A Nossa Universidade

Colégio dos Jesuítas
Rua dos Ferreiros - 9000-082, Funchal

Tel: +351 291 209400
Fax: +351 291 209410
Email: gabinetedareitoria@uma.pt

UNIVERSIDADE da MADEIRA

DM

Modelos para Acontecimentos Múltiplos

Ivo Miguel Sousa Ferreira

Modelos para Acontecimentos Múltiplos

DISSERTAÇÃO DE MESTRADO

Ivo Miguel Sousa Ferreira
MESTRADO EM MATEMÁTICA

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

julho | 2016

DM

UMa | Nome do Projecto/Relatório/Dissertação de Mestrado e/ou Tese de Doutoramento | Nome do Autor

DIMENSÕES: 45 X 29,7 cm

PAPEL: COUCHÊ MATE 350 GRAMAS

IMPRESSÃO: 4 CORES (CMYK)

ACABAMENTO: LAMINAÇÃO MATE

NOTA*

Caso a lombada tenha um tamanho inferior a 2 cm de largura, o logótipo institucional da UMa terá de rodar 90º , para que não perca a sua legibilidade|identidade.

Caso a lombada tenha menos de 1,5 cm até 0,7 cm de largura o laoyut da mesma passa a ser aquele que consta no lado direito da folha.

UNIVERSIDADE da MADEIRA

UNIVERSIDADE da MADEIRA

Modelos para Acontecimentos Múltiplos

DISSERTAÇÃO DE MESTRADO

Ivo Miguel Sousa Ferreira

MESTRADO EM MATEMÁTICA

ORIENTADORA

Ana Maria Cortesão Pais Figueira da Silva Abreu

Modelos para Acontecimentos Múltiplos

DISSERTAÇÃO DE MESTRADO

Ivo Miguel Sousa Ferreira

MESTRADO EM MATEMÁTICA

JÚRI

Maria Teresa Alves Homem de Gouveia

Cristina Maria Tristão Simões Rocha

Ana Maria Cortesão Pais Figueira da Silva Abreu

"You've got to find what you love,
and that is as true for your work as it is for your lovers.
Your work is going to fill a large part of your life,
and the only way to be truly satisfied is to do what you believe is great work.
And the only way to do great work is to love what you do.
If you haven't found it yet, keep looking. Don't settle.
As with all matters of the heart, you'll know when you find it.
And, like any great relationship,
it just gets better and better as the years roll on.
So keep looking until you find it. Don't settle."

Steve Paul Jobs – Stanford commencement speech, 2005

Agradecimentos

Em setembro de 2014, mergulhei num sonho que há muito tempo ambicionava seguir: o estudo da Matemática/Estatística. Era um sonho de infância, daqueles que nem sempre se dá a devida importância. Felizmente, a paixão pela área moldou sempre a minha vida, de tal forma que deixou de ser coincidência o modo como as oportunidades foram surgindo. Decidi então parar de fugir daquele que possivelmente é o meu futuro e tenho que confessar que desde então sinto-me profundamente realizado.

A composição de uma dissertação de mestrado resulta de uma intensa experiência pessoal que envolve inúmeros desafios, os quais requerem muitas vezes o nosso isolamento para que possam ser resolvidos. Por vezes, dou por mim a pensar em todas as pessoas que me são próximas e na forma egoísta como tive de prescindir da presença delas. Para mim, esses momentos são os mais difíceis de ultrapassar. Na verdade, o que mais me angustia não é o trabalho, mas sim todos aqueles que estupidamente deixo para trás. Embora durante algum tempo não pude estar presente fisicamente na vida de alguns familiares e amigos, nunca me esqueci deles e segurei-me com força nas memórias que trago, as quais permitiram indubitavelmente chegar à conclusão deste projeto. Assim sendo, sinto que nestas páginas devo de expressar todo o meu reconhecimento para com aqueles que, durante os últimos dois anos, acompanharam de perto o meu percurso, em especial:

À Professora Doutora Ana Maria Abreu, orientadora desta dissertação, pela dedicação, autonomia e confiança que constantemente depositou em mim. A sua forma de trabalhar e ensinar cativou-me desde o primeiro instante, contribuindo para que o meu interesse/curiosidade pela área da Estatística fosse cada vez maior, em particular pela Análise de Sobrevivência. Agradeço-lhe ainda pela nossa amizade, por ter tido sempre o cuidado de escutar os meus desabafos e por me ter apoiado em muitas outras situações ao longo do mestrado.

Aos professores do Departamento de Matemática, da Faculdade de Ciências Exatas e da Engenharias, pela sólida formação académica transmitida aos seus alunos. Um obrigado especial à professora Doutora Maria

Teresa Gouveia por ser uma das pessoas que mais se preocupa com o bem estar de qualquer aluno e à professora Doutora Rita Vasconcelos por ter sempre acreditado nas minhas capacidades desde a licenciatura em Gestão.

Ao Dr. António Freitas, técnico da biblioteca da Universidade da Madeira, pela sua assistência na pesquisa de vários artigos científicos.

Aos meus pais, Cora Maria e João Carlos, pelo amor incondicional que todos os dias fazem questão de transmitir. Agradeço-lhes profundamente por todos os sacrifícios que têm feito e por todas as oportunidades que me têm proporcionado. Por essa razão e muitas outras, quero dedicar-lhes este trabalho, pois reconheço que sem eles nada do que conquistei até hoje era possível.

Aos meus irmãos, Luís Carlos e Marília Joana, por estarem sempre presentes e por me apoiarem em qualquer decisão.

À minha avó materna, Julieta de Sousa, por ser uma das mulheres que mais adoro neste mundo e porque sei o quanto ambicionou assistir à conclusão deste curso.

Ao Pedro Diogo Ideia Freitas por me ter aconselhado e ajudado a superar diversas situações que eu próprio julgava que não iria ultrapassar. Nele vejo um exemplo a seguir a todos os níveis e, muitas vezes, é com base nessa imagem que me tento aperfeiçoar. Aproveito para agradecer à sua família, tanto pela amizade como pelo acolhimento com que sempre me receberam.

À Joana Pinto por todas as palavras de apoio que só ela sabe dar, as quais se revelaram cruciais para continuar motivado e, assim, terminar este objetivo. Agradeço ainda à Sandra Agrela e ao Diogo Abreu por me terem ajudado a abstrair, principalmente na fase final em que tudo parecia demasiado complicado.

Aos meus colegas de curso, em especial ao Vítor Jesus, Duarte Sousa e António Gomes, por me terem acompanhado desde o primeiro ano do Ensino Superior e com os quais pude crescer pessoal e profissionalmente.

Aos restantes familiares e amigos que, direta ou indiretamente, contribuíram para a realização desta grandiosa etapa: o meu sincero obrigado!

Resumo

Ao longo dos últimos anos tem existido um interesse crescente em estudar o tempo até à ocorrência de vários acontecimentos, os quais podem ser observados mais do que uma vez para um mesmo indivíduo. Os dados respeitantes a acontecimentos múltiplos têm como principal característica o facto de se registar mais do que um tempo de vida para cada indivíduo, o que inviabiliza a aplicação direta do modelo de regressão de Cox. Assim, surgiu a necessidade de desenvolver novas extensões deste modelo, sendo que as mais utilizadas na prática foram sugeridas por: Prentice, Williams e Peterson (PWP); Andersen e Gill (AG); Wei, Lin e Weissfeld (WLW); e Lee, Wei e Amato (LWA).

Um dos maiores obstáculos na aplicação destes modelos é a forte possibilidade de existir correlação intraindivíduos. Neste ponto, os quatro modelos referidos anteriormente são classificados como modelos marginais, uma vez que o vetor de parâmetros de regressão é estimado com base no ajustamento de um modelo que ignora a correlação entre acontecimentos. Para compensar esse facto, nestes modelos é usado um estimador robusto da matriz de covariância, o qual permite efetuar a correção necessária na estimativa da variância usual.

Após realizar uma descrição detalhada de cada modelo marginal procedeu-se à respetiva implementação através do *software* estatístico R. Para o efeito, recorreu-se à simulação de dados relativos a acontecimentos múltiplos do mesma natureza, isto é, a acontecimentos recorrentes. Os resultados obtidos permitiram realçar e confirmar as características dos modelos estudados.

Palavras-chave: acontecimentos múltiplos, Análise de Sobrevivência, modelo de regressão de Cox, modelos marginais, simulação no R.

Abstract

Over the last years there has been an increasing interest in studying the time until the occurrence of several events, which can be observed more than once for the same subject. The main feature of multiple events data is the fact that, for each subject, any event can occur multiple times in the course of the follow-up. This makes the direct application of the Cox regression model unfeasible. Thus, the development of new extensions of this model was needed. The most employed ones were suggested by: Prentice, Williams and Peterson (PWP); Andersen and Gill (AG); Wei, Lin and Weissfeld (WLW); and Lee, Wei and Amato (LWA).

One of the major obstacles in the application of these models is the strong possibility of existing within-subject correlation. At this point, the four models mentioned above are classified as marginal models, since the vector of regression parameters are estimated based on fitting a model that ignores the correlation between events. In order to overcome this fact, a robust covariance matrix estimator has been used, which allows the needed correction in the standard variance estimate.

After performing a detailed description of each marginal models, its implementation through the R statistical software was shown. For this purpose, the simulation of recurrent events data was performed. The results allowed to highlight and confirm the characteristics of these models.

Keywords: Cox regression model, multiple events, marginal models, R simulation, Survival Analysis.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Análise de Sobrevivência

1.1 Introdução

A Análise de Sobrevivência é um dos ramos mais antigos da Estatística, com origem no século XVII. Surgiu, essencialmente, como resposta à necessidade de obter métodos estatísticos que permitissem a resolução de problemas na área das ciências biomédicas, o que veio a influenciar toda a terminologia utilizada. O seu desenvolvimento começou a ser notório na segunda metade do século XX, graças ao trabalho pioneiro desenvolvido por Kaplan e Meier em 1958, onde propuseram um estimador não paramétrico para a função de sobrevivência [44]. Um outro marco histórico extremamente importante ocorreu em 1972, quando Sir David Cox propôs um modelo de regressão semiparamétrico capaz de englobar fatores que se presume afetarem o tempo de sobrevivência [27]. Este modelo veio revolucionar a Análise de Sobrevivência e, a partir dele, têm sido desenvolvidas novas extensões e abordagens que procuram dar resposta aos mais variados problemas. Consoante o acontecimento que se pretende estudar, tem sido aplicado em várias áreas, nomeadamente na medicina, economia, sociologia, engenharia e física.

A Análise de Sobrevivência permite analisar os tempos de vida dos indivíduos, desde um instante inicial bem definido (instante em que estes entram em estudo) até à ocorrência de um acontecimento de interesse. Tendo em conta que este ramo da estatística surge ligado à saúde, usualmente utiliza-se o termo morte ou falha para designar a ocorrência do acontecimento de interesse. Este acontecimento é definido à partida e pode assumir diversas formas como, por exemplo, a morte de um paciente, o ingresso no mercado de trabalho ou o pedido de insolvência de uma empresa.

A escolha deste tipo de análise estatística tem a particularidade de permitir a inclusão de dados censurados, que surgem quando não é possível observar a realização do acontecimento de interesse durante o período em que o indivíduo se encontra em observação. Algumas razões para isso acontecer residem no facto do indivíduo viver para além do tempo de estudo (se o acontecimento de interesse for a morte), ou por abandonar o tratamento que lhe foi atribuído. Deste modo, todas as observações podem e devem ser incorporadas na análise, não existindo perda de informação.

Durante a análise do tempo de vida de cada indivíduo, existem diversos fatores capazes de afetar a sobrevivência, conhecidos como fatores de risco ou prognóstico. Esses fatores são representados por variáveis chamadas explanatórias, preditoras ou covariáveis. A forma usual de englobar este tipo de variáveis no estudo é através da análise de regressão. No entanto, dadas as características que os dados de sobrevivência apresentam (sendo a principal a existência de dados censurados) não é possível aplicar a análise de regressão habitual, tendo sido desenvolvidos modelos adequados a esta situação, sendo o mais usual o modelo proposto por Cox [71].

O presente capítulo debruçar-se-á sobre os conceitos básicos de Análise de Sobrevivência. Neste contexto, serão introduzidas algumas definições essenciais para o estudo da variável aleatória que representa o tempo de vida, também conhecido por tempo de sobrevivência. Seguidamente, serão feitas algumas considerações acerca da função de risco e da sua importância para a análise de tempos de vida. A temática da censura também integra este capítulo, visto ser um conceito muito particular deste ramo da estatística. Será ainda construída a função de verosimilhança de um modelo paramétrico, para indivíduos sujeitos a um mecanismo de censura à direita. Por fim, serão apresentadas algumas abordagens não paramétricas para a estimação das funções de sobrevivência e de risco, assim como alguns testes não paramétricos utilizados para comparar as funções de sobrevivência de dois ou mais grupos de indivíduos.

1.2 Alguns conceitos básicos

Considere-se que o tempo de vida de um indivíduo pertencente a uma dada população homogénea[1] pode ser representado por uma variável aleatória (v.a.) T não negativa. Consoante a situação o tempo de vida pode ser discreto ou contínuo. Contudo, para o desenvolvimento dos capítulos que se seguem apenas será necessário analisar o caso contínuo.

[1] Não existem diferenças entre indivíduos, quanto a fatores que se supõe afetar o tempo de vida.

Em geral, define-se uma v.a. absolutamente contínua através da função densidade de probabilidade e da função de distribuição. Porém, na Análise de Sobrevivência a caracterização da v.a. T, que representa o tempo de vida de um indivíduo, é essencialmente feita a partir da respetiva função de sobrevivência e da função de risco (*hazard function*).

Em primeiro lugar, defina-se a função de sobrevivência da v.a. T como a probabilidade de um indivíduo sobreviver para além do instante t e represente-se por

$$S(t) = P(T > t) = 1 - F(t), \qquad t \geq 0,$$

onde $F(t)$ é a função de distribuição correspondente. A função de sobrevivência é monótona decrescente, contínua à esquerda e verifica as seguintes propriedades:

- $S(t) = 1$ quando $t = 0$;
- $S(+\infty) = \lim_{t \to +\infty} S(t) = 0$.

A função densidade de probabilidade num determinado instante t representa um valor aproximado da probabilidade de um indivíduo morrer num intervalo infinitesimal $[t;\ t + dt)$. Assim, esta função é designada por taxa instantânea de morte e é definida por

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

Posto isto, quando $F(t)$ e $S(t)$ são diferenciáveis, é possível verificar-se que $f(t) = F'(t) = -S'(t)$.

Uma outra forma de caracterizar a distribuição de T pode ser conseguida a partir da função de risco, também conhecida por função intensidade ou força de mortalidade, definida do seguinte modo

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

e que satisfaz as propriedades que se seguem:

- $h(t) \geq 0, \quad \forall t \geq 0$;
- $\int_0^{+\infty} h(t)dt = +\infty$.

Esta função representa a taxa instantânea de morte de um indivíduo no instante t, condicional a este não ter morrido até esse instante.

Com base nas definições anteriores é possível deduzir-se várias relações entre $f(t)$, $S(t)$ e $h(t)$, nomeadamente

$$h(t) = \frac{f(t)}{S(t)}, \quad t \geq 0; \tag{1.1}$$

$$S(t) = \exp\left(- \int_0^t h(u)du \right), \quad t \geq 0. \tag{1.2}$$

Além do mais, pode-se ainda definir a função de risco cumulativa ou integrada por

$$H(t) = \int_0^t h(u)du, \quad t \geq 0, \tag{1.3}$$

donde, a partir de (1.2), obtém-se

$$S(t) = \exp\big[- H(t)\big] \Leftrightarrow H(t) = - \log S(t), \quad t \geq 0. \tag{1.4}$$

Assim sendo, $H(t)$ mede o risco de ocorrência do acontecimento de interesse até ao instante t, pelo que é uma função não negativa, monótona crescente e $\lim\limits_{t \to +\infty} H(t) = +\infty$.

1.3 Função de risco

Quando se pretende estudar a evolução do risco de morte ao longo do tempo, a função de risco revela-se extremamente útil pois é a forma desta função que irá indicar o modo como o risco de morte evolui. Repare-se que, para esta questão, a função de sobrevivência seria pouco informativa visto que é sempre uma função monótona decrescente [71]. Deste modo, analise-se as possíveis formas que a função de risco pode tomar:

1. **monótona crescente** – é a forma mais comum na análise de tempos de vida, uma vez que corresponde a um risco de morte crescente ao longo do tempo. Esta situação ocorre quando a proporção dos indivíduos que morrem relativamente aos que sobrevivem vai aumentando;

2. **monótona decrescente** – é a forma menos habitual, pois encontra-se associada a uma diminuição do risco de morte ao longo do tempo. Como exemplo, considere-se o risco de desemprego que, em geral, diminui à medida que aumenta o tempo que um indivíduo permanece no mesmo trabalho;

3. **constante** – refere-se a tempos de vida com distribuição exponencial. Esta forma ocorre, por exemplo, quando os únicos riscos de morte são situações esporádicas como acidentes e doenças raras, ou quando o tempo de observação é demasiado curto para permitir que se verifiquem alterações no risco de morte;

4. ***bathtub-shaped*** – inicialmente verifica-se um risco de morte decrescente até que este estabiliza, ficando constante durante um longo período de tempo, após o qual começa a aumentar. Surge, por exemplo, em estudos sobre a mortalidade populacional, onde os indivíduos em observação são acompanhados desde o nascimento até à morte efetiva;

5. ***hump-shaped* ou unimodal** – inicialmente o risco de morte aumenta durante algum tempo e até certo ponto, a partir daí o mesmo começa a diminuir. Esta situação verifica-se, por exemplo, no caso de indivíduos submetidos a uma cirurgia, onde no início o risco de morte aumenta mas depois diminui dada a recuperação do indivíduo.

A representação gráfica da função de risco permite orientar a escolha da função de distribuição a aplicar. Em Análise de Sobrevivência, as distribuições paramétricas mais utilizadas são: distribuição exponencial, distribuição de Weibull, distribuição gama, distribuição loglogística e distribuição lognormal.

1.4 Censura

Uma observação é censurada quando não é possível observar a realização do acontecimento de interesse, durante o período em que o indivíduo se encontra em observação [46, 47]. Em tal caso, obtém-se uma informação parcial sobre o tempo de vida do indivíduo, sendo que esta deve ser englobada no estudo de modo a não existir perda de informação.

Existem várias razões que levam à existência de dados censurados, em particular: i) quando o período de observação termina sem que o indivíduo tenha experienciado o acontecimento de interesse; ii) quando se desconhece o estado em que o indivíduo se encontra porque se perdeu o contacto com o mesmo (nesta situação diz-se que o indivíduo ficou perdido para o *follow-up*, ficando apenas registada a data de último contacto); iii) quando existem fatores relacionados com o tempo de vida do indivíduo em estudo que impedem a ocorrência do acontecimento de interesse.

Para qualquer uma das três razões enunciadas, suponha-se que o indivíduo entra em estudo no instante t_0 e que a realização do acontecimento de interesse ocorre no instante $t_0 + t$, com t desconhecido. A única informação

disponível refere-se a um instante em que o acontecimento de interesse ainda não ocorreu, pelo que este pode ser definido por $t_0 + c$, $c < t$, onde c representa o tempo de vida censurado. Este mecanismo de censura é conhecido por censura à direita e é o mais comum na análise de dados de sobrevivência.

Existem outros mecanismos de censura, como a censura à esquerda e a censura intervalar. A censura à esquerda surge quando se tem conhecimento que o tempo de sobrevivência de um indivíduo é inferior ao tempo observado, isto é, aquando da observação verificou-se que o acontecimento de interesse já tinha ocorrido. É o mecanismo de censura menos habitual. Quanto à censura intervalar, esta acontece quando se sabe que já ocorreu o acontecimento de interesse, mas não se conseguiu observar o momento exato da sua realização, apenas se sabe que este ocorreu num determinado intervalo de tempo. Este mecanismo de censura é frequente em estudos onde os indivíduos são seguidos de forma periódica.

O desenvolvimento deste trabalho irá considerar apenas o mecanismo de censura à direita pois, como já foi referido, é o mais habitual. Neste contexto, é importante mencionar que existem três tipos de censura à direita:

- **censura de tipo I** – surge quando o período de observação, durante o qual irá decorrer o estudo, é definido previamente. Neste tipo de censura, a todos os indivíduos que não tenham experienciado o acontecimento de interesse durante esse período, irão corresponder observações censuradas. Deste modo, o número de acontecimentos de interesse observados é uma variável aleatória.

- **censura de tipo II** – acontece quando, para uma amostra de dimensão n, se define à partida que o estudo terminará assim que for observado o r-ésimo acontecimento de interesse ($r \leq n$). Neste tipo de censura, a duração do tempo de estudo é uma variável aleatória.

- **censura aleatória ou de tipo III** – surge quando o período de observação é definido inicialmente mas a entrada dos indivíduos no estudo ocorre de forma aleatória. Por esse motivo, o tempo que os indivíduos permanecem em observação é uma variável aleatória.

Ao longo de um estudo é preciso ter em consideração que a existência de censura pode provocar alterações no risco de morte. A validade dos métodos habitualmente utilizados em Análise de Sobrevivência pressupõe que a razão que leva à existência de censura é independente da razão que leva à ocorrência do acontecimento de interesse. Quando um mecanismo de censura satisfaz esta condição diz-se que é um mecanismo de censura independente. Associado a esta ideia surge muitas vezes o conceito de censura não informativa pois,

ao considerar uma população homogénea de indivíduos, é necessário garantir que os indivíduos censurados num dado instante, são representativos (quanto ao risco de ocorrência do acontecimento de interesse) de todos os indivíduos que sobreviveram até esse instante [11, 71].

1.5 Variáveis explanatórias

Uma vez que o objetivo principal da Análise de Sobrevivência consiste na análise do tempo de vida dos indivíduos, é razoável considerar que existem diversos fatores capazes de afetar a sobrevivência, conhecidos como fatores de risco ou prognóstico. Esses fatores podem ser agrupados em duas categorias: i) fatores intrínsecos, que dizem respeito a propriedades ou características diretamente relacionadas com o indivíduo, tais como a idade, o género e outros fatores inerentes à sua história individual; ii) fatores exógenos, que se referem a elementos externos ao indivíduo, nomeadamente ao ambiente onde este se encontra inserido, seja este laboral, social, familiar, entre outros.

Em meados da década de setenta do século XX, começou-se a dar uma maior importância aos fatores de risco, passando a representá-los por variáveis chamadas explanatórias, preditoras ou covariáveis [27, 71]. É através destas variáveis que se obtém informação sobre a heterogeneidade existente na população pois, de uma maneira geral, os indivíduos em estudo diferem entre si relativamente a fatores passíveis de influenciar a sua sobrevivência. Assim, sempre que for possível, deve-se registar os valores individuais observados para cada uma das covariáveis, com o objetivo de compreender e quantificar a relação existente entre estas e o tempo de vida de cada indivíduo.

Genericamente, as variáveis explanatórias podem ser classificadas como sendo constantes ou dependentes do tempo. Uma variável explanatória é definida como constante quando não ocorrem alterações no seu valor ao longo do período de estudo. Para este caso tem-se, por exemplo, a variável indicatriz que expressa o grupo a que o indivíduo pertence e as variáveis demográficas (como o género e a nacionalidade). Por outro lado, diz-se que uma variável explanatória é dependente do tempo se o seu valor sofre alterações ao longo do período em que indivíduo se encontra em observação como, por exemplo, as variáveis laboratoriais obtidas em análises clínicas periódicas (como a glicémia e o colesterol) e fatores que podem ser controlados pelo experimentador com a intenção de os fazer variar ao longo do estudo (como a alimentação e a administração de um dado medicamento). Segundo Kalbfleisch e Prentice [43] esta última classificação, relativa às covariáveis dependentes do tempo, pode ainda ser subdividida em:

- **externas** – Uma covariável diz-se externa quando não se encontra diretamente relacionada com o mecanismo que regula a ocorrência do acontecimento de interesse;

- **internas** – Uma covariável diz-se interna quando as suas alterações ao longo do tempo estão relacionadas com a sobrevivência do indivíduo, sendo que essas alterações transportam informação sobre o seu tempo de vida.

Habitualmente, representa-se por $\boldsymbol{z} = (z_1, \ldots, z_p)'$ o vetor de covariáveis associado a cada indivíduo, onde p é o número de covariáveis em estudo.

1.6 Função de verosimilhança

Muitos dos métodos de inferência estatística baseiam-se na teoria assintótica de máxima verosimilhança. Como referido na secção 1.4, os dados de sobrevivência têm a particularidade de poderem ser censurados, o que dificulta a obtenção de distribuições amostrais exatas [48, 71].

Considere-se que a v.a. T, que representa o tempo de vida, segue um determinado modelo paramétrico, ou seja, que a distribuição do tempo de vida é conhecida a menos de um vetor de parâmetros, designe-se por $\boldsymbol{\theta}$, sobre o qual se pretende realizar inferência.

Na construção da função de verosimilhança admite-se que os tempos de vida e os tempos de censura são independentes e que a censura é não informativa. Para além disso, a função de verosimilhança será escrita de forma diferente consoante o mecanismo de censura a que os indivíduos estão sujeitos. Dado que o desenvolvimento deste trabalho assenta sobre o mecanismo de censura à direita, será abordada apenas essa situação.

Por consequência, a construção da função de verosimilhança será relativa a situações em que existem observações não censuradas e observações censuradas à direita. Nesse sentido, importa salientar que uma observação não censurada corresponde a um tempo de vida exato pelo que, através da função densidade de probabilidade nesse instante, é possível obter informação sobre a probabilidade de ocorrer o acontecimento de interesse. Por outro lado, quando uma observação é censurada à direita sabe-se que o verdadeiro tempo de vida do indivíduo é superior àquele que foi observado, sendo que essa informação pode ser extraída da função de sobrevivência nesse instante.

Sejam X e C duas v.a.'s independentes que representam o tempo de vida exato e o tempo de censura, respetivamente. Assim, define-se que o tempo de vida observado t, para um dado indivíduo, é uma observação da v.a. $T = \min\{X, C\}$. Considere-se ainda uma variável indicatriz δ, que toma o

valor um quando o acontecimento de interesse é observado, $X \leq C$, e toma o valor zero quando o tempo de vida é censurado, $X > C$. Deste modo, tem-se a distribuição conjunta do par (T, δ), onde a sua função densidade de probabilidade pode ser obtida por intermédio da função densidade conjunta de X e C. Desta feita, a contribuição de um indivíduo cujo acontecimento de interesse foi observado é dada por

$$P(T = t, \delta = 1) = P(X = t, X \leq C) = P(X = t, t \leq C) = P(X = t)P(C \geq t).$$

De igual modo, obtém-se a contribuição de um indivíduo cujo tempo de vida foi censurado,

$$P(T = t, \delta = 0) = P(C = t, X > C) = P(C = t, X > t) = P(C = t)P(X > t).$$

Assumindo que as v.a.'s X e C são absolutamente contínuas, com funções densidade de probabilidade f e g e com funções de sobrevivência S e $1 - G$, respetivamente, tem-se que

$$P(T = t,\ \delta = 1) = f(t)\big[1 - G(t)\big],$$

e

$$P(T = t,\ \delta = 0) = g(t)S(t).$$

Por conseguinte, dada uma amostra aleatória de dimensão n, constituída pelos pares (t_i, δ_i), $i = 1, \ldots, n$, a função de verosimilhança virá escrita da seguinte forma

$$L = \prod_{i=1}^{n} \Big\{ f(t_i)\big[1 - G(t_i)\big] \Big\}^{\delta_i} \Big\{ g(t_i)S(t_i) \Big\}^{1-\delta_i}$$

ou, alternativamente,

$$L = \prod_{i=1}^{n} \Big\{ f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \Big\} \prod_{i=1}^{n} \Big\{ g(t_i)^{1-\delta_i} \big[1 - G(t_i)\big]^{\delta_i} \Big\}.$$

No caso de a censura ser não informativa, ou seja, de a distribuição do tempo de censura não depender do vetor de parâmetros $\boldsymbol{\theta}$, pode-se simplificar a expressão anterior, donde se obtém

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{1.5}$$

De forma equivalente, aplicando a relação (1.1), tem-se que

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

Perante condições de regularidade bastante gerais nos processos de morte e de censura, os resultados assintóticos usuais (baseados na teoria assintótica da máxima verosimilhança) continuam válidos. Consequentemente, o estimador de máxima verosimilhança $\hat{\boldsymbol{\theta}}$ é assintoticamente gaussiano multivariado com valor médio $\boldsymbol{\theta}$ e matriz de covariância $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$, onde $\boldsymbol{I}(\boldsymbol{\theta})$ denota a matriz de informação de Fisher.

1.7 Estimação não paramétrica

A estimação não paramétrica, tanto da função de sobrevivência como da função de risco cumulativa, é um tema muito abordado na literatura estatística para o desenvolvimento da investigação na área da saúde. Nesta secção, serão introduzidos alguns estimadores não paramétricos utilizados na estimação dessas funções. Diversos autores, como Collett [24], Cox e Oakes [29], Klein e Moeschberger [46], Lawless [48] e Marubini e Valsecchi [54] analisaram minuciosamente este assunto.

1.7.1 Função de sobrevivência

Em estudos relativos ao tempo de vida, é fundamental determinar a expressão que melhor represente a estimativa da probabilidade de um indivíduo sobreviver para além de um dado instante. Quando não existem observações censuradas, uma forma útil de retratar essa expressão é através da função de sobrevivência empírica [24, 48]. Para uma amostra de dimensão n, a função de sobrevivência empírica num dado instante t é dada pela proporção de indivíduos que sobreviveram para além desse instante, sendo definida por

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

Na presença de observações censuradas é necessário modificar a expressão anterior, pois não é possível determinar a quantidade exata de tempos de vida que são maiores do que t. Com o intuito de resolver essa particularidade, em 1958, Kaplan e Meier [44] apresentaram um estimador não paramétrico para a função de sobrevivência, conhecido por estimador produto-limite ou, simplesmente, estimador de Kaplan-Meier (KM).

Estimador de Kaplan-Meier

Suponha-se que para uma amostra aleatória de dimensão n, proveniente de uma população homogénea, os tempos de vida e de censura são conhecidos. Sejam $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, com $r \leq n$, os tempos de vida distintos para os quais se observou o acontecimento de interesse. Designe-se por d_k ($k = 1, \ldots, r$) o número de acontecimentos de interesse observados em $t_{(k)}$ e n_k o número de indivíduos em risco no instante imediatamente anterior, isto é, em $t_{(k)}^-$. Fazem parte do conjunto de indivíduos em risco, os indivíduos aos quais não se observou o acontecimento de interesse, nem a existência de censura. Caso um indivíduo seja censurado precisamente em $t_{(k)}$, convenciona-se que a censura ocorre imediatamente após esse instante.

Deste modo, Kaplan e Meier [44] propuseram o seguinte estimador não paramétrico para a função de sobrevivência:

$$\widehat{S}_{KM}(t) = \prod_{k:t_{(k)} \leq t} \frac{n_k - d_k}{n_k} = \prod_{k:t_{(k)} \leq t} \left(1 - \frac{d_k}{n_k}\right).$$

Tem-se que $\widehat{S}_{KM}(t) = 1$ enquanto $0 \leq t < t_{(1)}$, ou seja, este valor não se altera enquanto o primeiro acontecimento de interesse não for observado. É igualmente importante analisar, qual o valor que $\widehat{S}_{KM}(t)$ tomará assim que a maior observação for registada. Para isso é necessário considerar dois casos: i) se a maior observação registada for não censurada então $n_k = d_k$, pelo que $\widehat{S}_{KM}(t) = 0$ para $t \geq t_{(r)}$; ii) por outro lado, se a maior observação registada for censurada, designe-se por t^*, o estimador $\widehat{S}_{KM}(t)$ nunca poderá assumir o valor zero pois apenas se encontra definido até esse instante, isto é, $\widehat{S}_{KM}(t) = \widehat{S}_{KM}\big(t_{(r)}\big)$ para $t_{(r)} \leq t \leq t^*$ [24, 48, 71].

O estimador de Kaplan-Meier é uma função decrescente em forma de escada, onde cada degrau corresponde à observação de pelo menos um acontecimento de interesse. Um outro aspeto a salientar, é o facto deste estimador coincidir com a função de sobrevivência empírica quando não existem observações censuradas.

Para além disso, $\widehat{S}_{KM}(t)$ é um estimador consistente da função de sobrevivência e, perante determinadas condições de regularidade, também pode ser considerado um estimador de máxima verosimilhança não paramétrico.

Intervalo de confiança e estimação de quantis

Após a definição do estimador de Kaplan-Meier, interessa avaliar a sua precisão. Para isso, existe a possibilidade de se construir um intervalo de confiança para cada uma das estimativas obtidas, a partir dos quais se obtém

uma noção da precisão do estimador $\widehat{S}_{KM}(t)$ em cada um desses instantes. Todavia, no seguimento da construção desse intervalo, importa referir que a variância de $\widehat{S}_{KM}(t)$ será estimada pela fórmula de Greenwood [37], dada por

$$\widehat{\text{var}}\left\{\widehat{S}_{KM}(t)\right\} = \left[\widehat{S}_{KM}(t)\right]^2 \sum_{k:t_{(k)}\leq t} \frac{d_k}{n_k(n_k - d_k)},$$

uma vez que neste contexto existem observações censuradas.

Por conseguinte, dado que $\widehat{S}_{KM}(t)$ apresenta uma distribuição assintótica gaussiana de valor médio $S_{KM}(t)$, o intervalo de $100(1-\alpha)\%$ de confiança para a função de sobrevivência num dado instante t_0 é definido por

$$\widehat{S}_{KM}(t_0) \pm z_{1-\alpha/2}\widehat{\text{EP}}\left\{\widehat{S}_{KM}(t_0)\right\},$$

onde $\widehat{\text{EP}}\{\widehat{S}_{KM}(t_0)\}$ é a estimativa do erro padrão de $\widehat{S}_{KM}(t_0)$, obtida por intermédio da relação[2] entre a variância (resultante da fórmula de Greenwood) e o erro padrão, e $z_{1-\alpha/2}$ é o quantil $1-\alpha/2$ da distribuição gaussiana padrão. Os intervalos assim obtidos, para cada instante de morte, são designados por intervalos de confiança ponto-a-ponto (*pointwise*). Como exemplo, veja-se a Figura 1.1, obtida através do *software* estatístico R [70], em particular do *plug-in* RcmdrPlugin.KMggplot2 [82], que diz respeito ao tempo de vida de pacientes com leucemia mielóide aguda, os quais constam na base de dados *leukemia*, contida no *package* survival [78].

A maioria dos *softwares* estatísticos utilizados, determina este intervalo de confiança. No entanto, é preciso ter em atenção que as estimativas obtidas pelo estimador de Kaplan-Meier variam entre (0, 1) e, como o intervalo de confiança é simétrico, pode acontecer que os seus limites estejam fora desse intervalo. Na literatura existem algumas sugestões para a resolução deste problema [12, 24, 71]. Uma possível solução passa pela transformação das estimativas obtidas, por exemplo $\log\left[-\log \widehat{S}_{KM}(t_0)\right]$, sendo que após essa transformação calcula-se novamente o intervalo de confiança.

Em geral, a variável aleatória correspondente ao tempo de vida apresenta uma distribuição assimétrica positiva. Por essa razão, faz mais sentido utilizar a mediana para localizar o centro da distribuição (ver Figura 1.1). Assim, defina-se a estimativa da mediana do tempo de vida por

$$m = \min\left\{t_{(k)} : \widehat{S}_{KM}(t_{(k)}) \leq 0.5\right\},$$

[2] O erro padrão é estimado por $\widehat{\text{EP}}\left\{\widehat{S}_{KM}(t_0)\right\} = \widehat{S}_{KM}(t_0)\sqrt{\sum_{k:t_{(k)}\leq t_0} \frac{d_k}{n_k(n_k-d_k)}}$.

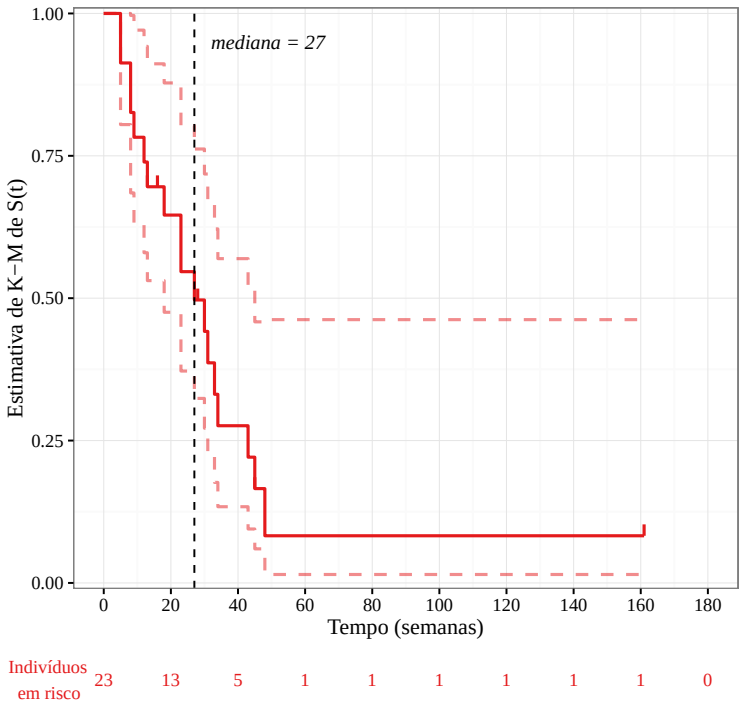


Figura 1.1: Estimativa de Kaplan-Meier da função de sobrevivência, respetivo intervalo de confiança e mediana.

onde $t_{(k)}$, $k = 1, \ldots, r$, representa o k-ésimo instante em que ocorreu algum acontecimento de interesse [24]. Note-se que podem acontecer situações em que a estimativa da função de sobrevivência mantém-se acima dos 0.5, qualquer que seja o valor de t. Este facto impossibilitará a estimação não paramétrica da mediana do tempo de vida. Para essas situações, será conveniente estimar outro quantil da distribuição. Então, generalizando a expressão anterior, a estimativa do quantil de probabilidade p é definida por

$$\widehat{Q}_p = \min \left\{ t_{(k)} : \widehat{S}_{KM}(t_{(k)}) \leq 1 - p \right\},$$

sendo $1 - p$ o respetivo percentil.

1.7.2 Função de risco cumulativa

Como referido anteriormente, a informação qualitativa sobre a função de risco pode ser preponderante na seleção de uma família de modelos para a caracterização da distribuição do tempo de vida. Por esse motivo, existem diversas abordagens reportadas na literatura para a estimação não paramétrica da

função de risco cumulativa, mas nesta secção serão analisadas apenas duas.

Uma das abordagens permite a construção de um estimador natural da função de risco cumulativa por intermédio do estimador de Kaplan-Meier da função de sobrevivência e da relação (1.4), de onde se obtém

$$\widehat{H}_{KM}(t) = -\log \widehat{S}_{KM}(t) = -\sum_{k:t_{(k)}\leq t} \log\left(1 - \frac{d_k}{n_k}\right). \tag{1.6}$$

Contudo, este estimador não é o mais apropriado para amostras de pequena dimensão. Para o efeito, Nelson e Aalen propuseram um outro estimador de $H(t)$, curiosamente em períodos diferentes [48].

Em 1969, Nelson [60] apresentou um estimador alternativo para a função de risco cumulativa. Esse mesmo estimador volta a ser sugerido por Aalen em 1972, aquando da publicação da sua tese de Mestrado [55]. Deste modo, considerando a notação utilizada no estimador de Kaplan-Meier, defina-se o estimador de Nelson-Aalen (NA) por

$$\widehat{H}_{NA}(t) = \sum_{k:t_{(k)}\leq t} \frac{d_k}{n_k},$$

também conhecido por função de risco cumulativa empírica. Para amostras de pequena dimensão é aconselhável estimar a variância de $\widehat{H}_{NA}(t)$ por

$$\widehat{\text{var}}\left\{\widehat{H}_{NA}(t)\right\} = \sum_{k:t_{(k)}\leq t} \frac{d_k}{n_k^2},$$

com o objetivo de se obter um viés reduzido [48].

Os dois estimadores apresentados atrás, para a estimação não paramétrica de $H(t)$, são assintoticamente equivalentes e originam resultados muito semelhantes, principalmente em amostras de grande dimensão. De facto, aplicando a expansão em série de $\log(1-x) = -x - {}^{x^2}/_2 - {}^{x^3}/_3 - \ldots$ ao estimador (1.6) e uma vez que $x = {}^{d_k}/_{n_k} < 1$, constata-se que

$$\widehat{H}_{KM}(t) = -\sum_{k:t_{(k)}\leq t} \log\left(1 - \frac{d_k}{n_k}\right) = \sum_{k:t_{(k)}\leq t} \left(\frac{d_k}{n_k} + \frac{d_k^2}{2n_k^2} + \frac{d_k^3}{3n_k^3} + \ldots\right),$$

pelo que o estimador de Nelson-Aalen pode ser encarado como uma aproximação de primeira ordem do estimador de Kaplan-Meier da função de risco cumulativa [24, 71].

Como é evidente, a partir do estimador de Nelson-Aalen também é possível obter um estimador não paramétrico da função de sobrevivência, designado por estimador de Breslow [17]. Assim, recorrendo novamente à relação (1.4) tem-se que

$$\widehat{S}_B(t) = \exp\left(-\widehat{H}_{NA}(t)\right) = \prod_{k:t_{(k)}\leq t} \exp\left(-\frac{d_k}{n_k}\right).$$

Tal como o estimador de Nelson-Aalen para a função de risco cumulativa, este estimador tem um melhor comportamento para pequenas amostras.

Em 1984, Fleming e Harrington [31] realizaram um estudo comparativo entre os estimadores de $S(t)$ propostos por Breslow e Kaplan-Meier, onde mostraram que, para amostras de grande dimensão, os dois estimadores são assintoticamente equivalentes. Realmente, se $\exp(-x) \approx 1 - x$ para valores pequenos de x, então para d_k/n_k pequeno, isto é, n_k grande[3] em relação a d_k, resulta que $\widehat{S}_B(t) \approx \widehat{S}_{KM}(t)$. Porém, na prática tem-se que $\exp(-x) \geq 1 - x$ e, por isso, a estimativa de Breslow é sempre superior ou igual à estimativa de Kaplan-Meier da função de sobrevivência, $\widehat{S}_B(t) \geq \widehat{S}_{KM}(t),\ \forall t > 0$. Note-se ainda que, no caso da última observação registada ser não censurada, $\widehat{S}_B(t) > 0$ enquanto que $\widehat{S}_{KM}(t) = 0$, como analisado na secção anterior.

1.8 Testes não paramétricos

Quando os indivíduos de uma dada população diferem entre si relativamente a algum fator, deve-se agrupá-los de forma homogénea de modo a determinar a estimativa de Kaplan-Meier para cada um dos grupos. Neste ponto, surge a necessidade de comparar as estimativas obtidas, com o objetivo de analisar o comportamento da distribuição do tempo de vida nos vários grupos e averiguar se existe algum grupo que evidencie uma maior sobrevivência.

Uma abordagem preliminar, que avalia de forma informal a existência de diferenças entre grupos, consiste em representar graficamente as estimativas de Kaplan-Meier de $S(t)$ num único referencial cartesiano, de onde é possível obter uma noção do padrão de sobrevivência. A Figura 1.2 ilustra essa situação, tendo-se agrupado os pacientes com leucemia mielóide aguda (da já referida base de dados *leukemia*) consoante o tipo de tratamento a que recorreram: sem quimioterapia (Grupo 1) ou com quimioterapia (Grupo 2). Contudo, esta abordagem não é a mais rigorosa, por isso, foram desenvolvidos testes de hipóteses que procuram avaliar a possibilidade de existirem diferenças significativas entre as curvas de sobrevivência [48, 54].

[3] Rocha e Papoila [71] sugerem como regra prática $n_k \geq 10d_k$.

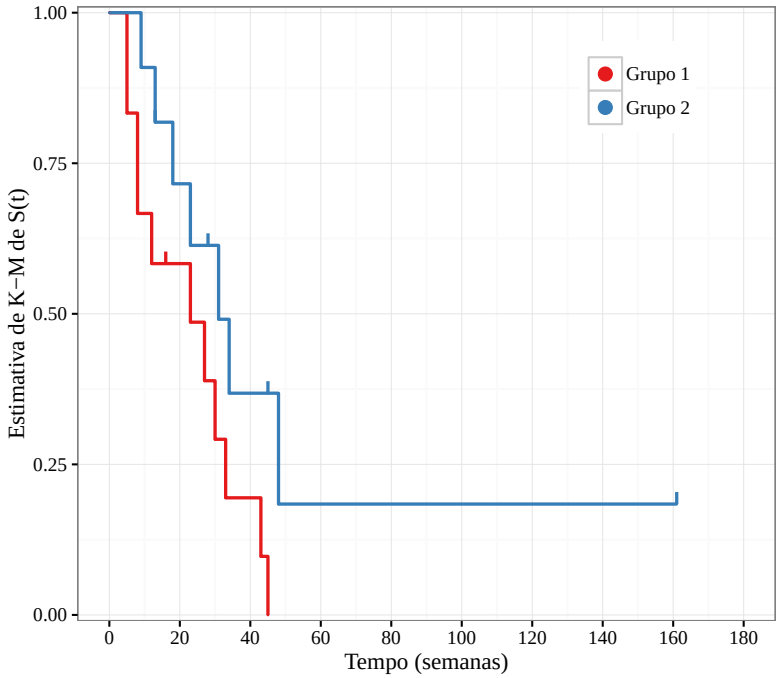


Figura 1.2: Representação das estimativas de Kaplan-Meier da função de sobrevivência.

Apesar de existirem vários testes não paramétricos para a comparação de funções de sobrevivência, não se pode afirmar que exista um que seja apropriado para todas as situações. Assim, a escolha encontra-se dependente de vários fatores como, por exemplo, dos padrões de ocorrência do acontecimento de interesse e de censura nos diferentes grupos, do efeito que estes provocam nas respetivas funções de risco[4] e, ainda, da hipótese alternativa de interesse. Assim, é a natureza da diferença que se espera detetar que determinará o teste mais potente para cada situação [71].

Inicialmente, considere-se que existem apenas dois grupos homogéneos de indivíduos ($g = 1, 2$) e represente-se por $S_g(t)$ a função de sobrevivência relativa ao g-ésimo grupo. Deste modo, a formulação das hipóteses de interesse é dada por

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

Quando não existem observações censuradas pode-se testar as hipóteses anteriores utilizando os habituais testes não paramétricos baseados em *ranks* como, por exemplo, o teste de Wilcoxon (para duas amostras emparelhadas)

[4]É aconselhável acompanhar a evolução do risco ao longo do tempo. Para isso, também se pode obter as estimativas de Nelson-Aalen de $H(t)$ e representá-las graficamente.

ou o teste de Mann-Whitney (para duas amostras independentes) [54]. No entanto, a existência de dados censurados compromete a aplicação destes testes, tendo sido desenvolvidos, para o efeito, uma extensão dos testes já existentes, o que deu origem a uma classe de testes em particular. De entre os vários testes pertencentes a essa classe, destaca-se o teste proposto por Mantel e Haenszel [53], usualmente designado por teste *log-rank*. Na prática, este é o teste mais utilizado e encontra-se presente na maioria dos *softwares* estatísticos.

1.8.1 Teste de Mantel-Haenszel (*log-rank*)

Considere-se uma amostra aleatória de dimensão $n = m_1 + m_2$, em que m_1 e m_2 representam as dimensões dos grupos 1 e 2, respetivamente. Represente-se por $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ os r instantes de morte distintos observados, relativos aos n indivíduos em estudo. Seja d_{gk}, $k = 1, \ldots, r$, o número de mortes observadas em $t_{(k)}$ no grupo g, $g = 1, 2$, e n_{gk} o número de indivíduos em risco no grupo g imediatamente antes desse instante, isto é, em $t_{(k)}^-$. Desta forma, $d_k = d_{1k} + d_{2k}$ é o número total de mortes observadas em $t_{(k)}$ e $n_k = n_{1k} + n_{2k}$ é o número total de indivíduos em risco em $t_{(k)}^-$.

Para cada instante $t_{(k)}$, a informação sobre os processos de morte nos dois grupos pode ser resumida numa tabela de contingência 2×2, como se observa na Tabela 1.1 (adaptada de Collett [24]).

Tabela 1.1: Tabela de contingência 2×2 para cada instante de morte $t_{(k)}$.

Grupo	Número de mortes em $t_{(k)}$	Número de sobreviventes em $t_{(k)}$	Número de indivíduos em risco em $t_{(k)}^-$
1	d_{1k}	$n_{1k} - d_{1k}$	n_{1k}
2	d_{1k}	$n_{2k} - d_{2k}$	n_{2k}
Total	d_k	$n_k - d_k$	n_k

Em 1959, Mantel e Haenszel [53] propuseram que se considerasse a distribuição condicional das frequências observadas em cada célula dados os seus totais marginais, sob a validade da hipótese nula relativa à igualdade das funções de sobrevivência. Para isso, apenas é necessário considerar a distribuição da frequência de uma única célula, por exemplo d_{1k}, uma vez que as restantes frequências ficam automaticamente determinadas pelos totais marginais fixos [54, 71].

Assim, admitindo que H_0 é verdadeira, a distribuição de d_{1k}, condicional aos totais marginais, tem a seguinte expressão

$$p(d_{1k}|d_k, n_k) = \frac{\binom{d_k}{d_{1k}}\binom{n_k - d_k}{n_{1k} - d_{1k}}}{\binom{n_k}{n_{1k}}},$$

ou seja, d_{1k} tem distribuição hipergeométrica, com valor médio e variância condicionais dadas por

$$E(d_{1k}) = n_{1k} \times \frac{d_k}{n_k} \qquad \text{e} \qquad \text{var}(d_{1k}) = \frac{n_{1k}n_{2k}d_k(n_k - d_k)}{n_k^2(n_k - 1)},$$

respetivamente. Repare-se que $E(d_{1k})$ é o número esperado de mortes em $t_{(k)}$ no grupo 1, pois n_{1k} é o número de indivíduos em risco no grupo 1 imediatamente antes do instante $t_{(k)}$ e d_k/n_k representa a probabilidade de ocorrer o acontecimento de interesse nesse instante, dado que, sob H_0, esta probabilidade não depende do grupo a que o indivíduo pertence [24].

Com o propósito de obter uma medida global que tenha em consideração os desvios dos valores observados d_{1k} relativamente aos valores esperados $E(d_{1k})$, é necessário combinar a informação compreendida nas r tabelas de contingência. Para o efeito, considere-se a estatística

$$U = \sum_{k=1}^{r} \left[d_{1k} - E(d_{1k})\right].$$

A estatística U tem como valor médio $E(U) = 0$ e, assumindo que $d_{11}, d_{12}, \ldots, d_{1r}$ são v.a.'s independentes, a sua variância corresponde à soma das variâncias de d_{1k}, isto é,

$$\text{var}(U) = \sum_{k=1}^{r} \text{var}(d_{1k}).$$

Para além disso, quando o número de acontecimentos de interesse observados é suficientemente grande, pode-se demonstrar que a estatística U tem distribuição assintoticamente gaussiana e, por conseguinte, $U/\sqrt{\text{var}(U)}$ segue uma distribuição assintótica gaussiana padrão, $N(0,1)$. Posto isto, através do quadrado desta última estatística, obtém-se a estatística de teste proposta por Mantel e Haenszel [53],

$$Q_{MH} = \frac{\left\{\sum_{k=1}^{r}\left[d_{1k} - E(d_{1k})\right]\right\}^2}{\sum_{k=1}^{r}\text{var}(d_{1k})} = \frac{U^2}{\text{var}(U)},$$

que se sabe ter distribuição assintótica qui-quadrado com um grau de liberdade, χ^2_1. Então, H_0 é rejeitada ao nível de significância α quando o valor observado de Q_{MH} é superior ao quantil de probabilidade $1-\alpha$ da distribuição qui-quadrado com um grau de liberdade, isto é, quando $Q_{MH} > \chi^2_{1-\alpha}$.

O motivo pelo qual este teste é habitualmente designado por teste *log-rank*, reside no facto de poder ser obtido pelos *ranks* dos tempos de vida nos dois grupos, sendo a estatística de teste resultante baseada no logaritmo[5] da estimativa da função de sobrevivência de Breslow [24].

Quando as funções de risco são proporcionais, o teste de Mantel-Haenszel é considerado o mais potente na deteção de diferenças significativas entre as funções de sobrevivência. Para além disso, também é bastante potente no caso de as funções de risco serem não proporcionais mas não se intersetarem. Com o intuito de avaliar a hipótese de riscos proporcionais, Rocha e Papoila [71] aconselham que, após determinar as estimativas de Kaplan-Meier da função de sobrevivência para cada um dos grupos, se construa o gráfico de $\log\left[-\log \widehat{S}_g(t)\right]$ *versus* t $(g = 1, 2)$, de onde é possível averiguar se as funções de risco se mantêm razoavelmente equidistantes.

1.8.2 Classe de testes Tarone-Ware

Como mencionado anteriormente, o teste de Mantel-Haenszel faz parte de uma classe específica de testes não paramétricos baseados em *ranks*, utilizados para testar a hipótese de igualdade de duas curvas de sobrevivência. Outros testes utilizados com alguma regularidade, e que são membros desta classe, foram sugeridos por Gehan [34], Tarone e Ware [77], Peto e Peto [66] e Prentice [67, 68]. Todos estes testes podem ser obtidos através de uma única estatística de teste geral, dada por

$$Q = \frac{\left\{\sum_{k=1}^{r} w_k\left[d_{1k} - E(d_{1k})\right]\right\}^2}{\sum_{k=1}^{r} w_k^2 \text{var}(d_{1k})}, \tag{1.7}$$

[5] Corresponde ao logaritmo de base natural.

em que $w_1, w_2, \ldots, w_r$ são constantes conhecidas, habitualmente designadas por pesos. Sob a validade de H_0, esta estatística de teste segue uma distribuição assintótica χ^2_1. Assim, de acordo com os valores atribuídos aos pesos w_k em (1.7), é possível obter uma vasta quantidade de testes diferentes, como sejam os referidos na Tabela 1.2 (adaptada de Marubini e Valsecchi [54]).

Tabela 1.2: Variantes dos testes não paramétricos baseados em *ranks*.

Designação	Pesos w_k	Estatística de teste
Teste *log-rank*	1	Q_{MH}
Teste de Gehan	n_k	Q_G
Teste de Tarone-Ware	$\sqrt{n_k}$	Q_{TW}
Teste de Peto-Prentice	$\prod_{l:t_{(l)}\leq t_{(k)}} \frac{n_l - d_l + 1}{n_l + 1}$	Q_{PP}

Em 1965, Gehan [34] propôs uma generalização do teste de Mann-Withney-Wilcoxon devido à existência de observações censuradas, pelo que também ficou conhecido como teste de Wilcoxon generalizado. Comparativamente ao teste *log-rank*, a estatística Q_G confere um maior peso às primeiras diferenças $\left[d_{1k} - E(d_{1k})\right]$, onde o número total de indivíduos em risco n_k é elevado, enquanto que Q_{MH} lhes atribui um peso constante ao longo dos r instantes de morte. Por esse motivo, o teste de Gehan é mais potente a encontrar diferenças significativas a curto prazo [71].

Contudo, quando os padrões de censura diferem consideravelmente entre os dois grupos de indivíduos, a estatística Q_G perde sensibilidade com o decorrer do tempo, podendo conduzir a conclusões inadequadas. De modo a contornar esse aspeto, Tarone e Ware [77], em 1977, sugeriram outros valores para os pesos w_k, que podem ser vistos como um compromisso entre os pesos utilizados em Q_{MH} e Q_G. Estes autores foram os primeiros a argumentar que os testes de Mantel-Haenszel e Gehan podem ser considerados variantes de uma mesma classe de testes, cuja estatística de teste tem a forma comum (1.7), levando a que esta classe seja frequentemente apelidada de classe de testes Tarone-Ware.

Alguns anos mais tarde, Peto e Peto [66] e Prentice [67], movidos pelo mesmo objetivo que Tarone e Ware, propuseram a estatística de teste Q_{PP}. É importante enfatizar que os pesos utilizados no teste de Peto-Prentice (também designado por teste de Peto-Peto) são bastante similares ao estimador de Kaplan-Meier da função de sobrevivência comum aos dois grupos de indivíduos [54]. Desta forma, os pesos utilizados em Q_{PP} dependem do

número total de indivíduos sobreviventes a cada instante $t_{(k)}$ e não são afetados por possíveis padrões de censura distintos, como acontece no teste de Gehan.

Como a estatística de teste (1.7) segue uma distribuição que é assintótica, a aplicação de um teste pertencente a esta classe requer algum cuidado na interpretação de resultados, principalmente em amostras de pequena dimensão e/ou quando existem muitas observações censuradas, nunca esquecendo o pressuposto de independência entre os tempos de vida e de censura.

1.8.3 Comparação de três ou mais grupos

Todos os testes atrás referidos podem ser generalizados com o intuito de comparar G grupos homogéneos de indivíduos, em que $G \geq 3$. Nesta situação, sem perda de generalidade, as hipóteses a testar são

$$H_0 : S_1(t) = S_2(t) = \ldots = S_G(t) \qquad \text{vs} \qquad H_1 : \exists g_1, g_2 : S_{g_1} \neq S_{g_2}, \;\; g_1 \neq g_2.$$

as quais podem ser testadas através de uma extensão da estatística de teste (1.7). Para tal, é necessário acomodar a informação pertinente sobre cada instante de morte $t_{(k)}$ numa tabela de contingência $G \times 2$.

Sob a validade de H_0, é possível demonstrar que o número de mortes observadas nos G grupos de indivíduos tem distribuição hipergeométrica multivariada de dimensão $G-1$ [54]. Analogamente à construção do teste *log-rank*, considere-se a distribuição da frequência d_{gk} (relativa ao g-ésimo grupo) condicional aos totais marginais da tabela $G \times 2$, cujo número esperado de mortes em $t_{(k)}$ é $E(d_{gk}) = n_{gk} \times {}^{d_k}/_{n_k}$.

Por conseguinte, combinando a informação contida nas r tabelas de contingência, constrói-se a seguinte estatística

$$U_g = \sum_{k=1}^{r} w_k \big[d_{gk} - E(d_{gk}) \big], \qquad g = 1, 2, \ldots, G-1,$$

de onde se obtém o vetor $\boldsymbol{U} = (U_1, U_2, \ldots, U_{G-1})$. Além do mais, para cada tabela de contingência, é preciso determinar a matriz de covariância $\boldsymbol{V}(t_{(k)})$ de dimensão $(G-1) \times (G-1)$, cujo elemento genérico (g_1, g_2) é dado por

$$\boldsymbol{V}(t_{(k)}) = \operatorname{cov}(d_{g_1 k}, d_{g_2 k}) = \begin{cases} w_k^2 \dfrac{n_{g_1 k}(n_k - n_{g_1 k}) d_k (n_k - d_k)}{n_k^2 (n_k - 1)}, & g_1 = g_2 \\ -w_k^2 \dfrac{n_{g_1 k} \times n_{g_2 k} \times d_k (n_k - d_k)}{n_k^2 (n_k - 1)}, & g_1 \neq g_2 \end{cases},$$

para $g_1, g_2 = 1, \ldots, G-1$, de onde se obtém a matriz de covariância global $\boldsymbol{V} = \sum_{k=1}^{r} \boldsymbol{V}(t_{(k)})$. Por último, deduz-se a estatística de teste $\boldsymbol{Q} = \boldsymbol{U}'\boldsymbol{V}^{-1}\boldsymbol{U}$ que, sob H_0, segue uma distribuição assintótica χ^2_{G-1}.

Capítulo 2

Modelo de regressão de Cox

2.1 Introdução

Em 1972, o estatístico britânico Sir David Cox [27] propôs um modelo de regressão que impulsionou o desenvolvimento da análise de dados de sobrevivência. Apesar da utilização deste modelo estar mais associada à análise de tempos de vida, a sua versatilidade permite a resolução de diversas situações práticas, não só na área da medicina, como também em economia, psicologia, ciências políticas, entre outras, razão pela qual este modelo se tornou rapidamente popular.

A aplicação do modelo de Cox a dados longitudinais procura avaliar o prognóstico dos indivíduos em estudo. Assim, neste modelo de regressão, considera-se como variável resposta o tempo de vida e os diversos fatores passíveis de a influenciar são representados por variáveis explanatórias, como referido na secção 1.5.

Tendo em conta que os indivíduos em estudo podem diferir entre si relativamente a fatores de risco, é natural que esse conjunto de indivíduos constitua um grupo heterogéneo. Neste ponto, a utilização deste modelo tem-se demonstrado uma mais-valia, quer pela estimação do efeito das covariáveis, quer pelo seu ajustamento perante a existência de fatores de confundimento.

Ao longo deste capítulo será realizada uma descrição detalhada do modelo de Cox. Inicialmente, será apresentada a sua definição seguida pela metodologia utilizada para a obtenção dos estimadores dos parâmetros de regressão. Posto isto, serão abordados alguns dos testes de hipóteses utilizados na sua validação, onde a análise de resíduos se revela uma etapa fundamental. Por fim, serão analisadas duas extensões deste modelo que, a seu tempo, revelar-se-ão úteis na definição de modelos para acontecimentos múltiplos e que serão abordados no capítulo seguinte.

2.2 Formulação do modelo de Cox

O modelo de regressão de Cox, permite estudar a relação existente entre o tempo de vida (variável dependente) e as variáveis explanatórias (variáveis independentes), numa população heterogénea [24]. Por conseguinte, seja T uma v.a. contínua, com vista a especificar um modelo para a distribuição do tempo de vida, dado um vetor $\boldsymbol{z}$ de covariáveis associado a cada indivíduo.

Face à importância da função de risco na Análise de Sobrevivência, Cox [27] introduziu um modelo de regressão tendo por base essa mesma função. Deste modo, seja n o número de indivíduos em estudo, sendo que para cada indivíduo i, com $i = 1, \ldots, n$, considera-se o vetor $(t_i, \delta_i, \boldsymbol{z}_i)$, onde t_i diz respeito ao tempo de vida observado, δ_i adverte para a ocorrência ou não do acontecimento de interesse e $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})'$ é o vetor que representa os valores observados para cada uma das p covariáveis. Consequentemente, a função de risco pode ser apresentada na forma

$$h(t; \boldsymbol{z}_i) = h_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}_i), \qquad i = 1, \ldots, n, \tag{2.1}$$

onde $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ designa o vetor de parâmetros ou coeficientes de regressão e corresponde ao efeito desconhecido das p covariáveis na sobrevivência de cada indivíduo. A função de risco subjacente, $h_0(t)$, representa uma função de risco arbitrária não negativa, à qual está associado o vetor de covariáveis nulo, correspondente à situação padrão.

Como é possível observar, a função de risco (2.1) estabelece uma relação entre duas componentes:

i) A primeira, $h_0(t)$, diz respeito a uma função de risco não especificada, dependente do tempo e comum a todos os indivíduos em estudo. É de salientar que, para um indivíduo na situação padrão, cujo vetor de covariáveis $\boldsymbol{z} = \mathbf{0}$, através da expressão (2.1) deduz-se que $h(t; \mathbf{0}) = h_0(t)$. Por esse motivo, esta componente é frequentemente denominada por função de risco subjacente (*hazard baseline function*) e daí dizer-se que $h_0(t)$ é a função de risco associada a um indivíduo padrão;

ii) No que diz respeito à segunda componente, $\exp(\boldsymbol{\beta}' \boldsymbol{z}_i)$, esta é conhecida como risco relativo e expressa a influência que as covariáveis exercem sobre a função de risco. O modelo de Cox clássico pressupõe que esta influência não sofre qualquer alteração durante o período de estudo. Note-se que as covariáveis têm um efeito multiplicativo na função de risco. Em contraste com a componente referida anteriormente, $\exp(\boldsymbol{\beta}' \boldsymbol{z}_i)$ não depende de t, mas sim dos valores observados para cada uma das p covariáveis e dos coeficientes de regressão $\boldsymbol{\beta}$.

Ao considerar dois indivíduos em estudo, a e b, a partir de (2.1) pode-se construir a razão dos riscos (*hazard ratio*), também designada por risco de morte ou risco relativo,

$$\frac{h(t;\boldsymbol{z}_a)}{h(t;\boldsymbol{z}_b)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{z}_a)}{h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{z}_b)} = \exp\left[\boldsymbol{\beta}'(\boldsymbol{z}_a - \boldsymbol{z}_b)\right], \tag{2.2}$$

de onde se conclui que esta não depende da função de risco subjacente $h_0(t)$ e consequentemente não depende de t. Além disso, as funções de risco de dois quaisquer indivíduos são proporcionais e, por este motivo, o modelo de Cox pode ser considerado um modelo de riscos proporcionais.

Note-se que, apesar do efeito das covariáveis $\boldsymbol{\beta}$ ser modelado parametricamente no modelo de regressão (2.1), a existência de uma função de risco subjacente não especificada $h_0(t)$ faz com que esta abordagem seja classificada como semiparamétrica.

Em análise de regressão, a formulação de um modelo consiste muitas vezes em modelar linearmente o efeito das covariáveis. O mesmo pode ser conseguido com o modelo de Cox. Para tal, considere-se dois indivíduos a que estão associados os vetores de covariáveis $\boldsymbol{z}$ e $\mathbf{0}$. Aplicando o logaritmo à razão dos riscos (2.2), tem-se que

$$\log\left(\frac{h(t;\boldsymbol{z})}{h(t;\mathbf{0})}\right) = \log h(t;\boldsymbol{z}) - \log h_0(t) = \beta_1 z_1 + \ldots + \beta_p z_p,$$

onde $\sum_{j=1}^{p}\beta_j z_j$ diz respeito à componente linear do modelo. É de salientar que, para um dado indivíduo i, $i = 1,\ldots,n$, a quantidade $\boldsymbol{\beta}'\boldsymbol{z}_i = \sum_{j=1}^{p}\beta_j z_{ij}$ é usualmente conhecida como índice de prognóstico ou *risk score*, pois β_j representa o efeito da covariável z_{ij} no tempo de vida do i-ésimo indivíduo. Através de uma análise mais atenta, conclui-se que a utilização desta escala faz com que o efeito produzido pelas covariáveis seja aditivo. Por essa razão, o modelo de Cox pode ser encarado como um modelo linear.

O modelo de Cox (2.1) pode ainda ser formulado através dos processos de contagem (consultar anexo A). Esta alternativa revela-se útil na análise de dados mais complexos, como acontece no âmbito dos acontecimentos múltiplos em dados longitudinais, que serão abordados no Capítulo 3. Todavia, para o desenvolvimento das secções seguintes não é importante considerar este tipo de formulação, dado que neste momento o interesse encontra-se direcionado para a análise do tempo até à ocorrência de um único acontecimento de interesse.

2.3 Estimação

Após a formulação do modelo semiparamétrico de Cox, pretende-se estimar o vetor de parâmetros de regressão e a função de sobrevivência. Para isso, é necessário proceder à construção da função de verosimilhança, a partir da qual se obtêm os estimadores dos parâmetros. As estimativas obtidas a partir destes estimadores permitem a realização de inferência estatística sobre os dados.

Como analisado na secção anterior, a função de risco subjacente $h_0(t)$ não é especificada parametricamente, pelo que a utilização da função de verosimilhança habitual (1.5) torna-se inadequada para a estimação do vetor de parâmetros $\boldsymbol{\beta}$. Perante esta situação, ainda em 1972, Cox [27] propôs uma função não dependente de $h_0(t)$, razão pela qual a estimação será feita em separado.

2.3.1 Parâmetros de regressão

Considere-se que existem n indivíduos em estudo, sujeitos a um mecanismo de censura não informativa, e que foram observados r tempos de vida distintos $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, $r \leq n$. Suponha-se que $R(t_{(k)}) = R_k$ representa o conjunto de índices relativos aos indivíduos em risco imediatamente antes do instante $t_{(k)}$ e seja $\boldsymbol{z}_k$ o vetor de variáveis explanatórias associado ao indivíduo que morre em $t_{(k)}$.

Em 1972, Cox [27] sugeriu a seguinte função de verosimilhança para a estimação de $\boldsymbol{\beta}$,

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{r} \frac{\exp\left(\boldsymbol{\beta}' \boldsymbol{z}_k\right)}{\sum\limits_{l \in R_k} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}, \tag{2.3}$$

a qual pode ser interpretada como uma verosimilhança parcial. Dois anos mais tarde, viria a mostrar que a função (2.3) permite a estimação dos parâmetros β_j $(j = 1, \ldots, p)$ na presença de fatores perturbadores[1] [28]. Como é possível observar-se, (2.3) não é uma função de verosimilhança no sentido usual, uma vez que a probabilidade de realização de um acontecimento observável não se encontra representada. Importa salientar que, apesar das observações censuradas não contribuírem para a estimação de $\boldsymbol{\beta}$, estas são imprescindíveis para a identificação do conjunto de risco, R_k.

Demonstrou-se, através dos processos de contagem, que para amostras de grande dimensão, o estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$, representado por $\widehat{\boldsymbol{\beta}}$, goza das mesmas propriedades que os obtidos pelo método de

[1] Nesta situação, a função de risco subjacente, $h_0(t)$, é vista como uma função perturbadora.

máxima verosimilhança [8]. Assim, $\widehat{\boldsymbol{\beta}}$ é consistente, assintoticamente gaussiano multivariado de dimensão p, com valor médio $\boldsymbol{\beta}$ e matriz de covariância $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})$, sendo a matriz de informação de Fisher dada por

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\left[E\left(\frac{\partial^2 \log L}{\partial\beta_j \partial\beta_m}\right)\right]_{p\times p}.$$

Os parâmetros de regressão, $\boldsymbol{\beta}$, são estimados a partir dos valores $\widehat{\boldsymbol{\beta}}$ que maximizam (2.3). Como é habitual, neste contexto, determina-se o máximo da função (2.3) aplicando o logaritmo, resultando na função logverosimilhança parcial,

$$\log L(\boldsymbol{\beta}) = \sum_{k=1}^{r}\left\{\boldsymbol{\beta}'\boldsymbol{z}_k - \log\left[\sum_{l\in R_k}\exp(\boldsymbol{\beta}'\boldsymbol{z}_l)\right]\right\}. \tag{2.4}$$

Assim, o estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}' = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$, é obtido derivando (2.4) em ordem a cada um dos parâmetros, β_j, $j = 1, \ldots, p$, e igualando essas derivadas a zero. Consequentemente, obtém-se a função *score*,

$$U_j(\boldsymbol{\beta}) = \frac{\partial \log L(\boldsymbol{\beta})}{\partial\beta_j} = \sum_{k=1}^{r}\left[z_{kj} - \frac{\sum\limits_{l\in R_k} z_{lj}\exp(\boldsymbol{\beta}'\boldsymbol{z}_l)}{\sum\limits_{l\in R_k}\exp(\boldsymbol{\beta}'\boldsymbol{z}_l)}\right], \quad j = 1, \ldots, p \tag{2.5}$$

a partir da qual também se obtém o vetor $\boldsymbol{U}(\boldsymbol{\beta}) = \big(U_1(\boldsymbol{\beta}), \ldots, U_p(\boldsymbol{\beta})\big)'$, designado por vetor *score*. Deste modo, o estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$ é a solução do sistema de equações $U_j(\boldsymbol{\beta}) = 0$, para $j = 1, \ldots, p$.

Apesar de se estar a considerar apenas o caso contínuo, mesmo neste caso, podem ocorrer tempos de vida iguais (empatados), dependendo da unidade de medida utilizada [54]. Para essas situações, vários autores sugeriram modificações à função de verosimilhança parcial (2.3), das quais apenas será desenvolvida aquela sugerida por Peto [65] e por Breslow [18]. Sejam d_k o número de mortes observadas no instante $t_{(k)}$ e $\widetilde{\boldsymbol{z}}_k = \sum_{l=1}^{d_k}\boldsymbol{z}_{kl}$ a soma dos vetores de covariáveis dos d_k indivíduos que morreram nesse instante. Então,

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{r}\frac{\exp\left(\boldsymbol{\beta}'\widetilde{\boldsymbol{z}}_k\right)}{\left[\sum\limits_{l\in R_k}\exp(\boldsymbol{\beta}'\boldsymbol{z}_l)\right]^{d_k}}, \tag{2.6}$$

donde, pelo mesmo raciocínio seguido anteriormente,

$$\log L(\boldsymbol{\beta}) = \sum_{k=1}^{r} \left\{ \boldsymbol{\beta}' \widetilde{\boldsymbol{z}}_k - d_k \log \left[\sum_{l \in R_k} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l) \right] \right\}.$$

Desta forma, quando não existem empates, isto é, quando $d_k = 1$ e $\widetilde{\boldsymbol{z}}_k = \boldsymbol{z}_k$, $L(\boldsymbol{\beta})$ e $\log L(\boldsymbol{\beta})$ reduzem-se a (2.3) e a (2.4), respetivamente. Tem-se também que

$$U_j(\boldsymbol{\beta}) = \sum_{k=1}^{r} \left[\widetilde{z}_{kj} - d_k \frac{\sum\limits_{l \in R_k} z_{lj} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_k} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)} \right], \quad j = 1, \ldots, p$$

e, para além disso, através da relação

$$I_{jm}(\boldsymbol{\beta}) = -\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_m} = -\frac{\partial U_j(\boldsymbol{\beta})}{\partial \beta_m},$$

para cada componente que se encontra na linha j e na coluna m da matriz de informação de Fisher tem-se que

$$I_{jm}(\boldsymbol{\beta}) = \sum_{k=1}^{r} d_k \left[\frac{\sum\limits_{l \in R_k} z_{lj} z_{lm} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_k} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)} - \frac{\sum\limits_{l \in R_k} z_{lj} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l) \sum\limits_{l \in R_k} z_{lm} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}{\left(\sum\limits_{l \in R_k} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l) \right)^2} \right],$$

com $j, m = 1, \ldots, p$.

Interpretação dos parâmetros de regressão

No decorrer de uma análise de regressão, a interpretação dos parâmetros é uma etapa fundamental para a compreensão da relação existente entre a variável resposta (neste caso o tempo de vida) e as variáveis explanatórias. A utilização do modelo de Cox sugere que a interpretação dos parâmetros de regressão seja feita em termos de $\exp(\beta_j)$, $j = 1, \ldots, p$, uma vez que esta quantidade está diretamente relacionada com o risco de morte.

Por conseguinte, considere-se novamente dois indivíduos, a e b, a que estão associados os vetores de covariáveis $\boldsymbol{z}_a$ e $\boldsymbol{z}_b$, respetivamente. Suponha-se ainda que para uma das p covariáveis, digamos c, esses indivíduos diferem apenas no valor observado dessa covariável, de tal forma que $z_{ac} - z_{bc} = 1$. Nesta situação, com base na expressão (2.2), a razão dos riscos é dada por

$$\frac{h(t;\boldsymbol{z}_a)}{h(t;\boldsymbol{z}_b)} = \frac{h_0(t)\exp(\beta_1 z_{a1} + \ldots + \beta_c z_{ac} + \ldots + \beta_p z_{ap})}{h_0(t)\exp(\beta_1 z_{b1} + \ldots + \beta_c z_{bc} + \ldots + \beta_p z_{bp})} = \exp(\beta_c).$$

Consequentemente, obtém-se a relação $h(t;\boldsymbol{z}_a) = \exp(\beta_c) \times h(t;\boldsymbol{z}_b)$ e assim a interpretação pode ser sumarizada em três situações possíveis, consoante o valor de β_c:

- se $\beta_c = 0 \Leftrightarrow \exp(\beta_c) = 1$, pelo que a covariável z_c não influencia significativamente o risco de morte dos indivíduos;
- se $\beta_c > 0 \Leftrightarrow \exp(\beta_c) > 1$, significa que o indivíduo a apresenta um risco de morte acrescido, comparativamente ao indivíduo b;
- se $\beta_c < 0 \Leftrightarrow \exp(\beta_c) < 1$, significa que o indivíduo a apresenta um risco de morte inferior ao do indivíduo b.

2.3.2 Função de sobrevivência

A caracterização do modelo de regressão de Cox também pode ser feita por intermédio da função de sobrevivência. Para isso, através de (1.2), (1.4) e (2.1), tem-se que

$$S(t;\boldsymbol{z}_i) = S_0(t)^{\exp(\boldsymbol{\beta}'\boldsymbol{z}_i)}, \qquad i = 1,\ldots,n, \tag{2.7}$$

onde $S_0(t)$ representa a função de sobrevivência para um indivíduo na situação padrão. Então, para se obter as estimativas de $S(t;\boldsymbol{z}_i)$, $i = 1,\ldots,n$, torna-se necessário estimar a função de risco subjacente, $h_0(t)$, e posteriormente $S_0(t)$. Existem diferentes abordagens que podem ser adotadas para a estimação da função de risco subjacente, ou de forma equivalente, para a estimação da sua função cumulativa, $H_0(t)$. O que se segue, refere-se apenas à abordagem seguida por Breslow [18, 19].

Considere-se que foram observados r tempos de vida distintos, $t_{(1)} < \ldots < t_{(r)}$, $r \leq n$, e que as observações se iniciam em $t_{(0)} = 0$. Assuma-se que entre dois quaisquer instantes de morte consecutivos, $t_{(k-1)}$ e $t_{(k)}$, a distribuição do tempo de vida tem uma função de risco constante e que, no caso de existirem observações censuradas entre esses instantes, essas observações ocorrem em $t_{(k-1)}$.

Tendo em conta que $\widehat{\boldsymbol{\beta}}$ foi obtido a partir da verosimilhança parcial (2.3), a estimativa de $h_0(t)$ para um dado intervalo de tempo $\left(t_{(k-1)},\ t_{(k)}\right]$ é dada por

$$\widehat{h}_k = \frac{d_k}{\left(t_{(k)} - t_{(k-1)}\right) \sum\limits_{i \in R_k} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i\right)}, \tag{2.8}$$

onde d_k diz respeito ao número de mortes observadas nesse intervalo. Assim, $\widehat{h}_k$ representa o rácio entre o número de acontecimentos observados e o número ponderado de indivíduos em risco (pertencentes ao conjunto de risco R_k), sendo que cada um desses indivíduos contribui com um peso $\exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i)$ para esse intervalo.

Com base na expressão (2.8), em 1974, Breslow [18] sugeriu o seguinte estimador para a função de risco cumulativa subjacente,

$$\widehat{H}_0(t) = \sum_{t_{(k)} \leq t} \frac{d_k}{\sum\limits_{i \in R_k} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i\right)}.$$

Este estimador é uma função em escada, onde cada salto corresponde a um instante de morte. Curiosamente, quando $\boldsymbol{\beta} = \mathbf{0}$ o estimador de Breslow coincide com o estimador de Nelson-Aalen, introduzido na secção 1.7.2.

Consequentemente, através das relações (1.2) e (1.3), o estimador de $S_0(t)$ é dado por

$$\widehat{S}_0(t) = \prod_{t_{(k)} \leq t} \exp\left(- \frac{d_k}{\sum\limits_{i \in R_k} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i\right)} \right).$$

Por último, pode-se estimar a função de sobrevivência para um dado indivíduo i com vetor de covariáveis $\boldsymbol{z}_i$,

$$\widehat{S}(t; \boldsymbol{z}_i) = \widehat{S}_0(t)^{\exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i\right)}, \qquad i = 1, \ldots, n.$$

2.4 Testes de hipóteses

Como analisado na secção 1.8, um dos propósitos da Análise de Sobrevivência é a comparação das distribuições do tempo de vida de dois ou mais grupos de indivíduos, que podem diferir numa ou mais características. Em geral, é raro conhecer-se previamente a forma funcional da função de sobrevivência teórica, daí essa comparação ser essencialmente baseada em métodos não paramétricos. Neste ponto, pode-se recorrer à utilização do modelo de Cox, desde que os grupos apresentem funções de risco proporcionais.

No ajustamento de um modelo de regressão procura-se incluir primordialmente as variáveis explanatórias que se revelem mais explicativas, de entre todas as que foram registadas, ou seja, aquelas que têm influência significativa na sobrevivência dos indivíduos. A seleção dessas variáveis é um processo bastante importante, visto que o modelo de regressão final deverá ser o mais parcimonioso possível.

Nesta secção serão introduzidos alguns dos testes de hipóteses mais frequentes para abordar essas questões.

2.4.1 Comparação de distribuições do tempo de vida

Considere-se unicamente o caso mais simples que se refere à comparação das distribuições do tempo de vida de dois grupos de indivíduos que diferem apenas numa característica. Quando a hipótese de riscos proporcionais é válida, o modelo semiparamétrico de Cox pode ser utilizado para testar a hipótese de igualdade de duas curvas de sobrevivência, contra a hipótese alternativa destas serem diferentes [48].

Desta forma, seja z uma covariável indicatriz que define o grupo a que o indivíduo pertence, tomando os valores: zero no caso de o indivíduo pertencer ao grupo 1; e um no caso de o indivíduo pertencer ao grupo 2. Então, sendo $S_1(t)$ e $S_2(t)$ as funções de sobrevivência correspondentes aos dois grupos, relacionadas por intermédio de (2.7), formalmente tem-se que

$$H_0 : S_2(t) = S_1(t) \qquad \text{vs} \qquad H_1 : S_2(t) = S_1(t)^{\exp(\beta)},$$

que é equivalente a testar

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0. \tag{2.9}$$

Considere-se uma amostra constituída por $n = m_1 + m_2$ indivíduos, em que m_1 e m_2 indicam o número de indivíduos pertencentes aos grupos 1 e 2, respetivamente. Sejam $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, $r \leq n$, os instantes de mortes distintos observados na amostra conjunta, n_{gk} o número de indivíduos em risco em $t^-_{(k)}$ no grupo g $(g = 1, 2)$ e d_{gk} o número de mortes observadas em $t_{(k)}$ no grupo g. Assim, o número total de indivíduos em risco imediatamente antes do instante $t_{(k)}$ é obtido por $n_k = n_{1k} + n_{2k}$ e o número total de mortes observadas em $t_{(k)}$ é obtido por $d_k = d_{1k} + d_{2k}$.

Quando existem poucas observações empatadas, particularizando (2.4) a esta situação, tem-se que

$$\log L(\beta) = \widetilde{d}_2\beta - \sum_{k=1}^{r} d_k \log\big(n_{1k} + n_{2k}\exp(\beta)\big),$$

onde $\widetilde{d}_2 = \sum_{k=1}^{r} d_{2k}$, visto que aos indivíduos do grupo 1 corresponde $z = 0$. Então,

$$U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = \widetilde{d}_2 - \sum_{k=1}^{r} \frac{d_k n_{2k} \exp(\beta)}{n_{1k} + n_{2k} \exp(\beta)}$$

e

$$I(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta^2} = \sum_{k=1}^{r} \frac{d_k n_{1k} n_{2k} \exp(\beta)}{\big(n_{1k} + n_{2k} \exp(\beta)\big)^2}.$$

Como $U(\beta)$ tem distribuição assintoticamente gaussiana de valor médio zero e variância $I(\beta)$, um teste bastante simples para testar (2.9) e que não implica o cálculo de $\widehat{\beta}$ é o teste *score*

$$Z = \frac{U(0)}{\sqrt{I(0)}}, \tag{2.10}$$

que segue uma distribuição assintótica $N(0,1)$, onde

$$U(0) = \sum_{k=1}^{r} \left(d_{2k} - \frac{d_k n_{2k}}{n_{1k} + n_{2k}} \right) \qquad \text{e} \qquad I(0) = \sum_{k=1}^{r} \frac{d_k n_{1k} n_{2k}}{n_k^2}.$$

Repare-se que, sob a validade de H_0, Z^2 apresenta uma distribuição assintótica χ^2_1, o que significa que este teste é equivalente ao teste *log-rank*. Por esse motivo, (2.10) também é designado por teste de Cox-Mantel.

No caso de existir um número considerável de observações empatadas, é necessário proceder a uma ligeira alteração em (2.10), mais precisamente em $I(0)$, de modo a que este teste tenha em conta a natureza discreta dos dados. Assim,

$$I(0) = \sum_{k=1}^{r} \frac{d_k n_{1k} n_{2k} \big(n_k - d_k\big)}{n_k^2 \big(n_k - 1\big)},$$

pelo que as estatísticas coincidem quando $d_k = 1$ para todos os instantes de morte $t_{(k)}$, $k = 1, \ldots, r$.

2.4.2 Seleção de variáveis

Numa análise de regressão procura-se construir um modelo que se ajuste o melhor possível aos dados disponíveis, sendo por isso necessário identificar quais as variáveis explanatórias que influenciam significativamente os tempos de vida dos indivíduos. Neste ponto, a contribuição dos profissionais ligados à área em se está a efetuar o estudo é fundamental pois, embora o modelo de regressão final deva ser parcimonioso, podem existir variáveis que não se tenham revelado estatisticamente significativas e que sejam relevantes para serem incluídas no modelo [22, 71].

Pretende-se então avaliar se existe evidência de que uma dada variável explanatória z_j $(j = 1, \ldots, p)$ tem influência significativa na sobrevivência dos indivíduos, na presença das restantes variáveis. Visto que o parâmetro de regressão β_j quantifica o efeito da variável explanatória z_j no modelo de Cox, pode-se testar

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

Para tal, é frequente utilizar-se o teste de Wald, cuja estatística de teste tem a forma

$$W = \frac{\widehat{\beta}_j^2}{\text{var}\big(\widehat{\beta}_j\big)}, \tag{2.11}$$

e que, sob H_0, segue uma distribuição assintótica χ_1^2. Quando se aplica este teste pressupõe-se que as estimativas $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ são independentes umas das outras. Porém, isso nem sempre se verifica, tornando difícil a interpretação dos resultados obtidos acerca dos parâmetros associados às variáveis explanatórias incluídas nesse modelo. Em alternativa, é comum recorrer-se a métodos que permitam comparar modelos encaixados (*nested models*), ou seja, que permitam comparar submodelos do modelo de regressão original [24].

Suponha que se pretende comparar dois modelos de Cox, um com u variáveis explanatórias (modelo 1) e outro em que estão incluídas v variáveis explanatórias adicionais (modelo 2). Desta forma, diz-se que o modelo 1 encontra-se encaixado no modelo 2, pois tem-se que

Modelo 1: $h_0(t)\exp(\beta_1 z_1 + \ldots + \beta_u z_u)$;

Modelo 2: $h_0(t)\exp(\beta_1 z_1 + \ldots + \beta_u z_u + \beta_{u+1} z_{u+1} + \ldots + \beta_{u+v} z_{u+v})$.

Tenciona-se então testar se as v parcelas adicionadas ao modelo 2 melhoram significativamente a capacidade explicativa desse modelo, comparativamente

ao modelo 1. Se isso não se verificar, conclui-se que o modelo 1 é aquele que melhor se ajusta aos dados disponíveis, pelo que as v parcelas podem ser omitidas. Formalmente, as hipóteses de interesse são

$$H_0 : \beta_{u+1} = \ldots = \beta_{u+v} = 0 \qquad \text{vs} \qquad H_1 : \exists j : \beta_j \neq 0,\ j = u+1, \ldots, u+v.$$

Uma medida adequada para mensurar a qualidade do ajustamento de um modelo aos dados e que, por sua vez, permite comparar modelos encaixados, é o valor da função de verosimilhança quando os parâmetros de regressão são substituídos pelas respetivas estimativas de máxima verosimilhança. De facto, toda a informação disponível a respeito dos parâmetros desconhecidos, encontra-se sintetizada nesta função. Por questões de interpretação[2], considera-se a estatística $-2\log\widehat{L}$, em que $\widehat{L}$ representa a função de verosimilhança maximizada. Assim, quanto maior for o valor de $\widehat{L}$ para um certo modelo, menor será o valor da estatística $-2\log\widehat{L}$ e, consequentemente, melhor será o seu ajustamento. Contudo, esta medida só pode ser utilizada para comparar modelos ajustados aos mesmos dados, uma vez que o cálculo de $\widehat{L}$ depende da dimensão da amostra. Desta forma, as hipóteses de interesse anteriores podem ser testadas através da diferença entre os valores da estatística $-2\log\widehat{L}$ calculada para cada modelo,

$$RV = -2\left(\log\widehat{L}_1 - \log\widehat{L}_2\right) = -2\log\left(\frac{\widehat{L}_1}{\widehat{L}_2}\right), \tag{2.12}$$

de onde se obtém o designado teste da razão de verosimilhanças. Sob a validade de H_0, esta estatística de teste tem distribuição assintótica qui-quadrado com número de graus de liberdade igual à diferença entre o número de covariáveis em cada modelo, ou seja, χ^2_v.

A comparação de modelos encaixados é por muitos considerada o ponto chave na seleção de variáveis a incluir no modelo final. Grande parte dos *softwares* estatísticos têm incorporados métodos que selecionam automaticamente essas variáveis, designadamente a seleção *forward*, a eliminação *backward* e, ainda, uma combinação entre estas duas, denominada por procedimento *stepwise*. Porém, Collett [24] alerta para algumas desvantagens na aplicação destes métodos como, por exemplo, o facto dos submodelos encontrados dependerem do método utilizado (muitas vezes originam submodelos diferentes, sendo difícil decidir qual o método mais apropriado) e do seu critério de paragem (usado para decidir se uma

[2]Note-se que, $\widehat{L}$ resulta da multiplicação de probabilidades condicionais pelo que o seu valor será sempre inferior à unidade e, por isso, $-2\log\widehat{L}$ tomará sempre um valor positivo.

dada variável deve ou não ser incluída no modelo). Consciente destas limitações, o autor sugeriu uma estratégia para a seleção do modelo que melhor se ajusta aos dados disponíveis e que está bem relatada no seu livro. Além do mais, recomenda que o critério de paragem para a inclusão ou omissão das variáveis, se baseie num nível de significância $\alpha \simeq 0.10$.

2.5 Análise de Resíduos

Posteriormente ao ajustamento de um modelo de regressão, surge a necessidade de avaliar a sua adequabilidade, o que requer que se efetue uma análise de resíduos. Quando se pensa em resíduos, a definição mais natural é a de diferença entre o valor observado da variável resposta e o valor predito pelo modelo ajustado, como ocorre na análise de regressão linear [71]. Contudo, os modelos da Análise de Sobrevivência têm determinadas características que os distinguem dos modelos de regressão habituais, nomeadamente a possibilidade de existirem observações censuradas (quando uma dada observação é censurada o seu resíduo também o é) e a própria configuração do modelo de Cox. Desta forma, a definição de resíduo deixa de ser tão óbvia e, por isso, têm sido sugeridas diferentes abordagens para o seu cálculo, dependendo do que se pretende avaliar.

No contexto da análise de tempos de vida, os resíduos podem ser utilizados para examinar diferentes aspetos relacionados com o ajustamento do modelo [54], tais como:

i) validação da hipótese de riscos proporcionais;

ii) identificação de observações influentes, isto é, valores atípicos mas, de certo modo, coerentes com a tendência geral das restantes observações;

iii) investigação da forma funcional do modo como cada covariável influencia a variável resposta (na presença das restantes covariáveis);

iv) identificação de *outliers*, isto é, observações discrepantes cujos valores foram mal preditos pelo modelo.

Habitualmente, na aplicação do modelo de Cox, as observações influentes ocorrem em indivíduos com tempos de vida longos, enquanto que os *outliers* surgem em indivíduos para os quais se observou o acontecimento de interesse demasiado cedo ou demasiado tarde, comparativamente aos indivíduos com características semelhantes.

Nesta secção serão então introduzidos quatro tipos de resíduos, desenvolvidos para diagnosticar cada um dos aspetos anteriores.

2.5.1 Resíduos de Schoenfeld

Em 1982, Schoenfeld [74] propôs um tipo de resíduos com o intuito de testar a proporcionalidade das funções de risco, tanto no geral como para cada covariável individualmente. Para isso, sugeriu que estes fossem obtidos por intermédio da função *score* (2.5) baseada na função de verosimilhança parcial, o que levou a que inicialmente se designassem de resíduos parciais [24]. Como no modelo de Cox a cada indivíduo corresponde um vetor de covariáveis, então para cada um deles também irá corresponder um conjunto de resíduos, tantos quanto o número de covariáveis incluídas no modelo.

Através de um olhar mais atento a (2.5) conclui-se que, para um indivíduo que morre em $t_{(k)}$, $k = 1, \ldots, r$, a função *score* representa a diferença entre o valor da variável aleatória z_{kj} ($j = 1, \ldots, p$) e o seu valor esperado condicional ao conjunto de indivíduos em risco R_k, pelo que esta função pode ser escrita da seguinte forma

$$U_j(\boldsymbol{\beta}) = \sum_{k=1}^{r} \Big[z_{kj} - E(z_{kj}|R_k) \Big], \text{ com } E(z_{kj}|R_k) = \frac{\sum\limits_{l \in R_k} z_{lj} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_k} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)},$$

onde os estimadores de máxima verosimilhança parcial de $\boldsymbol{\beta}$ são a solução do sistema de p equações $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$, como já foi referido na secção 2.3.1. Alternativamente, a função *score* pode ser expressa para todos os indivíduos em estudo, ou seja,

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \Big[z_{ij} - E(z_{ij}|R_i) \Big], \text{ com } E(z_{ij}|R_i) = \frac{\sum\limits_{l \in R_i} z_{lj} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}, \quad (2.13)$$

em que

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

Por conseguinte, Schoenfeld [74] sugeriu que os resíduos fossem obtidos substituindo $\boldsymbol{\beta}$ por $\widehat{\boldsymbol{\beta}}$ na expressão (2.13). Assim, para o i-ésimo indivíduo em estudo, o resíduo de Schoenfeld referente à covariável z_j é definido como

$$\widehat{r}_{P_{ij}} = \delta_i \left[z_{ij} - \widehat{E}(z_{ij}|R_i) \right], \quad (2.14)$$

onde

$$\widehat{E}(z_{ij}|R_i) = \frac{\sum\limits_{l \in R_i} z_{lj} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}.$$

Deste modo, para um indivíduo cuja morte foi observada em t_i, (2.14) representa a diferença entre o valor observado da covariável z_j para esse indivíduo e a média ponderada dos valores observados dessa covariável para todos os indivíduos pertencentes a R_i, sendo que $\exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)$ é o peso associado a cada um deles. Apesar destes resíduos serem sempre nulos para um indivíduo cujo tempo de vida não foi observado, o mais usual é que estes sejam representados por valores omissos, de modo a que não sejam confundidos com aqueles que, de facto, tomam o valor zero.

Para todos os indivíduos em estudo, a soma dos resíduos de Schoenfeld referentes a cada covariável é igual a zero. Além disso, quando n é suficientemente grande, estes resíduos são não correlacionados e têm valor esperado nulo. Então, quando um dado modelo ajustado é adequado, a representação gráfica dos resíduos de Schoenfeld *versus* os tempos de vida (ou as suas ordens) deve revelar uma nuvem aleatória de pontos em torno de zero.

Em 1994, Grambsch e Therneau [36] sugeriram uma modificação nos resíduos (2.14), com o objetivo de aumentar a sua capacidade em detetar afastamentos do modelo assumido, os quais se designam por resíduos de Schoenfeld padronizados ou ponderados. Designe-se por $\widehat{\boldsymbol{r}}_{P_i} = (\widehat{r}_{P_{i1}}, \widehat{r}_{P_{i2}}, \ldots, \widehat{r}_{P_{ip}})$ o vetor de resíduos de Schoenfeld correspondente ao i-ésimo indivíduo. Assim, os resíduos de Schoenfeld padronizados $\widehat{r}^{\,*}_{P_{ij}}$ podem ser definidos como sendo as componentes do vetor

$$\widehat{\boldsymbol{r}}^{\,*}_{P_i} = r \times \mathrm{var}\big(\widehat{\boldsymbol{\beta}}\big) \times \widehat{\boldsymbol{r}}_{P_i},$$

onde r é o número total de mortes observadas no estudo e $\mathrm{var}\big(\widehat{\boldsymbol{\beta}}\big)$ é a matriz de covariância do estimador $\widehat{\boldsymbol{\beta}}$ no modelo ajustado. Estes autores argumentaram que, para testar a hipótese de riscos proporcionais, era preciso considerar a possibilidade do efeito da covariável z_j variar ao longo do tempo, assumindo a seguinte forma

$$\beta_j(t) = \beta_j + \theta_j \left(g_j(t) - \overline{g}_j\right), \tag{2.15}$$

onde $g_j(t)$ é uma função conhecida calculada para cada instante t_i, θ_j é o seu declive e $\overline{g}_j$ é a média dos $g_j(t_i)$. Repare-se que, quando $\theta_j = 0$ o efeito da covariável z_j é constante, uma vez que $\beta_j(t) = \beta_j$ e, assim, a hipótese de riscos proporcionais é satisfeita.

Grambsch e Therneau [36] demonstraram que o valor médio do resíduo de Schoenfeld padronizado no instante t_i é dado por

$$E\big(\widehat{r}^{\,*}_{P_{ij}}\big) \approx \beta_j(t_i) - \widehat{\beta}_j,$$

em que $\beta_j(t_i)$ é o parâmetro da covariável z_j calculado para o instante t_i e $\widehat{\beta}_j$ é a estimativa de β_j do modelo de Cox ajustado. Deste modo, através da representação gráfica de $\widehat{r}^{\,*}_{P_{ij}} + \widehat{\beta}_j$ *versus* o tempo é possível averiguar se os resíduos apresentam algum tipo de padrão que demonstre a existência de não proporcionalidade. Para facilitar a interpretação, este gráfico é geralmente complementado por uma curva de suavização (por exemplo, o suavizador *LOWESS*[3]) e, ainda, por um teste formal às hipóteses: $H_0 : \theta_k = 0$ vs $H_1 : \theta_k \neq 0$, com base na estatística de teste (2.15) que, sob H_0, segue uma distribuição assintótica χ^2_1 (para mais detalhes consulte-se Therneau e Grambsch [79]).

2.5.2 Resíduos *score*

Os resíduos *score* foram sugeridos por Therneau *et al.* [80], em 1990, com base na teoria das martingalas e são utilizados para identificar a existência de observações influentes. Estes resíduos também são obtidos por intermédio da função *score* e permitem quantificar a influência que cada indivíduo exerce na estimativa de máxima verosimilhança parcial. Assim, é necessário determinar a diferença que ocorre na estimativa de $\boldsymbol{\beta}$ quando um dado indivíduo i é suprimido da análise, o que significa que para cada um deles é calculada a diferença: $\Delta\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}$, $i = 1, \ldots, n$. Quanto maior for essa diferença maior será a influência que o indivíduo i exerce na estimativa. Por outro lado, se $\Delta\widehat{\boldsymbol{\beta}}$ for igual a zero, tem-se que o i-ésimo indivíduo não tem qualquer influência [22, 79].

Nesta situação, para que não seja necessário estimar $n + 1$ modelos de Cox, em que n é o número total de observações, utiliza-se uma reformulação da função *score* (2.13), da qual resulta

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \delta_i \Big[z_{ij} - E(z_{ij}|R_i) \Big] + \exp(\boldsymbol{\beta}'\boldsymbol{z}_i) \sum_{t_{\tilde{r}} \leq t_i} \frac{\Big[E(z_{\tilde{r}j}|R_{\tilde{r}}) - z_{ij} \Big] \delta_{\tilde{r}}}{\sum\limits_{l \in R_{\tilde{r}}} \exp(\boldsymbol{\beta}'\boldsymbol{z}_l)} \right\},$$

[3] *Locally Weighted Scatterplot Smoother*

com

$$E(z_{ij}|R_i) = \frac{\sum\limits_{l \in R_i} z_{lj} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)},$$

em que δ_i é a variável indicatriz referente ao estado do indivíduo e $R_{\tilde{r}}$ é o conjunto de indivíduos em risco em $t_{\tilde{r}}$, onde $\tilde{r} \leq r$. De acordo com esta definição, o i-ésimo indivíduo apenas contribui para a derivada até ao instante t_i. Por outras palavras, se o estudo fosse concluído em t_i, a i-ésima componente da derivada não seria afetada [24].

Assim sendo, para o i-ésimo indivíduo em estudo, o resíduo *score* referente à covariável z_j é dado por

$$\widehat{r}_{S_{ij}} = \delta_i \left[z_{ij} - \widehat{E}(z_{ij}|R_i) \right] + \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i) \sum_{t_r \leq t_i} \frac{\left[\widehat{E}(z_{rj}|R_r) - z_{ij} \right] \delta_r}{\sum\limits_{l \in R_r} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}.$$

Estes tipo de resíduos também pode ser visto como uma modificação dos resíduos de Schoenfeld, $r_{P_{ij}}$, pois através de (2.14) tem-se que

$$\widehat{r}_{S_{ij}} = \widehat{r}_{P_{ij}} + \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i) \sum_{t_r \leq t_i} \frac{\left[\widehat{E}(z_{rj}|R_r) - z_{ij} \right] \delta_r}{\sum\limits_{l \in R_r} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}.$$

Tal como ocorre nos resíduos de Schoenfeld, a soma dos resíduos *score* correspondente a cada covariável é igual a zero. Todavia, para indivíduos cuja observação foi censurada, estes resíduos não têm que ser obrigatoriamente nulos, o que se revela uma vantagem quando a percentagem de censura é elevada. Além disso, também são úteis para situações em que pode ser observado mais do que um acontecimento de interesse para um mesmo indivíduo, uma vez que permitem uma estimação robusta da variância de $\widehat{\boldsymbol{\beta}}$, como será elucidado no capítulo 3.

Após determinar os valores de $\widehat{r}_{S_{ij}}$ para cada covariável z_j, procede-se à sua representação gráfica, em que os valores de $\widehat{r}_{S_{ij}}$ são representados no eixo das ordenadas e os valores de z_j no eixo das abcissas. Através deste gráfico é possível averiguar quais os valores extremos que influenciam fortemente a estimativa do parâmetro β_j. Para simplificar essa interpretação, também é comum representar os resíduos $\widehat{r}_{S_{ij}}$ ponderados pela estimativa do erro padrão de $\widehat{\beta}_j$.

2.5.3 Resíduos martingala

Os resíduos martingala, assim como os resíduos *score*, fazem parte de uma classe de resíduos inicialmente desenvolvida por Barlow e Prentice [10], em 1988. São assim designados pelo facto de poderem ser obtidos pela transformação de martingalas, no âmbito dos processos de contagem (ver Anexo A). Estes resíduos são úteis para determinar a forma funcional que deve ser considerada para uma dada covariável, de maneira a que o seu efeito na sobrevivência seja explicado o melhor possível, e ainda para detetar observações discrepantes (*outliers*).

Suponha-se que para um indivíduo i $(i = 1, \ldots, n)$ em estudo, $N_i(t)$, $t \geq 0$, é uma função que toma o valor zero até ao instante imediatamente anterior à ocorrência do acontecimento de interesse e que toma o valor um a partir desse instante. Assim, diz-se que $N_i(t)$ é um processo de contagem que indica se foi ou não observado o acontecimento para o i-ésimo indivíduo. Do mesmo modo, o processo $N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{t_i \leq t} \delta_i$, em que δ_i é a variável indicatriz usual, também é um processo de contagem que conta o número de acontecimentos observados na amostra até ao instante t. Em geral, este processo pode ser modelado por uma componente sistemática, que neste caso será a função de risco cumulativa do modelo de Cox, e por uma componente aleatória referente ao erro associado. Desta forma, tem-se que $N(t) = H(t) + M(t)$, donde se obtém a formulação tradicional de o valor observado ser igual à soma do seu valor esperado com o erro resultante da estimativa que, neste contexto, corresponde aos resíduos. Consequentemente, obtém-se o seguinte processo

$$M(t) = N(t) - H(t),$$

conhecido como processo de contagem martingala. Este processo apresenta um valor para cada instante t, mas para o cálculo de resíduos apenas será necessário determinar o seu valor para cada indivíduo no fim do período de *follow-up*.

Então, o resíduo martingala associado ao i-ésimo indivíduo, com vetor de covariáveis $\boldsymbol{z}_i$ e tempo de vida t_i, é definido como

$$\widehat{r}_{M_i} = \delta_i - H\big(t_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\beta}}\big), \tag{2.16}$$

em que $H\big(t_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\beta}}\big)$ é a função de risco cumulativa do modelo de Cox, obtida substituindo $\boldsymbol{\beta}$ pela sua estimativa de máxima verosimilhança parcial, uma vez que o seu verdadeiro valor é desconhecido. Estes resíduos exibem uma grande assimetria, tomando valores no intervalo $(-\infty, 1)$. Note-se que aos indivíduos com observações censuradas $(\delta_i = 0)$ correspondem resíduos negativos.

Os resíduos martingala, quando calculados para o verdadeiro valor de $\boldsymbol{\beta}$, têm valor esperado igual a zero $\big(E(r_{M_i}) = 0\big)$ e são não correlacionados $\big(\text{cov}(r_{M_i}, r_{M_j}) = 0\big)$. Contudo, os resíduos estimados a partir de $\widehat{\boldsymbol{\beta}}$ apresentam uma ligeira correlação negativa $\big(\text{cov}(\widehat{r}_{M_i}, \widehat{r}_{M_j}) < 0\big)$. Para além disso, a soma dos resíduos estimados baseados em $\widehat{\boldsymbol{\beta}}$ é igual a zero, ou seja, $\sum_{i=1}^{n} \widehat{r}_{M_i} = 0$.

Analisando atentamente a expressão (2.16), conclui-se que os resíduos martingala são na realidade estimados através da diferença entre o número observado de acontecimentos para um dado indivíduo e o respetivo número esperado, obtido com base no modelo ajustado. Os resíduos assim calculados irão revelar os indivíduos que se encontram mal ajustados pelo modelo, os quais se designam por *outliers*, sendo que estes podem surgir por duas razões: i) quando o acontecimento de interesse ocorre demasiado cedo, apesar do seu risco de ocorrência ser baixo; ou ii) quando o acontecimento de interesse ocorre demasiado tarde, apesar do seu risco de ocorrência ser elevado. Para verificar a sua existência, procede-se à representação gráfica dos resíduos $\widehat{r}_{M_i}$ *versus* o índice $\widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i$ de cada indivíduo.

A análise de resíduos martingala, para além de detetar *outliers*, também permite averiguar se uma dada covariável deve ser incluída no modelo tal como foi registada ou se deve ser efetuada alguma transformação na sua forma funcional. Para isso, representa-se graficamente os resíduos *versus* os valores de uma dada covariável. Porém, Therneau e Grambsch [79] apresentaram uma abordagem mais simples que consiste em representar, no eixo das ordenadas, os resíduos martingala estimados por um modelo sem covariáveis (modelo nulo) e no eixo das abcissas, os valores de cada uma das covariáveis inseridas no modelo. Estes autores também mostraram que se o modelo correto para uma dada covariável z_j é da forma $\exp\big(\beta_j f(z_j)\big)$, para uma certa função suave f, então a curva de suavização para essa covariável irá expressar, sob determinadas condições, a forma funcional de f. Quando essa curva for linear, significa que a covariável em questão não necessita de ser transformada. Caso contrário, procede-se à sua transformação, dependendo da forma sugestiva da curva de suavização resultante.

2.5.4 Resíduos *deviance*

Os resíduos *deviance* (por vezes designados de desvios residuais) também foram sugeridos no artigo de Therneau *et al.* [80], em 1990, e são obtidos através de uma normalização dos resíduos martingala. A razão que levou ao seu desenvolvimento prende-se com o facto de os resíduos martingala apresentarem uma distribuição bastante assimétrica em torno de zero, o que dificulta

a identificação de *outliers*. De forma a colmatar essa falta de simetria, para cada indivíduo i em estudo, o resíduo *deviance* é definido como

$$\widehat{r}_{D_i} = \text{sgn}(\widehat{r}_{M_i})\Big\{ -2\big[\widehat{r}_{M_i} + \delta_i \log\left(\delta_i - \widehat{r}_{M_i}\right)\big]\Big\}^{\frac{1}{2}},$$

onde $\text{sgn}(\widehat{r}_{M_i})$ é a função sinal do resíduo martingala $\widehat{r}_{M_i}$ associado ao i-ésimo indivíduo. Quando existem poucas observações censuradas, estes resíduos seguem uma distribuição assintótica gaussiana, o que se revela uma vantagem na deteção de *outliers*.

A motivação que está por detrás da forma como os resíduos *deviance* são obtidos tem a ver com o facto destes serem componentes da estatística *deviance*, definida por $D = -2(\log \widehat{L}_c - \log \widehat{L}_s)$, em que $\widehat{L}_c$ e $\widehat{L}_s$ são as verosimilhanças parciais maximizadas do modelo corrente (modelo que se pretende avaliar) e do modelo saturado (modelo que inclui todas as covariáveis), respetivamente. Como habitualmente, quanto menor for o valor da estatística D, melhor será o modelo em causa. Note-se que esta diferença entre dois modelos é a mesma introduzida na secção 2.4.2, aquando da utilização da estatística $-2\log \widehat{L}$.

Como os resíduos $\widehat{r}_{D_i}$ são tais que $D = \sum_{i=1}^{n} \widehat{r}_{D_i}^{\,2}$, os indivíduos mal ajustados pelo modelo serão aqueles que expressarem resíduos de valor absoluto muito elevados. Carvalho *et al.* [22] recomenda que para efetuar este diagnóstico seja utilizado um dos três gráficos seguintes: resíduos $\widehat{r}_{D_i}$ *versus* índice de prognóstico estimado ou *versus* valores preditos do modelo e o gráfico quantil-quantil (*Q-Q plot*).

2.6 Extensões do Modelo de Cox

O modelo de Cox referido até agora apresenta três aspetos que o caracterizam de forma determinante [54]: i) é um modelo de riscos proporcionais; ii) as variáveis explanatórias nele incluídas têm um valor fixo ao longo do estudo; e iii) as variáveis explanatórias têm um efeito multiplicativo na função de risco.

Nesta secção serão consideradas duas extensões do modelo de Cox (2.1), desenvolvidas com o objetivo de adequar este modelo a determinadas situações que podem surgir na prática, tais como: a violação da hipótese de riscos proporcionais e a eventual existência de uma ou mais variáveis explanatórias dependentes do tempo, ou seja, que se alteram ao longo do período de observação.

2.6.1 Estratificação

Suponha-se que para uma dada variável qualitativa, as funções de risco das suas diferentes categorias são nitidamente não proporcionais. De modo a contornar este problema, pode-se efetuar uma estratificação dos indivíduos, dividindo a amostra em subgrupos disjuntos com base nas categorias ou níveis dessa covariável, os quais se designam por estratos (*strata*). Esta estratificação é possível mesmo no caso em que a covariável é contínua embora a categorização de uma variável deste tipo tenha as suas dificuldades. Além disso, é preciso ter em consideração que o efeito da covariável utilizada para estratificar os indivíduos, designada por variável de estratificação, deve ser diminuto em relação às restantes covariáveis, uma vez que a estratificação impede a possibilidade de estimar o efeito dessa covariável na sobrevivência dos indivíduos [54].

Da necessidade de recorrer a uma extensão do modelo de Cox (2.1) que permita acomodar este tipo de situação, surge então o modelo de Cox estratificado. Para tal, considere-se que para uma variável de estratificação com S categorias, o modelo de Cox correspondente ao estrato s, $s = 1, \ldots, S$, é dado por

$$h(t; \boldsymbol{z}_{is}) = h_{0s}(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{is}), \qquad i = 1, \ldots, n_s < n,$$

onde n_s designa o número de indivíduos no estrato s, $\boldsymbol{z}_{is}$ é o vetor de covariáveis associado ao i-ésimo indivíduo pertencente ao estrato s e $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos. Este modelo assume que, para indivíduos de um mesmo estrato, as funções de risco são proporcionais. De facto, analisando a razão dos riscos de dois indivíduos, a e b, pertencentes ao s-ésimo estrato, com vetores de covariáveis $\boldsymbol{z}_{as}$ e $\boldsymbol{z}_{bs}$, respetivamente, confirma-se que existe proporcionalidade dos riscos:

$$\frac{h(t; \boldsymbol{z}_{as})}{h(t; \boldsymbol{z}_{bs})} = \frac{h_{0s}(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{as})}{h_{0s}(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{bs})} = \exp\left[\boldsymbol{\beta}'(\boldsymbol{z}_{as} - \boldsymbol{z}_{bs})\right].$$

Por outro lado, quando os indivíduos pertencem a estratos diferentes pode acontecer que as funções de risco sejam não proporcionais, dado que as funções de risco subjacentes de cada estrato, $h_{01}(t), h_{02}(t) \ldots, h_{0S}(t)$, são arbitrárias e não estão relacionadas entre si.

Nesta situação, para estimar os parâmetros de regressão é necessário adaptar a função de verosimilhança parcial (2.3), incorporando a informação proveniente dos vários estratos. Desta forma, os parâmetros $\boldsymbol{\beta}$ são estimados através da maximização da seguinte função de verosimilhança parcial

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{S} L_s(\boldsymbol{\beta}) = \prod_{s=1}^{S} \prod_{k=1}^{r_s} \frac{\exp\left(\boldsymbol{\beta}' \boldsymbol{z}_{ks}\right)}{\sum\limits_{l \in R_{ks}} \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ls})},$$

onde $L_s(\boldsymbol{\beta})$ é a verosimilhança parcial correspondente ao s-ésimo estrato, $r_1, r_2, \ldots, r_S$ indicam o número de mortes observadas em cada um dos estratos e R_{ks} representa o conjunto de indivíduos em risco no k-ésimo instante de morte referente ao estrato s. Analogamente ao exposto na secção 2.3.1, também aqui é possível obter a função de verosimilhança parcial para tempos de vida empatados, sendo dada por

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{S} \prod_{k=1}^{r_s} \frac{\exp\left(\boldsymbol{\beta}' \widetilde{\boldsymbol{z}}_{ks}\right)}{\left[\sum\limits_{l \in R_{ks}} \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ls})\right]^{d_{ks}}},$$

em que $\widetilde{\boldsymbol{z}}_{ks}$ e d_{ks} equivalem a $\widetilde{\boldsymbol{z}}_k$ e d_k na expressão (2.6), respetivamente, só que agora no estrato s.

Os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ obtidos desta forma gozam das mesmas propriedades assintóticas mencionadas anteriormente. Então, a inferência com base nos parâmetros de regressão pode ser feita como até agora. Deste modo, o modelo de Cox estratificado também pode ser utilizado para testar a validade da hipótese de riscos proporcionais. É de salientar que as estimativas obtidas através destes estimadores são iguais para todos os estratos. O coeficiente é o mesmo mas o valor da covariável não tem que ser (nem é) igual para todos os indivíduos, por isso, o efeito das covariáveis não é o mesmo para todos os indivíduos.

2.6.2 Variáveis explanatórias dependentes do tempo

Conforme referido na secção 1.5, uma variável explanatória dependente do tempo é uma variável cujo valor sofre alterações ao longo do período de observação e, por essa razão, para o i-ésimo indivíduo em estudo pode ser definida por $z_i(t)$, $t \geq 0$. De facto, existem situações em que não faz sentido considerar que o valor de uma covariável é constante ao longo do estudo, como é o caso da maioria das variáveis laboratoriais. Assim, na monitorização dos indivíduos deve-se registar regularmente, e sempre que for possível, as alterações que ocorrem nos valores dessas covariáveis. Um modelo de regressão que tenha em consideração essas alterações é, obviamente, um modelo mais satisfatório do que aquele que apenas inclua os valores observados no início do estudo.

Para este caso, torna-se necessário recorrer a uma extensão do modelo de Cox (2.1) que permita incluir este tipo de covariáveis. Desta forma, o modelo passa a depender dos valores observados dessas covariáveis em cada instante de tempo, pelo que pode ser definido do seguinte modo

$$h\big(t; \boldsymbol{z}_i(t)\big) = h_0(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_i(t)\big), \qquad i = 1, \ldots, n,$$

onde $\boldsymbol{z}_i(t)$ é o vetor de covariáveis associado ao i-ésimo indivíduo, podendo conter uma ou mais variáveis dependentes do tempo, enquanto as restantes se encontram fixas. Para que esta extensão seja utilizada corretamente é importante compreender qual o tipo de dependência que cada covariável apresenta. Como visto anteriormente, existem dois tipos de covariáveis dependentes do tempo: as externas e as internas. Ambos os tipos têm influência nos processos de morte e, consequentemente, afetam o prognóstico dos indivíduos em estudo [54]. Note-se que esta extensão já não pode ser encarada como um modelo de riscos proporcionais, pois a razão dos riscos de dois indivíduos a que correspondem os vetores de covariáveis $\boldsymbol{z}_a(t)$ e $\boldsymbol{z}_b(t)$ é dada por

$$\frac{h\big(t; \boldsymbol{z}_a(t)\big)}{h\big(t; \boldsymbol{z}_b(t)\big)} = \frac{h_0(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_a(t)\big)}{h_0(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_b(t)\big)} = \exp\Big[\boldsymbol{\beta}'\big(\boldsymbol{z}_a(t) - \boldsymbol{z}_b(t)\big)\Big],$$

donde se observa que esta se encontra dependente do instante t.

A realização de inferência também pode ser feita através desta extensão, mas para isso é necessário modificar a função de verosimilhança parcial (2.3), passando a representá-la do seguinte modo

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{r} \frac{\exp\Big(\boldsymbol{\beta}' \boldsymbol{z}_k\big(t_{(k)}\big)\Big)}{\sum\limits_{l \in R_k} \exp\Big(\boldsymbol{\beta}' \boldsymbol{z}_l\big(t_{(k)}\big)\Big)},$$

onde $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ são os tempos de vida observados distintos e R_k é o conjunto de indivíduos em risco em $t_{(k)}$. Repare-se que assim é necessário dispor do valor de cada variável explanatória inserida no modelo no instante $t_{(k)}$, para todos os indivíduos em risco nesse instante. Em tal caso, pode ser preciso utilizar valores aproximados como, por exemplo, o último valor registado de cada covariável dependente do tempo antes do instante em causa.

Na presença de empates, pode-se adaptar $L(\boldsymbol{\beta})$ pelo mesmo processo efetuado anteriormente. Além disso, esta extensão para variáveis explanatórias

dependentes do tempo pode ainda ser utilizada para testar a hipótese de riscos proporcionais, em que se estuda o impacto de uma covariável estar fixa enquanto as restantes covariáveis se encontram dependentes do tempo (para aprofundar este assunto veja-se Collett [24] e Rocha e Papoila [71]).

Capítulo 3

Modelos para acontecimentos múltiplos

3.1 Introdução

O modelo de Cox clássico (2.1) é apropriado para analisar o tempo decorrido desde um instante inicial, bem definido, até à observação de um único acontecimento de interesse. Além disso, os tempos de vida dos indivíduos em estudo não apresentam qualquer tipo de dependência entre si. Assim, para cada indivíduo regista-se um único tempo, sendo esse tempo determinado pela ocorrência do acontecimento que se pretende estudar ou pela censura. A partir desse instante o indivíduo é excluído do conjunto de indivíduos em risco, abandonando por completo o estudo. Na Figura 3.1 observa-se uma possível ilustração desta situação (adaptado de Carvalho *et al.* [22] e Therneau e Grambsch [79]). Alguns exemplos de acontecimentos que ocorrem uma única vez são: a morte, o diagnóstico de algumas doenças (como a doença de Alzheimer e a doença de Parkinson), a entrada no ensino superior, entre outros.

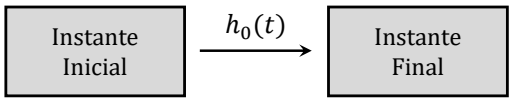


Figura 3.1: Representação esquemática do modelo de Cox clássico.

Contudo, nos últimos anos tem existido um interesse crescente em estudar o tempo até à observação de acontecimentos múltiplos para um mesmo indivíduo [23, 25, 38, 42]. De facto, o desenvolvimento da investigação na área da saúde assim como o aumento da qualidade de vida e das condições

hospitalares, tem vindo a incrementar consideravelmente a esperança média de vida dos indivíduos e a possibilitar que algumas situações clínicas ocorram mais do que uma vez. Deste modo, o interesse tem-se direcionado para a análise do tempo até à ocorrência de acontecimentos múltiplos por indivíduo. Por exemplo, em estudos oncológicos existe uma grande preocupação em estudar o tempo até à reincidência da doença; já em estudos farmacológicos, quando os investigadores testam um novo medicamento, muitas vezes procuram estudar o tempo até a ocorrência de diversos efeitos secundários.

Os dados relativos a acontecimentos múltiplos têm como principal característica o facto de se registar mais do que um tempo de vida para cada indivíduo, inviabilizando a aplicação direta do modelo de Cox. Neste tipo de dados, é legítimo considerar a possibilidade de os tempos observados estarem correlacionados entre si. A correlação entre os tempos de vida pode ocorrer em dois contextos diferentes: i) quando se observam acontecimentos múltiplos para alguns indivíduos; e ii) quando existe algum tipo de correlação entre indivíduos pertencentes a um mesmo grupo[1], embora só se observe um acontecimento para cada indivíduo. Repare-se que esta última situação pode ser interpretada como a observação de acontecimentos múltiplos dentro de cada grupo de indivíduos.

Uma outra característica própria destes dados diz respeito à constituição do conjunto de indivíduos em risco. Quando para um dado indivíduo se observam vários acontecimentos, significa que, após a observação de cada um deles, o indivíduo manteve-se em risco de sofrer um novo acontecimento (ou seja, manteve-se no conjunto de indivíduos em risco), contrariamente ao que sucede no modelo de Cox clássico e nas suas duas extensões referidas no capítulo anterior [9, 22].

Os acontecimentos múltiplos podem ser divididos em duas grandes categorias: acontecimentos da mesma natureza (designados por acontecimentos recorrentes) ou acontecimentos de natureza diferente [45, 85]. Os acontecimentos da mesma natureza ocorrem quando existe uma repetição ou recorrência do mesmo tipo de acontecimento (por exemplo, como acontece no estudo de múltiplos AVC's[2] por indivíduo), enquanto os acontecimentos de natureza diferente ocorrem quando se observam acontecimentos de tipos completamente diferentes (por exemplo, após um indivíduo ser diagnosticado com SIDA[3] pode-se estudar o tempo até à ocorrência de diferentes doenças oportunistas). É de salientar que quando se observam acontecimentos re-

[1] Os indivíduos estão agrupados por terem alguma característica em comum como, por exemplo, serem membros da mesma família (mãe, pai, irmãos, entre outros) ou porque assim foi determinado pelo *design* do estudo.

[2] Acidentes Vasculares Cerebrais.

[3] Síndrome de Imunodeficiência Adquirida.

correntes, estes não podem ocorrer em simultâneo, isto é, para um mesmo indivíduo não existem observações empatadas.

Um outro aspeto importante a identificar na análise de acontecimentos múltiplos é se a ordem pela qual os acontecimentos ocorrem deve ser ou não levada em consideração. De facto, como existem situações em que os acontecimentos ocorrem de forma sequencial, a ocorrência de um dado acontecimento pode aumentar ou diminuir o risco de ocorrência dos restantes acontecimentos, ou até mesmo impedir a sua observação (sendo este designado por acontecimento terminal) [35, 88]. Como exemplo, considere-se o caso de um indivíduo que sofreu um AVC, onde é plausível admitir que durante um certo período de tempo esse indivíduo encontra-se mais suscetível a sofrer um novo AVC; por outro lado, se o AVC for fatal, impedirá que esse indivíduo sofra um novo AVC.

Além do mais, os acontecimentos podem ser instantâneos ou duradouros, no sentido em que um indivíduo pode voltar a estar em risco imediatamente após a observação do acontecimento de interesse ou haver um certo período de tempo durante o qual esse indivíduo deixa de estar em risco, respetivamente. De forma a ilustrar esta última situação, imagine-se que num determinado desporto um atleta sofre uma lesão que o impede de voltar a treinar durante um certo período de tempo. Neste caso, enquanto o atleta está em período de recuperação, não se considera que esteja em risco de sofrer uma nova lesão.

No decorrer deste capítulo, serão discutidos alguns dos modelos mais utilizados para a análise de acontecimentos múltiplos em dados longitudinais. Em primeiro lugar, serão apresentadas algumas das abordagens possíveis para o tratamento deste tipo de dados, sendo a parte restante do capítulo puramente dedicada às extensões do modelo de Cox. Neste contexto, serão apresentados alguns conceitos fulcrais para uniformizar a caracterização dessas extensões e, consequentemente, facilitar a compreensão sobre a forma como estas diferem entre si. Posto isto, estão reunidas as condições para se proceder à formalização dos modelos. Para terminar, serão feitas algumas considerações acerca da potencial existência de correlação entre os tempos observados de um mesmo indivíduo e da sua influência, tanto na estimação dos parâmetros de regressão, como na aplicação dos testes de hipóteses.

3.2 Abordagens possíveis

Na literatura têm surgido várias abordagens para modelar o tempo até à ocorrência de acontecimentos múltiplos por indivíduo [26, 42, 79]. Inicialmente, analisou-se apenas o tempo até a ocorrência do primeiro acontecimento, ignorando a multiplicidade de acontecimentos observados. Mas isto implica que a informação relativa ao tempo até aos restantes acontecimentos

seja desperdiçada/desvalorizada, originando conclusões pouco abrangentes. Em alternativa, considerou-se um modelo para cada acontecimento. Porém, também esta abordagem apresenta as suas desvantagens, sendo a mais preocupante o facto de não ter em conta a potencial correlação entre os tempos de vida.

Uma forma mais adequada de analisar este tipo de dados, consiste em considerar um modelo de regressão onde a variável resposta segue uma distribuição de Poisson e é definida como sendo o número de acontecimentos observados por indivíduo ao longo do período de estudo. Assim, pressupõe-se que os acontecimentos ocorrem de forma contínua ao longo do tempo e são independentes entre si. A aplicação desta abordagem apresenta algumas restrições, nomeadamente por não ter em consideração a natureza dos acontecimentos, nem permitir que dois indivíduos que tenham sofrido o mesmo número de acontecimentos, mas com tempos de observação diferentes, sejam distinguidos [21] (por exemplo, um indivíduo que tenha sofrido 7 acontecimentos num período de 90 dias não é distinguido de outro que tenha sofrido igual número de acontecimentos mas apenas num período de 15 dias).

Outra abordagem surgiu associada à ideia de que, em qualquer estudo, está sempre presente uma certa heterogeneidade entre os indivíduos que não é observável. Desta forma, recorre-se a modelos de efeitos aleatórios, usualmente conhecidos neste ramo da estatística por modelos com fragilidade (*frailty models*). Estes modelos incorporam uma variável aleatória, designada por fragilidade, que tem como objetivo descrever a variabilidade que ficou por explicar entre os tempos de vida de um mesmo indivíduo ou entre os tempos de vida de um grupo de indivíduos [42]. Assim, a fragilidade irá contemplar a influência dos fatores de risco desconhecidos, ou não mensuráveis, sobre os tempos de vida observados. Neste tipo de modelos, assume-se que os vários acontecimentos de um mesmo indivíduo, condicionados à sua fragilidade, são independentes. Por esse motivo, são também designados de modelos condicionais. Nos últimos anos, esta abordagem tem sido alvo de uma investigação ativa, embora estes modelos só tenham sido aplicados a acontecimentos recorrentes.

Por fim, existe ainda uma abordagem diferente de todas aquelas que acabaram de ser referidas, que consiste em adaptar o modelo de Cox clássico às características dos acontecimentos múltiplos. Por conseguinte, têm sido sugeridas diversas extensões do modelo de Cox que procuram dar resposta às mais variadas particularidades que cada caso de estudo apresenta. Em todos estes modelos de regressão, a estimação dos parâmetros é feita através do ajustamento de um modelo que ignora a correlação entre os tempos de vida de um mesmo indivíduo, designada por correlação intraindivíduos (*within-subject/intra-subject correlation*), razão pela qual estes modelos são

habitualmente designados por modelos marginais [51, 79]. De forma a compensar esse facto, é necessário proceder a uma correção da estimativa da variância usual. Para o efeito, utiliza-se um estimador robusto da matriz de covariância e, com base neste estimador, é possível detetar a presença de tempos de vida correlacionados.

Os modelos marginais podem ser agrupados em duas classes, consoante os acontecimentos obedeçam ou não a uma determinada estrutura de ordenação. Alguns dos modelos mais usuais encontram-se mencionados na Figura 3.2. Quando os acontecimentos são ordenados, os tempos de vida de cada indivíduo seguem obrigatoriamente uma ordem, que pode ser dada tanto pela definição temporal, em que se registam as datas de entrada e de saída dos indivíduos no estudo, como por se assumir uma determinada ordem ao estratificar os indivíduos por acontecimento [22].

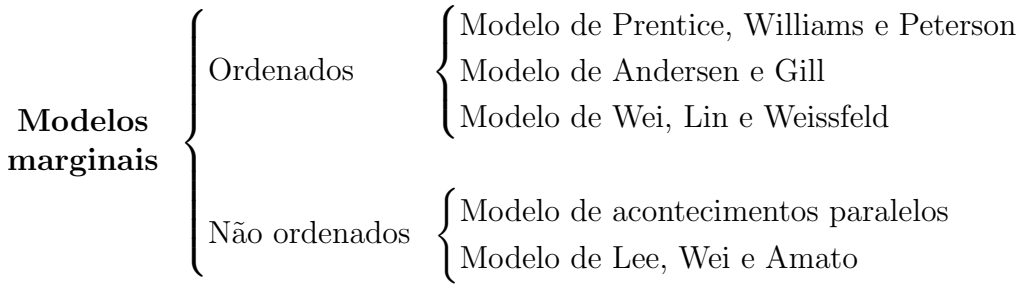


Figura 3.2: Modelos marginais classificados consoante os acontecimentos seguem ou não uma determinada ordem.

O modelo de acontecimentos paralelos aplica-se apenas a acontecimentos de natureza diferente, onde se considera que a partir do instante em que um indivíduo entra no estudo, este encontra-se simultaneamente/paralelamente em risco de sofrer qualquer um dos acontecimentos. Este modelo pressupõe que a ocorrência de um acontecimento não exclui que os restantes sejam observados, e os acontecimentos não decorrem segundo uma determinada ordem ou, pelo menos, não é importante considerar uma possível ordenação. Considerando novamente o exemplo de um indivíduo diagnosticado com SIDA, pode-se estudar o tempo até a ocorrência de várias doenças oportunistas, como a diarreia, o herpes, a tuberculose, entre outras, e nenhuma destas doenças elimina o risco de outra ocorrer. Dos cinco modelos apresentados na Figura 3.2, o modelo de acontecimentos paralelos pode ser considerado o mais simples quanto à sua construção, uma vez que a formulação deste modelo marginal consiste em considerar um modelo de Cox clássico para cada tipo de acontecimento. É importante chamar à atenção para a interpretação dos resultados pois, apesar de se estar a modelar cada acontecimento em

separado, a ocorrência de mais do que um acontecimento para um mesmo indivíduo terá implicações relevantes para a sua análise. Este modelo não será mais detalhado neste trabalho, uma vez que não é adequado para acontecimentos recorrentes (para mais detalhes consulte-se Carvalho *et al.* [22]).

As secções seguintes dedicar-se-ão exclusivamente ao estudo dos restantes quatro modelos, dado que a formulação de cada um deles requer um olhar mais cuidadoso sobre as diferentes especificidades que apresentam. Começa-se por introduzir alguns conceitos essenciais utilizados na caracterização dos mesmos, os quais serão úteis para compreender a forma como diferem entre si e, consequentemente, auxiliar na escolha do modelo mais adequado para cada situação.

3.3 Caracterização dos modelos marginais

A aplicação de modelos marginais para a análise de acontecimentos múltiplos é uma abordagem relativamente recente e tem sido alvo de vários desenvolvimentos ao longo dos últimos anos. Apesar de já existir um número considerável de modelos, não se pode apontar para um em especial e dizer que esse é o melhor modelo pois, para além de os dados apresentarem diversas características que influenciam essa decisão, o próprio planeamento do estudo pode permitir várias interpretações acerca do mecanismo que gera os acontecimentos, o que significa que pode haver situações em que é plausível aplicar mais do que um modelo. Desta forma, não existem modelos perfeitos mas sim modelos mais ou menos apropriados a uma determinada situação.

Kelly e Lim [45], conscientes da complexidade do problema, sugeriram uma forma sistemática para diferenciar este tipo de extensões do modelo de Cox, baseada essencialmente em quatro componentes chave: i) intervalo de risco; ii) função de risco subjacente; iii) conjunto de indivíduos em risco; e iv) estrutura de dependência entre acontecimentos.

Nesta secção, essas componentes serão analisadas com algum detalhe. Nesse sentido, interessa ter presente que, embora cada componente diga respeito a uma determinada característica, as componentes estão de algum modo ligadas entre si, isto é, a definição de cada uma delas influenciará as demais.

3.3.1 Intervalo de risco

O intervalo de risco é definido como o período de tempo durante o qual um indivíduo está em risco de sofrer algum acontecimento. A sua formulação pode ser feita de três formas distintas: tempo total (*total time*), tempo por intervalos (*gap time*) ou processo de contagem (*counting process*). De maneira a facilitar a compreensão sobre o modo como estas formulações diferem

entre si, considere-se o exemplo patente na Figura 3.3 (adaptado de Kelly e Lim [45]), referente a três indivíduos que entraram no estudo no mesmo instante de tempo, os quais estavam em risco de lhes serem observados acontecimentos múltiplos (os gráficos foram obtidos por intermédio do *package* Hmisc [41] do *software* estatístico R).

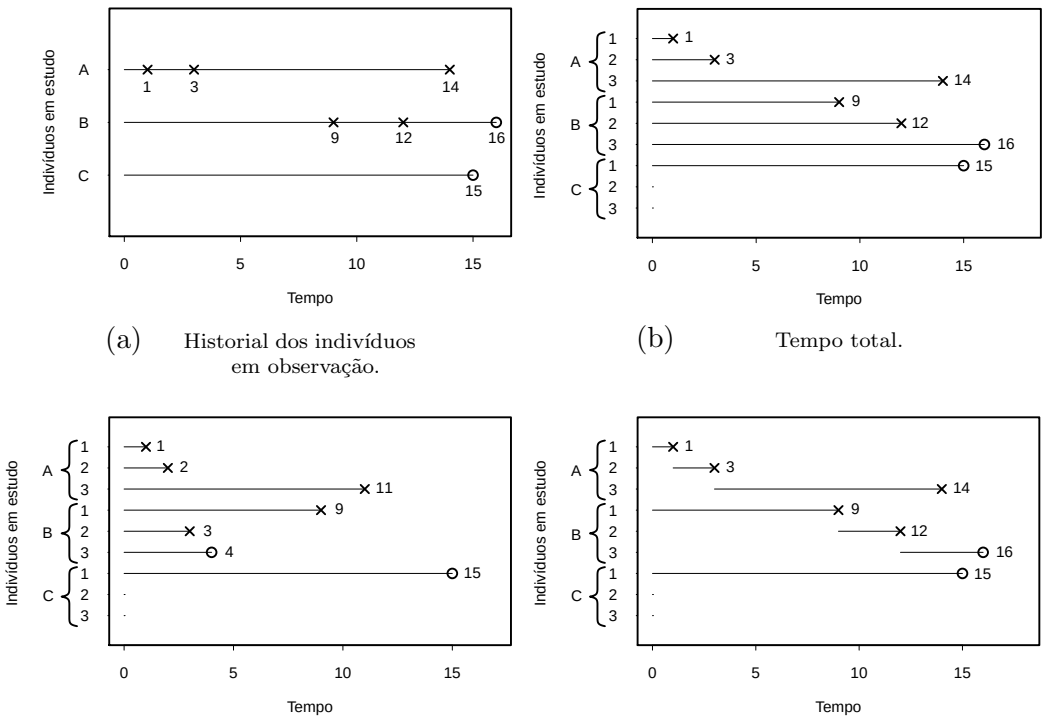


Figura 3.3: Ilustração das possíveis formulações do intervalo de risco, baseadas no exemplo hipotético da observação de três indivíduos, onde × e O representam a ocorrência do acontecimento de interesse e a ocorrência de censura, respetivamente.

Na formulação por tempo total, todos os tempos de vida de um determinado indivíduo (relativos à observação de cada acontecimento) iniciam-se num mesmo instante de tempo, sendo usual considerar como instante inicial a data de entrada do indivíduo no estudo (ver Figura 3.3(b)). No que diz respeito ao tempo por intervalos, o relógio reinicia a contagem do tempo após a ocorrência de cada acontecimento, fazendo com que o tempo de observação de cada indivíduo seja dividido em intervalos de tempo, ou seja, os tempos de vida de um mesmo indivíduo iniciam-se imediatamente após a observação do último acontecimento (ver Figura 3.3(c)). No caso da formulação por processo de contagem, o tempo é definido pela mesma escala utilizada no tempo

total, porém esta formulação reconhece a possibilidade do acontecimento ter uma determinada duração, o que significa que o indivíduo não está em risco para o próximo acontecimento até que o que está a decorrer tenha efetivamente terminado (ver Figura 3.3(d)). Esta particularidade representa uma vantagem sobre as outras duas formulações.

Para que as diferenças entre as formulações fiquem ainda mais percetíveis, analise-se quais os intervalos de tempo em que o indivíduo A se encontra em risco de sofrer um acontecimento, consoante o tipo de formulação adotada (consultar Tabela 3.1). Como é possível observar, qualquer que seja a formulação do intervalo de risco, o indivíduo A encontra-se em risco de sofrer o primeiro acontecimento durante o intervalo de tempo $(0, 1]$. Além disso, observe-se que nas duas últimas formulações (tempo por intervalos e processo de contagem) os intervalos de tempo apresentam a mesma amplitude, apesar das escalas de tempo serem diferentes. Isto acontece porque, neste exemplo, os acontecimentos ocorrem de forma instantânea.

Tabela 3.1: Intervalos de tempo em que o indivíduo A está em risco para cada acontecimento, consoante a formulação adotada.

Intervalo de risco	**Indivíduo A**		
	Acont. 1	Acont. 2	Acont. 3
Tempo total	$(0, 1]$	$(0, 3]$	$(0, 14]$
Tempo por intervalos	$(0, 1]$	$(0, 2]$	$(0, 11]$
Processo de contagem	$(0, 1]$	$(1, 3]$	$(3, 14]$

A decisão sobre qual das formulações do intervalo de risco será adotada é geralmente tomada aquando da construção da base de dados. Contudo, é necessário ter em consideração as implicações dessa escolha sobre as restantes componentes, principalmente para o conjunto de indivíduos em risco, uma vez que a identificação dos indivíduos que contribuem para o conjunto de risco processa-se de forma diferente em cada uma destas formulações, como será analisado mais à frente.

3.3.2 Função de risco subjacente

De acordo com a definição introduzida no modelo de Cox clássico (secção 2.2), a função de risco subjacente é uma função arbitrária não negativa, dependente do tempo e comum a todos os indivíduos em estudo. No contexto da análise de acontecimentos múltiplos por indivíduo, essa definição

mantém-se, mas agora a função de risco subjacente pode ser formulada de duas formas diferentes: pode ser comum a todos os acontecimentos ou pode ser específica para cada acontecimento.

Um modelo que incorpore uma função de risco subjacente comum a todos os acontecimentos, é um modelo onde se admite que, para um mesmo indivíduo, os acontecimentos têm igual risco de ocorrerem, seja qual for a ordem de observação.

Por outro lado, se a ordem pela qual os acontecimentos ocorrem influencia o risco de ocorrência dos acontecimentos seguintes, então faz mais sentido considerar uma função de risco subjacente específica para cada acontecimento. Nesta situação, torna-se necessário estratificar os indivíduos por acontecimento para acomodar essa diferença. A estratificação consiste essencialmente em ajustar um modelo para cada acontecimento. Deste modo, se tiverem ocorrido s acontecimentos de interesse, então existirão s funções de risco subjacentes diferentes, uma para cada estrato.

3.3.3 Conjunto de indivíduos em risco

Recorde-se que o conjunto de risco é definido como o conjunto de indivíduos que estão em risco imediatamente antes da observação do acontecimento de interesse. A ocorrência de acontecimentos múltiplos faz com que existam tantos conjuntos de risco quanto o número de acontecimentos que podem vir a ser observados. Deste modo, o s-ésimo conjunto de risco contém os indivíduos que estão em risco para o acontecimento de ordem s.

Existem três possibilidades para caracterizar o conjunto de indivíduos em risco, nomeadamente: não restritivo, restritivo e semirrestritivo. Esta caracterização incorpora a escolha da função de risco subjacente: um conjunto de risco não restritivo implica uma função de risco subjacente comum a todos os acontecimentos, enquanto um conjunto de risco restritivo ou semirrestritivo implica que a função de risco subjacente seja especificada para cada acontecimento. Esta ideia ficará mais clara no decorrer desta secção.

Por intermédio da combinação entre o conjunto de indivíduos em risco e a formulação do intervalo de risco (tempo total, tempo por intervalos e processo de contagem), é possível identificar quais os indivíduos em risco num certo instante, em particular, nos instantes em que ocorre algum acontecimento.

Quando o conjunto de risco é não restritivo, considera-se que todos os intervalos de risco, de todos os indivíduos em estudo, podem estar incluídos no conjunto de risco de qualquer acontecimento, seja qual for o número de acontecimentos observados para cada indivíduo. Com base no exemplo da Figura

3.3, analise-se quais os intervalos de risco que contribuem para o segundo acontecimento do indivíduo B. Na formulação por tempo total (ver Figura 3.3(b)), o segundo acontecimento do indivíduo B, que ocorre no instante 12, inclui informação dos seguintes intervalos de risco: terceiro acontecimento do indivíduo A, segundo e terceiro acontecimentos do indivíduo B e primeiro acontecimento do indivíduo C. Quando se considera o tempo por intervalos (ver Figura 3.3(c)), o segundo acontecimento do indivíduo B, que ocorre no instante 3, inclui a contribuição dos mesmos acontecimentos mencionados na formulação anterior, mas inclui ainda a informação do intervalo de risco correspondente ao primeiro acontecimento do indivíduo B. Se a formulação do intervalo de risco for feita por processo de contagem (ver Figura 3.3(d)), o segundo acontecimento do indivíduo B, que ocorre no instante 12, inclui a contribuição dos mesmos acontecimentos mencionados na formulação tempo total, à exceção do terceiro acontecimento do indivíduo B.

No caso de o conjunto de risco ser restritivo, considera-se que as contribuições para o conjunto de risco de ordem s estão restritas à inclusão de informação proveniente dos s-ésimos intervalos de risco, correspondentes aos indivíduos que sofreram exatamente $s-1$ acontecimentos. Por outras palavras, apenas os indivíduos que tenham sofrido $s-1$ acontecimentos estão incluídos no conjunto de risco relativo ao acontecimento de ordem s. Recorrendo ao exemplo da Figura 3.3, desta vez analise-se quais os intervalos de risco que contribuem para o segundo acontecimento do indivíduo A. Em ambas as formulações, tempo total e tempo por intervalos (ver Figuras 3.3(b) e 3.3(c)), o segundo acontecimento do indivíduo A, que ocorre nos instantes 3 e 2, respetivamente, inclui informação do segundo acontecimento do indivíduo A e do segundo acontecimento do indivíduo B. É de salientar que não importa o instante em que ocorre o primeiro acontecimento para um dado indivíduo (como acontece neste exemplo com o indivíduo B), pois o que determina a sua entrada no conjunto de risco do segundo acontecimento de outro indivíduo é apenas o facto de ter experienciado o primeiro acontecimento. Já na formulação por processo de contagem (ver Figura 3.3(d)), o segundo acontecimento do indivíduo A, que ocorre no instante 3, inclui apenas a contribuição do intervalo de risco do segundo acontecimento do indivíduo A, ou seja, inclui apenas o seu próprio tempo de vida correspondente a esse acontecimento.

Por último, num conjunto de risco semirrestritivo, considera-se que qualquer indivíduo que tenha sofrido $s-1$ acontecimentos ou menos (ou seja, que tenha sofrido no máximo $s-1$ acontecimentos) está incluído no conjunto de risco relativo ao s-ésimo acontecimento. Porém, a informação referente ao s-ésimo intervalo de risco não pode ser incorporada no conjunto de risco relativo ao acontecimento de ordem $s-1$ ou inferior, contrariamente ao que acontece

num conjunto de risco não restritivo. Novamente pelo exemplo da Figura 3.3, os intervalos de risco que contribuem para o segundo acontecimento do indivíduo A, na formulação por tempo total (ver Figura 3.3(b)), dizem respeito ao segundo acontecimento do indivíduo B e ao primeiro acontecimento do indivíduo C. Note-se que, como até ao final do estudo o indivíduo C mantém-se em risco apenas para o primeiro acontecimento, a função de risco que lhe está associada é sempre a mesma, permitindo que contribua para o conjunto de risco dos restantes indivíduos, seja qual for o acontecimento considerado. Relativamente ao indivíduo B, o seu primeiro intervalo de risco não pode contribuir para o segundo acontecimento do indivíduo A, visto que ao conjunto de risco semirrestritivo está necessariamente associado uma função de risco subjacente específica para cada acontecimento.

O conjunto de risco semirrestritivo só se aplica às formulações tempo total e processo de contagem, uma vez que encontra-se dependente do tempo entre os acontecimentos. Repare-se que tal não é conseguido através da escala utilizada na formulação tempo por intervalos (ver Figura 3.3(c)). Para além disso, apesar de ser possível associar um conjunto de risco semirrestritivo à formulação por processo de contagem, essa combinação não faz muito sentido. De facto, a formulação por processo de contagem é usada para prevenir que um indivíduo esteja em risco para um dado acontecimento sem que o anterior tenha efetivamente terminado, premissa essa que é violada quando se considera um conjunto de risco semirrestritivo.

3.3.4 Estrutura de dependência entre acontecimentos

Finalmente, no que se refere à estrutura de dependência entre acontecimentos, esta encontra-se ligada à forma como o modelo lida com a potencial correlação existente entre os tempos de vida de um mesmo indivíduo (correlação intraindivíduos). Nesse sentido, foram desenvolvidas três abordagens que contabilizam essa correlação de maneira diferente, designadamente: marginal, condicional e efeitos aleatórios.

Na abordagem marginal, considera-se que desde o instante inicial os indivíduos encontram-se simultaneamente em risco para qualquer um dos acontecimentos e a ocorrência de cada um deles não está dependente de que algum dos outros tenha que ser previamente observado, ou seja, assume-se que os acontecimentos de um mesmo indivíduo são independentes entre si. Para esta abordagem desenvolveu-se um estimador robusto da matriz de covariância que, quando comparado com o estimador usual, permite analisar a existência ou não de correlação intraindivíduos.

No que respeita à abordagem condicional, os indivíduos não são considerados em risco para um dado acontecimento sem que o anterior tenha ocorrido, verificando-se assim uma certa dependência entre os acontecimentos. Inicialmente, a única forma de captar a estrutura de dependência entre os acontecimentos era por intermédio de covariáveis dependentes do tempo como, por exemplo, o número de ocorrências observadas a cada instante de tempo. No entanto, após o aparecimento de um estimador robusto da matriz de covariância, muitos autores começaram por também aplicar este estimador aos modelos que apresentavam uma estrutura de dependência condicional.

Por último, existe ainda a abordagem que recorre a efeitos aleatórios, cujos modelos são denominados por modelos com fragilidade. Neste caso, nas extensões do modelo de Cox, considera-se uma variável aleatória com o objetivo de representar a estrutura de dependência entre os tempos de vida de um mesmo indivíduo ou, ainda, entre tempos de vida de indivíduos pertencentes a um mesmo grupo.

Identificadas as diferenças entre as abordagens marginal e condicional, constata-se que estas estão intimamente relacionadas com as formulações do intervalo de risco. Quando se constrói o intervalo de risco por intermédio das formulações tempo por intervalos ou processo de contagem, determina-se que o modelo segue uma abordagem condicional, no sentido em que um indivíduo não pode estar em risco para um determinado acontecimento sem que o anterior tenha terminado, isto é, está condicionado à ocorrência dos acontecimentos anteriores. Por outro lado, quando o intervalo de risco é construído utilizando a formulação tempo total, fica determinado que o modelo segue uma abordagem marginal, uma vez que os indivíduos são considerados em risco para a ocorrência de qualquer acontecimento desde o instante em que entram em observação e nenhum dos acontecimentos está dependente da ocorrência de outro.

3.4 Formulação dos modelos marginais

Nas últimas quatro décadas tem existido um interesse notório em desenvolver novas extensões do modelo de Cox que permitam analisar dados resultantes da observação de acontecimentos múltiplos por indivíduo. Como referido anteriormente, o modelo de Cox clássico não é apropriado para analisar este tipo de dados. Por um lado, devido à falta de independência entre as observações de um mesmo indivíduo e, por outro, porque os indivíduos mantêm-se no conjunto de risco após a observação de cada acontecimento, razões essas que levariam a que as estimativas dos parâmetros fossem pouco consistentes.

Uma vez introduzidos os conceitos essenciais que permitem caracterizar e diferenciar as extensões do modelo de Cox, estão reunidas as condições necessárias para que os quatro modelos referidos no final da secção 3.2 sejam definidos. Estes modelos serão apresentados por ordem cronológica, de forma a que se acompanhe o evoluir dos mesmos e, ainda, o porquê destes terem surgido. Importa chamar à atenção que, do ponto de vista da estimação dos parâmetros, todos estes modelos são classificados como modelos marginais [51]. Porém, no que se refere à estrutura de dependência entre acontecimentos, os modelos de Prentice, Williams e Peterson (PWP) e de Andersen e Gill (AG) seguem uma abordagem condicional, enquanto os modelos de Wei, Lin e Weissfeld (WLW) e de Lee, Wei e Amato (LWA) seguem uma abordagem marginal [45]. Para auxiliar o estudo destes modelos, recomenda-se que a Tabela B.1, que compila as características de cada um deles, acompanhe a leitura desta secção (consultar anexo B).

Antes de se definir cada modelo, é fundamental introduzir alguma notação. Considere-se que, numa amostra aleatória de dimensão n, o tempo de observação do indivíduo i ($i = 1, \ldots, n$) correspondente ao acontecimento s ($s = 1, \ldots, S$) é dado por $T_{is} = \min\left\{X_{is}, C_{is}\right\}$, onde X_{is} e C_{is} representam o seu verdadeiro tempo de vida e o seu tempo de censura, respetivamente. Para um acontecimento s qualquer, define-se o intervalo de tempo (*gap time*) entre dois acontecimentos consecutivos por $G_{is} = T_{is} - T_{i(s-1)}$, onde $T_{i0} = 0$. Denote-se por $\boldsymbol{z}_{is}(t) = \big(z_{is1}(t), \ldots, z_{isp}(t)\big)'$ o vetor de p covariáveis (possivelmente dependentes do tempo) associado ao i-ésimo indivíduo referente ao acontecimento s e $\boldsymbol{z}_i(t) = \big(\boldsymbol{z}'_{i1}(t), \ldots, \boldsymbol{z}'_{iS}(t)\big)$ o vetor global de covariáveis observadas para esse indivíduo, sendo S o número máximo de acontecimentos que podem ser observados no estudo (para um mesmo indivíduo). Seja $I(\cdot)$ uma variável indicatriz, onde $I(A) = 1$ quando A é verdadeira e $I(A) = 0$ caso contrário. Deste modo, a variável de censura que expressa o estado do i-ésimo indivíduo é definida por $\delta_{is} = I\big(X_{is} \leq C_{is}\big)$. Assume-se que os vetores $\boldsymbol{X}_i = \big(X_{i1}, \ldots, X_{iS}\big)'$ e $\boldsymbol{C}_i = \big(C_{i1}, \ldots, C_{iS}\big)'$ são condicionalmente independentes do vetor $\boldsymbol{z}_i(t)$. Se X_{is} ou $\boldsymbol{z}_{is}(t)$ são omissos, define-se que $C_{is} = 0$, o que garante que $T_{is}(t) = 0$ e $\delta_{is} = 0$. Além disso, pressupõe-se que os casos omissos ocorrem completamente ao acaso (*Missing Completely At Random* – MCAR). Define-se ainda $h\big(t; \boldsymbol{z}_{is}(t)\big)$ como sendo a função de risco do indivíduo i relativa ao acontecimento s, $h_0(t)$ a função de risco subjacente arbitrária comum a todos os indivíduos (seja qual for o acontecimento) e $h_{0s}(t)$ a função de risco subjacente específica para o s-ésimo acontecimento. Por fim, seja $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ o vetor global de parâmetros de regressão desconhecidos e $\boldsymbol{\beta}_s = (\beta_{s1}, \ldots, \beta_{sp})'$ o vetor de parâmetros específico do acontecimento s.

3.4.1 Modelo de Prentice, Williams e Peterson

Em 1981, Prentice, Williams e Peterson [69] propuseram uma das primeiras extensões do modelo de Cox para analisar acontecimentos múltiplos. Este modelo, usualmente designado por modelo PWP, é mais direcionado para o estudo de uma população constituída por um grande número de indivíduos, onde para cada um deles é observado um pequeno número de acontecimentos. A metodologia apresentada por estes autores pode ser interpretada como uma generalização das técnicas utilizadas até então, no sentido em que é permitido continuar a modelar a função de risco após a observação do primeiro acontecimento.

No modelo PWP, existem duas opções para construir o intervalo de risco: processo de contagem ou tempo por intervalos. Quando se constrói o intervalo de risco utilizando a formulação por processo de contagem tem-se o modelo PWP *counting process* (PWP-CP), enquanto que se essa construção for feita considerando o tempo por intervalos tem-se o modelo PWP *gap time* (PWP-GT). Deste modo, pode-se dizer que existem dois modelos PWP, que diferem apenas na escala de tempo utilizada.

Seja qual for a formulação considerada para a construção do intervalo de risco, o modelo PWP aplica-se a situações como aquela que se encontra esquematizada na Figura 3.4, ou seja, onde existe uma função de risco subjacente específica para cada acontecimento.

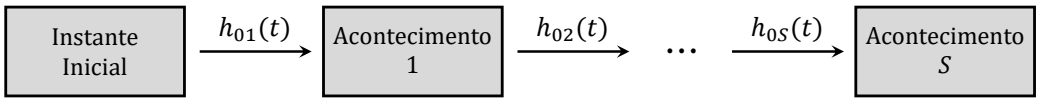


Figura 3.4: Representação esquemática do modelo de Prentice, Williams e Peterson.

A formulação que se segue diz respeito ao caso em que os acontecimentos são recorrentes. Importa salientar que, embora na literatura se encontre várias aplicações deste modelo a acontecimentos recorrentes [39, 40, 50], este também pode ser aplicado a acontecimentos de natureza distinta, mas que recorrem ao longo do tempo.

Seja n o número de indivíduos em estudo e $\boldsymbol{z}_{is}(t) = \big(z_{is1}(t), \ldots, z_{isp}(t)\big)'$ o vetor de p covariáveis associado ao indivíduo i $(i = 1, \ldots, n)$ correspondente ao acontecimento s $(s = 1, \ldots, S)$. Tendo em conta a formulação do modelo de Cox [27], o modelo PWP pode ser definido pelo produto de uma função arbitrária dependente do tempo (específica para cada acontecimento) por uma função exponencial que incorpora o efeito das covariáveis associadas aos indivíduos.

Por conseguinte, para o i-ésimo indivíduo em estudo, as funções de risco dos modelos semiparamétricos PWP-CP e PWP-GT são dadas por

$$h\big(t; \boldsymbol{z}_{is}(t)\big) = h_{0s}(t) \exp\big(\boldsymbol{\beta}_s' \boldsymbol{z}_{is}(t)\big), \qquad t \geq 0, \tag{3.1}$$

e

$$h\big(t; \boldsymbol{z}_{is}(t)\big) = h_{0s}\big(t - t_{i(s-1)}\big) \exp\big(\boldsymbol{\beta}_s' \boldsymbol{z}_{is}(t)\big), \qquad t \geq 0, \tag{3.2}$$

respetivamente, onde $h_{0s}(\cdot) \geq 0$ é a função de risco subjacente comum a todos os indivíduos pertencentes ao estrato s e $\boldsymbol{\beta}_s = (\beta_{s1}, \ldots, \beta_{sp})'$ é o vetor de parâmetros de regressão referente a esse estrato. Note-se que (3.1) e (3.2) correspondem às escalas de tempo mais usuais, embora seja possível a utilização de outras. Assim, $h_{0s}(t)$ corresponde à função de risco subjacente relativa ao tempo observado desde o início do estudo até ao instante t, enquanto que $h_{0s}\big(t - t_{i(s-1)}\big)$ corresponde à função de risco subjacente relativa ao tempo desde a ocorrência do último acontecimento. Para além disso, repare-se que o vetor de covariáveis $\boldsymbol{z}_{is}(t)$ depende do número de acontecimentos ocorridos ao indivíduo i. Deste modo, tal como acontece no modelo de Cox clássico, $\exp\big(\boldsymbol{\beta}_s' \boldsymbol{z}_{is}(t)\big)$ representa o risco relativo associado ao vetor de covariáveis.

Em Prentice *et al.* [69], o modelo PWP-CP encontra-se caracterizado como sendo um modelo que utiliza a formulação tempo total, porém essa designação não é a mais apropriada. De facto, conforme descrito na secção 3.3.1, as formulações tempo total e processo de contagem utilizam a mesma escala de tempo. No entanto, a construção do intervalo de risco segundo a formulação tempo total estabelece que os indivíduos estão em risco para qualquer acontecimento desde o instante inicial, seja qual for o número de acontecimentos observados, o que significa que esta formulação é adequada para modelos em que a estrutura de dependência entre acontecimentos é marginal, enquanto as outras duas são consideradas para modelos condicionais. Por esse motivo, atualmente o modelo PWP-CP deixou de ser designado por modelo PWP com tempo total [45].

De acordo com o exemplo utilizado anteriormente (Figura 3.3), quando os acontecimentos são instantâneos, pode-se dizer que os modelos PWP-CP (3.1) e PWP-GT (3.2) têm intervalos de risco com a mesma amplitude, embora utilizem escalas de tempo diferentes. A escala de tempo utilizada influenciará a interpretação dos riscos relativos. O modelo PWP-CP é adequado para estudar o efeito de uma covariável no tempo a partir do instante em que o indivíduo entra em observação, enquanto para o modelo

PWP-GT esse efeito diz apenas respeito ao tempo desde a ocorrência do último acontecimento.

Em ambos os modelos PWP, assume-se que o risco de ocorrência do acontecimento seguinte é alterado pela ocorrência do acontecimento anterior, o que implica considerar uma função de risco subjacente específica para cada acontecimento (ver Figura 3.4). Por conseguinte, é necessário estratificar os indivíduos segundo a ordem pela qual os acontecimentos ocorrem. Assim, se puderem ser observados s acontecimentos, existirão s estratos ordenados, sendo que a cada um deles estará associada a função de risco subjacente específica $h_{0s}(t)$, $s = 1, \ldots, S$. Como exemplo de aplicação, considere-se o caso em que, para cada indivíduo, podem ocorrer múltiplos enfartes do miocárdio. Nesta situação, é razoável admitir que a observação do enfarte seguinte apresenta um risco alterado/diferente comparativamente ao enfarte imediatamente anterior.

Relativamente ao conjunto de indivíduos em risco, considera-se que estão em risco para o s-ésimo acontecimento, apenas os indivíduos aos quais já foi observado o acontecimento de ordem $s-1$, o que se traduz num conjunto de risco restritivo. Quer isto dizer que um indivíduo transita para o estrato s imediatamente após a observação do $(s-1)$-ésimo acontecimento, permanecendo nesse estrato até a ocorrência do acontecimento de ordem s ou até ser censurado. Deste modo, diz-se que o risco de ocorrência de cada acontecimento está condicionado à ocorrência do acontecimento imediatamente anterior e, por esse motivo, este modelo também é conhecido por modelo de acontecimentos ordenados com risco condicional. Um outro aspeto a salientar neste modelo, é o facto de o conjunto de risco tornar-se cada vez mais pequeno à medida que o número de acontecimentos observados aumenta, o que pode originar estimativas pouco fiáveis. Assim, deve-se averiguar de que forma os indivíduos se distribuem ao longos de todos os estratos e, tendo em conta esse aspeto, escolher o número máximo de acontecimentos que deve ser considerado para a análise [21].

3.4.2 Modelo de Andersen e Gill

Um ano mais tarde, em 1982, Andersen e Gill [8] propuseram uma outra extensão do modelo de Cox, essencialmente na mesma linha de raciocínio do modelo PWP. Este modelo, habitualmente conhecido por modelo AG, é considerado por diversos autores o mais simples dos quatro modelos referidos anteriormente, mas também o que evidencia pressupostos mais fortes [21, 79].

O modelo AG foi proposto para situações em que os acontecimentos são da mesma natureza e ocorrem de forma ordenada. A sua representação es-

quemática encontra-se patente na Figura 3.5, onde se observa que existe apenas uma função de risco subjacente que é comum a todos os acontecimentos.

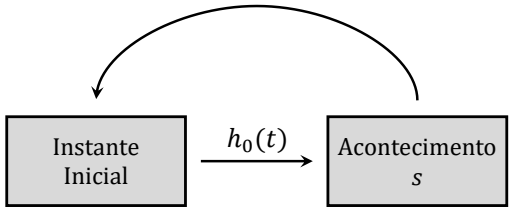


Figura 3.5: Representação esquemática do modelo de Andersen e Gill.

Neste modelo, a construção do intervalo de risco é feita através da formulação por processo de contagem, pelo que cada indivíduo contribui apenas uma vez para o conjunto de risco de um acontecimento. Seja n o número de indivíduos em estudo e $\boldsymbol{z}_{is}(t) = \big(z_{is1}(t), \ldots, z_{isp}(t)\big)'$ o vetor de covariáveis associado ao indivíduo i $(i = 1, \ldots, n)$ correspondente ao acontecimento s $(s = 1, \ldots, S)$. Por conseguinte, a função de risco do modelo semiparamétrico AG é definida por

$$h\big(t; \boldsymbol{z}_{is}(t)\big) = h_0(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{is}(t)\big), \qquad t \geq 0, \tag{3.3}$$

onde $h_0(\cdot) \geq 0$ é a função de risco subjacente comum a todos os indivíduos (seja qual for o acontecimento) e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ é o vetor global de parâmetros desconhecidos. Uma vez que a função de risco subjacente é comum a todos os acontecimentos, pode dizer-se que após a ocorrência de cada acontecimento o indivíduo regressa sempre ao seu instante inicial (Figura 3.5). Desta forma, o modelo AG (3.3) pode ser visto como um modelo PWP-CP (3.1) não estratificado.

Andersen e Gill [8] idealizaram este modelo para o caso em que a ocorrência de cada acontecimento não depende do tempo decorrido desde a última observação, nem do número de acontecimentos observados anteriormente. Quer isto dizer que, apesar da estrutura de dependência entre acontecimentos ser condicional à ocorrência dos acontecimentos anteriores, considera-se que os tempos entre os acontecimentos são independentes. Desta forma, o processo de contagem associado a cada indivíduo contabiliza incrementos independentes e, por esse motivo, o modelo AG também é designado por modelo de incrementos independentes.

Assim sendo, para um mesmo indivíduo, assume-se que os acontecimentos apresentam igual risco de ocorrerem, não existindo necessidade de considerar uma função de risco subjacente específica para cada acontecimento (ver Figura 3.5). Tendo em conta este aspeto, assim como a formulação do intervalo de risco, torna-se evidente que o conjunto de indivíduos em risco para cada acontecimento é não restritivo. Alguns exemplos de aplicação do modelo AG são: estudo do tempo até à ocorrência de outra gravidez, desde que o período de observação não seja demasiado longo para permitir que determinados fatores, como a idade e o número de gestações, alterem a fertilidade da mulher; estudo do tempo até à recorrência da perda de emprego, desde que determinados fatores, como as habilitações literárias e o estatuto social, não afetem o risco de ocorrência dos acontecimentos.

Therneau e Grambsch [79], admitiram que a construção do intervalo de risco também pode ser feita recorrendo à utilização de outra escala de tempo, nomeadamente o tempo por intervalos. Em 2000, também Kelly e Lim [45] referiram a existência desta possibilidade, a qual dá origem a um modelo AG *gap time*. Porém, esta variante não tem a capacidade de reconhecer que um indivíduo possa ter períodos de tempo em que não está em risco de sofrer qualquer acontecimento. Considerando o exemplo anterior, quando uma mulher está grávida, durante o tempo de gestação esta não se encontra em risco de voltar a engravidar, sendo que esse intervalo de tempo pode e deve ser incorporado na análise, o que não é possível quando se utiliza a formulação do tempo por intervalos.

3.4.3 Modelo de Wei, Lin e Weissfeld

Em 1989, Wei, Lin e Weissfeld [86] propuseram uma extensão do modelo de Cox para modelar o tempo desde o instante inicial até à ocorrência de acontecimentos múltiplos, qualquer que seja a sua natureza. O motivo que levou ao desenvolvimento deste modelo, denominado por modelo WLW, foi a falta de robustez demonstrada pelos modelos condicionais (PWP e AG) analisados anteriormente. A aplicação destes modelos deixa de ser adequada quando existe uma má especificação da estrutura de dependência entre os acontecimentos de um mesmo indivíduo [21], ou seja, quando os acontecimentos revelam uma estrutura de dependência não condicional.

No modelo WLW, a construção do intervalo de risco é feita através da formulação tempo total, uma vez que se considera que um indivíduo está simultaneamente em risco para a ocorrência de qualquer um dos acontecimentos desde o instante em que entra em observação. Por conseguinte, este modelo aplica-se a situações como aquela que se encontra ilustrada na Figura 3.6, onde existe uma função de risco subjacente para cada acontecimento.

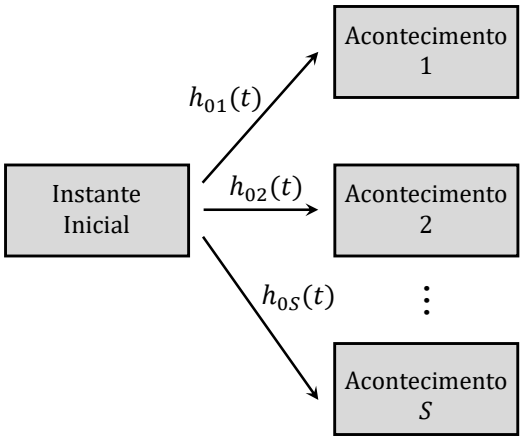


Figura 3.6: Representação esquemática do modelo de Wei, Lin e Weissfeld.

Considere-se uma amostra de dimensão n, em que a cada indivíduo i ($i = 1, \ldots, n$) corresponde o vetor de covariáveis $\boldsymbol{z}_{is}(t) = \big(z_{is1}(t), \ldots, z_{isp}(t)\big)'$, referente ao acontecimento de ordem s ($s = 1, \ldots, S$). Deste modo, pode-se definir a função de risco do modelo semiparamétrico WLW por

$$h\big(t; \boldsymbol{z}_{is}(t)\big) = h_{0s}(t) \exp\big(\boldsymbol{\beta}_s' \boldsymbol{z}_{is}(t)\big), \qquad t \geq 0, \tag{3.4}$$

onde $h_{0s}(\cdot) \geq 0$ é a função de risco subjacente específica do estrato s e $\boldsymbol{\beta}_s = (\beta_{s1}, \ldots, \beta_{sp})'$ é o vetor de parâmetros de regressão desconhecidos referente a esse estrato. Repare-se que, apesar de as funções de risco dos modelos PWP-CP (3.1) e WLW (3.4) serem iguais, estes diferem tanto na construção do intervalo de risco como na definição do conjunto de indivíduos em risco (consultar Tabela B.1).

Os modelos semiparamétricos PWP e AG, basicamente, permitem que as suas funções de risco continuem a modelar o tempo após a observação do primeiro acontecimento, pelo que podem ser vistos como generalizações do modelo de Cox clássico. Contrariamente ao que acontece nestes dois modelos, no modelo WLW não se exige qualquer estrutura de dependência entre os tempos de vida de um mesmo indivíduo, ou seja, o risco de ocorrência de um acontecimento não está condicionado à ocorrência de algum dos outros. Deste modo, Wei *et al.* [86] recorreram ao modelo de Cox clássico para modelar separadamente o tempo até a ocorrência de cada um dos acontecimentos, pelo que pode dizer-se que o risco de ocorrência de cada um deles é modelado de forma marginal e, por essa razão, este modelo é frequentemente designado por modelo marginal.

No modelo WLW, considera-se que qualquer indivíduo que ainda não tenha sofrido o acontecimento s está incluído no conjunto de risco relativo ao s-ésimo acontecimento, o que significa que o conjunto de indivíduos em risco é semirrestritivo. Deste modo, se S for o número máximo de acontecimentos que se pode observar num mesmo indivíduo, em determinadas situações, um indivíduo pode pertencer a S conjuntos de risco em simultâneo, aspeto este que apenas é possível no modelo WLW. Além disso, para cada um dos indivíduos serão registados S tempos de vida (correspondentes aos S acontecimentos), visto que estes se encontram em risco para todos os acontecimentos desde o instante inicial. Assim, um indivíduo deixa de estar em risco a partir do momento em que lhe é observado o acontecimento S ou quando, por algum motivo, deixa de estar em observação. Neste último caso, o seu tempo de vida passa a ser censurado.

Como é possível observar na Figura 3.6, o modelo WLW (3.4) considera uma função de risco subjacente específica para cada acontecimento de acordo com a sua ordenação. De facto, quer os acontecimentos sejam recorrentes ou não, o que importa é a ordem pela qual são observados. Significa então que, apesar da estrutura de dependência entre acontecimentos ser marginal, este modelo assume que os acontecimentos podem ter um risco de ocorrência diferente. Desta forma, tal como acontece no modelo PWP, define-se S estratos referentes a S acontecimentos, em que a cada um deles encontra-se associada a função de risco subjacente diferente. O modelo WLW pode ser aplicado, por exemplo, quando se pretende comparar o efeito de dois tratamentos relativos a uma determinada doença, no sentido de avaliar qual o tratamento que melhor consegue prolongar o tempo até a ocorrência/recorrência de efeitos secundários. Neste caso, considera-se que os acontecimentos têm um risco de ocorrência diferente, qualquer que seja a sua natureza.

O modelo WLW pode ser visto como um modelo de acontecimentos competitivos não ordenados, uma vez que não se sabe qual a ordem em que os acontecimentos irão ocorrer [23, 79]. Assim, após a observação dos vários acontecimentos, estes são inseridos na base de dados de forma ordenada, sem estarem condicionados à ocorrência do acontecimento anterior. Além disso, como se considera que os acontecimentos têm riscos de ocorrência diferentes, estes estão, de certo modo, a competir entre si, embora a ocorrência de um acontecimento não impeça que qualquer outro venha a ser observado.

Neste modelo, a interpretação dos parâmetros de regressão merece algum cuidado. Considerando o exemplo anterior, imagine-se que se estava a analisar o efeito do tratamento no tempo até ao segundo efeito secundário (acontecimento). Esse efeito, não diz apenas respeito ao tempo entre o primeiro e o segundo acontecimentos, pois agora a escala de tempo utilizada refere-se ao tempo desde o início do estudo [75].

3.4.4 Modelo de Lee, Wei e Amato

Em 1992, Lee, Wei e Amato [49] propuseram um modelo para a análise de acontecimentos múltiplos da mesma natureza, no seguimento do modelo WLW. Este modelo, conhecido por modelo LWA, surge numa ótica ligeiramente diferente dos modelos anteriores, tendo sido proposto para o caso em que a amostra é constituída por um grande número de grupos independentes de pequena dimensão, a qual pode diferir de grupo para grupo.

Os modelos apresentados até agora, permitem analisar acontecimentos múltiplos que tenham sido observados para um mesmo indivíduo. Todavia, o modelo LWA foi proposto para o caso em que os acontecimentos múltiplos ocorrem dentro de um mesmo grupo, o que significa que os tempos de vida são, por algum motivo, agrupados. Na prática, este modelo pode ser aplicado a duas situações: i) quando os grupos são compostos por um pequeno número de indivíduos que têm características semelhantes e para cada um deles observa-se um único acontecimento de interesse; ou ii) quando cada indivíduo por si só pode ser visto como um grupo e, neste caso, para um mesmo indivíduo observa-se um pequeno número de acontecimentos.

De forma a clarificar em que situações este modelo pode ser aplicado, considere-se os dois exemplos seguintes. Para a primeira situação, suponha que se pretende modelar o tempo até o diagnóstico de uma determinada doença[4] numa amostra constituída por grupos de indivíduos, onde cada grupo representa uma família. Neste caso, os indivíduos encontram-se agrupados porque têm em comum uma parte do material genético e para cada um deles é observado um único tempo de vida. Quanto à segunda situação, considere que se pretende estudar o efeito de um determinado medicamento no tratamento de uma doença renal, nomeadamente em retardar a perda total das funções renais. O medicamento, ao ser ingerido por um indivíduo, irá afetar todo o seu organismo, sendo que esse facto terá consequências para ambos os rins. Deste modo, é necessário observar em separado o tempo até a perda total das funções de cada rim. Assim, pode dizer-se que os tempos de vida encontram-se agrupados por indivíduo, uma vez que se acompanha a evolução de dois rins.

No modelo LWA, o intervalo de risco também é definido segundo a formulação tempo total, pelo que fica determinado que não existe qualquer imposição relativamente à estrutura de dependência entre acontecimentos (isto é, segue uma abordagem marginal). Desta forma, para ambas as situações enunciadas atrás, a representação esquemática pode ser feita tal como se encontra exposto na Figura 3.7, onde se observa apenas uma função de risco subjacente.

[4]Por exemplo, a doença de Alzheimer que pode surgir ligada a um fator hereditário.

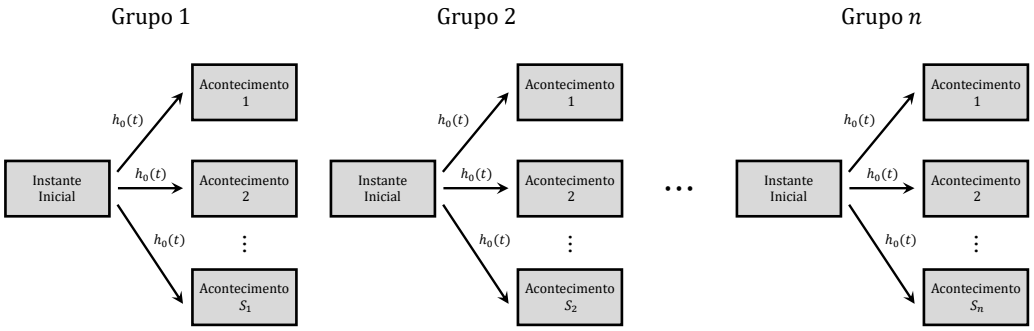


Figura 3.7: Representação esquemática do modelo de Lee, Wei e Amato.

Sem perda de generalidade, a formalização do modelo LWA será feita tendo por base a segunda situação. Considere-se que, numa amostra de dimensão n, cada indivíduo i $(i = 1, \ldots, n)$ representa um grupo. Seja S_i o número de acontecimentos observados em cada grupo i, que se assume ser relativamente pequeno quando comparado com n. Define-se t_{is} como sendo o tempo de observação do i-ésimo indivíduo correspondente ao acontecimento s e $\boldsymbol{z}_{is}(t) = \big(z_{is1}(t), \ldots, z_{isp}(t)\big)'$ o respetivo vetor de covariáveis. Por conseguinte, a função de risco do modelo semiparamétrico LWA pode ser definida por

$$h\big(t; \boldsymbol{z}_{is}(t)\big) = h_0(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{is}(t)\big), \qquad t \geq 0, \tag{3.5}$$

onde $h_0(\cdot) \geq 0$ é a função de risco subjacente comum a todos os acontecimentos e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ é o vetor global de parâmetros desconhecidos. Note-se que, embora as funções de risco dos modelos AG (3.3) e LWA (3.5) sejam iguais, estes diferem na construção do intervalo de risco e, consequentemente, na estrutura de dependência entre acontecimentos (consultar Tabela B.1).

O modelo LWA assume que os acontecimentos apresentam igual risco de ocorrerem, o que implica um conjunto de indivíduos em risco não restritivo. A combinação de um intervalo de risco construído segundo a formulação tempo total com um conjunto de risco não restritivo, permite que um indivíduo esteja em risco para vários acontecimentos em simultâneo. Significa então que, em qualquer instante de tempo, um indivíduo com s intervalos de risco pode ser incluído em s conjuntos de risco sem estar restringido à observação de algum acontecimento. Esta situação apenas é permitida neste modelo.

No geral, o modelo LWA considera um modelo WLW por grupo, mas onde o intuito é analisar apenas acontecimentos da mesma natureza (compare as Figuras 3.6 e 3.7). De acordo com Kelly e Lim [45], é possível

outra interpretação para relacionar estes dois modelos: o modelo WLW é um modelo LWA estratificado por acontecimento, no qual se considera um conjunto de risco semirrestritivo. De facto, ao olhar para um único grupo[5] do modelo LWA, conclui-se que estes dois modelos diferem essencialmente na definição da função de risco subjacente. Importa salientar que, na aplicação do modelo WLW, não é possível considerar uma função de risco subjacente comum a todos os acontecimentos, mesmo no caso em que estes são da mesma natureza, dada a definição do conjunto de risco que lhe está associada.

Um outro exemplo de aplicação deste modelo pode ser encontrado em Lee *et al.* [49], onde se testou o efeito do medicamento sorbinil no tratamento da retinopatia diabética. O objetivo deste estudo, consistiu em analisar o tempo até à perda severa de visão em cada um dos olhos. Antes do surgimento deste modelo, efetuava-se a análise estratificando os tempos observados por indivíduo, utilizando para o efeito a variável de identificação [79]. Porém, os autores do modelo LWA mostraram que é possível obter resultados mais fiáveis quando não é considerada qualquer estratificação, efetuando apenas uma correção na estimativa da variância usual. Assim, destacaram a importância de recorrer ao estimador robusto da matriz de covariância (que será introduzido na secção 3.6) para compensar o facto de os tempos de vida estarem agrupados por indivíduo. Desde então, têm surgido outros modelos semiparamétricos para a análise de acontecimentos múltiplos agrupados. Um modelo interessante, que foi proposto por Schaubel e Cai [73], em 2005, consiste em considerar uma função de risco subjacente específica para cada grupo, ou seja, os acontecimentos observados dentro de um mesmo grupo têm igual risco de ocorrerem, mas quando comparados entre grupos já apresentam um risco diferente.

3.5 Estimação dos parâmetros de regressão

Posteriormente à formulação dos modelos para a análise de acontecimentos múltiplos, pretende-se estimar o vetor de parâmetros de regressão $\boldsymbol{\beta}$. Para tal, é necessário adaptar a função de verosimilhança parcial (2.3) sugerida por Cox [27, 28], de modo a que seja possível acomodar mais do que um tempo de observação para um mesmo indivíduo. Assim, tal como já foi referido, o estimador de máxima verosimilhança parcial $\widehat{\boldsymbol{\beta}}$ é obtido ignorando a estrutura de dependência entre os tempos observados, isto é, a existência de correlação intraindivíduos. Por essa razão, do ponto de vista da estimação de

[5]Note-se que, em determinadas situações, um indivíduo pode ser considerado um grupo.

$\boldsymbol{\beta}$, os quatro modelos analisados na secção 3.4 são classificados como modelos marginais.

A construção da função de verosimilhança parcial varia consoante exista ou não estratificação nos modelos [45, 50, 51]. Para identificar rapidamente quais os modelos onde existe estratificação ou não, basta seguir a linha que diz respeito à função de risco subjacente da Tabela B.1 que se encontra em anexo. Num modelo em que a função de risco subjacente seja específica para cada acontecimento, os indivíduos são estratificados por acontecimento, enquanto que num modelo em que a função de risco seja comum a todos os indivíduos, essa estratificação não ocorre. Deste modo, tem-se que nos modelos PWP e WLW existe estratificação e nos modelos AG e LWA não existe estratificação.

Suponha que se encontram em estudo n indivíduos e que, para cada um deles, o vetor de p covariáveis correspondente ao s-ésimo acontecimento é dado por $\boldsymbol{z}_{is}(t) = \big(z_{is1}(t), \ldots, z_{isp}(t)\big)'$, onde $i = 1, \ldots, n$ e $s = 1, \ldots, S$. Seja t_{is} o tempo de observação do indivíduo i referente ao s-ésimo acontecimento. Consequentemente, a função de verosimilhança parcial para um modelo com estratificação pode ser definida por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left[\frac{\exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{is}(t_{is})\big)}{\sum_{j=1}^{n} Y_{js}(t_{is}) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{js}(t_{is})\big)} \right]^{\delta_{is}}, \tag{3.6}$$

enquanto para um modelo sem estratificação pode ser definida por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left[\frac{\exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{is}(t_{is})\big)}{\sum_{j=1}^{n} \sum_{l=1}^{S} Y_{jl}(t_{is}) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{jl}(t_{is})\big)} \right]^{\delta_{is}}, \tag{3.7}$$

onde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ é o vetor global de parâmetros desconhecidos, δ_{is} representa o estado do indivíduo i em relação ao s-ésimo acontecimento e $Y_{is}(t)$ é o indicador do conjunto de indivíduos que estão em risco para esse acontecimento. A aplicação de qualquer uma das funções de verosimilhança anteriores está condicionada à definição do conjunto de risco, que varia de acordo com a forma como o intervalo de risco é construído em cada modelo.

O modelo PWP considera que o intervalo de risco pode ser construído segundo a formulação por processo de contagem ou tempo por intervalos. Quando se considera o modelo PWP-CP (3.1) aplica-se a função de verosimilhança (3.6), definindo a variável indicadora do conjunto de risco por $Y_{is}(t) = I\big(t_{i(s-1)} < t \leq t_{is}\big)$. Por outro lado, quando se considera o modelo PWP-GT (3.2) também se aplica a função de verosimilhança (3.6), mas

é necessário substituir $Y_{js}(t_{is})$ por $Y_{js}(g_{is})$ e $\boldsymbol{z}_{js}(t_{is})$ por $\boldsymbol{z}_{js}\big(t_{i(s-1)} + g_{is}\big)$, onde $g_{is} = t_{is} - t_{i(s-1)}$ representa o intervalo de tempo observado entre dois acontecimentos consecutivos. Por sua vez, para este modelo, define-se $Y_{is}(t) = I\big(g_{i(s-1)} \geq t\big)$ como sendo a variável indicadora do conjunto de risco.

Relativamente ao modelo AG (3.3), o intervalo de risco é construído segundo a formulação por processo de contagem. Neste caso, recorre-se à função de verosimilhança (3.7), definindo a variável indicadora do conjunto de risco por $Y_{is}(t) = I\big(t_{i(s-1)} < t \leq t_{is}\big)$.

No que diz respeito ao modelo WLW (3.4), constrói-se o intervalo de risco através da formulação tempo total. Assim, tal como no modelo PWP, utiliza-se a função de verosimilhança (3.6), mas o conjunto de risco é definido por $Y_{is}(t) = I\big(t_{is} \geq t\big)$.

Por último, no que concerne ao modelo LWA, o intervalo de risco é construído segundo a formulação tempo total. Desta forma, tal como no modelo AG, recorre-se à função de verosimilhança (3.7), porém o conjunto de risco passa a ser definido por $Y_{is}(t) = I\big(t_{is} \geq t\big)$.

Todos os modelos marginais permitem obter um estimador global do vetor de parâmetros $\boldsymbol{\beta}$. Habitualmente, o estimador global $\widehat{\boldsymbol{\beta}}$ é obtido considerando um único vetor de covariáveis, $\boldsymbol{z}_i(t)$, no modelo que está a ser ajustado. Por conseguinte, as funções *score* correspondentes às funções de verosimilhança parciais (3.6) e (3.7), obtêm-se derivando os respetivos logaritmos em ordem aos parâmetros, sendo dadas por

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{s=1}^{S} \delta_{is} \left[\boldsymbol{z}_{is}(t_{is}) - \frac{\boldsymbol{Q}_s^{(1)}(\boldsymbol{\beta}, t_{is})}{\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta}, t_{is})} \right], \tag{3.8}$$

e

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{s=1}^{S} \delta_{is} \left[\boldsymbol{z}_{is}(t_{is}) - \frac{\overline{\boldsymbol{Q}}^{(1)}(\boldsymbol{\beta}, t_{is})}{\overline{\boldsymbol{Q}}^{(0)}(\boldsymbol{\beta}, t_{is})} \right], \tag{3.9}$$

respetivamente, onde $\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_{js}(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{js}(t)\big)$, $\boldsymbol{Q}_s^{(1)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_{js}(t) \boldsymbol{z}_{js}(t) \exp\big(\boldsymbol{\beta}' \boldsymbol{z}_{js}(t)\big)$ e $\overline{\boldsymbol{Q}}^{(r)}(\boldsymbol{\beta}, t) = \sum_{s=1}^{S} \boldsymbol{Q}_s^{(r)}(\boldsymbol{\beta}, t)$, para $r = 0, 1$. Em ambos os casos, o estimador global de máxima verosimilhança parcial $\widehat{\boldsymbol{\beta}}$ é a solução do sistema de equações $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$. Quando existe estratificação, para além do estimador global do vetor de parâmetros, também é possível obter os estimadores específicos dos vetores de parâmetros $\boldsymbol{\beta}_s$ $(s = 1, \ldots, S)$,

um para cada estrato s. Então, tanto para o modelo PWP como para o modelo WLW, os estimadores específicos são $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_S$. Para o efeito, é preciso ajustar os vetores de covariáveis específicos de cada estrato ao modelo considerado, de tal forma que $\boldsymbol{z}_i(t) = \big(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{z}_{is}(t), \mathbf{0}, \ldots, \mathbf{0}\big)'$, para $s = 1, \ldots, S$. Posto isto, é importante referir que no modelo WLW a estimação do vetor global de parâmetros $\boldsymbol{\beta}$ processa-se de forma distinta, comparativamente aos restantes três modelos. Neste caso, o estimador global $\widehat{\boldsymbol{\beta}}$ é determinado através da média ponderada dos estimadores específicos $\widehat{\boldsymbol{\beta}}_s$, de maneira que a respetiva média ponderada das variâncias robustas seja a menor possível [45] (a temática que envolve a estimação robusta da matriz de covariância será desenvolvida na próxima secção).

Os estimadores dos parâmetros de regressão, obtidos através das funções *score* (3.8) e (3.9), gozam das mesmas propriedades assintóticas enunciadas para o modelo de Cox clássico [27, 28] (estudadas na secção 2.3.1). Com efeito, vários autores [8, 14, 21, 86] mostraram que, sob determinadas condições de regularidade, $\widehat{\boldsymbol{\beta}}$ continua a ser consistente e com distribuição gaussiana p-variada, mesmo na presença de observações correlacionadas. Contudo, é preciso dar especial atenção à forma como o número de indivíduos em risco evolui ao longo do estudo, assim como ao número de indivíduos que sofre efetivamente o acontecimento em cada um dos estratos, principalmente quando se pretende estimar os vetores de parâmetros específicos $\boldsymbol{\beta}_s$.

Importa frisar que, de entre os quatro modelos, o modelo AG é o mais conservador quanto à violação do pressuposto de independência entre os tempos observados de um mesmo indivíduo. Quando não existe independência, a matriz de covariância dos estimadores dos parâmetros obtida através do estimador usual $\boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big)$ (estimador que assume que as observações são independentes) é subestimada. Nesta situação, deve-se recorrer ao estimador robusto da matriz de covariância que será definido na secção seguinte. Então, para testar a validade deste pressuposto, basta comparar as duas estimativas: usual *versus* robusta. O modelo AG é considerado válido, apenas quando a variância robusta é ligeiramente superior à variância usual. Caso a variância robusta seja muito superior à variância usual, a aplicação do modelo AG não é recomendada. Conforme discutido por Lin e Wei [52], se a condição de incrementos independentes não é satisfeita, o modelo AG pode deixar de ser um modelo de riscos proporcionais, tornando o valor assintótico de $\widehat{\boldsymbol{\beta}}$ difícil de interpretar. Ainda assim, para níveis baixos de correlação intraindivíduos, $\widehat{\boldsymbol{\beta}}$ é uma estimativa útil, mesmo quando os pressupostos subjacentes não são verificados.

3.6 Estimador robusto da matriz de covariância

Os quatro modelos analisados atrás recorrem ao modelo de Cox clássico para modelar o tempo até a ocorrência de acontecimentos múltiplos, o que implica que a estimação dos parâmetros seja feita assumindo que as observações são independentes. Este facto não tem qualquer influência para a estimação dos parâmetros de regressão, visto que a consistência e as propriedades assintóticas de $\widehat{\boldsymbol{\beta}}$ encontram-se salvaguardadas, tal como já foi visto anteriormente. Assim, pode-se utilizar as estimativas obtidas através desses estimadores para realizar inferência estatística sobre os dados.

No entanto, a matriz de covariância dos estimadores dos parâmetros obtida através do processo usual, isto é, da inversa da matriz de informação de Fisher $\boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big)$, não pode ser considerada por si só uma aproximação válida. Para este caso, tendo presente a relação $\boldsymbol{I}(\boldsymbol{\beta}) = -{}^{\partial^2 \log L(\boldsymbol{\beta})}\!/_{\partial \boldsymbol{\beta}^2}$, a matriz de informação de Fisher associada a um modelo estratificado é dada por

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{s=1}^{S}\delta_{is}\left[\frac{\boldsymbol{Q}_s^{(2)}(\boldsymbol{\beta},t_{is})}{\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta},t_{is})} - \frac{\boldsymbol{Q}_s^{(1)}(\boldsymbol{\beta},t_{is})\boldsymbol{Q}_s^{(1)}(\boldsymbol{\beta},t_{is})'}{\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta},t_{is})^2}\right], \tag{3.10}$$

e para um modelo não estratificado é dada por

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{s=1}^{S}\delta_{is}\left[\frac{\overline{\boldsymbol{Q}}^{(2)}(\boldsymbol{\beta},t_{is})}{\overline{\boldsymbol{Q}}^{(0)}(\boldsymbol{\beta},t_{is})} - \frac{\overline{\boldsymbol{Q}}^{(1)}(\boldsymbol{\beta},t_{is})\overline{\boldsymbol{Q}}^{(1)}(\boldsymbol{\beta},t_{is})'}{\overline{\boldsymbol{Q}}^{(0)}(\boldsymbol{\beta},t_{is})^2}\right], \tag{3.11}$$

onde $\boldsymbol{Q}_s^{(2)}(\boldsymbol{\beta},t) = \sum_{j=1}^{n} Y_{js}(t)\boldsymbol{z}_{js}(t)\boldsymbol{z}_{js}(t)'\exp\left(\boldsymbol{\beta}'\boldsymbol{z}_{js}(t)\right)$ e $\overline{\boldsymbol{Q}}^{(2)}(\boldsymbol{\beta},t) = \sum_{s=1}^{S}\boldsymbol{Q}_s^{(2)}(\boldsymbol{\beta},t)$. Note-se que as matrizes (3.10) e (3.11) são deduzidas a partir das funções de verosimilhança parciais (3.6) e (3.7), respetivamente, pelo que estas também não têm a capacidade de ter em conta a correlação entre os tempos observados de um mesmo indivíduo.

A potencial existência de correlação levou ao desenvolvimento de um outro estimador para a matriz de covariância. Em 1989, Lin e Wei [52] recorreram aos conhecidos estimadores "*sandwich*", com o objetivo de desenvolverem um estimador robusto para o modelo de Cox clássico. Ainda nesse ano, Wei *et al.* [86] generalizaram esse estimador robusto de forma a acomodar mais do que um tempo de vida para um mesmo indivíduo, no mesmo artigo em que também propuseram o modelo WLW. Estes autores demonstraram que, aproximando a estatística $\boldsymbol{U}(\boldsymbol{\beta})$ por uma soma de n vetores

aleatórios independentes e identicamente distribuídos, pode-se estabelecer a normalidade assintótica de $\boldsymbol{U}(\boldsymbol{\beta})$ e, consequentemente, obter a sua matriz de covariância limite [51]. Deste modo, quando n é suficientemente grande em relação a S, a estatística $\boldsymbol{U}(\boldsymbol{\beta})$ é assintoticamente gaussiana p-variada, com valor médio $\mathbf{0}$ e matriz de covariância definida por

$$\boldsymbol{V}\big(\widehat{\boldsymbol{\beta}}\big) = \sum_{i=1}^{n} \sum_{s=1}^{S} \sum_{l=1}^{S} \boldsymbol{W}_{is}\big(\widehat{\boldsymbol{\beta}}\big) \boldsymbol{W}_{il}\big(\widehat{\boldsymbol{\beta}}\big)',$$

em que a matriz $\boldsymbol{W}_{is}(\boldsymbol{\beta})$ varia consoante no modelo considerado exista ou não estratificação. Assim, para um modelo com estratificação utiliza-se

$$\begin{aligned} \boldsymbol{W}_{is}(\boldsymbol{\beta}) = \; & \delta_{is} \left[\boldsymbol{z}_{is}(t_{is}) - \frac{\boldsymbol{Q}_s^{(1)}(\boldsymbol{\beta}, t_{is})}{\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta}, t_{is})} \right] - \\ & \sum_{j=1}^{n} \frac{\delta_{js} Y_{is}(t_{js}) \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_{is}(t_{js})\right)}{\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta}, t_{js})} \left[\boldsymbol{z}_{is}(t_{js}) - \frac{\boldsymbol{Q}_s^{(1)}(\boldsymbol{\beta}, t_{js})}{\boldsymbol{Q}_s^{(0)}(\boldsymbol{\beta}, t_{js})} \right], \end{aligned}$$

enquanto para um modelo sem estratificação utiliza-se

$$\begin{aligned} \boldsymbol{W}_{is}(\boldsymbol{\beta}) = \; & \delta_{is} \left[\boldsymbol{z}_{is}(t_{is}) - \frac{\overline{\boldsymbol{Q}}^{(1)}(\boldsymbol{\beta}, t_{is})}{\overline{\boldsymbol{Q}}^{(0)}(\boldsymbol{\beta}, t_{is})} \right] - \\ & \sum_{j=1}^{n} \sum_{l=1}^{S} \frac{\delta_{jl} Y_{is}(t_{jl}) \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_{is}(t_{jl})\right)}{\overline{\boldsymbol{Q}}^{(0)}(\boldsymbol{\beta}, t_{jl})} \left[\boldsymbol{z}_{is}(t_{jl}) - \frac{\overline{\boldsymbol{Q}}^{(1)}(\boldsymbol{\beta}, t_{jl})}{\overline{\boldsymbol{Q}}^{(0)}(\boldsymbol{\beta}, t_{jl})} \right]. \end{aligned}$$

Foi ainda demonstrado que, se o modelo marginal for corretamente especificado (adequado ao caso em estudo) e se os tempos observados de um mesmo indivíduo forem independentes, então a matriz $\boldsymbol{V}\big(\widehat{\boldsymbol{\beta}}\big)$ é assintoticamente equivalente à matriz $\boldsymbol{I}\big(\widehat{\boldsymbol{\beta}}\big)$.

Por conseguinte, o estimador robusto da matriz de covariância de $\widehat{\boldsymbol{\beta}}$, habitualmente designado por estimador "*sandwich*", é definido por

$$\boldsymbol{D}\big(\widehat{\boldsymbol{\beta}}\big) = \boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big) \boldsymbol{V}\big(\widehat{\boldsymbol{\beta}}\big) \boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big). \tag{3.12}$$

Um aspeto que merece ser salientado neste estimador, é o facto de a sua expressão incorporar o estimador usual $\boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big)$. Quer isto dizer que, na

prática, o estimador "*sandwich*" permite introduzir a correção necessária sobre o estimador usual, de modo a que a potencial correlação existente entre as observações seja tida em consideração. Por essa razão, os modelos marginais são também conhecidos por modelos de variância corrigida (*variance-correction models*) [16].

Com base no estimador robusto da matriz de covariância (3.12), o investigador pode averiguar se os dados se encontram correlacionados, assim como avaliar corretamente a precisão das estimativas dos parâmetros de regressão. Quando existe correlação intraindivíduos, a estimativa da matriz de covariância de $\widehat{\boldsymbol{\beta}}$ obtida através do estimador $\boldsymbol{D}\big(\widehat{\boldsymbol{\beta}}\big)$, é consideravelmente maior do que aquela obtida através do estimador $\boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big)$. Porém, também podem existir situações em que o contrário acontece. De acordo com Kelly e Lim [45], se a estimativa robusta for inferior à estimativa usual, significa que se está perante uma situação em que existe maior variabilidade intraindivíduos do que entre indivíduos.

Inicialmente, o estimador robusto $\boldsymbol{D}\big(\widehat{\boldsymbol{\beta}}\big)$ foi proposto para os modelos WLW e LWA. De facto, quando os modelos PWP e AG foram sugeridos, não apresentaram nenhuma medida que ajustasse a correlação intraindivíduos. Contudo, os autores destes modelos estavam conscientes dessa forte possibilidade e recomendaram tentar captar essa correlação incluindo covariáveis dependentes do tempo no modelo. Alguns anos mais tarde, percebeu-se que era possível tirar partido do facto destes dois modelos também serem marginais do ponto de vista da estimação dos parâmetros e, da mesma forma, passou-se a aplicar-lhes este estimador.

Para qualquer um dos modelos marginais, pode-se aplicar os três testes de hipóteses enunciados na secção 2.4, mas é importante ter presente alguns aspetos acerca dos mesmos:

- O teste *score* (2.10) pode ser generalizado para comparar vários grupos de indivíduos que diferem em mais do que uma característica, sendo a respetiva estatística de teste dada por $\boldsymbol{U}'(\mathbf{0})\boldsymbol{I}^{-1}(\mathbf{0})\boldsymbol{U}(\mathbf{0})$. Este teste utiliza o estimador usual da matriz de covariância, pelo que assume que as observações são independentes. A sua aplicação deixa de ser apropriada quando existe correlação, o que fez com que fosse desenvolvida uma versão robusta deste teste, que consiste em substituir o estimador usual $\boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big)$ pela matriz de covariância limite $\boldsymbol{V}\big(\widehat{\boldsymbol{\beta}}\big)$ da função *score*. Assim, passa-se a aplicar o teste *score* robusto dado por $\boldsymbol{U}'(\mathbf{0})\boldsymbol{V}^{-1}(\mathbf{0})\boldsymbol{U}(\mathbf{0})$.

- O teste de Wald (2.11) também pode ser generalizado, com o intuito de testar se um subconjunto de covariáveis tem efeito significativo no tempo de vida dos indivíduos, sendo a respetiva estatística de teste dada por $\widehat{\boldsymbol{\beta}}' \, \boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big) \, \widehat{\boldsymbol{\beta}}$. Este teste também recorre ao estimador usual da matriz de covariância, pelo que a sua aplicação deixa de ser adequada na presença de dados correlacionados. Para este caso, foi desenvolvido o teste de Wald robusto, que consiste em substituir diretamente o estimador usual $\boldsymbol{I}^{-1}\big(\widehat{\boldsymbol{\beta}}\big)$ pelo estimador robusto $\boldsymbol{D}\big(\widehat{\boldsymbol{\beta}}\big)$, de onde resulta $\widehat{\boldsymbol{\beta}}' \, \boldsymbol{D}^{-1}\big(\widehat{\boldsymbol{\beta}}\big) \, \widehat{\boldsymbol{\beta}}$.

- O teste da razão de verosimilhanças (2.12), ao contrário dos outros dois testes, não introduz nenhuma modificação para compensar a possibilidade de existir correlação, dado que apenas se baseia na função de verosimilhança parcial. Deste modo, para aplicar este teste deve-se primeiro averiguar se existe correlação intraindivíduos, comparando as estimativas obtidas para a matriz de covariância: usual *versus* robusta. Só assim é possível realizar uma interpretação correta sobre o resultado obtido.

Tanto o teste *score* robusto como o teste de Wald robusto são mais conservadores[6] do que os respetivos testes que utilizam a estimativa usual da matriz de covariância. Em 2006, Boher e Cook [14] efetuaram um estudo, com base em dados simulados, para avaliar o comportamento do teste *score* utilizando a estimativa usual e a estimativa robusta. Nesse estudo, compararam o desempenho dos testes nos quatros modelos marginais. Em geral, os autores concluíram que a utilização apropriada da estimativa robusta garante a validade dos testes de hipóteses, no sentido em que leva a um bom controlo do erro de tipo I (incorreta rejeição da hipótese nula quando esta é verdadeira).

De igual modo, os quatro tipos de resíduos desenvolvidos no capítulo anterior podem e devem ser aplicados para avaliar a adequabilidade de qualquer um dos modelos. Para isso, é necessário efetuar as devidas adaptações de modo a que a análise de resíduos tenha em conta a ocorrência de vários acontecimentos para um mesmo indivíduo. Neste ponto, interessa referir que Lee *et al.* [49], quando construíram o modelo LWA, destacaram a importância da aplicação da matriz $\boldsymbol{D}\big(\widehat{\boldsymbol{\beta}}\big)$, mostrando que a estimativa robusta da matriz de covariância de $\widehat{\boldsymbol{\beta}}$ pode ser obtida utilizando apenas os resíduos *score* [79]. Recorde que a importância destes resíduos para a análise de acontecimentos múltiplos já havia sido referida na secção 2.5.2.

[6] Para mais informações acerca destes testes consulte-se Lin [51] e Therneau e Grambsch [79].

Capítulo 4

Acontecimentos múltiplos no R

4.1 Introdução

Nos últimos anos, as quatro extensões do modelo de Cox abordadas na secção 3.4 têm sido amplamente aplicadas, tanto a dados reais como a dados simulados. Um dos principais propósitos deste capítulo é apresentar o modo como esses modelos podem ser aplicados na prática, recorrendo para o efeito ao *software* estatístico R. Todavia, em primeiro lugar é necessário dispor de um conjunto de dados. Uma vez que não foi possível obter um conjunto real de dados, surgiu a ideia de proceder à sua simulação. Na literatura é possível encontrar diversos artigos [14, 45, 63, 84] que revelam que a simulação de dados é uma alternativa cada vez mais utilizada pelos investigadores. Por essa razão, também se pretende indicar algumas das ferramentas disponíveis neste *software* para simular acontecimentos múltiplos e, em particular, acontecimentos recorrentes.

Conforme mencionado por Carvalho *et al.* [22], Therneau e Grambsch [79] e Therneau e Hamilton [81], os quatro modelos marginais podem ser implementados com recurso a qualquer um dos *softwares* estatísticos: R, S-Plus ou SAS. Neste contexto, tem interesse referir que ambos os *softwares* R e S-Plus têm por base a linguagem de programação S, razão pela qual muitos dos códigos desenvolvidos para o S-Plus podem ser facilmente implementados no R e, por vezes, nem é necessário efetuar qualquer alteração a esses códigos.

No decorrer deste capítulo, assim como em todo o desenvolvimento desta dissertação, decidiu-se apenas explorar alguns dos *packages* disponíveis no R para abordar a temática dos acontecimentos múltiplos, uma vez que este *software* beneficia do facto de ser gratuito. Este aspeto revela-se uma mais-valia para toda a comunidade académica, em especial na área da investigação, pois permite que qualquer indivíduo/organização possa contribuir para o seu aperfeiçoamento.

Após a obtenção dos dados simulados, serão feitas algumas considerações acerca do cuidado que é necessário ter na construção e organização da base de dados, visto que a sua formatação deve ser adequada às características que cada modelo apresenta, onde a definição do intervalo de risco se revela uma componente decisiva. Posto isto, proceder-se-á à implementação dos modelos marginais, terminando com uma breve discussão dos resultados obtidos.

Importa referir que a simulação de dados, assim como a respetiva análise estatística, foi efetuada por intermédio da versão 3.2.5 do *software* estatístico R.

4.2 Simulação de dados

A simulação de dados é uma alternativa que tem sido frequentemente utilizada, não só na estatística como em muitas outras áreas das ciências exatas. Habitualmente, recorre-se a dados simulados quando se pretende efetuar estudos empíricos acerca das potencialidades de um novo método ou modelo. A simulação permite controlar, até certo ponto, algumas características que se deseja que estejam presentes no conjunto de dados. De facto, se houver um controlo prévio sobre determinadas condições, tais como: percentagem de indivíduos censurados; distribuição do tempo de vida associado a cada acontecimento observado; definição do efeito de cada covariável no tempo de vida dos indivíduos; entre outros, torna-se mais fácil avaliar com exatidão o desempenho do modelo que está a ser considerado. Além disso, a possibilidade de obter um conjunto de dados com as caraterísticas pretendidas também contribui para a simplificação de estudos onde se pretende comparar dois ou mais modelos. Assim, esta alternativa revela-se extremamente útil, permitindo antecipar determinadas situações que no futuro podem vir a tornar-se uma realidade.

Neste trabalho, a simulação de dados também vai de encontro aos compromissos assumidos, onde o principal objetivo é exemplificar de que forma os modelos marginais podem ser aplicados com recurso ao *software* estatístico R. Após alguma pesquisa na rede de arquivos do R (*Comprehensive R Archive Network* – CRAN), em particular na *task view* dedicada ao ramo da Análise de Sobrevivência [6], encontrou-se um *package* que foi expressamente desenvolvido para a simulação de dados de sobrevivência simples e complexos, o qual se designa por survsim [58]. Este *package* foi desenvolvido por Moriña e Navarro [57], em 2014, e possui a capacidade de simular acontecimentos múltiplos, podendo estes ser do mesmo tipo ou de tipos diferentes. Porém, para que os quatro modelos marginais possam ser aplicados ao mesmo conjunto de dados e, consequentemente, seja possível compará-los na

análise de resultados, apenas será abordada a situação de gerar acontecimentos múltiplos do mesmo tipo, ou seja, acontecimentos recorrentes. Para o efeito, utilizou-se a função rec.ev.sim contida nesse *package*. O procedimento efetuado para a simulação dos dados, assim como a descrição dos argumentos utilizados nessa função, encontram-se disponíveis no anexo C.1.

Na simulação do conjunto de dados, teve-se a preocupação de controlar apenas alguns dos argumentos que se podem definir na função rec.ev.sim, de maneira a que este procedimento fosse o mais simples possível. Essencialmente, gerou-se uma amostra aleatória constituída por $n = 1\,000$ indivíduos, para os quais se definiu um tempo máximo de *follow-up* igual a 1 825 dias (o que equivale a 5 anos). Considerou-se que o tempo até à ocorrência de cada acontecimento, assim como o tempo até à ocorrência de censura à direita, seguem uma distribuição de Weibull. Posto isto, verificou-se que neste *package* as covariáveis podiam ser geradas através de três distribuições distintas, tendo-se decidido explorar as três possibilidades. Simulou-se então uma variável categórica x com distribuição de Bernoulli com probabilidade de sucesso igual a 0.5, e duas variáveis contínuas, uma designada por x.1 com distribuição uniforme que toma valores no intervalo $[0, 1]$ e outra designada por x.2 com distribuição gaussiana padrão.

Após a realização de todo este procedimento, obteve-se o conjunto de dados denominado por dados1. Através da função head(dados1, 7) é possível visualizar as primeiras sete linhas deste conjunto de dados (arredondados), tal como se observa na Tabela 4.1.

Tabela 4.1: Visualização das primeiras sete entradas do conjunto de dados – dados1.

nid	obs.episode	status	start	stop	time	x	x.1	x.2
1	1	0	0.000	139.190	139.190	0	0.579	-0.087
2	1	1	0.000	261.651	261.651	1	0.446	1.309
2	2	0	261.651	424.657	163.006	1	0.446	1.309
3	1	0	0.000	1 428.484	1 428.484	1	0.072	-1.611
4	1	1	0.000	626.210	626.210	1	0.254	0.325
4	2	1	626.210	857.427	231.218	1	0.254	0.325
4	3	0	857.427	1 306.826	449.398	1	0.254	0.325

Na verdade, o conjunto de dados que foi simulado apresenta mais variáveis do que aquelas que estão na tabela anterior. Todavia, para o desenvolvimento das secções seguintes apenas é necessário considerar aquelas que se encontram aqui retratadas, as quais serão descritas em seguida:

nid – número de identificação do indivíduo;

obs.episode – número do acontecimento a que corresponde o tempo observado;

status – variável indicatriz que toma o valor 1 quando o acontecimento é observado e 0 quando não é;

start – instante de tempo a partir do qual um indivíduo passa a estar em risco de sofrer o acontecimento correspondente;

stop – instante de tempo em que um indivíduo deixa de estar em risco de sofrer o acontecimento correspondente;

time – tempo observado até à ocorrência do acontecimento de interesse ou até à censura;

x, x.1 e x.2 – valor de cada uma das covariáveis geradas aleatoriamente.

Importa referir que a partir da função rec.ev.sim é possível gerar acontecimentos instantâneos ou acontecimentos duradouros. Porém, optou-se novamente pela situação mais simples, pelo que foram simulados apenas acontecimentos instantâneos. Significa então que um indivíduo volta a estar em risco imediatamente após a ocorrência de cada acontecimento, a menos que seja censurado, tal como se observa nas colunas start e stop da Tabela 4.1. Repare-se que a coluna time pode ser obtida por intermédio da diferença stop − start. Além disso, um outro aspeto que é preciso ter presente daqui em diante é o facto de se estar a assumir que todos os indivíduos começam a ser seguidos no instante 0 (ver coluna start), o que quer dizer que não se considerou a possibilidade de os indivíduos terem uma entrada atrasada no estudo.

Tabela 4.2: Resumo da informação com maior relevância sobre os dados simulados.

Número do acontecimento	Indivíduos em risco	Acontecimentos observados	Mediana do tempo de *follow-up*[1]
1	1 000	365	383.427
2	365	162	184.971
3	162	87	107.623
4	87	48	18.385
5	48	27	33.954
> 5	27	34	17.727
Resultados globais	1 000	723	251.964

[1] Os valores indicados são diferentes dos valores da mediana do tempo de vida.

Para sumarizar toda a informação pertinente acerca do conjunto de dados que foi simulado, basta recorrer à função summary(dados1). Na Tabela 4.2

encontra-se compilada parte dessa informação, a qual será essencial para a análise de resultados. Assim, obtém-se de imediato uma visão global sobre as características dos dados, sem que ainda tenha sido aplicado qualquer modelo. Como se pode observar, essa informação encontra-se organizada por acontecimento, o que permite obter uma breve perspetiva sobre o que aconteceu em cada um deles. Note-se que a segunda coluna representa o número de indivíduos em risco em cada um dos acontecimentos, de onde é possível confirmar que os dados simulados correspondem a acontecimentos recorrentes. De facto, é bem visível que se está a assumir que os indivíduos só podem estar em risco para um acontecimento de cada vez, o que retrata uma das mais importantes características que este tipo de dados apresenta. Além do mais, os valores 365, 162, 87, 48 e 27, aparecem tanto na segunda como na terceira coluna, o que revela a natureza instantânea dos acontecimentos.

4.3 Construção da base de dados

A chave para que um modelo seja corretamente ajustado é a construção de uma base de dados apropriada às suas especificidades. Deste modo, a organização da base de dados é uma etapa fundamental na aplicação de qualquer modelo, que por vezes pode revelar-se desafiante na análise de dados mais complexos, como acontece no caso em que se regista mais do que um acontecimento para um mesmo indivíduo. Assim sendo, antes de implementar cada um dos modelos marginais é preciso perceber de que forma a base de dados deve ser organizada.

Em qualquer modelo marginal, a construção da base de dados é feita de forma a que cada entrada/linha diga respeito a um tempo observado, ao qual é necessário fazer corresponder um indivíduo. Para além disso, é preciso indicar se esse tempo foi ou não censurado e, ainda, a ordem pela qual este foi registado para aquele indivíduo em questão. O registo da ordenação dos tempos observados deve de ser efetuado com imenso cuidado, uma vez que essa variável será utilizada como variável de estratificação nos modelos que consideram uma função de risco subjacente específica para cada acontecimento. Posto isto, também é importante proceder ao registo de todas as covariáveis que se julgue poderem vir a afetar o tempo até a ocorrência de cada um dos acontecimentos.

A construção da base de dados adequada a cada modelo marginal difere essencialmente na forma como se regista os tempos observados. Quando se definiu formalmente estes modelos enfatizou-se para o facto dos modelos PWP-CP (3.1) e WLW (3.4), bem como dos modelos AG (3.3) e LWA (3.5), terem funções de risco semelhantes. Foi então referido que a grande diferença entre estes modelos está na formulação do intervalo de risco que

cada um considera, o que consequentemente terá influência na forma como os tempos devem ser registados. Para perceber facilmente como é que esse registo se processa em cada uma das formulações, considere-se o exemplo que se segue. Suponha-se que um dado indivíduo começou a ser seguido no instante 0 e sofreu acontecimentos múltiplos instantâneos nos instantes t_1 e t_2, acabando por ser censurado em t_3. Os tempos observados deste indivíduo, ou seja, os seus intervalos de risco, devem ser registados tal como se encontra indicado na Tabela 4.3.

Tabela 4.3: Influência da formulação do intervalo de risco na construção da base de dados.

Processo de contagem	Tempo por intervalos	Tempo total
$(0,\ t_1]$	$(0,\ t_1]$	$(0,\ t_1]$
$(t_1,\ t_2]$	$(0,\ t_2 - t_1]$	$(0,\ t_2]$
$(t_2,\ t_3]$	$(0,\ t_3 - t_2]$	$(0,\ t_3]$

Qualquer que seja a formulação adotada, o registo dos tempos observados deve ser feito indicando o instante de entrada do indivíduo no conjunto de risco e o instante em que ocorreu o acontecimento de interesse ou a saída do estudo. Alternativamente, quando se considera as formulações tempo por intervalos ou tempo total, pode-se registar apenas as amplitudes dos intervalos de risco, uma vez que nestas formulações os instantes iniciais mantêm-se constantes. Note-se que, em todos os casos, um indivíduo encontra-se em risco para o primeiro acontecimento no intervalo $(0,\ t_1]$, aspeto este que já havia sido referido na secção 3.3.1.

Voltando ao conjunto de dados que foi simulado na secção anterior (dados1), o qual se encontra representado na Tabela 4.1, constata-se que este foi construído por intermédio das formulações processo de contagem (ver colunas start e stop) e tempo por intervalos (ver coluna time). Deste modo, esta base de dados já se encontra em condições para que os modelos PWP-CP, PWP-GT e AG, possam ser implementados.

Embora numa rápida leitura da base de dados1 não seja possível verificar a formulação tempo total, esta também se encontra presente neste conjunto de dados. Ao examinar com melhor atenção a Tabela 4.3, verifica-se que a partir de qualquer uma das formulações do intervalo de risco é possível obter as outras duas. Para além disso, comparando as formulações processo de contagem e tempo total, chega-se à conclusão que os instantes finais dos respetivos intervalos de risco são os mesmos. Como a formulação tempo total permite que seja registada a amplitude dos intervalos de risco e, neste caso,

os instantes iniciais são todos iguais a 0, a coluna stop pode ser vista como o tempo decorrido desde o início do estudo até à ocorrência do acontecimento de interesse ou à censura. Por outras palavras, a coluna stop por si só representa a formulação tempo total.

Após identificar a formulação do intervalo de risco que será utilizada nos modelos WLW e LWA, é necessário referir que para estes modelos a construção da base de dados realiza-se de forma diferente. A razão para que isso aconteça está no facto de se assumir que, a partir do instante em que um indivíduo começa a ser seguido, ele encontra-se simultaneamente em risco para a ocorrência de qualquer um dos acontecimentos. Assim sendo, é fulcral que a base de dados reflita este aspeto, pois só assim é possível diferenciar um modelo marginal de um modelo condicional, relativamente à estrutura de dependência entre acontecimentos.

No caso do modelo WLW, se o maior número de acontecimentos que um indivíduo pode vir a sofrer for S, então para cada indivíduo em estudo terão que ser registados S tempos observados. Desta forma, haverá sempre n observações associadas a cada acontecimento e, consequentemente, a base de dados terá que ser constituída por $n \times S$ linhas. Já no caso do modelo LWA, a construção da base de dados processa-se de forma semelhante, o único pormenor a ter em conta é que os grupos não têm que ter obrigatoriamente a mesma dimensão, ou seja, pode haver situações em que o número de entradas é diferente de grupo para grupo. No entanto, como neste trabalho se está a efetuar um estudo com dados simulados, admite-se que se está perante a situação mais simples, isto é, que os grupos têm a mesma dimensão. Assim, pode-se utilizar a mesma base de dados para os modelos WLW e LWA.

Tendo em conta todos os aspetos que acabaram de ser mencionados, será necessário construir uma nova base de dados, a qual se designará por dados2. Executando no R o comando max(dados1$obs.episode), determinou-se que nesta base de dados cada indivíduo terá que ter $S = 10$ linhas. Porém, através do comando max(dados1[dados1$status==1, "obs.episode"]) obteve-se que o número máximo de acontecimentos observados para um mesmo indivíduo foi 9, o que significa que nenhum dos indivíduos em risco para o acontecimento de ordem 10 sofreu esse acontecimento. Em todo o caso, será necessário construir uma nova base de dados com $n \times S = 1\,000 \times 10 = 10\,000$ entradas. Esta nova base de dados é obtida por intermédio da base de dados1. Basicamente a ideia é repetir os valores da última linha de cada indivíduo, o número de vezes necessário até atingir 10 linhas. Depois, só é preciso garantir que a variável obs.episode toma, para todos os indivíduos, os valores consecutivos $1, 2, \ldots, 10$. Em relação às variáveis que definem o tempo observado, só é preciso englobar na base de dados2 a variável stop, mas sugere-se que a sua designação seja alterada para fulltime, de modo a evitar qualquer conflito.

Com o intuito de apresentar o aspeto final da base de dados2, representou-se na Tabela 4.4 apenas as entradas[1] do indivíduo 4. Tal como é possível observar, este indivíduo sofreu dois acontecimentos e foi censurado quando se encontrava em risco para o acontecimento de ordem 3. Repare-se que este indivíduo também se encontra representado na Tabela 4.1, onde apenas apresenta 3 entradas.

Tabela 4.4: Visualização das dez entradas do indivíduo 4 no conjunto de dados – dados2.

nid	obs.episode	status	fulltime	x	x.1	x.2
4	1	1	626.210	1	0.254	0.325
4	2	1	857.427	1	0.254	0.325
4	3	0	1 306.826	1	0.254	0.325
4	4	0	1 306.826	1	0.254	0.325
4	5	0	1 306.826	1	0.254	0.325
4	6	0	1 306.826	1	0.254	0.325
4	7	0	1 306.826	1	0.254	0.325
4	8	0	1 306.826	1	0.254	0.325
4	9	0	1 306.826	1	0.254	0.325
4	10	0	1 306.826	1	0.254	0.325

Após a construção das duas bases de dados é preciso verificar exaustivamente se todas as variáveis se encontram bem definidas, em particular a variável de estratificação (obs.episode), a variável que indica se foi observado o acontecimento ou a censura (status) e, por fim, aquelas que representam os tempos observados (start, stop, time e fulltime). Ainda que este seja um processo que requer tempo e rigor, é imprescindível para que as funções que serão implementadas no R permitam obter resultados fidedignos.

4.4 Implementação dos modelos marginais

Neste momento, pode dizer-se que estão reunidas as condições necessárias para se proceder à implementação dos modelos marginais. Interessa salientar que para o ajustamento dos modelos apresentados nesta secção utilizou-se o *package* survival[2] desenvolvido por Therneau [78].

[1]Para que seja possível visualizar no R as entradas de um único indivíduo, basta o aplicar comando que se segue: dados2[dados2$nid==4,].

[2]Para instalar e carregar o *package* survival utilizam-se os comandos install.packages("survival") e library(survival), respetivamente.

Antes de aplicar qualquer modelo marginal é usual efetuar a análise do tempo até à ocorrência do primeiro acontecimento. Embora exista uma grande possibilidade de se estar a desperdiçar informação, é importante tentar criar à partida uma base de comparação. O objetivo é permitir que, numa fase posterior, seja possível avaliar se a informação proveniente dos acontecimentos múltiplos acrescenta ou não algum proveito à análise. Desta forma, começou-se por implementar o modelo de Cox clássico:

```
> modeloCox <- coxph(Surv(start, stop, status) ~ as.factor(x) + x.1 + x.2 +
    cluster(nid), data = dados1, subset = (obs.episode == 1))
> summary(modeloCox)

n = 1000, number of events = 365

                    coef exp(coef) se(coef) robust se      z       p
as.factor(x)[T.1] 0.46477   1.59165 0.10648   0.10416  4.462 8.12e-06
x.1               0.01177   1.01184 0.18203   0.17741  0.066   0.947
x.2               1.06403   2.89802 0.06688   0.06445 16.510  <2e-16

Concordance = 0.736 (se = 0.016)
Rsquare = 0.25 (max possible = 0.99)
Likelihood ratio test = 288.3 on 3 df, p = 0
Wald test             = 284.7 on 3 df, p = 0
Score (logrank) test = 270.2 on 3 df, p = 0, Robust = 188.6 p = 0

(Note: the likelihood ratio and score tests assume independence of observations
    within a cluster, the Wald and robust score tests do not).
```

A função principal que permite ajustar os parâmetros de qualquer modelo é a função coxph(Surv(intervalo.risco, status) ~ covariáveis, data=base.dados), a qual está dividida em duas partes. Na primeira parte define-se os seguintes argumentos: a formulação do intervalo de risco associada ao modelo considerado; a variável que indica a ocorrência do acontecimento de interesse ou a censura; e as covariáveis[3] registadas no estudo. Quanto à segunda parte, esta é utilizada para designar a base de dados que foi previamente construída para o modelo em questão. Nesta situação adicionou-se o argumento subset porque se pretendia analisar apenas o primeiro acontecimento. Para além destes argumentos básicos, também se incluiu a função cluster(nid) que é, possivelmente, a maior novidade neste tipo de análise. Esta função estipula que as entradas da base de dados com o mesmo número de identificação, nid,

[3]Neste caso, utilizou-se a função as.factor(x) porque a covariável x é categórica.

pertencem a um único indivíduo. Assim, a função coxph procede à correção da estimativa da variância usual, pois quando se analisam acontecimentos múltiplos existe a forte possibilidade de existir correlação intraindivíduos. Note-se que, para analisar o tempo até à ocorrência do primeiro acontecimento utilizou-se a base de dados dados1, mas também se podia ter usado a base de dados dados2. De facto, pelo menos os intervalos de risco do primeiro acontecimento são iguais em ambas as bases de dados, o que faz com que, nesta situação, seja possível recorrer a qualquer uma delas. Além do mais, em vez da formulação por processo de contagem também se podia ter usado qualquer uma das outras duas formulações, uma vez que estas não apresentam diferenças no intervalo de risco para o primeiro acontecimento, tal como já foi visto anteriormente.

Em relação ao *output* que foi gerado, apresentou-se somente a parte que diz respeito à estimação dos parâmetros, tendo sido omitida a informação referente à interpretação das covariáveis. Na coluna coef observam-se as estimativas dos parâmetros obtidas através da função de verosimilhança parcial e na coluna exp(coef) tem-se o risco relativo associado às covariáveis. Seguem-se as colunas se(coef) e robust se, que representam as estimativas do erro padrão usual e robusta, respetivamente. As últimas duas colunas referem-se à estatística de Wald e ao correspondente valor-p, que servem para avaliar se uma dada covariável tem influência significativa no tempo de vida dos indivíduos, na presença das restantes covariáveis. É de salientar que quando se utiliza a função cluster(nid), o R emite sempre uma nota de aviso com o intuito de informar que apenas os testes de Wald e *score* robusto foram corrigidos, estando assim livres da existência de correlação.

Após analisar o primeiro acontecimento, procedeu-se à implementação dos modelos marginais. Em primeiro lugar, implementou-se o modelo PWP-CP do seguinte modo:

```
> modeloPWPCP <- coxph(Surv(start, stop, status) ~ as.factor(x) + x.1 + x.2 +
    strata(obs.episode) + cluster(nid), data = dados1)
> summary(modeloPWPCP)

n = 1723, number of events = 723

                    coef exp(coef) se(coef) robust se     z       p
as.factor(x)[T.1] 0.52806   1.69563  0.08076   0.08619 6.127 8.98e-10
x.1               0.08727   1.09119  0.13376   0.14708 0.593   0.553
x.2               0.67621   1.96640  0.04761   0.05002 13.519  <2e-16
```

```
Concordance = 0.722 (se = 0.017)
Rsquare = 0.129 (max possible = 0.983)
Likelihood ratio test = 237.9 on 3 df, p = 0
Wald test             = 205.4 on 3 df, p = 0
Score (logrank) test = 231.3 on 3 df, p = 0, Robust = 172 p = 0

(Note: the likelihood ratio and score tests assume independence of observations
    within a cluster, the Wald and robust score tests do not).
```

Tal como se observa, deixou de se utilizar o argumento subset, o que significa que todos os acontecimentos foram englobados na análise. Comparando o comando do modelo PWP-CP com o anterior, também se verifica que se passou a incluir a função strata(obs.episode). Recorde-se que o modelo PWP é um modelo estratificado e, por isso, utilizou-se esta função para permitir que exista uma função de risco subjacente específica para cada acontecimento, consoante a ordenação definida na variável obs.episode. O comando associado ao modelo PWP-GT é semelhante ao anterior, com a exceção das variáveis start e stop que são substituídas pela variável time, dado que a este modelo está associada a formulação tempo por intervalos. Em relação ao *output* que foi gerado, este apresenta uma estrutura idêntica àquela que foi examinada atrás e o mesmo acontece para os restantes modelos. Assim, daqui em diante apenas serão feitas algumas considerações acerca do código do R que cada modelo utiliza, sendo que os respetivos *outputs* podem ser consultados na secção C.2, do anexo C.

Relativamente ao restantes modelos marginais, estes foram implementados por intermédio dos comandos que se seguem:

```
> modeloAG <- coxph(Surv(start, stop, status) ~ as.factor(x) + x.1 + x.2 +
    cluster(nid), data = dados1)

> modeloWLW <- coxph(Surv(fulltime, status) ~ as.factor(x) + x.1 + x.2 +
    strata(obs.episode) + cluster(nid), data = dados2)

> modeloLWA <- coxph(Surv(fulltime, status) ~ as.factor(x) + x.1 + x.2 +
    cluster(nid), data = dados2)
```

O modelo AG é um modelo não estratificado que utiliza a formulação por processo de contagem; por esse motivo a única diferença que existe entre o seu comando e o do modelo PWP-CP é a ausência da função strata(obs.episode). No que concerne ao modelo WLW, este é um modelo estratificado mas que utiliza a formulação tempo total, o que significa que o intervalo de risco é definido pela variável fulltime. Além disso, este modelo é aplicado à base

de dados dados2, pelas mesmas razões enunciadas na secção anterior. Por último, o modelo LWA também considera a formulação tempo total, mas ao contrário do modelo WLW, não considera uma função de risco subjacente específica para cada acontecimento.

No geral, pode afirmar-se que os comandos dos modelos marginais são iguais dois a dois, isto é, os comandos dos modelos estratificados têm a mesma estrutura e, por sua vez, os comandos dos modelos não estratificados também apresentam a mesma estrutura. Este aspeto está novamente relacionado com o facto destes modelos exibirem funções de risco semelhantes dois a dois. Mas agora torna-se bem visível nestes comandos de que forma estes modelos diferem entre si.

Os comandos do R que acabaram de ser apresentados, permitem determinar as estimativas globais dos parâmetros de regressão, as quais dizem respeito ao efeito global de cada covariável no tempo de vida dos indivíduos. Todavia, para os modelos estratificados PWP e WLW existe ainda a possibilidade de se obter as estimativas específicas dos parâmetros, o que significa que se pode obter o efeito de cada covariável para cada um dos acontecimentos. Para que essas estimativas específicas sejam determinadas no R, é necessário efetuar uma ligeira alteração na disposição dos argumentos utilizados nos comandos dos modelos PWP e WLW. De modo a exemplificar como é que essa alteração se processa, considere-se o comando correspondente ao modelo PWP-CP:

```
> modeloPWPCPesp <- coxph(Surv(start, stop, status) ~ strata(obs.episode)/
    (as.factor(x) + x.1 + x.2) + cluster(nid), data = dados1, subset = (obs.episode
    <= 5))
```

Na determinação das estimativas específicas dos parâmetros de regressão é preciso ter algum cuidado, uma vez que os modelos PWP e WLW apresentam um conjunto de indivíduos em risco restritivo e semirrestritivo, respetivamente. Por conseguinte, o número de indivíduos em risco diminui à medida que o número de acontecimentos observados aumenta, aspeto este que se revela ainda mais crítico no caso do modelo PWP. Este facto faz com que as estimativas específicas dos últimos estratos sejam pouco fiáveis. Por essa razão, decidiu-se calcular apenas as estimativas específicas referentes aos cinco primeiros estratos (consultar secção C.2) e daí ter-se utilizado o argumento subset = (obs.episode <= 5) no comando anterior.

4.5 Discussão de resultados

Após o ajustamento de qualquer modelo deve-se avaliar a sua adequabilidade e depois interpretar os resultados obtidos. Todavia, como o principal objetivo

deste capítulo consiste em apresentar de que forma os modelos marginais podem ser implementados no R, a análise de resíduos foi remetida para o anexo C, secção C.3.

Na Figura 4.1, é possível visualizar a estimativa de Kaplan-Meier da função de sobrevivência para os vários acontecimentos, bem como as correspondentes funções de risco cumulativas. Em ambos os gráficos é notório que os acontecimentos não apresentam todos o mesmo risco de ocorrer, em especial os três primeiros. Note-se que esta característica já era esperada tendo em conta a forma como os dados foram simulados.

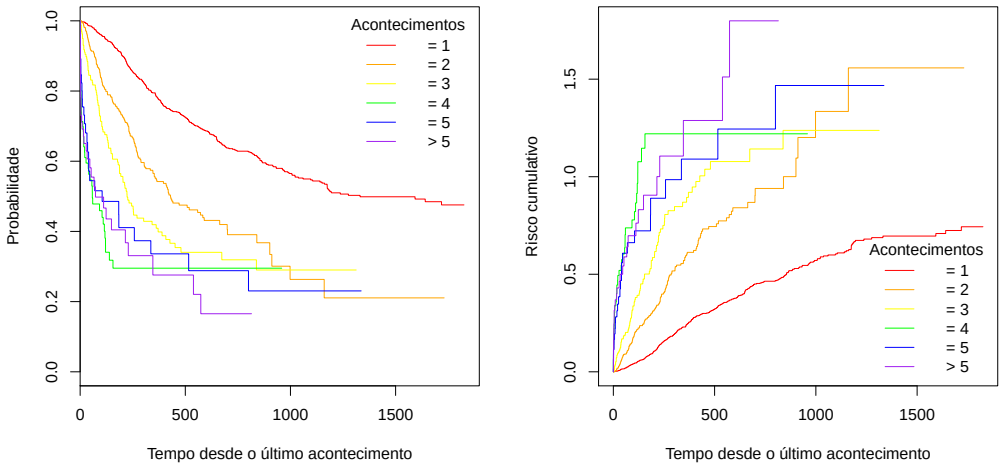


Figura 4.1: Representação das estimativas de Kaplan-Meier da função de sobrevivência e da função de risco cumulativa para cada acontecimento.

Com o intuito de comparar os vários modelos ajustados, compilou-se os resultados obtidos nas Tabelas 4.5 e 4.6. Com base na primeira tabela, constata-se que as covariáveis x e x.2 têm influência significativa para todos os modelos, enquanto a covariável x.1 não a tem para nenhum. Neste ponto, interessa notar que nos modelos PWP-CP e PWP-GT essas estimativas podem ser consideradas semelhantes. Além disso, é possível ainda verificar que nos modelos que apresentam uma estrutura de dependência entre acontecimentos marginal (WLW e LWA), o valor de $\widehat{\beta}_j$ tende a ser sobrestimado.

Posto isto, compare-se as estimativas do erro padrão: usual *versus* robusta. Relativamente ao modelo AG, constata-se que a estimativa usual é ligeiramente inferior à robusta, o que pode ser um alerta para a eventual falta de independência entre os tempos observados de um mesmo indivíduo. Ainda assim, neste caso, este aspeto pode ser desvalorizado não invalidando a aplicação do modelo AG. No que concerne aos modelos WLW e LWA, só

Tabela 4.5: Estimativas globais dos vários parâmetros associados a cada modelo.

Covariável/Modelo	$\widehat{\beta}_j$	$\exp(\widehat{\beta}_j)$	$EP(\widehat{\beta}_j)$	$EP_r(\widehat{\beta}_j)$	Valor-p
x					
Cox clássico	0.465	1.592	0.106	0.104	8.12e-06
PWP-CP	0.528	1.696	0.081	0.086	8.98e-10
PWP-GT	0.567	1.763	0.079	0.082	4.47e-12
AG	0.671	1.956	0.078	0.100	2.10e-11
WLW	0.815	2.259	0.078	0.125	7.07e-11
LWA	0.707	2.028	0.078	0.107	3.91e-11
x.1					
Cox clássico	0.012	1.012	0.182	0.177	0.947
PWP-CP	0.087	1.091	0.134	0.147	0.553
PWP-GT	0.172	1.188	0.129	0.139	0.215
AG	0.179	1.196	0.129	0.178	0.316
WLW	0.188	1.206	0.129	0.213	0.379
LWA	0.182	1.199	0.128	0.187	0.330
x.2					
Cox clássico	1.064	2.898	0.067	0.064	< 2e-16
PWP-CP	0.676	1.966	0.048	0.050	< 2e-16
PWP-GT	0.672	1.959	0.045	0.044	< 2e-16
AG	0.874	2.396	0.041	0.047	< 2e-16
WLW	1.190	3.287	0.047	0.071	< 2e-16
LWA	0.942	2.566	0.042	0.054	< 2e-16

é necessário considerar a estimativa robusta, a qual se revelou ser superior à usual, tal como se esperava. Nos restantes modelos (Cox clássico, PWP-CP e PWP-GT) as estimativas são da mesma ordem de grandeza, pelo que não se verifica correlação, tanto entre indivíduos como intraindivíduos.

Posteriormente à análise das estimativas globais dos parâmetros associados a cada modelo, importa efetuar alguns comentários acerca das estimativas específicas dos parâmetros de regressão (Tabela 4.6). Em primeiro lugar, note-se que para o primeiro acontecimento as estimativas específicas dos parâmetros são iguais, seja qual for o modelo (PWP-CP, PWP-GT e WLW), que por sua vez são ainda iguais às estimativas do modelo de Cox clássico apresentadas na Tabela 4.5. Assim sendo, para estudar o tempo até ao primeiro acontecimento, pode ser usado qualquer um destes modelos. Em relação ao efeito das covariáveis, com base na secção C.2, conclui-se que a covariável `x.1` foi a única que não se revelou significativa para nenhum dos

Tabela 4.6: Estimativas específicas dos parâmetros de regressão para cada modelo.

Covariável/Modelo	Estimativas específicas				
	estrato 1	estrato 2	estrato 3	estrato 4	estrato 5
x					
PWP-CP	0.465	0.675	0.848	-0.219	-0.625
PWP-GT	0.465	0.697	0.938	0.041	-0.429
WLW	0.465	1.003	1.529	1.252	0.972
x.1					
PWP-CP	0.012	0.325	-0.180	-0.218	-0.345
PWP-GT	0.012	0.265	0.078	-0.009	0.367
WLW	0.012	0.268	0.308	0.148	0.266
x.2					
PWP-CP	1.064	0.204	0.456	0.061	0.210
PWP-GT	1.064	0.225	0.516	0.147	0.286
WLW	1.064	1.076	1.386	1.368	1.586

acontecimentos. Nos dois modelos PWP, as covariáveis `x` e `x.2` apresentaram efeito significativo apenas para os três primeiros acontecimentos, enquanto no modelo WLW o efeito de cada uma destas foi sempre significativo.

Capítulo 5

Conclusão

5.1 Comentários sobre os modelos marginais

A presente dissertação teve como principal propósito dar a conhecer as principais abordagens, que se encontram disponíveis na literatura, para a modelação de acontecimentos múltiplos. Após transmitir uma visão geral sobre o assunto, direcionou-se a atenção para o estudo das extensões do modelo de Cox. Neste contexto, descreveu-se de forma minuciosa os quatro modelos mais utilizados na prática: o modelo PWP; o modelo AG; o modelo WLW; e o modelo LWA. Todos estes modelos são classificados como modelos marginais, dada a forma como se processa a estimação dos parâmetros de regressão. Além do mais, por intermédio de dados simulados, mostrou-se que a aplicação destes modelos pode ser facilmente conseguida recorrendo ao *software* estatístico R, onde se evidenciou a importância de construir uma base de dados adequada às especificidades de cada modelo. Todavia, existem ainda alguns aspetos a respeito destes quatro modelos que merecem ser discutidos.

Antes de tudo, importa referir que a análise de acontecimentos múltiplos não tem que ser sempre melhor do que a análise simples, onde se analisa apenas o tempo até à ocorrência do primeiro acontecimento. De facto, embora nesta última situação haja o risco de haver informação que não esteja a ser aproveitada, pode acontecer que esta seja irrelevante e, assim sendo, a sua inclusão no estudo não é necessária. Deste modo, antes da aplicação dos modelos marginais, deve-se começar por efetuar uma análise simples recorrendo ao modelo de Cox clássico. A seleção do modelo final será feita avaliando o ajustamento de cada um deles, optando sempre pelo modelo mais simples e parcimonioso possível.

Na secção 3.3, referiu-se que Kelly e Lim [45] sugeriram quatro componentes chave para diferenciar os modelos marginais: i) intervalo de risco; ii) função de risco subjacente; iii) conjunto de indivíduos em risco; e iv) estrutura de dependência entre acontecimentos. Cada uma destas componentes foi cuidadosamente estudada, de onde se chegou à conclusão de que existe uma certa ligação entre as mesmas. De acordo com estes autores, apenas duas das quatro componentes são essenciais para identificar a extensão do modelo de Cox mais adequada ao caso em estudo, nomeadamente o intervalo de risco e o conjunto de indivíduos em risco. De facto, tendo por base a Tabela B.1 em anexo, quando o intervalo de risco é definido segundo a formulação por processo de contagem ou tempo por intervalos, define-se indiretamente que a estrutura de dependência entre acontecimentos segue uma abordagem condicional, enquanto que se o intervalo de risco é definido segundo a formulação tempo total, define-se que a estrutura de dependência segue uma abordagem marginal. Analogamente, quando se define que o conjunto de indivíduos em risco é restritivo ou semirrestritivo, de certo modo está a ser assumido que os vários acontecimentos apresentam riscos de ocorrência diferentes, pelo que o modelo terá forçosamente de ter uma função de risco subjacente específica para cada acontecimento. Por outro lado, quando o conjunto de risco é não restritivo, deixa de fazer sentido considerar uma função de risco subjacente diferente para cada acontecimento. Assim sendo, basta definir o intervalo de risco e o conjunto de indivíduos em risco para que as outras duas componentes fiquem implicitamente determinadas.

Ao aplicar um determinado modelo marginal é preciso ter o cuidado de verificar qual a formulação do intervalo de risco que este considera e, em particular, qual a escala de tempo associada a essa formulação. Quando um modelo utiliza a formulação tempo total ou processo de contagem, considera que a escala de tempo se refere ao tempo desde o início do estudo. Apesar destas duas formulações terem a mesma escala de tempo, recorde-se que a formulação por processo de contagem tem a vantagem de reconhecer que podem haver certos períodos de tempo em que um indivíduo deixa de estar em risco de sofrer algum acontecimento (o que acontece quando os acontecimentos são não instantâneos). Por outro lado, quando um modelo utiliza a formulação do tempo por intervalos, a escala de tempo que lhe está associada diz respeito ao tempo desde a observação do último acontecimento, pois nesta formulação o relógio reinicia a sua contagem sempre que algum acontecimento é observado. Então, pode dizer-se que existem duas escalas de tempo e, obviamente, cada uma delas terá implicações na interpretação de resultados, principalmente quando se examina o efeito de cada covariável no tempo de vida dos indivíduos.

No decorrer dos últimos anos, vários autores realizaram estudos para comparar os quatro modelos marginais, na tentativa de perceberem em que situações a aplicação de cada um deles é mais apropriada [45, 50, 63, 72, 81, 84]. Essa comparação tem sido feita não só por intermédio de dados reais como também de dados simulados. Em todos esses estudos, os modelos marginais apresentaram resultados distintos para o mesmo conjunto de dados, o que não é uma descoberta inesperada, dado que cada um deles foi desenvolvido para tratar diferentes questões de investigação. Ainda assim, interessa efetuar uma análise comparativa entre os mesmos, de modo a enfatizar alguns dos pontos fortes e pontos fracos que cada modelo apresenta.

Como referido anteriormente, o modelo WLW foi sugerido como uma alternativa ao modelo PWP, com o intuito de ultrapassar algumas das suas limitações. Em ambos os modelos os indivíduos estão estratificados por acontecimento, pelo que possibilitam que se analise a forma como o efeito de cada covariável se altera ao longo dos mesmos. Este é um aspeto importante, dado que se o efeito de uma covariável não é constante ao longo do tempo, então tem mais interesse analisar a sua evolução. Uma das principais vantagens do modelo WLW é o facto de o seu conjunto de indivíduos em risco ter sempre, para todos os estratos, uma dimensão superior à do modelo PWP. Neste último, a dimensão do conjunto de risco diminui de cada vez que se passa ao estrato seguinte, levando a que as estimativas dos parâmetros de regressão sejam pouco precisas, ou mesmo pouco fiáveis, nos últimos estratos. Outra diferença marcante entre estes dois modelos, diz respeito à estrutura de dependência entre os acontecimentos. No modelo PWP a interpretação dos parâmetros de regressão é muitas vezes dificultada pela natureza condicional dos acontecimentos, ao contrário do que acontece no modelo WLW, onde não existe qualquer estrutura de dependência específica.

Não obstante as vantagens que o modelo WLW apresenta, a sua aplicação a acontecimentos recorrentes tem sido alvo de muitas críticas por parte de alguns investigadores [16, 21, 56]. Em primeiro lugar, é preciso ter presente que os dados relativos a acontecimentos recorrentes têm duas características que estão intimamente interligadas: i) os indivíduos só podem estar em risco para um acontecimento de cada vez; e ii) os acontecimentos ocorrem de forma ordenada. Ora, no modelo WLW, o conjunto de indivíduos em risco é semirrestritivo, o que significa que este modelo consente que um indivíduo seja incluído no conjunto de risco relativo ao s-ésimo acontecimento sem que ainda tenha experienciado o acontecimento de ordem $s-1$. Este aspeto revela-se um inconveniente para a análise de acontecimentos recorrentes, pois não permite acomodar a natureza ordenada desde tipo de dados. Uma outra crítica a respeito da aplicação deste modelo a acontecimentos recorrentes, refere-se à sua incapacidade em modelar a relação existente entre os aconte-

cimentos observados de um mesmo indivíduo, uma vez que a sua estrutura de dependência entre acontecimentos segue uma abordagem marginal. Assim, este modelo não é tão eficiente, quanto o modelo PWP, em estimar o efeito produzido pelas covariáveis. Por essa razão, Lin [51] defende que quando se está interessado em estudar o efeito de uma determinada covariável na sobrevivência dos indivíduos (por exemplo, o efeito de um tratamento) deve-se aplicar os dois modelos e analisá-los separadamente, de modo a ter uma imagem mais completa sobre o mesmo.

Além do mais, existe ainda outra crítica que tem sido apontada não só ao modelo WLW, como também a todos os modelos que tenham um conjunto de risco não restritivo ou semirrestritivo. Quando se recorre a um modelo que utiliza alguma das duas definições anteriores é preciso ter o cuidado de verificar qual a formulação do intervalo de risco que este considera. A combinação de um intervalo de risco definido segundo a formulação tempo total com um conjunto de risco não restritivo ou semirrestritivo, leva a que usualmente ocorra o designado efeito de arrastamento (*carry-over effect*) [21, 45]. Nesta formulação, os intervalos de risco de um mesmo indivíduo tendem a ser correlacionados, mesmo quando a análise baseada no tempo por intervalos não o é. Recorde-se que quando se ilustrou a formulação tempo total na secção 3.3.1 (consultar Figura 3.3(b)), observou-se que os intervalos de risco de um mesmo indivíduo encontravam-se sobrepostos. O que acontece é que o segundo intervalo de risco de um mesmo indivíduo inclui o seu primeiro intervalo de risco, o terceiro intervalo de risco inclui o primeiro e o segundo intervalos de risco, e assim por diante. Como exemplo, suponha-se que o efeito de uma certa covariável apenas é significativo na ocorrência do primeiro acontecimento e não nos restantes. A análise baseada nos tempos totais pode enganosamente arrastar esse efeito para os acontecimentos seguintes, fazendo com que este seja exageradamente maior nos últimos estratos.

Por conseguinte, pode dizer-se que o modelo PWP é um modelo adequado para analisar acontecimentos recorrentes, pois neste tipo de dados é fundamental que o conjunto de indivíduos em risco seja restritivo, principalmente quando o risco de ocorrência se altera após a observação de cada acontecimento. Só assim é que os parâmetros de regressão específicos de cada acontecimento são estimados corretamente. Além disso, um conjunto de risco restritivo impede que ocorra o efeito de arrastamento, pois para a análise do s-ésimo acontecimento apenas estão em risco os indivíduos que sofreram o acontecimento de ordem $s-1$. Os indivíduos que respeitarem essa condição contribuem com um único intervalo de risco, ao contrário do que acontece nas outras duas definições, em que um mesmo indivíduo pode contribuir com vários intervalos de risco. Note-se ainda que neste modelo

a formulação do intervalo de risco pode ser feita por processo de contagem (PWP-CP) ou tempo por intervalos (PWP-GT). Segundo Kelly e Lim [45], as estimativas específicas obtidas pelos modelos PWP-CP e PWP-GT para o efeito do tratamento, revelaram resultados semelhantes com um viés desprezável. Ainda assim, deve haver um esforço em utilizar a formulação que melhor represente a situação em estudo, pois só assim os resultados refletirão a realidade.

Embora o modelo PWP seja totalmente livre do efeito de arrastamento, a utilização de um conjunto de risco restritivo também tem os seus inconvenientes. Como os indivíduos que não sofreram exatamente $s-1$ acontecimentos são excluídos da análise referente ao s-ésimo acontecimento, o conjunto de risco irá progressivamente diminuir ao longo do estudo, sendo cada vez menos heterogéneo. Consequentemente, o número de indivíduos em risco nos últimos estratos tenderá a ser drasticamente pequeno, levando a que as estimativas específicas dos parâmetros de regressão sejam pouco fiáveis, como já havia sido referido. Para além da perda de heterogeneidade, a escolha dos indivíduos que vão estar em risco para um dado acontecimento não ocorre de forma aleatória, uma vez que é determinada pela observação do acontecimento anterior. Este facto leva à violação do pressuposto MCAR. Therneau e Grambsch [79] apresentaram duas possibilidades para tentar solucionar estes problemas. A primeira consiste em truncar os dados exatamente no acontecimento (estrato) em que se considera que o número de indivíduos em risco é demasiado pequeno, descartando a informação proveniente dos estratos seguintes. A segunda possibilidade consiste em aglomerar os últimos estratos num único, a partir daquele que se considera ter um número pequeno de indivíduos. Esta última possibilidade é considerada a mais apelativa, pois tem a vantagem de não desperdiçar informação que pode vir a revelar-se crucial para a análise.

Numa outra perspetiva, se a partir de um certo acontecimento as funções de risco forem semelhantes deixa de ter fundamento considerá-las distintas, o que uma vez mais conduz à junção dos últimos estratos. Note-se que, inicialmente, o modelo PWP tem um conjunto de risco bastante heterogéneo, pelo que as diferenças existentes entre as funções de risco dos vários indivíduos deve-se, em especial, ao efeito das diversas covariáveis com valores bastante distintos para cada um deles. Admita-se que o conjunto de risco relativo ao segundo acontecimento (que contém apenas os indivíduos que sofreram o primeiro acontecimento) deixa de conter os indivíduos que pertencem a uma certa categoria de uma covariável. Isto significa que, para além dos indivíduos serem menos heterogéneos, essa covariável deixa de ser importante para o modelo e o seu efeito na sobrevivência passa a estar incorporado na função de risco subjacente, que assim terá que ser necessariamente diferente

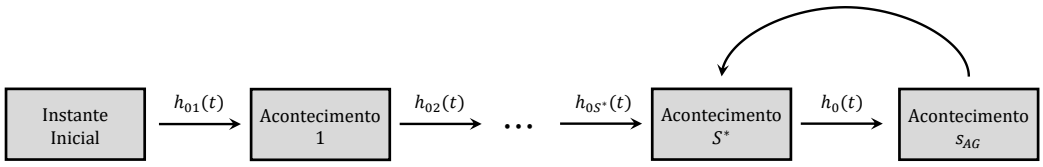


Figura 5.1: Representação esquemática do modelo híbrido PWP-AG.

da função de risco subjacente correspondente ao primeiro acontecimento. Suponha-se agora que a partir de um certo acontecimento, designe-se por S^*, os indivíduos que se encontram em risco serão mais homogéneos, de tal forma que pode não ser necessário considerar uma função de risco subjacente específica para cada acontecimento. Então, em vez de continuar a aplicar o modelo PWP, a partir do acontecimento S^* passa-se a aplicar o modelo AG. Para o efeito, aglomera-se os últimos estratos, considerando uma função de risco subjacente comum aos restantes acontecimentos. Esta formulação dá origem ao modelo híbrido PWP-AG, recentemente apresentado por Sousa-Ferreira e Abreu [76], cujo esquema se encontra na Figura 5.1. No geral, serão analisados $s_{PWP} = 1, 2, \ldots, S^*$ acontecimentos com o modelo PWP e $s_{AG} = S^* + 1, S^* + 2, \ldots, S$ acontecimentos com o modelo AG. É de salientar que esta possibilidade permite minimizar a violação do pressuposto MCAR, dado que a partir do acontecimento S^* passa-se a considerar que o conjunto de indivíduos em risco é não restritivo.

Os conjuntos de risco não restritivos são adequados apenas quando o risco de ocorrência não se altera à medida que os acontecimentos vão sendo observados, ou seja, quando não é necessário considerar uma função de risco subjacente específica para cada acontecimento, tal como acontece no modelo AG. Neste modelo, é importante verificar a validade do pressuposto de independência entre os tempos de vida de um mesmo indivíduo, de modo a garantir que este seja um modelo de riscos proporcionais. Quando este pressuposto é violado, em alternativa, deve-se recorrer ao modelo PWP ou a um modelo de efeitos aleatórios.

O modelo LWA, tal como o modelo AG, apresenta um conjunto de indivíduos em risco não restritivo, que por sua vez é apropriado ao facto de também considerar uma função de risco subjacente comum a todos os acontecimentos. Contudo, este aspeto não abona a favor do modelo LWA, uma vez que a combinação de um conjunto de risco não restritivo com um intervalo de risco definido segundo a formulação tempo total faz com que o efeito do tratamento seja subestimado, conforme referido por Kelly e Lim [45]. Recorde-se que este modelo pode ser visto como um derivado do modelo WLW, tendo sido sugerido para o caso em que se pretende analisar acontecimentos agru-

pados. Na literatura, existem poucos estudos a comparar o modelo LWA aos outros três modelos marginais, pois há quem defenda que este modelo não é apropriado para a análise de acontecimentos recorrentes [50]. De facto, apesar deste modelo ter sido sugerido para analisar acontecimentos múltiplos da mesma natureza, cada um deles é observado uma única vez, ou seja, não existe repetição do mesmo acontecimento. Por exemplo, no caso em que se está a testar o efeito de um determinado medicamento em retardar a perda total das funções renais, os tempos de vida estão agrupados por indivíduo porque possui dois rins. Porém, a partir do momento em que se observa o acontecimento de interesse, este não tem a possibilidade de se repetir, o que significa que um mesmo rim gera apenas um único acontecimento.

Apesar de ainda não existir um modelo adequado para todas as situações, de acordo com os estudos empíricos efetuados por Villegas *et al.* [84], existe uma tendência para que os modelos PWP e AG sejam preferíveis ao modelo WLW quando os níveis de correlação são baixos, sendo que essa situação reverte-se à medida que o nível de correlação aumenta. Conforme discutido por Lin [51], quando se pretende analisar a taxa global de acontecimentos da mesma natureza, o modelo mais simples de implementar é o modelo AG (englobando as covariáveis dependentes do tempo para captar a dependência entre acontecimentos), principalmente quando o número de acontecimentos múltiplos observados é baixo. Por outro lado, quando se pretende analisar o tempo entre acontecimentos deve-se recorrer ao modelo PWP, enquanto que se o interesse for analisar o tempo desde o instante em que o indivíduo entra em estudo (isto é, o tempo total), o modelo mais apropriado é o modelo WLW.

Resumidamente, a escolha do melhor modelo a aplicar é muitas vezes determinada pelos objectivos do investigador, que por sua vez devem ser coerentes com as especificidades que o conjunto de dados apresenta. Conclui-se então que não é possível afirmar que existe um modelo ótimo para todas as situações, pois cada um deles possui características particulares que fazem com que a sua aplicação seja mais direcionada para uma determinada situação. Com o intuito de facilitar/orientar essa escolha, pode-se e deve-se recorrer às quatro componentes chave analisadas anteriormente. Além disso, é preciso ter consciência de que o modo como se interpreta o caso em estudo também tem influência na escolha do modelo a ser aplicado e que muitas vezes um mesmo caso pode originar várias interpretações. Este facto permite que o problema seja encarado de diversas formas, o que possibilita que a análise de acontecimentos múltiplos seja feita recorrendo a mais do que um modelo marginal, nunca perdendo de vista as especificidades que fazem com que cada um deles seja único, uma vez que estas irão interferir diretamente na análise de resultados.

5.2 Perspetivas futuras

Futuramente, pretende-se estudar com mais detalhe o modelo híbrido PWP-AG [76] recém apresentado, com o intuito de avaliar o seu desempenho na análise de acontecimentos recorrentes.

Ao longo da realização deste trabalho, sentiu-se que a análise dos resíduos nos modelos com estratificação não está devidamente desenvolvida para a análise de acontecimentos múltiplos, pois a análise que se costuma efetuar permite estudar apenas os resíduos associados às estimativas globais.

Um tema que não foi abordado e que de acordo com alguns autores [13, 42, 83] é uma boa aposta, consiste em considerar um modelo com fragilidade, ou seja, um modelo condicional. Desta forma será possível ter em conta a heterogeneidade não observável através da variável aleatória que representa a fragilidade.

Por último, uma área ainda pouco explorada é a inclusão da fração de cura nos modelos para acontecimentos múltiplos [87]. De facto, podem existir situações em que, para alguns indivíduos, não se irá observar qualquer novo acontecimento.

Anexos

Anexo A

Processos de contagem

A.1 Contextualização

Em meados da década de setenta do século XX, o estatístico norueguês Odd Aalen introduziu pela primeira vez os processos de contagem e a teoria das martingalas na Análise de Sobrevivência [55]. Desde então, têm sido verificados desenvolvimentos relevantes neste ramo da estatística, onde importa salientar a notória intervenção dos processos de contagem na análise de resíduos, primeiro estudados por Barlow e Prentice [10] e, alguns anos mais tarde, por Therneau *et al.* [80]. O recurso à notação dos processos de contagem permitiu, essencialmente, solidificar a fundamentação teórica de diversos estimadores que haviam surgido na altura, como é o caso do estimador dos parâmetros do modelo de regressão de Cox [27, 28], cuja implementação foi bem fundamentada e estruturada por Andersen e Gill [8], em 1982. Em termos gerais, um processo de contagem é definido como sendo um processo estocástico que conta o número de acontecimentos de um certo tipo no intervalo de tempo $[0, t]$.

Antes do surgimento dos processos de contagem, era preciso assumir que todos os indivíduos entravam em estudo no mesmo instante zero, pelo que não era possível englobar na análise que um indivíduo tinha tido uma entrada atrasada no estudo. Para além disso, não havia forma de registar que um indivíduo pudesse sair temporariamente do estudo, voltando a fazer parte do mesmo algum tempo mais tarde. Assim, os processos de contagem têm vindo a revelar-se extremamente úteis na resolução de algumas limitações que são comuns na análise de dados mais complexos como, por exemplo, a análise de acontecimentos múltiplos em dados longitudinais. Um dos maiores obstáculos da análise deste tipo de dados é o facto de um indivíduo se manter em risco após a observação de cada acontecimento de interesse

(exceto quando esse acontecimento impede que os seguintes venham a ser observados). Porém, também esta limitação pode ser facilmente ultrapassada recorrendo aos processos de contagem. De seguida, serão introduzidos alguns conceitos básicos para ajudar a perceber melhor como se implementa esta formulação.

Em primeiro lugar, considere-se o caso mais simples, em que se pretende analisar o tempo até à ocorrência de um único acontecimento de interesse, onde os indivíduos estão sujeitos a um mecanismo de censura à direita. Até então, para uma amostra aleatória de dimensão n, considerou-se que o tempo observado do indivíduo i é uma observação da variável aleatória $T_i = \min\{X_i, C_i\}$, onde X_i e C_i denotam o seu verdadeiro tempo de vida e o seu tempo de censura, respetivamente. Assim, representou-se a variável de censura por $\delta_i = I(X_i \leq C_i)$, que toma os valores 1 quando ocorre o acontecimento de interesse e 0 quando ocorre a censura. Pode-se então dizer que, para cada indivíduo i, o que se observa na verdade é o par (T_i, δ_i), $i = 1, \ldots, n$. No caso mais simples, a formulação por processo de contagem consiste em substituir esse par de variáveis pelo par de funções $(N_i(t), Y_i(t))$, $t \geq 0$, onde

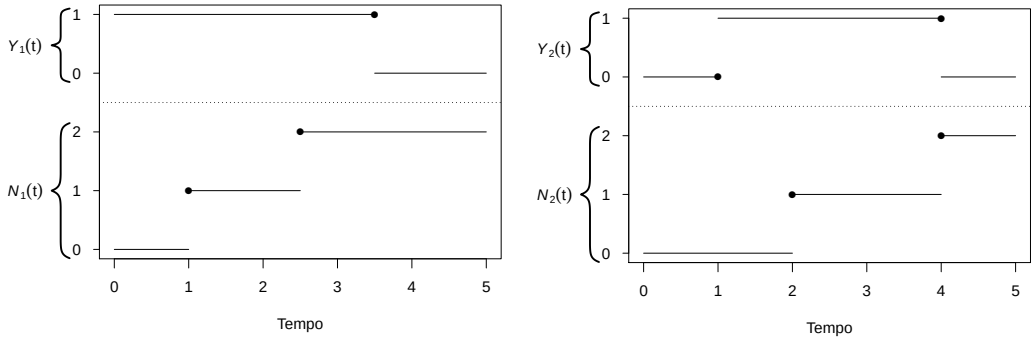
- $N_i(t) = I(T_i \leq t, \delta_i = 1)$ – é um processo de contagem que (neste caso) toma apenas os valores 0 ou 1. Quando o instante t é superior ou igual ao tempo observado do i-ésimo indivíduo e o acontecimento de interesse já tiver sido observado, tem-se que $N_i(t) = 1$. Caso contrário, tem-se que $N_i(t) = 0$;

- $Y_i(t) = I(T_i \geq t)$ – é um processo que, para um mesmo indivíduo, toma sempre os valores 0 ou 1. Quando o i-ésimo indivíduo está em risco imediatamente antes do instante t, isto é, ainda não foi observado nem o acontecimento de interesse nem a censura, tem-se que $Y_i(t) = 1$. Caso contrário, tem-se que $Y_i(t) = 0$.

Esta notação pode ser diretamente generalizada ao contexto de acontecimentos múltiplos. Nesse caso, o par de funções $(N_i(t), Y_i(t))$ passa a ser definido por

- $N_i(t)$ – representa o número de acontecimentos observados para o i-ésimo indivíduo no intervalo de tempo $[0, t]$;

- $Y_i(t)$ – indica se o i-ésimo indivíduo está sob observação e pertence ao conjunto de risco no instante t.



(a) Indivíduo 1. (b) Indivíduo 2.

Figura A.1: Ilustração dos processos $N_i(t)$ e $Y_i(t)$ associados a cada indivíduo i.

De forma a clarificar como funcionam os processos de contagem quando aplicados a acontecimentos múltiplos, considere-se os dois exemplos presentes na Figura A.1 (adaptados de Therneau e Grambsch [79]), cujos gráficos foram obtidos através do *software* estatístico R. Na Figura A.1(a), observa-se que o indivíduo 1 começa a ser seguido a partir do instante 0 e sofre dois acontecimentos de interesse, um no instante 1 e outro no instante 2.5, acabando por abandonar o estudo no instante 3.5. Já na Figura A.1(b), observa-se que o indivíduo 2 apenas é considerado em risco de sofrer algum acontecimento a partir instante 1, sofrendo acontecimentos nos instantes 2 e 4. Repare-se que, neste exemplo, o indivíduo 2 abandona o estudo exatamente no instante em que sofre o segundo acontecimento.

Um processo de contagem é um processo estocástico que se inicia em 0, cujas concretizações correspondem a uma função em escada, contínua à direita e com saltos de altura 1. Pela Figura A.1 observa-se que o processo $N_i(t)$ é um processo de contagem, dado que é contínuo à direita, o que significa que o seu valor é atualizado exatamente no instante em que ocorre o acontecimento de interesse. Como $Y_i(t)$ é um processo contínuo à esquerda, este não é um processo de contagem. Porém, $Y_i(t)$ é um exemplo de um processo previsível[1] (*predictable process*), uma vez que o seu valor é conhecido imediatamente antes do acontecimento ser observado. Basicamente, o que acontece é que um indivíduo só pode sofrer algum acontecimento no instante t se estiver em risco imediatamente antes desse instante, isto é, em t^-.

Como já foi referido, a notação dos processos de contagem pode ser utilizada tanto no caso em que são observados acontecimentos múltiplos por indivíduo, como no caso em que apenas se pretende analisar o tempo até à ocorrência do primeiro acontecimento. Assim sendo, os estimadores não

[1]Importa referir que o facto de $Y_i(t)$ ser contínuo à esquerda é uma condição suficiente (mas não necessária) para que seja considerado um processo previsível.

paramétricos e o modelo de regressão de Cox estudados nos dois primeiros capítulos, também podem ser formalizados através da notação dos processos de contagem. Para tal, é necessário introduzir alguma notação suplementar acerca dos processos $N_i(t)$ e $Y_i(t)$, como seja:

- $\overline{N}(t) = \sum_{i=1}^{n} N_i(t)$ – representa o número total de acontecimentos observados até ao instante t;
- $\overline{Y}(t) = \sum_{i=1}^{n} Y_i(t)$ – expressa o número de indivíduos em risco no instante t, mais concretamente no intervalo infinitesimal $(t - dt; t]$;
- $\Delta\overline{N}(t) = \overline{N}(t) - \overline{N}(t^-)$ – representa o número de acontecimentos observados precisamente no instante t.

A.2 Estimadores não paramétricos

Por conseguinte, pode-se reformular os estimadores não paramétricos estudados na secção 1.7, de modo a que estes sejam escritos com base na notação dos processos de contagem:

Estimador de Kaplan-Meier $$\widehat{S}_{KM}(t) = \prod_{k:t_{(k)} \leq t} \left(1 - \frac{\Delta\overline{N}\big(t_{(k)}\big)}{\overline{Y}\big(t_{(k)}\big)}\right);$$

Estimador de Nelson-Aalen $$\widehat{H}_{NA}(t) = \sum_{k:t_{(k)} \leq t} \frac{\Delta\overline{N}\big(t_{(k)}\big)}{\overline{Y}\big(t_{(k)}\big)};$$

Estimador de Breslow $$\widehat{S}_{B}(t) = \prod_{k:t_{(k)} \leq t} \exp\left(-\frac{\Delta\overline{N}\big(t_{(k)}\big)}{\overline{Y}\big(t_{(k)}\big)}\right).$$

A.3 Modelo de regressão de Cox

A estimação dos parâmetros do modelo de Cox [27] pode ser feita com recurso à notação dos processos de contagem da forma que se segue. Considere-se uma amostra aleatória de dimensão n, em que para cada indivíduo i se regista um vetor de p covariáveis possivelmente dependentes do tempo, denotado por $\boldsymbol{z}_i(t) = \big(z_{i1}(t), z_{i2}(t), \ldots, z_{ip}(t)\big)'$, a função de risco associada ao modelo de regressão de Cox é dada por $h\big(t; \boldsymbol{z}_i(t)\big) = h_0(t) \exp\big(\boldsymbol{\beta}'\boldsymbol{z}_i(t)\big)$, $i = 1, \ldots, n$.

A função de verosimilhança parcial passa a ser formulada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{t \geq 0} \left[\frac{Y_i(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_i(t)\right)}{\sum_{j=1}^{n} Y_j(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_j(t)\right)} \right]^{\Delta \overline{N}_i(t)},$$

pelo que a respetiva função logverosimilhança parcial é dada por

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t \geq 0} \Delta \overline{N}_i(t) \left[\log\left(Y_i(t)\right) + \boldsymbol{\beta}' \boldsymbol{z}_i(t) - \log\left(\boldsymbol{Q}^{(0)}(\boldsymbol{\beta}, t)\right) \right],$$

onde $\boldsymbol{Q}^{(0)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_j(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_j(t)\right)$. Consequentemente, pelo mesmo processo efetuado na secção 2.3.1, obtém-se a função *score*

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t \geq 0} \Delta \overline{N}_i(t) \left[\boldsymbol{z}_i(t) - \frac{\boldsymbol{Q}^{(1)}(\boldsymbol{\beta}, t)}{\boldsymbol{Q}^{(0)}(\boldsymbol{\beta}, t)} \right],$$

em que $\boldsymbol{Q}^{(1)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_j(t) \boldsymbol{z}_j(t) \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_j(t)\right)$. Desta forma, o estimador de máxima verosimilhança parcial $\widehat{\boldsymbol{\beta}}$ é obtido através da resolução do sistema de equações $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$. Além do mais, pode-se ainda obter a matriz de informação de Fisher, sendo esta dada por

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t \geq 0} \Delta \overline{N}_i(t) \left[\frac{\boldsymbol{Q}^{(2)}(\boldsymbol{\beta}, t)}{\boldsymbol{Q}^{(0)}(\boldsymbol{\beta}, t)} - \frac{\boldsymbol{Q}^{(1)}(\boldsymbol{\beta}, t) \boldsymbol{Q}^{(1)}(\boldsymbol{\beta}, t)'}{\boldsymbol{Q}^{(0)}(\boldsymbol{\beta}, t)^2} \right],$$

onde $\boldsymbol{Q}^{(2)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_j(t) \boldsymbol{z}_j(t) \boldsymbol{z}_j(t)' \exp\left(\boldsymbol{\beta}' \boldsymbol{z}_j(t)\right)$.

Anexo B

Características dos modelos marginais

Tabela B.1: Caracterização das extensões do modelo de Cox, desenvolvidas para o estudo de acontecimentos múltiplos.

Componentes chave	**Modelos Marginais**			
	PWP	AG	WLW	LWA
Intervalo de risco*	CP ou GT	CP	TT	TT
Função de risco subjacente	específica	comum	específica	comum
Conjunto de indivíduos em risco	restritivo	não restritivo	semirrestritivo	não restritivo
Estrutura de dependência entre acontecimentos	condicional	condicional	marginal	marginal

* Tempo total (*total time* – TT), tempo por intervalos (*gap time* – GT) e processo de contagem (*counting process* – CP).

Anexo C

Anexos do capítulo 4

C.1 Procedimento para a simulação de dados

O conjunto de dados utilizado neste trabalho foi simulado através do *package* survsim [58] do *software* estatístico R, onde explorou-se unicamente a simulação de acontecimentos recorrentes. Após a instalação[1] deste *package* e a ativação[2] das suas funcionalidades, realizou-se a simulação de dados pelo seguinte procedimento:

```
> set.seed(500)
> dist.ev <- c("weibull", "weibull", "weibull", "weibull", "weibull")
> anc.ev <- c(1.5, 1.2, 0.8, 0.5, 0.4)
> beta0.ev <- c(7.2, 6.5, 6.7, 6.4, 6.4)
> dist.cens <- c("weibull", "weibull", "weibull", "weibull", "weibull")
> anc.cens <- c(1.5, 1.1, 0.9, 0.5, 0.5)
> beta0.cens <- c(7.2, 6.6, 6.7, 6.4, 6.4)
> dados1 <- rec.ev.sim(n=1000, foltime=1825, dist.ev, anc.ev, beta0.ev,
    dist.cens, anc.cens, beta0.cens, x=list(c("bern", 0.5), c("unif", 0, 1),
    c("normal", 0, 1)), beta=list(c(-0.4, -0.5, -0.6, -0.7, -0.8), c(-0.07, -0.02,
    -0.06, -0.06, -0.06), c(-0.7, -0.2, -0.6, -0.6, -0.6)))
```

Embora seja possível simular dados com as características pretendidas, é preciso ter presente que estes são gerados de forma aleatória, o que significa que para as mesmas características pode-se obter uma variedade de conjuntos de dados diferentes. Assim, surgiu a necessidade de fixar uma semente para que no futuro seja possível reproduzir o mesmo conjunto de dados, razão pela

[1] install.packages("survsim")
[2] library(survsim)

qual se utilizou a função set.seed. De seguida, serão descritos os argumentos que foram utilizados na função rec.ev.sim, os quais definem as características dos dados simulados, nomeadamente:

- n – dimensão do conjunto de dados simulados;
- foltime – tempo máximo durante o qual os indivíduos se encontram em estudo;
- dist.ev – vetor de tamanho arbitrário onde se define a distribuição do tempo até à ocorrência do acontecimento de interesse. Existe a possibilidade de utilizar três distribuições diferentes: distribuição de Weibull ("weibull"), distribuição loglogística ("llogistic") e distribuição lognormal ("lnorm");
- anc.ev – vetor de tamanho arbitrário correspondente às componentes reais que contêm os parâmetros ancilares associados a cada uma das distribuições do tempo de vida. Quando na distribuição de Weibull se define anc.ev=1, obtém-se o caso particular correspondente à distribuição exponencial;
- beta0.ev – vetor de tamanho arbitrário correspondente às componentes reais que contêm os parâmetros beta0 associados a cada uma das distribuições do tempo de vida;
- dist.cens – vetor de tamanho arbitrário onde se define a distribuição do tempo até à ocorrência de censura à direita, sendo esta não informativa. Aqui é possível utilizar as mesmas distribuições enunciadas no argumento dist.ev;
- anc.cens – vetor de tamanho arbitrário correspondente às componentes reais que contêm os parâmetros ancilares associados a cada uma das distribuições do tempo de censura;
- beta0.cens – vetor de tamanho arbitrário correspondente às componentes reais que contêm os parâmetros beta0 associados a cada uma das distribuições do tempo de censura;
- x – lista de vetores que indicam a distribuição das covariáveis simuladas e os seus respetivos parâmetros. Existe a possibilidade de simular três distribuições diferentes: distribuição de Bernoulli ("bern"), distribuição uniforme ("unif") e distribuição de Gauss ("normal");
- beta – lista de vetores que indicam o efeito de cada covariável simulada. O número de vetores inserido deve corresponder ao número de covariáveis simuladas, que por sua vez deve ser coerente com o número de acontecimentos considerados em dist.ev.

Importa ainda fazer algumas considerações acerca dos argumentos que acabaram de ser apresentados. Os parâmetros beta0, que surgem nos argumentos beta0.ev e beta0.cen, são utilizados na parametrização[3] das distribuições escolhidas, tanto para o tempo de vida como para o tempo de censura. Relativamente aos vetores definidos no argumento x, estes devem conter o nome da distribuição, seguidos da definição do(s) parâmetro(s) correspondente(s): no caso da distribuição de Bernoulli define-se apenas probabilidade de sucesso; no caso da distribuição uniforme define-se primeiro o seu valor mínimo e depois o seu valor máximo; e no caso de uma distribuição de Gauss define-se primeiro a média e depois o desvio padrão.

Para além dos argumentos descritos anteriormente, existem outros contidos no *package* survsim que podem ser convenientemente definidos, como por exemplo: o argumento z que permite simular uma covariável onde é induzido o efeito aleatório associado à heterogeneidade individual ou à correlação intraindivíduos; o argumento lambda que serve para estipular a duração média que um indivíduo leva a sofrer cada acontecimento; e o argumento max.ep que é utilizado para indicar o número máximo de acontecimentos observados num mesmo indivíduo (quando este não é indicado, a função rec.ev.sim assume por padrão que esse número é ilimitado durante o tempo máximo que foi estipulado no argumento foltime). Além disso, tem interesse referir que se um indivíduo sofrer mais acontecimentos do que aqueles que se encontram definidos no argumento dist.ev, o procedimento de simulação utiliza sempre os últimos valores que foram definidos em cada argumento, de maneira a que seja possível gerar os acontecimentos seguintes.

Recentemente, Moriña e Navarro [59] desenvolveram uma nova função, designada por crisk.sim, que permite simular dados no contexto dos acontecimentos competitivos, onde incluíram a possibilidade de especificar a distribuição que se encontra associada ao risco de ocorrência de cada acontecimento.

C.2 *Outputs* do R resultantes da aplicação dos modelos marginais

O presente anexo compila os restantes *outputs* do R resultantes da implementação dos modelos marginais, cujos comandos base foram introduzidos na secção 4.4.

Em primeiro lugar, serão apresentados os *outputs* que dizem respeito ao cálculo da estimativa global dos parâmetros de regressão:

[3]Para mais informações acerca destes parâmetros consulte-se Moriña e Navarro [57].

- **Modelo PWP-GT (tempo por intervalos)**

```
> modeloPWPGT <- coxph(Surv(time, status) ~ as.factor(x) + x.1 + x.2 +
    strata(obs.episode) + cluster(nid), data = dados1)
> summary(modeloPWPGT)
```

n = 1723, number of events = 723

	coef	exp(coef)	se(coef)	robust se	z	p
as.factor(x)[T.1]	0.56684	1.76269	0.07934	0.08190	6.921	4.47e-12
x.1	0.17226	1.18798	0.12944	0.13906	1.239	0.215
x.2	0.67220	1.95855	0.04541	0.04374	15.368	<2e-16

Concordance = 0.707 (se = 0.018)
Rsquare = 0.139 (max possible = 0.989)
Likelihood ratio test = 258.7 on 3 df, p = 0
Wald test = 266 on 3 df, p = 0
Score (logrank) test = 253 on 3 df, p = 0, Robust = 175.5 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

- **Modelo AG (processo de contagem)**

```
> modeloAG <- coxph(Surv(start, stop, status) ~ as.factor(x) + x.1 + x.2 +
    cluster(nid), data = dados1)
> summary(modeloAG)
```

n = 1723, number of events = 723

	coef	exp(coef)	se(coef)	robust se	z	p
as.factor(x)[T.1]	0.67075	1.95571	0.07770	0.10013	6.699	2.1e-11
x.1	0.17860	1.19554	0.12890	0.17805	1.003	0.316
x.2	0.87392	2.39628	0.04121	0.04664	18.737	<2e-16

Concordance = 0.736 (se = 0.011)
Rsquare = 0.263 (max possible = 0.995)
Likelihood ratio test = 526.8 on 3 df, p = 0
Wald test = 381.9 on 3 df, p = 0
Score (logrank) test = 516 on 3 df, p = 0, Robust = 114.8 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

- **Modelo WLW (tempo total)**

```
> modeloWLW <- coxph(Surv(fulltime, status) ~ as.factor(x) + x.1 + x.2 +
    strata(obs.episode) + cluster(nid), data = dados2)
> summary(modeloWLW)
```

n = 10000, number of events = 723

	coef	exp(coef)	se(coef)	robust se	z	p
as.factor(x)[T.1]	0.81485	2.25885	0.07821	0.12500	6.519	7.07e-11
x.1	0.18763	1.20639	0.12852	0.21332	0.880	0.379
x.2	1.18989	3.28672	0.04666	0.07139	16.667	<2e-16

Concordance = 0.769 (se = 0.026)
Rsquare = 0.076 (max possible = 0.598)
Likelihood ratio test = 785.6 on 3 df, p = 0
Wald test = 300.7 on 3 df, p = 0
Score (logrank) test = 739.7 on 3 df, p = 0, Robust = 115.1 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

- **Modelo LWA (tempo total)**

```
> modeloLWA <- coxph(Surv(fulltime, status) ~ as.factor(x) + x.1 + x.2 +
    cluster(nid), data = dados2)
> summary(modeloLWA)
```

n = 10000, number of events = 723

	coef	exp(coef)	se(coef)	robust se	z	p
as.factor(x)[T.1]	0.70723	2.02837	0.07779	0.10704	6.607	3.91e-11
x.1	0.18172	1.19928	0.12848	0.18672	0.973	0.33
x.2	0.94243	2.56622	0.04209	0.05365	17.566	<2e-16

Concordance = 0.742 (se = 0.011)
Rsquare = 0.057 (max possible = 0.716)
Likelihood ratio test = 586.9 on 3 df, p = 0
Wald test = 332.4 on 3 df, p = 0
Score (logrank) test = 571.8 on 3 df, p = 0, Robust = 110.5 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

Relativamente ao cálculo das estimativas específicas dos parâmetros, estas foram obtidas apenas para os 5 primeiros estratos, uma vez que os modelos PWP e WLW consideram um conjunto de indivíduos em rico restritivo e semirrestritivo, respetivamente. Este aspeto faz com que existam poucos indivíduos em risco nos últimos estratos, tornando as estimativas pouco fiáveis. De seguida, serão apresentados os *outputs* resultantes:

- **Modelo PWP-CP (processo de contagem)**

```
> modeloPWPCPesp <- coxph(Surv(start, stop, status) ~ strata(obs.episode)/
    (as.factor(x) + x.1 + x.2) + cluster(nid), data = dados1, subset = (obs.episode
    <= 5))
> summary(modeloPWPCPesp)

n = 1662, number of events = 689
```

	coef	exp(coef)	se(coef)	robust se
obs.episode=1:as.factor(x)[T.1]	0.46477	1.59165	0.10648	0.10416
obs.episode=2:as.factor(x)[T.1]	0.67469	1.96343	0.17062	0.17152
obs.episode=3:as.factor(x)[T.1]	0.84802	2.33502	0.26499	0.27377
obs.episode=4:as.factor(x)[T.1]	-0.21876	0.80352	0.36074	0.41553
obs.episode=5:as.factor(x)[T.1]	-0.62475	0.53539	0.46729	0.53307
obs.episode=1:x.1	0.01177	1.01184	0.18203	0.17741
obs.episode=2:x.1	0.32468	1.38359	0.26572	0.27883
obs.episode=3:x.1	-0.18043	0.83491	0.39788	0.44334
obs.episode=4:x.1	-0.21836	0.80383	0.67545	0.71853
obs.episode=5:x.1	-0.34528	0.70802	0.74243	0.88276
obs.episode=1:x.2	1.06403	2.89802	0.06688	0.06445
obs.episode=2:x.2	0.20387	1.22614	0.09562	0.09423
obs.episode=3:x.2	0.45594	1.57766	0.13862	0.13060
obs.episode=4:x.2	0.06057	1.06244	0.22123	0.26327
obs.episode=5:x.2	0.21026	1.23400	0.27520	0.28532

	z	p
obs.episode=1:as.factor(x)[T.1]	4.462	8.12e-06
obs.episode=2:as.factor(x)[T.1]	3.934	8.37e-05
obs.episode=3:as.factor(x)[T.1]	3.098	0.001951
obs.episode=4:as.factor(x)[T.1]	-0.526	0.598572
obs.episode=5:as.factor(x)[T.1]	-1.172	0.241200

	z	p
obs.episode=1:x.1	0.066	0.947116
obs.episode=2:x.1	1.164	0.244244
obs.episode=3:x.1	-0.407	0.684019
obs.episode=4:x.1	-0.304	0.761204
obs.episode=5:x.1	-0.391	0.695696
obs.episode=1:x.2	16.510	<2e-16
obs.episode=2:x.2	2.164	0.030499
obs.episode=3:x.2	3.491	0.000481
obs.episode=4:x.2	0.230	0.818035
obs.episode=5:x.2	0.737	0.461164

Concordance = 0.727 (se = 0.017)
Rsquare = 0.181 (max possible = 0.985)
Likelihood ratio test = 331.3 on 15 df, p = 0
Wald test = 359.7 on 15 df, p = 0
Score (logrank) test = 312.3 on 15 df, p = 0, Robust = 204.6 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

- **Modelo PWP-GT (tempo por intervalos)**

```
> modeloPWPGTesp <- coxph(Surv(time, status) ~ strata(obs.episode)/
  (as.factor(x) + x.1 + x.2) + cluster(nid), data = dados1, subset = (obs.episode
  <= 5))
> summary(modeloPWPGTesp)
```

n = 1662, number of events = 689

	coef	exp(coef)	se(coef)	robust se
obs.episode=1:as.factor(x)[T.1]	0.464772	1.591651	0.106478	0.104162
obs.episode=2:as.factor(x)[T.1]	0.697006	2.007732	0.169849	0.169465
obs.episode=3:as.factor(x)[T.1]	0.937714	2.554135	0.258861	0.237793
obs.episode=4:as.factor(x)[T.1]	0.040976	1.041827	0.343059	0.331713
obs.episode=5:as.factor(x)[T.1]	-0.429100	0.651095	0.450705	0.424490
obs.episode=1:x.1	0.011767	1.011837	0.182029	0.177407
obs.episode=2:x.1	0.264855	1.303242	0.265663	0.269657
obs.episode=3:x.1	0.078131	1.081264	0.364851	0.366002
obs.episode=4:x.1	-0.008917	0.991122	0.521155	0.510255
obs.episode=5:x.1	0.366515	1.442698	0.706291	0.738952

	coef	exp(coef)	se(coef)	robust se
obs.episode=1:x.2	1.064029	2.898023	0.066876	0.064447
obs.episode=2:x.2	0.225091	1.252437	0.091024	0.088383
obs.episode=3:x.2	0.515921	1.675180	0.134017	0.117054
obs.episode=4:x.2	0.146846	1.158176	0.183432	0.143523
obs.episode=5:x.2	0.286088	1.331210	0.257810	0.215441

	z	p
obs.episode=1:as.factor(x)[T.1]	4.462	8.12e-06
obs.episode=2:as.factor(x)[T.1]	4.113	3.91e-05
obs.episode=3:as.factor(x)[T.1]	3.943	8.03e-05
obs.episode=4:as.factor(x)[T.1]	0.124	0.9017
obs.episode=5:as.factor(x)[T.1]	-1.011	0.3121
obs.episode=1:x.1	0.066	0.9471
obs.episode=2:x.1	0.982	0.3260
obs.episode=3:x.1	0.213	0.8310
obs.episode=4:x.1	-0.017	0.9861
obs.episode=5:x.1	0.496	0.6199
obs.episode=1:x.2	16.510	<2e-16
obs.episode=2:x.2	2.547	0.0109
obs.episode=3:x.2	4.408	1.05e-05
obs.episode=4:x.2	1.023	0.3062
obs.episode=5:x.2	1.328	0.1842

Concordance = 0.715 (se = 0.018)
Rsquare = 0.186 (max possible = 0.99)
Likelihood ratio test = 342.2 on 15 df, p = 0
Wald test = 372.9 on 15 df, p = 0
Score (logrank) test = 322.7 on 15 df, p = 0, Robust = 205.8 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

- **Modelo WLW (tempo total)**

```
> modeloWLWesp <- coxph(Surv(fulltime, status) ~ strata(obs.episode)/
  (as.factor(x) + x.1 + x.2) + cluster(nid), data = dados2, subset = (obs.episode
  <= 5))
> summary(modeloWLWesp)
```

n = 5000, number of events = 689

	coef	exp(coef)	se(coef)	robust se
obs.episode=1:as.factor(x)[T.1]	0.46477	1.59165	0.10648	0.10416
obs.episode=2:as.factor(x)[T.1]	1.00340	2.72754	0.16984	0.16739
obs.episode=3:as.factor(x)[T.1]	1.52923	4.61463	0.25824	0.24448
obs.episode=4:as.factor(x)[T.1]	1.25188	3.49692	0.32844	0.31159
obs.episode=5:as.factor(x)[T.1]	0.97192	2.64303	0.41483	0.39702
obs.episode=1:x.1	0.01177	1.01184	0.18203	0.17741
obs.episode=2:x.1	0.26820	1.30761	0.26915	0.27210
obs.episode=3:x.1	0.30820	1.36097	0.36705	0.36589
obs.episode=4:x.1	0.14826	1.15981	0.49880	0.49264
obs.episode=5:x.1	0.26630	1.30513	0.66651	0.66077
obs.episode=1:x.2	1.06403	2.89802	0.06688	0.06445
obs.episode=2:x.2	1.07589	2.93260	0.09626	0.09091
obs.episode=3:x.2	1.38577	3.99790	0.13467	0.11543
obs.episode=4:x.2	1.36754	3.92570	0.17405	0.15464
obs.episode=5:x.2	1.58613	4.88479	0.23605	0.22231

	z	p
obs.episode=1:as.factor(x)[T.1]	4.462	8.12e-06
obs.episode=2:as.factor(x)[T.1]	5.995	2.04e-09
obs.episode=3:as.factor(x)[T.1]	6.255	3.98e-10
obs.episode=4:as.factor(x)[T.1]	4.018	5.88e-05
obs.episode=5:as.factor(x)[T.1]	2.448	0.0144
obs.episode=1:x.1	0.066	0.9471
obs.episode=2:x.1	0.986	0.3243
obs.episode=3:x.1	0.842	0.3996
obs.episode=4:x.1	0.301	0.7635
obs.episode=5:x.1	0.403	0.6869
obs.episode=1:x.2	16.510	<2e-16
obs.episode=2:x.2	11.835	<2e-16
obs.episode=3:x.2	12.005	<2e-16
obs.episode=4:x.2	8.843	<2e-16
obs.episode=5:x.2	7.135	9.69e-13

Concordance = 0.767 (se = 0.023)
Rsquare = 0.136 (max possible = 0.825)
Likelihood ratio test = 730.2 on 15 df, p = 0
Wald test = 415.6 on 15 df, p = 0
Score (logrank) test = 687.3 on 15 df, p = 0, Robust = 198.5 p = 0

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

C.3 Resíduos

O principal objetivo deste trabalho consistiu essencialmente em apresentar o modo como os modelos marginais são implementados no R. Contudo, é preciso consciencializar que após o ajustamento de qualquer modelo deve-se avaliar a sua adequabilidade, o que implica que se efetue a análise dos resíduos. Assim, nesta secção pretende-se exemplificar, através do modelo de Cox clássico e do modelo PWP-CP, como é que essa análise deve ser feita. Além disso, de entre os quatro tipos de resíduos apresentados na secção 2.5, decidiu-se apenas analisar os resíduos de Schoenfeld e os resíduos martingala, uma vez que estes são aqueles que podem, até certo ponto, inviabilizar a interpretação dos resultados obtidos.

- **Modelo de Cox clássico**

Geralmente, começa-se por analisar a influência das covariáveis na proporcionalidade dos riscos, tanto no global como individualmente. Para isso, através da função cox.zph do *package* survival, determinou-se os resíduos de Schoenfeld padronizados, tendo-se efetuado o seguinte procedimento:

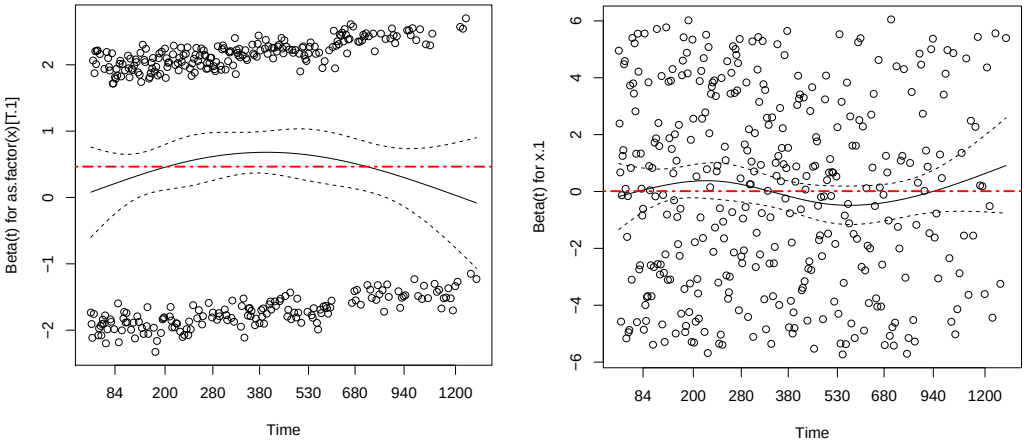
```
> sch.modeloCox <- cox.zph(modeloCox)
> sch.modeloCox
```

	rho	chisq	p
as.factor(x)[T.1]	0.00882	0.0271	0.869
x.1	-0.02637	0.2482	0.618
x.2	0.06762	1.6780	0.195
GLOBAL	NA	1.8808	0.598

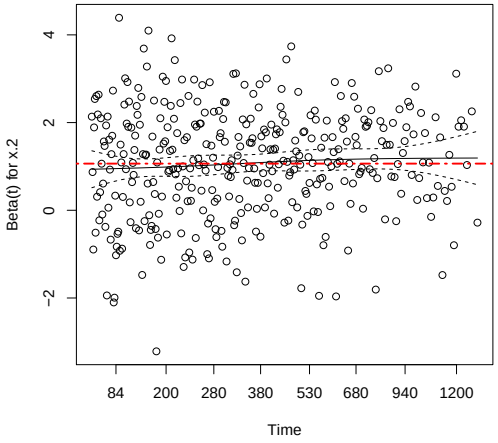
Na tabela que foi gerada, a coluna rho representa o coeficiente de regressão linear entre os resíduos e o tempo de sobrevivência, e as colunas chisq e p representam o valor da estatística qui-quadrado e o correspondente valor-p. Tal como se observa, nenhum dos testes rejeita a hipótese de proporcionalidade dos riscos.

Ainda assim, os testes referentes a cada covariável podem ser acompanhados por um gráfico de resíduos, onde se representa a curva de suavização *spline* dos mesmos e o respetivo intervalo de confiança. Para além disso, também se engloba uma linha horizontal (neste caso a vermelho) que diz respeito ao efeito constante da covariável que foi estimado por esse modelo. No caso de não existir violação da proporcionalidade, espera-se que a linha horizontal se mantenha dentro do intervalos de confiança. Na Figura C.1 é

possível visualizar os gráficos resultantes, os quais podem ser simplesmente obtidos através da função plot(sch.modeloCox).



(a) Covariável x. (b) Covariável x.1.



(c) Covariável x.2.

Figura C.1: Gráficos dos resíduos de Schoenfeld padronizados do modelo de Cox clássico.

Através da análise gráfica conclui-se que o efeito das covariáveis x.1 e x.2 podem ser considerados constantes. Em relação à covariável x, o seu efeito revela ser diferente tanto no início como no fim do período de observação, mas isto acontece porque existem poucas observações, o que significa que estas variações podes ser desvalorizadas. Além do mais, o teste global não rejeita a proporcionalidade dos riscos.

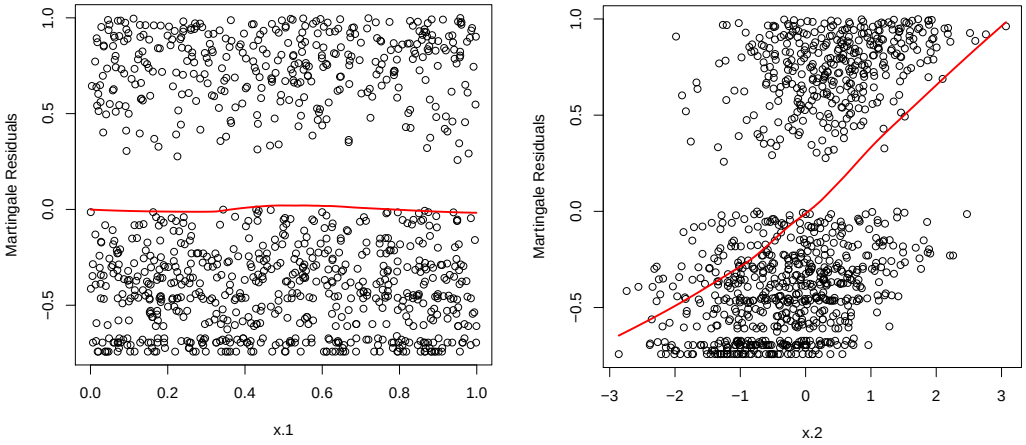
No que toca aos resíduos martingala, estes podem ser utilizados tanto para averiguar a existência de pontos influentes, como para explorar se a forma

funcional de uma covariável contínua é linear ou necessita de alguma transformação (por exemplo, através do logaritmo). O que se segue diz apenas respeito à segunda situação, que consiste em analisar os resíduos martingala do modelo nulo, isto é, do modelo sem covariáveis. Após o ajustamento deste modelo, procede-se à construção dos gráficos de resíduos, onde no eixo das abcissas se colocam os valores da covariável que se pretende estudar e no eixo das ordenadas representam-se os resíduos martingala do modelo nulo. Usualmente, para facilitar esta análise, também se representa a curva *lowess*. No R, este procedimento pode ser conseguido através dos seguintes passos:

```
> modeloCox.nulo <- coxph(Surv(start, stop, status) ~ 1, data=dados1, subset
    = (obs.episode == 1))
> martmodeloCox.nulo <- resid(modeloCox.nulo, type = "martingale")

> plot(dados1[dados1$obs.episode == 1, "x.1"], martmodeloCox.nulo, xlab =
    "x.1", ylab="Martingale Residuals")
> lines(lowess(dados1[dados1$obs.episode == 1, "x.1"], martmodeloCox.nulo,
    iter = 0), lwd=2, col="red")
```

Tal como é possível observar, encontra-se exemplificado o modo como se obtém o gráfico de resíduos martingala da covariável x.1, sendo que para a covariável x.2 apenas é necessário efetuar as devidas adaptações.



(a) Covariável x.1. (b) Covariável x.2.

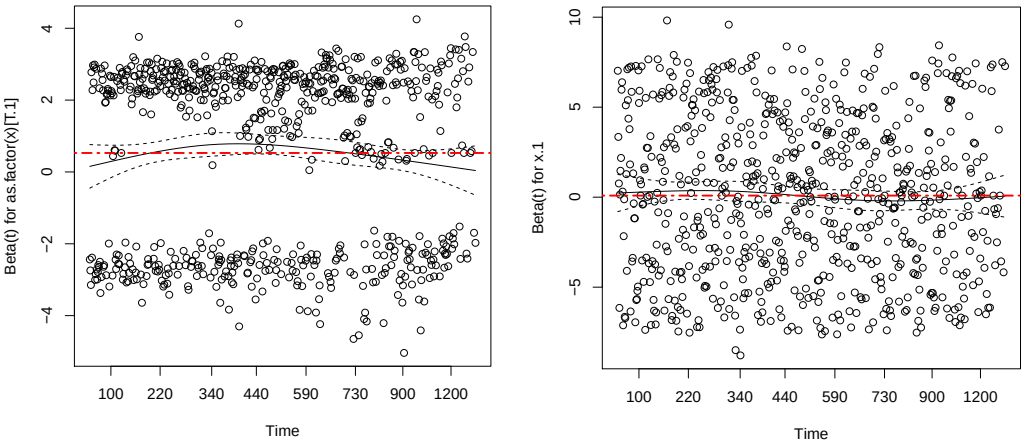
Figura C.2: Gráficos dos resíduos martingala do modelo de Cox clássico.

Por conseguinte, obteve-se os gráficos representados na Figura C.2, de onde se conclui que não existem dúvidas de que a forma funcional de am-

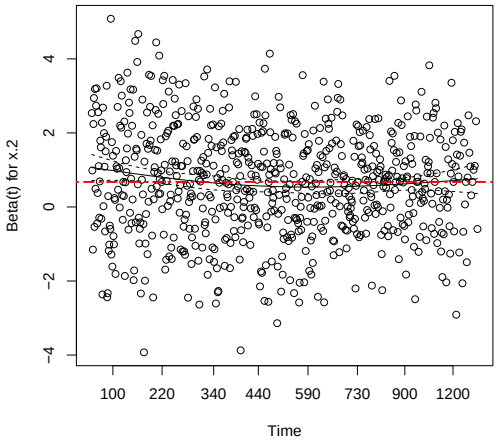
bas as covariáveis é linear, o que significa que o efeito destas foi estimado corretamente. Assim não é necessário proceder a qualquer transformação.

- **Modelo PWP-CP**

No que diz respeito aos restantes modelos, a análise dos resíduos de Schoenfeld e dos resíduos martingala baseia-se essencialmente no mesmo procedimento efetuado para o modelo anterior. Como exemplo, analise-se os resíduos do modelo PWP-CP.



(a) Covariável x. (b) Covariável x.1.



(c) Covariável x.2.

Figura C.3: Gráficos dos resíduos de Schoenfeld padronizados do modelo PWP-CP.

Os resíduos de Schoenfeld padronizados do modelo PWP-CP podem ser determinados do seguinte modo:

```
> sch.modeloPWPCP <- cox.zph(modeloPWPCP)
> sch.modeloPWPCP
```

	rho	chisq	p
as.factor(x)[T.1]	-0.0282	0.665	0.4149
x.1	-0.0417	1.560	0.2116
x.2	-0.0613	3.229	0.0723
GLOBAL	NA	5.211	0.1570

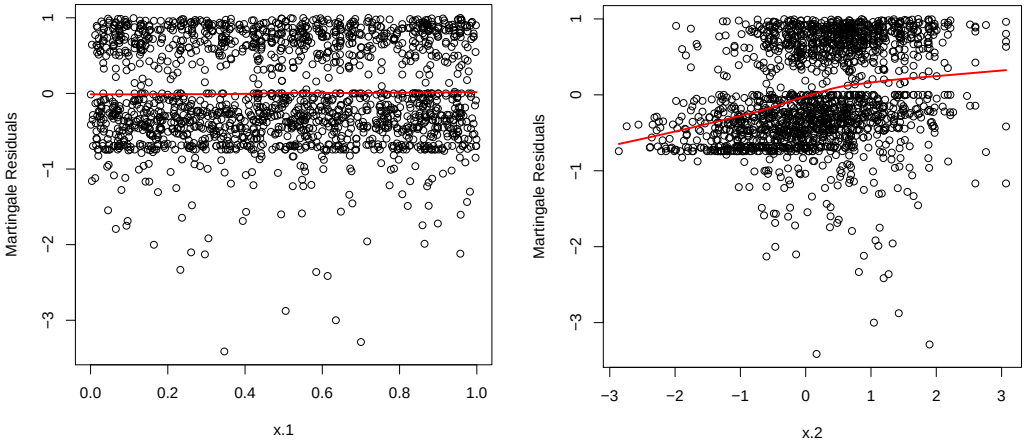
De acordo com a tabela que foi gerada, não existe evidência para rejeitar a proporcionalidade dos riscos, tanto a nível global como a nível individual. Para complementar a análise, pode-se construir os respetivos gráficos de resíduos, utilizando a função plot(sch.modeloPWPCP). Assim, obteve-se os gráficos que constam na Figura C.3. As variações observadas na covariável x, à semelhança do que aconteceu no modelo anterior, podem ser desvalorizadas. Em relação à covariável x.1, esta revela claramente um efeito constante ao longo do tempo. Por fim, na covariável x.2 observa-se uma variação nos instantes iniciais, mas esta pode também ser desvalorizada, uma vez que no global não existe violação da hipótese de riscos proporcionais.

Quanto aos resíduos martingala do modelo nulo, estes podem ser calculados por intermédio dos comandos que se seguem:

```
> modeloPWPCP.nulo <- coxph(Surv(start, stop, status ) ~ 1 + strata(
    obs.episode), data=dados1)
> martmodeloPWPCP.nulo <- resid(modeloPWPCP.nulo, type = "martingale")

> plot(dados1$x.1, martmodeloPWPCP.nulo, xlab = "x.1", ylab="Martingale
    Residuals")
> lines(lowess(dados1$x.1, martmodeloPWPCP.nulo, iter=0), lwd=2, col=
    "red")
```

Novamente, apenas foi apresentado o modo como se obtém o gráfico de resíduos da covariável x.1, sendo que para a outra efetua-se o mesmo procedimento. Com base na Figura C.4, concluí-se que a forma funcional da covariável x.1 e da covariável x.2 pode ser considerada linear. Assim sendo, também aqui não é necessário transformar as covariáveis.



(a) Covariável x.1. (b) Covariável x.2.

Figura C.4: Gráficos dos resíduos martingala do modelo PWP-CP.

Bibliografia

[1] Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, Vol. 6, No. 4, p. 701–726.

[2] Abreu, A. M. (1997). *Modelos de Sobrevivência para Populações Heterogéneas*. Tese de Mestrado. Faculdade de Ciências da Universidade de Lisboa.

[3] Abreu, A. M. (2004). *Modelos de Sobrevivência para Populações com Indivíduos Imunes*. Tese de Doutoramento. Universidade da Madeira.

[4] Abreu, A. M. (2014/2015). Apontamentos das aulas de "Complementos de Estatística". Universidade da Madeira.

[5] Abreu, A. M. (2014/2015). Sebenta das aulas de "Estatística Computacional". Universidade da Madeira.

[6] Allignol, A. e Latouche, A. (2016). *CRAN Task View: Survival Analysis*. Consultado a 6 de julho de 2016, através do URL: `https://cran.r-project.org/web/views/Survival.html`.

[7] Andersen, P. K., Borgan, O., Gill, R. D. e Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York. ISBN: 978-0-387-94519-4.

[8] Andersen, P. K. e Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, Vol. 10, No. 4, p. 1100–1120.

[9] Andreozzi, V. e Carvalho, M. S. (2011). Sobrevivência de múltiplos eventos. Em *Boletim da Sociedade Portuguesa de Estatística: Análise de Sobrevivência*, Editor: Rosado, F., p. 73–80. Edições SPE, Lisboa.

[10] Barlow, W. E. e Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, Vol. 75, No. 1, p. 65–74.

[11] Bastos, J. e Rocha, C. (2006). Análise de Sobrevivência: Conceitos Básicos. *Arquivos de Medicina*, Vol. 20, No. 5–6, p. 185–187.

[12] Bastos, J. e Rocha, C. (2007). Análise de Sobrevivência: Métodos Não Paramétricos. *Arquivos de Medicina*, Vol. 21, No. 3–4, p. 111–114.

[13] Bijwaard, G. E., Franses, P. H. e Paap, R. (2006). Modeling purchases as repeated events. *Journal of Business & Economic Statistics*, Vol. 24, No. 4, p. 487–502.

[14] Boher, J. e Cook, R. J. (2006). Implications of model misspecification in robust tests for recurrent events. *Lifetime Data Anal*, Vol. 12, No. 1, p. 69–95.

[15] Borges, A. I. (2014). *Análise de Sobrevivência com o R*. Tese de Mestrado. Universidade da Madeira.

[16] Box-Steffensmeier, J. M. e Zorn, C. (2002). Duration models for repeated events. *The Journal of Politics*, Vol. 64, No. 4, p. 1069–1094.

[17] Breslow, N. E. (1972). Contribution to discussion of paper by D. R. Cox. *Journal of Royal Statistical Society, Series B*, Vol. 34, No. 2, p. 216–217.

[18] Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, Vol. 30, No. 1, p. 89–99.

[19] Breslow, N. E. e Crowley, J. (1974). A Large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, Vol. 2, No. 3, p. 437–453.

[20] Cabete, A. B. (2012). *Análise de Sobrevivência com acontecimentos múltiplos: Aplicação ao estudo do tempo até à ocorrência de enfarte do miocárdio*. Tese de Mestrado. Faculdade de Ciências da Universidade de Lisboa.

[21] Cai, J., e Schaubel, D. E. (2004). Analysis of recurrent event data. Em Balakrishnan, N. e Rao, C. R. (Eds.), *Handbook of Statistics 23: Advances in Survival Analysis*, Elsevier, North Holland, p. 603–623. ISBN: 978-0-444-50079-3.

[22] Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S. e Shimakura, S. E. (2011). *Análise de Sobrevivência: teoria e aplicações em saúde*, 2.ª Edição. Fiocruz, Rio de Janeiro, Brasil. ISBN: 978-85-7541-216-9.

[23] Castañeda, J. e Gerritse, B. (2010). Appraisal of several methods to model time to multiple events per subject: modelling time to hospitalizations and death. *Revista Colombiana de Estadística*, Vol. 33, No. 1, p. 43–61.

[24] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2.ª Edição. Chapman and Hall/CRC, Boca Raton. ISBN: 978-1-584-88325-8.

[25] Cook, R. J. e Lawless, J. F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research*, Vol. 11, No. 2, p. 141–166.

[26] Cook, R. J. e Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media. ISBN: 978-0-387-69809-0.

[27] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, Vol. 34, No. 2, p. 187–220.

[28] Cox, D. R. (1975). Partial Likelihood. *Biometrika*, Vol. 62, No. 2, p. 269–276.

[29] Cox, D. R. e Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London, New York. ISBN: 978-0-412-24490-2.

[30] Cui, J., Forbes, A., Kirby, A., Marschner, I., Simes, J., Hunt, D., West, M. and Tonkin, A. (2010). Semi-parametric risk prediction models for recurrent cardiovascular events in the LIPID study. *BMC Medical Research Methodology*, Vol. 10, No. 1, p. 1–9.

[31] Fleming, T. R. e Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics - Theory and Methods*, Vol. 13, No. 20, p. 2469–2486.

[32] Fox, J. e Carvalho, M. S. (2012). The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis. *Journal of Statistical Software*, Vol. 49, No. 7, p. 1–32.

[33] García-Mora, B. e Santamaría, C., Rubio, G. e Pontones, J. L. (2008). Modeling the recurrence-progression process in bladder carcinoma. *Computers and Mathematics with Applications*, Vol. 56, No. 3, p. 619–630.

[34] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, Vol. 52, No. 1-2, p. 203–223.

[35] Ghosh, D. e Lin, D. Y. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica*, Vol. 12, No. 3, p. 663–688.

[36] Grambsch, P. M. e Therneau, T. M. (1994). Proportional hazards tests and diagnostic based on weighted residuals. *Biometrika*, Vol. 81, No. 3, p. 515–526.

[37] Greenwood, M. (1926). The natural duration of cancer. *Reports of Public Health and Medical Subjects*, HMSO, London, Vol. 33, p. 1–26.

[38] Guo, S. e Zeng, D. (2014). An overview of semiparametric models in survival analysis. *Journal of Statistical Planning and Inference*, Vol. 151–152, p. 1–16.

[39] Guo, Z., Gill, T. M. e Allore, H. G. (2008). Modeling repeated time-to-event health conditions with discontinuous risk intervals: an example of longitudinal study of functional disability among older persons. *Methods of Information in Medicine*, Vol. 47, No. 2, p. 1–16.

[40] Gutiérrez, E., Lozano, S. e González, J. R. (2011). A recurrent-events survival analysis of the duration of olympic records. *Journal of Management Mathematics*, Vol. 22, No. 2, p. 115–128.

[41] Harrell Jr, F. E. (2016). *Hmisc: Harrell Miscellaneous. Package* do R versão 3.17-3. URL: `https://CRAN.R-project.org/package=Hmisc`.

[42] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York. ISBN: 978-1-4612-7087-4.

[43] Kalbfleisch, J. D. e Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2.ª Edição. John Wiley & Sons, New York. ISBN: 978-0-471-36357-6.

[44] Kaplan, E. L. e Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, Vol. 53, No. 282, p. 457–481.

[45] Kelly, P. J. e Lim, L. L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, Vol. 19, No. 1, p. 13–33.

[46] Klein, J. P. e Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, 2.ª Edição. Springer-Verlag, New York. ISBN: 978-0-387-95399-1.

[47] Kleinbaum, D. G. e Klein, M. (2012). *Survival Analysis. A Self-Learning Text*, 3.ª Edição. Springer, New York. ISBN: 978-1-441-96645-2.

[48] Lawless, J. F. (2003). *Statistical Models and Methods for lifetime Data*. 2.ª Edição. John Wiley & Sons, New York. ISBN: 978-0-471-37215-8.

[49] Lee, E. W., Wei, L. J. e Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. Em Klein, J. P. e Goel, P. K. (Eds.), *Survival Analysis: State of the Art*, Kluwer Academic Publisher, Dordrecht, p. 237–247. ISBN: 978-90-481-4133-3.

[50] Lim, H. J., Liu, J. e Melzer-Lange, M. (2007). Comparison of methods for analyzing recurrent events data: application to the emergency department visits of pediatric firearm victims. *Accident Analysis and Prevention*, Vol. 39, No. 2, p. 290–299.

[51] Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, Vol. 13, No. 21, p. 2233–2247.

[52] Lin, D. Y. e Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, Vol. 84, No. 408, p. 1074–1078.

[53] Mantel, N. e Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, Vol. 22, No. 4, p. 719–748.

[54] Marubini, E. e Valsecchi, M. G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley & Sons, Ltd. Chichester. ISBN: 978-0-471-93987-0.

[55] McKeague, I. W. (1997). Introduction to Aalen (1978) 'Nonparametric Inference for a Family of Counting Processes'. Em Kotz, S. e Johnson, N. L. (Eds.), *Breakthroughs in Statistics. Springer Series in Statistics*, Springer New York, Vol. 3, p. 347–367. ISBN: 978-1-4612-0667-5.

[56] Metcalfe, C. e Thompson, S. G. (2007). Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: Should it be applied to a recurrent events outcome? *Statistical Methods in Medical Research*, Vol. 16, No. 2, p. 103–122.

[57] Moriña, D. e Navarro, A. (2014). The R Package survsim for the Simulation of Simple and Complex Survival Data. *Journal of Statistical Software*, Vol. 59, No. 2, p. 1–20.

[58] Moriña, D. e Navarro, A. (2015). *survsim: Simulation of Simple and Complex Survival Data. Package* do R versão 1.1.4. URL: `https://CRAN.R-project.org/package=survsim`.

[59] Moriña, D. e Navarro, A. (2016). Competing risks simulation with the survsim R package. *Communications in Statistics – Simulation and Computation*. Retirado a partir do URL do jornal: `http://dx.doi.org/10.1080/03610918.2016.1175621`.

[60] Nelson, W. B. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, Vol. 1, p. 27–52.

[61] Nelson, W. B. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, Vol. 14, No. 4, p. 945–966.

[62] Pandeya, N., Purdie, D. M., Green, A. e Williams, G. (2005). Repeated occurrence of basal cell carcinoma of the skin and multifailure survival analysis: follow-up data from the nambour skin cancer prevention trial. *American Journal of Epidemiology*, Vol. 161, No. 8, p. 748–754.

[63] Parpia, S., Thabane, L., Julian, J. A., Whelan, T. J., Levine, M. N. (2013). Empirical comparison of methods for analyzing multiple time-to-event outcomes in non-inferiority trial: a breast cancer study. *BMC Medical Research Methodology*, Vol. 13, No. 1, p. 1–7.

[64] Paulino, C. D., Pestana, D., Branco, J., Singer, J., Barroso, L. e Bussab, W. (2015). *Glossário Inglês-Português de Estatística*. 3.ª Edição. Sociedade Portuguesa de Estatística e Associação Brasileira de Estatística.

[65] Peto, R. (1972). Contribution to discussion of paper by D. R. Cox. *Journal of Royal Statistical Society, Series B*, Vol. 34, No. 2, p. 205–207.

[66] Peto, R. e Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, Vol. 135, No. 2, p. 185–207.

[67] Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, Vol. 65, No. 1, p. 167–179.

[68] Prentice, R. L. (1983). Amendments and corrections: 'Linear rank tests with right censored data'. *Biometrika*, Vol. 70, No. 1, p. 304.

[69] Prentice, R. L., Williams, B. J. e Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, Vol. 68, No. 2, p. 373–379.

[70] R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org/`.

[71] Rocha, C. e Papoila, A. P. (2009). *Análise de Sobrevivência*. XVII Congresso da Sociedade Portuguesa de Estatística. SPE. ISBN: 978-972-8890-22-3.

[72] Sagara, I., Giorgi, R., Doumbo, O. K., Piarroux, R. e Gaudart, J. (2014). Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. *Malaria Journal*, Vol. 13, No. 1, p. 1–9.

[73] Schaubel, D. E. e Cai, J. (2005). Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients. *Biostatistics*, Vol. 6, No. 3, p. 404–419.

[74] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, Vol. 69, No. 1, p. 239–241.

[75] Signorovitch, J. E. e Wei, L-J. (2008). Wei-Lin-Weissfeld method for multiple times to events. Em Ralph, B., D'Agostino, R. B., Sullivan, L. e Massaro, J. (Eds.), *Wiley Encyclopedia of Clinical Trials*, John Wiley & Sons, Ltd. ISBN: 978-0-4713-5203-7.

[76] Sousa-Ferreira, I. e Abreu, A. M. (2016). Hybrid model for recurrent event data. *The 25th International Workshop on Matrices and Statistics (IWMS'2016)*. Madeira, Portugal.

[77] Tarone, R. E. e Ware, J. H. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, Vol. 64, No. 1, p. 156–160.

[78] Therneau, T. M. (2015). *survival: A package for survival analysis in S. Package* do R versão 2.38-3. URL: `http://CRAN.R-project.org/package=survival`.

[79] Therneau, T. M. e Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer-Verlag, New York. ISBN: 978-1-4419-3161-0.

[80] Therneau, T. M., Grambsch, P. M. e Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, Vol. 77, No. 1, p. 147–160.

[81] Therneau, T. M. e Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, Vol. 16, No. 18, p. 2029–2047.

[82] Triad, sou. e Nagashima, K. (2015). *RcmdrPlugin.KMggplot2: A Rcmdr Plug-in for Kaplan-Meier Plots and Other Plots by Using the ggplot2 Package. Package* do R versão 0.2-3. URL: `https://CRAN.R-project.org/package=RcmdrPlugin.KMggplot2`.

[83] Ullah, S., Gabbett, T. J. e Finch, C. F. (2014). Statistical modelling for recurrent events: an application to sports injuries. *British Journal of Sports Medicine*, Vol. 48, No. 17, p. 1287–1293.

[84] Villegas, R., Julià, O. e Ocaña, J. (2013). Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring. *BMC Medical Research Methodology*, Vol. 13, No. 1, p. 1–10.

[85] Wei, L. J. e Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine*, Vol. 16, No. 8, p. 833–839.

[86] Wei, L. J., Lin, D. Y. e Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, Vol. 84, No. 408, p. 1065–1073.

[87] Xu, Y., Lam, K. F. e Cheung, Y. B. (2014). Estimation of intervention effects using recurrent event time data in the presence of event dependence and a cured fraction. *Statistics in Medicine*, Vol. 33, No. 13, p. 2263–2274.

[88] Ye, Y., Kalbfleisch, J. D. e Schaubel, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*, Vol. 63, No. 1, p. 78–87.