \left(\frac{\theta}{\theta + s}\right)^{\delta} = (1 + s/\theta)^{-\delta}.$$

Então, por (7.3) e (7.4), a função de sobrevivência e a função de risco populacionais são dadas respectivamente por

$$S(t) = \left(\frac{\theta}{\theta + \Lambda(t)}\right)^{\delta}$$

e

$$h(t) = \frac{\delta}{\theta + \Lambda(t)} \lambda(t).$$

Distribuição da fragilidade entre os indivíduos sobreviventes no instante t

Para a distribuição condicional de W dado $T \geq t$, obtemos a seguinte função densidade de probabilidade:

$$g(w|T \geq t) = \frac{(\theta + \Lambda(t))^{\delta}}{\Gamma(\delta)} w^{\delta-1} \exp\{-(\theta + \Lambda(t))w\} \qquad w > 0.$$

Portanto, entre os sobreviventes no instante t, W tem uma distribuição gama de parâmetros δ e $\theta + \Lambda(t)$, i.e., a distribuição da fragilidade entre os sobreviventes num dado instante é novamente gama com o mesmo parâmetro de forma mas um parâmetro de escala diferente.

Temos também que $E(W|T \geq t) = \delta/(\theta + \Lambda(t))$. Então, o coeficiente de variação entre os sobreviventes é constante e igual a $\delta^{-1/2}$. Portanto, o grau de heterogeneidade existente na população sobrevivente não se altera ao longo do tempo.

Distribuição da fragilidade entre os mortos no instante t

Se W seguir uma distribuição gama de parâmetros δ e θ, então a distribuição da fragilidade entre os mortos no instante t é ainda gama, de parâmetros $\delta + 1$ e $\theta + \Lambda(t)$.

7.2.2 Distribuição Gaussiana inversa

Suponhamos que a fragilidade W segue uma distribuição Gaussiana inversa com parâmetros $\theta > 0$ e $\mu > 0$, cuja função densidade de probabilidade é dada por

$$f(w) = (\mu/\pi)^{1/2} \exp[(4\mu\theta)^{1/2}] w^{-3/2} \exp(-\theta w - \mu/w) \qquad w > 0.$$

A sua transformada de Laplace é

$$L(t) = \exp\{(4\mu\theta)^{1/2} - [4\mu(\theta + t)]^{1/2}\}.$$

O valor médio da distribuição é $(\mu/\theta)^{1/2}$ e $\text{var}(W) = \frac{1}{2}\mu^{1/2}\theta^{-3/2}$. Portanto, o coeficiente de variação é igual a $2^{-1/2}(\mu\theta)^{-1/4}$.

Então, por (7.3) e (7.4), a função de sobrevivência e a função de risco populacionais são dadas respectivamente por

$$S(t) = \exp\{(4\mu\theta)^{1/2} - [4\mu(\theta + \Lambda(t))]^{1/2}\}$$

e

$$h(t) = [\mu/(\theta + \Lambda(t))]^{1/2} \lambda(t).$$

Distribuição da fragilidade entre os indivíduos sobreviventes no instante t

A distribuição condicional de W dado $T \geq t$ tem a seguinte função densidade de probabilidade:

$$g(w|T \geq t)$$
$$= (\mu/\pi)^{1/2} \exp\{[4\mu(\theta + \Lambda(t))]^{1/2}\} w^{-3/2} \exp\{-(\theta + \Lambda(t))w - \mu/w\}.$$

Logo, a distribuição da fragilidade entre os sobreviventes no instante t é também uma distribuição Gaussiana inversa de parâmetros $\theta + \Lambda(t)$ e μ. A fragilidade média entre os sobreviventes no instante $t, \forall t > 0$, é dada por

$$E(W|T \geq t) = [\mu/(\theta + \Lambda(t))]^{1/2}.$$

Ao propôr a utilização da distribuição Gaussiana inversa como alternativa à distribuição gama, Hougaard (1984) referiu que neste modelo a população sobrevivente torna-se mais homogénea com o decorrer do tempo, o que é consistente com uma população sujeita a um fenómeno de selecção em que os indivíduos mais frágeis vão sendo eliminados. De facto, se a distribuição de W for Gaussiana inversa de parâmetros θ e μ, o coeficiente de variação entre os sobreviventes é dado por $2^{-1/2}\{\mu(\theta + \Lambda(t))\}^{-1/4}$, portanto decresce com o tempo.

Concluimos assim que as distribuições gama e Gaussiana inversa são fechadas para a selecção induzida pela mortalidade, i.e., a distribuição da variável fragilidade, entre os sobreviventes num determinado instante, pertence ainda à família de distribuições da fragilidade.

7.2.3 Distribuição de Poisson composta

Em medicina, constata-se por vezes que alguns indivíduos na população não são susceptíveis a determinada doença, enquanto que os

restantes apresentam um grau variável de susceptibilidade, possivelmente de natureza genética. Nesta situação, é adequado considerar a variável fragilidade como uma v.a. mista com massa de probabilidade não nula em zero. Aalen (1988, 1992) propôs a utilização da distribuição de Poisson composta gerada por variáveis aleatórias gama. Esta distribuição pode ser expressa como soma de um número aleatório de variáveis aleatórias gama independentes e identicamente distribuídas, em que o número de parcelas segue uma distribuição de Poisson.

A probabilidade $P(W = 0) > 0$ corresponde à hipótese de não susceptibilidade a determinado acontecimento para um certo grupo de indivíduos. Se estes representarem a maior parte da população, a heterogeneidade pode ter um impacto considerável, ainda que o acontecimento em causa seja raro. Notemos que a função de risco populacional tem, neste caso, integral finito em $(0, \infty)$, portanto não se trata de uma função de risco no sentido usual. De facto, a função de distribuição correspondente é imprópria, pois $F(\infty) < 1$.

7.3 Escolha da função de risco subjacente

Consideremos agora a questão da escolha da distribuição condicional do tempo de vida T dado $W = w$, o que é equivalente à escolha de uma forma paramétrica para a função $\lambda(t)$. Pelas razões que adiante referimos, os modelos mais utilizados são os seguintes:

- **Exponencial**

  A função de risco subjacente é neste caso constante, portanto cada indivíduo tem uma probabilidade instantânea de morte que não se altera ao longo do tempo. Se existir heterogeneidade de risco entre os indivíduos, a população apresenta função de risco decrescente, resultante de uma forte selecção dos indi-

víduos de alto risco. Trata-se, no entanto, de uma hipótese algo restritiva, que é razoável apenas em determinadas situações.

- **Weibull**

  A utilização de uma função de risco subjacente de tipo Weibull pode por vezes ser justificada pelo facto de ser uma distribuição de valores extremos, o que a torna apropriada, por exemplo, para estudo da incidência de cancro.

- **Gompertz**

  A sua utilização tem sido justificada por várias teorias biológicas de envelhecimento. De facto, ao tentar modelar o tempo de vida de indivíduos que se encontram na fase adulta, verificou-se a necessidade de utilizar uma função de risco que apresentasse um crescimento mais rápido do que o apresentado, por exemplo, pela distribuição de Weibull.

Uma questão importante é a da sensibilidade das estimativas dos parâmetros dos modelos de heterogeneidade à escolha, de certo modo arbitrária, da distribuição da fragilidade e da função de risco subjacente.

Manton *et al.* (1986) analisaram tabelas de mortalidade utilizando quatro modelos, envolvendo as combinações da distribuição da fragilidade gama e Gaussiana inversa com uma função de risco subjacente Weibull e Gompertz. O melhor ajustamento foi obtido com o modelo gama/Weibull. Neste contexto, concluiram que as estimativas dos parâmetros eram menos sensíveis à escolha da distribuição da fragilidade do que à da função de risco subjacente. Heckman e Singer (1984b) argumentaram que, como a teoria económica oferece, por vezes, alguma orientação quanto à forma funcional da distribuição condicional de T dado $W = w$, apenas a escolha da distribuição da fragilidade é arbitrária.

7.4 Função de verosimilhança

Admitindo para a variável fragilidade uma das distribuições referidas e também um modelo paramétrico para a função de risco subjacente, teremos então, para uma amostra de dimensão n, a seguinte função de verosimilhança conjunta para T e W

$$L = \prod_{i=1}^{n} [f(t_i, w_i)]^{\delta_i} [S(t_i, w_i)]^{1-\delta_i},$$

onde t_i representa o tempo observado e w_i o valor da fragilidade para o i-ésimo indivíduo, sendo $\delta_i = 1$ se t_i é uma observação não censurada e $\delta_i = 0$ no caso contrário. Para estimação dos parâmetros da distribuição da fragilidade e da distribuição do tempo de vida, uma opção consiste em utilizar o algoritmo EM, proposto por Dempster *et al.* (1977). Para tal é necessário construir a verosimilhança completa, assim designada pelo facto dos valores da fragilidade serem encarados como observados. Esta função de verosimilhança é da forma

$$\begin{aligned} L_C &= \prod_{i=1}^{n} [\mu(t_i|w_i) S(t_i|w_i) f(w_i)]^{\delta_i} [S(t_i, |w_i) f(w_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(w_i) \prod_{i=1}^{n} \mu(t_i|w_i)^{\delta_i} S(t_i|w_i). \end{aligned}$$

7.5 Identificabilidade do modelo

A questão da identificabilidade é extremamente importante na teoria dos modelos com fragilidade, pelo que iremos referir, de forma sucinta, de que modo diversos autores têm abordado este problema no âmbito do modelo multiplicativo.

Comecemos por considerar a situação em que não existem covariáveis observadas. Como vimos anteriormente, a função de sobrevivência

populacional é da forma

$$S(t) = L(\Lambda(t)),$$

onde $L(s) = \int_0^\infty \exp(-ws)dF(w)$ e F é a função de distribuição da fragilidade W. Esta equação coloca-nos precisamente um problema de identificação. Com efeito, se não dispomos de informação prévia sobre a distribuição de W e sendo $S(t)$ a função de sobrevivência observada (e que pode ser estimada a partir dos dados), então a verdadeira função $\Lambda(t)$ pode ser qualquer função da forma

$$\Lambda(t) = L^{-1}(S(t)),$$

onde L é qualquer função completamente monótona, com $L(0) = 1$. Lancaster e Nickell (1980) apresentam um exemplo em que duas combinações de distribuições da fragilidade e funções de risco subjacentes diferentes dão origem à mesma função de sobrevivência observada. Concluem então que, na prática, parece ser muito difícil distinguir entre o efeito da heterogeneidade não observada e a dependência do tempo, especificada pela função $\lambda(t)$.

Elbers e Ridder (1982) abordam a questão da identificabilidade no contexto de um modelo de riscos proporcionais com fragilidade, em que a função de risco é da forma

$$h(t|\mathbf{z}, w) = w\lambda_0(t)\phi(\mathbf{z}, \boldsymbol{\beta}),$$

onde $\mathbf{z}$ é um vector de covariáveis constantes. Seja $F(w)$ a função de distribuição da fragilidade W e $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. Elbers e Ridder (1982) estabelecem a identificabilidade das funções ϕ, Λ_0 e F sob as hipóteses de existência do valor médio da v.a. W e de que o modelo contém pelo menos uma covariável que toma valores num certo intervalo.

Heckman e Singer (1984a) mostraram também que, para uma certa classe de modelos paramétricos, a especificação da forma funcional da função de risco subjacente, até um número finito de parâmetros desconhecidos, bem como algumas restrições nos momentos da distribuição da fragilidade, são suficientes para a identificabilidade do modelo, mesmo que não existam covariáveis observadas. Os autores demonstram este facto para uma classe que contém, entre outros, os modelos Weibull e Gompertz.

7.6 Exemplos práticos

Vamos referir nesta secção alguns exemplos de situações práticas para as quais os modelos univariados de heterogeneidade não observável são adequados.

1. **Expulsão do dispositivo intra-uterino (DIU)**

   No início, o risco de expulsão é elevado e depois decresce rapidamente com o tempo decorrido desde a inserção do DIU. Uma explicação possível é que cada mulher apresenta um risco de expulsão constante, mas o nível do risco varia muito de mulher para mulher. A explicação biológica defende que o organismo adapta-se ao uso do DIU.

   Notemos que de novo se verifica um antagonismo entre as explicações biológica e estatística. No entanto, dado que é difícil negar a existência de alguma heterogeneidade entre os indivíduos, um modelo com fragilidade será certamente um bom contributo para a compreensão do problema.

   Aalen (1987) considerou um modelo com função de risco individual constante e em que a variável fragilidade segue uma distribuição gama.

2. **Transplante renal**

Uma situação em que as funções de risco não são proporcionais é aquela em que se verifica o chamado declínio no efeito do tratamento: inicialmente o novo tratamento é superior ao tratamento habitual mas, após algum tempo, a razão das funções de risco correspondentes aos dois grupos de tratamento tende para um, portanto as funções de risco são convergentes.

Este efeito foi encontrado por Dabrowska *et al.*(1992) na análise de dados obtidos num estudo para comparação de dois medicamentos imuno-supressores usados no tratamento de doentes submetidos a transplantes renais. Em nossa opinião, o modelo (7.2) com fragilidade gama seria adequado a esta situação.

Com efeito, dado o vector de covariáveis observadas $\mathbf{z}$ se admitirmos que W tem uma distribuição gama de valor médio 1 e variância γ, a função de sobrevivência populacional é dada por

$$\begin{aligned} S(t|\mathbf{z}) &= L\{\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t)\} \\ &= (1+\gamma\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t))^{-1/\gamma}. \end{aligned}$$

A correspondente função de risco é então da forma

$$h(t|\mathbf{z}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{z})\lambda_0(t)}{1+\gamma\exp(\boldsymbol{\beta}'\mathbf{z})\Lambda_0(t)}.$$

Seja $\mathbf{z}_j, j=1,2$ o vector de covariáveis associado a um indivíduo pertencente ao grupo de tratamento j. Então, temos que

$$\begin{aligned} \lim_{t\to 0^+}\frac{h(t|\mathbf{z}_1)}{h(t|\mathbf{z}_2)} &= \lim_{t\to 0^+}\frac{\exp(\boldsymbol{\beta}'\mathbf{z}_1)[1+\gamma\exp(\boldsymbol{\beta}'\mathbf{z}_2)\Lambda_0(t)]}{\exp(\boldsymbol{\beta}'\mathbf{z}_2)[1+\gamma\exp(\boldsymbol{\beta}'\mathbf{z}_1)\Lambda_0(t)]} \\ &= \exp[\boldsymbol{\beta}'(\mathbf{z}_1-\mathbf{z}_2)], \end{aligned}$$

enquanto que, por outro lado,

$$\lim_{t\to\infty} \frac{h(t|\mathbf{z}_1)}{h(t|\mathbf{z}_2)} = 1.$$

3. **Nefropatia diabética**

   Trata-se de uma grave doença renal que provoca elevada mortalidade nos diabéticos. Dado que um grande número de diabéticos nunca desenvolve esta doença, podemos considerar que estes indivíduos apresentam susceptibilidade nula. A função de risco do tempo até à ocorrência de nefropatia cresce até um valor máximo e depois decresce para zero. Assim sendo, uma escolha natural para a fragilidade é a distribuição de Poisson composta, podendo a função de risco subjacente ser de tipo Weibull ou Gompertz.

Capítulo 8

Aplicações

8.1 Estudo do tempo até à recidiva de cancro da bexiga

Os dados utilizados neste exemplo prático encontram-se disponíveis em Collett (2003) e provêm de um ensaio clínico aleatorizado e controlado com placebo, em doentes com cancro da bexiga. O estudo teve como objectivo comparar duas abordagens terapêuticas através do estudo do tempo até à primeira recidiva, tendo os dados sido obtidos através do seguimento de 86 doentes com tumores superficiais da bexiga, submetidos previamente a cirurgia para a sua remoção. As observações censuradas correspondem a doentes para os quais não foi observada a recidiva durante o período de observação. Além do tempo até à recidiva (em meses), foram consideradas as seguintes variáveis: grupo de tratamento (placebo/quimioterapia), número inicial de tumores e diâmetro do maior tumor inicial (em cm). Após uma análise descritiva inicial, constatou-se que a mediana do número de tumores era igual a 1 (min = 1, max = 8), a mediana do diâmetro do maior tumor era igual a 1 cm (min = 1, max = 7) e que 38 (44.2%) doentes foram submetidos a quimioterapia. A mediana do tempo de seguimento dos doentes foi de 12.5 meses (min = 1, max = 59) e 47 (54.7%) tiveram recaída. A distribuição do tempo de seguimento dos

doentes que recidivaram e dos que não recidivaram pode ser observada na Figura 8.1.

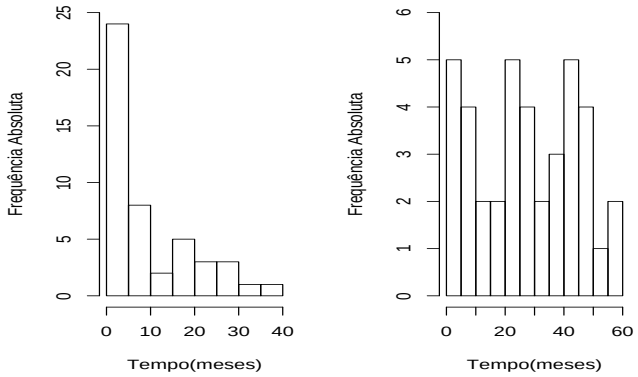


Figura 8.1 *Histogramas da distribuição do tempo de seguimento dos doentes que recidivaram (à esquerda) e dos que não recidivaram (à direita)*

8.1.1 Abordagem não paramétrica

A estimativa de Kaplan-Meier da função de sobrevivência encontra-se representada na Figura 8.2. Podemos então obter a estimativa da mediana do tempo até à recaída, que foi de 22.0 meses, sendo o intervalo de 95% de confiança (12.8, 31.2). Seguiu-se a estimação da função de risco cumulativa através do estimador de Nelson-Aalen (Figura 8.3). Ao comparar a função de sobrevivência estimada a partir do estimador de Kaplan-Meier com a obtida a partir do estimador de Nelson-Aalen (Figura 8.4),constata-se, como esperado, que as duas estimativas são muito semelhantes, embora os valores dos saltos da última função sejam ligeiramente inferiores aos da primeira (indicador

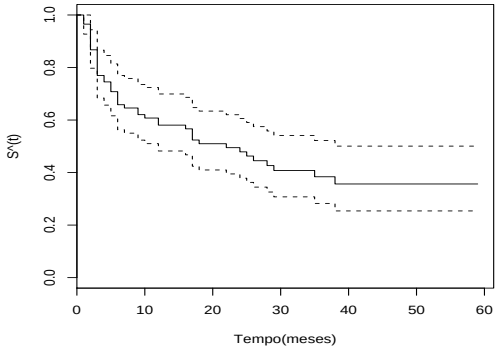


Figura 8.2 *Estimativa de Kaplan-Meier da função de sobrevivência*

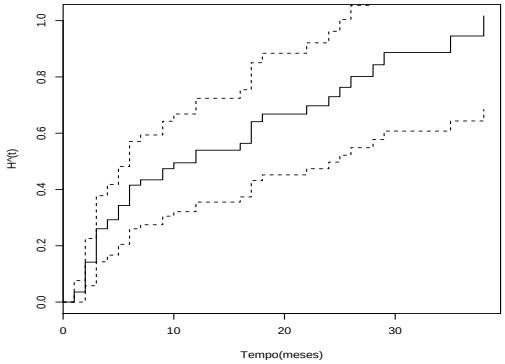


Figura 8.3 *Estimativa de Nelson-Aalen da função de risco cumulativa*

de uma menor variância).

Tendo em conta que o objectivo principal do estudo é comparar a eficácia dos dois tratamentos (cirurgia ou cirurgia seguida de quimiote-

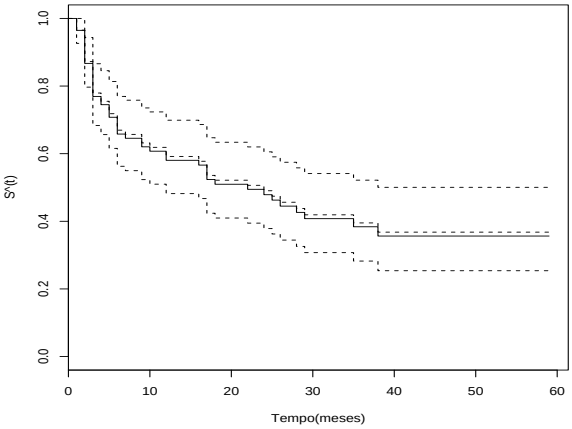


Figura 8.4 *Estimativa de Kaplan-Meier (linha contínua) com intervalos de 95% de confiança e estimativa de Nelson-Aalen para a função de sobrevivência (linha tracejada)*

rapia) no que diz respeito ao tempo livre de doença, obtivemos as curvas de sobrevivência para cada um dos grupos, utilizando o estimador de Kaplan-Meier. Como podemos observar, o gráfico da Figura 8.5 sugere que, qualquer que seja o instante considerado, a probabilidade de não recidivar pelo menos até esse instante é sempre superior no grupo da quimioterapia, com excepção de um período inicial de aproximadamente 7 meses. Para concluir se essa maior eficácia é significativa, começámos por utilizar os testes não paramétricos disponíveis no R. A partir dos resultados obtidos pelo teste log-rank (valor-p = 0.219) e pelo teste de Peto-Peto (valor-p = 0.340), será de concluir que não existe evidência de que o tempo livre de doença dependa da opção terapêutica escolhida após a cirurgia, considerando, é claro, o tratamento como único factor de prognóstico. No entanto, tendo to-

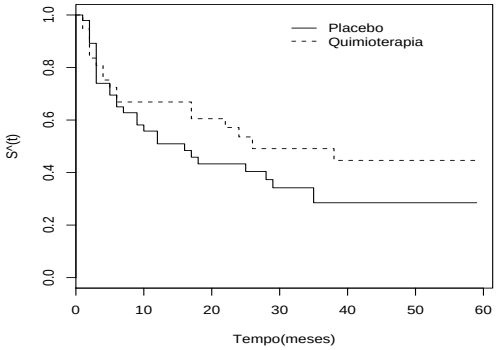


Figura 8.5 *Estimativa de Kaplan-Meier da função de sobrevivência, estratificada por tratamento*

dos os doentes sido submetidos a cirurgia para remoção dos tumores, é razoável admitir que a quimioterapia, como tratamento suplementar, não revele maior eficácia (se tal existir) a curto prazo. De facto, pela observação do gráfico da Figura 8.5, é visível o período inicial já referido em que as curvas de sobrevivência estão praticamente sobrepostas. Portanto, sería mais adequado, nesta situação, aplicar um teste em que fosse dado maior peso às diferenças a longo prazo como o teste de Fleming-Harrington com $p = 0, q = 1$. Para tal, foi utilizado o *software* STATA, tendo sido obtido o valor-$p = 0.122$. Embora não significativo, houve, de facto, um decréscimo do valor-p.

8.1.2 Abordagem paramétrica

Para ilustrar a estimação paramétrica da função de sobrevivência, vamos considerar a distribuição de Weibull para modelar o tempo até à recidiva. A adequabilidade do modelo pode ser avaliada através de

um gráfico onde se representa o $\log(-\log(\hat{S}(t)))$ *versus* $\log t$. Como se pode observar pela Figura 8.6, embora tal não se verifique na totalidade, o gráfico é razoalmente linear. Depois de ajustar a distribuição de Weibull aos dados, obtivemos as seguintes estimativas de máxima verosimilhança dos parâmetros: $\hat{\lambda} = 0.027$ e $\hat{\gamma} = 0.73$. Através da

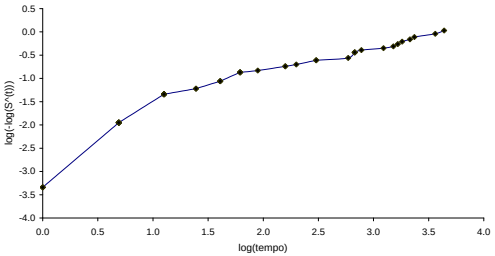


Figura 8.6 *Gráfico da função* $log(-log(\hat{S}(t)))$ *versus logt*

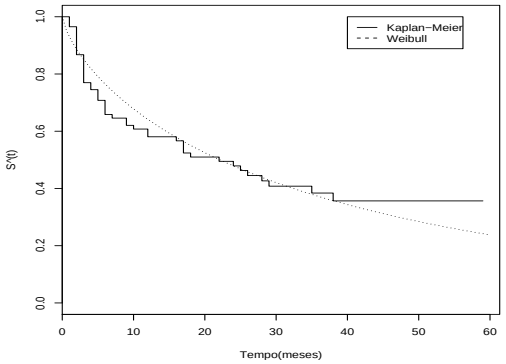


Figura 8.7 *Estimativas não paramétrica e paramétrica (Weibull) da função de sobrevivência*

comparação das estimativas paramétrica e não paramétrica (Kaplan-Meier) da função de sobrevivência, constatamos a plausibilidade da utilização da distribuição de Weibull. Como se pode observar pela Figura 8.7, as duas funções estão razoavelmente próximas.

A distribuição de Weibull parece, então, ser uma alternativa paramétrica possível para modelar o tempo até à recidiva de cancro da bexiga.

8.1.3 O modelo de regressão de Cox

Vejamos então se as conclusões obtidas anteriormente, em relação ao efeito da variável tratamento, se mantêm ao ser considerada a influência das restantes variáveis sobre o tempo até à recidiva. Para isso, utilizar-se-á uma abordagem semi-paramétrica, em que se recorrerá ao modelo de Cox. Começámos por ajustar um modelo de regressão simples incluindo cada uma das seguintes variáveis: tratamento, número inicial de tumores e diâmetro do maior tumor inicial. Apenas a primeira covariável é categórica, sendo o indivíduo padrão aquele que não fez quimioterapia. Como se pode observar na Tabela

Tabela 8.1 *Resultados do ajustamento do modelo de regressão de Cox simples*

Variáveis	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\exp(\hat{\beta})$	valor-p (Wald)	IC (95%)
tratamento	-0.369	0.303	0.69	0.220	(0.382,1.250)
número	0.201	0.071	1.22	0.004	(1.060,1.400)
diâmetro	0.032	0.101	1.03	0.750	(0.847,1.260)

8.1, os resultados que dizem respeito ao tratamento são concordantes com os já obtidos anteriormente pelos testes log-rank e de Peto-Peto.

Em relação às outras duas covariáveis, apenas o número inicial de tumores se revelou fortemente influente no risco de recidiva. No entanto, embora o efeito do tratamento não seja estatísticamente significativo, é necessário incluir esta variável no modelo para atingirmos o objectivo do estudo. Justifica-se, assim, estimar o efeito da quimioterapia ajustado pelo número inicial de tumores, além de que uma redução no risco de recidiva de cerca de 30% já tem alguma relevância clínica. Assim sendo, procedeu-se ao ajustamento do modelo de regressão de Cox com estas duas variáveis (Tabela 8.2).

Tabela 8.2 *Resultados do ajustamento do modelo de regressão de Cox múltiplo*

Variáveis	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\exp(\hat{\beta})$	valor-p (Wald)	IC (95%)
tratamento	-0.515	0.313	0.60	0.100	(0.325,1.110)
número	0.231	0.075	1.26	0.002	(1.087,1.460)

É interessante observar que, com este modelo, podemos concluir que passa a existir alguma evidência, embora fraca, de que a quimioterapia é mais eficaz. De facto, a realização de quimioterapia leva a um decréscimo de 40% do risco estimado de recidiva, para indivíduos com o mesmo número de tumores. Além disso, ter mais um tumor aumenta o risco estimado de recidiva em 26%, para indivíduos submetidos ao mesmo tratamento. Após a obtenção deste modelo de efeitos principais, foi introduzido um termo de interacção entre as duas covariáveis, o qual não revelou significância estatística (valor-p = 0.830).

8.1.4 Estudo da proporcionalidade das funções de risco

Para uma utilização correcta do modelo de regressão de Cox há que examinar a validade da hipótese de riscos proporcionais, o que pode ser feito através de um método gráfico e de um teste de hipóteses, baseados nos resíduos de Schoenfeld padronizados. Os gráficos

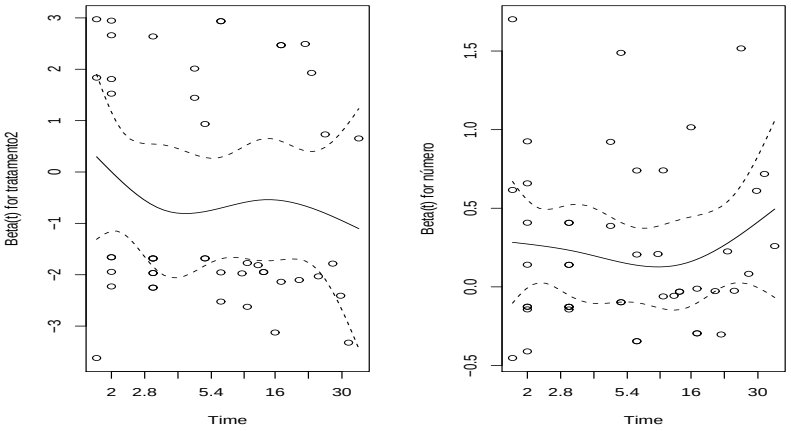


Figura 8.8 *Resíduos de Schoenfeld para o modelo de regressão de Cox múltiplo*

construídos constituem diagramas de dispersão, sobrepostos de uma função suavizadora gerada pelo LOWESS. Ao analisar estes gráficos, constata-se que o efeito de ambas as covariáveis se mantém razoavelmente constante ao longo do tempo, garantindo assim a proporcionalidade das funções de risco. De facto, se tentarmos ajustar uma recta à núvem de resíduos representada no gráfico, constamos que o seu declive será, seguramente, muito próximo de zero. Ao utilizar o teste

de hipóteses proposto por Grambsch e Therneau (1994), obteve-se uma conclusão semelhante, como se pode observar pela Tabela 8.3.

Tabela 8.3 *Resultados do teste de hipóteses para verificar a proporcionalidade das funções de risco*

Variáveis	$\hat{\rho}$	valor-p
tratamento	-0.125	0.396
número	0.021	0.884
Global		0.696

8.2 Análise de tempos de recidiva de doentes com leucemia

Com a investigação que deu origem aos dados que apresentamos neste segundo exemplo prático, pretendeu-se averiguar, entre outros aspectos, quais os factores que pudessem influenciar o tempo até à recidiva de doentes com leucemia linfoblástica aguda, após terem sido submetidos a determinado tratamento de quimioterapia que induziu um estado de remissão. O estudo decorreu no Instituto Português de Oncologia em Lisboa (entre Janeiro de 1994 e Dezembro de 2001), onde foram seguidos 150 doentes em remissão da referida patologia. O objectivo principal deste estudo de coorte prospectivo era avaliar a importância, como factor de prognóstico, da doença residual mínima (DRM) após a indução de remissão, medida através da percentagem de células malignas (blastos) que resistiram ao tratamento. Naturalmente, o estudo desta variável teria que ser feito tendo também em consideração outras variáveis de reconhecida importância clínica na recidiva da doença em causa.

Assim, no que diz respeito aos dados recolhidos, foi tida em conta toda a informação correspondente a características demográficas, clínicas e biológicas consideradas relevantes para o estudo e, como é óbvio, foi registado o tempo que decorreu entre a data de entrada em remissão (data do último tratamento) e a data do diagnóstico de recaída para os doentes que recidivaram, ou o tempo que decorreu entre a data de entrada em remissão e a data do fim de estudo para os doentes que não recidivaram. Concretamente, foi registada a seguinte informação relativa aos 150 doentes: raça (caucasiana/outra), idade antes do tratamento, sexo, fenotipo de mau prognóstico (não/sim), número de leucócitos na data do diagnóstico da doença, percentagem de células malignas no final da terapêutica de indução, existência de alterações citogenéticas de mau prognóstico(não/sim), ploidia de mau prognóstico (não/sim), realização de transplante de medula óssea (não/alogénico/autólogo), tratamento (protocolo terapêutico efectuado), tempo de seguimento (em semanas) e estado final (recaída, remissão mantida).

Além dos elementos atrás referidos, alguma informação suplementar foi fornecida posteriormente pelo clínico. Assim, no que diz respeito à idade, tinham sido formados dois grupos: um grupo padrão que incluía doentes com idade compreendida entre os 2 e os 9 anos e um outro grupo de alto risco que incluía os restantes doentes; em relação à percentagem de blastos resistentes ao tratamento, foram constituídos três grupos: um grupo de baixo risco DRM(1)(% de blastos $< 0.01\%$), um grupo de risco intermédio DRM(2)(% de blastos compreendida entre 0.01% e 1%) e finalmente um grupo de alto risco DRM(3)(% de blastos $> 1\%$). Quanto ao número de leucócitos registados na altura do diagnóstico, havia por parte do investigador um desconhecimento total acerca da existência de qualquer categorização dos dados que permitisse classificar os doentes em grupos de risco

diferentes, existindo apenas algumas referências a estudos anteriores que atribuíam alguma importância a esta variável como factor de prognóstico.

Foi efectuada inicialmente uma análise descritiva de todas as variáveis. Assim, 106 (70%) doentes eram de raça caucasiana, 81 (54%) eram do sexo feminino e a idade dos doentes variou entre 0.1 e 50.3 anos com uma mediana de 6.44 anos. Ainda no que diz respeito à idade, se considerarmos as duas classes etárias sugeridas pelo clínico, constatamos que existe um número semelhante de doentes nos grupos de alto e baixo risco, 72 (48%) e 78 (52%), respectivamente. A mediana do número de leucócitos antes do tratamento foi de $18250/mm^3$, com um mínimo de $500/mm^3$ e um máximo de $1000000/mm^3$ (o valor normal varia entre $5000/mm^3$ e $11000/mm^3$). No que diz respeito à percentagem de blastos, 49 (33%) doentes encontravam-se no grupo de baixo risco, 75 (50%) no grupo de risco intermédio e 26 (17%) no grupo de alto risco. Com alterações citogenéticas havia apenas 12 (8%) doentes (23% de valores omissos) e com um fenotipo de mau prognóstico 33 (22%) doentes. Em relação à ploidia, 89 (59.3%) doentes pertenciam ao grupo de mau prognóstico (dos restantes, 30.7% tinham valores omissos). A maioria dos doentes (86.7%) não efectuou transplante de medula óssea, tendo os restantes sido submetidos a transplante alogénico (8%) e autólogo (5.3%). Foram constituídos dois grupos terapêuticos com 128 (85.3%) e 22 (14.7%) doentes e, finalmente, dos 150 doentes, 45 (30%) recidivaram. A mediana do tempo de seguimento foi de 101 semanas (min = 8, max = 416). Para os indivíduos que recidivaram, obteve-se uma mediana de 65.3 semanas (min = 8, max = 211.3), enquanto que para os indivíduos que não recidivaram, a mediana foi de 118.7 (min = 9.7, max = 416).

8.2.1 Análise univariável

Prosseguimos a análise destes dados utilizando as abordagens não paramétrica e semi-paramétrica. Estimámos, então, a função de sobrevivência através do estimador de Kaplan-Meier. Como se pode observar pela Figura 8.9, não foi possível estimar não parametricamente o tempo mediano até à recidiva visto que a estimativa da função de sobrevivência toma valores superiores a 0.5 em todos os instantes; de facto, $\hat{S}(t) = 0.564$ para $211.3 \leq t \leq 416$. Notemos ainda que o nivelamento da estimativa de Kaplan-Meier, após 211.3 semanas, é uma indicação da provável existência de indivíduos "curados" na população. Para atingir os objectivos deste estudo, foram efectuadas

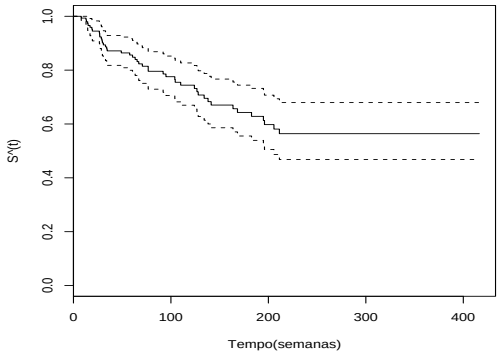


Figura 8.9 *Estimativa de Kaplan-Meier da função de sobrevivência*

algumas comparações entre grupos de doentes definidos pelos vários níveis das covariáveis. É de recordar que o interesse principal recaía sobre a estimativa do valor prognóstico da doença residual mínima (DRM), ajustada por outras variáveis de reconhecida importância clínica na recidiva da leucemia linfoblástica aguda. Assim sendo, es-

tas variáveis foram identificadas a partir de um estudo univariável em que recorremos ao estimador de Kaplan-Meier, aos testes log-rank e de Peto-Peto e também ao modelo de regressão de Cox. Na Figura 8.10 encontram-se os gráficos das curvas de sobrevivência estratificadas pela DMR e pelas alterações citogenéticas. A partir desta

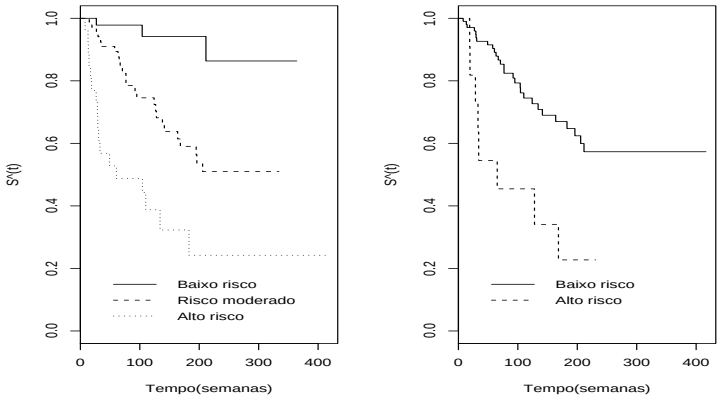


Figura 8.10 *Estimativa de Kaplan-Meier da função de sobrevivência, estratificada pela DRM (à esquerda) e pelas alterações citogenéticas (à direita)*

análise exploratória inicial, parece plausível não duvidar da proporcionalidade das funções de risco na medida em que as estimativas das funções de sobrevivência correspondentes às categorias de cada covariável não se cruzam. As curvas de sobrevivência são concordantes com o esperado pelo clínico e as diferenças encontradas pelos testes de hipóteses log-rank e de Peto-Peto são estatísticamente significativas (valor-$p < 0.001$ para ambos os testes). Em relação aos resultados obtidos pelo modelo de regressão de Cox simples, as variáveis que revelaram, por si só, estar relacionadas com o tempo até

à recidiva, encontram-se na Tabela 8.4. Note-se que as categorias de referência (padrão) utilizadas para as variáveis categóricas foram as que o clínico classificou como de baixo risco. Como se pode observar,

Tabela 8.4 *Resultados do ajustamento do modelo de regressão de Cox simples*

Variáveis	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\exp(\hat{\beta})$	valor-p (Wald)	IC (95%)
Idade	0.935	0.309	2.546	0.003	(1.389,4.668)
Nº leuc./5000	0.013	0.005	1.013	0.009	(1.003,1.024)
DRM(2)	1.762	0.611	5.823	0.004	(1.758,19.290)
DRM(3)	2.816	0.627	16.707	< 0.001	(4.886, 57.132)
Alter. citog.	1.210	0.402	3.354	0.003	(1.524,7.379)
Tratamento	1.134	0.343	3.109	0.001	(1.589, 6.085)

alguns dos intervalos de confiança têm amplitudes maiores do que o desejável, revelando alguma falta de precisão do estudo devido à baixa casuística que originou uma falta de representatividade em algumas categorias das covariáveis. Também é de referir que o número de leucócitos foi estudado após a sua divisão por 5000 dado ser este o número que, segundo o clínico, faz sentido utilizar aquando da interpretação do risco relativo (no nosso caso, o valor de 1.013 significa que, por cada aumento de 5000 no número de leucócitos, existe um acréscimo de 1.3% do risco estimado de recidiva).

8.2.2 Análise multivariável

Após esta primeira análise, foi ajustado um modelo de regressão múltipla em que, para apurar quais as variáveis que no final per-

maneceriam no modelo, foi utilizado o método de selecção progressiva. Aliás, foram utilizados outros métodos de selecção de variáveis, mas os resultados obtidos foram os mesmos e encontram-se resumidos na Tabela 8.5. Como se pode observar, o modelo não inclui o número

Tabela 8.5 *Resultados do ajustamento do modelo de regressão de Cox múltiplo (1º modelo)*

Variáveis	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\exp(\hat{\beta})$	p-value (Wald)	IC (95%)
Idade	1.038	0.359	2.82	0.004	(1.396,5.711)
DRM(2)	1.609	0.622	4.50	0.010	(1.476,16.921)
DRM(3)	2.832	0.645	16.98	0.000	(4.798, 60.122)
Alter.Citog.	1.034	0.421	2.81	0.014	(1.233,6.412)

de leucócitos e o tratamento. No que diz respeito à primeira destas variáveis, a sua não inclusão no modelo, apesar de isoladamente se ter revelado estatísticamente significativa, pode dever-se a vários motivos: uma explicação plausível consiste no facto da variável poder ser realmente irrelevante devido ao aumento da eficácia dos tratamentos. Em relação às duas abordagens terapêuticas utilizadas, já era esperado pelo clínico que a sua influência sobre o tempo até à recidiva fosse semelhante. Depois de obtido este modelo de efeitos principais, tentámos introduzir algumas interacções mas a baixa casuística conduziu a problemas de sobreajustamento.

Análise dos resíduos

Nesta fase do estudo, vamos começar por analisar os resíduos de Schoenfeld padronizados com vista a investigar a proporcionalidade das funções de risco. Para isso, fizemos a representação dos resí-

duos *versus* o tempo para cada uma das covariáveis incluídas no modelo, sobrepondo uma curva de suavização obtida pelo LOWESS. Ao analisarmos o gráfico da Figura 8.11, constatamos que a variável respeitante à idade, tal como foi discretizada pelo clínico, não parece verificar o pressuposto em estudo, contrariamente às restantes covariáveis. De facto, o efeito da idade não aparenta ser constante ao longo do tempo. Para complementar a observação dos gráficos, recorremos

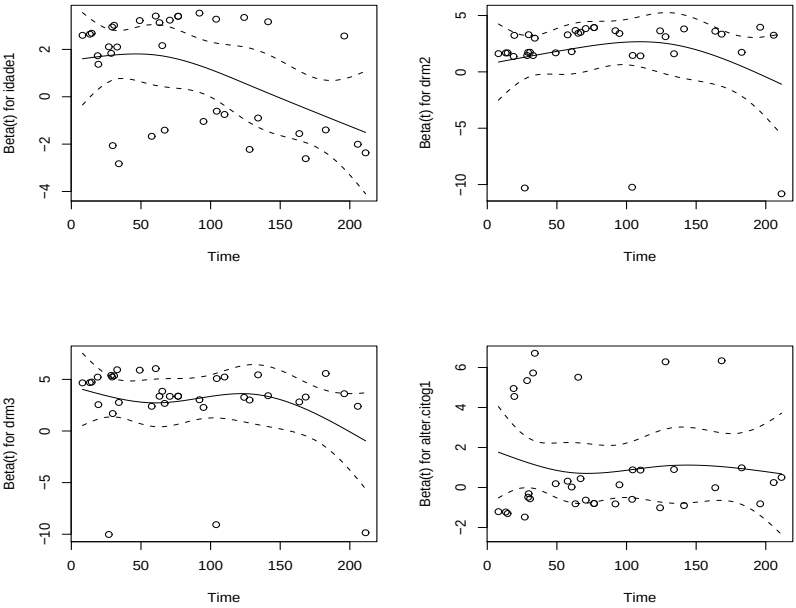


Figura 8.11 *Resíduos de Schoenfeld para o modelo de regressão de Cox múltiplo (1º modelo)*

ao teste de hipóteses cujos resultados se encontram na Tabela 8.6. Como se pode observar, as conclusões são concordantes com as obtidas a partir dos gráficos dos resíduos de Schoenfeld, sendo a idade

Tabela 8.6 *Resultados do teste de hipóteses para verificar a proporcionalidade das funções de risco (1º modelo)*

Variáveis	$\hat{\rho}$	valor-p
Idade	-0.4203	0.008
DRM(2)	-0.0672	0.685
DRM(3)	-0.2065	0.206
Alter.Citog.	-0.0593	0.715
Global		0.050

a única variável que viola o pressuposto da proporcionalidade das funções de risco. Nesta altura, devíamos então pensar nas estratégias alternativas para abordar esta questão. No entanto, como dispomos do valor da idade para cada indivíduo, vamos averiguar qual a forma funcional adequada para esta variável, eventualmente encontrando novos pontos de corte que levem à definição de uma nova variável para a qual se verifique a proporcionalidade dos riscos. Assim sendo, passaremos a analisar os resíduos martingala não só para determinar a forma funcional da variável idade e da percentagem de blastos, como também para identificar observações mal ajustadas pelo modelo. Em relação ao primeiro destes pontos, veremos se a influência destas variáveis no tempo até à recidiva é melhor modelada por uma forma funcional diferente da que foi considerada e se, portanto, é necessário transformar as variáveis referidas. Procedemos então à construção de um diagrama de dispersão em que os valores da covariável em estudo são registados no eixo das abcissas e os resíduos martingala, obtidos a partir do modelo nulo, são registados no eixo das ordenadas. Adicionalmente, foi sobreposto ao gráfico anterior uma curva de suavização obtida pelo LOWESS. Ao analisarmos o gráfico da Figura 8.12, são sugeridos novos pontos de corte para a idade. Assim,

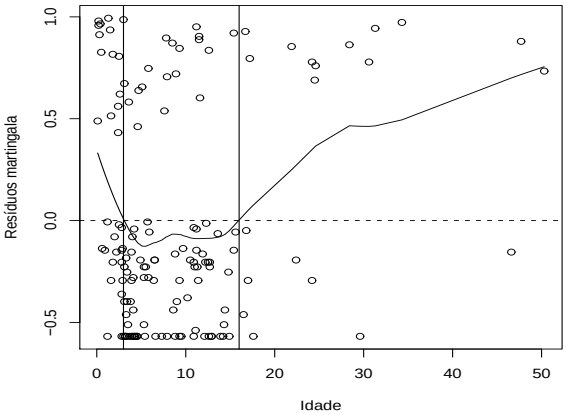


Figura 8.12 *Estudo da forma funcional da variável idade*

no grupo de baixo risco passam a estar incluídos os doentes com idade compreendida entre os 3 e os 16 anos e os restantes ficarão inseridos no grupo de alto risco. No que diz respeito à variável DRM, como se pode observar pela Figura 8.13, também são sugeridos novos pontos de corte. Assim, de acordo com a figura, seriam considerados apenas dois grupos: um de baixo risco com a percentagem de blastos inferior a 1.8% e outro de alto risco, com uma percentagem de blastos superior a 1.8%. No entanto, era importante para o clínico considerar o grupo de baixo risco indicado inicialmente e, por este motivo, mantiveram-se os três níveis para este factor. Assim, apenas foram alterados o grupo de risco intermédio, em que passaram a estar incluídos os doentes com uma percentagem de blastos compreendida entre 0.01% e 1.8%, e o grupo de alto risco incluindo os doentes com uma percentagem de blastos superior a 1.8%. Após esta análise, procedemos a um novo ajustamento do modelo de regressão de Cox. Como

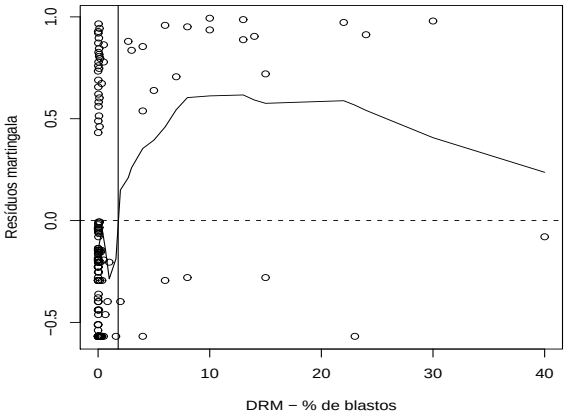


Figura 8.13 *Estudo da forma funcional da Doença Residual Mínima*

Tabela 8.7 *Resultados do ajustamento do modelo de regressão de Cox múltiplo (2º modelo)*

Variáveis	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\exp(\hat{\beta})$	p-value (Wald)	IC (95%)
Idade	1.25	0.379	3.47	0.001	(1.651,7.310)
DRM(2)	1.16	0.634	3.19	0.067	(0.921,11.060)
DRM(3)	2.81	0.647	16.67	< 0.001	(4.688,59.290)
Alter.Citog.	0.76	0.439	2.14	0.083	(0.904,5.050)

se pode observar pela Tabela 8.7, a idade é, sem dúvida, um importante factor de prognóstico para a recidiva da leucemia linfoblástica aguda. Ao procedermos, de novo, à verificação da proporcionalidade das funções de risco, obtivemos melhores resultados do que com o primeiro modelo que ajustámos aos dados, como se pode observar

pela Figura 8.14 e pela Tabela 8.8.

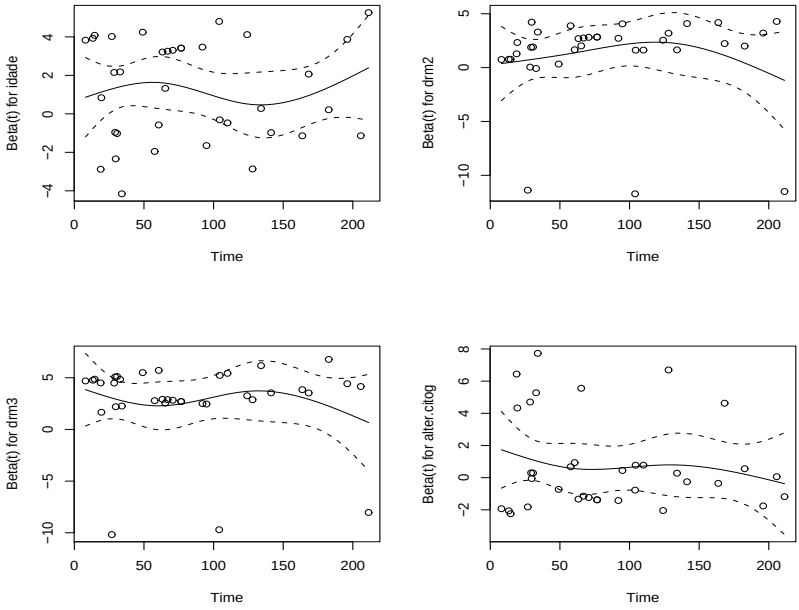


Figura 8.14 *Resíduos de Schoenfeld para o modelo de regressão de Cox múltiplo (2º modelo)*

Ainda recorrendo aos resíduos martingala, vejamos se existem observações mal ajustadas pelo modelo. Para isso, analisemos o gráfico da Figura 8.15. No gráfico à esquerda, destacam-se algumas observações com resíduos negativos no canto inferior direito da figura. Estas correspondem a doentes que se mantiveram em remissão durante mais tempo do que o esperado, embora apresentassem características indicadoras de mau prognóstico. Por outro lado, surgem ainda alguns resíduos iguais ou muito próximos da unidade (valor máximo que este tipo de resíduos pode atingir), correspondentes a doentes que

Tabela 8.8 *Resultados do teste de hipóteses para verificar a proporcionalidade das funções de risco (2º modelo)*

Variáveis	$\hat{\rho}$	valor-p
Idade	0.0145	0.921
DRM(2)	-0.0151	0.924
DRM(3)	-0.0813	0.627
Alter.Citog.	-0.1385	0.374
Global		0.864

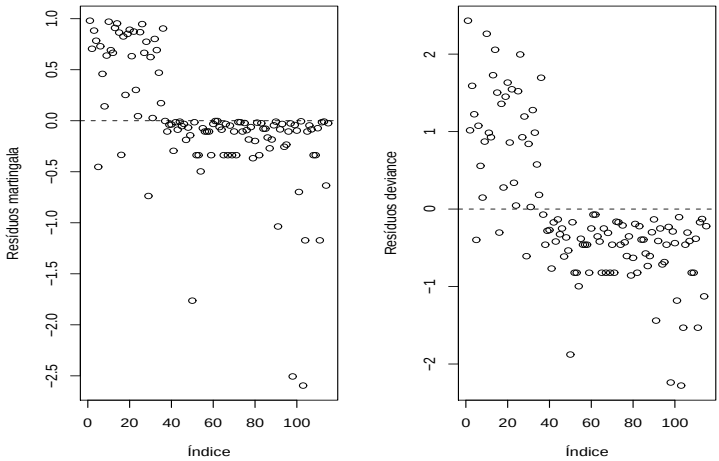


Figura 8.15 *Resíduos martingala versus índice (à esquerda), desvios residuais versus índice (à direita), para o modelo de regressão de Cox múltiplo (2º modelo)*

recidivaram muito cedo, de uma forma inesperada.

No que diz respeito à análise dos desvios residuais, também ilustrados na Figura 8.15, podemos constatar a sua distribuição mais simétrica em torno de zero. Basicamente, a informação contida nos dois gráficos é semelhante.

Os resultados encontrados no estudo da forma funcional das variáveis idade e percentagem de blastos foram ligeiramente discordantes da opinião expressa pelo clínico. Quanto à contagem leucocitária, a opinião recente de alguns especialistas aponta para um decréscimo da importância desta variável na ocorrência da recidiva dos doentes, devido a um aumento da eficácia de tratamentos utilizados actualmente. De facto, a variável não se revelou influente no tempo até à recidiva. Devemos também salientar que a data exacta da recidiva é desconhecida e está compreendida num intervalo delimitado pela data da última consulta de *follow-up* em que o doente ainda não recaiu e pela data da primeira consulta de *follow-up* em que o doente já recaiu. Embora habitualmente se considere que a data de recidiva coincide com a data do seu diagnóstico e que portanto, o tempo até à recidiva é exacto (não censurado), trata-se na realidade de um problema que envolve censura intervalar e cuja solução passa pela utilização de modelos próprios para a situação em causa.

Agradecimentos: As autoras agradecem ao Dr. Paulo Lúcio do Instituto Português de Oncologia de Francisco Gentil, a cedência dos dados sobre leucemia linfoblástica aguda.

Apêndice A

O modelo de Cox e os processos de contagem

Nos últimos anos, o recurso aos processos de contagem e teoria de martingalas para o estudo do modelo de Cox tornou-se uma realidade. Entre outros aspectos, os processos de contagem intervêm de uma forma relevante em novos desenvolvimentos na teoria de resíduos, tendo sido inicialmente usados para este fim por Barlow e Prentice (1988) e mais tarde por Therneau *et al.* (1990). Vamos em seguida começar por referir alguns conceitos básicos sobre processos de contagem (ver, por exemplo, Klein e Moeschberger, 1997; Therneau e Grambsch, 2000). Diz-se que $N(t)$, $t \geq 0$, representa um processo de contagem se for um processo estocástico que começa em zero ($N(0) = 0$), que verifica $N(t) < \infty$ com probabilidade 1 e cujas trajectórias são funções em escada, contínuas à direita e com saltos de amplitude +1.

Se estivermos perante uma amostra com dados censurados à direita, o processo $N_i(t) = I[T_i \leq t, \delta_i = 1]$, $i = 1, ..., n$, com $T_i = \min(X_i, C_i)$, em que X_i e C_i representam respectivamente, o tempo de vida e o tempo de censura do i-ésimo indivíduo e δ_i é igual a 1 se o i-ésimo indivíduo morrer e 0 se for censurado, é um processo de contagem que toma o valor 0 até ao instante de morte do i-ésimo indivíduo, a partir do qual passa a tomar o valor 1. De igual modo, o processo $N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{t_i \leq t} \delta_i$ também é um processo de contagem que conta o número de mortes que ocorreram na amostra até ao

instante t.

A toda a informação que caracteriza os indivíduos, acumulada até a um determinado instante t, chama-se história do processo de contagem, denota-se por $\{F_t; t \geq 0\}$ e trata-se da σ-álgebra gerada pelas variáveis aleatórias que caracterizam os indivíduos até esse instante; como é natural, para $s \leq t$, $F_s \subseteq F_t$, reflectindo o aumento de informação com o decorrer do tempo.

Se considerarmos F_{t-} como sendo a história de um indivíduo até imediatamente antes do instante t, então a $P[t \leq T_i < t + dt, \delta_i = 1|F_{t-}]$, que é igual a $h(t)dt$ (em que $h(t)$ representa a função de risco) se $T_i \geq t$ e igual a 0 caso contrário, representará a probabilidade do acontecimento de interesse ocorrer no intervalo $[t, t + dt)$, conhecida a respectiva história até ao instante t.

Definamos agora $dN(t)$ como sendo o incremento que o processo $N(t)$ sofre num intervalo de tempo infinitesimal $[t, t + dt)$. Como é óbvio, $dN(t)$ é igual 1 se ocorreu uma morte nesse intervalo e é igual a 0 caso contrário.

Considerando $Y(t)$ como sendo o processo que conta o número de indivíduos que se encontram em risco no instante t ($T_i \geq t$), então $E[dN(t)|F_{t-}]$=E[nº de observações com $t \leq X_i \leq t + dt, C_i > t + dt|F_{t-}]$=$Y(t)h(t)dt$ e a $\lambda(t)$=$Y(t)h(t)$ dá-se o nome de processo de intensidade do processo de contagem. Ao processo

$$\Lambda(t) = \int_0^t \lambda(s)ds, t \geq 0$$

chama-se processo de intensidade cumulativa e o processo $M(t) = N(t) - \Lambda(t)$ é conhecido como processo de contagem martingala. Este processo possui a propriedade de que o valor esperado dos seus incrementos é igual a zero, dada a história anterior F_{t-}, ou seja, $E[dM(t)|F_{t-}] = 0$. Ao recordarmos que um processo estocástico $M(t)$ é uma martingala se $E[M(t)|F_s] = M(s)$, $\forall s < t$, demonstra-se

que um processo de contagem martingala é efectivamente uma martingala. Segundo o teorema de decomposição de Doob-Meyer, qualquer processo de contagem se pode decompor univocamente como a soma de uma martingala e de um processo chamado compensador que, além de ser contínuo à direita e de assumir o valor 0 no instante 0, é também um processo predizível, ou seja, o seu valor no instante $t-$ é conhecido, dada a história passada F_{t-}. Referindo de novo os processos de contagem martingala, tem-se que $M_i(t) = N_i(t) - \Lambda_i(t)$ para o i-ésimo indivíduo logo, reescrevendo esta equação, obtém-se $N_i(t) = \Lambda_i(t) + M_i(t) = \int_0^t Y_i(s)h_i(s)ds + M_i(t)$, ou seja, "processo de contagem = compensador + martingala" análoga à decomposição, "valor observado = valor esperado + resíduos". Se considerarmos agora $\int_0^t Y_i(s)h_i(s)ds$ como um valor esperado $E_i(t)$, então a decomposição Doob-Meyer resulta em $N_i(t) = E_i(t) + M_i(t)$, decomposição esta que vai inspirar parte da teoria de resíduos.

O modelo de regressão de Cox assume que a função de risco é da forma $h_i(t) = h_0(t)e^{X_i(t)\boldsymbol{\beta}}$, onde $h_0(t)$ é uma função não especificada e não negativa, $X_i(t)$ representa o vector de covariáveis associado ao i-ésimo indivíduo no instante t e $\boldsymbol{\beta}$ é o vector de coeficientes de regressão.

No que diz respeito a este modelo, salientam-se, entre outros, os resíduos martingala que passamos a definir. Assim, como já foi visto anteriormente, o processo de contagem martingala para o i-ésimo indivíduo vem dado por $M_i(t) = N_i(t) - E_i(t)$ e transforma-se em $M_i(t) = N_i(t)$ - $\int_0^t Y_i(s)e^{X_i(s)\boldsymbol{\beta}}ds$ para o modelo de regressão de Cox. Naturalmente, o processo martingala residual virá definido por $\hat{M}_i(t) = N_i(t) - \hat{E}_i(t) = \int_0^t Y_i(s)h_0(s)e^{X_i(s)\hat{\boldsymbol{\beta}}}d\hat{H}_0$ em que $\hat{\boldsymbol{\beta}}$ é a estimativa de máxima verosimilhança parcial de $\boldsymbol{\beta}$ e $\hat{H}_0$ é a estimativa da função de risco cumulativa subjacente. Na realidade, os resíduos

martingala, correspondentes a determinado indivíduo, representam a diferença entre o número de acontecimentos observados para esse indivíduo e o número de acontecimentos esperados, dado o modelo que foi ajustado. Estes resíduos possuem algumas propriedades tais como: o valor esperado de cada resíduo é zero, quando calculado para o verdadeiro, mas desconhecido, vector de parâmetros $\boldsymbol{\beta}$ ($E(M_i) = 0$), a soma dos resíduos observados baseados em $\hat{\boldsymbol{\beta}}$ é zero ($\sum \hat{M}_i = 0$), os resíduos calculados para o verdadeiro vector de parâmetros $\boldsymbol{\beta}$ são não correlacionados ($Cov(M_i, M_j) = 0$) e finalmente, os resíduos observados são correlacionados negativamente embora essa correlação seja muito fraca ($Cov(\hat{M}_i, \hat{M}_j) < 0$) .

Referências

1. Aalen, O.O. (1987). Two examples of modelling heterogeneity in survival analysis. *Scandinavian Journal of Statistics*, **14**, 19-25.
2. Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine*, **7**, 1121-1137.
3. Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, **2**, 951-972.
4. Andersen, P.K. (2005). Censored data. Em *Encyclopedia of Biostatistics*, (eds. P. Armitage e T. Colton), Vol. 1, 578-582. London: Wiley.
5. Anderson, P.K. e Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, **10**, 1100-1120.
6. Barlow, W.E. e Prentice, R.L.(1988). Residuals for relative risk regression. *Biometrika*, Vol. 75, p. 65-74.
7. Bennett, S. (1983a). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273-277.
8. Bennett, S. (1983b). Log-logistic regression models for survival data. *Applied Statistics* **32**, 165-171.
9. Bergsten-Brucefors, A. (1976). A note on the accuracy of recalled age at menarche. *Annals of Human Biology*, **3.1**, 71-3.
10. Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**,89-99.
11. Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829-836.

12. Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2nd edition, Chapman & Hall/CRC, Boca Raton.
13. Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.
14. Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.
15. Cox, D.R. e Snell, E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series A*, **30**, 248-275.
16. Dabrowska, D.M., Doksum, K.A., Feduska, N.J., Husing, R. e Neville, P. (1992). Methods for comparing cumulative hazard functions in a semi-proportional hazard model. *Statistics in Medicine*, **11**, 1465-1476.
17. Dempster, A.P., Laird, N.M. e Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* , **39**, 1-38.
18. Diamond, I.D., McDonald, J.W. e Shah, I.H. (1986). Proportional hazard models for current status data: application to the study of differentials in age at weaning at Pakistan. *Demography*, **23**, 607-620.
19. Diamond, I.D. e McDonald, J.W. (1991). The analysis of current status data. Em *Demographic Applications of Event History Analysis*. J. Trussel, R. Hankinson e J. Tilton (eds.), Oxford: Oxford University Press.
20. Efron, B. (1967). The two sample problem with censored data. Em *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 831-853. New York: Prentice-Hall.
21. Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557-565.
22. Elbers, C. e Ridder, G. (1982). True and spurious duration dependence:The identifiability of the proportional hazard model. *Rev. Econ. Studies*, **49**, 403-409.
23. Fleming, T.R. e Harrington, D.P. (1981). A class of hypothesis tests for one and two samples of censored survival data. *Communications in Statistics*, **10**, 763-794.
24. Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203-223.

25. Grambsch P.M. e Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.
26. Hall, W.J. e Wellner, J.A. (1980). Confidence bands for a survival curve from censored data. *Biometrika*, **67**, 133-143.
27. Heckman, J. e Singer, B. (1984a). The identifiability of the proportional hazard model. *Rev. Econ. Studies*, **51**, 231-241.
28. Heckman, J. e Singer, B. (1984b). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271-320.
29. Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, **71**, 75-83.
30. Hougaard, P. (1991). Modelling heterogeneity in survival data. *Journal of Applied Probability*, **28**, 695-701.
31. Jewell, N.P. e Shiboski, S.C. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics*, **46**, 1133-1150.
32. Kalbfleisch, J.D. e Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267-278.
33. Kalbfleisch, J.D. e Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
34. Kaplan, E.L. e Meier, P.(1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.
35. Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series B*, **154**, 371-412.
36. Keiding, N., Begtrup, K., Scheike, T.H. e Hasibeder, G. (1996). Estimation from current-status data in continuous time. *Lifetime Data Analysis*, **2**, 119-129.
37. Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795-806.
38. Klein, J.P e Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.
39. Lancaster, T. e Nickell, S. (1980). The analysis of re-employment probabilities for the unemployed (with discussion). *Journal of the Royal*

Statistical Society, Series A, **143**, 141-165.

40. Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data*. 2nd edition, New York: Wiley.
41. Mantel, N. (1967). Ranking procedures for arbitrarily restricted observations. *Biometrics*, **23**, 65-78.
42. Mantel, N. e Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748.
43. Manton, K.G., Stallard, E. e Vaupel, J.W. (1981). Methods for comparing the mortality experience of heterogeneous populations. *Demography*, **18**, 384-410.
44. Manton, K.G., Stallard, E. e Vaupel, J.W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*, **81**, 635-644.
45. Marubuni, E. e Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. New York: Wiley.
46. Peto, R. (1972). Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society, Series B*, **34**, 205-207.
47. Peto, R. e Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, **135**, 185-206.
48. Prentice, R.L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167-179, Correction **70**:304 (1983).
49. Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Jr, Flournoy, N., Farewell, V.T. e Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**, 541-554.
50. Quandt, S. (1987). Material recall accuracy for dates of infant feeding transitions. *Human Organization*, **46**, 152-159.
51. Royston, P., Altman, D.G. e Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, **25**, 127-141.
52. Schoenfeld, D.A. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.
53. Shiboski, S.C. (1998). Generalized additive models for current status data.

Lifetime Data Analysis, **4**, 29-50.

54. Tarone, R.E. e Ware, J.H. (1977). On distribution-free tests for equality for survival distributions. *Biometrika*, **64**, 156-160.
55. Therneau, T.M. e Grambsch, P.M. (2000). *Modelling Survival Data: Extending the Cox Model*. New York: Springer.
56. Therneau, T.M., Grambsch, P.M. e Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147-160.
57. Tsiatis, A.A. (1975). A non-identifiability aspect of the problem of competing risks. Em *Proceedings of the National Academy of Science*, **72**, 20-22.
58. Vaupel, J.W., Manton, K.G. e Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

Agradecemos às seguintes entidades o apoio concedido na realização do XVII Congresso da Sociedade Portuguesa de Estatística:

Banco Millennium

Banco de Portugal

Câmara Municipal de Sesimbra

Centro de Matemáticas e Aplicações da UNL

Departamento de Matemática da FCT/UNL

Fundação para a Ciência e Tecnologia

Sesimbra SPA & Hotel

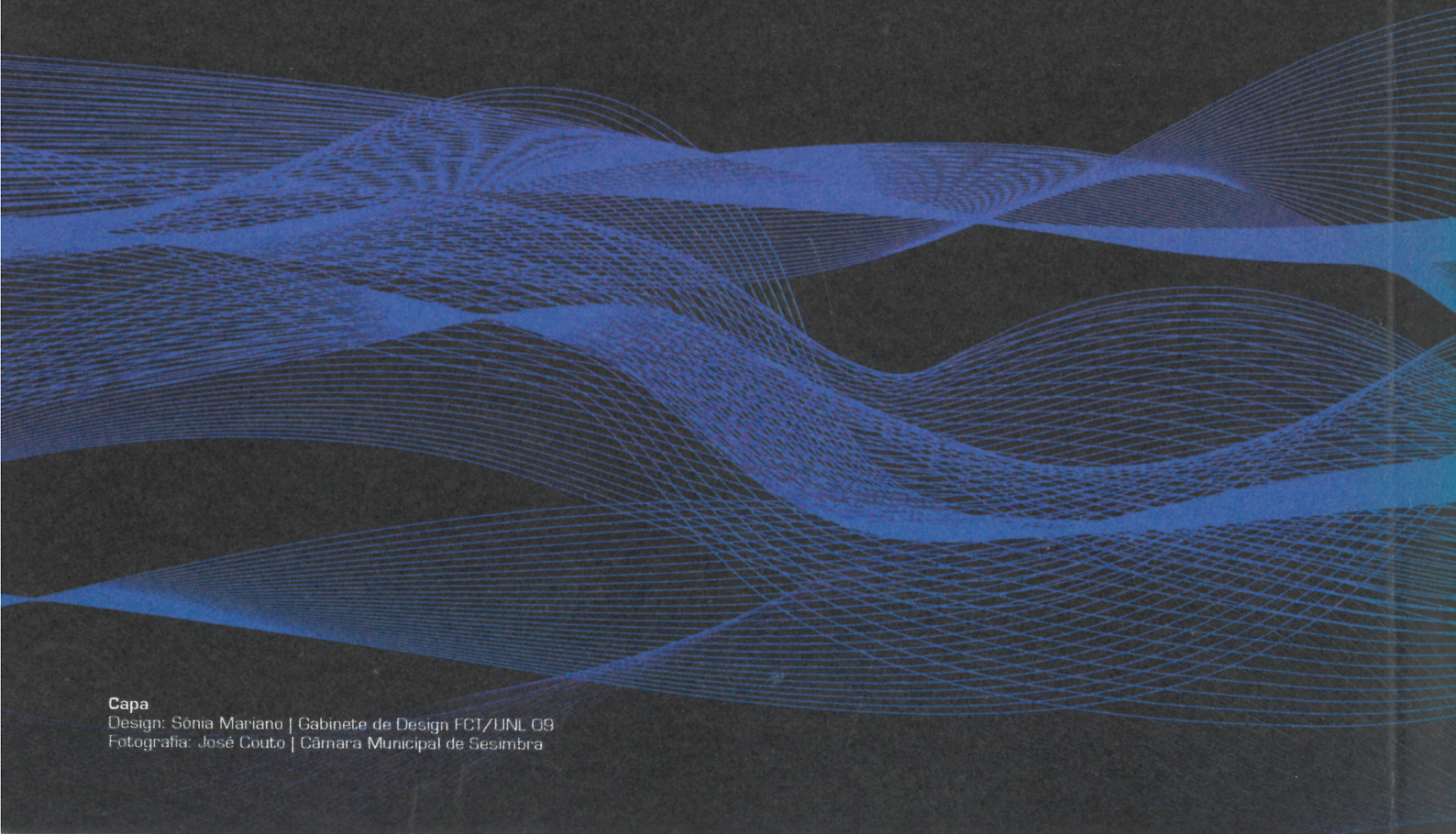
Instituto Nacional de Estatística

Livraria Escolar Editora

PSE - Produtos e Serviços de Estatística, Lda.

Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

ISBN 978-972-8890-22-3
Depósito Legal n.º 297954/09



Capa
Design: Sónia Mariano | Gabinete de Design FCT/UNL 09
Fotografia: José Couto | Câmara Municipal de Sesimbra