View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Repositório Digital da Universidade da Madeira

DM

Uma Incursão pela Sobrevivência Relativa

DISSERTAÇÃO DE MESTRADO

Tatiana Filipa Fernandes Temtem Nunes

MESTRADO EM MATEMÁTICA

UNIVERSIDADE da MADEIRA

A Nossa Universidade

www.uma.pt

dezembro | 2017

Uma Incursão pela Sobrevivência Relativa

DISSERTAÇÃO DE MESTRADO

Tatiana Filipa Fernandes Temtem Nunes

MESTRADO EM MATEMÁTICA

ORIENTADORA

Ana Maria Cortesão Pais Figueira da Silva Abreu

Uma Incursão pela Sobrevivência Relativa

DISSERTAÇÃO DE MESTRADO

Tatiana Filipa Fernandes Temtem Nunes

MESTRADO EM MATEMÁTICA

JÚRI

Maribel Gomes Gonçalves Gordon
Rita Maria César e Sá Fernandes de Vasconcelos
Ana Maria Cortesão Pais Figueira da Silva Abreu

"Para ser grande, sê inteiro.
Nada teu exagera ou exclui.
Sê todo em cada coisa.
Põe quanto és no mínimo que fazes.
Assim em cada lago a lua toda brilha porque alta vive."

Fernando Pessoa

Agradecimentos

Este espaço foi reservado para agradecer a todos aqueles que me apoiaram e ajudaram a concretizar mais uma etapa da minha vida.

Aos meus pais, pelo amor, apoio, carinho, confiança, força, persistência, sacrifício e por tudo aquilo que me ofereceram e ensinaram que me ajudou a construir a pessoa que sou hoje. Um agradecimento especial à minha mãe, por seres o meu grande alicerce na minha vida, pelo incentivo e inspiração que me incutiste desde a minha infância a lutar pelos objetivos e sonhos, e por ensinares a nunca desistir independentemente do grau de dificuldade ou exigência.

À minha orientadora, Professora Doutora Ana Maria Abreu, um especial e enorme agradecimento, pelo apoio, compreensão, dedicação, disponibilidade como me acompanhou nesta caminhada e ainda pela oportunidade que me concedeu para realizar este estudo na área da Análise de Sobrevivência.

Aos meus irmãos, pelo apoio, incentivo, confiança e por acreditarem sempre em mim, e ainda pela compreensão nos momentos em que não pude estar presente.

À Esmeralda Faria pela amizade e apoio.

Aos docentes do Departamento de Matemática, da Faculdade de Ciências Exatas e da Engenharias, pelo conhecimento transmitido.

A todos os outros familiares, amigos e colegas, que me incentivaram e que contribuíram para a pessoa que sou hoje, pelos conselhos transmitidos e pelos momentos vividos.

Aos restantes que, direta ou indiretamente, contribuíram para a concretização desta etapa tão importante.

A TODOS, O MEU MUITO OBRIGADO!!

Resumo

O principal objetivo desta dissertação é dar a conhecer a Análise de Sobrevivência Relativa, que é uma subárea da Análise de Sobrevivência muito utilizada em Oncologia.

Na Análise de Sobrevivência estuda-se o tempo de vida entre o instante inicial e o acontecimento de interesse do estudo, sendo este último, frequentemente, a morte por causa da doença. Porém, o que acontece muitas vezes é que a causa de morte é desconhecida ou a informação na certidão de óbito não especifica se foi ou não devido à doença em estudo, sendo então necessário recorrer à Análise de Sobrevivência Relativa. A diferença nesta última abordagem reside no facto de, na estimação do tempo de vida, não ser necessário informação sobre a causa de morte.

Assim, iniciámos esta dissertação com alguns conceitos essenciais da Análise de Sobrevivência, depois apresentámos a Análise da Sobrevivência Relativa e os correspondentes modelos de regressão. Através do programa estatístico R e de uma das suas bases de dados, exemplificamos os procedimentos habituais nesta área. A base de dados referia-se a 5971 observações relativas a doentes diagnosticados com cancro do colón ou do reto entre 1994 e 2000 e a tábua de mortalidade utilizada foi referente a Portugal. Consideramos um período de *follow-up* de dez anos.

Palavras-chave: Análise de Sobrevivência, Análise de Sobrevivência Relativa, modelos de regressão, programa estatístico R.

Abstract

The main goal of this dissertation is to present the fundamental concepts of Relative Survival Analysis, which is a subarea of Survival Analysis often used in Oncology.

In Survival Analysis we study the lifetime between an initial time and the event of interest of the study, the latter being often death due to the disease. However, what often happens is that the cause of death is unknown or the information on therefore death certificate does not specify if it was due to the disease under study and it is necessary to use the Relative Survival Analysis. The difference in this latter approach lies in the fact that, in estimation of the lifetime, no information on the cause of death is required.

Thus, we started this dissertation with some essential concepts of the Survival Analysis, then we presented the Relative Survival Analysis and the corresponding regression models. Through the statistical program R and one of its databases, we exemplify the usual procedures in this area. The database referred to 5971 observations concerning patients diagnosed with colon or reto cancer between 1994 and 2000 and the mortality table used was from Portugal. We considered a follow-up period of ten years.

Key words: Survival Analysis, Relative Survival Analysis, regression models, statistical program R.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Análise de Sobrevivência

1.1 Introdução

A Análise de Sobrevivência surge na área da Estatística direcionada para a Biologia e Medicina. Contudo, tem aplicações em diversas outras áreas, nomeadamente nas Ciências Sociais, nas Económicas e nas Engenharias [1].

A Análise de Sobrevivência engloba métodos e modelos que permitem analisar estatisticamente dados de sobrevivência [42]. O tempo de vida é uma variável que permite registar o tempo de sobrevivência, nomeadamente o tempo até ocorrer o acontecimento interesse ou falha. Este tempo inicia quando o indivíduo é selecionado para o estudo e termina quando é observado o acontecimento de interesse ou quando ocorre a censura e pode ser medido em anos, meses, semanas ou dias.

O acontecimento de interesse do estudo é definido, inicialmente, pelo investigador e pode ser: morte, ocorrência da doença, recuperação ou recaída durante o tratamento. Contudo, estes exemplos são direcionados para a área da medicina, nomeadamente estudar a eficácia de um ou mais tratamentos nos indivíduos com determinada patologia. No caso de ser um estudo direcionado para a área das ciências económicas pode ser, por exemplo, a duração que um país se encontra numa situação de depressão económica, ou tempo que uma população jovem, após terminar terminar o ensino superior, fica no desemprego.

1.2 Função de Sobrevivência

Seja T uma variável aleatória não negativa, que representa o tempo de sobrevivência e que varia de 0 a $+\infty$.

A função de sobrevivência é denotada por $S(t)$ e indica a probabilidade da variável aleatória T ultrapassar o tempo t, ou seja, indica a probabilidade do tempo de vida do indivíduo ser maior que t [11, 13].

Assim sendo, esta função é dada por

$$S(t) = P(T > t), \quad t \geq 0 \tag{1.1}$$

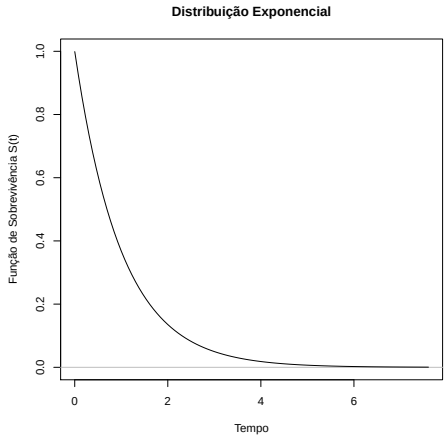

Figura 1.1: Curva da função de sobrevivência.

A Figura 1.1 é um exemplo desta função e foi gerada no programa estatístico R, através do *package R Commander* de uma amostra de dimensão 100 da distribuição exponencial de parâmetro 1.

As propriedades da função de sobrevivência são (ver Figura 1.1)

- função monótona decrescente e contínua a esquerda;
- $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to +\infty$.

A função de sobrevivência pode ser obtida através da função de distribuição $F(t)$, pois como $F(t) = P(T \leq t)$ então $S(t) = 1 - F(t)$.

1.3 Função de Risco

A função de risco (*hazard function*), também é conhecida como taxa ou função de incidência, taxa de falha, força de mortalidade ou força de mortalidade condicional [11, 13].

A função de risco associada ao acontecimento em estudo, é dada com base na taxa instantânea de ocorrência do acontecimento de interesse, ou seja, é definida por:

$$h(t) = \lim_{\epsilon \to 0} \frac{P[(t \leq T < t + \epsilon) \mid T \geq t]}{\epsilon}, \quad t \geq 0 \tag{1.2}$$

Assim, a função de risco representa o risco momentâneo da ocorrência do acontecimento de interesse, sabendo que o evento em estudo ainda não ocorreu.

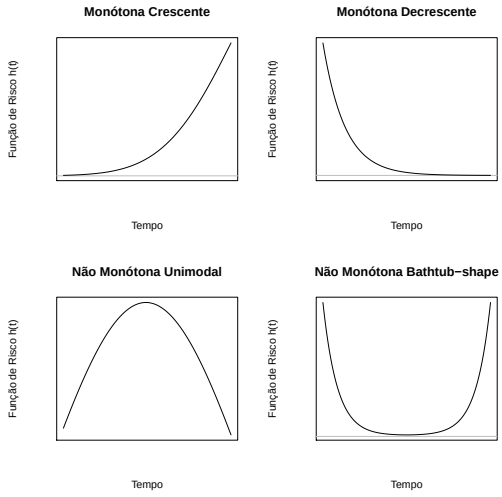

Figura 1.2: Funções de Risco.

A função de risco goza das seguintes propriedades:

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

Note-se que esta função pode tomar qualquer valor real maior ou igual a zero, ou seja, não está limitada ao intervalo $[0, 1]$.

A função de risco pode ter várias formas (ver Figura 1.2) pois pode ser monótona (crescente, decrescente ou constante) ou não monótona (*bathtub-shaped*; unimodal). Contudo, está sempre associada a uma função de sobrevivência decrescente.

A função de sobrevivência e a função de risco são inversamente proporcionais, no sentido em que quando a probabilidade de sobrevivência aumenta o risco diminui e vice-versa.

A função densidade de probabilidade da variável aleatória T está relacionada com a função de sobrevivência da seguinte forma:

$$f(t) = -S'(t) = \lim_{\epsilon \to 0} \frac{P(t \leq T < t + \epsilon)}{\epsilon} \tag{1.3}$$

Podemos constatar algumas relações entre a função de risco, a função densidade de probabilidade e a função de sobrevivência, como sejam:

$$h(t) = \frac{f(t)}{S(t)} \tag{1.4}$$

$$S(t) = \exp(-\int_0^t h(x)dx) \tag{1.5}$$

$$f(t) = h(t)\exp(-\int_0^t h(x)dx) \tag{1.6}$$

A função de risco cumulativa define-se da forma que se segue:

$$H(t) = \int_0^t h(x)dx, \quad t \geq 0 \tag{1.7}$$

de onde se deduz que

$$S(t) = \exp(-H(t)) \Leftrightarrow H(t) = -\log S(t) \tag{1.8}$$

Portanto, $H(t)$ é uma função não negativa, monótona crescente e mede o risco de ocorrência do acontecimento de interesse até ao momento t.

1.4 Mecanismos de Censura

A censura verifica-se quando não sabemos exatamente o tempo de sobrevivência, ou seja, quando não é possível observar o acontecimento de interesse durante o decorrer do estudo. Neste caso o verdadeiro tempo de vida é desconhecido, existindo apenas uma informação parcial [11, 33].

Desta forma, a censura é verificada quando:

- o estudo termina, mas o acontecimento de interesse não é observado nesse indivíduo;
- é perdido o contato com o indivíduo durante o decorrer do estudo (perdido para o *follow-up*);
- ocorre a morte do indivíduo, mas não relacionada com o evento de interesse;
- o indivíduo sai do estudo, devido a uma reação adversa ou por risco de morte, mas não relacionado com o estudo.

Existem vários tipos de censura, que serão descritos em seguida.

Censura à direita

A censura à direita ocorre quando o tempo de observação é inferior ao tempo de sobrevivência, ou seja, o tempo até à ocorrência do acontecimento de interesse é um valor desconhecido mas maior do que o tempo de observação. No entanto, o tempo deve ser registado, apesar do acontecimento em estudo não ter ocorrido no período do estudo. O tempo de vida é $t_0 + t$, onde t_0 indica o instante em que o indivíduo entrou no estudo e t é um valor desconhecido, contudo superior ao tempo em que o indivíduo está em observação. Se a censura ocorrer em $t_0 + a$, sendo $a < t$, então a indica um tempo de vida censurado.

Este tipo de censura surge quando o indivíduo: sobrevive até ao fim do estudo; é perdido para o *follow-up* ou morre por outra causa diferente ao estudo.

Um exemplo de censura à direita: quando queremos estudar o tempo de vida de um grupo de indivíduos de uma faixa etária onde foi diagnosticado cancro pulmonar, num determinado período de estudo. As observações correspondentes aos indivíduos que sobreviverem para além do período estipulado para o estudo, os indivíduos que morreram por outra causa e não devido ao cancro pulmonar ou os que forem perdidos para o *follow-up*, são observações censuradas à direita.

Censura à esquerda

Se o tempo de sobrevivência (desconhecido) for inferior ao tempo de observação do indivíduo significa que existe censura à esquerda. Tal situação surge quando o acontecimento de interesse ocorreu num instante anterior ao tempo de observação. A censura à esquerda é menos frequente do que a censura à direita. O tempo de vida é $t_0 + t$, onde t_0 representa o momento em que o indivíduo entrou no estudo e t é um valor desconhecido, mas que se sabe ser inferior ao tempo observado. Se a censura ocorrer em $t_0 + a$, sendo $t < a$, então a indica um tempo de vida censurado. Por exemplo, se o acontecimento de interesse for o aparecimento de metástases, quando o doente vai à consulta se já tiver metástases então existe censura à esquerda uma vez que o tempo até ao seu aparecimento (t) é inferior ao tempo até à consulta (a).

Censura intervalar

A censura intervalar verfica-se quando o acontecimento de interesse ocorreu num determinado intervalo de tempo conhecido. O acontecimento de interesse ocorreu entre t_1 e t_2, mas não sabemos o momento exato, por isso, o verdadeiro tempo de vida é desconhecido.

Um exemplo da censura intervalar ocorre quando fazemos vigilância de pessoas até se tornarem HIV positivas, onde no primeiro momento (t_1) apresentam um teste negativo, mas que no segundo momento (t_2) já é positivo. Assim, o tempo de sobrevivência é superior a t_1, mas inferior a t_2.

A censura pode ser ainda considerada não informativa ou informativa.

A **censura não informativa ou independente** acontece quando esta não está relacionada com o acontecimento de interesse. Assim sendo, a censura acontece por acaso e não associada a alguma informação relacionada com o acontecimento de interesse.

Já a **censura informativa** ocorre, por exemplo, quando o indivíduo em estudo abandona o tratamento devido a algumas complicações.

Portanto, ao iniciar um estudo, na fase de recrutamento deve-se evitar viés de seleção e, ao longo do estudo, é importante diminuir situações de perda de indivíduos, nomeadamente tendo em conta os seguintes aspetos [11]:

- uma criteriosa seleção dos indivíduos;
- o conhecimento de potenciais causas de censura.

1.5 Mecanismos de Truncatura

A truncatura acontece quando apenas são observados indivíduos a quem ocorreu determinado acontecimento ou vai ocorrer. Este tipo de mecanismo acontece por causa de um processo de seleção específico ao planeamento do estudo.

A ocorrência de truncatura ou de censura faz com que os dados sejam incompletos, apesar de serem de natureza distintas. O mecanismo de truncatura é composto por dois tipos: truncatura à esquerda e truncatura à direita.

Truncatura à esquerda

Na truncatura à esquerda os indivíduos selecionados para o estudo são apenas aqueles que satisfazem determinada condição. A condição é pré-estabelecida e deve ser verificada antes do acontecimento de interesse.

Um exemplo deste tipo de truncatura acontece quando estudamos indivíduos que estão a fazer hemodiálise. Os indivíduos selecionados são aqueles que estão a usufruir do tratamento à data de início do estudo e aqueles

que iniciarão o tratamento posteriormente. No entanto, os indivíduos são excluídos do estudo no caso de já lhes ter sido observado o acontecimento de interesse antes do início do estudo.

Truncatura à direita

A truncatura à direita verifica-se quando os indivíduos selecionados para o estudo já sofreram o acontecimento de interesse. O acontecimento em estudo é verificado durante o período de observação, antes de uma data específica, e o risco pode ser sobrestimado.

O tempo de sobrevivência (T) é inferior ao limite superior (T_D) do período estipulado, sendo T um valor conhecido e T_D um valor pré-definido no início do estudo.

Os dados com truncatura à direita não têm censura e os indivíduos que não sofrerem o acontecimento de interesse no período estipulado não serão incluídos no estudo, apesar de terem os fatores de risco. Normalmente, o critério de seleção dos indivíduos no estudo inicia na ocorrência do acontecimento interesse, mas antes do fim do período de observação.

Um exemplo da truncatura à direita é a transmissão do VIH por via materna (mãe-filho), pois os filhos selecionados para o estudo são aqueles que foram infetados e a data de infeção é a data de nascimento. Contudo, o período de observação até ao diagnóstico do vírus é o tempo decorrido desde o nascimento até ao aparecimento dos primeiros sintomas.

1.6 Função de Verosimilhança

Na Análise de Sobrevivência os métodos de inferência estatística utilizados, em geral, são baseados na teoria assintótica da máxima verosimilhança, este facto deve-se à existência de observações sujeitas aos processos de morte e de censura.

O tempo de vida T é uma variável aleatória que acompanha um modelo paramétrico, no qual a distribuição do tempo de vida T é conhecida, indexado por um vetor de parâmetros θ, sobre o qual se pretende efetuar inferência

estatística.

A construção da função de verosimilhança para observações sujeitas a censura, admite-se normalmente que os tempos de censura e os tempos de vida são independentes e que a censura é não informativa. Contudo, para a construção da função é necessário ter em atenção o tipo de censura: direita, esquerda ou intervalar, a que os indivíduos estão sujeitos, porque para cada mecanismo de censura a função de verosimilhança é definida de forma diferente.

Assim, quando estamos com tempos de vida exatos, ou seja, observações não censuradas, a informação fornecida é sobre a probabilidade de ocorrer o acontecimento de interesse nesse instante, o que corresponde à função densidade da variável tempo de vida, $f(t)$, nesse instante. Para observações censuradas à direita, a informação dada é que o verdadeiro tempo de sobrevivência é superior ao tempo de observação, que é transmitida através da função de sobrevivência, $S(t)$, nesse instante. Por outro lado, para observações censuradas à esquerda, a informação dada é que o tempo de sobrevivência (desconhecido) é inferior ao tempo de observação, que é transmitida pela função de distribuição, $1 - S(t)$, nesse instante. E, por fim, para observações sujeitas a censura intervalar, a informação disponibilizada é que o tempo de vida pertence a um determinado intervalo de tempo conhecido, que é dada pela função de sobrevivência desse intervalo de tempo.

A função de verosimilhança para uma amostra aleatória de dimensão n, $(t_1, \delta_1), (t_2, \delta_2), ..., (t_n, \delta_n)$, e assumindo que a censura é não informativa, ou seja, que a distribuição do tempo de censura não depende do vetor de parâmetros de interesse θ, podemos basear toda a inferência sobre este vetor a função é dada por:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{1.9}$$

o que é equivalente a

$$L = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \tag{1.10}$$

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

1.7 Estimador de Kaplan-Meier

Em 1958, Kaplan e Meier propuseram um estimador não paramétrico para a função de sobrevivência, designado por estimador de Kaplan-Meier ou estimador produto-limite. Este estimador permite a inclusão das observações censuradas [13, 34].

O estimador de Kaplan-Meier pressupõe que a ocorrência dos acontecimentos de interesse são independentes. Devido a este facto, a função de sobrevivência é estimada através do produto das probabilidades de sobrevivência até ao instante t.

Sejam, $t_{(1)}, t_{(2)}, ..., t_{(J)}$, tempos de mortes distintos de uma amostra dimensão $n, (J \leq n)$, de uma população homogénea. Seja n_j o número de indivíduos em risco no instante $t_{(j)}$, ou seja, os indivíduos para os quais ainda não foi observado o acontecimento de interesse ou cujos tempos de vida não foram censurados. Designe-se por d_j o número de acontecimentos ocorridos em $t_{(j)}$.

O estimador de Kaplan-Meier da função de sobrevivência é da forma:

$$\widehat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j} \right) \Leftrightarrow \widehat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j} \right) \tag{1.11}$$

Em contrapartida, a fórmula para calcular a função de sobrevivência com observações sem censura, é dada por

$$\widehat{S}(t) = \frac{\text{número de indivíduos com um tempo de vida} \geq t}{\text{número total de indivíduos de uma amostra}}$$

e designa-se por função de sobrevivência empírica.

O estimador de Kaplan-Meier da função de sobrevivência, $\widehat{S}(t)$, verifica as seguintes propriedades:

1. $\widehat{S}(t) = 1$ para $0 \leq t < t_{(1)}$;

2. $\widehat{S}(t) = 0$ para $t \geq t_{(J)}$, se o maior tempo de vida observado, $t_{(J)}$, não for um tempo censurado;

3. No caso de a maior observação, t^*, ser censurada, para $t_{(J)} \leq t \leq t^*$, $\widehat{S}(t)$ nunca é zero, estando definida apenas até t^*;

4. No caso de não existir censura, a função de sobrevivência empírica e o estimador de Kaplan-Meier da função de sobrevivência coincidem;

5. A função $\widehat{S}(t)$ é representada em escada, onde os degraus indicam os instantes onde se observaram o acontecimento de interesse (ver Figura 1.3);

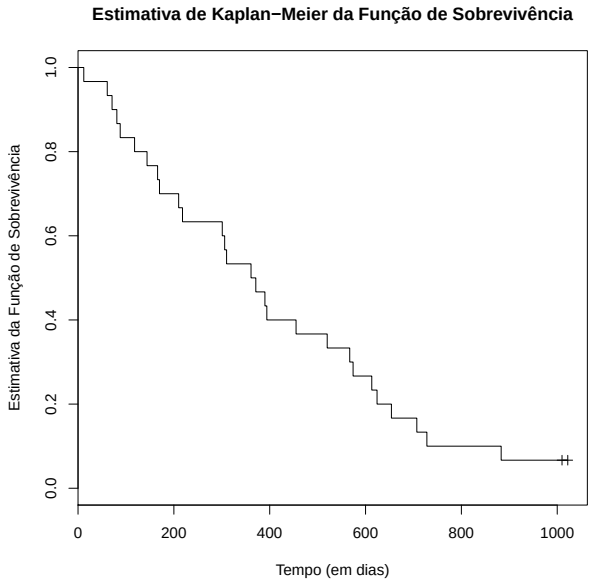

Figura 1.3: Estimativa de Kaplan-Meier da função de sobrevivência.

Para a representação gráfica da estimativa de Kaplan-Meier da função de sobrevivência $\widehat{S}(t)$ utilizamos o programa estatístico R, através do

package R Commander, onde foram selecionados os trinta primeiros indivíduos da base de dados *lung* do *package survival* (Figura 1.3).

6. $\widehat{S}(t)$ pode ainda ser visto como um estimador de máxima verosimilhança não paramétrico de $S(t)$.

O estimador da variância de $\widehat{S}(t)$, conhecido por fórmula de Greenwood, é da forma:

$$\widehat{var}(\widehat{S}(t)) = (\widehat{S}(t))^2 \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{1.12}$$

1.8 Modelos de Regressão

Os modelos de regressão permitem estudar a relação entre o tempo de vida e as covariáveis (variáveis independentes).

A especificação do modelo para a distribuição do tempo de sobrevivência T para determinado indivíduo engloba o vetor $\mathbf{z} = (z_1, z_2, ..., z_p)$ que representa as variáveis explanatórias, que correspondem a fatores de risco conhecidos.

Os modelos de regressão podem ser: paramétricos ou semiparamétricos. Os modelos de regressão paramétricos são implementados com base em distribuições, nomeadamente a distribuição exponencial, Weibull e log-logística, entre outras. Por outro lado, os modelos de regressão semiparamétricos não usam uma distribuição paramétrica para o tempo de sobrevivência, sendo o efeito das covariáveis a parte paramétrica do modelo.

Os modelos de regressão mais utilizados na Análise de Sobrevivência são: modelos com funções de risco proporcionais, modelos de tempo de vida acelerado e modelos de possibilidades proporcionais.

Modelos com Funções de Risco Proporcionais

A função de risco para um indivíduo com vetor de covariáveis $\mathbf{z} = (z_1, z_2, ..., z_p)$ é da forma

$$h(t;\mathbf{z}) = h_0(t)g(\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}) \tag{1.13}$$

sendo $h_0(t)$ a função de risco subjacente, ou seja, a correspondente a um indivíduo com vetor de covariáveis $\mathbf{z} = 0$ e $g(\mathbf{z})$ uma função das covariáveis. Habitualmente, $g(\mathbf{z}) = \exp(\beta'\mathbf{z})$, onde $\beta'\mathbf{z} = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p$ e $\beta_1, \beta_2, ..., \beta_p$ designam os coeficientes de regressão.

Nos modelos com funções de risco proporcionais a razão das funções de risco $\frac{h(t;\mathbf{z}_1)}{h(t;\mathbf{z}_2)}$ não depende de t, logo as funções de risco para dois indivíduos com vetores de covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, são proporcionais ao longo do tempo. As covariáveis têm um efeito multiplicativo na função de risco.

A função de sobrevivência é dada por

$$S(t;\mathbf{z}) = S_0(t)^{g(\mathbf{z})} = S_0(t)^{\exp(\beta'\mathbf{z})} \tag{1.14}$$

Quando se usa a distribuição exponencial ou, mais geralmente, a distribuição de Weibull para modelar o tempo de vida dos indivíduos, o modelo resultante é um modelo de riscos proporcionais. Outro exemplo deste tipo de modelos é o modelo de regressão de Cox.

Modelos de Tempo de Vida Acelerado

Os modelos de tempo de vida acelerado, também são conhecidos por modelos de localização-escala para $\log T$ ou modelos log-lineares para T. O modelo log-linear é definido do seguinte modo

$$\log T = \mu + \beta'\mathbf{z} + \sigma\varepsilon = \mu + (\beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p) + \sigma\varepsilon \tag{1.15}$$

onde μ é o termo independente, β é o vetor de parâmetros de regressão, $z_1, z_2, ..., z_p$ são as p covariáveis, σ é o parâmetro de escala ($\sigma > 0$) e ε é uma variável aleatória com uma distribuição de probabilidade que não depende de $\mathbf{z}$.

A função de risco para um indivíduo com vetor de covariáveis $\mathbf{z}$ é definida do seguinte modo

$$h(t;\mathbf{z}) = \alpha(\mathbf{z})h_0(t\alpha(\mathbf{z})) = \exp(-\beta'\mathbf{z})h_0(t\exp(-\beta'\mathbf{z})) \tag{1.16}$$

e a função de sobrevivência é

$$S(t;\mathbf{z}) = S_0(t\alpha(\mathbf{z})) = S_0(t\exp(-\beta'\mathbf{z})) \tag{1.17}$$

onde $\alpha(\mathbf{z}) = \exp(-\beta'\mathbf{z})$ e $\beta'\mathbf{z} = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p$ é a componente linear do modelo das z_j variáveis explanatórias para $j = 1, 2, ..., p$.

Neste tipo de modelo, as covariáveis têm um efeito multiplicativo em T. Este efeito pode ser de aceleração ou de desaceleração do tempo de sobrevivência até à ocorrência do acontecimento de interesse, nomeadamente o factor $\alpha(\mathbf{z})$ é denominado como factor de aceleração. No caso de $0 < \alpha(\mathbf{z}) < 1$ o efeito das covariáveis é de desaceleração do tempo de vida até à ocorrência do acontecimento de interesse, já no caso de $\alpha(\mathbf{z}) > 1$ o efeito das covariáveis é de aceleração do tempo de vida até à ocorrência do acontecimento de interesse.

As distribuições de Weibull e log-logística dão origem a modelos de regressão de tempo de vida acelerado.

Modelos de Possibilidades Proporcionais

Os modelos de possibilidades proporcionais são definidos pela possibilidade (*odds*) de sobrevivência para além do instante t, como sendo

$$\frac{S(t)}{1 - S(t)} \tag{1.18}$$

A função de sobrevivência, nesta classe de modelos, é definida do seguinte modo

$$\frac{S(t;\mathbf{z})}{1 - S(t;\mathbf{z})} = \exp(\eta_i)\frac{S_0(t)}{1 - S_0(t)} \tag{1.19}$$

que ainda pode ser escrita na forma

$$\begin{aligned} S(t;\mathbf{z}) &= S_0(t)\{\exp(-\eta_i) + (1 - \exp(-\eta_i))S_0(t)\}^{-1} \\ &= \frac{S_0(t)}{\exp(-\eta_i) + (1 - \exp(-\eta_i))S_0(t)} \end{aligned} \tag{1.20}$$

sendo $\eta_i = \beta_1 z_{1i} + \beta_2 z_{2i} + ... + \beta_p z_{pi}$ uma combinação linear das p covariáveis $z_1, z_2, ..., z_p$ associadas ao i-ésimo indivíduo e $S_0(t)$ é a função de sobrevivência subjacente correspondente a um indivíduo com vetor de covariáveis $\mathbf{z=0}$.

A função de risco para um indivíduo com vetor de covariáveis $\mathbf{z} = (z_1, z_2, ..., z_p)$ é da forma

$$h(t; \mathbf{z}) = h_0(t) - \frac{(1 - \exp(-\eta_i))h_0(t)S_0(t)}{\exp(-\eta_i) + (1 + \exp(-\eta_i))S_0(t)} \tag{1.21}$$

que após alguns cálculos pode ser escrita do seguinte modo

$$\frac{h(t; \mathbf{z})}{h_0(t)} = [1 + (\exp(\eta_i) - 1)S_0(t)]^{-1} \Leftrightarrow h(t; \mathbf{z}) = \frac{h_0(t)}{1 + (\exp(\eta_i) - 1)S_0(t)} \tag{1.22}$$

Neste modelo, as variáveis explanatórias atuam de forma multiplicativa na possibilidade de um indivíduo sobreviver para além do momento t. As funções de risco dos indivíduos convergem ao fim de um certo tempo, de acordo com a equação (1.22).

A distribuição log-logística dá origem a um modelo de possibilidades proporcionais.

1.8.1 Modelos Semiparamétricos

Num modelo semiparamétrico a função $h_0(t)$ não é especificada. Um exemplo deste tipo de modelo é o modelo de riscos proporcionais de Cox. Este modelo é muito utilizado na área da medicina.

Modelo de Cox

O modelo de regressão de Cox (1972) é um modelo semiparamétrico que permite uma análise de dados de sobrevivência com uma ou mais variáveis, designadas de covariáveis. É o modelo mais utilizado na regressão para análise do tempo de sobrevivência, devido à sua flexibilidade e versatilidade.

Seja T uma variável aleatória contínua que representa o tempo de sobrevivência e $\mathbf{z} = (z_1, z_2, ..., z_p)$ um vetor de covariáveis de determinado indivíduo. No instante t, o modelo de Cox, escrito com base nas funções de

risco, é dado por:

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p) \tag{1.23}$$

onde,

- $\beta_1, ..., \beta_p$ são coeficientes de regressão (desconhecidos) que representam o efeito das covariáveis na sobrevivência;
- $h_0(t)$ é uma função arbitrária não negativa, também conhecida por função de risco subjacente. Representa a função de risco para um indivíduo com vetor de covariáveis nulo ($\mathbf{z=0}$).

Como este é um modelo de riscos proporcionais, a função de risco para dois indivíduos com covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, é da forma:

$$\frac{h(t;\mathbf{z}_1)}{h(t;\mathbf{z}_2)} = \exp[\beta'(\mathbf{z}_1 - \mathbf{z}_2)] \tag{1.24}$$

ou seja, não depende de t.

1.8.2 Modelos Paramétricos

Os modelos paramétricos são modelos que têm um número finito de parâmetros e a distribuição do tempo até ao acontecimento de interesse é caracterizada em termos de paramêtros desconhecidos, ou seja, é considerada uma distribuição de probabilidade para o tempo de sobrevivência.

Nos modelos paramétricos, as distribuições mais comuns para o tempo de vida são a distribuição exponencial, a de Weibull e a log-logística, as quais serão descritas em seguida.

Modelo de Regressão Exponencial

O modelo de regressão exponencial caracteriza-se por considerar a distribuição exponencial para modelar o tempo de vida dos indivíduos em estudo.

- **Distribuição Exponencial**

Seja T uma variável aleatória com uma distribuição exponencial a qual só tem um único parâmetro, λ, que pode tomar qualquer valor positivo. Esta distribuição é a mais simples dos modelos paramétricos, onde a função densidade de probabilidade é dada por:

$$f(t) = \lambda \exp(-\lambda t), \quad \lambda > 0, \quad t \geq 0 \tag{1.25}$$

Na distribuição exponencial a função de risco é constante. A função de risco e a função de sobrevivência, podem ser determinadas através das equações (1.3), (1.4) e (1.25) e são, respetivamente,

$$h(t) = \lambda, \quad t \geq 0 \tag{1.26}$$

$$S(t) = \exp(-\lambda t), \quad t \geq 0 \tag{1.27}$$

Como a distribuição apresenta uma função de risco constante, o risco de morte é igual em qualquer momento, independentemente do tempo decorrido. Isto deve-se ao facto de a distribuição ter falta de memória.

- **Modelo de Regressão**

O modelo de regressão exponencial pode ser formulado como um modelo de riscos proporcionais ou como um modelo de tempo de vida acelerado.

A função de risco, pelo modelo de riscos proporcionais, para um indivíduo com vetor de covariáveis $\mathbf{z}$, é expressa a partir das equações (1.13) e (1.26)

$$h(t; \mathbf{z}) = h_0(t) \exp(\beta' \mathbf{z}) = \lambda \exp(\beta' \mathbf{z}) \tag{1.28}$$

e a função de sobrevivência é definida a partir das equações (1.14) e (1.27)

$$S(t; \mathbf{z}) = S_0(t)^{\exp(\beta' \mathbf{z})} = \exp(-\lambda t \exp(\beta' \mathbf{z})) \tag{1.29}$$

Por outro lado, a função de risco, com base no modelo de tempo de vida acelerado, para um indivíduo com vetor de covariáveis $\mathbf{z}$, é definida com base

nas equações (1.16) e (1.26)

$$h(t;\mathbf{z}) = \exp(-\beta'\mathbf{z})h_0(t\exp(-\beta'\mathbf{z})) = \exp(-\beta'\mathbf{z})\lambda = \lambda\exp(-\beta'\mathbf{z}) \quad (1.30)$$

e a função de sobrevivência é escrita com base nas equações (1.17) e (1.27)

$$S(t;\mathbf{z}) = S_0(t\exp(-\beta'\mathbf{z})) = \exp(-\lambda t\exp(-\beta'\mathbf{z})) \quad (1.31)$$

Modelo de Regressão de Weibull

O modelo de regressão de Weibull utiliza a distribuição de Weibull para modelar o tempo de vida.

- **Distribuição de Weibull**

A distribuição de Weibull contém dois parâmetros, o parâmetro de escala $\lambda > 0$ e o parâmetro de forma $\gamma > 0$, com a função densidade de probabilidade dada por

$$f(t) = \lambda\gamma(t)^{\gamma-1}\exp(-\lambda t^{\gamma}), \quad t \geq 0 \quad (1.32)$$

A função de risco e a função de sobrevivência são, respetivamente,

$$h(t) = \lambda\gamma t^{\gamma-1}, \quad t \geq 0 \quad (1.33)$$

$$S(t) = \exp(-\lambda t^{\gamma}), \quad t \geq 0 \quad (1.34)$$

No caso de $\gamma = 1$ a distribuição de Weibull corresponde à distribuição exponencial, logo a função de risco é constante; quando $\gamma > 1$ a função de risco é monótona crescente, e, por fim, quando $0 < \gamma < 1$ a função de risco é monótona decrescente.

A distribuição de Weibull é a distribuição mais utilizada na Análise de Sobrevivência, especialmente nas áreas da medicina e da biologia, devido à flexibilidade da sua função de risco.

- **Modelo de Regressão**

O modelo de regressão de Weibull pode ser formulado como um modelo de riscos proporcionais ou como um modelo de tempo de vida acelerado.

Na formulação de riscos proporcionais, a função de risco, para um indivíduo com vetor de covariáveis $\mathbf{z}$, pode ser escrita da forma:

$$h(t;\mathbf{z}) = h_0(t)\exp(\beta'\mathbf{z}) = \lambda\gamma t^{\gamma-1}\exp(\beta'\mathbf{z}) = \lambda\exp(\beta'\mathbf{z})\gamma t^{\gamma-1} \tag{1.35}$$

Assim, o tempo de sobrevivência do indivíduo tem uma distribuição de Weibull e o parâmetro de escala e de forma são, respetivamente, $\lambda\exp(\beta'\mathbf{z})$ e γ. É no parâmetro de escala que o efeito das covariáveis é sentido, já o parâmetro de forma mantém-se inalterado. A função de sobrevivência é da forma

$$S(t;\mathbf{z}) = S_0(t)^{\exp(\beta'\mathbf{z})} = \exp(-\lambda t^{\gamma})^{\exp(\beta'\mathbf{z})} = \exp(-\lambda t^{\gamma}\exp(\beta'\mathbf{z})) \tag{1.36}$$

Já na formulação do tempo de vida acelerado, a função de risco, para um indivíduo com vetor de covariáveis $\mathbf{z}$, é

$$h(t;\mathbf{z}) = \exp(\beta'\mathbf{z})\lambda\gamma(t\exp(\beta'\mathbf{z}))^{\gamma-1} = \lambda(\exp(\beta'\mathbf{z}))^{\gamma}\gamma t^{\gamma-1} \tag{1.37}$$

A função de sobrevivência é

$$S(t;\mathbf{z}) = \exp[-\lambda(\exp(\beta'\mathbf{z})t)^{\gamma}] = \exp[-\lambda(\exp(\beta'\mathbf{z}))^{\gamma}t^{\gamma}] \tag{1.38}$$

O modelo de regressão de Weibull é o único modelo que pode ser formulado em termos de modelo de riscos proporcionais ou de modelo de tempo de vida acelerado. Note-se que a distribuição de exponencial é um caso particular da distribuição de Weibull pois corresponde a considerar $\gamma = 1$.

Modelo de Regressão Log-logístico

O modelo de regressão log-logístico usa a distribuição log-logística para modelar o tempo de vida e tanto pode ser formulado em termos de tempo de vida acelerado como em termos de possibilidades proporcionais.

- **Distribuição Log-logística**

A distribuição log-logística é uma alternativa à distribuição de Weibull, visto que esta distribuição pode considerar a função de risco unimodal. Esta distribuição é caracterizada por dois parâmetros, o parâmetro de escala $\lambda > 0$ e o parâmetro de forma $k > 0$.

A função densidade de probabilidade, a função de risco e a função de sobrevivência são, respetivamente, as seguintes,

$$f(t) = \frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}, \quad t \geq 0 \tag{1.39}$$

$$h(t) = \frac{\lambda k t^{k-1}}{1+\lambda t^k}, \quad t \geq 0 \tag{1.40}$$

$$S(t) = \frac{1}{1+\lambda t^k}, \quad t \geq 0 \tag{1.41}$$

Para $k > 1$ a função de risco é unimodal, sendo monótona descrescente para $0 < k \leq 1$.

- **Modelo de Regressão**

O modelo de regressão log-logístico é uma alternativa ao modelo de regressão de Weibull, nomeadamente para situações onde as funções de risco são não monótonas ou para modelos de possibilidades proporcionais.

A função de risco para um indivíduo com vetor de covariáveis $\mathbf{z}$, com base no modelo de possibilidades proporcionais, pode ser escrita a partir das equações (1.22), (1.40) e (1.41):

$$\begin{aligned} h(t;\mathbf{z}) &= \frac{h_0(t)}{1+(\exp(-\eta_i)-1)S_0(t)} = \frac{\frac{\lambda k t^{k-1}}{1+\lambda t^k}}{1+(\exp(\eta_i)-1)\frac{1}{1+\lambda t^k}} \\ &= \frac{\frac{\lambda k t^{k-1}}{1+\lambda t^k}}{1+\frac{\exp(\eta_i)-1}{1+\lambda t^k}} = \frac{\lambda k t^{k-1}}{\lambda t^k + \exp(\eta_i)} \end{aligned} \tag{1.42}$$

A função de sobrevivência é baseada nas equações (1.20) e (1.41):

$$\begin{aligned} S(t;\mathbf{z}) &= \frac{S_0(t)}{\exp(-\eta_i)+(1-\exp(-\eta_i))S_0(t)} = \frac{\frac{1}{1+\lambda t^k}}{\exp(-\eta_i)+\frac{(1-\exp(-\eta_i))}{1+\lambda t^k}} \\ &= \frac{\frac{1}{1+\lambda t^k}}{\frac{(1+\lambda t^k)\exp(-\eta_i)+1-\exp(-\eta_i)}{1+\lambda t^k}} = \frac{1}{1+\lambda\exp(-\eta_i)t^k} \end{aligned} \tag{1.43}$$

O tempo de vida tem uma distribuição log-logística, onde o parâmetro de escala é $\lambda\exp(-\eta_i)$ e o parâmetro de forma é k.

Por outro lado, na formulação através do tempo de vida acelerado, a função de risco, para um indivíduo com vetor de covariáveis $\mathbf{z}$, pode ser escrita na forma

$$\begin{aligned} h(t;\mathbf{z}) &= \exp(-\beta'\mathbf{z})h_0(t\exp(-\beta'\mathbf{z})) = \exp(-\beta'\mathbf{z})\frac{\lambda k(t\exp(-\beta'\mathbf{z}))^{k-1}}{1+\lambda(t\exp(-\beta'\mathbf{z}))^k} \\ &= \frac{\lambda(\exp(-\beta'\mathbf{z}))^k k t^{k-1}}{1+\lambda(\exp(-\beta'\mathbf{z}))^k t^k} \end{aligned} \tag{1.44}$$

A função de sobrevivência é da forma

$$S(t;\mathbf{z}) = S_0(t\exp(-\beta'\mathbf{z})) = \frac{1}{1+\lambda(t\exp(-\beta'\mathbf{z}))^k} = \frac{1}{1+\lambda(\exp(-\beta'\mathbf{z}))^k t^k} \tag{1.45}$$

Capítulo 2

Análise de Sobrevivência Relativa

2.1 Introdução

A Análise de Sobrevivência Relativa é uma subárea da Análise de Sobrevivência muito usada nas Ciências Biomédicas, especialmente na oncologia, que permite monitorizar a atividade de controlo do cancro e estimar o tempo de sobrevivência dos indivíduos com determinado diagnóstico de neoplasia [31].

Como a Análise de Sobrevivência clássica estuda o tempo de vida dos indivíduos, também a Análise da Sobrevivência Relativa estuda o tempo decorrido desde a data de diagnóstico de cancro até a data da morte do indivíduo, sendo o acontecimento de interesse a morte devido ao cancro. Este tempo apresenta-se sob a forma de taxa, a qual indica a percentagem de indivíduos vivos durante o período de estudo.

Por vezes a causa de morte do indivíduo é desconhecida ou a informação na certidão de óbito não especifica se a morte foi devida ao cancro. Então, para estimar a probabilidade do indivíduo estar vivo no instante t sendo o acontecimento de interesse a morte por causa do cancro, utiliza-se a Análise de Sobrevivência Relativa. Esta análise surge da dificuldade em determinar a principal causa de morte do indivíduo, visto que a morte do indivíduo

pode ser por outras causas que não a do cancro, como por exemplo: enfarte de miocárdio, pneumonia, acidente vascular cerebral, entre outros, sendo a causa específica de morte desconhecida.

A Sobrevivência Relativa é uma medida objetiva da sobrevivência que não necessita de informação sobre a causa concreta de morte e define-se pelo rácio entre a função de sobrevivência observada e a função de sobrevivência esperada. O resultado deste rácio é uma estimativa do tempo de vida dos indivíduos em estudo, no caso de a patologia em análise ser a única causa de morte possível e permite analisar o impacto das mortes causadas pela doença em estudo em relação à mortalidade da população em geral.

Em 2012, Pohar Perme, Stare e Estève, [37], propuseram um método, designado por método de Pohar Perme que permite calcular a probabilidade de sobreviver ao cancro, na situação hipotética de não ser possível morrer de outras causas. Assim sendo, este método também permite calcular a net survival (sobrevivência parcial). Além do mais, não compara globalmente a proporção da sobrevivência observada e a proporção da sobrevivência esperada como na Sobrevivência Relativa; essa comparação é feita indivíduo a indivíduo.

Segundo os autores Dickman e Coviello,[17], a sobrevivência observada e a esperada são representadas por proporções e não por taxas. Assim, segundo Ederer et al., [19], a sobrevivência observada diz respeito a um grupo de indivíduos com diagnóstico de neoplasia ou de outra doença específica durante o período de *follow-up*. Já a sobrevivência esperada diz respeito a um grupo de indivíduos semelhante ao grupo da sobrevivência observada, mas sem diagnóstico da doença e pertencentes à população em geral. Este grupo, também conhecido como grupo de correspondência ou de referência, é constituído por indivíduos sem diagnóstico da doença mas com características semelhantes aos indivíduos com diagnóstico da doença em estudo, nomeadamente pertencer à mesma faixa etária, género, raça, entre outras.

No entanto, devido à dificuldade em obter uma coorte de pessoas sem diagnóstico da doença, a sobrevivência esperada é estimada com base nas tábuas de mortalidade utilizando o método Ederer I, ou método Ederer II ou método Hakulinen. As tábuas de mortalidade apresentam dados relativos à

sobrevivência da população em geral seguida ao longo do tempo, designado por coorte (*cohort*). Em Portugal, estas tábuas são divulgadas pelo Instituto Nacional de Estatística (INE) podendo ser obtidas através do endereço eletrónico (https://www.ine.pt) e contêm dados sobre Portugal Continental e sobre as Regiões Autonómas dos Açores e da Madeira. A nível global, podem ser obtidas através do projeto do Departamento de Demografia da Universidade da Califórnia, Berkeley, EUA, e do Instituto Max Planck de Pesquisa Demográfica, em Rostock, Alemanha, a *Human Mortality Database* (HMD) cujo endereço eletrónico é http://www.mortality.org/ e que contém informação sobre 38 países, entre eles Portugal.

As tábuas de mortalidade podem ser: por coorte (*cohort*) ou por período. As tábuas de mortalidade por coorte apresentam a probabilidade de morte de coortes de indivíduos. Porém não são muito utilizadas, porque as estimativas da sobrevivência esperada são calculadas com base em tábuas de mortalidade relativa de anos anteriores distantes, sendo o resultado das taxas de sobrevivência pouco recentes, não traduzindo os avanços científicos e tecnológicos e de tratamento.

Por outro lado, as tábuas de mortalidade por período incluem todos os indivíduos com doença de todas as faixas etárias no período estabelecido, não se limitando a uma geração, dando estimativas de sobrevivência mais recentes. A estimação do período de análise deve ser realizada com base na informação mais recente disponibilizada pelo registo oncológico.

No entanto, a Sobrevivência Relativa tem algumas limitações, [27], nomeadamente quando há uma elevada proporção de mortes devido a determinada doença específica no grupo de correspondência. Este excesso de mortalidade será provavelmente subestimado, levando a sobrestimativas da Sobrevivência Relativa dos indivíduos com cancro. Porém, segundo Ederer et al, [19], a proporção de mortes devido a uma causa específica é tão pequena em comparação com a mortalidade global que se torna irrelevante ajustar os valores da tábua de mortalidade, a fim de eliminar as mortes para determinar as estimativas da Sobrevivência Relativa. No entanto, esta suposição é questionável quando o estudo é sobre cancros mais comuns (e.g. cancro do cólon ou da mama), particularmente em idades mais avançadas.

Portanto, a Sobrevivência Relativa permite: conhecer o impacto da neoplasia no tempo de vida dos indivíduos; fazer comparações entre o grupo de indivíduos em estudo com o grupo de correspondência e identificar diferenças na análise de sobrevivência relativa ao nível da idade, sexo, estadio da doença, data do diagnóstico, tipo de cancro, tipo de tratamento, localização geográfica, entre outros.

2.2 Tábuas de Mortalidade

Em 1693, Halley, [14], editou a primeira tábua de mortalidade que foi construída com base em registos de óbitos classificados por idade entre 1687 e 1691 ocorridos na cidade de Breslau, em Inglaterra. No entanto, só em 1815, Milne divulga a primeira tábua científica, por meio de técnicas estatísticas e demográficas, baseada na população e nos óbitos classificados por idade.

As tábuas de mortalidade, normalmente, são calculadas para homens, mulheres e para ambos os géneros, que integram indicadores que permitem calcular as probabilidades de sobrevivência, medir o fenómeno da mortalidade e divulgar a esperança média de vida. O período de referência para as tábuas completas de mortalidade é de três anos consecutivos.

A tábua de mortalidade é um método de apresentação de dados relativos à sobrevivência de um conjunto de indivíduos seguidos ao longo do tempo, designado por coorte. Este método é utilizado para a estimação não paramétrica da função de sobrevivência, sendo muito utilizada na área da investigação médica.

Como o conjunto de indivíduos constituem uma amostra aleatória proveniente de uma dada população, é possível com a tábua de mortalidade estimar a probabilidade de sobrevivência para além de um determinado intervalo e a probabilidade condicional de morte num intervalo, dada a sobrevivência no início desse intervalo.

Seja então um conjunto de dados relativos à sobrevivência da população em geral seguida ao longo do tempo, ou seja, uma coorte de n indivíduos provenientes da população em estudo. O intervalo $[0, \infty)$ é dividido em $K+1$

intervalos adjacentes de amplitude fixa, ou seja, $l_k = [a_{k-1}, a_k)$, $k = 1, ..., K+1$, sendo $a_0 = 0$ o instante de início do estudo e $a_K = L$ o limite superior de observação e $a_{K+1} = \infty$.

Os dados são formados pelo número de indivíduos vivos no início de cada intervalo e pelo número de indivíduos que morrem ou são censurados em cada intervalo, onde n_k é o número de indivíduos em risco (isto é, vivos e não censurados) no instante a_{k-1}, d_k é o número de mortes observadas em l_k e w_k é o número de observações censuradas em l_k.

Seja P_k a probabilidade de um indivíduo sobreviver para além de l_k e q_k a probabilidade de um indivíduo morrer em l_k sabendo que sobreviveu para além de l_{k-1}. A representação da tábua de mortalidade é apresentada na Tabela 2.1.

Tabela 2.1: Tábua de Mortalidade

k	nº indivíduos em risco (n_k)	nº de mortes observadas (d_k)	nº de observações censuradas (w_k)	estimativa q_k	estimativa P_k
1	$n_1 = n$	d_1	w_1	$\frac{d_1}{n-\frac{w_1}{2}}$	$\widehat{p}_1\widehat{P}_0$
...	...	...	...	...	...
K	$n_{K-1}-d_{K-1}-w_{K-1}$	d_K	w_K	$\frac{d_K}{n_K-\frac{w_K}{2}}$	$\widehat{p}_K\widehat{P}_{K-1}$

A tábua de mortalidade é uma tabela onde, para cada intervalo l_k são representados todos os valores de n_k, d_k e w_k e as estimativas de q_k e P_k. Note-se que no primeiro intervalo estão em risco todos os indivíduos ($n = n_1$) e para os restantes intervalos, o número de indivíduos em risco é calculado através do número de indivíduos em risco no intervalo anterior (n_{k-1}) menos o número de mortes (d_{k-1}) e o número de observações censuradas (w_{k-1}) nesse intervalo, para $k = 2, ..., K+1$.

Define-se que $p_k = 1 - q_k = \frac{P_k}{P_{k-1}}$, onde $P_0 = 1$, $P_{K+1} = 0$ e $q_{K+1} = 1$. Então para obter a estimativa de $P_k = p_1p_2...p_k$ para $k = 1, ..., K+1$ é necessário primeiro estimar q_k usando o estimador atuarial que é da forma

$$\widehat{q}_k = \begin{cases} 1, & \text{se } n_k = 0 \\ \frac{d_k}{n_k - \frac{w_k}{2}}, & \text{se } n_k > 0 \end{cases}$$

Assim, $\widehat{p}_k = 1 - \widehat{q}_k$, o estimador de P_k é da forma

$$\widehat{P}_k = \widehat{p}_1 ... \widehat{p}_k, \quad k = 1, ..., K+1$$

ou, alternativamente,

$$\widehat{P}_k = \widehat{p}_k \widehat{P}_{k-1}$$

O cálculo das probabilidades de sobrevivência e da esperança de vida são estimados através de uma análise das tábuas de mortalidade por período ou por coorte.

2.2.1 Análise por Coorte

A análise por coorte é um método tradicional utilizado na análise de sobrevivência. Porém, muitas vezes, os resultados estão desatualizados no momento em que se tornam disponíveis. Isto deve-se ao facto de as estimativas da sobrevivência esperada serem baseadas exclusivamente em indivíduos diagnosticados há muito tempo, não fornecendo as tendências mais recentes das taxas de sobrevivência como reflexo do progresso de diagnóstico e prognóstico, do avanço científico e tecnológico e de tratamento, o que torna uma desvantagem em comparação com a análise por período que será abordada na secção seguinte.

Na análise por coorte, para a determinação das estimativas de sobrevivência apenas são incluídos os indivíduos que foram seguidos durante todo o período de *follow-up* [9]. Assim, as tábuas de mortalidade por coorte apresentam a probabilidade de morte de indivíduos que pertencem à mesma geração ou coorte (e.g. todos os indivíduos nascidos em 1970), o que implica que os indivíduos sejam acompanhados desde o nascimento e ao longo da vida.

2.2.2 Análise por Período

Em 1996, surge a primeira publicação que propõe o uso da análise de período para a determinação das estimativas de sobrevivência de longo prazo. Esta metodologia é descrita em detalhe por Brenner et al. [10].

A análise por período é um método muito utilizado na análise de sobrevivência de indivíduos com doenças crónicas, como por exemplo o cancro. Como a maioria das mortes devido ao cancro ocorrem durante os primeiros anos após o diagnóstico, é importante obter probabilidades de sobrevivência de indivíduos recentemente diagnosticados mais atualizadas, de modo a ser possível monitorizar o progresso no tratamento do cancro ao longo do tempo e comparar a qualidade de tratamento entre as diferentes populações [29].

Este método limita a análise a um período de tempo mais recente, sendo os tempos de sobrevivência truncados à esquerda no início do período de interesse e censurados à direita no fim do período. Os indivíduos selecionados são de todas as idades num período fixo, não se limitando a uma geração, o que permite avaliar as condições de mortalidade durante o período selecionado.

A estimação do período de análise deve ser realizada tendo em conta a informação mais recente disponibilizada pelo registo oncológico e que permita maximizar o número de anos de diagnóstico [9, 10, 29]. Normalmente, o período de referência é de um a três anos. As estimativas de sobrevivência são mais atualizadas, revelando tendências recentes e com maior precisão em comparação à análise por coorte.

Segundo os autores Holleczek e Brenner, [29], a análise por período foi introduzida como modelo de regressão usando os modelos lineares generalizados (GLM) que surgiram no início de 1970. A estrutura dos modelos lineares generalizados na análise por período veio permitir uma maior precisão das estimativas de sobrevivência, avaliar e ter em conta os efeitos das covariáveis adicionais, nomeadamente a idade, ano de diagnóstico, tipo de neoplasia, género, região geográfica, entre outras.

O programa estatístico R fornece uma implementação muito flexível da estrutura dos modelos lineares generalizados através da função `glm`. Para a análise por período, o R teve disponível temporariamente (em 2007, pelo

menos) o *package periodR*. Entretanto, esse *package* já não se encontra disponível, razão pela qual não foi possível exemplificar a implementação deste método.

2.3 Sobrevivência Relativa

Seja T uma variável aleatória, que representa o tempo decorrido desde a data de diagnóstico do cancro até à data da morte do indivíduo. A função de sobrevivência relativa $S_R(t)$ é da forma

$$S_R(t) = \frac{S_O(t)}{S_E(t)} = \frac{\frac{1}{n}\sum_{i=1}^{n} S_{Oi}(t)}{\frac{1}{n}\sum_{i=1}^{n} S_{Ei}(t)} \tag{2.1}$$

onde $S_O(t)$ representa a função de sobrevivência observada de um grupo de n indivíduos com diagnóstico de cancro e $S_E(t)$ representa a função de sobrevivência esperada do grupo de correspondência.

A sobrevivência observada pode ser estimada a partir do método atuarial ou através da função de risco cumulativa e a sobrevivência esperada é calculada através das tábuas de mortalidade. Contudo, segundo Ederer et al., [19], o facto de as tábuas de mortalidade apresentarem informação sobre a mortalidade global que inclui as mortes relacionadas com o cancro e por outras causas não consitui um problema uma vez que a mortalidade por causa do cancro torna-se irrelevante face à mortalidade global.

A Figura 2.1 apresenta um exemplo da estimação da curva de Sobrevivência Relativa para indivíduos do sexo masculino com 65 anos e com data de diagnóstico a 1 de julho 1982. Esta figura foi executada no programa estatístico *R*, através do *package R Commander* com a base de dados *rdata* e as tábuas de mortalidade *slopop* do *package relsurv*. A base de dados *rdata* é constituída por 1040 observações com seis variáveis: `time` (tempo de sobrevivência em dias), `cens` (indicador de estado: 0=censura e 1=morte), `age` (idade em anos), `sex` (sexo: 1=masculino e 2=feminino), `year` (data do diagnóstico em formato de data) e `agegr` (grupo etário) e as tábuas de mortalidade foram realizadas com dados do censo da população da Eslovénia entre 1930 e 2014.

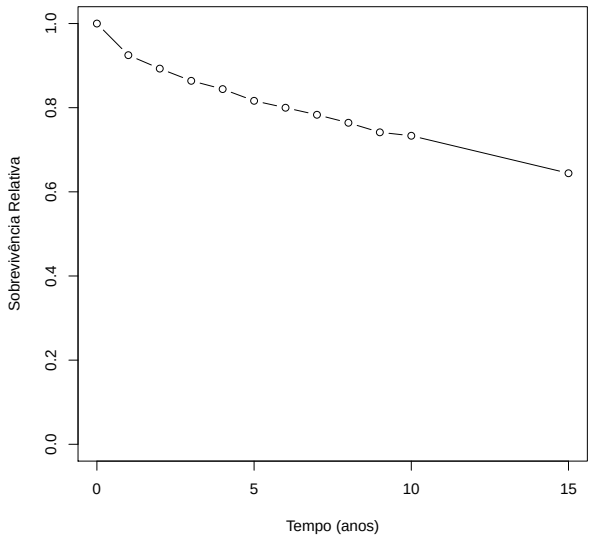

Figura 2.1: Curva da Sobrevivência Relativa

As propriedades da função de sobrevivência relativa são:

- função poderá ser ou não monótona decrescente;
- resultado é sempre positivo, mas pode ser superior a 1.

Quanto ao resultado da função de sobrevivência relativa (equação 2.1), normalmente, a função regista valores com percentagem inferior a 100% o que representa um excesso de mortalidade dos indivíduos com diagnóstico de cancro em relação à mortalidade global de coorte. Por outro lado, mas pouco frequente, se a percentagem for superior a 100% podemos estar perante não só uma potencial cura mas também uma melhoria geral no estado de saúde, resultante de um maior acompanhamento médico.

2.3.1 Estimação da Sobrevivência Observada

Segundo Ederer et al., [19], a sobrevivência observada é calculada a partir de um conjunto de dados correspondente ao tempo de vida de um grupo de doentes com cancro, seguidos durante o período de *follow-up*.

A sobrevivência observada, é a probabilidade de sobrevivência, a qualquer causa de morte, para os indivíduos com diagnóstico de cancro [12]. Pode ser estimada através das tábuas de mortalidade utilizando os métodos atuarial ou a função de risco cumulativa [17].

O método atuarial é um método muito antigo da estatística que permite calcular a probabilidade de sobrevivência observada, construir as tábuas de mortalidade e estimar as características de distribuição dos tempos de morte. Segundo os autores Parkin e Hakulinen, [35], o método atuarial fornece informação sobre o padrão de sobrevivência e sobre as mudanças no risco de morrer nos sucessivos instantes de observação.

Seja $t_{(j)}$, $j = 1, ..., J$ os instantes de observação de uma coorte de dimensão n, $(J \leq n)$, indivíduos com diagnóstico do cancro em estudo. A probabilidade da sobrevivência observada no instante $t_{(j)}$, através do método atuarial, é da forma:

$$p_j = 1 - q_j = 1 - \left(\frac{d_j}{n_j - \frac{w_j}{2}} \right) = 1 - \left(\frac{d_j}{l_j} \right) \tag{2.2}$$

onde,

- q_j é a probabilidade de morte em $t_{(j)}$;
- d_j é o número de acontecimentos de interesse (mortes) ocorridos, podendo ser justificados pela doença em estudo ou por outra causa de morte em $t_{(j)}$;
- n_j o número de indivíduos em risco (isto é, vivos e não censurados) em $t_{(j)}$;
- w_j é o número de observações censuradas em $t_{(j)}$;
- $l_j = n_j - \frac{w_j}{2}$ é o número ajustado de indivíduos em risco em $t_{(j)}$.

Portanto, no método atuarial o número de instantes é fixado à partida e o número de indivíduos expostos ao risco corresponde aos indivíduos vivos ao início de cada intervalo. Contudo, este método é muito semelhante ao método de Kaplan-Meier pois neste último o número de instantes de tempo

corresponde ao número de instantes de ocorrência do acontecimento de interesse.

Outra forma de estimar a sobrevivência observada é a através da função de risco cumulativa. A função de risco cumulativa mede o risco de ocorrência do acontecimento de interesse até um determinado momento, ou seja, é a soma dos riscos dos vários instantes de ocorrência do acontecimento de interesse. Como vimos na secção 1.3 a função de risco cumulativa (ver equação 1.7 e 1.8) é da seguinte forma:

$$H(t) = \int_0^t h(x)dx, \quad t \geq 0 \quad e \quad H(t) = -log(S(t)) \tag{2.3}$$

O risco médio para o instante $t_{(j)}$ é do seguinte modo:

$$\lambda_j = \frac{d_j}{y_j} \tag{2.4}$$

onde d_j representa o número de mortes em $t_{(j)}$ e y_j é o tempo total em risco do indivíduo (*person-time at risk*) em $(t_{j-1}, t_j]$, que é da forma $y_j = n_j - \left(\frac{w_j + d_j}{2}\right)$.

A função de risco cumulativa é utilizada quando sabemos o tempo exato do indivíduo em estudo, portanto a probabilidade do tempo de sobrevivência do indivíduo no instante $t_{(j)}$ é dada da forma:

$$p_j = \exp(-H(j)) = \exp\left(- k_j\left(\frac{d_j}{y_j}\right)\right) \tag{2.5}$$

onde k_j é a amplitude de $(t_{j-1}, t_j]$. No caso do risco ser constante, então o risco cumulativo é $H(j) = k_j\left(\frac{d_j}{y_j}\right)$.

Portanto, a sobrevivência observada $S_O(t)$ de um grupo de indivíduos com diagnóstico de doença em estudo ao fim do intervalo j e é representada do seguinte modo (os procedimentos utilizados para o cálculo, encontram-se disponíveis no Anexo A.1):

$$S_O(t) = \prod_{j:t_{(j)} \leq t} p_j \tag{2.6}$$

O estimador da variância da sobrevivência observada, com recurso à fórmula de Greenwood, no instante $t_{(j)}$ pode ser escrito da forma:

$$\widehat{\text{var}}(p_j) = (p_j)^2 \frac{d_j}{l_j(l_j - d_j)} \tag{2.7}$$

Já o estimador da variância da sobrevivência observada ao fim do intervalo j é definido por:

$$\widehat{\text{var}}(S_O(t)) = (S_O(t))^2 \sum_{j:t_{(j)} \leq t} \frac{d_j}{l_j(l_j - d_j)} \tag{2.8}$$

2.3.2 Estimação da Sobrevivência Esperada

Segundo Ederer et al., [19], a sobrevivência esperada é calculada com base num conjunto de indivíduos semelhantes àqueles para os quais foi calculada a sobrevivência observada, semelhança essa verificada ao nível do género, idade e raça, mas que não têm o diagnóstico de cancro (ou, mais geralmente, da doença em estudo).

No entanto, para a estimação da sobrevivência esperada é necessário uma coorte de indivíduos sem cancro, mas é difícil. Então, a solução é recorrer às tábuas de mortalidade através dos métodos, Ederer I ou Ederer II ou Hakulinen, os quais diferem entre si quanto ao tempo em que cada indivíduo é considerado estar em risco. E, apesar da tábua de mortalidade incluir todo o tipo de morte, os autores Ederer et al., [19], dizem que a mortalidade por causa do cancro é irrelevante face à mortalidade global.

Em termos gerais, os métodos de Ederer têm por base o cálculo de uma sobrevivência populacional média para cada instante em que ocorreu um acontecimento (morte). A diferença entre os dois métodos Ederer reside no número de indivíduos que se considera em cada instante. Ederer I considera para cada intervalo todos os pacientes, enquanto que Ederer II considera apenas os que estão em risco nesse intervalo.

De seguida serão apresentados os três métodos para estimação da sobrevivência esperada: Ederer I, Ederer II e Hakulinen.

Método Ederer I

Segundo Ederer, Axtell e Cutler (1961),[19], o método Ederer I determina a sobrevivência esperada considerando que cada paciente é um membro da população em geral desde a data do diagnóstico até ao fim do *follow-up*, de modo que os indivíduos do grupo de correspondência estão sempre em risco. Assim, o facto de ocorrer a morte ou censura no indivíduo com cancro não vai ter qualquer efeito sobre o indivíduo do grupo de correspondência e, consequentemente, sobre a sobrevivência esperada. Deste modo, não se considera o facto de poder haver tempos de *follow-up* heterogéneos, originando potenciais estimativas enviesadas da sobrevivência relativa.

Sejam $t_{(1)}, t_{(2)}, ..., t_{(J)}$ os instantes de morte da coorte em estudo. Seja ainda $\tilde{p}_{ij}$ a probabilidade esperada (calculada a partir das tábuas de mortalidade, ver Anexo A.1) do i-ésimo paciente sobreviver ao fim do instante $t_{(j)}$. Então, a probabilidade de sobrevivência esperada para o instante $t_{(j)}$ é da forma:

$$p_j^E = \frac{1}{l_j} \sum_{i=1}^{l_j} \tilde{p}_{ij} \tag{2.9}$$

Note-se que, para efeitos do cálculo da sobrevivência esperada, o i-ésimo indivíduo é selecionado da população em geral de modo a ter as mesmas características do i-ésimo paciente no que diz respeito à idade e ao sexo.

A probabilidade de sobrevivência esperada cumulativa de sobreviver ao fim do intervalo j é dada por:

$$S_E(t) = \frac{1}{l_1} \sum_{i=1}^{l_1} \left(\prod_{j=1}^{J} \tilde{p}_{ij} \right) \tag{2.10}$$

onde o l_1 representa o número de indivíduos com cancro vivos no início do período de *follow-up*.

Método Ederer II

Em 1959, Ederer e Heise, propuseram o método Ederer II [18]. Este método fornece outra forma de calcular a sobrevivência esperada que permite

tempos de *follow-up* heterogéneos e é mais flexível do que o método Ederer I.

Os indivíduos do grupo de correspondência só estão em risco até ao momento em que ocorra morte ou censura dos indivíduos do grupo com cancro. A média da sobrevivência esperada é determinada para os indivíduos com o mesmo período de *follow-up*. Contudo, a sobrevivência esperada depende da mortalidade observada, dependência essa que leva a estimativas potencialmente enviesadas da Sobrevivência Relativa.

Assim, para o cálculo da probabilidade de sobrevivência esperada, apenas teremos que ter conta os indivíduos com diagnóstico de cancro que estejam em risco no início do intervalo. Então, a probabilidade de sobrevivência esperada cumulativa de sobreviver ao fim do intervalo j é dado por:

$$S_E(t) = \prod_{j=1}^{J} \left(\frac{1}{l_j} \sum_{i=1}^{l_j} \tilde{p}_{ij} \right) \tag{2.11}$$

ou seja, no início de cada intervalo j, para todos os indivíduos do grupo de correspondência, determina-se a média das probabilidades esperadas de sobreviver ao fim do intervalo j, para $j = 1, ..., J$ e depois multiplicam-se todas estas médias.

De acordo com a literatura o método Ederer II é o mais recomendado para o cálculo da sobrevivência esperada. Os resultados apresentados por este método são semelhantes ao método Ederer I, só diferem quando o período de *follow-up* é superior a dez anos [26].

Método de Hakulinen

Em 1982, Hakulinen propôs um novo método para calcular a sobrevivência esperada [24]. Neste método, atualmente designado por método Hakulinen, se o tempo de sobrevivência do indivíduo com diagnóstico de cancro for censurado, também o tempo do indivíduo correspondente do grupo de correspondência será. Porém, se o indivíduo com cancro morre, o indivíduo do grupo de correspondência permanece em risco até ao final do período de *follow-up*.

Este método é semelhante ao método de Kaplan-Meier, no sentido em que também tem em conta as observações censuradas. O número de indivíduos com diagnóstico de cancro em risco para cada instante é calculado pelo número de indivíduos vivos e não censurados (indivíduos em risco) esperados, pelo número de observações censuradas esperadas e pelo número de mortes esperadas.

A estimação da Sobrevivência Relativa, [46], é feita através do ajustamento dos tempos de *follow-up* potencialmente heterogéneos pelo potencial tempo de *follow-up*. A sobrevivência esperada não depende da mortalidade observada, mas tem em consideração os tempos de censura possivelmente heterogéneos.

Este método é ideal quando não sabemos o tempo exato de sobrevivência do indivíduo em estudo devido a alguma situação, ou seja, quando estamos sob censura informativa, quando a sobrevivência relativa é constante ao longo dos tempos. Porém, este método é utilizado mesmo quando esta suposição de constância não é verificada, visto que na maioria das situações de cancro isto não acontece [26].

A probabilidade de sobrevivência esperada para o instante $t_{(j)}$, é da forma:

$$p_j^E = 1 - \frac{d_j^E}{n_j^E - \frac{w_j^E}{2}} \tag{2.12}$$

onde, d_j^E é o número total de mortes esperadas durante o instante $t_{(j)}$; n_j^E é o número de indivíduos vivos e não censurados (em observação) esperados no início do instante $t_{(j)}$ e w_j^E é o número de observações censuradas esperadas durante o instante $t_{(j)}$. Os pormenores do cálculo da probabilidade de sobrevivência esperada podem ser consultados no apêndice do artigo [17].

A sobrevivência esperada cumulativa de um grupo de indivíduos com diagnóstico da doença em estudo com as mesmas características do grupo de indivíduos do grupo de correspondência ao fim do intervalo j é dado por:

$$S_E(t) = \prod_{j=1}^{J} p_j^E \tag{2.13}$$

2.4 *Net Survival*

A *net survival*, que pode ser designada por sobrevivência parcial, consiste no cálculo da sobrevivência em que os riscos de morrer devido a outras causas e não devido ao cancro, são excluídas, sendo o estimador determinado através do método Pohar Perme.

O método Pohar Perme surgiu em 2012 e é da autoria de Pohar Perme, Stare e Estève [37]. Este método é imparcial e não enviesado, em relação aos métodos da Análise de Sobrevivência Relativa,[47], uma vez que, a *net survival* é da forma:

$$Net\ Survival = \frac{1}{n}\sum_{i=1}^{n}\frac{S_{Oi}(t)}{S_{Ei}(t)} \tag{2.14}$$

o que não é igual à equação da Sobrevivência Relativa (ver a equação 2.1), porque este método não se baseia numa comparação global entre a sobrevivência observada e a sobrevivência esperada. Contudo, utiliza as tábuas de mortalidade para estimar a mortalidade devido a outras causas, pelo que não necessita de informação sobre a causa de morte.

Assim, a *net survival* pode ser dividida em duas categorias, a *cause-specific setting* quando a causa de morte é explícita, ou seja, é conhecida e perfeitamente identificada e a *relative survival setting* quando a causa de morte é desconhecida ou quando não temos informação na certidão de óbito se a morte foi devido ao cancro.

O método de Pohar Perme foi desenvolvido para tempos de sobrevivência contínuos, contudo nos registos de cancro os tempos de sobrevivência são discretos, nomeadamente os tempos de sobrevivência são em períodos de meses ou anos completos. No entanto, quando implementamos este método em tempos de sobrevivência discretos obtêm-se resultados como se fosse considerado tempos de sobrevivência contínuos [17].

Para estimar a probabilidade de sobrevivência parcial para indivíduos com diagnóstico de cancro, no intervalo j, o método Pohar Perme utiliza o

método atuarial que é dada pela expressão:

$$S_{Lj} = \frac{1 - \frac{d_j^P}{n_j^P - \frac{w_j^P}{2}}}{\exp\left\{ - \frac{\sum_i^{nj} \lambda_j^* - \frac{\sum_i^{wj} \lambda_j^*}{2} - \frac{\sum_i^{dj} \lambda_j^*}{2}}{n_j^P - \frac{\left(d_j^P + w_j^P\right)}{2}} \right\}} \tag{2.15}$$

onde n_j^P corresponde ao número ponderado de indivíduos vivos no início do intervalo j, d_j^P é o número ponderado de mortes durante o intervalo j e w_j^P é o número ponderado de censuras durante o intervalo j. O risco esperado ponderado é representado por λ_j^*. Os valores ponderados são o inverso da probabilidade da sobrevivência esperada acumulada e são calculados no ponto médio de cada intervalo j [17].

Portanto, através do estimador da *net survival* é possível estudar a proporção de indivíduos que morreram por consequências ligadas diretamente ou indiretamente ao diagnóstico de neoplasia ou de outra doença em estudo e fazer comparações de sobrevivência entre países [47].

2.5 Exemplo de Aplicação

Nesta secção vamos exemplificar as abordagens descritas anteriormente sobre a Análise de Sobrevivência Relativa, nomeadamente a estimação da sobrevivência observada e esperada, a análise por coorte e por período.

A Análise de Sobrevivência Relativa foi realizada no programa estatístico R, versão 3.2.5, com recurso ao *package R Commander*. Para proceder à implementação da análise é necessário instalar [1] e carregar [2] o *package relsurv*. A documentação mais recente sobre o *package relsurv* é da autoria de Maja Pohar Perme e Klemen Pavlic de abril de 2016, [38].

Para a realização do exemplo de aplicação é necessário uma base de dados com as características da Sobrevivência Relativa e uma tábua de mortalidade.

A tábua de mortalidade utilizada é referente a Portugal entre os anos 1940 a 2012 e foi obtida através da *Human Mortality Database* (HMD). Para ter acesso à tábua de mortalidade é necessário inscrever-se no site [3] e depois descarregar. Ao descarregar deve ter em atenção alguns procedimentos básicos, como: descarregar as tábuas do sexo masculino e feminino separadamente e salvar em formato `.txt`, e, posteriormente, eliminar a primeira linha do ficheiro. O passo a seguir é carregar no programa estatístico R [4] e está pronta a ser utlizada.

A base de dados selecionada foi a *colrec* do *package relsurv* que corresponde a pacientes com cancro do cólon e do reto, diagnosticados entre 1994 e 2000 e que foi fornecida pelo Slovene Cancer Registry. Esta base de dados é composta por 5971 observações com sete variáveis: `sex` (sexo: 1=masculino e 2=feminino), `age` (idade em dias), `diag` (data do diagnóstico no formato de data), `time` (tempo de sobrevivência em dias), `stat` (indicador de estado: 0=censura e 1=morte), `stage` (estadio do cancro: toma valores 1 a 3 e o valor 99 significa desconhecido) e `site` (local do cancro: rectum=reto ou colon=cólon). Com o intuito de dar uma noção das variáveis desta base de dados, apresenta-se a correspondente estatística descritiva na Tabela 2.2.

[1] `>install.packages("relsurv")`

[2] `>library(relsurv)`

[3] www.mortality.org/

[4] Por exemplo: `porttab<-transrate.hmd(male="mltper_1x1.txt",female="fltper_1x1.txt")`

Tabela 2.2: Estatística descritiva das variáveis

Variável	**Estatística**
Sexo	Masculino: 3289 Feminino: 2682
Idade (em dias)	Mínimo: 4559 ($\approx$ 12 anos) 1º Quartil: 21973 ($\approx$ 60 anos) Mediana: 24864 ($\approx$ 68 anos) Média: 24565 ($\approx$ 67 anos) 3º Quartil: 27493 ($\approx$ 75 anos) Máximo: 35325 ($\approx$ 97 anos)
Data de diagnóstico	01/01/1994 a 30/12/2000
Tempo de Sobrevivência (em dias)	Mínimo: 1 Mediana: 872,0 ($\approx$ 2 anos) Média: 2181,7 ($\approx$ 6 anos) Máximo: 8148 ($\approx$ 22 anos)
Estado	Censura: 992 Morte: 4979
Estadio do cancro	I: 889 II: 3328 III: 1361 Desconhecido: 393
Local do cancro	Reto: 2434 Cólon: 3537
Total de observações	5971

A variável `stage`, estadio do cancro, descreve o grau de severidade do cancro baseado na magnitude do tumor original e na dispersão que possa ter ocorrido pelo corpo. Segundo o Sistema TNM (Tumor, Nódulo, Metástase), a classificação dos tumores e a descrição da extensão anatómica da doença é avaliada em três aspetos, o tumor, os nódulos (ou gânglios linfáticos) e as metástases. O cancro do cólon é um tumor maligno, invasivo, que tem origem

nas células que formam a camada epitelial da parede do intestino grosso e o cancro do reto tem início no reto que corresponde à última porção do intestino grosso. A classificação do estadio varia de acordo com a localização do cancro sendo, para o tumor no cólon e reto, a apresentada na Tabela 2.3.

Tabela 2.3: Classificação clínica do tumor no cólon e reto

Estadio	**T**	**N**	**M**
0	Tis	N0	M0
I	T1-T2	N0	M0
II	T3-T4	N0	M0
III	Qualquer T	N1-N2	M0

A descrição do Sistema TNM (Tumor, Nódulo, Metástase) que se segue foi baseada em [45].

A descrição do tumor, representada pela letra (**T**), dá informação sobre o tamanho e grau de desenvolvimento do tumor primário nos tecidos, no local onde se iniciou:

- Tx - O tumor primário não pode ser avaliado.
- T0 - Não há evidência de tumor primário (não pode ser encontrado).
- Tis - Carcinoma in situ (ou seja, as células do cancro só crescem na camada mais superficial do tecido sem invandir os tecidos mais internos).
- T1, T2, T3 e T4 - Tamanho crescente e/ ou extensão local do tumor primário. Quando maior o número, maior é o tamanho e/ ou extensão.

A descrição do nódulo, representada pela letra (**N**), refere a ausência ou presença e a quantidade de metástases em gânglios regionais:

- Nx - Os gânglios regionais não podem ser avaliados.
- N0 - Ausência de metástases em gânglios regionais.

- N1, N2, N3 - Número crescente dos gânglios regionais afetados. Quanto maior o número, maior é a invasão dos gânglios regionais afetados pelo cancro.

A descrição das metástases, representada pela letra (**M**), consiste na ausência ou presença de metástases em locais mais distantes do corpo:

- Mx - A presença de metástases à distância não pode ser avaliada.

- M0 - Ausência de metástases à distância.

- M1 - Presença de metástases à distância, ou seja, o cancro alastrou a outros órgãos ou tecidos.

A váriavel `diag`, data do diagnóstico, permite determinar o tempo de sobrevivência dos indivíduos. Normalmente, esta data corresponde à data do diagnóstico clínico, mas pode ser a data da primeira consulta ou a data de admissão no hospital, ou a data da primeira confirmação histológica, entre outras. Para o resultado ficar no formato de data só é necessário saber a origem correspondente à tábua de mortalidade que é 01-01-1960 e depois converter utilizado a linha de comando *as.Date*[5] .

Após isto, passamos ao cálculo da sobrevivência esperada, observada e relativa dos 5971 indivíduos com diagnóstico de cancro do reto ou do cólon entre 1994 a 2000. A sobrevivência esperada apresentada nos gráficos é estimada pelo método Ederer II, sendo os resultados para os métodos Ederer I e Hakulinen apresentados nas tabelas do Anexo A.3. A sobrevivência observada é estimada através do estimador de Kaplan-Meier.

[5]Por exemplo: `as.Date(diag,origin="1960-01-01")`

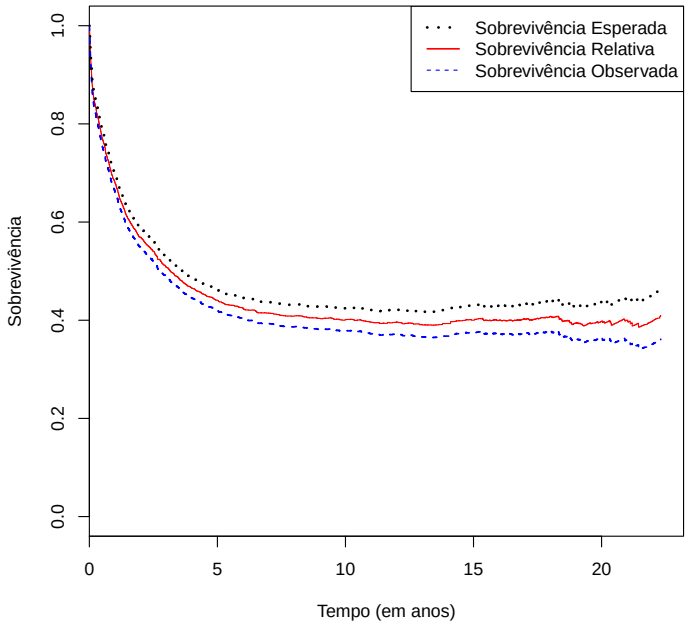

Figura 2.2: Sobrevivência Esperada, Observada e Relativa para o sexo Feminino

Pela Tabela A.1, que encontra-se disponível no Anexo A.2, constata-se que, do total de observações do sexo feminino com diagnóstico de cancro do cólon e reto, foi observado o acontecimento de interesse em 2175 (81%) indivíduos e 507 (19%) tiveram um tempo de vida censurado. Pela Figura 2.2, a estimativa da mediana do tempo de sobrevivência é aproximadamente 3 anos (1155 dias), o que significa que metade dos indivíduos do sexo feminino com este diagnóstico têm pouco mais de 3 anos de vida após o diagnóstico. O intervalo de confiança de 95% para o tempo médio de sobrevivência é de 1023 dias ($\approx$ 2,8 anos) a 1329 dias ($\approx$ 3,64 anos). Para um período de 20 anos, as curvas de sobrevivência esperada e observada apresentam um comportamento semelhante, sendo que nos primeiros 5 anos as curvas de sobrevivência decrescem rapidamente. Após os 5 anos e até aos 14 anos aproximadamente, temos uma estabilização das curvas, a partir desse período a curva de sobrevivência esperada começa a aumentar ligeiramente com algumas oscilações, enquanto que a curva de sobrevivência observada apresenta uma melhoria

até aos 18 anos aproximadamente e depois diminui com algumas variações positivas.

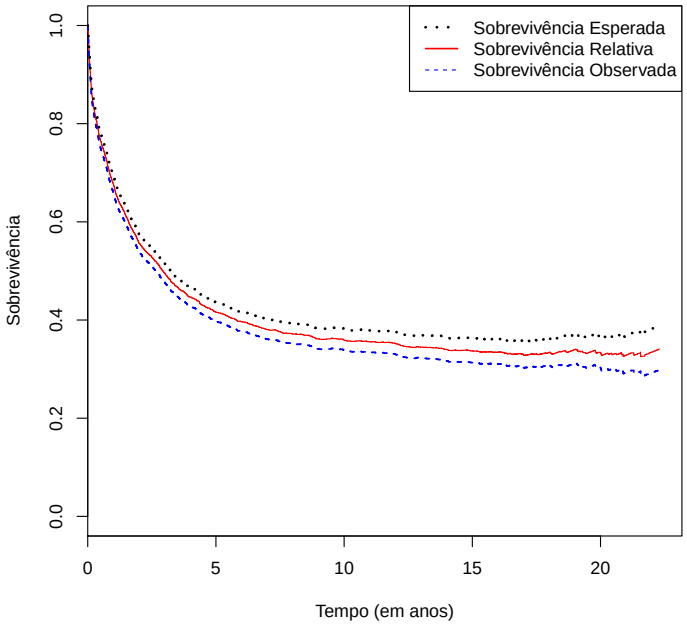

Figura 2.3: Sobrevivência Esperada, Observada e Relativa para o sexo Masculino

Ainda pela mesma tabela, verifica-se que, dos indivíduos diagnosticados com cancro do cólon e reto do sexo masculino, em 2804 (85%)foi observado o acontecimento de interesse e 485 (15%) tiveram um tempo de vida censurado. No que diz respeito à estimativa da mediana do tempo de sobrevivência, pela Figura 2.3 observa-se que é ligeiramente inferior a 3 anos (1072 dias), o que significa que metade dos indivíduos têm menos de 3 anos de vida após o diagnóstico. O intervalo de confiança de 95% para o tempo médio de sobrevivência é de 976 dias ($\approx$ 2,67 anos) a 1179 dias ($\approx$ 3,23 anos). Numa análise às curvas de sobrevivência, observamos que nos primeiros 18 anos a sobrevivência esperada e observada têm um comportamento semelhante, sendo que a partir daí a sobrevivência esperada aumenta muito ligeiramente, enquanto que a sobrevivência observada continua diminuir.

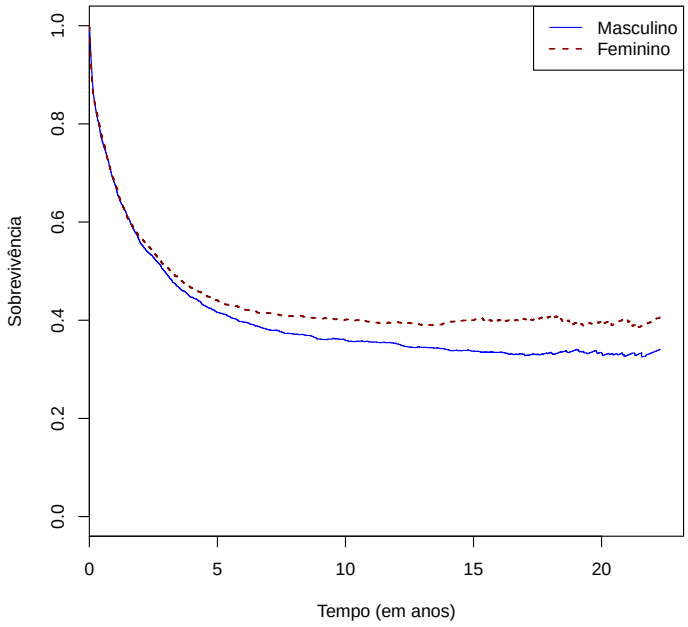

Figura 2.4: Sobrevivência Relativa para ambos os Sexos

Na Figura 2.4, observa-se que a probabilidade de sobrevivência relativa diminui à medida que o tempo aumenta, sendo que nos primeiros três anos as curvas de sobrevivência relativa para ambos os sexos coincidem. Após este período as curvas continuam a diminuir, porém para o sexo feminino entre os 5 e os 14 anos após o diagnóstico a curva apresenta uma estabilização e a partir dos 14 anos, a curva apresenta uma ligeira melhoria assim como algumas oscilações no fim do tempo de sobrevivência. Portanto, a sobrevivência relativa é melhor no sexo feminino do que no sexo masculino. Segundo o Instituto CUF de Oncologia (I.C.O.), o cancro do cólon ou reto é mais comum nos homens, e é o terceiro mais comum a nível global, a seguir ao cancro da mama e da próstata.

Pela Tabela A.3, que encontra-se disponível no Anexo A.2, constata-se que, do total de indivíduos com diagnóstico de cancro do cólon, em 2905 (82%) foi observado o acontecimento de interesse e 632 (18%) tiveram um tempo de vida censurado. Pela Figura 2.5, a estimativa da mediana do

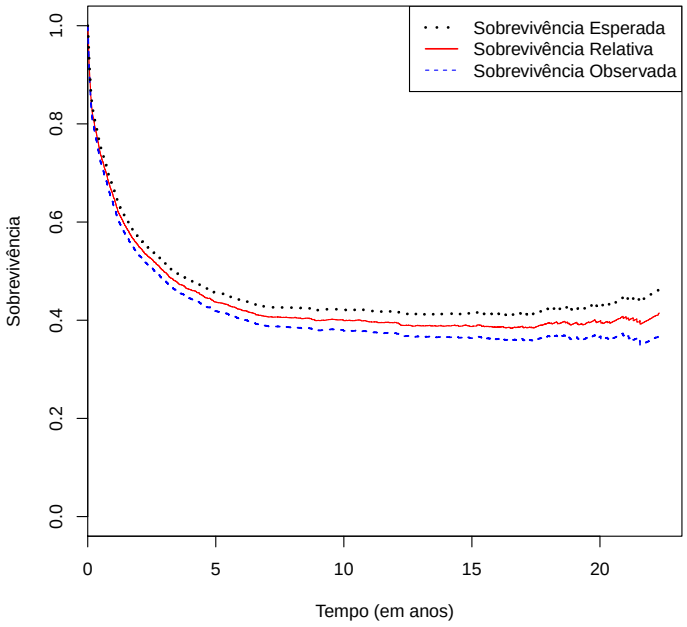

Figura 2.5: Sobrevivência Esperada, Observada e Relativa para o Cólon

tempo de sobrevivência é de aproximadamente 3 anos (1083 dias). O intervalo de confiança de 95% para o tempo médio de sobrevivência é de 968 dias (≈ 2,65 anos) a 1228 dias (≈ 3,36 anos). Nos primeiros anos as curvas de sobrevivência diminuem e no período 6 a 18 anos após o diagnóstico, a sobrevivência esperada e a sobrevivência observada apresentam uma estabilização.

Ainda pela Tabela A.3, verifica-se que, dos 2434 indivíduos com diagnóstico de cancro do reto, em 2074 (85%) foi observado o acontecimento de interesse e 360 (15%) tiveram um tempo de vida censurado. Pela Figura 2.6, a estimativa da mediana do tempo de sobrevivência é de 3 anos (1124 dias). O intervalo de confiança de 95% para o tempo médio de sobrevivência é de 1024 dias (≈ 2,80 anos) a 1242 dias (≈ 3,40 anos). Num período de 20 anos, as curvas de sobrevivência esperada e observada diminuem, a seguir a esse período há uma melhoria.

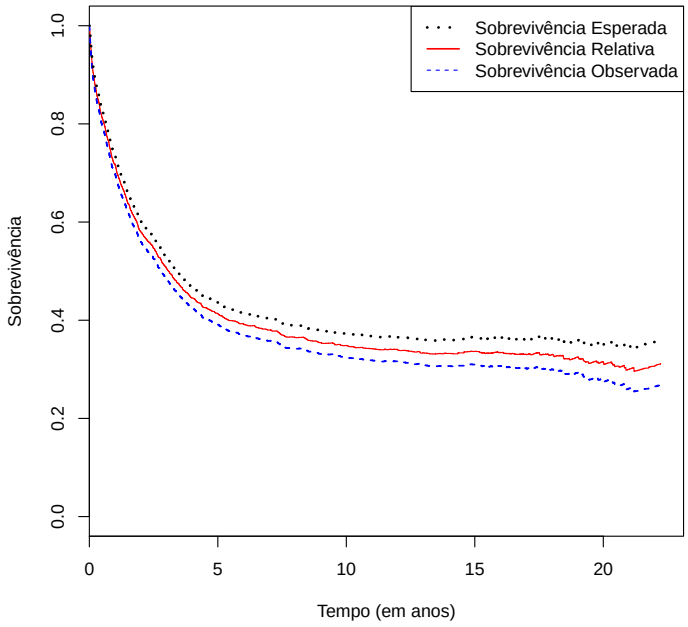

Figura 2.6: Sobrevivência Esperada, Observada e Relativa para o Reto

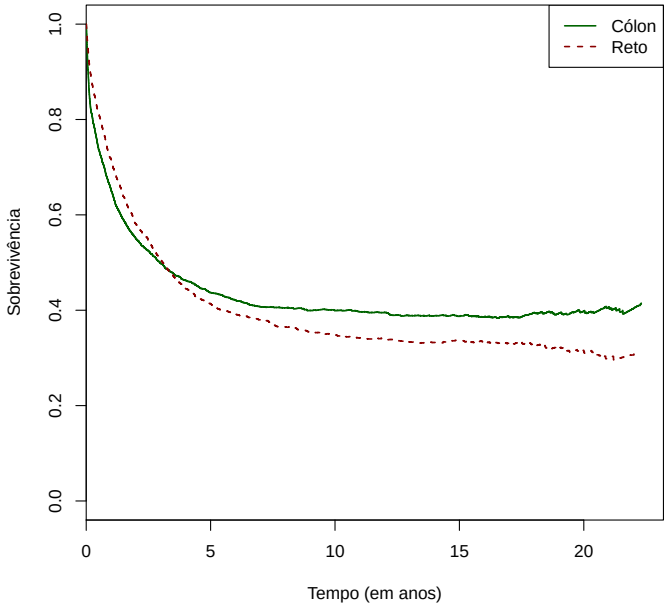

Figura 2.7: Sobrevivência Relativa para o Cólon e o Reto

Pode-se constatar pela Figura 2.7 que nos primeiros anos após o diagnóstico a sobrevivência relativa é melhor para os indivíduos com diagnóstico do reto. A partir dos três anos, a sobrevivência relativa apresenta melhores resultados para os indivíduos com diagnóstico do cólon.

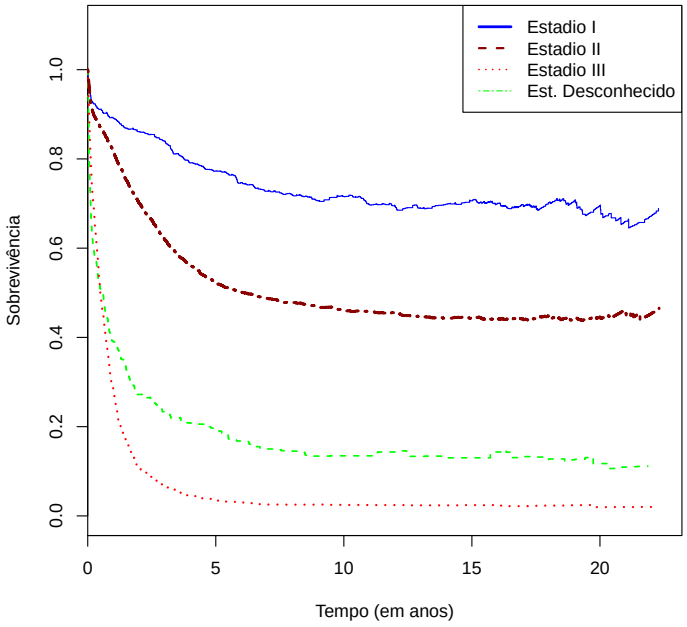

Figura 2.8: Sobrevivência Relativa segundo o estadio

Segundo a Figura 2.8 a sobrevivência relativa é tanto melhor quanto mais cedo ocorrer o diagnóstico, ou seja, quanto menor for o valor do estadio. O estadio III e o estadio desconhecido (ou sem informação) apresentam valores da sobrevivência relativa inferiores aos relativos ao estadio I e estadio II.

A Figura 2.9 representa a sobrevivência esperada para o sexo feminino, tendo em conta os estimadores Ederer I, Ederer II e Hakulinen. Para um período de 20 anos as curvas de sobrevivência esperada para os diferentes métodos são comparadas, sendo que as curvas do método Ederer I e método Hakulinen coincidem. Isto deve-se ao facto de no método Ederer I se considera que os indivíduos do grupo de correspondência estão sempre em risco e no caso de ocorrer morte ou censura do indivíduo com cancro não tem

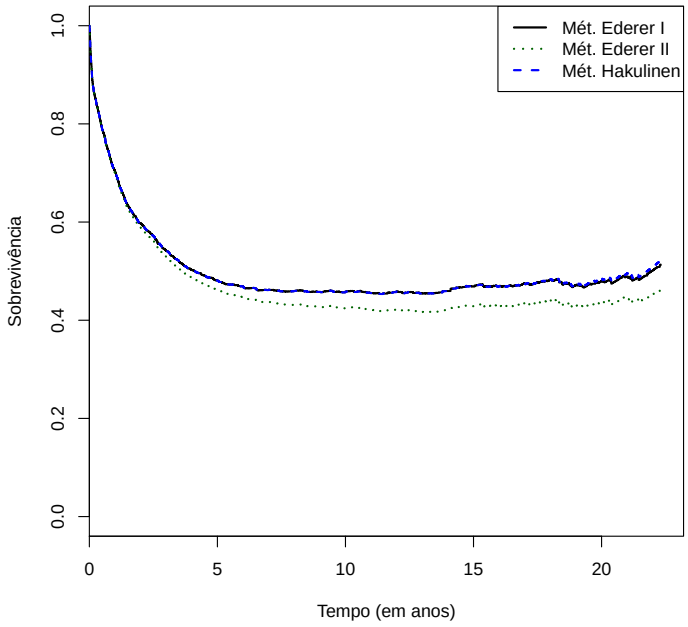

Figura 2.9: Sobrevivência Esperada para o Sexo Feminino com os diferentes métodos

qualquer efeito sobre o indivíduo do grupo de correspondência e, consequentemente, na sobrevivência esperada. Por outro lado, no método de Hakulinen se o tempo de sobrevivência do indivíduo com cancro for censurado, também o tempo do indivíduo do grupo de correspondência será, e para o caso do indivíduo com cancro morrer, o indivíduo do grupo de correspondência permanece em risco até ao final do período de *follow-up*.

No método Ederer II, os resultados são diferentes, visto que os indivíduos do grupo de correspondência só estão em risco até ao momento em que ocorra morte ou censura dos indivíduos com cancro. Para as restantes variáveis em estudo, os resultados dos diferentes métodos podem ser consultados nas tabelas do Anexo A.3.

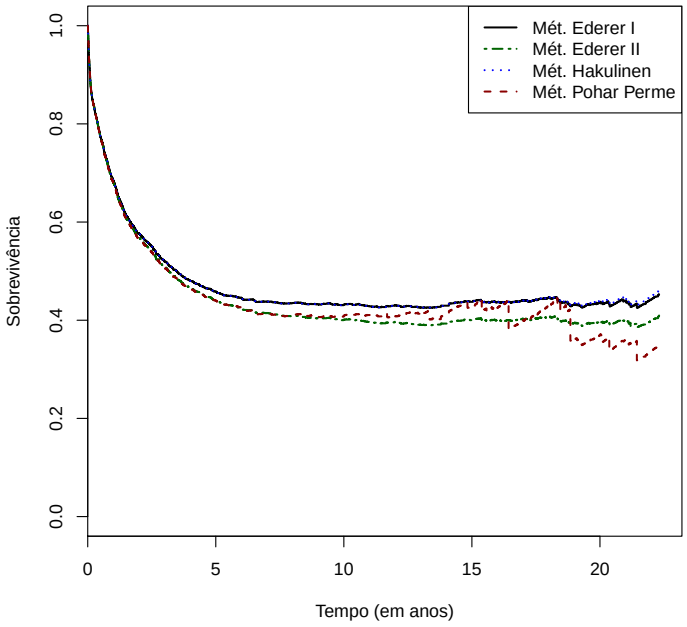

Figura 2.10: Sobrevivência Relativa e *Net Survival* do Sexo Feminino

Na Figura 2.10 visualizamos as curvas de sobrevivência relativa para o sexo feminino, tendo em conta os estimadores para a sobrevivência esperada, nomeadamente, método Ederer I, Ederer II e Hakulinen e a curva da *Net Survival* com o método Pohar Perme.

Constata-se que, a probabilidade de sobrevivência relativa diminui à medida que o tempo aumenta, sendo que o método Ederer I e Hakulinen têm um comportamento idêntico e o método Ederer II também tem um comportamento semelhante, mas com uma sobrevivência inferior. A curva *Net Survival* que corresponde ao método Pohar Perme tem um comportamento idêntico às curvas de sobrevivência relativa, mas entre 6 a 14 anos após o diagnóstico apresenta ligeiro crescimento, e a partir desse período apresenta algumas oscilações sendo estas mais acentuadas entre 16 a 18 anos.

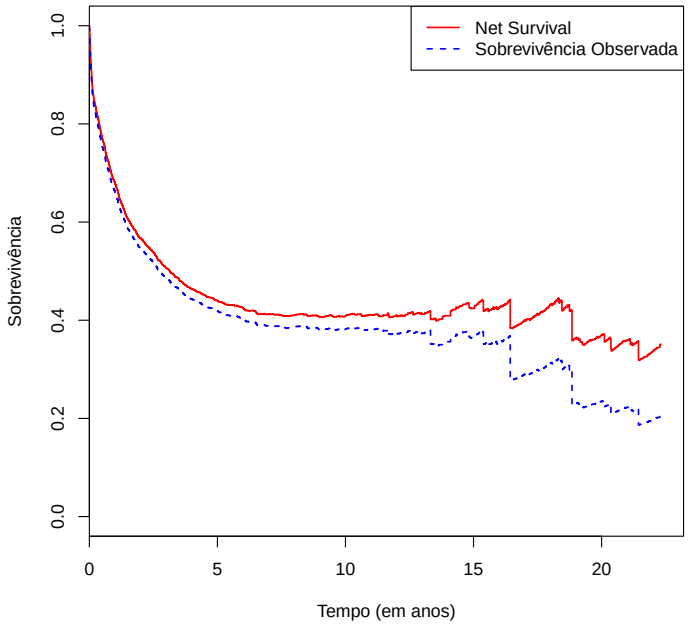

Figura 2.11: *Net Survival* e Sobrevivência Observada para sexo Feminino

A Figura 2.11 representa a curva *Net Survival* pelo método Pohar Perme e a curva de sobrevivência observada para o sexo feminino. Num período de 20 anos verificamos que a curva de sobrevivência observada é muito semelhante à curva *Net Survival*.

Tabela 2.4: Sobrevivência Relativa para os diferentes métodos da Sobrevivência Esperada e *Net Survival* por Sexo aos 10 anos

Sexo	**SR Ederer I**	**SR Ederer II**	**SR Hakulinen**	***Net Survival***
Feminino	0.432	0.401	0.432	0.410
Masculino	0.379	0.359	0.379	0.353

Através da Tabela 2.4 verifica-se que os valores da Sobrevivência Relativa aos 10 anos são iguais para o método Ederer I ou para o método de Hakuli-

nen, sendo que o sexo feminino apresenta melhores resultados do que o sexo masculino. Através destes resultados da Sobrevivência Relativa conseguimos concluir se determinado diagnóstico de cancro ou de outra doença em estudo tem uma elevada taxa de mortalidade em relação à população em geral.

Capítulo 3

Modelos de Regressão da Sobrevivência Relativa

3.1 Introdução

Os modelos de regressão para a Sobrevivência Relativa permitem estudar o efeito das covariáveis (variáveis independentes) na estimação da sobrevivência relativa, sendo que os modelos podem ser de natureza aditiva ou multiplicativa.

Os modelos de regressão aditivos que serão abordados são: o modelo de Hakulinen-Tenkanen [25], de Estève [20] e de Poisson [16]. O modelo de regressão de Andersen [6] será o único modelo multiplicativo considerado. Será ainda referido um modelo que considera uma transformação dos tempos de vida.

O modelo de Hakulinen-Tenkanen é um modelo linear generalizado para dados de sobrevivência agrupados, onde se admite que o número de mortes observadas em cada intervalo de tempo pode ser modelado por uma distribuição binomial. O modelo de Estève é um modelo de regressão para dados de sobrevivência individual e utiliza uma abordagem de verosimilhança baseada em tempos de vida exatos. O modelo de Poisson é um modelo linear generalizado para dados de sobrevivência agrupados ou individuais, no qual o número de mortes utiliza uma estrutura de erro de Poisson.

Os modelos de regressão referidos anteriormente produzem resultados muito semelhantes, porém o modelo de Estève é teoricamente superior ao modelo de Hakulinen-Tenkanen e ao de Poisson, porque o modelo de Estève utiliza os tempos de sobrevivência exatos e não depende de uma aproximação a um modelo binomial ou de Poisson [16]. O modelo de Andersen é um modelo de regressão multiplicativo semelhante ao modelo de Cox.

3.2 Modelo de Regressão Aditivo

O modelo aditivo, também conhecido por modelo de excesso de mortalidade, é formulado pela soma dos riscos,

$$h_O(t) = h_P(t) + h_E(t) \tag{3.1}$$

onde,

- $h_O(t)$ é o risco observado, fornecido pela base de dados;
- $h_P(t)$ corresponde ao risco esperado ou previsto da população de acordo com o risco de mortalidade obtido diretamente das tábuas de mortalidade para um conjunto de indivíduos com as mesmas características do grupo que constitui a base de dados, ao nível do género, faixa etária, ano do diagnóstico, entre outras;
- $h_E(t)$ indica o risco específico ou excesso de risco para certo tipo de neoplasia em estudo, ou seja, representa a função de mortalidade resultante de causa específica.

O risco específico ou excesso de risco é da forma:

$$h_E(t) = h_0(t) \exp(\beta' \mathbf{z}) \tag{3.2}$$

sendo $h_0(t)$ a função de risco subjacente. O risco esperado, $h_P(t)$, é geralmente dependente apenas de algumas covariáveis, tais como idade, sexo e período de diagnóstico, e não depende de covariáveis específicas do tumor, como histologia ou estadio.

Os modelos de regressão aditivos, segundo a literatura, [36, 39], são os modelos mais utilizados na investigação da área da oncologia, e os mais comuns serão descritos em seguida.

3.2.1 Modelo de Hakulinen-Tenkanen

Em 1987, Hakulinen e Tenkanen propuseram um modelo de regressão aditivo para a sobrevivência relativa [25], que é designado por modelo de Hakulinen-Tenkanen. Neste modelo, os indivíduos são agrupados em C estratos, indexados por c, sendo que cada estrato corresponde a uma combinação das covariáveis (idade, sexo, ano de diagnóstico, tipo de neoplasia) e o tempo de *follow-up* é dividido em K intervalos de tempo predefinidos, indexados por k.

No modelo de Hakulinen-Tenkanen, a função de risco observada, condicional ao intervalo $(t_{k-1}, t_k]$, onde $t_0 = 0$ e $t_K = \infty$, é dada por:

$$h_O(t_k; \mathbf{z}) = h_P(t_k; \mathbf{z}_1) + \exp(\alpha_k(t) + \beta' \mathbf{z}) \tag{3.3}$$

onde,

- $\mathbf{z}$ é um vetor de covariáveis;
- $\mathbf{z}_1$ é um subvetor do vetor de covariáveis (por exemplo: idade e sexo);
- β' é o vetor dos coeficientes de regressão;
- $\alpha_k(t)$ é uma função dependente do tempo.

Neste modelo, o risco específico pode não ser constante em cada intervalo, contrariamente à abordagem de Poisson e o número de mortes segue uma distribuição binomial. Se considerarmos as funções de sobrevivência observada e de sobrevivência esperada, $S_O(t_k, \mathbf{z})$ e $S_E(t_k, \mathbf{z}_1)$ respetivamente, para cada intervalo k, obtemos:

$$\ln \left[-\ln \left(\frac{S_O(t_k, \mathbf{z})}{S_E(t_k, \mathbf{z}_1)} \right) \right] = \beta' \mathbf{z} + \gamma_k \tag{3.4}$$

sendo,

$$\gamma_k = \ln\left[\int_{t_{k-1}}^{t_k} \exp(\alpha_k(t))dt\right] \tag{3.5}$$

o que resulta um modelo linear generalizado com estrutura de erro binomial e com uma função de ligação $\ln(-\ln)$ combinada com a divisão por $S_E(t_k, \mathbf{z}_1)$.

O modelo de Hakulinen-Tenkanen caracteriza-se por considerar a não proporcionalidade dos riscos, especialmente em estudos com pouca informação [22]. Segundo Giorgi et al., [23], este modelo é adequado para os tempos de vida agrupados num intervalo de tempo $(t_k, t_{k+1}]$, ou seja, quando os tempos de acompanhamento não são conhecidos com exatidão.

3.2.2 Modelo de Estève

O modelo de Estève é um modelo de regressão que permite estimar a Sobrevivência Relativa com base nos valores individuais. O risco observado, no modelo de Estève, é dado por:

$$\begin{aligned} h_O(t; \mathbf{z}, x) =& h_P(x+t; \mathbf{z}_1) + \sum_{k=1}^{K} \tau_k I_k \exp(\beta' \mathbf{z}) \\ =& h_P(x+t; \mathbf{z}_1) + \sum_{k=1}^{K} h_0(t_k) \exp(\beta' \mathbf{z}) \end{aligned} \tag{3.6}$$

onde,

- x corresponde à idade dos indivíduos no momento em que entram para o estudo;

- $\mathbf{z}_1$ é um subvetor do vetor das covariáveis (por exemplo, formado pelas covariáveis idade e sexo);

- τ_k é a taxa de mortalidade específica que é assumida como constante em cada intervalo de tempo k correspondente aos indivíduo com vetor de covariáveis nulo;

- I_k é a função indicatriz que toma o valor 1 quando t pertence ao intervalo k ou o valor zero, caso contrário;

- $h_0(t_k)$ é a função de risco subjacente, ou seja, representa a função de risco correspondente a um indivíduo com vetor de covariáveis nulo ($\mathbf{z}=\mathbf{0}$), que se assume constante em cada intervalo de tempo k.

O risco esperado, $h_P(x+t;\mathbf{z}_1)$, é determinado para cada indivíduo, sendo que é considerado constante em cada intervalo k. Este risco coincide com os valores das tábuas de mortalidade para indivíduos com subvetor do vetor das covariáveis ($\mathbf{z}_1$), idade ($x+t$) e o tempo de vida representado por t.

No modelo de Estève, os coeficientes de regressão são obtidos através do método de máxima verosimilhança (ver mais pormenores em [16, 39]). Este modelo é uma alternativa ao modelo de regressão de Hakulinen-Tenkanen, nomeadamente quando a base de dados em estudo é pequena ou quando o número de indivíduos por estrato é pequeno [23]. Uma vantagem do modelo de Estève é que ao nível teórico é superior ao restantes modelos de regressão aditivos, porque utiliza tempos de vida exatos. Quando existem covariáveis dependentes do tempo, algo comum em oncologia, para este modelo ainda não existe um *software* que englobe esta situação, o que constitui uma desvantagem, em especial se os riscos não são proporcionais [16].

O modelo de Estève difere do modelo de Hakulinen-Tenkanen, [23], no modo como estima os parâmetros, ou seja, o modelo de Estève usa a verosimilhança total com base nos valores individuais, enquanto que o modelo de Hakulinen-Tenkanen, no contexto dos modelos lineares generalizados, usa os dados agrupados. Neste último modelo pode ocorrer pouca precisão na estimação dos paramêtros e pouca potência quando o número de indivíduos por estrato é pequeno. Por outro lado, no modelo de Estève, os indivíduos podem ser ajustados individualmente, como no modelo de Cox, o que elimina os possíveis problemas resultantes do agrupamento de indivíduos heterogéneos em termos da sobrevivência quando a análise é realizada em dados agrupados. Portanto, a escolha do modelo de regressão depende da estrutura dos dados, ou seja, se os dados forem registados em intervalos de tempo optamos pelo o modelo de Hakulinen-Tenkanen, mas, por outro lado, se os dados

forem individuais devemos escolher o modelo de Estève.

3.2.3 Modelo de Poisson

O modelo de Poisson é um modelo linear generalizado que assume que a função de risco é constante em cada intervalo e que o número de mortes segue uma distribuição de Poisson. O risco específico pode alterar-se, especialmente no primeiro ano de *follow-up*, porque na maioria dos casos de neoplasia a maior mudança ocorre no início do período de *follow-up*. Tipicamente os intervalos são anuais, algo comum com a escala de tempo das estimativas da tábua de mortalidade, embora seja possível ajustar o modelo em intervalos mensais (por exemplo, no primeiro ano de *follow-up*) e depois em intervalos anuais.

O modelo de Poisson é semelhante ao modelo de Hakulinen-Tenkanen quando os dados são agrupados onde o risco específico é igual (ver equação (3.3)). Porém, quando os dados são individuais, ou seja, os tempos de sobrevivência são exatos os resultados são semelhantes ao modelo de Estève.

O número observado de mortes, d_{ck}, no estrato c no intervalo k segue uma distribuição de Poisson, $d_{ck} \sim$ Poisson (μ_{ck}), onde:

$$\mu_{ck} = h_{E,ck} y_{ck} \tag{3.7}$$

onde y_{ck} corresponde ao tempo em risco do indivíduo (*person-time at risk*) do grupo c no intervalo k e $h_{E,ck}$ é o risco específico. Sendo o número esperado de mortes d^E_{ck}, a equação (3.1) é escrita na forma:

$$\frac{\mu_{ck}}{y_{ck}} = \frac{d^E_{ck}}{y_{ck}} + \exp(\beta' \mathbf{z}) \tag{3.8}$$

ou ainda,

$$\ln(\mu_{ck} - d^E_{ck}) = \ln(y_{ck}) + \beta' \mathbf{z} \tag{3.9}$$

O modelo de Poisson é um modelo que pode ser estimado através de tempos de sobrevivência individuais ou agrupados. Porém, devemos dar preferência, sempre que possível, aos tempos de sobrevivência individuais e

não aos tempos de sobrevivência agrupados [16].

3.3 Modelo de Regressão Multiplicativo

O modelo multiplicativo é formulado pelo produto das funções de riscos, a esperada e a relativa, ou seja, é dado por:

$$h_O(t) = h_P(t)h_E(t) = h_P(t)h_R(t) \tag{3.10}$$

onde $h_R(t)$ é o risco relativo que corresponde ao quociente entre o risco observado e o risco esperado, $h_R(t) = \frac{h_O(t)}{h_P(t)}$. Este modelo não assume que o risco observado é sempre maior do que o risco esperado.

3.3.1 Modelo de Andersen

Em 1985, Andersen et al., [6], propuseram um modelo multiplicativo, designado por modelo de Andersen, onde o risco específico é representado do seguinte modo:

$$h_E(t) = h_0(t)\exp(\beta'\mathbf{z}) \tag{3.11}$$

ou seja, tem uma representação análoga à do modelo de Cox (ver equação 1.23).

Assim sendo, o risco observado é definido por:

$$h_O(t) = h_P(t)h_E(t) = h_P(t)h_0(t)\exp(\beta'\mathbf{z}) \tag{3.12}$$

podendo ser reescrito da seguinte forma:

$$h_O(t) = h_0(t)\exp[\beta'\mathbf{z} + \ln(h_P(t))] \tag{3.13}$$

o que constitui um modelo de Cox com uma variável adicional dependente do tempo. Assim, o ajustamento do modelo é feito através de procedimentos do modelo de Cox, onde os dados podem ser divididos em intervalos de um ano, com $h_P(t)$ a ser atualizado em cada intervalo.

3.4 Modelo de transformação dos tempos de vida

Uma outra abordagem consiste em transformar os tempos de vida dos indivíduos da seguinte forma:

$$y = F_P(t) \tag{3.14}$$

onde $F_P(t)$ representa a função de distribuição cumulativa de um indivíduo com determinada idade, género e ano de diagnóstico, de modo a ter as mesmas características do indivíduo da população em geral e os valores de y correspondem ao tempo de vida transformado para cada indivíduo. A função de distribuição cumulativa $F_P(t)$ é calculada a partir das tábuas de mortalidade.

Com esta transformação, é possível remover as diferenças entre a sobrevivência devidas à idade, sexo e ano de diagnóstico, ficando apenas o risco específico por determinar. Assim sendo, podemos modelar diretamente o risco específico e uma possibilidades é através do modelo de Cox, ou seja,

$$h_E(y) = h_0(y) \exp(\beta' \mathbf{z}) \tag{3.15}$$

Esta abordagem é recente [39] e, devido à sua simplicidade, os autores consideram que no futuro se pode tornar popular, em especial em estudos com observações de longo prazo.

Segundo Stare et al., [48], o modelo de transformação dos tempos de vida apresenta algumas vantagens em relação aos modelos de regressão aditivos (Hakulinen-Tenkanen, Estève e Poisson) e ao modelo multiplicativo, nomeadamente, não necessita de um *software* especial para ser aplicado, a comparação entre a base de dados e a população em geral é mais fácil e o uso da metodologia de regressão, em particular, do modelo de Cox é direto. Porém, uma possível desvantagem é a interpretação dos resultados, porque após a transformação dos tempos de vida as conclusões não são tão diretas.

3.5 Exemplo de Aplicação

Nesta secção vamos exemplicar as abordagens descritas para os modelos de regressão aditivos, multiplicativo e transformado, com a mesma base de dados e com o programa estatístico usado no exemplo de aplicação da secção 2.5.

Os modelos de regressão permitem determinar quais as covariáveis que são significativas. Inicialmente, vamos considerar todas as covariáveis: `sex` (Sexo), `stage` (Estadio) e `site` (Local), no entanto, as covariáveis `age` (Idade) e `diag` (Diagnóstico) serão recodificadas em covariáveis categóricas (os procedimentos estão disponíveis no Anexo B, secção B.1).

Após isto, começamos com a implementação dos modelos de regressão aditivos, sendo que dos três modelos abordados iniciámos com o modelo de Estève (os procedimentos podem ser consultados no Anexo B, subsecção B.1.1).

Na Tabela 3.1, as categorias de referência para cada covariável são: Idade - 12,5-60,2 anos; Sexo - masculino; Diagnóstico - 01-01-1994 a 30-09-1995; Estadio - I e Local - colón. No modelo de Estève, de todas as covariáveis consideradas, apenas duas não foram significativas para o nível de significância a 5%: Sexo (valor p=0.155) e Local (valor p=0.085). Relativamente, à covariável Local vamos incluir na análise, porque consideramos que é uma covariável de interesse para o estudo. Quanto à covariável Sexo, não iremos incluir no estudo, porque ao realizarmos o teste t de Student para as observações correspondentes aos tempos de vida observados, com a finalidade de comparar as diferenças entre o sexo feminino e sexo masculino em relação ao tempo de sobrevivência, concluímos que não existe diferenças estatisticamente significativas pois obteve-se o valor p=0.369. Contudo, a covariável Sexo apresenta um coeficiente de (- 0.053), o que é uma indicação de uma possível tendência de que o tempo de sobrevivência seja melhor para sexo o feminino do que para o masculino.

O próximo passo consiste em implementar o modelo de Estève com as covariáveis que foram significativas e com a covariável Local (ver a Tabela B.2, disponível no Anexo B, subsecção B.1.1). Após isto, voltamos a repetir os mesmos procedimentos para os restantes modelos de regressão aditivos: o

modelo de Poisson e o modelo de Hakulinen-Tenkanen, onde os resultados são apresentados no Anexo B, subsecções B.1.2 e B.1.3, com as respetivas Tabelas B.3, B.4, B.5 e B.6. Estas tabelas dão informação sobre as estimativas dos coeficientes, desvio padrão e valor p. Relativamente ao modelo de Poisson, as covariáveis não significativas são as mesmas do modelo de Estève, sendo a covariável Sexo (valor p=0.151) e Local (valor p=0.085). Por outro lado, no modelo de Hakulinen-Tenkanen a única covariável não significativa foi a covariável Sexo (valor p=0.190), enquanto as restantes foram significativas. Note-se que, com este modelo, a covariável Local passou a ser significativa (valor p=0.002), o que veio a reforçar termos feito uma boa opção em não eliminar a covariável do estudo.

A Tabela 3.2 (construída a partir das Tabelas B.2, B.4 e B.6, disponíveis no Anexo B, subsecções B.1.1, B.1.2 e B.1.3, respetivamente) apresenta os resultados das estimativas dos coeficientes de regressão e os desvios padrão para os três modelos de regressão. As estimativas obtidas através dos três modelos aditivos são muito próximas, o que faz com que as conclusões sejam as mesmas. No entanto, pode-se observar que, em geral, os desvios padrão obtidos com o modelo de Estève têm um valor inferior, o que está de acordo com o facto de este modelo ser aquele que apresenta estimativas mais precisas. Relativamente às covariáveis verificámos que a idade é uma covariável significativa para o estudo ($p \leq 175e-07$) [1] e, além disso, o tempo de vida diminui à medida que a idade aumenta. Por exemplo, se compararmos a faixa etária 75,4-96,7 com a faixa etária de referência [2] verifica-se que o risco de morte é cerca de 2,208 (=exp(0.792))[1] do risco de morte dos indivíduos mais jovens. Quanto ao diagnóstico os resultados são estatisticamente significativos a partir de 01-10-1995, ou seja, 21 meses depois da data do primeiro diagnóstico (01-01-1994), o que pode representar uma melhoria nos tratamentos. No estadio, verificámos que o estadio I apresenta melhores resultados em relação aos restantes estadios, sendo que os piores resultados encontram-se no estadio III, depois no estadio desconhecido e, por fim, no estadio II. Relativamente, ao tipo de cancro, constatámos que o cancro do cólon manifesta

[1] Modelo de Estève
[2] Idade 12,5-60,2

uma tendência para melhores resultados do que o cancro do reto, apesar de não ser significativo. As estimativas dos coeficientes do período de *follow-up* são semelhantes para os três modelos, sendo que o risco é maior no primeiro ano (exp(fu [0,1])=0.075) e depois vai diminuindo ao longo do período de seguimento.

Após a aplicação dos modelos de regressão aditivos, vamos agora aplicar o modelo de regressão multiplicativo, ou seja, o modelo de Andersen. Para começar vamos considerar todas as covariáveis e verificar quais as que são significativas para o modelo (os procedimentos podem ser consultados no Anexo B, secção B.2). Na Tabela 3.3, as categorias de referência para cada covariável são: Idade - 12,5-60,2 anos; Sexo - masculino; Diagnóstico - 01-01-1994 a 30-09-1995; Estadio - I e Local - colón. Neste modelo, todas as covariáveis são significativas, contrariamente ao que ocorreu nos modelos de regressão aditivos. Em relação à covariável Idade, verificámos que os indivíduos que pertencem à faixa etária 60,3-68,1 anos têm um risco de morte superior aos indivíduos mais velhos ($\geq$ 68,2 anos), ou seja, o risco de morte diminui com idade. Relativamente, à covariável Sexo constatámos que é significativa ($p <$ 2e-16) e a estimativa do coeficiente (0.444) permite afirmar que o sexo masculino tem uma tendência para um tempo de vida melhor do que no sexo feminino, contrariamente ao que aconteceu no modelo de Estève. No que diz respeito à covariável Diagnóstico podemos afirmar que à medida que a data do diagnóstico aumenta, o risco de morte diminui. Quanto à covariável Estadio o risco aumenta à medida que o estadio aumenta, sendo o estadio III o que apresenta maior risco de morte. Na covariável Local, uma vez que a categoria de referência é o cólon, a estimação do coeficiente obtida permite afirmar que esta apresenta menor risco de morte. O risco de morte associado ao cancro do reto é de 1.103 em relação ao do cancro do cólon.

Constatámos ainda pela análise dos diferentes *outputs* do modelo de Andersen para os diferentes períodos de *follow-up* que a dimensão da amostra alterava-se em cada período, sendo que em cada período de *follow-up* a base de dados (n) é formada pelo número inicial de observações da base de dados (5971) mais a soma das observações que não verificam o acontecimento de interesse no período anterior.

Depois do estudo dos modelos de regressão aditivos (Estève, Poisson, Hakulinen) e do modelo de regressão multiplicativo de Andersen, vamos agora estudar o modelo Transformado. Iniciámos o estudo com todas as covariáveis para verificar quais as covariáveis que são significativas (os procedimentos podem ser consultados no Anexo B, secção B.3). No modelo Transformado todas as covariáveis são significativas e as maiores mudanças estão associadas às covariáveis presentes nas tábuas de mortalidade, nomeadamente a covariável sexo, idade e ano de diagnóstico.

Segundo a Tabela 3.4, as categorias de referência para cada covariável são: Idade - 12,5-60,2 anos; Sexo - masculino; Diagnóstico - 01-01-1994 a 30-09-1995; Estadio - I e Local - colón. Assim, comparando o modelo Transformado com o modelo de regressão de Andersen verificámos que na covariável Sexo o nível de significância diminui e o valor da estimativa de coeficiente mostra que o risco de morte é menor no sexo masculino. Quanto à covariável Idade, a faixa etária 60,3-68,1 tornou-se menos significativa em relação às restantes faixas etárias que mantiveram o mesmo nível de significância. Em relação ao risco de morte, as correspondentes estimativas aumentaram em todas as faixas etárias, porém mantém-se a conclusão de que à medida que a idade aumenta, o risco de morte diminui. Relativamente, à covariável Diagnóstico tornou-se mais significativa e houve uma melhoria no risco, que pode ser justificada pela remoção do risco associado a outras doenças, como por exemplo, diabetes, hipertensão arterial, acidente vascular cerebral, entre outros, visto que o acontecimento de interesse é a morte por causa do cancro.

Após a análise dos modelos de regressão da Análise de Sobrevivência Relativa, vamos agora estudar a proporcionalidade das funções de risco, através da função `rs.br` disponível no *package relsurv*, onde podemos testar se as funções de risco são ou não proporcionais. Os resultados dos testes dos modelos de regressão aditivos (Estève, Poisson, Hakulinen), do modelo de regressão multiplicativo de Andersen e do modelo transformado, disponíveis no Anexo B, secção B.4, permitem concluir que não existe proporcionalidade. Note-se que no modelo transformado, para estudar a hipótese de riscos proporcionais as únicas covariáveis que podemos utilizar são: idade, ano de diagnóstico e sexo, porque são as covariáveis presentes nas tábuas de mortalidade.

Tabela 3.1: Modelo de Estève com todas as variáveis da base de dados

	Est. dos coeficientes	Desvio padrão	Valor p
Idade 60,3-68,1	0.266	0.051	2.33e-07
Idade 68,2-75,3	0.382	0.052	1.81e-13
Idade 75,4-96,7	0.799	0.053	< 2e-16
Sexo Feminino	- 0.053	0.037	0.155
Diagnóstico: 01-10-1995 a 30-06-1997	-0.106	0.052	0.042
Diagnóstico: 01-07-1997 a 31-03-1999	-0.206	0.052	6.59e-05
Diagnóstico: 01-04-1999 a 30-12-2000	-0.247	0.051	1.60e-06
Estadio II	0.878	0.087	< 2e-16
Estadio III	2.670	0.089	< 2e-16
Estadio Desc.	1.922	0.103	< 2e-16
Local Reto	0.065	0.037	0.085
fu [0,1]	- 2.515	0.114	< 2e-16
fu (1,2]	- 2.893	0.116	< 2e-16
fu (2,3]	- 3.263	0.123	< 2e-16
fu (3,4]	- 3.367	0.131	< 2e-16
fu (4,5]	- 3.733	0.146	< 2e-16
fu (5,6]	- 4.083	0.171	< 2e-16
fu (6,7]	- 4.368	0.200	< 2e-16
fu (7,8]	- 4.700	0.247	< 2e-16
fu (8,9]	- 4.809	0.272	< 2e-16
fu (9,10]	- 5.259	0.365	< 2e-16

Tabela 3.2: Comparação das estimativas dos coeficientes e dos desvios padrão para os três modelos de regressão aditivos

	Estève	**Poisson**	**Hakulinen**
Idade 60,3-68,1	0.269 (0.051)	0.267 (0.051)	0.257 (0.052)
Idade 68,2-75,3	0.378 (0.052)	0.378 (0.052)	0.349 (0.053)
Idade 75,4-96,7	0.792 (0.053)	0.797 (0.053)	0.728 (0.056)
Diagnóstico: 01-10-1995 a 30-06-1997	-0.105 (0.052)	-0.105 (0.052)	-0.104 (0.054)
Diagnóstico: 01-07-1997 a 31-03-1999	-0.204 (0.052)	-0.203 (0.052)	-0.184 (0.053)
Diagnóstico: 01-04-1999 a 30-12-2000	-0.244 (0.051)	-0.242 (0.051)	-0.253 (0.053)
Estadio II	0.878 (0.087)	0.889 (0.088)	0.901 (0.088)
Estadio III	2.674 (0.089)	2.673 (0.089)	2.690 (0.091)
Estadio Desc.	1.925 (0.103)	1.927 (0.103)	1.857 (0.105)
Local Reto	0.067 (0.037)	0.066 (0.037)	0.122 (0.039)
fu [0,1]	-2.593 (0.100)	-2.592 (0.099)	-2.663 (0.101)
fu (1,2]	-2.972 (0.102)	-2.982 (0.103)	-2.996 (0.104)
fu (2,3]	-3.342 (0.110)	-3.346 (0.111)	-3.341 (0.112)
fu (3,4]	-3.447 (0.118)	-3.457 (0.117)	-3.465 (0.119)
fu (4,5]	-3.814 (0.134)	-3.827 (0.135)	-3.835 (0.136)
fu (5,6]	-4.165 (0.161)	-4.170 (0.161)	-4.185 (0.162)
fu (6,7]	-4.453 (0.192)	-4.452 (0.192)	-4.474 (0.194)
fu (7,8]	-4.785 (0.241)	-4.779 (0.241)	-4.779 (0.241)
fu (8,9]	-4.894 (0.266)	-4.868 (0.264)	-4.867 (0.264)
fu (9,10]	-5.346 (0.361)	-5.390 (0.393)	-5.383 (0.389)

Tabela 3.3: Modelo de Regressão Multiplicativo - Modelo de Andersen

	Est. coeficientes	Risco	Desvio padrão	Valor p
Idade 60,3-68,1	-0.695	0.499	0.046	$<$ 2e-16
Idade 68,2-75,3	-1.246	0.288	0.045	$<$ 2e-16
Idade 75,4-96,7	-1.807	0.164	0.045	$<$ 2e-16
Sexo Feminino	0.444	1.560	0.031	$<$ 2e-16
Diagnóstico: 01-10-1995 a 30-06-1997	-0.086	0.918	0.044	0.050638
Diagnóstico: 01-07-1997 a 31-03-1999	-0.113	0.893	0.043	0.008599
Diagnóstico: 01-04-1999 a 30-12-2000	-0.149	0.862	0.043	0.000581
Estadio II	0.533	1.704	0.052	$<$ 2e-16
Estadio III	2.128	8.399	0.058	$<$ 2e-16
Estadio Desc.	1.220	3.386	0.072	$<$ 2e-16
Local Reto	0.098	1.103	0.031	0.001680

Tabela 3.4: Modelo Transformado

	Est. dos coeficientes	Risco	Desvio padrão	Valor p
Idade 60,3-68,1	-0.304	0.738	0.047	8.21e-11
Idade 68,2-75,3	-0.509	0.601	0.047	$<$ 2e-16
Idade 75,4-96,7	-0.583	0.558	0.048	$<$ 2e-16
Sexo Feminino	0.181	1.199	0.031	7.78e-09
Diagnóstico: 01-10-1995 a 30-06-1997	-0.086	0.918	0.044	0.05193
Diagnóstico: 01-07-1997 a 31-03-1999	-0.136	0.872	0.043	0.00153
Diagnóstico: 01-04-1999 a 30-12-2000	-0.173	0.841	0.043	6.11e-05
Estadio II	0.528	1.696	0.052	$<$ 2e-16
Estadio III	2.067	7.900	0.057	$<$ 2e-16
Estadio Desc.	1.272	3.568	0.072	$<$ 2e-16
Local Reto	0.091	1.095	0.031	0.00355

Capítulo 4

Conclusões e considerações finais

Esta dissertação teve como principal objetivo reunir conceitos, modelos e técnicas fundamentais para aplicar a Análise de Sobrevivência Relativa.

Como a Análise de Sobrevivência Relativa é uma subárea da Análise de Sobrevivência, optámos por começar com uma explicação dos conceitos e modelos fundamentais da Análise de Sobrevivência, com intuito de reunir toda a informação importante para a compreensão da Análise de Sobrevivência Relativa. Esta última foi abordada no segundo e terceiro capítulos e terminou com um exemplo de aplicação. Este exemplo teve a finalidade de colocar em prática os conceitos mencionados na Análise de Sobrevivência Relativa, e assim fornecer aos leitores desta dissertação um documento de apoio à realização de um estudo desta natureza. Com este objetivo, apresentámos os procedimentos teóricos e práticos necessários para a realização do estudo, os quais podem ser adaptados para qualquer outra base de dados.

Mas, antes de começar com o exemplo tivemos que encontrar uma base de dados e uma tábua de mortalidade. A seleção da base de dados foi uma tarefa que demorou algum tempo, visto que tinhamos que encontrar um conjunto de dados que fosse adequada às características da Sobrevivência Relativa, nomeadamente, a data do diagnóstico tinha que ser em formato de data, a idade e o tempo de vida em dias, e também para ser diferente do exemplo

apresentado pelo *package relsurv*. Neste sentido, a opção de trabalhar no programa estatístico R foi uma vantagem, porque apesar de ser gratuito e de acesso fácil, disponibiliza diversas bases de dados que permite ao utilizador selecionar a mais apropriada para o estudo, e ainda, fornece documentação sobre os códigos e instruções necessárias para a Análise de Sobrevivência Relativa. A escolha incidiu sobre a base de dados *colrec*, encontrada na biblioteca *relsurv* do programa estatístico R, que corresponde aos indivíduos com diagnóstico de cancro do reto e do cólon entre 1994 a 2000.

Quanto à tábua de mortalidade foi obtida do site da *Human Mortality Database* (HMD), sendo que a escolha incidiu sobre Portugal e referente aos anos 1940 a 2012. A utilização deste *site* foi um ponto forte no decorrer do trabalho, porque permitiu-nos um acesso fácil e gratuito às tábuas de mortalidade relativas a 38 países. Atualmente, novembro de 2017, o *site* já tem informação de 39 países (entretanto foi acrescentado a Croácia) e as tábuas de mortalidade de Portugal já se encontram atualizadas até 2015 (anteriormente estavam só até 2012).

Depois disto, começámos com o estudo da Análise de Sobrevivência Relativa, sendo que após uma exploração da base de dados e de algumas aplicações surge a primeira limitação ao nível da análise por período, na qual não foi possível realizar, uma vez que, o programa estatístico R, atualmente, não tem disponível o *package periodR*. No entanto, tentámos superar esta dificuldade com várias pesquisas para encontrar o código, até instalámos versões mais antigas do programa estatístico R, visto que o artigo [28] apresenta os procedimentos de execução, mas sem sucesso. Contudo, esta análise pode ser implementada com recurso a *software* estatístico não gratuito, o que nós consideramos uma desvantagem.

A segunda limitação surge nos modelos de regressão da Análise de Sobrevivência Relativa que foi verificarmos a não existência de riscos proporcionais. Assim, apenas apresentamos os procedimentos e os resultados dos modelos de regressão aditivos, do modelo multilplicativo e do modelo de transformação dos tempos de vida.

Portanto, com a realização desta dissertação desejámos contribuir para a compilação da informação sobre este tema e auxiliar no desenvolvimento de

trabalhos futuros sobre Análise de Sobrevivência Relativa.

Futuramente, seria interessante explorar um pouco mais esta base de dados, por exemplo, aplicando outro tipo de modelos de regressão que não de riscos proporcionais, de modo a conseguir obter resultados mais satisfatórios. Além disso, também seria relevante aplicar estes conceitos numa base dados reais, pois assim poderíamos obter conclusões mais interessantes e dessa forma contribuir para a descoberta de possíveis razões para as eventuais elevadas taxas de mortalidade associadas à doença em estudo ou contribuir para a prevenção e tratamento.

Anexos

Anexo A

Análise de Sobrevivência Relativa - Capítulo 2

A.1 Estimação da Sobrevivência Observada

Tabela para o Cálculo da Sobrevivência Observada (Método Atuarial)

Tabela do Cálculo da Sobrevivência Observada (Método Atuarial)							
1- Anos após o diagnóstico (j)	2- nº indivíduos vivos no início do ano (n_j)	3- nº de mortes observadas durante o ano (d_j)	4- nº de indivíduos vistos pela última vez durante o ano (w_j)	5- nº de indivíduos efetivos expostos ao risco de morrer ($l_j = n_j - \frac{w_j}{2}$)	6- proporção de mortes durante o ano ($q_j = \frac{d_j}{l_j}$)	7- proporção de sobreviventes durante o ano ($p_j = 1 - q_j$)	8- proporção de sobreviventes desde o primeiro tratamento até ao fim do ano ($\prod p_j$)
1	$n_1 = n$	d_1	w_1	$l_1 = n_1 - \frac{w_1}{2}$	$q_1 = \frac{d_1}{l_1}$	$p_1 = 1 - q_1$	p_1
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
J	$n_{J-1} - d_{J-1} - w_{J-1}$	d_J	w_J	$l_J = n_J - \frac{w_J}{2}$	$q_J = \frac{d_J}{l_J}$	$p_J = 1 - q_J$	$\prod p_J$

A.2 Cruzamento de Variáveis

Tabela A.1: Cruzamento de variáveis Sexo e Local

Sexo	**Reto**	**Cólon**	**Totais**	
Feminino	844	1331	2175	Mortes
	163	344	507	Censura
	1007	1675	2682	Total
Masculino	1230	1574	2804	Mortes
	197	288	485	Censura
	1427	1862	3289	Total
Total de Mortes	2074	2905	4979	
Total de Censura	360	632	992	
Total	2434	3537	5971	

Tabela A.2: Cruzamento de variáveis Sexo e Estadio

Sexo	**Estadio I**	**Estadio II**	**Estadio III**	**Estadio Desc.**	**Totais**	
Feminino	281	1150	548	196	2175	Mortes
	134	354	10	9	507	Censura
	415	1504	558	205	2682	Total
Masculino	341	1488	794	181	2804	Mortes
	133	336	9	7	485	Censura
	474	1824	803	188	3289	Total
Total de Mortes	622	2638	1342	377	4979	
Total de Censura	267	690	19	16	992	
Total	889	3328	1361	393	5971	

Tabela A.3: Cruzamento de variáveis Local e Estadio

Local	**Estadio I**	**Estadio II**	**Estadio III**	**Estadio Desc.**	**Totais**	
Reto	345	1073	460	196	2074	Mortes
	137	211	5	7	360	Censura
	482	1284	465	203	2434	Total
Cólon	277	1565	882	181	2905	Mortes
	130	479	14	9	632	Censura
	407	2044	896	190	3537	Total
Total de Mortes	622	2638	1342	377	4979	
Total de Censura	267	690	19	16	992	
Total	889	3328	1361	393	5971	

Tabela A.4: Cruzamento de variáveis Idade e Estadio

Idade	**Estadio I**	**Estadio II**	**Estadio III**	**Estadio Desc.**	**Totais**	
12,5-60,2	80	512	366	33	991	Mortes
	138	348	10	7	503	Censura
	218	860	376	40	1494	Total
60,3-68,1	151	643	326	51	1171	Mortes
	80	226	8	7	321	Censura
	231	869	334	58	1492	Total
68,2-75,3	175	768	339	69	1351	Mortes
	38	101	1	1	141	Censura
	213	869	340	70	1492	Total
75,4-96,7	216	715	311	224	1466	Mortes
	11	15	0	1	27	Censura
	227	730	311	225	1493	Total
Total de Mortes	622	2638	1342	377	4979	
Total de Censura	267	690	19	16	992	
Total	889	3328	1361	393	5971	

Tabela A.5: Cruzamento de variáveis Idade e Diagnóstico

Idade	**Diagnóstico 01-01-1994 a 30-09-1995**	**Diagnóstico 01-10-1995 a 30-06-1997**	**Diagnóstico 01-07-1997 a 31-03-1999**	**Diagnóstico 01-04-1999 a 30-12-2000**	**Totais**	
12,5-60,2	270	216	244	261	991	Mortes
	100	116	142	145	503	Censura
	370	332	386	406	1494	Total
60,3-68,1	309	285	295	282	1171	Mortes
	54	63	99	105	321	Censura
	363	348	394	387	1492	Total
68,2-75,3	300	319	383	349	1351	Mortes
	16	22	42	61	141	Censura
	316	341	425	410	1492	Total
75,4-96,7	298	381	394	393	1466	Mortes
	1	4	9	13	27	Censura
	299	385	403	406	1493	Total
Total de Mortes	1177	1201	1316	1285	4979	
Total de Censura	171	205	292	324	992	
Total	1348	1406	1608	1609	5971	

Tabela A.6: Cruzamento de variáveis Estadio e Diagnóstico

Estadio	**Diagnóstico 01-01-1994 a 30-09-1995**	**Diagnóstico 01-10-1995 a 30-06-1997**	**Diagnóstico 01-07-1997 a 31-03-1999**	**Diagnóstico 01-04-1999 a 30-12-2000**	**Totais**	
I	167	160	171	124	622	Mortes
	45	62	73	87	267	Censura
	212	222	244	211	889	Total
II	601	631	716	690	2638	Mortes
	118	132	212	228	690	Censura
	719	763	928	918	3328	Total
III	296	309	339	398	1342	Mortes
	4	5	4	6	19	Censura
	300	314	343	404	1361	Total
Desconhecido	113	101	90	73	377	Mortes
	4	6	3	3	16	Censura
	117	107	93	76	393	Total
Total de Mortes	1177	1201	1316	1285	4979	
Total de Censura	171	205	292	324	992	
Total	1348	1406	1608	1609	5971	

Tabela A.7: Cruzamento de variáveis Sexo e Diagnóstico

Sexo	**Diagnóstico 01-01-1994 a 30-09-1995**	**Diagnóstico 01-10-1995 a 30-06-1997**	**Diagnóstico 01-07-1997 a 31-03-1999**	**Diagnóstico 01-04-1999 a 30-12-2000**	**Totais**	
Feminino	528	555	562	530	2175	Mortes
	88	102	152	165	507	Censura
	616	657	714	695	2682	Total
Masculino	649	646	754	755	2804	Mortes
	83	103	140	159	485	Censura
	732	749	894	914	3289	Total
Total de Mortes	1177	1201	1316	1285	4979	
Total de Censura	171	205	292	324	992	
Total	1348	1406	1608	1608	5970	

A.3 Estimadores da Sobrevivência Esperada

Os procedimentos para estimar a sobrevivência relativa utilizando os estimadores da sobrevivência esperada, o método Ederer I, Ederer II e Hakulinen no programa estatístico R versão 3.2.5 são:

Variável `sex`

- Codificação da variável:
  `sex=1` (Masculino)
  `sex=2` (Feminino)

- Código para o cálculo da curva de Sobrevivência Relativa:
  `rs.surv(Surv(time,stat)~ sex`[1]`+ratetable(age=age,sex=sex,year=diag), data=colrec, ratetable=portpop, method`[2]`="ederer1")`

Tabela A.8: Sobrevivência Relativa para Sexo Masculino e Feminino

	Sex=1			**Sex=2**		
Method	Median	0.95LCL	0.95UCL	Median	0.95LCL	0.95UCL
Ederer I	1112	1014	1236	1273	1092	1508
Ederer II	1072	976	1179	1155	1023	1329
Hakulinen	1112	1014	1236	1269	1092	1507
Pohar Perme	1061	957	1173	1149	1013	1322

[1] As variáveis que podem ser utilizadas: `Sex` ou `Site` ou `Stage`
[2] `Method`: Ederer I = `"ederer1"`, Ederer II = `"ederer2"`, Hakulinen = `"hakulinen"`

Variável `site`

- Codificação da variável:
  `site[T.rectum]=0` (Cólon)
  `site[T.rectum]=1` (Reto)

Tabela A.9: Sobrevivência Relativa para Cólon e Reto

	site[T.rectum]=0			**site[T.rectum]=1**		
Method	Median	0.95LCL	0.95UCL	Median	0.95LCL	0.95UCL
Ederer I	1155	1026	1386	1177	1056	1309
Ederer II	1083	968	1228	1124	1024	1242
Hakulinen	1155	1026	1386	1177	1056	1309
Pohar Perme	1074	944	1218	1106	1015	1233

Variável `stage`

- Codificação da variável:
  `stage[T.2]=0, stage[T.3]=0, stage[T.99]=0` (Estadio I)
  `stage[T.2]=1, stage[T.3]=0, stage[T.99]=0` (Estadio II)
  `stage[T.2]=0, stage[T.3]=1, stage[T.99]=0` (Estadio III)
  `stage[T.2]=0, stage[T.3]=0, stage[T.99]=1` (Estadio Desc.)

Tabela A.10: Sobrevivência Relativa segundo o Estadio

	Estadio I			**Estadio II**		
Method	Median	0.95LCL	0.95UCL	Median	0.95LCL	0.95UCL
Ederer I	NA	NA	NA	2778	2187	NA
Ederer II	NA	NA	NA	2227	1865	2799
Hakulinen	NA	NA	NA	2778	2187	NA
Pohar Perme	NA	5972	NA	2285	1854	3075

	Estadio III			**Estadio Desc.**		
Method	Median	0.95LCL	0.95UCL	Median	0.95LCL	0.95UCL
Ederer I	180	168	201	218	139	279
Ederer II	179	168	199	216	139	279
Hakulinen	180	168	201	218	139	279
Pohar Perme	179	167	199	216	138	279

A.4 *Net Survival* pelo método Pohar Perme

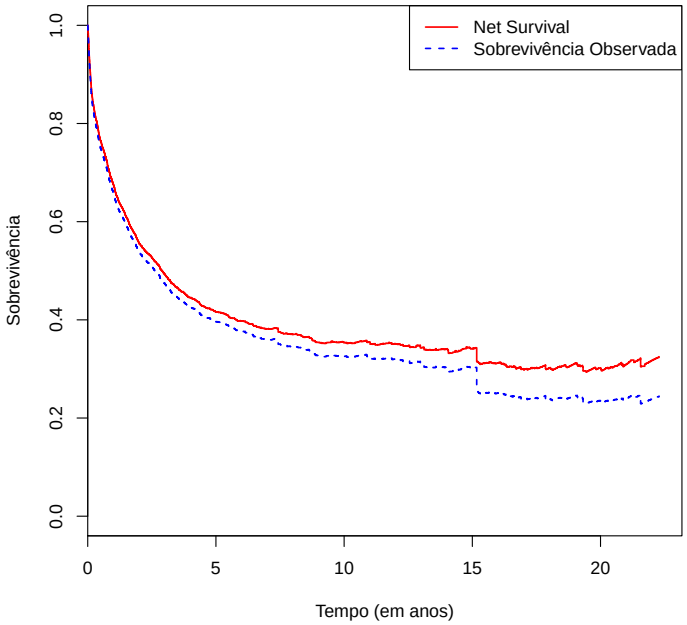

Figura A.1: *Net Survival* e Sobrevivência Observada para sexo Masculino

Anexo B

Modelos de Regressão da Sobrevivência Relativa - Capítulo 3

B.1 Modelos de Regressão Aditivos

Apresentam-se em seguida os procedimentos para calcular os modelos de regressão aditivos na Análise de Sobrevivência Relativa no programa estatístico R versão 3.2.5.

Para começar temos que recodificar as covariáveis `age` e `diag` em covariáveis categóricas.

Covariável `age`:

- Primeiro passo é o estudo dos quartis para a covariável `age` (ver as Figuras B.1; B.2):
  `Statistics-Summaries-Numerical Summaries`

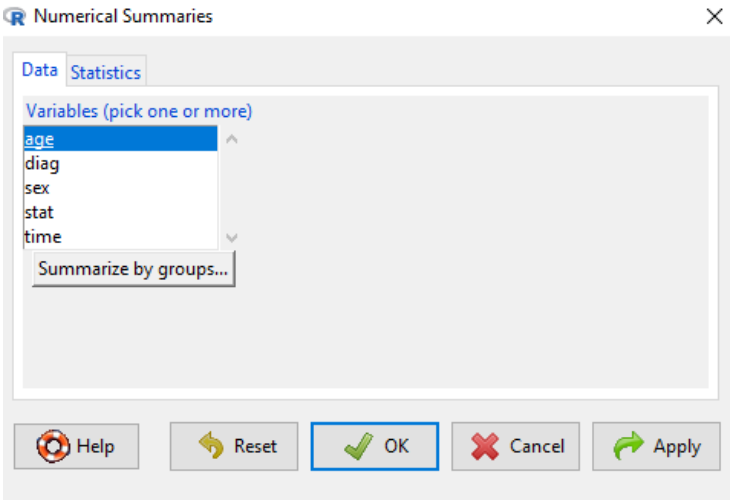

Figura B.1: Estudo dos Quartis - age

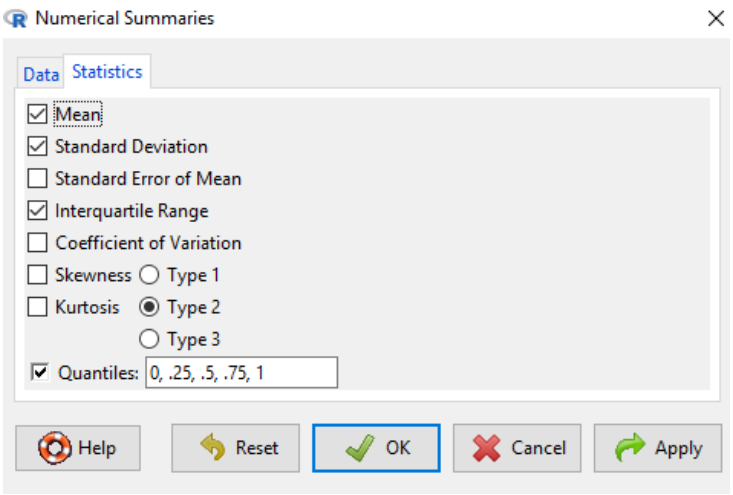

Figura B.2: Continuação do estudo dos Quartis - age

- Segundo passo é recodificar a covariável(ver as Figuras B.3; B.4):
  `Data-Manage Variables Inactive Data Set -`
  `Bin a Numeric Variable`

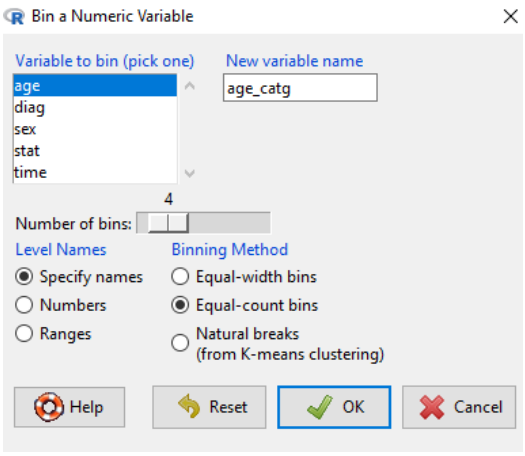

Figura B.3: Recodificação da covariável age

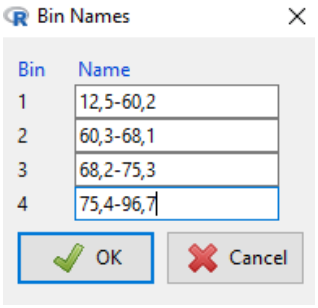

Figura B.4: Categorização da covariável age

Covariável `diag`:

As datas do diagnóstico são de 01-01-1994 a 30-12-2000.

- Primeiro passo é dividir a covariável em 4 categorias, ou seja, entre o início e o fim do estudo decorreram 7 anos e, sabendo que cada ano tem 12 meses, então fizemos $\frac{7\times12}{4} = 21$ meses.

- Segundo passo é recodificar a covariável (ver a Figura B.5):
  `Data-Manage Variables in Active Data Set-Recode Variables`

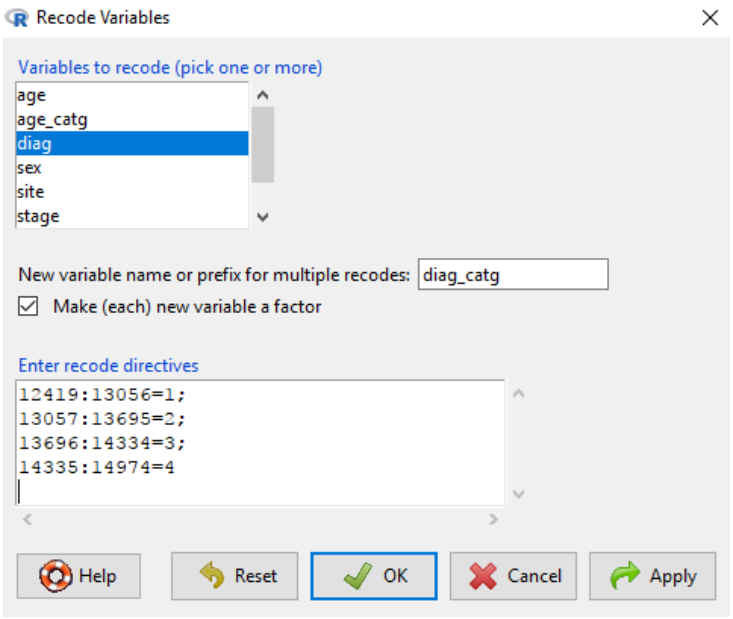

Figura B.5: Recodificação da covariável diag

Tabela B.1: Categorização da covariável diag

Categoria	Data de diagnóstico
I	01-01-1994 a 30-09-1995
II	01-10-1995 a 30-06-1997
III	01-07-1997 a 31-03-1999
IV	01-04-1999 a 30-12-2000

O próximo passo é estudar os modelos de regressão aditivos da Análise de Sobrevivência Relativa. Vamos iniciar pelo modelo de Estève, depois o modelo de Poisson e, por fim, o modelo Hakulinen-Tenkanen.

B.1.1 Modelo de Estève

- O primeiro passo é calcular o modelo para todas as covariáveis do estudo:

```
esteve<-rsadd(Surv(time,stat)∼sex+age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="max.lik")
```

- Para obter os resultados da Tabela 3.1, disponível na secção 3.5, só temos que introduzir o código seguinte:

```
summary(esteve)
```

De todas as covariáveis introduzidas do modelo apenas duas não foram significativas: Sexo e Local, porém incluiremos a covariável Local por ser uma covariável de interesse para o estudo.

- De seguida é calcular o modelo só para as covariáveis significativas do estudo (ver Tabela B.2):

```
esteve<-rsadd(Surv(time,stat)∼age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="max.lik")
summary(esteve)
```

Tabela B.2: Modelo de Estève com as covariáveis significativas

	Est. dos coeficientes	Risco	Desvio padrão	Valor p
Idade 60,3-68,1	0.269	1.309	0.051	1.75e-07
Idade 68,2-75,3	0.378	1.459	0.052	3.13e-13
Idade 75,4-96,7	0.792	2.208	0.053	$<$ 2e-16
Diagnóstico: 01-10-1995 a 30-06-1997	-0.105	0.900	0.052	0.0451
Diagnóstico: 01-07-1997 a 31-03-1999	-0.204	0.815	0.052	7.55e-05
Diagnóstico: 01-04-1999 a 30-12-2000	-0.244	0.783	0.051	2.18e-06
Estadio II	0.878	2.406	0.087	$<$ 2e-16
Estadio III	2.674	14.498	0.089	$<$ 2e-16
Estadio Desc.	1.925	6.855	0.103	$<$ 2e-16
Local Reto	0.067	1.069	0.037	0.0735
fu [0,1]	- 2.593	0.075	0.100	$<$ 2e-16
fu (1,2]	- 2.972	0.051	0.102	$<$ 2e-16
fu (2,3]	- 3.342	0.035	0.110	$<$ 2e-16
fu (3,4]	- 3.447	0.032	0.118	$<$ 2e-16
fu (4,5]	- 3.814	0.022	0.134	$<$ 2e-16
fu (5,6]	- 4.165	0.016	0.161	$<$ 2e-16
fu (6,7]	- 4.453	0.012	0.192	$<$ 2e-16
fu (7,8]	- 4.785	0.008	0.241	$<$ 2e-16
fu (8,9]	- 4.894	0.007	0.266	$<$ 2e-16
fu (9,10]	- 5.346	0.005	0.361	$<$ 2e-16

B.1.2 Modelo de Poisson

Para o estudo do modelo de Poisson, voltamos a repetir os mesmos passos que realizamos no modelo de Estève.

- O primeiro passo é calcular o modelo para todas as covariáveis do estudo (ver a Tabela B.3):

```
poisson<-rsadd(Surv(time,stat)∼sex+age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10[1],method="glm.poi")
summary(poisson)
```

- O segundo passo é calcular o modelo para as covariáveis significativas do estudo (ver a Tabela B.4):

```
poisson<-rsadd(Surv(time,stat)∼age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="glm.poi")
summary(poisson)
```

As conclusões são iguais ao modelo de Estève.

[1] período de *follow-up* considerado neste estudo

Tabela B.3: Modelo de Poisson com todas as variáveis da base de dados

	Est. dos coeficientes	Desvio padrão	Valor p
Idade 60,3-68,1	0.266	0.051	2.16e-07
Idade 68,2-75,3	0.380	0.052	2.61e-13
Idade 75,4-96,7	0.805	0.053	< 2e-16
Sexo Feminino	-0.053	0.037	0.1511
Diagnóstico: 01-10-1995 a 30-06-1997	-0.107	0.052	0.0410
Diagnóstico: 01-07-1997 a 31-03-1999	-0.208	0.052	5.55e-05
Diagnóstico: 01-04-1999 a 30-12-2000	-0.243	0.051	2.21e-06
Estadio II	0.868	0.086	< 2e-16
Estadio III	2.652	0.088	< 2e-16
Estadio Desc.	1.907	0.102	< 2e-16
Local Reto	0.064	0.0372	0.0852
fu [0,1]	-2.495	0.113	< 2e-16
fu (1,2]	-2.883	0.116	< 2e-16
fu (2,3]	-3.245	0.123	< 2e-16
fu (3,4]	-3.363	0.129	< 2e-16
fu (4,5]	-3.731	0.146	< 2e-16
fu (5,6]	-4.075	0.170	< 2e-16
fu (6,7]	-4.366	0.201	< 2e-16
fu (7,8]	-4.665	0.245	< 2e-16
fu (8,9]	-4.742	0.265	< 2e-16
fu (9,10]	-5.323	0.403	< 2e-16

Tabela B.4: Modelo de Poisson com as covariáveis significativas

	Est. dos coeficientes	Risco	Desvio padrão	Valor p
Idade 60,3-68,1	0.267	1.306	0.051	2.02e-07
Idade 68,2-75,3	0.378	1.459	0.052	2.76e-13
Idade 75,4-96,7	0.797	2.219	0.053	$<$ 2e-16
Diagnóstico: 01-10-1995 a 30-06-1997	-0.105	0.900	0.052	0.0449
Diagnóstico: 01-07-1997 a 31-03-1999	-0.203	0.816	0.052	8.02e-05
Diagnóstico: 01-04-1999 a 30-12-2000	-0.242	0.785	0.051	2.52e-06
Estadio II	0.889	2.433	0.088	$<$ 2e-16
Estadio III	2.673	14.483	0.089	$<$ 2e-16
Estadio Desc.	1.927	6.869	0.103	$<$ 2e-16
Local Reto	0.066	1.068	0.037	0.0748
fu [0,1]	-2.592	0.075	0.099	$<$ 2e-16
fu (1,2]	-2.982	0.051	0.103	$<$ 2e-16
fu (2,3]	-3.346	0.035	0.111	$<$ 2e-16
fu (3,4]	-3.457	0.032	0.117	$<$ 2e-16
fu (4,5]	-3.827	0.022	0.135	$<$ 2e-16
fu (5,6]	-4.170	0.015	0.161	$<$ 2e-16
fu (6,7]	-4.452	0.012	0.192	$<$ 2e-16
fu (7,8]	-4.779	0.008	0.241	$<$ 2e-16
fu (8,9]	-4.868	0.008	0.264	$<$ 2e-16
fu (9,10]	-5.390	0.005	0.393	$<$ 2e-16

B.1.3 Modelo de Hakulinen-Tenkanen

Como já se vem a repetir:

- O primeiro passo é calcular o modelo para todas as covariáveis do estudo (ver a Tabela B.5):

```
hakulinen<-rsadd(Surv(time,stat)~sex+age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="glm.bin")
summary(hakulinen)
```

Concluímos que de todas as covariáveis, apenas a covariável Sexo é não significativa, por isso vamos excluir do modelo.

- O próximo passo é calcular o modelo com as covariáveis significativas (ver a Tabela B.6):

```
hakulinen<-rsadd(Surv(time,stat)~age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="glm.bin")
summary(hakulinen)
```

Tabela B.5: Modelo de Hakulinen-Tenkanen com todas as variáveis da base de dados

	Est. dos coeficientes	Desvio padrão	Valor p
Idade 60,3-68,1	0.256	0.052	8.61e-07
Idade 68,2-75,3	0.350	0.053	4.30e-11
Idade 75,4-96,7	0.736	0.056	$<$ 2e-16
Sexo Feminino	-0.050	0.038	0.189982
Diagnóstico: 01-10-1995 a 30-06-1997	-0.106	0.054	0.051795
Diagnóstico: 01-07-1997 a 31-03-1999	-0.189	0.053	0.000402
Diagnóstico: 01-04-1999 a 30-12-2000	-0.255	0.053	1.71e-06
Estadio II	0.882	0.087	$<$ 2e-16
Estadio III	2.670	0.090	$<$ 2e-16
Estadio Desc.	1.839	0.104	$<$ 2e-16
Local Reto	0.121	0.039	0.001729
fu [0,1]	-2.572	0.115	$<$ 2e-16
fu (1,2]	-2.905	0.118	$<$ 2e-16
fu (2,3]	-3.246	0.125	$<$ 2e-16
fu (3,4]	-3.377	0.131	$<$ 2e-16
fu (4,5]	-3.745	0.148	$<$ 2e-16
fu (5,6]	-4.098	0.172	$<$ 2e-16
fu (6,7]	-4.391	0.203	$<$ 2e-16
fu (7,8]	-4.672	0.245	$<$ 2e-16
fu (8,9]	-4.752	0.266	$<$ 2e-16
fu (9,10]	-5.319	0.398	$<$ 2e-16

Tabela B.6: Modelo de Hakulinen-Tenkanen com as covariáveis significativas

	Est. dos coeficientes	Risco	Desvio padrão	Valor p
Idade 60,3-68,1	0.257	1.293	0.052	7.98e-07
Idade 68,2-75,3	0.349	1.418	0.053	4.53e-11
Idade 75,4-96,7	0.728	2.071	0.056	$<$ 2e-16
Diagnóstico: 01-10-1995 a 30-06-1997	-0.104	0.901	0.054	0.056150
Diagnóstico: 01-07-1997 a 31-03-1999	-0.184	0.832	0.053	0.000568
Diagnóstico: 01-04-1999 a 30-12-2000	-0.253	0.776	0.053	2.00e-06
Estadio II	0.901	2.462	0.088	$<$ 2e-16
Estadio III	2.690	14.732	0.091	$<$ 2e-16
Estadio Desc.	1.857	6.404	0.105	$<$ 2e-16
Local Reto	0.122	1.130	0.039	0.001542
fu [0,1]	-2.663	0.070	0.101	$<$ 2e-16
fu (1,2]	-2.996	0.050	0.104	$<$ 2e-16
fu (2,3]	-3.341	0.035	0.112	$<$ 2e-16
fu (3,4]	-3.465	0.031	0.119	$<$ 2e-16
fu (4,5]	-3.835	0.022	0.136	$<$ 2e-16
fu (5,6]	-4.185	0.015	0.162	$<$ 2e-16
fu (6,7]	-4.474	0.011	0.194	$<$ 2e-16
fu (7,8]	-4.779	0.008	0.241	$<$ 2e-16
fu (8,9]	-4.867	0.008	0.264	$<$ 2e-16
fu (9,10]	-5.383	0.005	0.389	$<$ 2e-16

B.2 Modelo de Regressão Multiplicativo

Os procedimentos para o modelo de regressão multiplicativo são muito semelhantes aos usados nos modelos de regressão aditivos.

B.2.1 Modelo de Andersen

Para o estudo do modelo de regressão de Andersen começamos com todas as covariáveis e verificamos quais as covariáveis que são significativas para o estudo. O código é seguinte:

```
andersen<-rsmul(Surv(time,stat)∼sex+age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10)
summary(andersen)
```

B.3 Modelo de Transformação dos tempos de vida

Para o estudo do modelo de transformação o código é da forma:

```
transf<-rstrans(Surv(time,stat)∼sex+age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10)
summary(transf)
```

B.4 Proporcionalidade das funções de risco

Para verificar a hipótese de riscos proporcionais para as covariáveis em estudo nos modelos de regressão utilizámos a função `rs.br` disponível no *package relsurv*.

B.4.1 Modelo de Estève

O código para o modelo de Estève é da forma:

```
esteve<-rsadd(Surv(time,stat)~age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="max.lik")
rs.br(esteve)
```

Tabela B.7: Teste de proporcionalidade das funções de risco para o modelo de Estève

	max	***p***
Idade 60,3-68,1	1.488	2.39e-02
Idade 68,2-75,3	0.827	5.00e-01
Idade 75,4-96,7	3.343	3.90e-10
Diagnóstico: 01-10-1995 a 30-06-1997	0.930	3.53e-01
Diagnóstico: 01-07-1997 a 31-03-1999	0.493	9.68e-01
Diagnóstico: 01-04-1999 a 30-12-2000	0.992	2.79e-01
Estadio II	1.733	4.92e-03
Estadio III	4.145	2.33e-15
Estadio Desc.	3.526	3.18e-11
Local Reto	4.090	5.88e-15
GLOBAL	9.478	0.00e+00

B.4.2 Modelo de Poisson

Para o modelo de Poisson é da seguinte forma:

```
poisson<-rsadd(Surv(time,stat)~age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="glm.poi")
rs.br(poisson)
```

Tabela B.8: Teste de proporcionalidade das funções de risco para o modelo de Poisson

	max	***p***
Idade 60,3-68,1	1.485	2.44e-02
Idade 68,2-75,3	0.831	4.95e-01
Idade 75,4-96,7	3.334	4.41e-10
Diagnóstico: 01-10-1995 a 30-06-1997	0.930	3.53e-01
Diagnóstico: 01-07-1997 a 31-03-1999	0.493	9.68e-01
Diagnóstico: 01-04-1999 a 30-12-2000	0.989	2.82e-01
Estadio II	1.751	4.34e-03
Estadio III	4.136	2.78e-15
Estadio Desc.	3.517	3.61e-11
Local Reto	4.093	5.55e-15
GLOBAL	9.339	0.00e+00

B.4.3 Modelo de Hakulinen-Tenkanen

Para o modelo de Hakulinen-Tenkanen é escrito na forma:

```
hakulinen<-rsadd(Surv(time,stat)~age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10,method="glm.bin")
rs.br(hakulinen)
```

Tabela B.9: Teste de proporcionalidade das funções de risco para o modelo de Hakulinen-Tenkanen

	max	***p***
Idade 60,3-68,1	1.502	2.19e-02
Idade 68,2-75,3	0.784	5.70e-01
Idade 75,4-96,7	3.426	1.28e-10
Diagnóstico: 01-10-1995 a 30-06-1997	0.942	3.38e-01
Diagnóstico: 01-07-1997 a 31-03-1999	0.538	9.34e-01
Diagnóstico: 01-04-1999 a 30-12-2000	1.016	2.53e-01
Estadio II	1.758	4.13e-03
Estadio III	4.166	1.67e-15
Estadio Desc.	3.702	2.49e-12
Local Reto	4.162	1.78e-15
GLOBAL	10.277	0.00e+00

B.4.4 Modelo de Andersen

Para o modelo de Andersen é definido do seguinte modo:

```
andersen<-rsmul(Surv(time,stat)~sex+age_catg+diag_catg+stage+site
+ratetable(age=age,sex=sex,year=diag),data=colrec,ratetable=portpop,
int=10)
rs.br(andersen)
```

Tabela B.10: Teste de proporcionalidade das funções de risco para o modelo de Andersen

	max	***p***
Idade 60,3-68,1	1.015	2.55e-01
Idade 68,2-75,3	1.102	1.76e-01
Idade 75,4-96,7	2.122	2.46e-04
Sexo	1.908	1.38e-03
Diagnóstico: 01-10-1995 a 30-06-1997	1.112	1.69e-01
Diagnóstico: 01-07-1997 a 31-03-1999	0.669	7.62e-01
Diagnóstico: 01-04-1999 a 30-12-2000	1.051	2.19e-01
Estadio II	1.771	3.77e-03
Estadio III	3.607	9.98e-12
Estadio Desc.	3.133	5.93e-09
Local Reto	3.678	3.55e-12
GLOBAL	4.293	2.22e-16

B.4.5 Modelo Transformado

Para o modelo transformado é escrito do seguinte modo:

```
transF<-rstrans(Surv(time,stat)∼sex+age+diag+ratetable(age=age,sex=sex,
year=diag),data=colrec,ratetable=portpop,int=10)
rs.br(transF)
```

Tabela B.11: Teste de proporcionalidade das funções de risco para o modelo Transformado

	max	***p***
Idade	3.725	1.79e-12
Diagnóstico	0.539	9.34e-01
Sexo	2.363	2.82e-05
GLOBAL	3.775	8.40e-13

Bibliografia

[1] Aalen, O. O., Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (2009) - History of Applications of Martingales in Survival Analysis. *Electronic Journal for History of Probability and Statistics*, Vol.5, Nº1.

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

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

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

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

[6] Andersen, P. K., Borch-Johnsen, K., Deckert, T., Green, A., Hougaard, P., Keiding, N. e Kreiner, S. (1985) - A Cox Regression Model for the Relative Mortality and Its Application to Diabetes Mellitus Survival Data. *Internacional Biometric Society*, Vol.41, Nº4, p.921-932.

[7] Bastos, J., Rocha, C. (2006) - Análise de Sobrevivência - Conceitos Básicos. *Arquivos de Medicina*, Vol.20, Nº5-6.

[8] Berkson, J. (1942) - The Calculation of Survival Rates. *In: Carcinoma and Other Malignant Lesions of the Stomach*. Edited by Walters, W., Gray, H. K., and Priestly. Philadelphia: Sanders.

[9] Brenner, H., Gefeller, O., Hakulinen, T. (2002) - A Computer Program for Period Analysis of Cancer Patient Survival. *European Journal of Cancer*, Vol.38, p.690-695.

[10] Brenner, H., Gefeller, O., Hakulinen, T. (2004) - Period Analysis for 'up-to-date' Cancer Survival Data: theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer*, Vol.40, p.326-335.

[11] Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S., Shimukura, S.E. (2011) - *Análise de Sobrevivência - Teoria e Aplicações em Saúde*. 2ª Edição, Rio de Janeiro: Editora Fiocruz. ISBN:978-85-7541-216-9.

[12] Cho, H., Howlader, N., Mariotto, A. B., Cronin, K. A. (2011) - Estimating Relative Survival for Cancer Patients from the SEER Program using expected rates based on Ederer I versus Ederer II method. Surveillance Research Program, NCI, Technical Report.

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

[14] Carrilho, M. J., Patrício, L. (2004) - Tábuas de Mortalidade em Portugal. *Revista de Estudos Demográficos*, Nº36. Instituto Nacional de Estatística (INE), Departamento de Estatística Sociais.

[15] Cox, D. R. (1972) - Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society*. Series B, Vol.34, Nº2, p.187-220.

[16] Dickman, P. W., Sloggett, A., Hills, M., Hakulinen, T. (2004) - Regression Models for Relative Survival. *Statistics in Medicine*, Vol.23, p.51-64.

[17] Dickman, P. W., Coviello, E. (2015) - Estimating and Modeling Relative Survival. *The Stata Journal*, Vol.15, Nº1, p.186-215.

[18] Ederer, F., Heise, H. (1959) - Instructions to IBM 650 Programmers in Processing Survival Computations. Technical, End Results Evaluation Section, National Cancer Institute.

[19] Ederer, F., Axtell, L. M., Cutler, S. J. (1961) - The Relative Survival Rate: a Statistical Methodology. *National Cancer Institute Monograph*, Vol.6, p. 101-121.

[20] Estève, J., Benhamou, E., Croasdale, M., Raymond, L. (1990) - Relative Survival and the Estimation of Net Survival: Elements for Further Discussion. *Statistics in Medicine*, Vol.9, N°5, p. 529-538.

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

[22] Giorgi, R., Hédelin, G., Schaffer, P. (2001) - Relative Survival: Comparison of Regressive Models and Advice for the User. *Journal of Epidemiology and Biostatistics*, Vol.6, N°6, p.455-462.

[23] Giorgi, R., Armanet, A., Gouvernet, J., Bonnier, P., Fieschi, M. (2005) - Revue Comparative des Modèles Régressifs de Survie Brute et de Survie Relative. *Rev Epidemiol Sante*, Vol.53, p.409-417.

[24] Hakulinen, T. (1982) - Cancer Survival Corrected for Heterogeneity in Patient Withdrawal. *Biometrics*, Vol.38, p.933-942.

[25] Hakulinen, T., Tenkanen, L. (1987) - Regression Analysis of Relative Survival Rates. *Journal of the Royal Statistical Society*. Applied Statistics. Vol.36, N°3, p.309-317.

[26] Hakulinen, T., Seppa, K., Lambert, P. C. (2011) - Choosing the Relative Survival Method for Cancer Survival Estimation. *European Journal of Cancer*, Vol.47, p.2202-2210.

[27] Hinchliffe, S. R., Dickman, P. W., Lambert, P. C. (2012) - Adjusting for the Proportion of Cancer Deaths in the General Population when Using Relative Survival: A Sensitivity Analysis. *The International Journal of Cancer Epidemiology, Detection, and Prevention*, Vol.36, p.148-152.

[28] Holleczek, B., Gondos, A., Brenner, H. (2009) - periodR - an R package to calculate long term cancer survival estimates using period analysis. *Methods of Information in Medicine*, Vol.48, N°2, p.123-128.

[29] Holleczek, B., Brenner, H. (2013) - Model based period analysis of absolute and relative survival with R: Data preparation, model fitting and derivation of survival estimates. *Computer Methods and Programs in Biomedicine*, Vol.110, p.192-202.

[30] Hougaard, P. (2001) - *Analysis of Multivariate Survival Data*. New York: Springer. ISBN:0-387-98873-4.

[31] João, R. S., Papoila, A. L., Miranda, A. (2013) - *Sobrevivência Relativa do Cancro-rectal e do Estômago no Sul de Portugal*. Estatística: A ciência da incerteza - Atas do XXI Congresso Sociedade Portuguesa de Estatística. ISBN:978-972-8890-35-3.

[32] Kaplan, E. L., Meier, P. (1958) - Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, Vol.53, Nº282, p.457-481.

[33] Kleinbaum, D. G, Klein, M. (2012) - *Survival Analysis: A Self-Learning Text*. 2ª Edição. New York: Springer. ISBN:0-387-23918-9.

[34] Le, Chap T.(1997) - *Applied Survival Analysis*. New York: John Wiley & Sons. ISBN:0-417-17085-2.

[35] Parkin, D. M, Hakulinen, T. (1991) - Analysis of Survival. In: Cancer Registration Principles and Methods (Jensen, O. M., Parkin, D. M., Maclennan, R., Muir, C. S. & Skeet, R. G.), Lyon: *International Agency for Research on Cancer (IARC), Scientific Publications*, Vol.95, p.159-176.

[36] Perme, M. P., Henderson, R., Stare, J. (2009) - An Approach to Estimation in Relative Survival Regression. *Biostatistics*, Vol.10, Nº1, p.136-146.

[37] Perme, M. P., Stare, J., Estève, J. (2012) - On Estimation in Relative Survival. *The International Biometric Society*, Vol.68, p.113-120.

[38] Perme, M. P., Pavlic, K. (2016) - Relsurv: A package for relsurv - Relative Survival. Versão 2.0-9. URL:http://CRAN.R-project.org/package=relsurv.

[39] Pohar, M., Stare, J. (2006) - Relative Survival Analysis in R. *Computer Methods and Programs in Biomedicine*, Vol.81, p.272-278.

[40] Pohar, M., Stare, J. (2007) - Making Relative Survival Analysis Relatively Easy. *Computers in Biology and Medicine*, Vol.37, p.1741-1749.

[41] Roch, G., Payan, J., Gouvernet, J. (2005) - RSURV: A Function to Perform Relative Survival Analysis with S-PLUS or R. *Computer Methods and Programs in Biomedicine*, Vol.78, p.175-178.

[42] Rocha, C. S. (1995) - *Modelos de Sobrevivência*. Departamento de Estatística e Investigação Operacional. Faculdade de Ciências da Universidade de Lisboa.

[43] Rocha, C., Papoila, A. P. (2009) - *Análise de Sobrevivência. XVII Congresso da Sociedade Portuguesa de Estatística*. SPE. ISBN: 978-972-8890-22-3.

[44] Roche, L., Danieli, C., Belot, A., Grosclaude, P., Bouvier, A., Velten, M., Iwaz, J., Remontet, L., Bossard, N. (2013) - Cancer net survival on registry data: Use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classic methods. *International Journal of Cancer*, Vol.132, p.2359-2369.

[45] Rodrigues, Mariana C. F. (2012) - *Análise de Sobrevivência Aplicada ao Estudo dos Tumores Malignos do Aparelho Digestivo na RAM*. Dissertação de Mestrado. Universidade da Madeira.

[46] Rutherford, M. J., Dickman, P. W., Lambert, P. C. (2012) - Comparison of Methods for Calculating Relative Survival in Population-based Studies. *The International Journal of Cancer Epidemiology, Detection, and Prevention*, Vol.36, p.16-21.

[47] Seppa, K., Hakulinen, T., Pokhrel, A. (2015) - Choosing the Net Survival Method for Cancer Survival Estimation. *European Journal of Cancer*, Vol.51, p.1123-1129.

[48] Stare, J., Henderson, R., Pohar, M. (2005) - An Individual Measure of Relative Survival. *Royal Statistical Society*, Vol. 54, p.115-126.