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

Análise de Sobrevivência com o R

Alexandra Isabel Monteiro Borges

UNIVERSIDADE da MADEIRA

Análise de Sobrevivência com o R

DISSERTAÇÃO DE MESTRADO

Alexandra Isabel Monteiro Borges
MESTRADO EM MATEMÁTICA

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

Setembro | 2014

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

Análise de Sobrevivência com o R

DISSERTAÇÃO DE MESTRADO

Alexandra Isabel Monteiro Borges
MESTRADO EM MATEMÁTICA

ORIENTAÇÃO
Ana Maria Cortesão Pais Figueira da Silva Abreu

Análise de Sobrevivência com o R

Alexandra Isabel Monteiro Borges

Setembro 2014

"De tudo ficaram três coisas...
A certeza de que estamos começando...
A certeza de que é preciso continuar...
A certeza de que podemos ser interrompidos antes de terminar...
Façamos da interrupção um caminho novo...
Da queda, um passo de dança...
Do medo, uma escada...
Do sonho, uma ponte...
Da procura, um encontro!"

Fernando Sabino - Trecho de "III – O Escolhido", do livro "O Encontro Marcado"

Agradecimentos

É com muita alegria e com sentimento de realização pessoal que termino esta etapa da minha vida. Não teria sido possível terminá-la se não tivesse tido a colaboração das pessoas que fazem parte do meu dia-a-dia. E, não poderia deixar de as agradecer por, directa ou indirectamente, me influenciarem nesta jornada.

O meu primeiro agradecimento é dirigido a Deus, pois tem-me sempre guiado num bom caminho, sempre com saúde, vontade de viver e sobretudo com vontade de superar-me todos os dias como ser humano.

Um especial e enorme agradecimento à minha Orientadora, a Professora Dr.ª Ana Maria Abreu, que sem ela, nada disto teria sido possível, pois soube sempre como me ajudar a ultrapassar os obstáculos que foram surgindo, incentivou-me, mostrou sempre dar valor ao meu trabalho e esforço, teve muita paciência e foi sem dúvida muito dedicada a este trabalho. Por tudo isto e muito mais, agradeço do fundo do meu coração.

Como não podia deixar de ser, um especial agradecimento aos meus Pais, Isabel e José, que sempre tentaram dar todas e as melhores condições para que conseguisse estudar e ser a pessoa que sou hoje. Acrescentando o meu querido e amigo Irmão, João Pedro que, juntamente com os meus Pais, sempre me apoiou, incentivou e que às vezes só recebeu as minhas rabugices. A eles um grande obrigado por tornarem o nosso ambiente familiar feliz, divertido e com muito amor. Sem o seu apoio e colaboração incondicional teria sido bem mais difícil.

Ao João, o meu Companheiro, amigo e confidente, o meu obrigado pela paciência, incentivo, colaboração e fazer-me sempre sorrir mesmo quando às vezes parecia impossível.

Às minhas queridas amigas de infância, Rita Mourato e Sara Santos, que, apesar da distância que nos separa, a amizade sempre se manteve e sempre nos apoiamos umas às outras. Aos meus queridos amigos de longa data, Catarina Teixeira, Carlos Quintal, Pedro Rocha, Joana Gomes e Filipa Costa, que sempre me fizeram sentir eu própria, apoiaram e tornaram o

meu dia mais alegre e divertido. Às minhas queridas amigas e colegas, Eva Henriques, Fábia Camacho, Helena Teixeira, Carla Spínola, Graça Paulo e Érica Serrão, um obrigado por sempre poder contar com elas, simplesmente para me fazerem sorrir ou pelo conforto de um ombro.

À Carina Alves e Mariana Rodrigues, que começaram por ser minhas colegas e que acabaram por se tornar minhas tutoras e amigas. Agradeço também às minhas Chefes e aos meus colegas de trabalho que me ajudaram e apoiaram, de alguma forma.

Aos meus Professores que me acompanharam ao longo destes anos académicos, obrigada pelo conhecimento que me transmitiram e alguns até, carinho e amizade.

A todos os meus familiares, amigos, professores e colegas de escola e de curso que contribuíram para a pessoa que sou hoje e que de alguma maneira me ajudaram no decorrer da minha vida.

Agradeço-vos a todos, do fundo do meu coração!

Resumo

O principal objectivo desta dissertação é dar a conhecer as potencialidades da linguagem R pois ainda existem algumas reservas quanto à sua utilização. E nada melhor que a análise de sobrevivência, por ser um tema da estatística com grande impacto no mundo das doenças e novas curas, para mostrar como este programa apresenta grandes vantagens.

Esta dissertação é então composta por quatro capítulos.

No primeiro capítulo introduzimos alguns conceitos fundamentais da análise de sobrevivência, os quais servirão de suporte para o terceiro capítulo. Assim sendo, apresentamos um pouco da sua história, conceitos básicos, conceitos novos numa perspectiva de regressão diferente da que estamos habituados, tendo como objectivo a construção de modelos de regressão tendo sempre em conta métodos para averiguar se o modelo é o mais adequado ou não.

No segundo capítulo apresentamos o R, o *package R Commander* (que já tem um interface mais amigável), o *package survival* (talvez o mais importante na análise de sobrevivência clássica), bem como outros *packages* que poderão ser úteis para quem quiser aprofundar o seu uso nesta área.

O terceiro capítulo é o que aplica os conhecimentos dos dois anteriores e no qual pretendemos dar a conhecer algumas das muitas possibilidades de utilização deste *software* nesta área da Estatística. Este é dividido em três, ou seja, está dividido consoante as etapas que vamos precisando para trabalhar a nossa base de dados, começando pela análise descritiva, para conhecermos os dados que temos, depois a função de sobrevivência, por ser um conceito importante e por fim, a construção de modelos de regressão, não paramétricos e paramétricos.

Por último, apresentamos as nossas conclusões deste trabalho.

Palavras-Chave: Análise de Sobrevivência, linguagem R, *package survival*, *R Commander*.

Abstract

The main goal of this dissertation is to show the potentials of the R language in order to overtake some reservations in terms of it's usage. Due of the great impact in the world of diseases and new ways of healing, survival analysis is the best way to show the potential and advantages of this program.

This dissertation has four chapters.

In the first chapter we will introduce some fundamental concepts of the survival analysis, which will serve as a support to the third chapter. We will present some of his history, basic concepts, and new concepts in a different perspective of regression, having in mind methods to evaluate if the model fits the data.

In the second chapter we introduce the R, the package R Commander (has a friendly interface), the package survival (the most important in the classical survival analysis), like some other packages that could be useful to whom would like to improve their knowledge in this area.

On the third chapter we apply the knowledge of the previous chapters and the usage of this software in this statistical area. It is divided in three, according to the stages needed to work with the data, beginning by descriptive analysis, to know the data we have, then the survival functions, because it's an important concept and, at the end, by constructing regression models, parametric non-parametric.

To finish we will present the conclusions of this work.

Key-words: Package survival, R Commander, R language, Survival Analysis.

Índice

1 Análise de Sobrevivência **1**
- 1.1 Introdução . . . 1
- 1.2 Conceitos básicos . . . 2
- 1.3 Censura . . . 4
- 1.4 Estimador de Kaplan-Meier . . . 8
- 1.5 Variáveis explanatórias . . . 10
- 1.6 Modelos de Regressão . . . 11
  - 1.6.1 Introdução . . . 11
  - 1.6.2 Modelo de Cox . . . 13
  - 1.6.3 Modelos Paramétricos . . . 16
- 1.7 Resíduos de Schoenfeld . . . 19

2 A linguagem R **21**
- 2.1 Noções gerais sobre o R . . . 21
- 2.2 Alguns *packages* úteis para a Análise de Sobrevivência . . . 24
  - 2.2.1 *R Commander* . . . 25
  - 2.2.2 *survival* . . . 30
  - 2.2.3 Outros *packages* . . . 30

3 Análise de Sobrevivência com o R **35**
- 3.1 Análise descritiva . . . 37
- 3.2 Função de sobrevivência . . . 38
  - 3.2.1 Algumas variantes . . . 42
- 3.3 Modelos de regressão . . . 44
  - 3.3.1 Modelo de Cox . . . 44
  - 3.3.2 Modelos paramétricos . . . 51
  - 3.3.3 Algumas variantes . . . 56

4 Conclusão **57**

Bibliografia **63**

Lista de Figuras

1.1 Monotonia da Função de Risco. 4
1.2 Vários tipos de censura à direita. 5

2.1 Janela do R. 22
2.2 Janela de ajuda do comando *RSiteSearch*. 23
2.3 Janela do *R Commander*. 25
2.4 Importação de ficheiros de texto, do *clipboard* ou da *internet*. . 27
2.5 Interface do *RcmdrPlugin.EZR* 29

3.1 Análise descritiva das variáveis numéricas. 38
3.2 Tabela de frequências e de percentagens para o tratamento. . . 38
3.3 Comandos e respectivos *outputs* para a estimativa de Kaplan--Meier da função de sobrevivência. 40
3.4 Estimativa de Kaplan-Meier para a função de sobrevivência derivada do comando original *versus* Estimativa de Kaplan-Meier para a função de sobrevivência derivada de modificações no comando original. 41
3.5 Diferenças entre as curvas de sobrevivência para o tipo de tratamento. 41
3.6 Estimativa de Kaplan-Meier para a função de sobrevivência para cada um dos grupos de tratamento. 42
3.7 Estimativa de Kaplan-Meier para a função de sobrevivência de uma sub-amostra através do *plug-in RcmdrPlugin.KMggplot2*. 43
3.8 Modelo de Cox com todas as variáveis da base de dados. . . . 45
3.9 Modelo de Regressão de Cox apenas com as variáveis significativas. 46
3.10 Curvas de Kaplan-Meier para as covariáveis significativas no modelo de Cox para testar a proporcionalidade das funções de risco. 47
3.11 *Output* gerado para testar a proporcionalidade das funções de risco dos vários passos para a construção do modelo de Cox com o respectivo coeficiente de determinação. 48

3.12 Teste de independência do Qui-quadrado para testar se as variáveis 4 ou mais nódulos (*node4*) e recorrência (*rec*) são independentes. 49
3.13 Resíduos de Schoenfeld para as variáveis idade (*age*) e extensão do tumor (*extent*). 50
3.14 Resíduos de Schoenfeld para a variável recorrência (*rec*). . . . 50
3.15 Modelo de Cox final com as covariáveis idade (*age*), extensão do tumor (*extent*) e recorrência (*rec*). 51
3.16 Gráfico de $\log\left[-\log \widehat{S}_0(t)\right]$ versus o logaritmo do tempo de vida. 52
3.17 Modelo de Weibull sem covariáveis. 52
3.18 Comando que fornece os valores da função especificada, neste caso, a função que gerou o modelo de regressão de Weibull, mas com mais casas decimais. 53
3.19 Obtenção dos parâmetros da recta através da função *Convert-Weibull* do *package SurvRegCensCov*. 53
3.20 Modelo Weibull com as covariáveis idade (*age*), extensão do tumor (*extent*) e recorrência (*rec*). 54
3.21 Modelo de regressão Log-logístico sem covariáveis. 55
3.22 Modelo de regressão log-logístico com as covariáveis idade (*age*), extensão do tumor (*extent*) e recorrência (*rec*). 55
3.23 Função de risco. 56

Capítulo 1

Análise de Sobrevivência

1.1 Introdução

A análise de sobrevivência é um procedimento estatístico implementado com maior frequência a partir de meados do século XX, atingindo o seu maior desenvolvimento e popularidade por volta da década de oitenta desse século. A medicina surge fortemente ligada à análise de sobrevivência, apesar deste procedimento estatístico poder ser usado nas mais variadas áreas, tais como na engenharia, sociologia, psicologia, educação, etc.. Pode ter como objecto de estudo, por exemplo, o tempo até que um automóvel tem a sua primeira avaria mecânica após a sua venda; o tempo desde que um criminoso é solto da prisão até reincidir no crime ou até mesmo o tempo de vida de uma máquina após substituição de uma componente mecânica.

A análise de sobrevivência consiste, entre outras coisas, em analisar os tempos de vida dos indivíduos desde o seu momento de entrada no estudo, até ao momento em que ocorre o acontecimento de interesse, acontecimento esse que é definido à partida. Este acontecimento é geralmente definido como uma falha, que poderá ser morte, recaída de uma doença ou até mesmo quando um tratamento começa a fazer efeito no paciente.

Tem a particularidade de lidar com dados censurados, ou seja, em alguns dos indivíduos pode não chegar a ocorrer o acontecimento de interesse durante o período de observação. Indivíduos que continuam vivos após o término do estudo ou que abandonam o tratamento são exemplos de indivíduos com tempo de vida censurado.

Outro aspecto importante a ter em conta, é o uso das variáveis explanatórias, ou covariáveis. O tempo de vida dos indivíduos é afectado por estas variáveis. Este tipo de dados sugere uma análise de regressão, mas não poderia ser uma regressão habitual devido às particularidades que este tipo

de dados apresenta. É precisamente devido às observações censuradas que a análise de sobrevivência se distingue.

Neste capítulo iremos apresentar os conceitos essenciais da análise de sobrevivência. Depois de algumas definições básicas, formalizaremos a noção central de censura. Apresentaremos o estimador de Kaplan-Meier para o tempo de sobrevivência e suas propriedades, e alguns modelos de regressão habitualmente utilizados para modelar a influência de variáveis explanatórias no tempo de sobrevivência. Estes modelos poderão ser paramétricos ou não paramétricos, dependendo do conhecimento que se tenha sobre a distribuição do tempo de vida dos indivíduos.

1.2 Conceitos básicos

Comecemos por definir que o tempo de vida de um determinado indivíduo, de uma população homogénea, será representado por uma variável aleatória T, não negativa e absolutamente contínua. Podemos então definir a função de sobrevivência desse indivíduo como sendo a probabilidade dele sobreviver para além de um instante t e que vamos representar por:

$$S\left(t\right) = P\left(T > t\right), t \geqslant 0$$

que é uma função monótona, não crescente e contínua e que tem as seguintes propriedades:

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

A função densidade de probabilidade num instante t é a taxa instantânea de morte nesse instante:

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

Uma função igualmente utilizada neste âmbito, é a função de risco (*hazard function*), também conhecida por função intensidade, taxa de falha, ou força de mortalidade, descrita como a taxa instantânea de morte de um indivíduo, que sobreviva até ao instante t, dada por:

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

e que satisfaz as seguintes condições:

1. $h(t) \geq 0$;
2. $\int_0^\infty h(t)dt = \infty$.

A partir das definições anteriores, é possível obter-se relações que poderão tornar-se úteis, tais como:

$$h(t) = \frac{f(t)}{S(t)}$$

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \tag{1.1}$$

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right)$$

A função de risco cumulativa, que é uma função não negativa e monótona crescente, é definida através de:

$$H(t) = \int_0^t h(u)du, t \geqslant 0$$

donde, por (1.1), tem-se:

$$S(t) = \exp\left[-H(t)\right] \iff H(t) = -\log S(t)$$

isto é, mede o risco de ocorrência do acontecimento de interesse até ao instante t.

Podemos também definir a função de distribuição, $F(t)$, como sendo a probabilidade de ocorrer o acontecimento de interesse até ao instante t, ou seja:

$$F(t) = P(T \leq t), 0 \leq t < \infty$$

Visto a função de sobrevivência ser decrescente, não permite uma leitura directa de como evolui o risco de morte ao longo do tempo. Essa evolução fica mais patente na forma da função de risco. Deste modo, temos cinco formas possíveis para a função de risco, como ilustra a Figura 1.1. Assim, esta função pode ser:

1. monótona crescente: é a forma da função de risco mais comum na análise de sobrevivência pois corresponde a um risco crescente;

2. monótona decrescente: mais raro, uma vez que corresponde a um risco decrescente;

3. constante: acontece quando ou o período de observação é curto, ou surge uma "situação imprevista", uma doença, ou um acidente e só é registado esse tempo;

4. *bathtub-shaped*: adequada para caracterizar a mortalidade populacional, pois no período inicial as mortes resultam essencialmente de doenças infantis, após o que se segue uma fase em que o risco de morte decresce e se mantém baixo até haver um novo aumento devido ao envelhecimento;

5. *hump-shaped* ou unimodal: acontece, por exemplo, devido a uma intervenção cirúrgica, onde o paciente aumenta o risco de morte no momento da operação, mas que ocorre sem complicações.

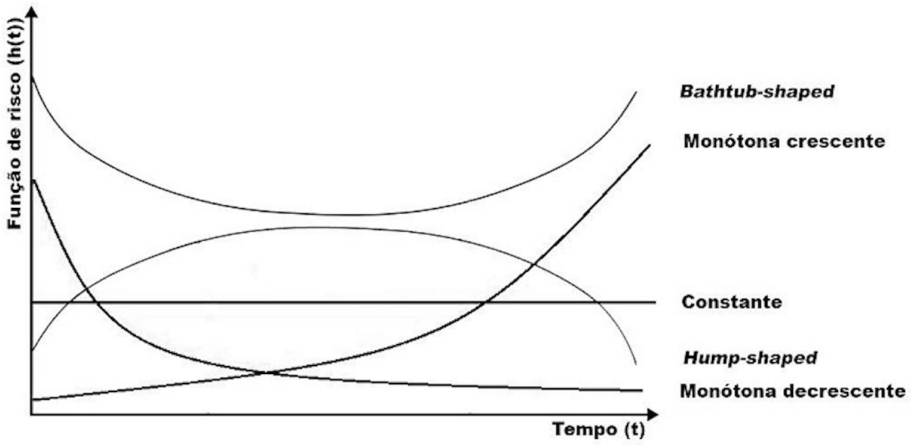

Figura 1.1: Monotonia da Função de Risco.

1.3 Censura

Uma característica própria dos dados de sobrevivência é o facto de o evento de interesse não ser experienciado em todas as observações do estudo. A esta característica é dada o nome de censura [1]. A censura pode advir dos limites de tempo ou de outro tipo de restrições, dependendo da natureza do estudo. Dizemos que uma variável aleatória é censurada quando não é

possível observar o seu valor exacto, mas se consegue obter um limite inferior para esse valor (censura à direita), ou um limite superior (censura à esquerda), ou ambos (censura intervalar).

Existem vários tipos de censura, como já referimos, que podem ocorrer, mas a mais comum, é a censura à direita.

A censura ocorre devido, essencialmente, a três motivos:

- o estudo chegou ao fim sem que fosse observado o acontecimento de interesse;
- o indivíduo em estudo fica perdido para *follow-up*;
- o indivíduo é retirado do estudo por algum motivo, relacionado com o tempo de vida.

A Figura 1.2 reflete o que pode acontecer com os indivíduos no estudo, onde Y representa o tempo de vida dos indivíduos durante o período de observação.

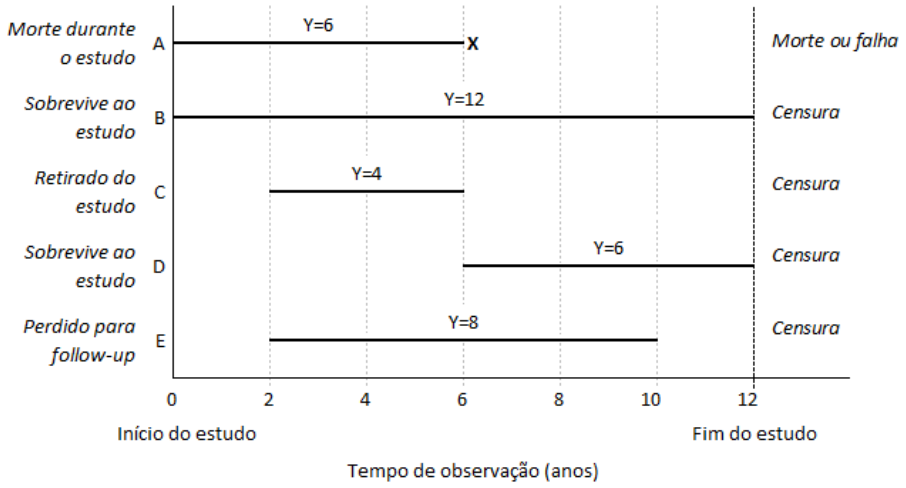

Figura 1.2: Vários tipos de censura à direita.

Na primeira situação, A, o indivíduo entra no início do estudo e experiencia o evento, ou a morte, no ano 6, o que significa ter um tempo de vida de 6 anos. No caso do indivíduo B, entra no início do estudo e continua vivo no término do mesmo, conferindo-lhe um tempo de vida de, pelo menos 12 anos, isto é, tem um tempo censurado de 12 anos. O indivíduo C, entra mais tarde no estudo, no 3^o ano, mas é retirado do estudo no 5^o ano porque o

tratamento deixou de ser eficaz, conferindo–lhe um tempo de vida de, pelo menos 4 anos. Devido ao término do estudo, o indivíduo D, não experiencia o evento ou a morte, conferindo-lhe um tempo de vida de pelo menos 6 anos. Por último, temos o caso do indivíduo E, que é uma situação semelhante à do C, possuindo um tempo de vida de pelo menos 8 anos, pois a partir desse ano, o indivíduo é perdido para *follow-up*, por algum motivo perdeu-se o contacto.

Existem essencialmente cinco tipos de censura:

1. Censura à direita: Estamos perante censura à direita quando apenas sabemos que o tempo de vida do indivíduo excede um determinado valor. Pode acontecer pelo motivo dos indivíduos não quererem permanecer no estudo e por isso a informação fica incompleta; ou porque o indivíduo não experienciou o evento ou a morte antes do fim do estudo; ou até mesmo que o contacto foi perdido com este. Nestas situações ficamos com informação parcial e sabemos que o evento ocorreu (ou irá ocorrer) algures depois da data do último *follow-up*. Apesar disso não podemos ignorar estas observações visto que nos fornecem alguma informação acerca da sobrevivência, não sabemos a data exacta da sua morte, mas sabemos que foi após certo instante. Ainda podemos dividir este tipo de censura em três:

   - Tipo I: este tipo de censura ocorre quando o estudo é concebido para acabar após x anos de *follow-up*. Neste caso, todos os indivíduos em que não se tenha observado o acontecimento de interesse durante o estudo, são considerados tempos censurados no ano x.
   - Tipo II: o estudo termina quando um número de eventos predefinido acontece;
   - Censura aleatória: o estudo é desenhado para acabar ao fim de x anos, ou seja, o indivíduo entra no estudo aquando da sua data de diagnóstico, e devido ao facto do término do estudo ter sido fixado previamente, o tempo que os indivíduos permaneceram no estudo é aleatório.

2. Censura à esquerda: acontece quando o tempo de vida é inferior ao tempo observado, ou seja, o evento de interesse já ocorreu nalgum instante anterior ao da observação. É um tipo de censura menos comum.

3. Censura intervalar: neste tipo de censura, sabemos que o evento de interesse aconteceu, mas não sabemos ao certo quando, apenas que aconteceu num certo intervalo de tempo.

4. Censura independente: acontece quando a razão para haver censura é independente da razão que leva à morte.

5. Censura não informativa: Neste tipo de censura temos de garantir que um indivíduo censurado é representativo de todos os indivíduos que sobreviveram e que tenham as mesmas características ou covariáveis.

Os estudos devem ser concebidos de forma a que a censura seja não informativa, ou seja, de forma a que a censura seja causada por algo que não seja o fracasso iminente.

Na análise de sobrevivência, os métodos de inferência estatística mais utilizados são, de um modo geral, baseados na teoria assintótica da máxima verosimilhança pois, como temos o factor de censura, é dificultada a obtenção de distribuições de amostragem exactas.

Pretende-se realizar uma inferência sobre o modelo paramétrico indexado por um vector de parâmetros θ que a distribuição do tempo de vida T segue.

Para construir a função de verosimilhança vamos considerar T e C duas variáveis aleatórias (v.a.'s) que representam o tempo de vida total e o tempo até a censura (à direita), respectivamente. O tempo de vida observado para um indivíduo é a v.a. $Y = \min\{T, C\}$. Os instantes de morte e censura poderiam em princípio coincidir, mas nesse caso convenciona-se que a censura ocorre depois da morte. Quando se observa a morte do indivíduo ($\delta = 1$) num intervalo suficientemente pequeno $[t, t + dt[$ então, admitindo independência entre T e C,

$$P[t \leq Y < t + dt, \delta = 1] = P[t \leq T < t + dt]P[C \geq t + dt]$$

e quando o tempo de vida é censurado ($\delta = 0$) temos:

$$P[t \leq Y < t + dt, \delta = 0] = P[t \leq C < t + dt]P[T \geq t + dt]$$

Se T e C forem absolutamente contínuas, com funções densidade f e g e com funções de sobrevivência S e $1 - G$, dividindo as duas igualdades anteriores por dt e fazendo $dt \to 0^+$ fica

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

Assim, para uma amostra aleatória de dimensão n, $(t_1, \delta_1), ..., (t_n, \delta_n)$, temos a função de verosimilhança:

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

Se tivermos o caso de censura não informativa (que é a situação mais comum), a expressão anterior pode ainda ser simplificada, obtendo-se:

$$L = \prod_{i=1}^{n} \left[f\left(t_i\right)\right]^{\delta_i} \left[S\left(t_i\right)\right]^{1-\delta_i} \iff L = \prod_{i=1}^{n} \left[h\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \tag{1.2}$$

visto que a distribuição do tempo de censura não depende do vector de parâmetro de interesse θ.

Os resultados assintóticos usuais da teoria da máxima verosimilhança continuam válidos, sob condições de regularidade bastante gerais nos processos de morte e censura, pelo que, o estimador de máxima verosimilhança, $\widehat{\theta}$, tem distribuição assintótica normal multivariada com valor médio θ e matriz de covariância $I(\theta)^{-1}$, sendo $I(\theta)$ a matriz de informação de Fisher.

1.4 Estimador de Kaplan-Meier

Se não houvesse censura, a função de sobrevivência seria estimada pela proporção de indivíduos que sobreviveram além do instante t:

$$\widehat{S}(t) = \frac{\text{número de tempos de vida} > t}{n}, t \geq 0$$

mas na presença de censura nem todos os tempos de vida serão observados na sua totalidade. Para ultrapassar esta dificuldade, Kaplan e Meier, [2] propuseram em 1958 uma generalização da função de sobrevivência empírica, conhecida como estimador de Kaplan-Meier (KM) ou estimador "produto--limite", que passamos a descrever.

Numa amostra com n tempos, suponha-se que $r \leqslant n$ correspondem a observações do acontecimento de interesse (que para concretizar ideias vamos designar por "morte") e os restantes $n-r$ tempos são censurados. Se $t_{(1)}, ...,$ $t_{(r)}$ são os instantes de morte, n_i é o número de indivíduos em risco imediatamente antes do instante $t_{(i)}$, e d_i é o número de mortes nesse instante, então o estimador de Kaplan-Meier para a função de sobrevivência define-se como:

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

onde $\widehat{S}(t) = 1$ para $0 \leq t \leq t_{(1)}$. Os instantes de morte e censura poderiam em princípio coincidir, mas nesse caso convenciona-se que a censura ocorre depois da morte.

Note-se que, quando não existe censura, a função de sobrevivência empírica e o estimador de Kaplan-Meier coincidem. Podemos também afirmar

que, se a maior observação registada for não censurada ($t \geq t_{(r)}$), então $\widehat{S}(t) = 0$. Mas, se a maior observação registada for censurada, t^*, então $\widehat{S}(t) = \widehat{S}(t_{(r)})$ para $t_{(r)} \leq t \leq t^*$, pois a estimativa só está definida até esse instante e não atinge o valor zero.

Uma outra propriedade deste estimador é ser uma função em escada, em que os "saltos" são os instantes onde a morte é observada. $\widehat{S}(t)$ é um estimador consistente de $S(t)$ e, sob certas condições de regularidade, pode ser considerado como um estimador de máxima verosimilhança não paramétrico de $S(t)$.

É possível obter uma estimativa da variância de $\widehat{S}(t)$, através da expressão:

$$\widehat{var}\left\{\widehat{S}(t)\right\} = \left[\widehat{S}(t)\right]^2 \sum_{i:t_{(i)}\leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{1.3}$$

que é conhecida por fórmula de Greenwood.

Podemos então estabelecer o intervalo de confiança para o verdadeiro valor da função de sobrevivência no instante t_0. Visto $\widehat{S}(t)$ ter uma distribuição assintótica normal de valor médio $S(t)$ e variância dada por (1.3), um intervalo de $100(1-\alpha)\%$ de confiança para $S(t_0)$, é dado por:

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

Apesar de ser o intervalo mais usado, apresenta alguns problemas, nomeadamente devido ao facto de ser simétrico, pois quando a estimativa $\widehat{S}(t_0)$ estiver próxima de 0 ou 1, os seus limites podem estar fora do intervalo $(0, 1)$. Como alternativa a este intervalo, pode-se usar uma transformação, por exemplo, $\log\left[-\log \widehat{S}(t_0)\right]$ e calcular o seu intervalo de confiança. Aos intervalos de confiança obtidos desta forma dá-se o nome de intervalos de confiança ponto-a-ponto (*pointwise*), por dizerem respeito a instantes específicos.

A distribuição do tempo de vida é, geralmente, assimétrica positiva, sendo preferível usar a mediana como medida central de localização. Então, sendo t_i o i-ésimo instante de morte com $i = 1, ..., r$, a estimativa da mediana do tempo de vida é dada por:

$$m = \min\left\{t_{(i)} : \widehat{S}(t_{(i)}) \leq 0.5\right\}$$

onde $\widehat{S}(t)$ representa a estimativa de Kaplan-Meier da função de sobrevivência.

$\widehat{S}(t)$ pode ser superior a 0.5 para todos os valores de t e, nesses casos, utiliza-se outra medida de localização (já não central), como por exemplo, a

estimativa de outro quantil mais conveniente. Temos então a estimativa do quantil de probabilidade p:

$$\widehat{\chi}_p = \min\left\{t_{(i)} : \widehat{S}(t_{(i)}) \leq 1 - p\right\}$$

1.5 Variáveis explanatórias

O tempo de vida de um indivíduo é afectado por diversos factores de risco ou de prognóstico. Podemos dividir esses factores em dois grupos:

- intrínsecos: são por exemplo as variáveis do tipo idade, género, história clínica, etc., ou seja, são os factores que são inerentes aos indivíduos;

- exógenos: são aqueles factores que resultam de elementos externos ao indivíduo, como por exemplo a história familiar, factores ambientais, sociais, etc..

Este factores de risco são designados por variáveis explanatórias ou covariáveis.

Existe, de modo geral, uma classificação das covariáveis em constantes ou dependentes do tempo:

1. Constantes: Quando o seu valor não se altera durante todo o período em que o indivíduo se encontra em observação. São exemplos deste tipo de covariáveis uma variável indicatriz do tipo tratamento a que o indivíduo é sujeito, variáveis demográficas como sejam o género ou a nacionalidade, variáveis clínicas cujos valores sejam registados uma única vez ao longo do estudo, etc.;

2. Dependentes do tempo: Quando o seu valor varia ao longo do estudo. Isto acontece quando, por exemplo, existem variáveis clínicas para as quais há vários registos ao longo do estudo como, por exemplo, a pressão arterial ou o peso. Podem ainda existir factores que estão sob controlo do experimentador, variando ao longo do estudo de forma pré-determinada, como seja a dieta a que um indivíduo é sujeito. Estas covariáveis podem ainda ser divididas em:

    - Externas: São consideradas covariáveis externas todo o tipo de covariável que não está directamente relacionada com o mecanismo que regula a morte dos indivíduos;

- Internas: São aquelas onde a mudança da covariável ao longo do tempo está relacionada com a sobrevivência do indivíduo, os valores observados levam informação sobre o seu tempo de sobrevivência.

É habitual representar as covariáveis por $z_1, ..., z_p$, ou seja, designar por $\mathbf{z} = (z_1, ..., z_p)'$ o vector de covariáveis.

1.6 Modelos de Regressão

1.6.1 Introdução

Para modelar o tempo de vida de uma população homogénea, em geral, são utilizadas distribuições contínuas univariadas. Porém, a existência de heterogeneidade entre os indivíduos no que toca a factores de risco, é comum. Para que possamos incorporar esses factores, que supomos que afectam o tempo de vida, temos de recorrer a um modelo de regressão, onde o tempo de vida é a variável dependente ou de resposta e as covariáveis são as variáveis independentes. Precisamos ainda especificar um modelo para a distribuição do tempo de vida, T, o qual pode ser obtido a partir do vector $\mathbf{z} = (z_1, ..., z_p)'$ de covariáveis de um indivíduo e de algumas famílias paramétricas ou semi-paramétricas, que iremos descrever no desenvolvimento da secção.

Podemos dividir os modelos de regressão utilizados na análise de sobrevivência em três classes:

- Modelo com funções de riscos proporcionais: Apesar dos indivíduos terem diferenças nos valores das covariáveis, a proporcionalidade entre as funções de risco é mantida. A função de risco e a função de sobrevivência são, respectivamente:

$$h(t; \mathbf{z}) = h_0(t)\, \varphi(\mathbf{z}), \qquad S(t; \mathbf{z}) = [S_0(t)]^{\varphi(\mathbf{z})}$$

  com a exigência de que $\varphi(\mathbf{0}) = 1$, onde $\varphi(\mathbf{z})$ representa o risco relativo. Neste modelo, as covariáveis têm um efeito multiplicativo na função de risco.

  Um exemplo, é o modelo de Cox (modelo semi-paramétrico), mas também existem modelos paramétricos, consoante a distribuição de probabilidade que seja usada para modelar o tempo de vida.

- Modelo de tempo de vida acelerado: Este modelo, em termos de variáveis aleatórias é dado por $T = T_0/\psi(\mathbf{z})$, onde a T_0 corresponde a

função de sobrevivência S e $\psi(\mathbf{z})$é tal que $\psi(\mathbf{0}) = 1$. As funções de risco e de sobrevivência são dadas por:

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

As covariáveis têm um efeito multiplicativo no tempo. Podemos ter modelos paramétricos ou semi-paramétricos.

Este tipo de modelo é também conhecido por modelo de localização-escala para $\log T$ ou modelo log-linear para T. De facto, se representarmos o tempo na forma logarítmica, temos $\log T = \mu + \boldsymbol{\alpha}'\mathbf{z} + \sigma\varepsilon$. A função de sobrevivência é dada por:

$$S(t;\mathbf{z}) = S_0(t/\exp(\boldsymbol{\alpha}'\mathbf{z}))$$

onde μ é o termo independente; $\boldsymbol{\alpha}$ um vector de parâmetros de regressão; σ um parâmetro de escala; ε uma v.a. que representa o erro e cuja distribuição não depende de $\mathbf{z}$ e $\exp(-\boldsymbol{\alpha}'\mathbf{z})$ é designado por factor de aceleração.

- Modelo de possibilidades proporcionais: Para este tipo de modelo, a possibilidade (*odds*) de um indivíduo com vector de covariáveis $\mathbf{z}$ sobreviver para além de um determinado instante t é dado por:

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

onde $\boldsymbol{\eta} = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p$ e z_j representa o valor da j-ésima covariável com $j = 1, ..., p$.

Também conseguimos estimar a função de risco subjacente, não parametricamente, tal como calculámos para o primeiro modelo. Obtemos assim:

$$\frac{h(t;\mathbf{z})}{h_0(t)} = \{1 + (e^{\boldsymbol{\eta}} - 1)S_0(t)\}^{-1} \tag{1.4}$$

e, quando $t = 0$, a razão das funções de risco é $e^{-\boldsymbol{\eta}}$ e quando $t \to \infty$, converge para um.

Visto que os resultados obtidos ao ajustar este modelo são semelhantes aos obtidos utilizando o modelo de regressão de Cox com covariáveis dependentes do tempo, este modelo não tem muita utilização prática (Collett [3]).

1.6.2 Modelo de Cox

Em 1972, Cox [4], propôs um modelo de regressão para a análise de dados com observações censuradas que rapidamente, se tornou o mais utilizado devido à sua flexibilidade e versatilidade. Este modelo abrangia um grande número de situações práticas onde podia ser utilizado. Prova disso são os inúmeros artigos publicados, sendo possível ser usado nas mais variadas áreas, desde a medicina à engenharia. Cox deu um grande contributo e consequente desenvolvimento na análise de sobrevivência.

Tem havido vários estudos posteriores que se têm baseado no modelo de Cox, quer através de aplicações, quer através de extensões ou generalizações. Destaca-se neste modelo o facto de ser baseado na relação entre a função de risco e as covariáveis, como veremos de seguida.

Vamos considerar uma v.a. contínua T, que representa o tempo de vida de um indivíduo com vector de covariáveis associadas, $\mathbf{z} = (z_1, ..., z_p)'$, num determinado instante t. Sejam $\boldsymbol{\beta}' = (\beta_1, ..., \beta_p)$ os coeficientes de regressão (desconhecidos), que representam o efeito das covariáveis na sobrevivência e $h_0(t)$ a função de risco subjacente (função arbitrária não negativa), ou seja, aquela que corresponde a um indivíduo com vector de covariáveis nulo. Então o modelo tem a seguinte expressão:

$$h(t; \mathbf{z}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{z}) = h_0(t) \exp(\beta_1 z_1 + ... + \beta_p z_p) \tag{1.5}$$

Deste modo, o efeito das covariáveis é modelado parametricamente, mas o mesmo não acontece em relação à função de risco subjacente, pelo que o modelo de Cox é um modelo de regressão semi-paramétrico, [1].

A razão das funções de risco para dois indivíduos com covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, escreve-se da seguinte forma:

$$\frac{h(t; \mathbf{z}_1)}{h(t; \mathbf{z}_2)} = \exp\{\beta'(\mathbf{z}_1 - \mathbf{z}_2)\} \tag{1.6}$$

ou seja, não depende do tempo, t.

Tendo em conta a expressão (1.6), conclui-se que, para dois quaisquer indivíduos, as correspondentes funções de risco são proporcionais, razão pela qual este é considerado um modelo de riscos proporcionais. Ainda na mesma expressão, (1.6), se $\mathbf{z}_2 = 0$, então a razão das funções de risco é apenas $\exp(\beta'\mathbf{z}_1)$, designado por risco relativo. Verifica-se assim que as covariáveis têm um efeito multiplicativo na função de risco.

O modelo de Cox assenta no princípio que, durante o tempo de observação dos indivíduos, a influência das covariáveis na função de risco não se altera.

Habitualmente, o vector de covariáveis não é nulo, reservando-se esse valor para identificar a situação padrão. Quando uma covariável é contínua

(por exemplo, a idade), pode não fazer sentido admitir que o caso padrão corresponde a considerar que a covariável é nula. Nesta situação, é usual convencionar que o caso padrão corresponde à média dessa covariável. Por exemplo, no caso da covariável z_j a função de risco para o i-ésimo indivíduo é escrita na forma:

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

Contudo, mesmo neste caso, é possível trabalhar com as covariáveis não transformadas e com a função de risco escrita na forma habitual, uma vez que a alteração referida não modifica a inferência sobre a influência das covariáveis no risco de morte.

Sejam $\mathbf{z}_1$ e $\mathbf{z}_2$ vectores de covariáveis de dois indivíduos que apenas diferem nos valores da covariável z_j; então:

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

por isso,para interpretar os coeficientes de regressão é preferível usar $\exp(\beta_j)$, que representa o efeito multiplicativo da diferença $z_{1j} - z_{2j}$ no risco de morte.

Para melhor percebermos como funciona, vejamos dois exemplos, semelhantes aos referidos em [5].

Exemplo 1.6.1 *Temos uma covariável dicotómica (ou binária),*

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

num estudo onde se pretende averiguar o tempo em remissão, ou seja, desde o último tratamento até ao reaparecimento da doença. A função de risco será:

$$h(t;z) = \begin{cases} h(t;z=0) = h_0(t) & \textit{se o indivíduo pertence ao grupo 1} \\ h(t;z=1) = h_0(t)e^{\beta} & \textit{se o indivíduo pertence ao grupo 2} \end{cases}$$

Para o caso de $\beta < 0 \iff e^{\beta} < 1$, os indivíduos do grupo 2 irão ter melhor prognóstico que no grupo 1, mas se $\beta > 0 \iff e^{\beta} > 1$, então observa-se o contrário.

Exemplo 1.6.2 *Existem três covariáveis em que uma corresponde ao grupo de tratamento e as outras duas são potenciais factores de risco para os indivíduos;*

$$\begin{cases} z_1\textit{: tratamento (0=tradicional; 1=novo)} \\ z_2\textit{: peso no início do estudo (kg)} \\ z_3\textit{: glicose no início do estudo (mg/dL)} \end{cases}$$

e queremos estudar o efeito de um novo tratamento face ao tratamento tradicional. Os indivíduos foram distribuídos de forma aleatória pelos dois grupos de tratamento e registou-se o tempo até à obtenção de valores normais para a glicose. Temos então que:

- e^{β_1} *representa o risco (propensão) para atingir os níveis normais de glicose num indivíduo a que foi administrado o novo tratamento, face a um indivíduo com valores idênticos de peso e glicose a que tenha sido administrado o tratamento tradicional, visto que*

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

- e^{β_2} *representa o efeito de cada kg de peso a mais no tempo até se obterem níveis normais de glicose, mantendo-se idênticas as outras covariáveis, visto que*

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

- e^{β_3} *será o efeito de cada mg/dL de glicose a mais no tempo até se obterem níveis normais de glicose, fixadas as restantes covariáveis,*

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

Cox [4], para a inferência sobre β, baseou-se na função de verosimilhança parcial, dada por:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' \mathbf{z}_{(i)})}{\sum\limits_{l \in R_i} \exp(\beta' \mathbf{z}_l)} \tag{1.7}$$

onde $R_i = R(t_{(i)}) = \{j : t_j \geq t_{(i)}\}$ é o conjunto de risco no instante $t_{(i)}$, ou seja, o conjunto de índices associados aos indivíduos em observação imediatamente antes do instante $t_{(i)}$ e $t_{(1)} < ... < t_{(k)}$, $k < n$ são os k tempos de vida distintos.

A função $L(\beta)$ considerada por Cox, não é a verosimilhança habitual (1.2), que para o modelo de Cox tomaria a forma:

$$\begin{aligned} L[\beta, h_0(t)] &= \prod_{i=1}^{n} [h_0(t_i) \exp(\beta' \mathbf{z}_i) S_0(t_i)^{\exp(\beta' \mathbf{z}_i)}]^{\delta_i} [S_0(t_i)^{\exp(\beta' \mathbf{z}_i)}]^{1-\delta_i} \\ &= \prod_{i \in D} \frac{\exp(\beta' \mathbf{z}_i)}{\sum_{l \in R_i} \exp(\beta' \mathbf{z}_l)} \prod_{i \in D} \left(h_0(t_i) \sum_{l \in R_i} \exp(\beta' \mathbf{z}_l) \right) \prod_{i=1}^{n} S_0(t_i)^{\exp(\beta' \mathbf{z}_i)} \end{aligned} \tag{1.8}$$

onde D representa o conjunto de indivíduos cuja morte foi observada.

Mas uma vez que $L(\beta)$ coincide com o primeiro fator de $L[\beta, h_0(t)]$, pode ser interpretada como uma verosimilhança parcial. Como $L(\beta)$ não depende de $h_0(t)$, permite realizar inferência sobre o vector de parâmetros β sem especificar $h_0(t)$.

Em 1982, autores como Andersen e Gill [6], concluíram que, sob condições de regularidade bastante gerais, o estimador de máxima verosimilhança parcial de β é consistente e assintoticamente normal com valor médio β e matriz de covariância $I(\beta)^{-1}$, onde:

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

Para a construção de $L(\beta)$, apenas foram consideradas observações distintas uma vez que observações empatadas têm probabilidade nula sob um modelo contínuo. No entanto, em estudos práticos, é possível obter observações empatadas, essencialmente devido à escala de medida utilizada. Nesses casos, é necessário usar uma aproximação da função de verosimilhança proposta por Peto [7] e Breslow [8].

Kalbfleisch e Prentice [9], obtiveram um estimador não paramétrico de $S_0(t)$ uma vez obtido $\widehat{\beta}$ a partir da verosimilhança parcial. Quando não há observações empatadas, este reduz-se a:

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

com:

$$\widehat{\alpha}_i = \left(1 - \frac{\exp\left(\widehat{\beta}'\mathbf{z}_{(i)}\right)}{\sum\limits_{l\in R_i}\exp(\widehat{\beta}'\mathbf{z}_l)}\right)^{\exp(-\widehat{\beta}'\mathbf{z}_{(i)})}$$

donde é possível obter estimativas de $S(t, \mathbf{z})$ para qualquer $\mathbf{z}$.

1.6.3 Modelos Paramétricos

Apesar do modelo de Cox ser o mais utilizado na análise de sobrevivência, Efron [10] mostrou que se consegue mais eficiência na obtenção dos estimadores de parâmetros de regressão em modelos paramétricos, sob certas circunstâncias, do que no modelo de Cox.

Por essa razão, vamos apresentar algumas distribuições contínuas univariadas, as mais utilizadas na análise de sobrevivência, e com elas construir alguns modelos de regressão.

– Distribuição exponencial: Com T, uma v.a. com distribuição exponencial de parâmetro $\lambda > 0$ e função de densidade de probabilidade $f(t) = \lambda \exp(-\lambda t)$ com $t \geq 0$, então:

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

Esta distribuição é adequada quando o risco de morte é sempre igual ao longo do tempo.

– Distribuição de Weibull: É a mais usada na análise de sobrevivência, pois apresenta uma razoável flexibilidade para a função de risco e por a função de risco e de sobrevivência poderem ser representadas através de expressões analíticas simples. Com parâmetro de escala $\lambda > 0$ e de forma $\gamma > 0$, para $t \geq 0$, tem a função de densidade de probabilidade:

$$f(t) = \lambda\gamma t^{\gamma-1} \exp(-\lambda t^{\gamma})$$

e à custa desta, obtêm-se as funções de risco e de sobrevivência:

$$h(t) = \lambda\gamma t^{\gamma-1}, \qquad S(t) = \exp(-\lambda t^{\gamma})$$

Quando $\gamma > 1$, a função de risco é monótona crescente; quando $0 < \gamma < 1$, a função de risco é monótona decrescente e quando $\gamma = 1$, a função de risco é constante e neste caso, obtém-se a distribuição exponencial.

– Distribuição Gama: Com parâmetro de escala $\lambda > 0$ e de forma $\alpha > 0$, para $t \geq 0$, tem a função de densidade de probabilidade

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

A função de sobrevivência exprime-se como $S(t) = 1 - I(\alpha, \lambda t)$, onde $I(\alpha, x)$ é a função gama incompleta e que se define:

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

A função de risco é monótona crescente quando $\alpha > 1$ com $h(0) = 0$ e $\lim_{t\to 0^+} h(t) = \lambda$; monótona decrescente quando $0 < \alpha < 1$ com $\lim_{t\to 0^+} h(t) = \infty$ e $\lim_{t\to\infty} h(t) = \lambda$ e constante quando $\alpha = 1$ (distribuição exponencial).

Esta distibuição é menos usada que a anterior.

– Distribuição log-logística: Com parâmetro de escala $\lambda > 0$ e de forma $\alpha > 0$, para $t \geq 0$, tem a função de densidade de probabilidade:

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

e:

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}, \qquad S(t) = \frac{1}{1+\lambda t^{\alpha}}$$

como funções de risco e de sobrevivência, respectivamente. A função de risco é monótona decrescente para $0 < \alpha \leq 1$.

Para $\alpha > 1$, serve de alternativa à distribuição de Weibull quando é necessário considerar uma função de risco unimodal. É crescente desde a origem até ao valor máximo, no instante $t = \left(\frac{\alpha-1}{\lambda}\right)^{1/\alpha}$, decrescendo a partir desse instante, com $\lim_{t\to\infty} h(t) = 0$.

Após referirmos as distribuições de tempo de vida mais usadas, vamos apresentar dois dos modelos de regressão mais usados: o modelo de regressão Weibull e o modelo de regressão log-logístico.

- Modelo de regressão Weibull:

  Este é o único modelo que pode ser considerado tanto um modelo de riscos proporcionais como de tempo de vida acelerado. Consideremos a sua formulação apenas em termos de modelo de riscos proporcionais, que é a mais usual. Um indivíduo com vector de covariáveis $\mathbf{z}$ apresenta a função de risco:

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}) = \lambda\gamma t^{\gamma-1}\exp(\beta'\mathbf{z}) \tag{1.9}$$

  onde o tempo de vida desse indivíduo tem uma distribuição de Weibull com parâmetro de escala $\lambda\exp(\beta'\mathbf{z})$ e parâmetro de forma γ. Conseguimos observar que as covariáveis só afectam o parâmetro de escala. A função de sobrevivência é dada por:

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

- Modelo de regressão log-logístico:

  Este modelo de regressão pode ser utilizado em alternativa ao modelo de Weibull, quando este não se adequa, ou seja, quando temos uma função de risco não monótona ou um modelo de possibilidades proporcionais ou até um modelo de tempo de vida acelerado.

Um indivíduo com vector de covariáveis $\mathbf{z}$ apresenta a função de sobrevivência:

$$S(t;\mathbf{z}) = \frac{1}{1+\lambda\exp(\beta'\mathbf{z})t^{\alpha}}$$

onde o tempo de vida desse indivíduo tem uma distribuição log-logística com parâmetro de escala $\lambda\exp(\beta'\mathbf{z})$ e parâmetro de forma α.

Modelos que sejam ajustados aos mesmos dados, podem ser comparados através da diferença entre os valores da estatística $-2\log\widehat{L}$ para cada modelo. Ou seja, faz-se um teste de razão de verosimilhanças para testar $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, $j = 1, ..., p$, pois sob H_0, a estatística $-2\log\left(\widehat{L}_{p-1}/\widehat{L}_p\right)$ tem distribuição assintótica de um qui-quadrado com 1 grau de liberdade.

Alguns *packages* estatísticos apresentam formas automáticas para a selecção das covariáveis, mas pode não ser a melhor opção. Collett [3] mencionou este facto e propôs um outro método de selecção.

1.7 Resíduos de Schoenfeld

Os resíduos são uma ferramenta importante para testar se o modelo de regressão é adequado ou não.

O resíduo de Schoenfeld [11], para o i-ésimo indivíduo com covariável z_j, é $r_{ji} = \delta_i\{z_{ji} - a_{ji}\}$ onde:

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é não censurado} \\ 0 & \text{se } t_i \text{ é censurado} \end{cases} \text{ e } a_{ji} = \frac{\sum_{l\in R_i} z_{jl}\exp\left(\widehat{\beta}'z_l\right)}{\sum_{l\in R_i}\exp\left(\widehat{\beta}'z_l\right)}$$

com $j = 1, ..., p$ e como já vimos anteriormente, R_i é o conjunto dos indivíduos em risco no instante t_i.

Quando uma observação é censurada, o resíduo tem o valor zero, por definição. Para distinguir estes dois casos em que o tempo de vida observado coincide verdadeiramente com o previsto pelo modelo, é habitual assinalar como valores omissos os resíduos nulos associados a observações censuradas.

Para o caso da morte ser observada no instante t_i, o resíduo associado a esse indivíduo pode ser interpretado como a diferença entre o valor da covariável z_j e a média ponderada dos valores dessa covariável, para todos os indivíduos em risco nesse instante. O peso associado a cada um desses indivíduos é $\exp\left(\widehat{\beta}'z_l\right)$.

A verosimilhança parcial $L(\beta)$ verifica a igualdade:

$$\frac{\partial \log L}{\partial \beta_j}(\widehat{\beta}) = \sum_{i=1}^{n} r_{ji} = 0$$

onde $\widehat{\beta}' = (\widehat{\beta}_1, ..., \widehat{\beta}_p)'$ são as estimativas de máxima verosimilhança dos coeficientes β_j das covariáveis. Assim, a soma dos resíduos de todos os indivíduos em estudo é sempre nula para cada covariável. Quando as amostras são grandes, os resíduos são não correlacionados e o valor esperado de r_{ji} é zero.

Quando interpretamos o gráfico dos resíduos de Schoenfeld *versus* o tempo de vida (ou as ordens dos tempos de vida), se os dados se dispuserem numa nuvem de pontos aleatória, centrada em zero, então estamos perante um modelo adequado para os dados.

Uns anos mais tarde, Grambsch e Therneau [12] propuseram uma versão padronizada destes resíduos que se revelaram mais eficazes para verificar o modelo de riscos proporcionais após o ajustamento do modelo de Cox.

Capítulo 2

A linguagem R

2.1 Noções gerais sobre o R

O R surge pela criação do *R Foundation for Statistical Computing* com o objectivo de ser uma ferramenta gratuita e de utilização livre. É uma linguagem computacional formal desenhada para ser utilizada na manipulação e análise de dados, possuindo uma forte componente gráfica e estatística. Tem por base a linguagem S que foi desenvolvida em 1976 em conjunto por John Chambers e seus colaboradores. Em 1995, Robert Gentleman e Ross Ihaka, dão a conhecer o R e transformam-no em *Open Source*, ou seja, é criada a possibilidade de qualquer utilizador poder programar, interagindo com o que já existe ou criando novas funções, a fim de melhor resolver o seu problema, pois o código é aberto [13]. Esta característica confere-lhe versatibilidade e isto é possível graças aos *packages* (ou livrarias) que são as contribuições dos utilizadores de toda a parte do mundo e que qualquer utilizador pode aceder.

Para utilizar este *software* é necessário programar, pois a interação é feita através de uma janela de comandos. Todavia já estão disponíveis *packages* gráficos, nomeadamente o *R Commander*, que tornam a interface mais amigável através da utilização de menus.

Apresenta compatibilidade com diversas plataformas como o Linux, Unix, Windows, Mac Os X, etc. e estabelece ligação com interfaces como o Excel, Minitab, S-PLUS, SAS, SPSS, Stata, Systat, entre outros.

Para obtenção do *software*, a página do *R* (http://www.r-project.org/), fornece o *download* da aplicação (clicando em CRAN no menu do lado esquerdo) e todo o processo de instalação.

Após concluída a instalação, ao abrirmos o programa, é-nos apresentada uma janela, Figura 2.1, onde consta a informação da versão instalada, informações gerais sobre o programa, alguns comandos úteis de obtenção de

ajuda e o espaço para introdução da(s) linha(s) de comando, que se inicia pelo símbolo >.

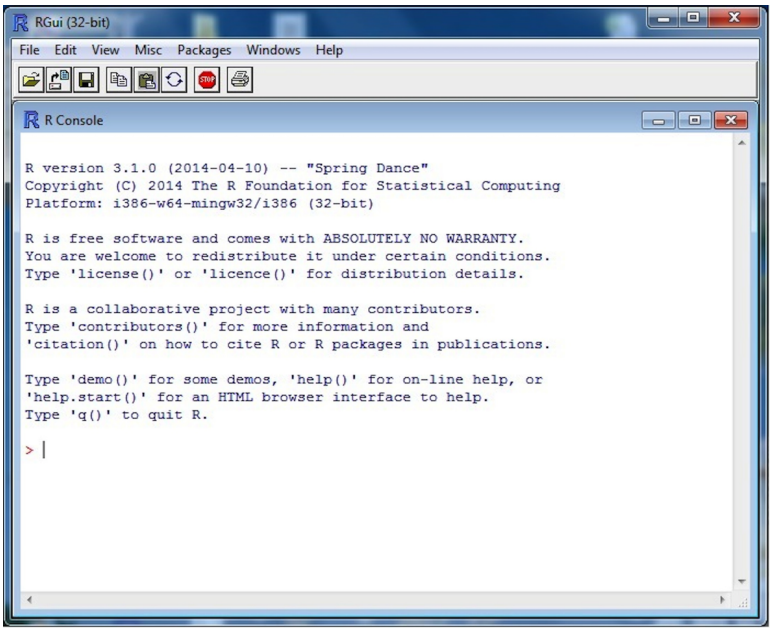

Figura 2.1: Janela do R.

Devido ao facto do R ser uma linguagem *case sensitive* (sensível às letras maiúsculas e minúsculas), é importante ter em atenção a forma como são escritos os comandos.

É importante saber quais os *packages* (ou livrarias) que o *software* possui e para isso utiliza-se o comando >library(), onde nos aparece uma nova janela com todos os *packages* disponíveis. O R possui milhares de *packages* (à data, 22/05/2014, 5566) e que, na instalação inicial do R, apenas alguns destes são instalados.

Uma potencialidade muito útil é a diversidade de formas de ajuda que possui:

- >help();
- >help.start(): dá-nos várias hiperligações, por exemplo manuais, com os mais variados tipos de ajuda. Como primeiro impacto ao *software* é bastante útil;
- >help("function") ou >?function: dá-nos informação acerca da construção de uma função. Podemos também substituir a palavra "func-

tion"por uma outra que represente uma função em concreto e nesse caso, fornece informação específica dessa função;

- >apropos("function") ou >help.search("function"): presta-nos auxílio quando desconhecemos o nome exacto da função pela qual procuramos. Se digitarmos o primeiro comando, conseguimos saber quantos tipos de funções existem daquele género e se digitarmos o segundo comando conseguimos saber em que livrarias é que a função se encontra;

- >RSiteSearch("tópico"): dá-nos uma forte ferramenta de ajuda, como se pode constatar pela Figura 2.2, onde é compilada toda a informação através das *mailing list* e outros documentos.

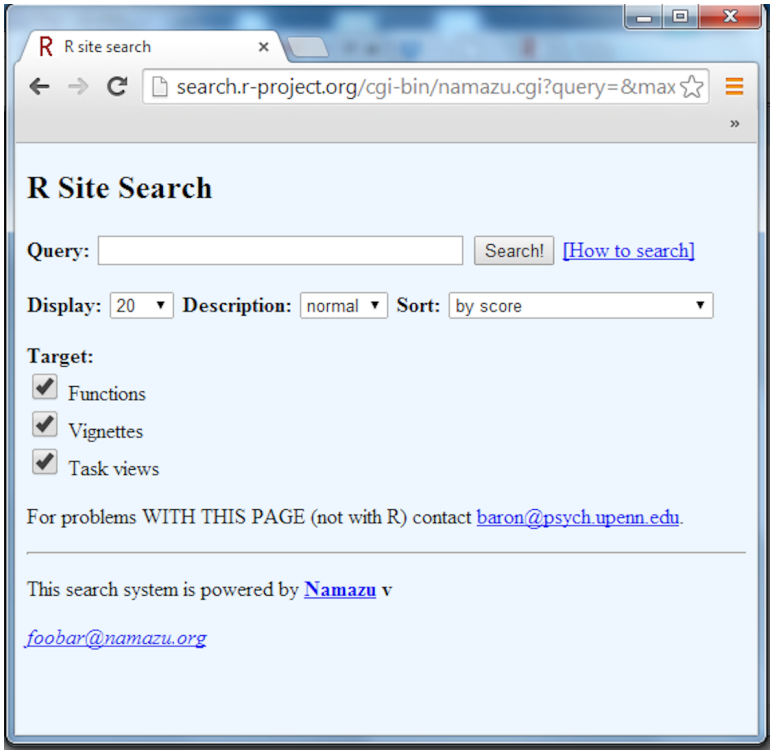

Figura 2.2: Janela de ajuda do comando *RSiteSearch*.

Neste trabalho foi utilizada a versão 3.1.0 do R. Para que todas as funcionalidades estejam operacionais para futuros trabalho a partir deste, é necessário que a versão do R seja igual à mencionada, pois, se for inferior ou superior, nem todas as funções vão estar disponíveis ou apresentam-se de forma diferente.

2.2 Alguns *packages* úteis para a Análise de Sobrevivência

Uma das vantagens deste *software* é o facto desta ferramenta não ocupar muito espaço na memória do computador, pois, todos os seus recursos, bases de dados ou as próprias funções estão disponíveis nos *packages* que têm de ser sempre "carregados" quando os desejamos usar. O comando que se utiliza é >library("nome do *package*") e a partir dele ficam disponíveis todas as funcionalidades desse *package* as quais podem ser consultadas através do comando >help(*package*="nome do *package*") ou então, através do comando >help("tópico"), dando uma informação mais detalhada.

Como existem inúmeros *packages* instalados, para que os possamos usar, é necessário fazer a sua instalação prévia a partir do comando >install.packages ("*package*") e só depois serem carregados.

Os *packages* estão disponíveis no sítio do R, clicando em CRAN no menu do lado esquerdo. Após essa selecção aparece uma lista de países, onde podemos escolher Portugal. De volta ao menu do lado esquerdo, selecionamos *packages* e aparecem duas hiperligações: *Table of available packages, sorted by date of publication* e *Table of available packages, sorted by name*. Se optarmos, pela primeira, podemos verificar que praticamente todos os dias aparecem novas publicações. Por exemplo, só no dia 20 de Maio de 2014 (data da pesquisa), existem 12 novos *packages*.

Visto ser um programa que está em constante actualização, os *packages* também sofrem modificações, e para que não estejamos sempre a instalá-los, pode-se recorrer ao comando >update.packages(). Se apenas for necessário saber qual a versão que foi instalada, utiliza-se >installed.packages().

Em seguida iremos descrever sucintamente alguns *packages* do R importantes para a análise de dados de sobrevivência. Achou-se interessante subdividir esta secção, dando ênfase ao *R Commander* por ser o ponto de partida para os menus e por haver *plug-ins* disponíveis; ao *package survival*, por ser sobre ele que assentam outros *packages* e por fim agrupamos outros que se achou ter relevância para a análise de sobrevivência.

Os *packages* podem sofrer mudanças ao longo do tempo, dando origem a novas versões e consequentemente a estrutura das funções pode mudar, devido a isso é preciso chamar à atenção que todos os *packages* consultados e instalados nesta dissertação datam de Maio de 2014.

2.2.1 *R Commander*

O *R Commander* é um *package* do R que possui menus e caixas de diálogo, desenvolvido por John Fox em 2003 com base no package *tcltk*. Assim, o R passou a ter uma interface mais amigável.

Para instalar o *R Commander* temos de recorrer ao comando: >install.packages("Rcmdr"). Sempre que pretendermos usá-lo, teremos de o carregar, através do comando >library(Rcmdr). A primeira vez que o carregarmos, surge uma janela onde nos é perguntado se queremos instalar aquela lista de *packages* que são necessários à utilização do *R Commander*, se aceitarmos a sua sugestão, temos a facilidade de que os *packages* já ficam disponíveis.

A janela do *R Commander*, é composta por sete partes (assinaladas pelas setas), como mostra a Figura 2.3.

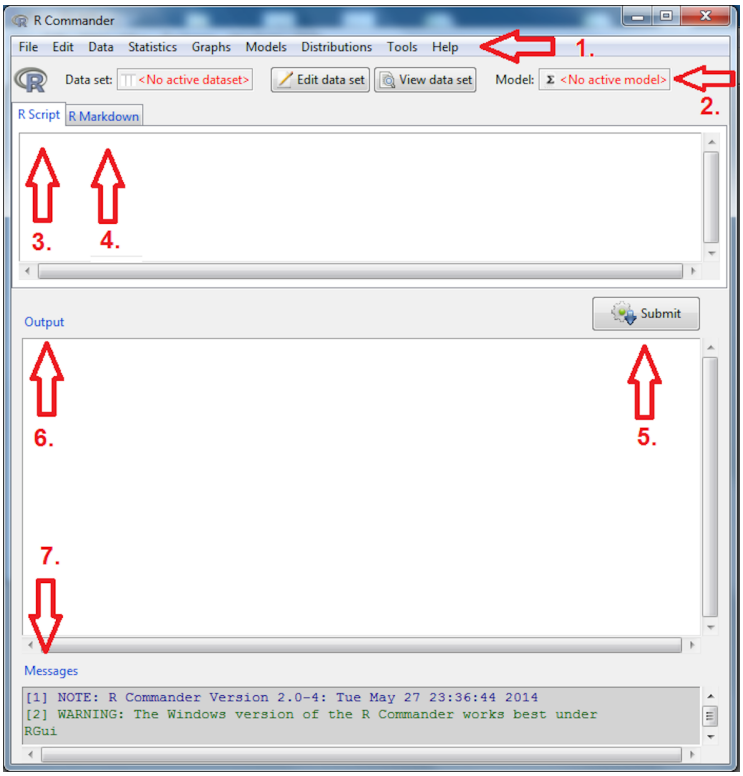

Figura 2.3: Janela do *R Commander*.

Temos assim:

1. menus: onde é possível fazer os habituais procedimentos estatísticos, comuns aos vários *softwares* estatísticos. Se alguma das funções dos

menus estiver a cinzento, isso quer dizer que essa função não está disponível. Isso acontece, ou por não haver dados para executar essas acções, ou por os dados não serem adequados para esses procedimentos;

2. barra de ferramentas: funções relacionadas com a base de dados que está activa. O primeiro item indica qual a base de dados que está activa; o segundo item serve para alterar ou acrescentar algum valor; o terceiro item serve apenas para visualizar a base de dados e o último item indica qual o modelo que está disponível;

3. R *Script*: ao utilizarmos os menus, irá aparecer todo o código que envolveu a operação realizada. Embora existam menus, pode-se introduzir o código manualmente e, nesse caso, tem de ser introduzido nesta janela;

4. R *Markdown*: quando estamos neste separador, aparece o botão *Generate HTML report* e se clicarmos, é gerado um documento numa página de html no *browser*, com o *input* e o *output*. Mais informações sobre este separador, encontram-se no menu *Help*;

5. *Submit*: este botão serve para dar o ok no comando introduzido manualmente. Quando o comando implica mais de uma linha, é preciso selecioná-las todas primeiro e só depois clicar no botão, pois caso contrário, apenas é submetido o comando da linha onde se encontra o cursor;

6. *Output*: todos os comandos introduzidos no R *Script* serão reproduzidos novamente neste espaço (a vermelho), acrescido do resultado que o comando implique (a azul);

7. *Messages*: são reportadas as mensagens de erro (a vermelho), informativas (a azul) ou apenas de aviso (a verde).

Uma componente muito importante é a compilação/obtenção da base de dados. Existem três possibilidades para o fazer:

1. Criação de uma base directamente no *R Commander*: Se seleccionarmos no menu *Data*$\longrightarrow$*New data set...*, aparece uma janela onde podemos introduzir o nome do ficheiro (sem espaços). Uma nova janela com aspecto de uma folha de cálculo, chamada *Data Editor,* fica activa, onde somos livres de introduzir toda a informação que pretendemos. Se quisermos alterar o nome das variáveis, dando dois cliques em cima da mesma, aparece uma janela onde podemos escolher o *type*, *numeric*

(variável quantitativa) ou *character* (variável qualitativa), consoante os dados;

2. Importação a partir dos *packages* disponíveis no R: Se seleccionarmos *Data*⟶*Data in packages*⟶*List data set in packages* iremos encontrar a lista com os *packages* disponíveis, assim como uma breve descrição. Para então importarmos uma base de dados seleccionamos *Data*⟶*Data in packages*⟶*Read data set from an attached package...* e depois escolhemos então o *package* (*car* ou *dataset* ou outro que tenha sido carregado) e depois um dos ficheiros disponíveis;

3. Importação de dados de outros programas de texto: Seleccionamos *Data*⟶*Import data*⟶*from text file, clipboard, or URL...* e aparece-nos uma janela, como mostra a Figura 2.4.

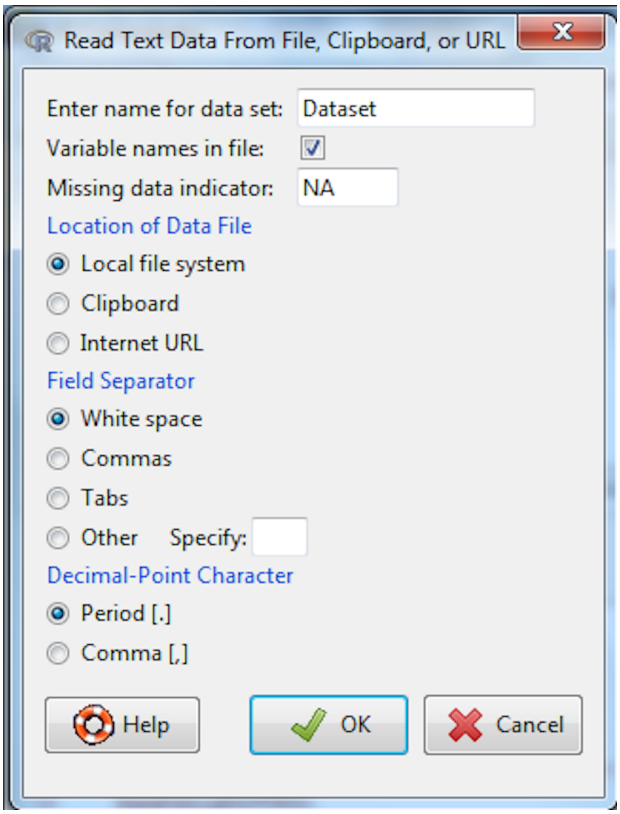

Figura 2.4: Importação de ficheiros de texto, do *clipboard* ou da *internet.*

No item *Missing data indicator* consta *NA* (*Not available*) pois é o código que irá ser atribuído aos valores omissos, caso existam na base de dados. No

item *Location of Data file*, indicamos a localização do ficheiro que queremos importar. No item *Field Separator* indicamos como é feita a separação dos dados. E por fim, no item *Decimal-Point Character* escolhemos a forma como estão separados os valores decimais.

Para o caso de o ficheiro ser em *Excel*, seleccionamos *Data*⟶*Import data*⟶*from Excel, Access or dBase data set...* abre-se uma janela para colocarmos o nome do documento e pressionamos ok, aparece a janela para escolher o directório do documento.

Plug-ins

Para este tipo de análise de dados, de sobrevivência, considerámos importantes três *packages* e que são *plug-ins* do *R Commander*. São eles, o *RcmdrPlugin.EZR*, o *RcmdrPlugin.KMggplot2* e o *RcmdrPlugin.survival*.

Todos estes *plug-ins* precisam de ser instalados para serem usados, da mesma forma que já foi mencionado no início da secção (>install.packages("nome do *plug-in*")) e escolhemos o País que pretendemos. Após a instalação, não há necessidade de carregarmos o *plug-in*, mas a próxima vez que iniciarmos a sessão do *R Commander* precisamos de o fazer, e nesse caso, vamos ao menu *Tools*→*Load Rcmdr plug-in(s)...*→*Plug-in* e seleccionamos o que pretendemos carregar.

- *RcmdrPlugin.EZR*

  O *package* EZR (*Easy* R), adiciona uma variedade de funções estatísticas, incluindo na análise de sobrevivência, nomeadamente, a análise de curvas ROC, meta-análises, cálculo da dimensão da amostra.

  O EZR tem disponível o fácil acesso de apontar e clicar para as funções estatísticas, especialmente para a estatística com aplicação médica. É uma plataforma independente e funciona nos variados sistemas operativos. O manual completo, apenas está disponível em japonês, mas foi publicado um artigo, em 2003, na revista *Bone Marrow Transplantation*, que serve de manual breve, [14].

  Este *package* deve ser instalado no R *Commander*, visto também ser um *plug-in* deste. Associado a este *package* estão outros que devem ser instalados, para isso usa-se o comando >install.packages(pkgs="Rcmdr Plugin.EZR", dependencies=TRUE).

  No caso do EZR, após o carregamento tem de ser reiniciado. Abre então uma nova janela, que irá ser a interface do EZR, que se parece com o R Commander, como podemos verificar pela Figura 2.5.

  Os menus disponíveis são:

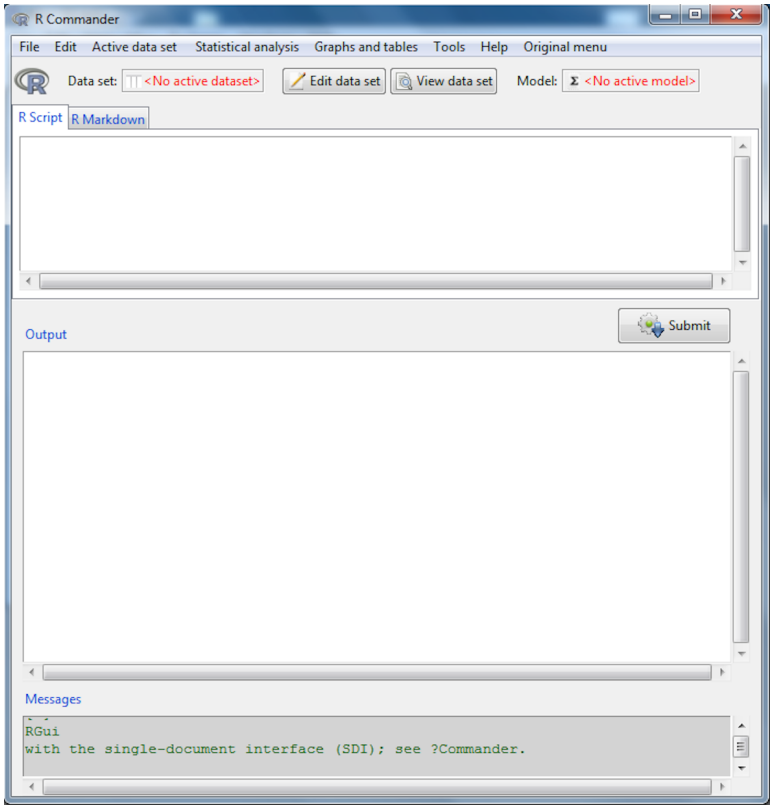

Figura 2.5: Interface do *RcmdrPlugin.EZR*

- *Active data set* - possui funções relacionadas com os dados, desde criar variáveis, a exportar dados;
- *Statistical analysis* - possui funções estatísticas, quer seja para dados discretos ou contínuos, teste não paramétricos ou o cálculo da dimensão de amostras;
- *Graphs and Tables* - consegue-se fazer vários tipos de gráficos e resumos das variáveis das tabelas;
- *Tools* - é a partir deste menu que podemos importar os *packages* ou *plug-in(s)* que já estão instalados;
- *Help* - menu de ajuda, tanto para o *R Commander* como para este *plug-in* (EZR);
- *Original Menu* - dá-nos todos os menus originais do *R Commander*, na mesma disposição que existiam sem termos o *plug-in* activo.

Neste momento, a versão deste *package* é a 1.24.

- *RcmdrPlugin.KMggplot2*

  Este *package* é essencialmente gráfico, útil sobretudo para fazer os gráficos da estimativa de Kaplan-Meier da função de sobrevivência. Os gráficos podem ser mais elaborados pois, para a sua construção estão disponíveis as funcionalidades do *package* ggplot2 [15]. É um *package* recente (23 de Janeiro de 2013) e detém a versão 0.2-0.

- *RcmdrPlugin.survival*

  Este *package* é uma extensão do que já existe no *R Commander*, acrescentando novos itens aos menus já existentes. Existem itens específicos para o modelo de Cox, modelo de regressão paramétrico, estimação de curvas de sobrevivência, juntamente com facilidades no manuseamento dos dados, testa diferenças entre as curvas de sobrevivência e possui uma variedade de testes, diagnósticos e gráficos.

  Nesta data, a versão disponível é a 1.0-4 e foi criada a 17 de Janeiro de 2007. Para mais informações sobre este *package* e as suas funcionalidades, consulte-se [16].

2.2.2 *survival*

Este *package* é o mais importante no domínio da análise de sobrevivência pois serve de base para muitos outros. É uma ferramenta para dados de sobrevivência onde podemos fazer análises descritivas, testes para duas amostras, modelos de tempo de vida acelerado paramétricos, modelo de Cox, conseguimos ter observações censuradas em todos os modelos, intervalos censurados para modelos paramétricos e *Case-cohort designs* (estudo coorte).

Não é necessário que este *package* esteja a ser usado com o *R Commander*, mas é uma clara vantagem se for assim usado, pois o acesso às funções é facilitado através dos menus.

Neste momento, a versão disponível é a 2.37-7. Consultar [17].

Para instalar e correr o *package*, o procedimento é idêntico aos outros.

2.2.3 Outros *packages*

Existem imensos *packages*, tanto para a análise de sobrevivência, como para as mais variadas áreas. Listamos alguns *packages* que se destacaram pelo nome e descrição, de maneira que vamos referir alguns, por ordem alfabética e que achámos interessantes para este tema:

- eha

  Das várias funções que este *package* possui, destaca-se a função coxreg, a qual é uma generalização da função coxph do *package survival*. Permite também o uso de modelos de tempo de vida acelerado com as distribuições de Weibull, Gompertz, log-logística, log-normal e de valores extremos. Possui a versão 2.4-1. Consultar [18].

- *KMsurv*

  Este *package* é, essencialmente, a compilação das bases de dados utilizadas no trabalho de Klein e Moeschberger [19]. É possível obter-se tabelas de mortalidade. Neste momento, possui a versão 0.1-5.

- muhaz

  Este *package* possui funções que permitem obter estimativas da função de risco para dados que possuam censura. Neste momento, a versão disponível é a 1.2.5. Consultar [20].

- *pec*

  Na análise de sobrevivência, um par de indivíduos é designado de concordante se o risco de ocorrer o acontecimento de interesse previsto pelo modelo é inferior para o indivíduo no qual esse acontecimento foi observado mais tarde. A probabilidade de concordância (índice-C) é a frequência de pares concordantes entre todos os pares de indivíduos. Este índice pode ser usado para medir e comparar a potência discriminante entre vários modelos de risco. Este *package* permite o cálculo deste índice na presença de observações censuradas à direita. Neste momento, a versão disponível é a 2.2.9. Consultar [21].

- *prodlim*

  É uma implementação fácil e amigável para estimadores não paramétricos com historial de eventos censurados de análise de sobrevivência. Implementa um algoritmo rápido e alguns recursos que não estão incluídos na função *survfit* (cria curvas de sobrevivência a partir de fórmulas, ou seja, KM, modelo de Cox ajustado previamente, ou modelo acelerado de tempos de falha) do *package survival*. Possui a versão 1.4.3. Consultar [22].

- *relsurv*

  Este *package* é adequado para o cálculo da sobrevivência relativa. Engloba a regressão com modelos aditivos (os mais usuais), modelos

multiplicativos e modelos em que os tempos de vida dos indivíduos são previamente transformados (*transformation models*) [23]. Possui neste momento a versão 2.0-4.

- *riskRegression*

  Este *package* é indicado para modelos de regressão de risco para análise de sobrevivência com e sem riscos competitivos. Possui a versão 0.0.8. Consultar [24].

- *rms*

  Permite a estimação dos parâmetros para uma grande variedade de modelos de regressão, embora tenha sido desenvolvido especialmente para modelos de regressão binários, de Cox, de tempo de vida acelerado, entre outros. Possui neste momento a versão 4.2-0. Consultar [25].

- *simPH*

  Simula e projecta quantidades de interesse (risco relativo, primeiras diferenças, taxa de risco) para coeficientes lineares, interações multiplicativas, polinómios, *splines* penalizados e riscos não proporcionais, bem como curvas de sobrevivência estratificadas a partir do modelo de Cox de riscos proporcionais. Projecta também efeitos marginais para interações multiplicativas. Possui a versão 1.2.1. Consultar [26].

- *smcure*

  Este *package* serve para ajustar modelos semiparamétricos de cura de mistura quer usando o modelo de riscos proporcionais quer o modelo de tempo de vida acelerado. Possui neste momento a versão 2.0. Consultar [27].

  Uma aplicação prática com utilização deste *package* pode ser consultada em [28].

- *survcomp*

  Este *package* tem funções que comparam a qualidade do ajustamento de vários modelos. A instalação deste *package* processa-se de maneira diferente dos outros, recorremos aos comandos:

  > source("http://bioconductor.org/biocLite.R")

  > biocLite("survcomp")

  Para correr o *package*, será feito da forma que já estamos habituados. Possui neste momento a versão 3.1.0. Consultar [29].

- *survMisc*

  Possui essencialmente funções para análise de sobrevivência, fazendo uma extensão do *package survival*. Por exemplo, através da função *lrSS* é possível determinar a dimensão da amostra por forma a detectar diferenças entre as funções de sobrevivência e a função *genSurv* permite gerar amostras aleatórias. Possui a versão 0.4.2. Consultar [30].

- *SurvRegCensCov*

  Este *package* destina-se a permitir a estimação de um modelo de regressão paramétrico em que se usa a distribuição de Weibul para a variável que representa o tempo de vida dos indivíduos. Tem várias novidades que não se encontram noutros *packages* que englobam este modelo, como sejam obter o valor das estimativas nas várias parametrizações do modelo Weibul, permitir considerar uma covariável com informação omissa, na presença de outras com informação completa, entre outras. Possui a versão 1.3. Consultar [31].

Todos estes *packages* funcionam sobre o *R Commander*, ou simplesmente sobre o R. São instalados todos da mesma forma (à excepção do *survcomp*), assim como precisam de ser todos carregados quando se pretende utilizá--los. Para mais informações sobre outros *packages* com funções usuais na análise de sobrevivência, consultar o excelente resumo elaborado por Allignol e Latouche [32]. Existe também um outro trabalho realizado que poderá servir de manual para trabalhar com o R com dados de sobrevivência, [33].

Capítulo 3

Análise de Sobrevivência com o R

O primeiro passo para iniciarmos esta análise consiste em ter uma base de dados. Visto que não possuíamos nenhuma e que, juntamente com o *package survival* são instaladas algumas, optamos por usar a base *colon* que, por ter bastantes dados e variáveis, considerámos ser uma boa aposta para apresentarmos várias situações. Após carregarmos o *package survival* e o *plug-in RcmdrPlugin.survival*, escolhemos a base de dados através do menu *Data→Data in packages→Read data set from an attached package...*, escolhemos o *package survival* e a *Data set colon*. Se em vez de clicarmos em OK, clicarmos em *Help on selected data set*, abre uma janela no *browser* com a informação sobre os dados que escolhemos.

Realmente o R possui uma compilação de base de dados sólida, mas muitas vezes tem pouca informação, ou poucos casos ou até mesmo poucas variáveis. Devido a essas razões, esta base de dados foi cuidadosamente escolhida, pois preenchia todos esses requisitos, acrescentando o facto de que, na informação facultada possuía *links* com os artigos que foram publicados na altura.

Esta base de dados é composta por 929 indivíduos com cancro do cólon, o qual foi classificado no estadio IIIB (T3-T4, N1 e M0) ou IIIC (qualquer T, N2 e M0), onde T representa o tamanho do tumor, N o número de nódulos positivos e M a presença (M1) ou ausência (M0) de metástases.

Os doentes foram classificados em três grupos, consoante o tipo de tratamento adjuvante utilizado no combate ao cancro, ou seja, o grupo de observação, o grupo ao qual foi administrada a toxina *levamisole* e o grupo ao qual foi administrada a combinação da toxina de *levamisole* e *fluorouracil* (*5-FU*).

Os doentes que entraram no estudo, [34], estavam inscritos entre Março

de 1984 e Outubro de 1987. O estudo foi interrompido após uma análise preliminar em Setembro de 1989, quando a combinação das duas toxinas foi considerada altamente eficaz no aumento do tempo de sobrevivência e redução do risco de recorrência do cancro.

De facto, no final do estudo, dos 315 pacientes que pertenciam ao grupo de observação, 155 tiveram recorrência e 114 morreram. No caso da administração da toxina *levamisole*, 310, 144 tiveram recorrência e 109 morreram. No caso da administração da combinação das duas toxinas, 304, 103 tiveram recorrência e 78 morreram.

Cada paciente possui dois registos: o relativo ao tempo até à recorrência e o relativo ao tempo até à morte pelo cancro.

Existem dezasseis variáveis neste estudo:

1. *id* - identificação do indivíduo;
2. *study* - 1 para todos os pacientes;
3. *rx* - tipo de tratamento (Obs - observado/ Lev - administração de *Levamisole*/ Lev+5-FU - administração de *Levamisole* e 5-FU);
4. *sex* - sexo (1 - masculino/ 0 - feminino);
5. *age* - idade (em anos);
6. *obstruct* - obstrução do cólon pelo tumor (1 - obstruído/ 0 - não obstruído);
7. *perfor* - perfuração do cólon (1 - perfurado/ 0 - não perfurado);
8. *adhere* - aderência aos orgãos vizinhos (1 - sim/ 0 - não);
9. *nodes* - número de nódulos linfáticos positivos;
10. *time* - dias até ao evento ou censura;
11. *status* - censura (1 - censurado/ 0 - observado);
12. *differ* - diferenciação do tumor (1 - bom/ 2 - moderado/ 3 - fraco);
13. *extent* - extensão da disseminação local (1 - submucosa/ 2 - muscular/ 3 - serosa/ 4 - estruturas contíguas);
14. *surg* - tempo até cirurgia (0 - pouco/ 1 - muito);
15. *node4* - mais de quatro nódulos linfáticos positivos;

16. *etype* - tipo de evento (1 - recorrência; 2 - morte).

Informação sobre a base de dados pode ser consultada através da ajuda que se encontra disponibilizada quando escolhemos a base de dados (referido anteriormente o procedimento) e o estudo realizado na altura, através de [34].

Visto que já carregámos o *package survival* e o *plug-in RcmdrPlugin.survival* para obtenção da base de dados, já não precisamos de os carregar novamente. Note-se que, se instalarmos o *plug-in RcmdrPlugin.EZR*, algumas das funções deixam de estar activas.

3.1 Análise descritiva

Antes de começarmos a fazer a análise descritiva, dado cada indivíduo possuir dois registos (duas linhas), para conseguirmos trabalhar os dados da melhor forma, é preciso que apenas exista um registo por indivíduo, por isso, temos de "transformar" a base de dados de modo a que só exista um indivíduo por linha. Para tal, a variável relativa ao tempo (*time*) dá origem a duas variáveis: *time1*, onde constará o tempo até à recorrência e *time2*, onde constará o tempo até à morte, as quais são identificadas a partir da variável *etype*. Comparando os tempos destas duas novas variáveis, é possível obter uma variável (*rec*) indicatriz da existência ou não da recorrência: quando os tempos são iguais não há recorrência e quando são diferentes há. Por último, também a variável relativa ao estado (*status*) dá origem a outras duas: *status1*, que nos dará o estado (indivíduo censurado - 1 ou observado - 0), se o acontecimento de interesse for a recaída, ou seja, quando *rec=1*; e *status2* dá-nos o estado quando o acontecimento de interesse é a morte, neste caso, *rec=2*.

Resumindo, passou-se a ter dezoito variáveis, onde as variáveis *time* e *status* dividiram-se em duas, aparece uma nova variável *rec* e a variável *etype* desaparece.

Começamos então por fazer um resumo das variáveis numéricas através do comando *Statistics→Summaries→Numerical summaries* e escolhemos as variáveis *age*, *time1* e *time2*. No separador *Statistics*, podemos escolher os parâmetros de interesse. Neste caso, obtivemos as estatísticas que se podem observar na Figura 3.1.

Conseguimos então saber a média (*mean*), o desvio padrão (*sd*), a amplitude interquartis (*IQR*), os quantis (0%, 25%, 50%, 75% e 100%) e a dimensão da amostra (n). Podemos fazer o mesmo através do comando *Statistical analysis→Continuous variables→Numerical summaries* se o *plug-in RcmdrPlugin.EZR* estiver carregado. Um comando análogo a este é *Statistics→Summaries→Active data set*, mas teremos a estatística descritiva de

```
> numSummary(BaseColon[,c("age", "time1", "time2")], statistics=c("mean",
+   "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
            mean        sd  IQR 0% 25%  50%  75% 100%   n
age     59.75457  11.94889   16 18  53   61   69   85 929
time1 1405.13563 998.90076 1919  8 370 1548 2289 3329 929
time2 1669.95587 872.09896 1558 23 806 1976 2364 3329 929
```

Figura 3.1: Análise descritiva das variáveis numéricas.

todas as variáveis da base de dados, o que não terá muito interesse pois a maior parte das variáveis não são numéricas.

Numa outra perspectiva, podemos fazer as frequências da variável *rx*, tipos de tratamentos, onde ficamos a saber a quantidade de indivíduos (e correspondente percentagem) atribuídos a cada um dos tratamentos, como mostra a Figura 3.2, através do menu *Statistics*→*Summaries*→*Frequency distribution...*.

```
> .Table <- table(BaseColon$rx)
> .Table  # counts for rx
    Lev Lev+5FU     Obs
    310     304     315
> round(100*.Table/sum(.Table), 2)  # percentages for rx
    Lev Lev+5FU     Obs
  33.37   32.72   33.91
> remove(.Table)
```

Figura 3.2: Tabela de frequências e de percentagens para o tratamento.

O primeiro comando cria uma tabela com as categorias da variável *rx*, o segundo, faz as contagens, o terceiro, as percentagens e o quarto remove a tabela pois a função *.Table* foi criada apenas para este contexto, não precisa de ser exibida novamente. Podemos fazer o mesmo através do comando *Statistical analysis*→*Discrete variables*→*Frequency distribution* (se tivermos o *plug-in RcmdrPlugin.EZR* carregado).

3.2 Função de sobrevivência

Antes de estimarmos a função de sobrevivência, precisamos indicar qual a variável que define o tempo e qual a variável que define se o indivíduo tem um tempo censurado ou não (evento). Recorre-se ao menu *Data*→*Survival data*→*Survival data definition...*, gerando as linhas de comando:

```
> attr(BaseColon, "time1") <- "time1"
> attr(BaseColon, "time2") <- "time2"
> attr(BaseColon, "event") <- "status2"
```

definimos desde já as variáveis *time1* e *time2*, pois podemos, mais para a frente, precisar do *time1.* Para a variável do estado, apenas podemos escolher uma de cada vez, neste caso, definimos o *status2* pois, vamos considerar a morte como o evento, em primeira instância pelo menos. Este comando não é estritamente necessário, apenas facilita a escolha das variáveis cada vez que temos de fazer algo novo utilizando as variáveis do tempo e *status*.

Ao executarmos o menu *Statistics→Survival analysis→Estimate survival function...*, obtemos a curva da estimativa de Kaplan-Meier da função de sobrevivência com o respectivo intervalo de confiança. Uma das barreiras que se encontrou ao realizar este gráfico, foi o facto do eixo não ser editável pois, por exemplo, era mais conveniente conseguir ter o eixo em anos. Uma forma de contornar esta questão é criar uma nova variável, mas com a escala que desejamos, neste caso, em anos. Através de *Data→Manage variables in active data set→Compute new variable*, escolhemos a variável a modificar, neste caso *time2*, damos um novo nome, por exemplo, *time2_anos* e na *Expression to compute* colocamos *time*2/365 e depois voltamos a fazer e assim já se obtém a escala em anos. Apesar de não conseguirmos através dos menus do *R Commander* contornar toda esta questão, e uma vez que este *software* é bastante versátil, conseguimos complementar os comandos que obtivemos originalmente. O comando que mostramos de seguida é uma adaptação do que já existe mas com uns melhoramentos, de forma a mostrarmos a estimativa de Kaplan-Meier, tanto com o comando original (primeiro gráfico da Figura 3.4), como as alterações que pretendíamos, onde tivemos de acrescentar, manualmente, a nova função (segundo gráfico da Figura 3.4):

```
.Survfit <- survfit(Surv(time2, status2) ~1, conf.type="log", conf.int=0.95,
type="kaplan-meier", error="greenwood", data=BaseColon)
.Survfit
.Survfit2 <- survfit(Surv(time2_anos, status2) ~1, conf.type="log",
conf.int=0.95, type="kaplan-meier", error="greenwood", data=BaseColon)
.Survfit2
mf<-par(mfrow=c(1,2))
plot(.Survfit, conf.int=TRUE, mark.time=TRUE)
plot(.Survfit2, conf.int=TRUE, mark.time=TRUE, xlab="Tempo (anos)",
ylab="Probabilidade")
par(mf)
remove(.Survfit)
remove(.Survfit2)
```

As primeiras duas linhas servem para construir a função *.Survfit* (*.Survfit* é uma função do *package survival* que nos fornece a estimativa de Kaplan-Meier para a função de sobrevivência) a partir do *time2* e *status2*; a terceira, dá-nos uma estatística descritiva da função definida na linha anterior; a quarta e quinta linha servem para a construção da função *.Survfit2* a partir do *time2_anos* e *status2*; a sexta dá-nos uma estatística descritiva da função definida na linha anterior; a sétima linha cria uma função a partir da qual conseguimos dividir a janela gráfica numa linha e duas colunas, de forma a que os dois gráficos apareçam em simultâneo na mesma janela, fechando este comando, mais abaixo com o comando *par(mf)*. Os comandos começados por *plot*, servem para fazer o gráfico, sendo que o primeiro é referente à função *.Survfit* e o segundo referente à função *.Survfit2* acrescidos com o nome dos novos eixos, que neste caso o tempo será em anos. O comando *remove* serve só para que as funções não apareçam, pois as funções (*.Survfit* e *.Survfit2*) foram criadas apenas para este contexto, não precisam de ser exibidas novamente. Obtém-se o *output* que podemos observar na Figura 3.3 e que dá origem aos gráficos que estão representados na Figura 3.4.

```
> .Survfit <- survfit(Surv(time2, status2) ~1, conf.type="log", conf.int=0.95,
+  type="kaplan-meier", error="greenwood", data=BaseColon)
> .Survfit
Call: survfit(formula = Surv(time2, status2) ~ 1, data = BaseColon,
    conf.type = "log", conf.int = 0.95, type = "kaplan-meier",
    error = "greenwood")
records   n.max n.start  events  median 0.95LCL 0.95UCL
    929     929     929     452    2552    2171      NA
> .Survfit2 <- survfit(Surv(time2_anos, status2) ~1, conf.type="log",
+ conf.int=0.95, type="kaplan-meier", error="greenwood", data=BaseColon)
> .Survfit2
Call: survfit(formula = Surv(time2_anos, status2) ~ 1, data = BaseColon,
    conf.type = "log", conf.int = 0.95, type = "kaplan-meier",
    error = "greenwood")
records   n.max n.start  events  median 0.95LCL 0.95UCL
 929.00  929.00  929.00  452.00    6.99    5.95      NA
> mf<-par(mfrow=c(1,2))
> plot(.Survfit, conf.int=TRUE, mark.time=TRUE)
> plot(.Survfit2, conf.int=TRUE, mark.time=TRUE, xlab="Tempo (anos)",
+ ylab="Probabilidade")
> par(mf)
> remove(.Survfit)
> remove(.Survfit2)
```

Figura 3.3: Comandos e respectivos *outputs* para a estimativa de Kaplan-Meier da função de sobrevivência.

Podemos calcular as diferenças entre as curvas de sobrevivência e assim saber se essas diferenças são significativas de grupo para grupo. Neste caso, a

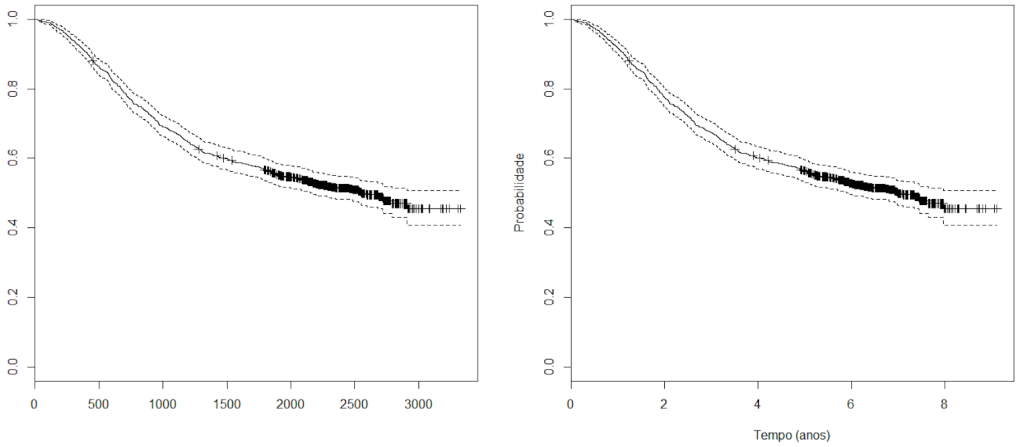

Figura 3.4: Estimativa de Kaplan-Meier para a função de sobrevivência derivada do comando original *versus* Estimativa de Kaplan-Meier para a função de sobrevivência derivada de modificações no comando original.

nossa variável do tipo de tratamento (rx) já está definida como categórica na base de dados original e assim podemos obter a diferença através do menu *Statistics*→*Survival analysis*→*Compare survival functions...* e obtemos o *output*, que se encontra na Figura 3.5.

```
> survdiff(Surv(time2_anos,status2) ~ rx, rho=0, data=BaseColon)
Call:
survdiff(formula = Surv(time2_anos, status2) ~ rx, data = BaseColon,
    rho = 0)

             N Observed Expected (O-E)^2/E (O-E)^2/V
rx=Lev     310      161      146      1.52      2.25
rx=Lev+5FU 304      123      157      7.55     11.62
rx=Obs     315      168      148      2.58      3.85

 Chisq= 11.7  on 2 degrees of freedom, p= 0.0029
```

Figura 3.5: Diferenças entre as curvas de sobrevivência para o tipo de tratamento.

Visto o $p-value$ ser muito inferior a 0.05, podemos concluir que existem diferenças entre os grupos de tratamento. Se quisermos podemos obter as curvas de Kaplan-Meier para cada um dos grupos de tratamento, através do comando anterior, mas com a diferença de que, na opção *Strata*, seleccionamos a variável rx.

Conseguimos averiguar, através da observação da Figura 3.6, que a curva dos doentes que pertencem ao grupo de observação e dos que pertencem ao

grupo onde foi administrado a *Levamisole*, têm formas semelhantes, o que sugere que fazer esse tratamento talvez não seja tão vantajoso. A curva do outro grupo está nitidamente acima das duas anteriores, indicando uma sobrevivência maior para os doentes sujeitos ao tratamento *Levamisole* e 5-FU.

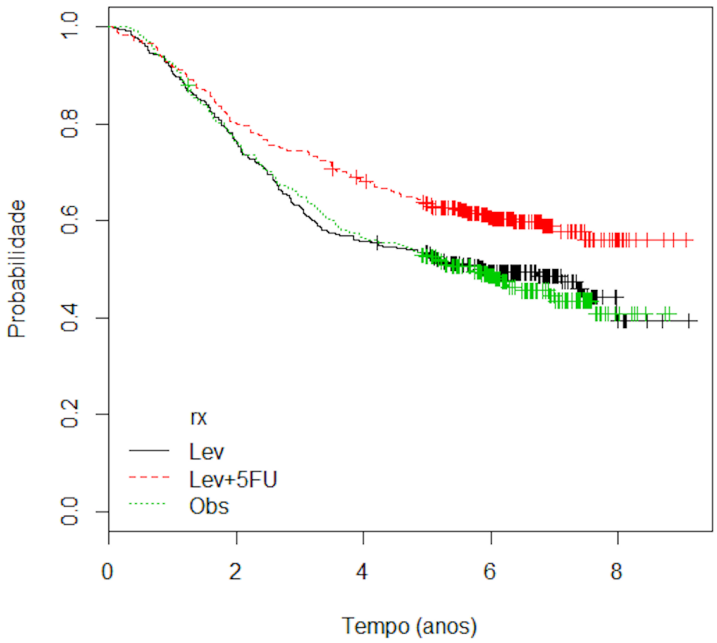

Figura 3.6: Estimativa de Kaplan-Meier para a função de sobrevivência para cada um dos grupos de tratamento.

3.2.1 Algumas variantes

Como já verificámos, ainda existe um leque abrangente de *packages* disponíveis para trabalhar dados de sobrevivência. Por ser um conceito importante, escolheu-se o *package RcmdrPlugin.KMggplot2*, para podermos explorar melhor o estimador de Kaplan-Meier. Já referimos que este *package* tem muito potencial no que diz respeito a gráficos. Vamos apresentar um exemplo.

Depois do carregamento, aparece uma nova janela após aceitarmos a reiniciação, semelhante à do *R Commander*, mas com um novo menu chamado *KMggplot2*. Para obtermos a estimação da função de sobrevivência, usamos os menus *KMggplot2*→*Kaplan-Meier plot...* e aparece uma janela onde pode-

mos escolher a variável do tempo (*Time variable - time2_anos*), o evento (*Event variable - status2*) e podemos escolher estratificar a curva por uma variável (*Stratum variable - rx*). Podemos definir o nome dos eixos, o título da legenda e o título do gráfico. Para além de termos a opção para que o intervalo de confiança (*Confidence interval*) apareça tal como nos *packages* que já mencionámos, as grandes vantagens deste *package* são o facto de no gráfico da função de sobrevivência poder constar o valor do *p-value* correspondente ao teste *log-rank* (*log-rank test*) e uma linha para localizar o valor da mediana do tempo de vida (*Reference line at median survival*). Também possibilita modificar o símbolo dos valores correspondentes a observações censuradas (*Dot censored symbol*). É de salientar que este *package* tem limite reduzido para o número de observações por variável (linhas na base de dados). Por essa razão, optou-se por gerar uma amostra aleatória a partir da nossa base de dados original. Sendo que o máximo de linhas que este *plug- -in* permite é de 80, construiu-se uma amostra com uma linha destinada ao nome das variáveis e mais 79 indivíduos. Claramente é uma barreira, mas a fim de testar as suas potencialidades utilizámos essa amostra e obtivemos as estimativas de Kaplan-Meier para as funções de sobrevivência, no primeiro gráfico, sem estratificar os dados, e no segundo gráfico, estratificando para a variável *rx*, como podemos constatar pela Figura 3.7.

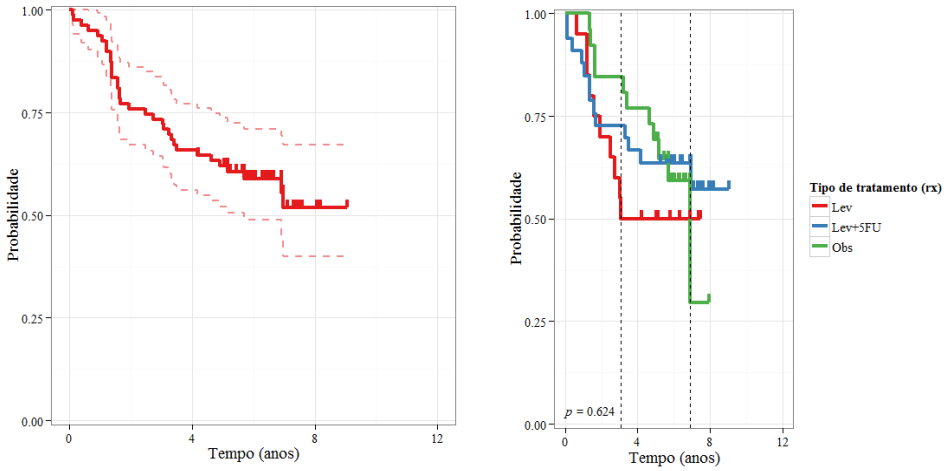

Figura 3.7: Estimativa de Kaplan-Meier para a função de sobrevivência de uma sub-amostra através do *plug-in RcmdrPlugin.KMggplot2*.

Note-se que só aparecem duas linhas para a mediana, uma vez que não é possível ser calculada para o grupo que foi sujeito ao tratamento Lev+5FU.

3.3 Modelos de regressão

A construção de um modelo de regressão é um passo bastante importante, pois é a partir dele que ficamos a conhecer que factores (covariáveis) influenciam o tempo de vida de um indivíduo. Consoante a distribuição que vamos usar para o tempo de vida dos indivíduos, iremos ter um modelo de regressão paramétrico ou não paramétrico.

Em primeiro lugar iremos considerar o modelo de Cox, ou seja, um modelo não paramétrico e, em seguida, dois modelos paramétricos, utilizando as distribuições de Weibull e Log-logística para o tempo de vida dos indivíduos.

3.3.1 Modelo de Cox

Para construirmos este modelo, consideramos que as variáveis são significativas para entrar no modelo se $\alpha = 0.10$.

Numa primeira instância, todas as variáveis entraram no modelo a fim de testar como se comporta o modelo de Cox. Obteve-se os comandos através do menu *Statistics→Fit models→Cox regression model...*.Podemos ver o *output* deste comando na Figura 3.8.

De todas as variáveis que foram introduzidas, podemos destacar as seis que foram significativas: *age* ($p - value = 0.000348$), *extent* ($p - value = 0.032605$), *node4* ($p - value = 0.000317$), *nodes* ($p - value = 0.093575$), *obstruct* ($p - value = 0.024627$) e *rec* ($p - value = 2e - 16$).

Note-se que a variável *nodes* só é significativa se considerarmos um nível de significância de 0.1. Atendendo a que esta variável é muito parecida à variável *node4* (que já se revelou significativa para o modelo) e juntando o facto de que o intervalo de confiança associado conter o valor 1 $(0.9956; 1.058)$, essa covariável não será considerada no modelo. Por outro lado, a variável *rec*, que acrescentámos à base de dados inicial, mostrou ser bastante importante para o modelo, pois não só o seu *p-value* é extremamente pequeno como o seu valor de $\exp(\beta)$ distancia-se muito de 1 (20.2226) e o intervalo de confiança é $(14.5180; 28.169)$.

Repare-se que, uma vez que a variável *rx* tem três categorias, apenas existem duas linhas para esta variável: a que diz respeito ao grupo que foi administrado *Levamisole* e 5-FU e a que diz respeito ao grupo de observação, não sendo visível o grupo que foi administrado apenas *Levamisole*. Isto deve-se ao facto de este grupo ser o de controlo (tratamento padrão), pois, por um lado pretendemos saber se existem diferenças entre os dois tipos de tratamento e se há diferenças entre fazer ou não fazer o tratamento padrão (objectivo do estudo inicial).

O próximo passo consiste em construir o modelo apenas com as covariáveis

```
> CoxModel.1 <- coxph(Surv(time2_anos,status2) ~ adhere + age + differ +
+   extent + node4 + nodes + obstruct + perfor +rec + rx + sex + surg,
+   method="efron", data=BaseColon)

> summary(CoxModel.1)
Call:
coxph(formula = Surv(time2_anos, status2) ~ adhere + age + differ +
    extent + node4 + nodes + obstruct + perfor + rec + rx + sex +
    surg, data = BaseColon, method = "efron")

  n= 888, number of events= 430
   (41 observations deleted due to missingness)
                    coef exp(coef)  se(coef)      z Pr(>|z|)
adhere          0.103418  1.108955  0.130883  0.790 0.429438
age             0.014678  1.014786  0.004104  3.577 0.000348 ***
differ          0.147961  1.159468  0.107488  1.377 0.168654
extent          0.251471  1.285916  0.117679  2.137 0.032605 *
node4           0.526537  1.693060  0.146206  3.601 0.000317 ***
nodes           0.026118  1.026462  0.015576  1.677 0.093575 .
obstruct        0.271957  1.312531  0.121021  2.247 0.024627 *
perfor         -0.213912  0.807420  0.268397 -0.797 0.425452
rec             3.006800 20.222590  0.169092 17.782  < 2e-16 ***
rx[T.Lev+5FU]   0.048010  1.049181  0.126891  0.378 0.705167
rx[T.Obs]      -0.020756  0.979458  0.113926 -0.182 0.855437
sex             0.040493  1.041324  0.098797  0.410 0.681911
surg            0.035155  1.035780  0.107183  0.328 0.742918
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
               exp(coef) exp(-coef) lower .95 upper .95
adhere            1.1090    0.90175    0.8580     1.433
age               1.0148    0.98543    1.0067     1.023
differ            1.1595    0.86246    0.9392     1.431
extent            1.2859    0.77766    1.0210     1.620
node4             1.6931    0.59065    1.2712     2.255
nodes             1.0265    0.97422    0.9956     1.058
obstruct          1.3125    0.76189    1.0354     1.664
perfor            0.8074    1.23851    0.4771     1.366
rec              20.2226    0.04945   14.5180    28.169
rx[T.Lev+5FU]     1.0492    0.95312    0.8182     1.345
rx[T.Obs]         0.9795    1.02097    0.7835     1.225
sex               1.0413    0.96032    0.8580     1.264
surg              1.0358    0.96546    0.8395     1.278

Concordance= 0.83  (se = 0.014 )
Rsquare= 0.547   (max possible= 0.998 )
Likelihood ratio test= 703  on 13 df,   p=0
Wald test            = 410.6  on 13 df,   p=0
Score (logrank) test = 709.5  on 13 df,   p=0
```

Figura 3.8: Modelo de Cox com todas as variáveis da base de dados.

que foram significativas anteriormente e ir retirando uma de cada vez, por forma a verificar se a significância do modelo aumentou ou não com a retirada dessa covariável. Tem-se assim o modelo patente na Figura 3.9.

```
> CoxModel.2 <- coxph(Surv(time2_anos,status2) ~ age + extent + node4
+   +obstruct + rec, method="efron", data=BaseColon)

> summary(CoxModel.2)
Call:
coxph(formula = Surv(time2_anos, status2) ~ age + extent + node4 +
    obstruct + rec, data = BaseColon, method = "efron")

  n= 929, number of events= 452

              coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.013487  1.013579  0.004007  3.366 0.000763 ***
extent    0.255666  1.291321  0.111790  2.287 0.022195 *
node4     0.689864  1.993445  0.098445  7.008 2.42e-12 ***
obstruct  0.275340  1.316978  0.115478  2.384 0.017109 *
rec       3.001904 20.123818  0.163798 18.327  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
age           1.014    0.98660     1.006     1.022
extent        1.291    0.77440     1.037     1.608
node4         1.993    0.50164     1.644     2.418
obstruct      1.317    0.75931     1.050     1.651
rec          20.124    0.04969    14.598    27.742

Concordance= 0.824  (se = 0.014 )
Rsquare= 0.544   (max possible= 0.998 )
Likelihood ratio test= 729.1  on 5 df,   p=0
Wald test            = 417.5  on 5 df,   p=0
Score (logrank) test = 730.7  on 5 df,   p=0
```

Figura 3.9: Modelo de Regressão de Cox apenas com as variáveis significativas.

Agora que ficámos só com as variáveis que tiveram significado para o modelo, precisamos de verificar os pressupostos do modelo de Cox. Para testar a proporcionalidade das funções de risco vamos começar por representar as curvas das estimativas de Kaplan-Meier da função de sobrevivência para cada uma das covariáveis discretas que não se deverão cruzar.

Visto as variáveis serem numéricas, temos de convertê-las em categóricas (ou factores) através do comando *Data→Manage variables in active data set→Convert numeric variables to factors.....* Obtêm-se assim as curvas exibidas na Figura 3.10.

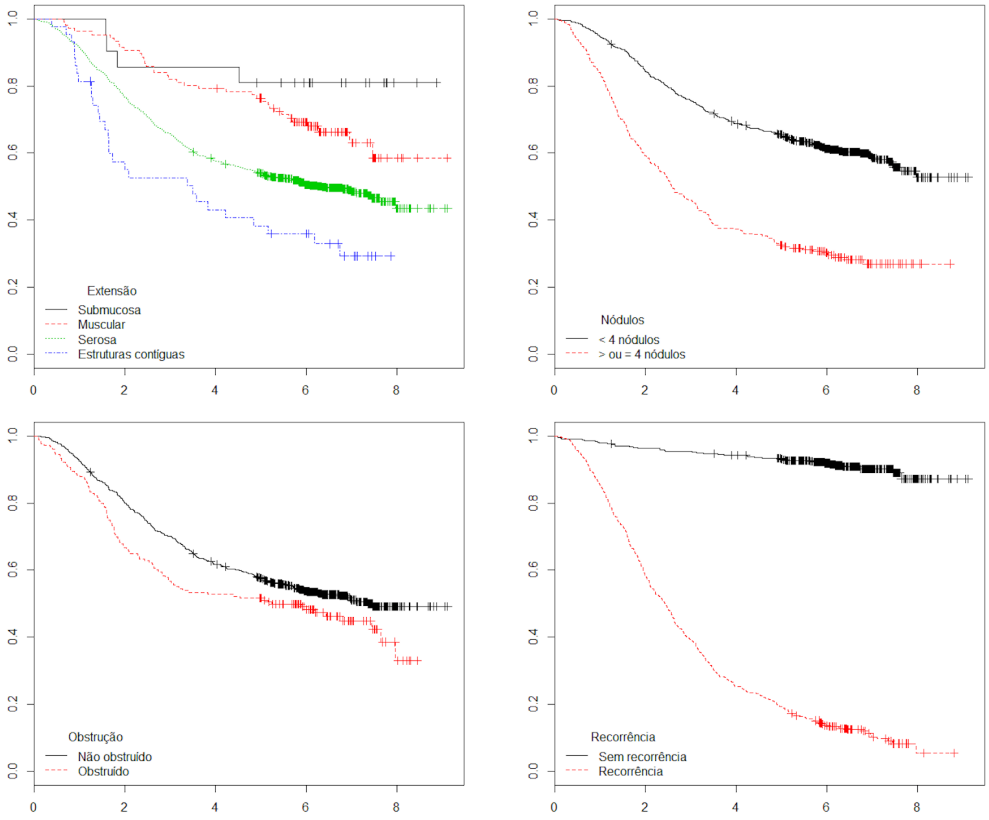

Figura 3.10: Curvas de Kaplan-Meier para as covariáveis significativas no modelo de Cox para testar a proporcionalidade das funções de risco.

Podemos observar que na variável *extent* (Extensão) há cruzamento das curvas de sobrevivência, ao contrário do que sucede nas restantes variáveis, violando ligeiramente o princípio da proporcionalidade das funções de riscos.

Em relação à variável *rec* nota-se que as curvas se vão afastando com o decorrer do tempo, o que também indicia uma possível violação do princípio da proporcionalidade das funções de risco.

Uma forma mais eficaz de verificar a proporcionalidade das funções de risco, consiste na utilização dos resíduos de Schoenfeld. Existe no menu a opção *Models*→*Numerical diagnostics*→*Test proportional hazards*, onde podemos testar se as funções de risco são ou não proporcionais. Os *outputs* gerados pelo comando anterior encontram-se na Figura 3.11, onde ainda é possível obter o valor de R^2 para cada um dos modelos de Cox considerados.

Este comando ajuda também a tomar a decisão da retirada das variáveis no modelo.

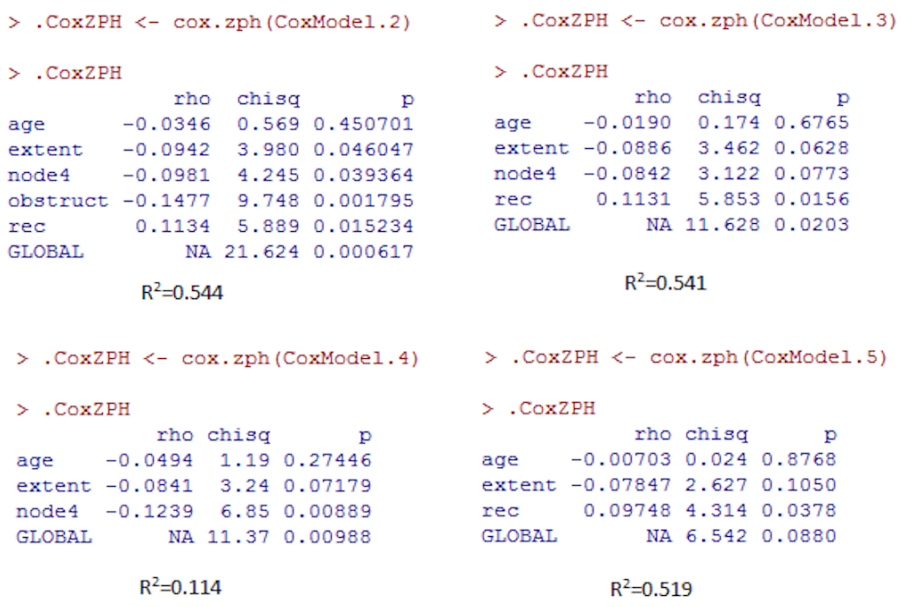

Figura 3.11: *Output* gerado para testar a proporcionalidade das funções de risco dos vários passos para a construção do modelo de Cox com o respectivo coeficiente de determinação.

No primeiro modelo (*CoxModel.2*), observamos que, a um nível de significância de 0.05, o pressuposto de proporcionalidade das funções de risco é violado em todas as variáveis, à excepção da idade (*age*), pois é a única variável em que o *p-value* (p) é superior a 0.05. Para que mais variáveis possam entrar no modelo, vamos considerar para critério de entrada um nível de significância de 0.10 e assim incluem-se as variáveis *extent* e *node4*, para as quais não há violação do pressuposto de riscos proporcionais. Facilmente reparamos que este não é o modelo mais adequado, pois o *p-value Global* é muito pequeno. Verificou-se que o tempo de sobrevivência era explicado em 54.4% ($R^2 = 0.544$).

No modelo *CoxModel.3* retirámos a variável com o *p-value* mais pequeno no teste à proporcionalidade das funções de risco, ou seja, a variável *obstruct*. O modelo continua a não ser o melhor, apesar do aumento do *p-value Global*, que passou para 0.0203 e os dados continuam a ser bem explicados só com estas variáveis ($R^2 = 0.541$).

Para construir o modelo seguinte (*CoxModel.4*), a próxima variável a retirar é a *rec* ($p < 0.05$). O valor de *p-value Global* baixa significativamente, o que implica que essa variável tem interesse para o modelo. E até mesmo

pelo valor de R^2 (0.114) se chega à conclusão que este não é um bom modelo.

Tendo em conta o que significam as variáveis *node4* e *rec*, é de esperar que elas estejam relacionadas. De facto, através do teste de independência do Qui-quadrado, verifica-se que estas variáveis não são independentes (Figura 3.12). Assim, optou-se por retirar a variável *node4* do modelo *CoxModel.3*, do que resultou o modelo *CoxModel.5*.

```
           Recorrência_Sim Recorrência_Não
Node4_Sim              176              79
Node4_Não              285             389
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
        Pearson's Chi-squared test
data:  .Table
X-squared = 52.8958, df = 1, p-value = 3.517e-13
```

Figura 3.12: Teste de independência do Qui-quadrado para testar se as variáveis 4 ou mais nódulos (*node4*) e recorrência (*rec*) são independentes.

Através da observação gráfica dos resíduos de Schoenfeld (Figura 3.13) para estas duas variáveis também se pode concluir que a proporcionalidade das funções de risco não é violada.

De facto, os resíduos têm um padrão aleatório (com algumas, mas poucas, observações isoladas) em torno do zero.

Quanto à variável *rec*, os resíduos que se situam sensivelmente a partir do tempo 3.4 (anos), exibem um padrão crescente (Figura 3.14), o que sugere violação do princípio de proporcionalidade das funções de risco. A acompanhar esta conclusão está o valor obtido do *p-value* desta variável no modelo *CoxModel.5*. Assim sendo, para melhorar o modelo seria preferível considerar a covariável *rec* dependente do tempo, eventualmente com corte nos 3.4 anos.

Então o modelo de Cox final tem a seguinte expressão:

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta_1 age + \beta_2 extent + \beta_3 rec)$$

ou ainda:

$$\frac{h(t;\mathbf{z})}{h_0(t)} = \exp(\beta_1 age + \beta_2 extent + \beta_3 rec)$$

Assim, substituindo os parâmetros pelas respectivas estimativas (Figura 3.15), obtém-se:

$$\frac{h(t;\mathbf{z})}{h_0(t)} = \exp(0.009853 * age + 0.251529 * extent + 3.062296 * rec)$$

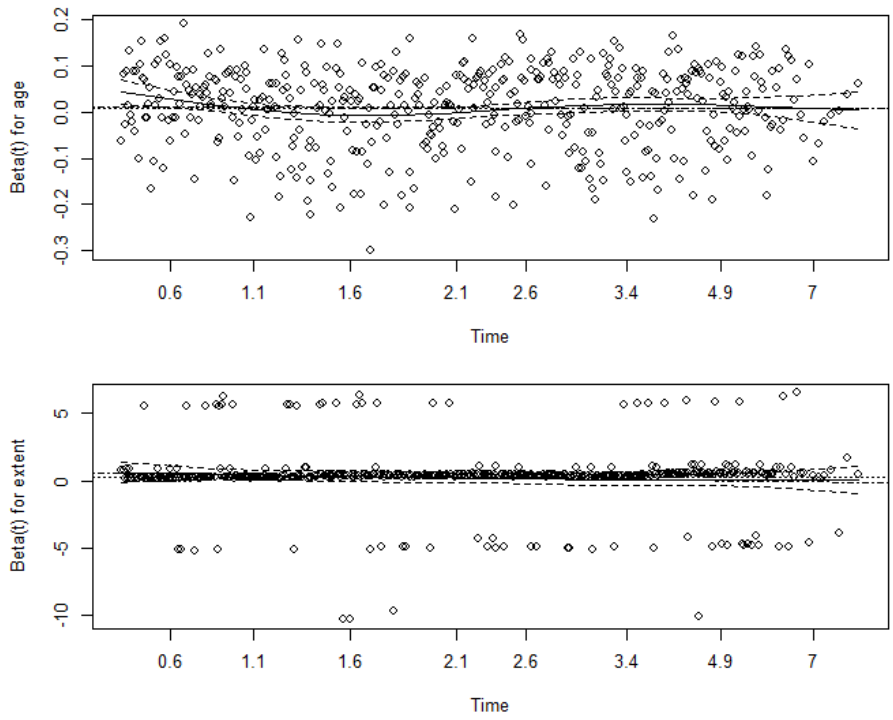

Figura 3.13: Resíduos de Schoenfeld para as variáveis idade (*age*) e extensão do tumor (*extent*).

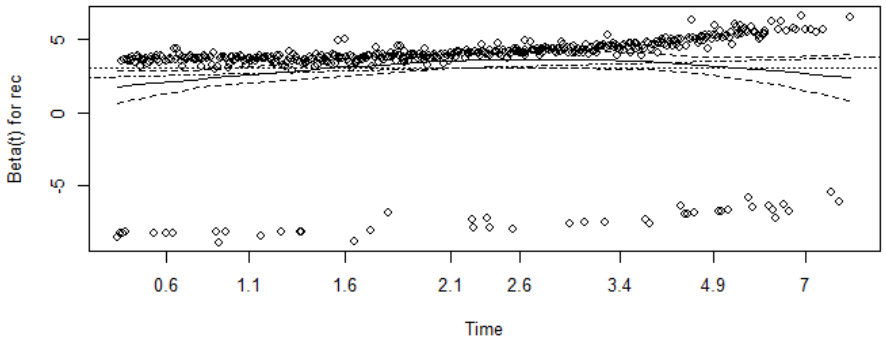

Figura 3.14: Resíduos de Schoenfeld para a variável recorrência (*rec*).

```
> CoxModel.6 <- coxph(Surv(time2_anos,status2) ~ age + extent + rec,
+   method="efron", data=BaseColon)

> summary(CoxModel.6)
Call:
coxph(formula = Surv(time2_anos, status2) ~ age + extent + rec,
    data = BaseColon, method = "efron")

  n= 929, number of events= 452

            coef exp(coef)  se(coef)      z Pr(>|z|)
age     0.009853  1.009901  0.004008  2.458    0.014 *
extent  0.251529  1.285991  0.109017  2.307    0.021 *
rec     3.062296 21.376571  0.163396 18.742   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
age        1.010    0.99020     1.002     1.018
extent     1.286    0.77761     1.039     1.592
rec       21.377    0.04678    15.519    29.446

Concordance= 0.796  (se = 0.014 )
Rsquare= 0.519   (max possible= 0.998 )
Likelihood ratio test= 679.3  on 3 df,   p=0
Wald test            = 366.8  on 3 df,   p=0
Score (logrank) test = 678.9  on 3 df,   p=0
```

Figura 3.15: Modelo de Cox final com as covariáveis idade (*age*), extensão do tumor (*extent*) e recorrência (*rec*).

3.3.2 Modelos paramétricos

Como distribuição para o tempo de vida dos indivíduos, escolhemos as distribuições paramétricas de Weibull e log-logística.

Comecemos então pelo modelo de Weibull. Este modelo pode ser visto como uma alternativa ao modelo de Cox.

Em primeiro lugar, temos de testar a hipótese de riscos proporcionais e isso já foi realizado aquando da construção do modelo de Cox, Figura 3.10. Agora falta-nos testar se o tempo de vida segue esta distribuição. No caso de termos covariáveis dicotómicas, através da expressão 1.9, tem-se que os tempos de vida de um indivíduo padrão seguem uma distribuição de Weibull com parâmetros λ e γ, e, para os restantes indivíduos, o tempo de vida segue uma distribuição de Weibull com parâmetros $\psi\lambda$ e γ, onde $\psi = e^{\beta}$.

Tem de ser verificado se o tempo de vida dos nossos indivíduos segue uma distribuição de Weibull e, para isso, uma possibilidade consiste em fazer a representação gráfica do logaritmo da função de risco cumulativa *versus* o logaritmo do tempo [35]. Se houver um bom ajustamento, a recta com declive $\widehat{\gamma}$ e ordenada na origem $\log \widehat{\lambda}$ deve-se ajustar também aos dados, (Figura 3.16).

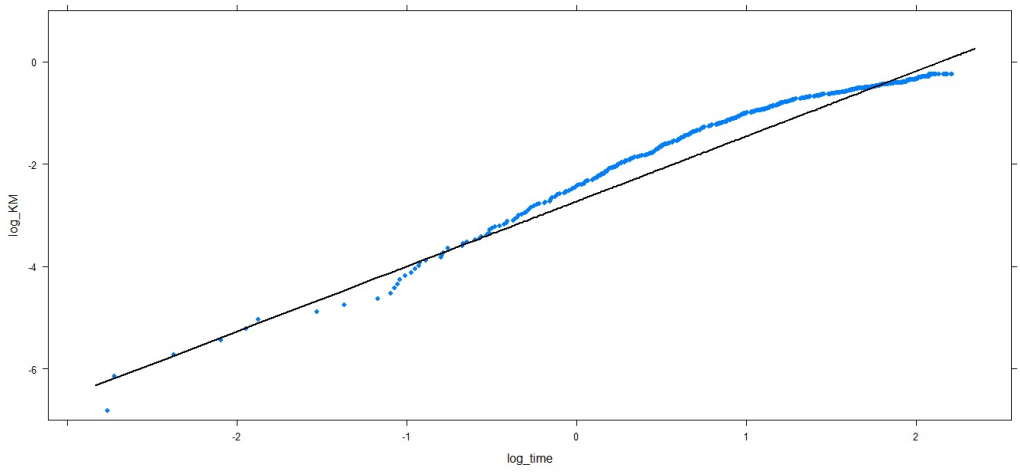

Figura 3.16: Gráfico de $\log\left[-\log \widehat{S}_0(t)\right]$ versus o logaritmo do tempo de vida.

Para a construção deste gráfico, não foi encontrado um *package* que o fizesse de modo "automático", de forma que se construiu por etapas. Primeiro começou-se por estimar os parâmetros do modelo Weibull (Figura 3.17).

```
> summary(SurvregModel.1)

Call:
survreg(formula = Surv(time2_anos, status2) ~ 1, data = BaseColon,
    dist = "weibull")
              Value Std. Error       z     p
(Intercept) 2.241560     0.0528 42.4657 0.000
Log(scale)  0.000856     0.0426  0.0201 0.984

Scale= 1

Weibull distribution
Loglik(model)= -1465   Loglik(intercept only)= -1465
Number of Newton-Raphson Iterations: 5
n= 929
```

Figura 3.17: Modelo de Weibull sem covariáveis.

A estimativa apresentada inicialmente para o parâmetro de escala (*scale*) é 1, mas pelo valor do logaritmo, $\log(scale)$, vemos que o valor exacto tem mais casas decimais. Podemos pedi-las fazendo *Submit* com o nome do modelo gerado (Figura 3.18). Obtemos então $\sigma = 1.000856$ e a ordenada na origem (*Intercept*) $\alpha_0 = 2.241560$.

```
> SurvregModel.1
Call:
survreg(formula = Surv(time2_anos, status2) ~ 1, data = BaseColon,
    dist = "weibull")

Coefficients:
(Intercept)
    2.24156

Scale= 1.000856

Loglik(model)= -1465   Loglik(intercept only)= -1465
n= 929
```

Figura 3.18: Comando que fornece os valores da função especificada, neste caso, a função que gerou o modelo de regressão de Weibull, mas com mais casas decimais.

Como referido em [35], para se obter as estimativas dos parâmetros da recta, utiliza-se a parametrização:

$$\sigma = \frac{1}{\gamma}, \; -\frac{\alpha_0}{\sigma} = \log \lambda$$

Substituindo os parâmetros pelas suas estimativas, obtém-se:

$$\sigma = \frac{1}{\gamma} \Rightarrow \widehat{\gamma} = 0.999144732$$
$$-\frac{\alpha_0}{\sigma} = \log \lambda \Rightarrow \widehat{\lambda} = 0.1064965$$

o que permite fazer a representação da recta.

Uma alternativa mais simples para obter directamente as estimativas consiste em usar a função *ConvertWeibull* do *package SurvRegCensCov* (Figura 3.19).

```
> ConvertWeibull(SurvregModel.1 <- survreg(Surv(time2_anos, status2) ~ 1,
+  dist="weibull", data=BaseKM))
$vars
        Estimate          SE
lambda 0.1064965 0.009128873
gamma  0.9991447 0.042612721
```

Figura 3.19: Obtenção dos parâmetros da recta através da função *ConvertWeibull* do *package SurvRegCensCov*.

Em seguida calculou-se o $\log t$ através do menu *Data*→ *Manage variable in active data set*→ *Compute new variable...*→ Seleccionar a variável

time2_anos, denominar por *log_time* e na *Expression to compute*, escrever log(*time*2_*anos*) e assim obteve-se o logaritmo do tempo.

Por último executou-se o código que se segue:

```
xyplot(log_KM ~log_time, type="p", pch=16,
auto.key=list(border=TRUE), par.settings=simpleTheme(pch=16),
scales=list(x=list(relation='same'), y=list(relation='same')),
data=BaseColon, ylim=c(-7,1))
lines(BaseColon$log_time,BaseColon$log_S, lwd=2)
```

onde *log_KM* indica as estimativas do logaritmo da função de risco cumulativa e *log_S* a variável com os valores para a representação da recta.

A Figura 3.16 permite concluir que o ajustamento à recta nunca é muito satisfatório, indicando que o modelo de regressão de Weibull pode não ser a melhor opção.

De modo a podermos comparar o modelo de Cox com o modelo de Weibull, vamos considerar neste último as mesmas covariáveis obtidas no modelo de Cox final, ou seja, *age*, *extent* e *rec* (Figura 3.20).

```
> SurvregModel.2 <- survreg(Surv(time2_anos, status2) ~ age + extent + rec,
+   dist="weibull", data=BaseColon)
> summary(SurvregModel.2)
Call:
survreg(formula = Surv(time2_anos, status2) ~ age + extent +
    rec, data = BaseColon, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  4.68960    0.33956  13.81 2.20e-43
age         -0.00786    0.00307  -2.56 1.05e-02
extent      -0.19708    0.08269  -2.38 1.71e-02
rec         -2.40588    0.14384 -16.73 8.53e-63
Log(scale)  -0.27030    0.03856  -7.01 2.39e-12

Scale= 0.763
Weibull distribution
Loglik(model)= -1097.1   Loglik(intercept only)= -1465
        Chisq= 735.66 on 3 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 929
```

Figura 3.20: Modelo Weibull com as covariáveis idade (*age*), extensão do tumor (*extent*) e recorrência (*rec*).

Observando os *outputs* do modelo de Cox (Figura 3.15) e do modelo de Weibull (Figura 3.20), podemos constatar que os $p - values$ associados às covariáveis são bastante mais pequenos no caso do modelo de Weibull, o que reforça a preferência por este modelo.

À semelhança do modelo de regressão de Weibull, estimámos o modelo de regressão log-logístico sem covariáveis, Figura 3.21.

```
> SurvregModel.1 <- survreg(Surv(time2_anos, status2) ~ 1, dist="loglogistic",
+    data=BaseColon)
> summary(SurvregModel.1)
Call:
survreg(formula = Surv(time2_anos, status2) ~ 1, data = BaseColon,
    dist = "loglogistic")
             Value Std. Error    z         p
(Intercept)  1.845     0.0546 33.8 1.01e-250
Log(scale)  -0.184     0.0408 -4.5  6.80e-06

Scale= 0.832
Log logistic distribution
Loglik(model)= -1449.6   Loglik(intercept only)= -1449.6
Number of Newton-Raphson Iterations: 4

n= 929
```

Figura 3.21: Modelo de regressão Log-logístico sem covariáveis.

Novamente com o intuito de podermos comparar os dois modelos de regressão paramétricos, vamos considerar o modelo log-logístico com as covariáveis *age*, *extent* e *rec* (Figura 3.22).

```
> SurvregModel.2 <- survreg(Surv(time2_anos, status2) ~ age + extent + rec,
+   dist="loglogistic", data=BaseColon)
> summary(SurvregModel.2)
Call:
survreg(formula = Surv(time2_anos, status2) ~ age + extent +
    rec, data = BaseColon, dist = "loglogistic")
              Value Std. Error      z        p
(Intercept)  4.40303    0.34281  12.84 9.28e-38
age         -0.00938    0.00318  -2.96 3.13e-03
extent      -0.25774    0.08982  -2.87 4.11e-03
rec         -2.21227    0.10858 -20.37 2.79e-92
Log(scale)  -0.61948    0.03989 -15.53 2.22e-54
Scale= 0.538

Log logistic distribution
Loglik(model)= -1092.3   Loglik(intercept only)= -1449.6
        Chisq= 714.59 on 3 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 929
```

Figura 3.22: Modelo de regressão log-logístico com as covariáveis idade (*age*), extensão do tumor (*extent*) e recorrência (*rec*).

Comparando os dois *outputs* das Figuras 3.20 e Figura 3.22, verifica-se que os $p - values$ associados às covariáveis são inferiores no caso do modelo log-logístico, o que significa que este é o melhor modelo, dos analisados, para estes dados.

3.3.3 Algumas variantes

Visto a função de risco ser um conceito de relevância no que toca a dados de análise de sobrevivência, usámos o *package muhaz* [20] para estimá-la, com o seguinte código:

```
data(BaseColon, package="survival")
attach(BaseColon)
fit2 <- muhaz(time2_anos, status2)
plot(fit2)
summary(fit2)
```

Quer a função de sobrevivência (Figura 3.4), quer a função de risco (Figura 3.23) permitem concluir que o risco de morte aumenta progressivamente desde o início do estudo. Mas na segunda é mais evidente que, mais ou menos ao fim de quatro anos, o risco de morte começa a estabilizar, assim como a sobrevivência.

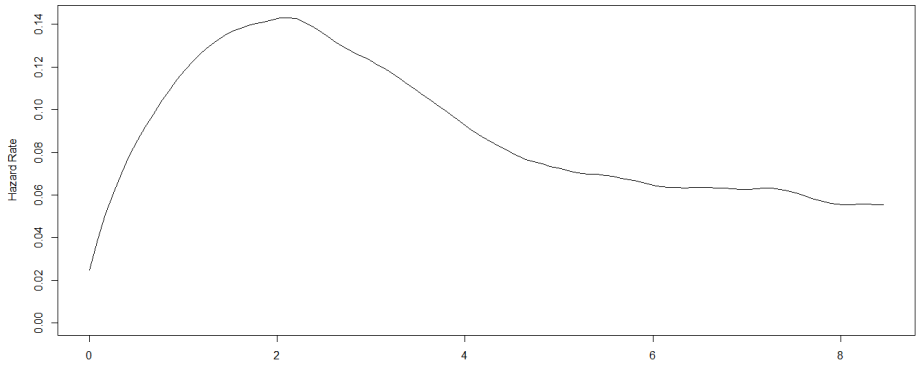

Figura 3.23: Função de risco.

Note-se ainda que o facto de podermos representar graficamente a função de risco e de, neste caso, constatarmos que é unimodal constitui uma mais valia na escolha da distribuição do tempo de vida dos indivíduos. De facto, se inicialmente tivessemos feito esta representação, teríamos logo optado pela distribuição log-logística visto que é uma das distribuições adequadas para este tipo de função de risco, como referimos na subsecção 1.6.3.

Capítulo 4

Conclusão

Nesta dissertação foi lançado o desafio de estudar a análise de sobrevivência num *software* que, à primeira vista, poderá não ser muito convidativo a experimentar. Após uma busca sobre este tema, deparámo-nos com muitos outros trabalhos académicos na mesma área, mas com realização noutros *softwares* mais apelativos, nem que seja devido à existência de menus e por já serem familiares de alguma forma. Todos os programas têm as suas vantagens e desvantagens e a verdade é que, no meio académico, o factor da disponibilidade das licenças gratuitas para os alunos e docentes é uma clara vantagem, mas quando passamos à vida real, deparamo-nos com um factor económico que pode não ser assim tão fácil de contornar. Assim, achou-se que seria bom explorar uma ferramenta que está ao alcance de todos nós, gratuitamente e com a grande vantagem que podemos sempre melhorá-lo.

Optou-se então por estudar a análise de sobrevivência, que foi um tema que nos cativou devido à sua importância na estatística e ao tipo de dados com que trabalha, e que foi fundamental apresentá-lo no primeiro capítulo, dando a conhecer os conceitos básicos, conceitos novos, tendo como objectivo a construção de modelos de regressão e optou-se por dar a conhecer melhor o *software* de estatística R, que o apresentámos no segundo capítulo, dando a conhecer a sua origem, a sua evolução e como se encontra no momento actual e que, sem estes conceitos, seria difícil mostrar a sua aplicabilidade.

Encontram-se artigos das mais variadas partes do mundo com a utilização do R, mas de Portugal pouca coisa existe, o que sugere que esta ferramenta é ainda pouco usada.

Como já vimos no decorrer desta dissertação, este *software* apresenta vantagens e desvantagens. Em termos de vantagens, destacam-se três que são realmente muito importantes: o facto de ser gratuito; o facto de ser de código aberto, onde podemos criar as nossas próprias funções, modelar as que já existem às novas situações, fazendo do R, uma ferramenta muito

versátil e o facto de ser um programa que ocupa pouco espaço na memória do computador, podendo ser uma vantagem para alguns utilizadores. Em termos de desvantagens, a principal é a de termos de saber a linguagem para conseguir trabalhar mas, com a introdução do *package Rcmdr*, o R tornou-se um ambiente mais aceitável, até amigável e quebrou-se a barreira que inicialmente se pôs.

Os *packages* são claramente uma vantagem, como já foi referido, existem muitos, mas uma das dificuldades encontradas, foi na procura de *packages* que tivessem funções que fossem possíveis usarmos na nossa análise. A pesquisa foi extensa. Alguns, por terem nomes sugestivos, induzem o acesso à sua informação, mas o utilizador perde-se na lista exaustiva de nomes. Para nossa sorte, existe uma *task view* no sítio do R, exclusivamente para a análise de sobrevivência onde, para além de nomear os *packages*, especifica as funções e o que elas fazem (ver [32]). À data de 14 de Agosto de 2014, existem 33 *task view* e têm todas uma data de elaboração relativamente recente; a mais antiga com mais ou menos um ano e a mais recente apenas com sensivelmente um mês, que é de sobrevivência. Juntamente com as *task view*, no fim da página do R, é apresentado um *package*, *ctv*, que se o instalarmos, e o correrrmos, o R instala todos os *packages* que estão disponíveis nessa *task view*, o que realmente apresenta uma clara vantagem, pois não precisamos de o fazer um a um, mas precisamos de dispender algum tempo, pois, como é de esperar, é muita informação a carregar.

Mais uma vez o R tornou-se útil neste trabalho, pois não possuíamos nenhuma base de dados para trabalhar. Existe uma diversidade de base de dados e após debruçarmo-nos sobre elas, encontramos uma que realmente era muito boa, pois possuía muitos indivíduos e muitas variáveis que podíamos explorar. O primeiro problema surge quando a informação disponibilizada é pouca na parte que descreve as variáveis, o que suscitou algumas dúvidas, mas ultrapassando esse problema, encontrou-se outro, achou-se mais prático trabalhar a base no *Excel* primeiro e então depois exportar para o R, pois a divisão da base em dois, devido ao facto de haver duas linhas para cada indivíduo, e depois a junção, foi impraticável.

O *R Commander* foi realmente uma grande ajuda para trabalhar os dados, não foi preciso perder muito tempo com a aprendizagem da linguagem, pois os menus estão bastante completos e toda a programação complementar aos códigos gerados foram coisas simples e que facilmente se encontra na documentação.

Foi no terceiro capítulo que se fez a ligação dos conhecimentos que adquirimos no primeiro e segundo capítulo, sendo um capítulo mais prático, de aplicação directa dos conhecimentos, onde realmente pudemos mostrar as vantagens e desvantagens do R. É de destacar que tivemos a necessidade

de incorporar uma subsecção chamada outras variantes por ser algo não necessário à análise de sobreviência "básica", mas como uma alternativa ao que já tinhamos, numa versão melhorada ou até mesmo diferentes coisas que se pode aplicar nessa matéria.

Com o R conseguiu-se fazer um estudo de uma base de dados de sobrevivência com alguma profundidade, tão bom como se tivessemos utilizado qualquer outro *software* estatístico, porventura até mesmo uma versão mais elaborada, pois conseguiu-se tirar partido da sua versatilidade e isto foi possível de realizar-se porque o R possui um sistema de ajuda muito bom.

É sem dúvida um *software* recomendável, pelo menos para análise de sobrevivência.

Bibliografia

[1] Marubini, E., Valsecchi, M.G. (1995) - *Analysing Survival Data from Clinical Trials and Observational Studies*. New York: Wiley.

[2] Kaplan, E.L., Meier, P. (1958) - Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

[3] Collett, D. (2003) - *Modelling Survival Data in Medical Research.* 2nd edition, Chapman & Hall/CRC, Boca Raton.

[4] Cox, D.R. (1972) - Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* Series B **34**, 187-220.

[5] Rocha C., Papoila A.L. (2009) - *Análise de Sobrevivência*, XVII Congresso da Sociedade Portuguesa de Estatística SPE.

[6] Andersen, P.K., Gill, R.D. (1982) - Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, **10**, 1100-1120.

[7] Peto, R., Peto, J. (1972) - Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society*, Series A, **135**, 185-206.

[8] Breslow, N.E. (1970) - A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, **57**, 579-594.

[9] Kalbfleisch, J.D., Prentice, R.L. (1973) - Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60, 267-278.

[10] Efron, B. (1977) - The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557-565.

[11] Schoenfeld, D.A. (1982) - Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

[12] Grambsch, P.M., Therneau, T.M. (1994) - Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

[13] Gentleman, R., Ihaka, R. (1997) - *The R Project for Statistical Computing*, University of Auckland. URL: http://www.r-project.org/.

[14] Kanda, Y. (2013) - Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplantation*, **48**, 452-458.

[15] Wickham, H. (2009) - *ggplot2: elegant graphics for data analysis*. Springer, New York.

[16] Fox, J., Sá Carvalho, M., (2012) - The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis. *Journal of Statistical Software*, **49** (7), 1-32.

[17] http://cran.dcc.fc.up.pt/web/packages/survival/survival.pdf. Consultado a 28/07/2014.

[18] http://cran.dcc.fc.up.pt/web/packages/eha/eha.pdf. Consultado a 28/07/2014.

[19] Klein, J.P., Moeschberger, M.L. (1997) - *Survival Analysis Techniques for Censored and Truncated Data*, Springer.

[20] http://cran.dcc.fc.up.pt/web/packages/muhaz/muhaz.pdf. Consultado a 28/07/2014.

[21] Mogensen, U.B., Ishwaran H., Gerds, T.A. (2012) - Eva- luating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, **50** (11), 1-23.

[22] http://cran.dcc.fc.up.pt/web/packages/prodlim/prodlim.pdf. Consultado a 28/07/2014.

[23] Pohar M., Starde J. (2006) - Relative survival analysis in R. *Computer methods and programs in biomedicine,* **81,** 272–278.

[24] http://cran.dcc.fc.up.pt/web/packages/riskRegression/riskRegression.pdf. Consultado a 28/07/2014.

[25] http://biostat.mc.vanderbilt.edu/rms. Consultado a 28/07/2014.

[26] http://christophergandrud.github.io/simPH/. Consultado a 28/07/2014.

[27] http://cran.dcc.fc.up.pt/web/packages/smcure/smcure.pdf. Consultado a 28/07/2014.

[28] Alves, A.C. (2012) - *Modelos de Cura: Aplicação ao Cancro da Mama Feminino* Dissertação de Mestrado. Centro de Competência de Ciências Exactas e da Engenharia. Universidade da Madeira.

[29] Schroeder, M.S., Culhane, A.C., Quackenbush, J., Haibe-Kains, B. (2011) - survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, **27** (22), 3206-3208.

[30] http://cran.dcc.fc.up.pt/web/packages/survMisc/survMisc.pdf. Consultado a 28/07/2014.

[31] http://cran.dcc.fc.up.pt/web/packages/SurvRegCensCov/SurvRegCensCov.pdf. Consultado a 28/07/2014.

[32] Allignol A., Latouche A., Task view: Survival Analysis. URL: http://cran.r-project.org/web/views/Survival.html. Consultado a 16/06/2014.

[33] Peña, R.E.B. (2005) *Análisis de Sobrevivencia utilizando el lenguaje R,* Simposio de Estadística, Paipa, Boyacá, Colombia.

[34] Lin. D.Y. (1994) - Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, **13**, 2233-2247.

[35] Abreu, A.M. (1997) - *Modelos de Sobrevivência para Populações Heterogéneas*. Dissertação de mestrado. Departamento de Estatística e Investigação Operacional. Faculdade de Ciências da Universidade de Lisboa.