DM

A Nossa Universidade

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

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

UNIVERSIDADE da MADEIRA

Modelos de Cura:
Aplicação ao Cancro da Mama Feminino
Ana Carina Fernandes Alves

DM

Modelos de Cura:
Aplicação ao Cancro da Mama Feminino
Ana Carina Fernandes Alves

Modelos de Cura:
Aplicação ao Cancro da Mama Feminino

DISSERTAÇÃO DE MESTRADO

Ana Carina Fernandes Alves
MESTRADO EM MATEMÁTICA

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

setembro | 2012

DM

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

DIMENSÕES: 45 X 29,7 cm
PAPEL: COUCHÊ MATE 350 GRAMAS
IMPRESSÃO: 4 CORES (CMYK)
ACABAMENTO: LAMINAÇÃO MATE

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

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

UNIVERSIDADE da MADEIRA

UNIVERSIDADE da MADEIRA

Modelos de Cura:
Aplicação ao Cancro da Mama Feminino

DISSERTAÇÃO DE MESTRADO

Ana Carina Fernandes Alves
MESTRADO EM MATEMÁTICA

ORIENTAÇÃO
Ana Maria Cortesão Pais Figueira da Silva Abreu

Modelos de Cura: Aplicação ao Cancro da Mama Feminino

Ana Carina Fernandes Alves

setembro 2012

aos meus pais

Agradecimentos

Com a realização deste trabalho alcancei um objetivo do qual me orgulho. O tema, já de si interessante, tornou-se muito envolvente, por se tratar de uma realidade próxima de todos nós, e despertou-me a atenção para esta doença que, ainda que curável, continua a vitimar muitas mulheres.

Quero aqui deixar expressa a minha gratidão a todas as pessoas que me ajudaram e incentivaram ao longo deste percurso:

Agradeço à Professora Doutora Ana Maria Abreu, minha orientadora. A sua dedicação e compromisso foram fulcrais para a execução desta dissertação. Obrigada pelo tempo que dispendeu, tirando dúvidas e oferecendo ideias e sugestões. Foi uma honra trabalhar consigo.

Aos meus pais, João Manuel e Maria do Carmo, por quererem dar aos filhos aquilo que não puderam ter, mesmo com grande esforço e sacrifício. Espero que estejam comigo por muitos e bons anos.

Ao Emanuel, meu companheiro, pela compreensão, apoio e incentivo. E pelo teu coração de ouro.

Aos meus irmãos, Luís, Rita e Joana, pelo apoio e pelas nossas nostálgicas discussões e brincadeiras.

À Márcia, minha *alter idem*, pela amizade e incentivo.

À Mariana e à Susana pela ajuda e pelos momentos bem passados nos estudos e trabalhos que realizámos.

Ao Registo Oncológico da Região Sul, pela disponibilização dos dados. Agradeço, em particular, à Dra. Cláudia Fraga, responsável pelo Registo Oncológico da RAM, pelo apoio prestado.

A todos os outros familiares, amigos e colegas, do trabalho e da tuna, que me incentivaram e elevaram a moral quando foi preciso.

Resumo

A Análise de Sobrevivência tem como objetivo o estudo do tempo desde um instante inicial bem definido até ao acontecimento de determinado evento. Por exemplo, poderá ser o tempo de vida de um indivíduo desde o momento em que lhe é diagnosticada uma doença até a sua morte ou cura.

Com a evolução da medicina, começou a se verificar a existência de indivíduos para os quais nunca se observava o acontecimento de interesse e designaram-se esses indivíduos por curados, imunes, ou não suscetíveis. Assim, da Análise de Sobrevivência clássica surgem os modelos de cura.

Neste trabalho, aplicaram-se estes conceitos a uma base de dados referentes a 833 mulheres diagnosticadas com cancro da mama, entre 1998 e 2005.

Verificou-se a existência de um risco de morte maior em mulheres na faixa etária dos 50 a 59 anos. Comprovou-se que o estadiamento tem um papel preponderante em relação ao prognóstico, sendo que, quanto mais avançado o estadio pior o prognóstico.

Dos tratamentos a que os doentes foram submetidos, a realização de cirurgia é indicativa de um melhor prognóstico, assim como a realização de hormonoterapia e de radioterapia. No entanto, este último tratamento não se revelou estatisticamente significativo para o modelo de regressão de Cox. A realização de quimioterapia apenas reflete um melhor prognóstico nos primeiros dois anos, o que já não acontece a partir daí. Esta caraterística inesperada ficou-se a dever à esperança de vida que o tratamento oferece aos doentes no estadio IV e à associação entre a existência de gânglios metastizados e o agravamento do prognóstico, no caso do estadio II.

O modelo de cura foi aplicado apenas ao grupo de mulheres no estadio IV, pois só neste caso se admitiu que o tempo de *follow-up* era suficiente, obtendo-se uma taxa de cura de $7,4\%$.

Palavras-chave: Análise de Sobrevivência, distribuição de Chen, estimador de Kaplan-Meier, modelo de cura, regressão de Cox.

Abstract

Survival Analysis focus on the study of time since a well-defined initial time until the ocurrence of the event of interest. For example, it may be the lifetime measured from the date of diagnosis until the death or cure of the patient.

Nowadays, with the evolution of medicine, some individuals can be considered cured, i.e., the event of interest is never observed for them. These individuals are designated cured, immune or not susceptible. Therefore, the classical methods of Survival Analysis evolved giving rise to the cure models.

In this dissertation, these concepts are applied to a database of 833 women diagnosed with breast cancer, between 1998 and 2005.

It was found that there is a higher risk of death among women in the age group 50 to 59 years old. It was proved that the cancer staging has an important role in the prognosis: the more advanced the cancer stage, the worse the prognosis.

Of all the treatments that the patients were submited, the surgery was indicative of a better prognosis, as well as radiotherapy and hormone therapy. However, the last treatment wasn't statistically significant for the Cox regression model. The chemotherapy reflects a better prognosis only in the first two years. This unexpected behaviour was due to the life expectancy that this treatment offers to stage IV patients and to the association between the presence of positive lymph nodes and the worsening of prognosis, in the case of stage II.

The cure model was applied just to the group of women in stage IV, because this was the only case where it was admitted that the follow-up time was sufficient, obtaining a cure rate of $7,4\%$.

Key words: Chen distribution, Cox regression, cure model, Kaplan-Meier estimator, Survival Analysis.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Análise de Sobrevivência

1.1 Introdução

A Análise de Sobrevivência é um ramo da Estatística que surgiu no fim da primeira metade do século vinte e que estuda o tempo de vida de indivíduos desde um instante inicial bem definido (por exemplo, data do diagnóstico de determinada doença) até à ocorrência do designado acontecimento de interesse. Este acontecimento de interesse pode ser a morte ou, num sentido mais lato, outro acontecimento definido pelo investigador, como seja o tempo que decorre até que surja uma determinada patologia numa população, a cura da doença, ou a ocorrência de uma recaída. Por exemplo, neste trabalho, serão analisados os tempos de vida de mulheres diagnosticadas com cancro da mama, onde o instante inicial é a data do diagnóstico, e o acontecimento de interesse é a morte pela doença.

A Análise de Sobrevivência permite estudar a distribuição do tempo de vida num só grupo de indivíduos, comparar a distribuição do tempo de vida em dois ou mais grupos de indivíduos bem como modelar e determinar a relação entre a distribuição do tempo de vida e as covariáveis associadas a cada indivíduo.

No entanto, a aplicação da Análise de Sobrevivência não se limita à área da saúde. Poderá ser utilizada para estudar o tempo que um aluno leva para concluir um curso, o tempo entre a obtenção de um cartão de crédito até à sua utilização, o tempo que um ex-recluso permanece em liberdade até reincidir no crime ou o tempo que um casal leva até se divorciar.

A par da medicina, também a informática superou-se e, atualmente, é uma preciosa aliada dos investigadores para a análise dos dados.

Seja T uma variável aleatória (v.a.) contínua que representa o tempo de vida de um indivíduo. A função de sobrevivência (f.s.) da v.a. T é

representada por

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

e traduz a probabilidade do tempo de vida de um indivíduo ser maior que t ou, por outras palavras, a probabilidade de um indivíduo sobreviver para além do instante t.

É então uma função contínua e monótona não crescente, onde

$$S(0) = 1 \qquad \text{e} \qquad S(+\infty) = \lim_{t\to\infty} S(t) = 0$$

Outra função importante neste contexto é a função de risco, também conhecida por função *hazard*, que é definida por

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

e que tem as seguintes propriedades

$$h(t) \geq 0, \forall t \geq 0 \qquad \text{e} \qquad \int_0^\infty h(t)dt = \infty$$

Esta função é contínua e pode ser crescente, decrescente, *bathtub-shaped*, constante ou unimodal.

A função de risco cumulativa é definida da seguinte forma

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

As funções de sobrevivência, densidade, de risco e de risco cumulativo verificam as seguintes relações

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

$$\iff H(t) = -\log S(t)$$

$$f(t) = h(t)\exp(-\int_0^t h(u)du) \tag{1.2}$$

$$h(t) = \frac{f(t)}{S(t)} \tag{1.3}$$

$$f(t) = -S'(t) \tag{1.4}$$

Frequentemente, no período estudado, o acontecimento de interesse não é observado em alguns indivíduos. Tal poderá acontecer por vários motivos, tais como a perda do rasto do doente, não se conhecendo a situação em que este se encontra, ou porque até a data do fim do estudo não foi verificada a falha nesses indivíduos. Estas observações parciais, designadas por observações censuradas, são sempre consideradas na Análise de Sobrevivência, a qual valoriza o contributo das mesmas, o que não acontece se for utilizada, por exemplo, a análise de regressão.

Nos exemplos referidos anteriormente, os tempos de vida desses indivíduos são considerados censurados à direita pois apenas se desconhece o tempo de vida desses indivíduos para além do instante de censura. A censura à direita é motivada normalmente por três razões: o fim do estudo (sem que se observe o acontecimento de interesse), a perda do indivíduo para o *follow-up* e a exclusão do indivíduo por alguma razão, como por exemplo, ter morrido por outra causa que não a doença em estudo. Existem outros tipos de censura, como sejam, censura à esquerda (o instante em que ocorre o acontecimento de interesse é desconhecido, apenas se sabe que é inferior ao tempo de censura) e censura intervalar (apenas se conhece o intervalo aleatório de tempo em que ocorre o acontecimento de interesse e não o momento exato), mas que não serão consideradas neste trabalho.

Existe ainda a classificação da censura como informativa ou não informativa. No primeiro caso, à censura está associada algum tipo de informação, por exemplo, a ocorrência de reações adversas a um novo medicamento, enquanto que a última é assumida quando a distribuição do tempo de censura não depende do vetor de parâmetros de interesse, ou seja, a sua ocorrência não está relacionada de forma alguma com o(s) fator(es) que provocam o evento de interesse.

A partir deste ponto, entenda-se por censura a censura não informativa e a censura à direita.

1.2 Algumas distribuições para o tempo de vida

1.2.1 Distribuição de Weibull

A distribuição de Weibull é a mais utilizada como modelo de distribuição de tempos de vida, abrangendo os casos onde existem eventos recorrentes e/ou de longa duração. A sua aplicação está presente em várias áreas como, por exemplo, em estudos de biologia ou em estudos médicos. O facto de apresentar expressões simples para as funções de risco, de sobrevivência e de den-

sidade e de contemplar situações com função de risco crescente, decrescente ou constante, são dois motivos preponderantes para a sua vasta aplicação. No entanto, tem como desvantagem o facto de não ser flexível com funções de risco não monótonas, tais como, a unimodal ou a *bathtub-shaped*.

A função de risco da distribuição de Weibull é

$$h(t) = \lambda\beta(\lambda t)^{\beta-1}$$

onde $\lambda > 0$ e $\beta > 0$ são os parâmetros. O seu comportamento varia consoante o valor de β, sendo monótona crescente quando $\beta > 1$, decrescente quando $\beta < 1$ e constante para $\beta = 1$.

Tendo em contas as igualdades (1.1) e (1.2), obtém-se a função de sobrevivência desta distribuição

$$S(t) = \exp[-(\lambda t)^{\beta}], \qquad t > 0$$

e a função densidade

$$f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}], \qquad t > 0$$

1.2.2 Distribuição log-logística

A distribuição log-logística é uma alternativa à distribuição de Weibull, quando se considera um modelo com função de risco unimodal. Assim, a função de risco é dada por

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}, \qquad t > 0$$

onde α, $\lambda > 0$ são os parâmetros de forma e de escala, respetivamente.

A função de sobrevivência é

$$S(t) = \frac{1}{1+\lambda t^{\alpha}}, \qquad t > 0$$

sendo a função de densidade a seguinte

$$f(t) = \frac{\alpha\lambda t^{\alpha-1}}{(1+\lambda t^{\alpha})^2}, \qquad t > 0$$

1.2.3 Distribuição de Chen

Chen (2000), [11], desenvolveu uma distribuição biparamétrica para o tempo de vida que tem como vantagem a flexibilidade da correspondente função de risco. Assim, esta poderá ser *bathtub-shaped* ou monótona crescente, dependendo do valor de um dos seus parâmetros.

A função de distribuição proposta por Chen é

$$F(t) = 1 - [\exp(\lambda(1 - \exp(t^{\beta}))], \qquad t > 0$$

onde λ e β são os parâmetros da distribuição, sendo λ o parâmetro de escala e β o parâmetro de forma.

A função de sobrevivência é

$$S(t) = \exp[\lambda(1 - \exp(t^{\beta}))], \qquad t > 0$$

e a função de risco correspondente

$$h(t) = \lambda\beta t^{\beta-1} \exp(t^{\beta}), \qquad t > 0$$

Como $h'(t) = \lambda\beta t^{\beta-2} \exp(t^{\beta})((\beta - 1) + \beta t^{\beta})$, então $h(t)$ pode ser *bathtub-shaped* para $\beta < 1$ e, quando $\beta \geq 1$, a função de risco é crescente.

1.3 Modelos de regressão

Os modelos de regressão são amplamente utilizados em Análise de Sobrevivência devido à sua capacidade de relacionar os tempos de sobrevivência com as covariáveis, permitindo avaliar a influência destas no prognóstico do indivíduo.

Estes modelos podem ser paramétricos ou não paramétricos consoante os pressupostos admitidos. Em termos genéricos, os modelos de regressão paramétricos são aqueles em que se admite que a variável correspondente ao tempo de vida, T, segue uma determinada distribuição paramétrica (exponencial, de Weibull,...), donde, quando esta premissa não for válida teremos os modelos não paramétricos. No entanto, se não for admitida uma distribuição paramétrica para T, mas forem feitos alguns pressupostos relativamente à distribuição de T, é habitual designar por modelo de regressão semiparamétrico, de que é exemplo o conhecido modelo de Cox (1972).

Os modelos mais utilizados são os modelos de

- riscos proporcionais - as funções de risco de dois indivíduos são proporcionais e as covariáveis têm um efeito multiplicativo na função de risco;

- tempo de vida acelerado - as covariáveis têm um efeito multiplicativo no tempo de vida, ou seja, o seu efeito é travar ou acelerar o tempo até a ocorrência do acontecimento de interesse;

- possibilidades proporcionais - as covariáveis atuam de forma multiplicativa na possibilidade de um indivíduo sobreviver para além de um determinado instante t, e as funções de risco correspondentes a indivíduos com diferentes valores das covariáveis convergem ao fim de um certo tempo.

1.3.1 Modelos de regressão paramétricos

Existem vários modelos de regressão paramétricos. Uma das caraterísticas que os distingue é a distribuição usada para o tempo de vida. As mais comuns são a distribuição de Weibull e a distribuição log-logística, as quais são apresentadas em seguida. Uma sugestão que não será aqui abordada é a utilização da distribuição de Chen.

Modelo de regressão Weibull

O modelo de Weibull, como modelo de riscos proporcionais, tem a seguinte função de risco, para um indivíduo com vetor de covariáveis $\mathbf{z}$

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

donde se conclui que o tempo de vida desse indivíduo tem distribuição de Weibull com parâmetro de escala $\lambda\exp(\beta'\mathbf{z})$ e parâmetro de forma γ. Verifica-se também que o efeito das covariáveis somente modifica o parâmetro de escala, ficando o parâmetro de forma inalterado.

A função de sobrevivência é

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

A utilização da distribuição de Weibull é adequada quer quando se tem um modelo de regressão de riscos proporcionais, quer quando se trata de um modelo de tempo de vida acelerado.

Modelo de regressão log-logístico

Nas situações onde o modelo de Weibull não é adequado (quando temos funções de risco não monótonas ou um modelo de possibilidades proporcionais), poderá ser utilizado o modelo log-logístico. Considerando este modelo como modelo de possibilidades proporcionais temos, para um indivíduo com vetor de covariáveis $\mathbf{z}$, a função de sobrevivência

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

o que significa que o tempo de vida de um indivíduo com vetor de covariáveis $\mathbf{z}$ segue uma distribuição log-logística com parâmetro de escala $\lambda\exp(\beta'\mathbf{z})$ e parâmetro de forma k.

O uso da distribuição log-logística é apropriado para modelos de tempo de vida acelerado e para modelos de possibilidades proporcionais.

1.3.2 Modelos de regressão não-paramétricos

Modelo de regressão de Cox

Em 1972, Cox, [15], apresentou um modelo de regressão que é amplamente utilizado na análise de tempos de vida e que permite identificar diferenças na sobrevivência causadas pelo tratamento e pelas variáveis prognóstico em estudos clínicos (Marubini e Valsecchi, 1996). Este é o modelo de regressão mais utilizado na análise de tempos de vida, devido à sua flexibilidade e versatilidade, que permite o seu uso num grande número de situações reais. As suas mais de 26000 citações no *ISI Web of Knowledge* (abril 2012) são a prova disso.

Seja T uma v.a. contínua que representa o tempo de vida. Cox (1972) propôs um modelo em que, no instante t e para um indivíduo a que esteja associado o vetor de covariáveis $\mathbf{z} = (z_1, \ldots, z_p)\prime$, a função de risco é da forma

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

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

Trata-se de um modelo de riscos proporcionais, visto que as funções correspondentes a dois indivíduos com covariáveis z_1 e z_2 são proporcionais. De facto,

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

não depende de t.

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

O modelo de Cox também pode ser escrito através da função de sobrevivência, da seguinte forma

$$S(t;\mathbf{z}) = S_0(t)^{\exp(\beta'\mathbf{z})} = S_0(t)^{\exp(\beta_1 z_1 + \ldots \beta_p z_p)} \tag{1.6}$$

1.4 Modelos de Mistura

Lawless (2002), [34] faz uma introdução aos modelos de mistura discretos. Só se considera os modelos de mistura quando se suspeita que a população é heterogénea, ou seja, que é formada por vários grupos com diferentes caraterísticas. Os modelos de mistura discretos permitem ter em conta esta heterogeneidade, considerando uma função de sobrevivência para cada grupo. Assim, cada indivíduo da população pertence a um dos k diferentes grupos, com uma probabilidade p_i, $i = 1, ..., k$, onde $0 < p_i < 1$ e $\sum p_i = 1$.

Assume-se que os indivíduos do tipo i têm uma distribuição de tempo de vida com função de sobrevivência $S_i(t)$. Um indivíduo selecionado aleatoriamente a partir desta população tem então função de sobrevivência

$$S(t) = p_1 S_1(t) + ... + p_k S_k(t)$$

Os modelos deste tipo são designados por modelos de mistura discretos e são úteis nas situações onde a população é heterogénea mas não é possível fazer a distinção dos diferentes tipos de indivíduos. Frequentemente, presume-se que as funções de sobrevivência $S_i(t)$, $i = 1, ..., k$, sejam da mesma família paramétrica, embora tal não seja necessário. As propriedades de um modelo de mistura derivam das propriedades das k distribuições, ou componentes, envolvidas na mistura. A estimação dos parâmetros envolvidos pode ser difícil, motivo pelo qual raramente são utilizados modelos com k maior que 3.

Na lista de modelos apresentados por Lawless, surge em primeiro lugar um importante modelo particular com $k = 2$, e com uma componente degenerada. Assim a função de sobrevivência para este modelo é

$$S(t) = pS_1(t) + 1 - p \qquad t \geq 0$$

onde $0 < p < 1$, e $S_1(t)$ é uma função de sobrevivência tal que $S_1(\infty) = 0$. O autor acrescenta que o modelo é utilizado quando uma fração $1 - p$ de indivíduos na população tem tempos de vida longos, por conveniência assumidos como infinitos. Conclui comentando que, em estudos médicos que envolvem o tratamento de doenças, a função de sobrevivência anteriormente apresentada é por vezes designada por modelo de cura.

1.5 Estimação

Os métodos de inferência estatística utilizados em Análise de Sobrevivência são, de um modo geral, baseados na teoria assintótica da máxima verosimilhança, apresentando resultados corretos sob condições de regularidade bastante gerais nos processos de morte e de censura.

1.5.1 Estimação paramétrica

Função de verosimilhança

Suponha-se que a distribuição do tempo de vida T segue um determinado modelo paramétrico, indexado por um vetor de parâmetros θ, sobre o qual pretendemos realizar inferência. Admita-se que as observações são não censuradas ou censuradas à direita. Seja T a v.a. que representa o tempo de vida e C a v.a. que representa o tempo de censura. Então o tempo observado t para um indivíduo é uma observação da v.a. $Y = \min\{T, C\}$.

Considere-se que se encontram em estudo n indivíduos. Seja t_i, $i = 1, ..., n$ o tempo de vida,

$$\delta_i = \begin{cases} 1, \text{ se } t_i \text{ observado} \\ 0, \text{ se } t_i \text{ censurado} \end{cases}$$

e $\mathbf{z}_i$ um vetor de covariáveis fixas. Suponha-se que, dado $\mathbf{z}$, a distribuição do tempo de vida, T, é conhecido e é indexado por um vetor de parâmetros θ, sobre o qual se pretende realizar inferência. Então a função de sobrevivência para o i-ésimo indivíduo é $S(t_i; \mathbf{z}_i, \theta)$ e a correspondente função densidade $f(t_i; \mathbf{z}_i, \theta)$. Assim a função de verosimilhança é dada por

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

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

1.5.2 Estimação não-paramétrica

Estimador de Kaplan-Meier

Em 1958, Kaplan e Meier propuseram um estimador não paramétrico da f.s., quando existem observações censuradas. Este estimador é designado por estimador de Kaplan-Meier ou estimador produto-limite.

Sejam $t_{(1)}, ..., t_{(r)}$ os instantes de morte distintos numa amostra de dimensão n ($r \leq n$), d_i o número de mortes ocorridas em $t_{(i)}$ e n_i o número de indivíduos em risco em $t_{(i)}$. O estimador de Kaplan-Meier da f.s. é dado por

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

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

$\hat{S}(t)$ é uma função em escada, com saltos nos instantes de morte observados. Sob certas condições de regularidade, $\hat{S}(t)$ pode ser considerado como um estimador de máxima verosimilhança não paramétrico de $S(t)$. Tem-se ainda que $\hat{S}(t) = 0$ para $t \geq t_{(r)}$ se $t_{(r)}$ for a maior observação registada, isto é, se a maior observação for não censurada. Se a maior observação registada t^* for censurada, então $\hat{S}(t)$ nunca toma o valor zero e considera-se que a estimativa está definida apenas até esse instante, sendo $\hat{S}(t) = \hat{S}(t_{(r)})$ para $t_{(r)} \leq t \leq t^*$.

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

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

Capítulo 2

Estado da Arte

As primeiras abordagens aos modelos de cura foram feitas por Boag (1949), [8], no Reino Unido e por Berkson e Gage (1952), [7], nos Estados Unidos da América. Estes autores apresentaram formulações específicas de modelos para dados que pudessem conter uma componente de cura, focando o problema da estimação da taxa de cura de indivíduos com cancro. No entanto, admitem a existência de apenas uma causa para a ocorrência do evento.

Em seguida surge Haybittle (1959), [24], que ajustou um modelo de Gompertz biparamétrico a três bases de dados que continham tempos de sobrevivência de indivíduos diagnosticados com cancro localizado do *cervix* ou da mama.

Em 1977, Farewell, [18], aplicou um modelo de mistura de Weibull que permitia a presença de indivíduos imunes, num estudo prospetivo sobre o cancro da mama.

Langlands *et al.* (1979), [32], efetuaram um estudo retrospetivo sobre a sobrevivência de mulheres diagnosticadas com cancro da mama, analisando-as separadamente consoante o seu estadio e comparando a sua sobrevivência com a sobrevivência esperada das mulheres da população em geral.

Em 1982, Farewell, [19], utiliza os dados de um estudo sobre toxicologia para fazer uma abordagem através dos modelos de mistura, pressupondo a presença de indivíduos com tempos de vida longos. Originalmente, os autores tinham utilizado o modelo de Cox.

Em 1983, Haybittle, [25], analisa o conceito de cura de acordo com três diferentes interpretações: *cura clínica*, *cura pessoal* e *cura estatística*. Cada uma é vista em termos do seu método de demonstração e em termos de evidência dos seus resultados. Neste artigo foram usados dados sobre o cancro da mama.

Segundo Goldman (1984), [21], a análise paramétrica da sobrevivência em estudos clínicos envolve frequentemente o pressuposto de que a função de

risco é constante ao longo do tempo. Quando a curva da função de sobrevivência empírica estabiliza num valor superior a zero, é possível obter um modelo em que o valor da função de risco diminui ao longo do tempo utilizando as distribuições de Gompertz ou de Weibull. Assim, em vez de todos os indivíduos terem o mesmo risco, decrescente com o tempo, um modelo biologicamente mais apropriado pressupõe que uma proporção desconhecida $(1-\pi)$ tem risco elevado constante, enquanto a restante proporção (π) basicamente tem risco nulo. Segundo refere a autora, este tipo de modelo é adequado para tratamentos oncológicos em que uma parte dos doentes são curados num curto espaço de tempo.

A partir de 1990 observou-se uma proliferação do número de artigos sobre os modelos de cura, e, com estes, o aprofundamento e abrangência das suas aplicações. De entre eles destaca-se Gordon (1990), [22], Sposto *et al.* (1992), [44], Maller e Zhou (1992), [37], Kuk e Chen (1992) [31], Laska e Meisner (1992) [33], Yakovlev e Tsodikov (1996) [48] e Broet *et al.* (2001) [9], cujos resumos podem ser encontrados em Abreu (2004).

Maller e Zhou, [38], publicaram um livro em 1996 onde referem a utilização do modelo de cura para modelar dados relativos a instantes de falha em diferentes tipos de cancro, como sejam o cancro da mama, o linfoma não-Hodgkins, a leucemia, o cancro da próstata, os melanomas e o cancro da cabeça e do pescoço, uma vez que nestas patologias uma proporção considerável de indivíduos são considerados curados. Os autores tiveram por objetivo sugerir e exemplificar uma metodologia sistemática para a análise de dados de sobrevivência que contenham indivíduos imunes ou curados, designados genericamente por *sobreviventes a longo prazo.* Neste trabalho são propostos e desenvolvidos quer os métodos não paramétricos quer os paramétricos.

Tsodikov (2001), [46], continuando a trabalhar com modelos de cura de não mistura, propôs um algoritmo para ajustar um modelo de riscos proporcionais condicional a um valor fixo da taxa de cura. Para estimar a proporção de sobreviventes de longa duração é utilizado um modelo de cura paramétrico. Para aliar a estabilidade do método paramétrico com a flexibilidade do não paramétrico, a função de sobrevivência é estimada não parametricamente condicionada pelas taxas de cura fornecidas pela análise paramétrica. O algoritmo foi aplicado a dados relativos à doença de Hodgkin's.

Em Portugal, Abreu (2004), [1], apresenta a sua tese de doutoramento sobre modelos de sobrevivência para populações com indivíduos imunes, utilizando pela primeira vez a distribuição de Chen neste contexto. A autora elaborou com pormenor uma revisão histórica sobre os modelos de cura até essa data.

Uma questão importante na abordagem aos modelos de cura diz respeito

ao *follow-up* mínimo necessário para se chegar a conclusões acerca da cura, por exemplo, de determinado cancro. Yu *et al.* (2004), [49], concluem que o tempo mínimo de *follow-up* num estudo sobre cura tem de ser maior do que a mediana dos tempos observados.

Frequentemente são utilizadas as taxas de sobrevivência a 5 anos como sinónimo de cura. Tai *et al.* (2005), [45], rebateram esta prática, afirmando não ser adequada à representação de cura em termos estatísticos. No seu trabalho, analisaram os registos de uma base de dados (SEER[1]) que contém vasta informação sobre muitos tipos de cancro. Começaram por considerar 49 cancros diferentes (de acordo com a localização ou com o orgão), mas apenas 42 foram estudados por serem aqueles cujo tempo de vida dos doentes seguia uma distribuição lognormal. Apesar de terem 27 anos de *follow-up* disponível, tal não foi considerado suficiente para determinar o tempo mínimo de *follow-up* necessário para assegurar a cura estatística no caso do cancro da mama e da tiróide.

Li *et al.* (2005), [35], propõem um modelo de cura generalista que permite a inclusão de censura dependente, o que preencheu uma falha, já que até então os estudos realizados com modelos de sobrevivência de mistura com existência de fração de cura pressuponham a independência entre o tempo de sobrevida e o tempo de censura.

Abreu e Rocha (2006), [3], apresentaram um novo modelo de cura paramétrico baseado numa distribuição proposta por Chen (2000). Este modelo revelou-se vantajoso em relação a outros devido à flexibilidade da função de risco, facilidade de tratamento matemático e qualidade de ajustamento a dados reais.

Durante a construção dos modelos de regressão, os indivíduos com pelo menos uma covariável omissa são geralmente excluídos. Neste âmbito, Paes (2007), [41], salienta a importância de desenvolver métodos que incorporem na análise informações sobre os casos incompletos. Para tal, propõe uma adaptação dos modelos de sobrevivência com omissão nas covariáveis aos casos onde existem indivíduos para os quais o acontecimento de interesse nunca será observado. Sugere a existência de relação entre a omissão de covariáveis e um pior prognóstico do indivíduo.

Klebanov e Yakovlev (2007), [30], também abordam a questão do *follow-up* mínimo necessário para a estimação da taxa de cura. Os autores propõem um teste estatístico interligado com a determinação da probabilidade de cura através do estimador de Kaplan-Meier no instante da última observação, para ajudar os investigadores a decidir se o período de observação foi suficientemente longo de modo a detetar a presença de indivíduos imunes na população

[1]Surveillance, Epidemiology and End Results

em estudo.

Segundo Cooner *et al.* (2007), [14], os desenvolvimentos observados nos modelos de cura hierárquicos bayesianos (*Bayesian hierarchical cure models*) levaram ao enfoque na relação entre os modelos de mistura clássicos de Berkson e Gage (1952), [7], e os modelos estocásticos para tumores introduzidos por Yakovlev *et al.* (1993), [47], e na preferência entre um e outro. Assim, os autores propõem uma classe unificada de modelos de cura que facilitam a construção hierárquica flexivel de modelos (*flexible hierarchical model-building*) incluindo ambas as classes de modelos de cura como casos particulares. Ilustraram a aplicação do modelo através de uma base de dados sobre melanoma e outra sobre o cancro da mama.

Em 2008, Yu e Peng, [50], generalizam o modelo marginal de cura de mistura a dados de sobrevivência multivariada através da modelação das distribuições marginais, pois até então só existiam para dados bivariados. O novo modelo é aplicável a dados de sobrevivência com *clusters* de dimensão variada e censura intervalar, e permite a inclusão de covariáveis quer na taxa de cura quer na distribuição do tempo de vida dos indivíduos suscetíveis.

Em 2009, Zhang e Peng, [51], propuseram um modelo de cura de mistura de risco acelerado, onde o efeito das covariáveis na distribuição dos instantes de falha dos indivíduos não curados é nulo no instante zero e aumenta com o passar do tempo. Desenvolveram um método de estimação semiparamétrico baseado em *ranks* para obter as estimativas de máxima verosimilhança dos parâmetros do modelo e aplicaram-no a uma base de dados sobre o cancro da mama, onde puderam verificar a utilidade desse novo modelo.

Kim *et al.* (2009), [29], tendo em vista o problema da classificação de risco na prática clínica, propõem um novo modelo de mistura para dados de sobrevivência com fração de cura através do que designaram por *latent cure rate markers*. Nesse modelo, os *latent cure rate markers* são modelados através da regressão logística multinomial, e os indivíduos que têm a mesma taxa de cura são classificados como estando no mesmo grupo de risco. Os autores referem que este novo modelo é vantajoso em relação a outros existentes para dados sobre o tempo de vida de indivíduos com cancro da próstata pois pode ser utilizado para determinar não só o número de grupos de risco como também para desenvolver um algoritmo de classificação preditivo.

Rodrigues *et al.* (2009), [43], propõem a unificação dos modelos de sobrevivência de longa duração de Berkson e Gage (1952) e de Chen *et al.* (1999), [10].

Ainda em 2009, Zhao *et al.,* [52], sugerem a aplicação de um modelo de cura que incorpore a abordagem dos modelos com ponto de mudança (*change-point models)* a dados de sobrevivência.

Em 2010, Huang *et al.,* [26], analisam um conjunto de dados relativos ao

cancro da mama, examinando a proporção de indivíduos curados sujeitos a cirurgia e relacionando-os com o(s) tratamento(s) a que foram submetidos, nomeadamente a quimioterapia e a hormonoterapia.

Gouveia (2010), [23], introduziu o conceito das variáveis aleatórias ponderadas no âmbito dos modelos de longa duração. O autor foi motivado pela probabilidade de se estar a cometer um erro de planeamento experimental ao considerar que a amostra obtida tem as mesmas caraterísticas da população, quando nestas existem indivíduos imunes. Isto porque os indivíduos com tempos de vida mais longos terão mais probabilidade de serem incluídos na amostra e isso sugere que a amostra obtida não é uma amostra aleatória simples, mas sim uma amostra viciada.

Aljawadi *et al.* (2011a), [4], apresentam um estudo com a estimação da taxa de cura baseada em dados censurados à direita, sendo desconhecida a forma da função de sobrevivência. Consideram o modelo de cura de não mistura e utilizam o método da estimação não-paramétrica da máxima verosimilhança via algoritmo EM para estimar a taxa de cura. Paralelamente, Aljawadi *et al.* (2011b), [5], apresentam uma abordagem paramétrica da estimação da taxa de cura, supondo a censura à esquerda e utilizando um modelo de cura de não-mistura.

Capítulo 3

Modelos de Cura

3.1 Introdução

Ao considerarmos um grupo de indivíduos com determinada patologia, é correto assumir que, após o respetivo tratamento, alguns desses indivíduos estejam curados. Por outras palavras, podemos admitir que os indivíduos são imunes ao acontecimento de interesse, o qual poderá ser, por exemplo, a morte ou a recaída da doença. Como os modelos clássicos para dados de sobrevivência não distinguiam esses casos, alguns investigadores, tais como Boag (1949), Berkson e Gage (1952), Chen *et al.* (1999), entre outros, dedicaram-se ao desenvolvimento de modelos que explicassem melhor o mecanismo biológico envolvido, ou seja, aos posteriormente designados por modelos de cura (ou também por modelos de longa duração).

Ao contrário da Análise de Sobrevivência clássica, onde se assume que até para os indivíduos com tempo de vida censurado será observado o acontecimento de interesse (só que para além do período em estudo), nos modelos de cura assume-se que uma parte dos indivíduos com tempo de vida censurado corresponderá aos indivíduos imunes, ou seja, para esses indivíduos o acontecimento de interesse nunca será observado. Boag (1949), Berkson e Gage (1952) e Farewell (1982), entre outros, modelaram exaustivamente populações onde havia suspeita da existência de uma subpopulação imune. Esta suspeita surge quando, por exemplo, se verifica o nivelamento da curva da estimativa de Kaplan-Meier da função de sobrevivência, num valor superior a zero.

Os modelos de cura de mistura preconizam que uma fração dos indivíduos está curada da doença e que ao tempo de falha dos indivíduos suscetíveis corresponde uma função de sobrevivência própria. No entanto, a f.s. geral é imprópria, sendo então limitada a função cumulativa. A propriedade "função

de sobrevivência imprópria"está relacionada com os modelos de cura de mistura enquanto que a propriedade "função de risco cumulativa limitada"está relacionada com os modelos de cura de não-mistura. Atualmente, qualquer um destes modelos permite englobar covariáveis.

O conceito de cura é, no entanto, subjectivo. Em 1983, Haybittle debruça-se sobre o significado de cura em indivíduos com cancro. A "cura clínica" fundamenta-se no seguinte raciocínio: uma pessoa cujo cancro tenha sido completamente eliminado, não deverá esperar um maior risco de morte devido a esse cancro do que as outras pessoas com a mesma idade e sexo, na população em geral. Quando um indivíduo não apresenta mais sintomas do cancro até o fim da sua vida - o que não quer dizer que a doença tenha sido erradicada ou que haja cura clínica - para ele é um resultado satisfatório pelo que se designa a cura como "cura pessoal". O conceito de cura de cancro mais geralmente aceite é aplicado a um grupo de doentes em vez de um indivíduo: se um grupo de doentes apresenta uma taxa de morte semelhante à de uma população normal, com a mesma idade e distribuição de sexo, então este grupo pode ser considerado curado. Trata-se da "cura estatística", que não é o mesmo que cura clínica ou pessoal.

3.2 Modelos de Cura

3.2.1 Modelos não paramétricos

Um estimador não paramétrico da taxa de cura, p, é o valor mínimo da curva de Kaplan-Meier: $\hat{p} = \min\{\hat{S}(t)\}$. Este estimador de p é conhecido por estimador de Kaplan-Meier. Contudo, Maller e Zhou (1996) afirmam que o valor da taxa de cura que este estimador apresenta não é necessariamente preferível ao estimador de máxima verosimilhança obtido a partir de um modelo paramétrico, mas que pode ser um valor base para técnicas mais elaboradas.

3.2.2 Modelos Paramétricos

Farewell (1982) recomenda que se utilize os modelos de mistura apenas quando exista prova científica da existência de uma proporção de indivíduos imunes pois quando é assumida a existência das duas subpopulações - indivíduos imunes e indivíduos suscetíveis - as inferências serão baseadas nessa premissa, quer seja verdadeira ou não. O mesmo autor alerta também para uma cuidadosa análise da função de verosimilhança, uma vez que em muitas aplicações torna-se menos convexa à medida que a taxa de cura p se altera

pelo que o máximo pode ser local e não global.

Assim, poderá ser difícil identificar corretamente se estamos perante uma proporção de indivíduos curados ou perante uma cauda longa da distribuição dos tempos de vida da subpopulação de indivíduos susceptíveis.

Modelos de mistura

Os modelos de cura mais comuns são os de mistura, onde se considera que a população é composta por indivíduos curados e indivíduos não curados.

Abreu (2004) escreve o modelo de cura de mistura da seguinte forma

$$S(t) = p + (1-p)S_d(t) \tag{3.1}$$

sendo $S(t)$ a função de sobrevivência populacional da v.a. T, que representa o tempo de vida de um indivíduo numa população, na qual se admite existirem indivíduos imunes e indivíduos suscetíveis, $S_d(t)$ a função de sobrevivência do tempo de vida dos indivíduos suscetíveis e p a proporção de indivíduos imunes (ou curados) na população. O parâmetro p também se designa por taxa de cura e define-se por $p = \lim_{t\to\infty} S(t)$.

Outra formulação possível para a função de sobrevivência, nos modelos de cura, é através da definição de uma variável aleatória binária Y, onde $Y = 1$ para os indivíduos suscetíveis e $Y = 0$ para os indivíduos imunes. A função de sobrevivência da v.a. T é

$$S(t) = p + (1-p)S(t|Y=1)$$

ou, dado que a função de sobrevivência condicional a $Y = 0$ é sempre 1, para qualquer instante t, $S(t|Y=0) = 1, \forall t$, esta função pode ser escrita à custa das funções de sobrevivência das respetivas subpopulações

$$S(t) = pS(t|Y=0) + (1-p)S(t|Y=1)$$

Tendo em conta (1.3) e (1.4), tem-se que $h(t) = \frac{-(0+(1-p)S'(t|Y=1))}{S(t)}$, ou seja, pode-se definir o modelo de cura através da função de risco da seguinte forma

$$h(t) = \frac{(1-p)f(t|Y=1)}{S(t)}$$

Esta função de risco pode igualmente ser escrita utilizando as respetivas funções de risco das subpopulações

$$h(t) = w_c(t)h(t|Y=0) + w_d(t)h(t|Y=1)$$

onde

$$w_d(t) = 1 - w_c(t) = \frac{(1-p)S(t|Y=1)}{pS(t|Y=0) + (1-p)S(t|Y=1)} = \frac{(1-p)S(t|Y=1)}{S(t)}$$

Considerando que $h(t) = -\frac{d \log S(t)}{dt}$, temos $h(t|Y=0) = 0, \forall t$, pois trata-se da derivada de uma constante. Temos então

$$h(t) = \frac{(1-p)S(t|Y=1)h(t|Y=1)}{S(t)}$$

Estimação dos parâmetros

Tal como em Abreu (2004), [1], considere-se o modelo de mistura

$$S(t) = p + (1-p)S_d(t) \tag{3.2}$$

sendo $S_d(t) = S(t|Y=1)$. De modo a realizar uma abordagem paramétrica, admita-se que o tempo de vida de um indivíduo doente tem uma distribuição contínua. Steele (2003) estimou os parâmetros do modelo, embora sendo para tempos de vida discretos, e Abreu (2004) fez o mesmo mas para tempos de vida contínuos. É esta última situação que será considerada neste trabalho, pelo que se passa a descrever o processo em termos gerais.

Seja uma amostra de dimensão n e $t_1, ..., t_n$ os tempos de vida. Suponha-se que os primeiros m $(m < n)$ tempos de vida são observados. Sejam $\delta_1, ..., \delta_n$ tais que

$$\delta_i = \begin{cases} 1 \text{ se } 1 \leq i \leq m \\ 0 \text{ se } m+1 \leq i \leq n \end{cases}$$

e $y_1, ..., y_n$ tais que

$$y_i = \begin{cases} 0 \text{ se o indivíduo é imune} \\ 1 \text{ se o indivíduo é suscetível} \end{cases}$$

As situações possíveis para o par (δ_i, y_i) e os casos a que correspondem estão descritos na Tabela 3.1. Note-se que o par (1,0) não é possível pois isso corresponderia a um tempo de vida observado para um indivíduo imune.

Quando existem observações censuradas, não é possível, regra geral, identificar inequivocamente quais os indivíduos imunes. O facto de não serem

Tabela 3.1: Valores/casos para (δ_i, y_i).

Valores possíveis para (δ_i, y_i)	Descrição da situação
$(0, 0)$	Censurado, Imune
$(0, 1)$	Censurado, Suscetível
$(1, 1)$	Observado, Suscetível

observados todos os $y_i\prime s$ (aqueles correspondentes aos tempos de vida censurados), conduz a uma situação de dados incompletos. Para este caso específico, um método adequado é o algoritmo EM ([39]), uma vez que é um método iterativo que permite obter as estimativas de máxima verosimilhança dos parâmetros em situações em que existem observações omissas ([38]).

Admitindo que a censura é não informativa e à direita, a função de verosimilhança para uma amostra de dimensão n é dada por

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

No caso do modelo (3.2), a verosimilhança total observada é

$$L_O = \prod_{i=1}^{n} [(1-p) f_d(t_i)]^{\delta_i} [p + (1-p) S_d(t_i)]^{1-\delta_i}$$

Se todos os $y_i\prime s$ fossem observados, de acordo com a Tabela 3.1, teríamos a seguinte verosimilhança completa

$$L_C = \prod_{i=1}^{n} [[(1-p) f_d(t_i)]^{y_i}]^{\delta_i} [p^{1-y_i} [(1-p) S_d(t_i)]^{y_i}]^{1-\delta_i}$$

Tendo em conta que $f_d(t_i) = h_d(t_i) S_d(t_i)$ e que, dados os valores possíveis para (δ_i, y_i), $(1-\delta_i)(1-y_i) = 1 - y_i$, a verosimilhança anterior pode ser fatorizada em

$$L_C = \prod_{i=1}^{n} (1-p)^{y_i} p^{1-y_i} \prod_{i=1}^{n} h_d(t_i)^{y_i \delta_i} S_d(t_i)^{y_i} = L_{C_1} L_{C_2}$$

Conclui-se assim que a contribuição das diferentes possibilidades do par (δ_i, y_i) para a função de verosimilhança é a que se exibe na Tabela 3.2.

A etapa E do algoritmo EM consiste em determinar o valor esperado do logaritmo da verosimilhança completa em relação à distribuição dos $Y\prime s$

Tabela 3.2: Resultados.

Valores possíveis para (δ_i, y_i)	Contribuição para a verosimilhança
$(0,0)$	p
$(0,1)$	$(1-p)S_d(t_i)$
$(1,1)$	$(1-p)h_d(t_i)S_d(t_i)$

não observados, condicional aos valores atuais dos parâmetros e aos dados observados O, onde $O = \{y_i \text{ observado}, (t_i, \delta_i),\ i = 1, ..., n\}$. No entanto, como em relação às observações censuradas o logaritmo de L_C é linear em Y, para calcular o valor esperado de $\log L_C$ basta substituir, na verosimilhança completa, os valores não observados de Y pelos respetivos valores esperados, denotados por τ_i. Então, temos

$$\tau_i = E(Y|O) = P(Y_i = 1|T_i > t_i, \delta_i = 0) = \frac{(1-p)S_d(t_i)}{S(t_i)}$$

Assim, na verosimilhança completa, cada y_i é substituído por ω_i, onde ω_i é definido da forma que se segue

$$\omega_i = \begin{cases} 1 \text{ se } \delta_i = 1 \\ \tau_i \text{ se } \delta_i = 0 \end{cases}$$

Na etapa E, cada observação censurada i é atribuída à subpopulação $Y = 1$ com probabilidade τ_i e à subpopulação $Y = 0$ com probabilidade $1 - \tau_i$.

Depois de substituir y_i por ω_i em $L_{C_1}L_{C_2}$, obtém-se a seguinte verosimilhança "esperada"

$$L_E = \prod_{i=1}^{n}(1-p)^{\omega_i}p^{1-\omega_i}\prod_{i=1}^{n}h_d(t_i)^{\omega_i\delta_i}S_d(t_i)^{\omega_i} = L_{E_1}L_{E_2}$$

que também pode ser escrita como

$$L_E = \prod_{i=1}^{n}q^{\omega_i}(1-q)^{1-\omega_i}\prod_{i=1}^{n}h_d(t_i)^{\omega_i\delta_i}S_d(t_i)^{\omega_i} = L_{E_1}L_{E_2}$$

se atendermos a que $q = 1 - p$.

Para efetuar a etapa M, é necessário maximizar as duas componentes do logaritmo da função de verosimilhança

$$\log L_{E_1} = \sum_{i=1}^{n} [\omega_i \log q + (1-\omega_i)\log(1-q)]$$

$$\iff \log L_{E_1} = \sum_{i=1}^{m} [\omega_i \log q + (1-\omega_i)\log(1-q)] + \sum_{i=m+1}^{n} [\omega_i \log q + (1-\omega_i)\log(1-q)]$$

(como para $i = 1, ..., m \implies \delta_i = 1 \implies \omega_i = 1$ e para $i = m+1, ..., n \implies \delta_i = 0 \implies \omega_i = \tau_i$)

$$\iff \log L_{E_1} = m \log q + \sum_{i=m+1}^{n} [\tau_i \log q + (1-\tau_i)\log(1-q)]$$

$$\iff \log L_{E_1} = m \log q + \sum_{i=m+1}^{n} \tau_i (\log q - \log(1-q)) + \sum_{i=m+1}^{n} \log(1-q)$$

$$\iff \log L_{E_1} = m \log q + (n-m)\log(1-q) + \sum_{i=m+1}^{n} \tau_i [\log q - \log(1-q)]$$

Na segunda componente

$$\log L_{E_2} = \sum_{i=1}^{n} [\delta_i \omega_i \log h_d(t_i) + \omega_i \log S_d(t_i)]$$

$$\iff \log L_{E_2} = \sum_{i=1}^{m} [\delta_i \omega_i \log h_d(t_i) + \omega_i \log S_d(t_i)] + \sum_{i=m+1}^{n} [\delta_i \omega_i \log h_d(t_i) + \omega_i \log S_d(t_i)]$$

$$\iff \log L_{E_2} = \sum_{i=1}^{m} [\log h_d(t_i) + \log S_d(t_i)] + \sum_{i=m+1}^{n} \tau_i \log S_d(t_i)]$$

É possível obter uma expressão explícita para o estimador de q, como se indica de seguida.

$$\frac{\partial \log L_{Ei}}{\partial q} = \frac{m}{q} - \frac{(n-m)}{1-q} + \sum_{i=m+1}^{n} \tau_i (\frac{1}{q} + \frac{1}{1-q}) = 0$$

$$\iff m(1-q) - q(n-m) + [(1-q)+q] \sum_{i=m+1}^{n} \tau_i = 0$$

$$\iff \widehat{q} = \frac{1}{n}[m + \sum_{i=m+1}^{n} \tau_i]$$

A expressão a usar no algoritmo EM será

$$\therefore q^{(k+1)} = \frac{1}{n}[m + \sum_{i=m+1}^{n} \tau_i^{(k)}]$$

Modelos de não mistura

Nos modelos de cura de não mistura mantém-se o facto da função de sobrevivência associada aos indivíduos imunes ser uma função imprópria e daí existir limite superior na função de risco cumulativa.

Seja $S(t)$ a função de sobrevivência de T e $H(t)$, tal que $S(t) = \exp[-H(t)]$. Se $S(\infty) > 0$ então existe $\theta < \infty$, tal que $H(\infty) = \theta$. Considerando $F(t)$ como a função de distribuição (própria) de uma variável aleatória não negativa, Yakovlev *et al.* (1993) considera que $H(t) = \theta F(t)$ como uma forma possível de caraterizar esta propriedade. Então o modelo de cura de não mistura pode ser escrito na forma

$$S(t) = \exp[-\theta F(t)]$$

A correspondente função de risco é

$$h(t) = \theta f(t)$$

onde $f(t)$ é a função densidade correspondente a $F(t)$.

Assim, a probabilidade de cura é $P(cura) = e^{-\theta}$, o que corresponde a $S(\infty)$. A probabilidade de cura a cada instante é determinada como se indica

$$P(Cura|T \geq t) = \exp[-\theta(1 - F(t))]$$

Com covariáveis

Os modelos de cura permitem a inclusão de covariáveis de modo a contemplar a existência de determinadas caraterísticas dos indivíduos na população que poderão influenciar a ocorrência do acontecimento de interesse.

Sejam $\mathbf{x}$ e $\mathbf{z}$ vetores de covariáveis associados a determinado indivíduo.

Assumindo a existência de covariáveis que influenciam o tempo de vida dos indivíduos doentes tem-se o seguinte modelo de cura

$$S(t;\mathbf{x}) = p + (1-p)S_d(t;\mathbf{x})$$

Se as covariáveis forem aplicadas apenas na taxa de cura, o que significa que existem fatores que influenciam a imunidade, vem que

$$S(t;\mathbf{z}) = p(\mathbf{z}) + [1-p(\mathbf{z})]S_d(t)$$

ou ainda

$$S(t;\mathbf{z}) = p(\mathbf{z}) + [1-p(\mathbf{z})]S(t|Y=1)$$

Se se admitir que as covariáveis influenciam quer o tempo de vida dos indivíduos suscetíveis quer a taxa de cura, o modelo é apresentado da seguinte forma

$$S(t;\mathbf{x};\mathbf{z}) = p(\mathbf{z}) + [(1-p(\mathbf{z})]S_d(t;\mathbf{x})$$

ou ainda

$$S(t;\mathbf{x};\mathbf{z}) = p(\mathbf{z}) + [1-p(\mathbf{z})]S(t|Y=1;\mathbf{x})$$

Neste último caso, os vetores $\mathbf{x}$ e $\mathbf{z}$ podem ter componentes em comum.

Relativamente à taxa de cura, fazendo uso da regressão logística, a variável Y é caracterizada por

$$P(Y=1|\mathbf{z}) = 1 - p(\mathbf{z}) = \frac{\exp(\gamma_0 + \mathbf{z}\gamma)}{1+\exp(\gamma_0 + \mathbf{z}\gamma)}$$

onde $\gamma = (\gamma_1, \gamma_2, ...\gamma_k)'$ e $\gamma_0, \gamma_1, ..., \gamma_k$ são os coeficientes de regressão.

Podem ser consideradas outras funções de ligação como a função complementar log-log e as funções de ligação *probit* nos modelos lineares generalizados para dados binários, *[51]* .

Seja $S_{d0}(t)$ função de sobrevivência subjacente. Tal como acontece com o modelo de riscos proporcionais na Análise de Sobrevivência usual, pode-se assumir que

$$S_d(t|x) = S_{d0}(t)^{\exp(\beta' x)}$$

e

$$h_d(t|x) = h_{d0}(t)\exp(\beta' x)$$

onde $h_d(t|x)$ e $h_{d0}(t)$ são as funções de risco associadas a $S_d(t)$ e $S_{d0}(t)$, respetivamente. Este modelo é conhecido como o modelo de cura de mistura de riscos proporcionais.

Capítulo 4

Aplicação ao Cancro da Mama Feminino

Segundo a Liga Portuguesa Contra o Cancro [36], o cancro da mama é um problema de saúde pública que apesar de não ser dos mais letais, tem uma alta incidência e uma alta mortalidade, sobretudo na mulher (apenas 1 em cada 100 cancros se desenvolve no homem). Atualmente em Portugal, com uma população feminina de 5 milhões, surgem 4500 novos casos de cancro da mama por ano, ou seja, 11 novos casos por dia, morrendo por dia 4 mulheres com esta doença.

Em 2012, o INE[1] publicou um relatório [27], que indica que no ano de 2010 morreram 30 mulheres por cancro da mama, em cada 100 mil. Acrescenta ainda que a taxa de mortalidade feminina por cancro da mama apresenta uma tendência de crescimento, tendo passado de $26,6$ óbitos em 2006, para $30,3$ em 2010, por cada cem mil mulheres.

O Ministério da Saúde [40] define o cancro da mama como um tumor maligno que se desenvolve nas células do tecido mamário. Alerta que o diagnóstico precoce do cancro da mama é fundamental, pois aumenta as hipóteses de cura já que evita que o cancro se espalhe para outras partes do corpo, favorecendo o prognóstico, a recuperação e a reabilitação. A escolha entre as diversas opções de tratamento depende do estadio da doença, do tipo de tumor e do estado geral de saúde do indivíduo. Dependendo das necessidades de cada doente, o médico opta por um dos tratamentos que se lista de seguida ou pela combinação dos mesmos.

- Cirurgia: é o tratamento inicial mais comum e o principal tratamento local. O tumor da mama será removido, assim como os gânglios linfáticos

[1] Instituto Nacional de Estatística

da axila. Estes gânglios filtram a linfa que flui da mama para outras partes do corpo e é através deles que o cancro pode alastrar.

- Radioterapia: utiliza raios de alta energia que têm a capacidade de destruir as células cancerosas e impedir que elas se multipliquem.
- Quimioterapia: é a utilização de drogas que agem na destruição das células malignas.
- Hormonoterapia: tem como finalidade impedir que as células malignas continuem a receber a hormona que estimula o seu crescimento.

Segundo a *American Joint Committe of Cancer* (AJCC) (2002) [6], a glândula mamária, situada na parede torácica anterior, é constituída por tecido glandular com um estroma fibroso denso. O tecido glandular consiste em lóbulos que se agrupam em 15-25 lobos dispostos aproximadamente num padrão *piramidal*. Vários canais *major* e *minor* ligam as glândulas lácteas ao mamilo. Existem pequenos ductos lactíferos na mama que convergem em canais coletores até à base do mamilo. A maioria dos cancros formam-se inicialmente nos ductos lobulares da mama, e, neste caso, o tumor designa-se por carcinoma ductal. O tecido glandular é mais abundante na parte superior externa da mama; como consequência, metade de todos os cancros da mama ocorrem nessa área.

Tabela 4.1: Tumor primário (T).

TX	Não é possível determinar a extensão do tumor primário
Tis	Carcinoma in situ
T0	Não existe evidência do tumor primário
T1	Tamanho do tumor menor ou igual a 2 cm
T2	Tamanho do tumor entre 2 e 5 cm
T3	Tamanho do tumor maior que 5 cm
T4	Tumor com qualquer tamanho, prolongando-se até à parede torácica (não incluindo o músculo peitoral) e/ou até à pele (edema ou ulceração da pele da mama ou nódulos satélite confinados à mesma mama)

O estadiamento aqui considerado é o apresentado no livro [6], segundo o tamanho do tumor primário (T), o número de gânglios linfáticos invadidos

(N) e a presença de metástases à distância (M). A categorização de T é apresentada na Tabela 4.1. A determinação de N pode ser feita por meio de avaliação clínica, com recurso a exame médico ou imagiológico, ou através de análise anátomo-patológica (identificada com a letra p). Na Tabela 4.2 está um excerto da informação sobre os gânglios publicada por [6]. Na Tabela 4.3 encontra-se a classificação referente às metástases (M).

Tabela 4.2: Gânglios linfáticos (N).

NX	Os gânglios linfáticos regionais não estão acessíveis
pN0	Não existe metastização dos gânglios linfáticos regionais
pN1	Entre 1 a 3 gânglios linfáticos axilares metastizados
pN2	Entre 4 a 9 gânglios linfáticos axilares metastizados
pN3	Metástases em 10 ou mais gânglios linfáticos axilares

Tabela 4.3: Metástases à distância (M).

MX	Não é possível determinar a existência de metástases à distância
M0	Ausência de metástases à distância
M1	Presença de metástases à distância

A obtenção do estadio é feita através da conjugação de T, N e M, como descrito na Tabela 4.4

4.1 Caraterização da base de dados.

4.1.1 Metodologia

Trata-se de um estudo prospetivo que utiliza dados fornecidos pelo Registo Oncológico - Sul. A base de dados tem 833 registos, os quais correspondem a mulheres diagnosticadas com cancro da mama (não inflamatório) entre 1998 e 2005, na Região Autónoma da Madeira (RAM).

4.1.2 Seleção dos indivíduos

Os indivíduos foram selecionados de acordo com os critérios que se indica em seguida.

Critérios de inclusão

Mulheres diagnosticadas com cancro da mama, entre janeiro de 1998 e dezembro de 2005, na RAM.

Critérios de exclusão

Mulheres diagnosticadas com cancro da mama do tipo inflamatório.

Mulheres cuja entrada na base de dados foi devida a uma recidiva, ou seja, o tumor primário foi diagnosticado antes do período em estudo.

Tabela 4.4: Classificação dos estadios.

Estadio 0	Tis	N0	M0
Estadio I	T1	N0	M0
Estadio IIA	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
Estadio IIB	T2	N1	M0
	T3	N0	M0
Estadio IIIA	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Estadio IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
Estadio IIIC	Qualquer T	N3	M0
Estadio IV	Qualquer T	Qualquer N	M0

Ferramentas informáticas

A análise dos dados foi realizada através dos programas estatísticos *PASW Statistics*, versão 18 e R, versão 2.11.1. No R foram ainda utilizadas as

bibliotecas específicas *survival*, versão 2.36-5 e *KMsurv*, versão 0.1-4.

4.1.3 Variáveis da base de dados

A informação solicitada estava compilada em 11 variáveis, que por sua vez deram origem a mais duas, nomeadamente, os dias de sobrevivência e o tipo de evento. Assim, as variáveis que compõem a base de dados são:

1. Data do diagnóstico
2. Data da morte / *follow-up*
3. Dias de sobrevivência
4. Estado do indivíduo
5. Idade
6. Estadiamento
7. Cirurgia
8. Radioterapia
9. Quimioterapia
10. Tipo de Quimioterapia
11. Hormonoterapia
12. Causa da Morte
13. Tipo de Evento

As variáveis referidas do ponto 7 ao ponto 11 dizem respeito ao tratamento a que os indivíduos foram submetidos.

4.1.4 Codificação das variáveis

As variáveis "Cirurgia", "Radioterapia", "Quimioterapia"e "Hormonoterapia"estão codificadas com 0 - "Não fez"e 1 - "Fez".

A variável "Tipo de Quimioterapia"está codificada da seguinte forma:

$$\begin{cases} 1 - \text{"Adjuvante"}, \\ 2 - \text{"Adjuvante e neoadjuvante"}, \\ 3 - \text{"Neoadjuvante"}, \\ 4 - \text{"Só quimioterapia (sem cirurgia)"}, \\ 999 - \text{"Desconhecido"}. \end{cases}$$

A variável Tipo de Evento foi criada a partir da variável Estado do indivíduo. Os indivíduos com o estado "Vivo"foram codificados com o Tipo de evento "Censurado". Os indivíduos com o estado "Falecido"e Causa da morte "Cancro da mama"foram codificados com o tipo de evento "Observado". Os falecidos com outra causa de morte ou com causa de morte desconhecida foram codificados como censurados. Assim, a variável "Tipo de evento"assume o valor 1 para os indivíduos com o acontecimento de interesse observado e, para os indivíduos com tempo de vida censurado, assume o valor 0.

Tabela 4.5: Distribuição do estadiamento inicial.

		N.º casos	%
Estadiamento	0	4	0,5
	I	112	13,4
	II	253	30,4
	III	78	9,4
	IV	52	6,2
	Desc.	334	40,1
	Total	833	100,0

Inicialmente a variável relativa ao estadiamento (Tabela 4.5) era composta por cinco categorias, uma por cada estadio. Contudo, devido à escassez de indivíduos no estadio 0 (apenas 4), optou-se por juntá-los aos do estadio I (ver Tabelas 4.6 e 4.7).

4.1.5 Análise descritiva dos dados

Embora os dados digam respeito a mulheres diagnosticadas com cancro da mama entre 1998 e 2005, o *follow-up* foi feito até fevereiro de 2012. O objetivo foi ter um período mínimo de 5 anos para todos os indivíduos.

Começou-se a analisar a base de dados explorando as suas variáveis. A média das idades situa-se nos 60 anos, com um desvio padrão de 14 anos (ver Figura 4.1). Estudando o comportamento da idade em relação ao estadiamento, quando este é conhecido, observa-se na Figura 4.2 que existem

diferenças significativas ($p < 0,001$) entre as médias das idades das mulheres nos diversos estadios. Verifica-se que, à medida que o estadio se agrava, a média das idades aumenta. Por exemplo, nos estadios 0 ou I, a média das idades é de 54 anos, com desvio padrão de 12, enquanto no último estadio a média é de 62 anos, tendo, no entanto, um menor desvio padrão ($s = 11$).

Descriptives

IDADE

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
0 ou I	116	53,55	11,992	1,113	51,35	55,76	32	80
II	253	57,60	12,710	,799	56,02	59,17	27	88
III	78	58,44	13,316	1,508	55,43	61,44	31	82
IV	52	61,88	10,675	1,480	58,91	64,86	39	80
Desc.	334	63,74	13,915	,761	62,24	65,24	25	93
Total	833	59,84	13,539	,469	58,92	60,76	25	93

Test of Homogeneity of Variances

IDADE

Levene Statistic	df1	df2	Sig.
2,484	4	828	,042

Figura 4.1: Comparação da média das idades por estadiamento.

De seguida optou-se por categorizar as idades em quatro grupos etários, semelhante ao que foi feito em Fields *et al.* (2010), [20]. Nas Tabelas resumo 4.6 e 4.7 verifica-se a diminuição da proporção de elementos nos estadios 0 ou I à medida que o grupo etário vai aumentando: <40 anos – $25,9\%$; 40 a 49 anos – $25,0\%$; 50 a 59 anos – $14,6\%$ e >=60 anos – $8,0\%$. No último estadio verifica-se o inverso: <40 anos – $3,7\%$; 40 a 49 anos – $3,8\%$; 50 a 59 anos – $6,5\%$ e >=60 anos – $7,3\%$. A distribuição dos indivíduos com estadio desconhecido também regista um maior número de casos nos grupos etários mais elevados.

Todas as mulheres nos estadios 0, I ou II foram tratadas com cirurgia. Reciprocamente, no estadio IV, o último estadio da doença, a maioria das mulheres não fez cirurgia (67,3%).

Na Tabela 4.6, pode-se observar que, tratando-se de estadios conhecidos, a aplicação de radioterapia é mais frequente nos primeiros estadios (0 ou I – $22,5\%$; II – $36,8\%$), apesar do facto de a maior proporção de mulheres que não foi submetida a este tratamento estar no estadio II ($26,6\%$). Note-se que o estadio II é o que tem maior número de casos ($30,4\%$), pelo que apresenta naturalmente maiores percentagens. Em relação aos indivíduos que não fizeram radioterapia, quase metade ($47,3\%$) tinha estadio desconhecido.

Descriptives

IDADE

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
0 ou I	116	53,55	11,992	1,113	51,35	55,76	32	80
II	253	57,60	12,710	,799	56,02	59,17	27	88
III	78	58,44	13,316	1,508	55,43	61,44	31	82
IV	52	61,88	10,675	1,480	58,91	64,86	39	80
Total	499	57,23	12,639	,566	56,12	58,35	27	88

Test of Homogeneity of Variances

IDADE

Levene Statistic	df1	df2	Sig.
1,435	3	495	,232

ANOVA

IDADE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2843,513	3	947,838	6,116	,000
Within Groups	76712,054	495	154,974		
Total	79555,567	498			

Figura 4.2: Comparação das médias das idades por estadiamento (conhecido).

Na Tabela 4.7 verifica-se que há uma maior percentagem de mulheres a não fazer radioterapia, em especial no estadio IV $(88,5\%)$ e no estadio desconhecido $(74,6\%)$. Apenas nos estadios 0 ou I observa-se o contrário $(40,5\%)$.

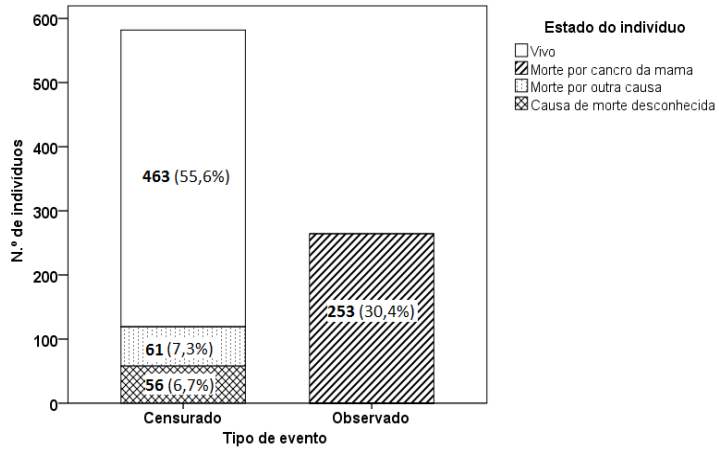

Figura 4.3: Relação entre o Tipo de evento e o Estado do indivíduo.

Em relação à quimioterapia, mais de metade ($57,0\%$) dos indivíduos que não fizeram esta terapêutica apresentavam estadio desconhecido (Tabela 4.6). Os indivíduos no estadio II são os que mais são submetidos a este tratamento ($41,2\%$). Quanto ao tipo de quimioterapia, verifica-se que $60,0\%$ dos indivíduos que fizeram quimioterapia sem terem sido submetidos a cirurgia, estão no estadio IV. Isto deve-se em parte à variedade de objetivos deste tratamento, que poderão ser desde a diminuição do tamanho do tumor até à prestação de cuidados paliativos. É este último o que mais se adequa aos indivíduos no último estadio.

Tabela 4.6: Caraterísticas das mulheres diagnosticadas com cancro da mama entre 1998 e 2005. Proporção dentro das variáveis.

			Estadiamento									
	Total	%	0 ou I	%	II	%	III	%	IV	%	Desc.	%
Dim. da amostra	846	100,0	116	13,9	253	30,4	78	9,4	52	6,2	334	40,1
Grupo Etário*												
<40 anos	54		14	25,9	19	35,2	6	11,1	2	3,7	13	24,1
40 a 49 anos	156		39	25,0	50	32,1	16	10,3	6	3,8	45	28,8
50 a 59 anos	199		29	14,6	72	36,2	18	9,0	13	6,5	67	33,7
>=60 anos	424		34	8,0	112	26,4	38	9,0	31	7,3	209	49,3
Cirurgia*												
Fez	745		116	15,6	253	34,0	70	9,4	17	2,3	289	38,8
Não fez	88		0	0,0	0	0,0	8	9,1	35	39,8	45	51,1
Radioterapia*												
Fez	307		69	22,5	113	36,8	34	11,1	6	2,0	85	27,7
Não fez	526		47	8,9	140	26,6	44	8,4	46	8,7	249	47,3
Quimioterapia*												
Fez	456		51	11,2	188	41,2	68	14,9	30	6,6	119	26,1
Não fez	377		65	17,2	65	17,2	10	2,7	22	5,8	215	57,0
Tipo de quimio*												
Adjuvante	292		45	15,4	156	53,4	28	9,6	8	2,7	55	18,8
Adj. e neoadj.	103		6	5,8	31	30,1	30	29,1	4	3,9	32	31,1
Neoadjuvante	19		0	0,0	1	5,3	5	26,3	0	0,0	13	68,4
Só quimio#	38		0	0,0	0	0,0	4	10,5	18	47,4	16	42,1
Hormonoterapia*												
Fez	361		64	17,7	134	37,1	37	10,2	5	1,4	121	33,5
Não fez	472		52	11,0	119	25,2	41	8,7	47	10,0	213	45,1

#Sem cirurgia

Desc. = Desconhecido; Dim. = Dimensão

*p<0,001.

Dadas as caraterísticas da doença na fase inicial, todos os indivíduos nos estadios 0 ou I são submetidos a cirurgia e frequentemente esse tratamento basta, pelo que apenas $44,0\%$ fazem quimioterapia (Tabela 4.7). Já no último

estadio, inverte-se a situação: a cirurgia é realizada apenas em $32,7\%$ dos casos, enquanto a quimioterapia é aplicada em $57,7\%$ das mulheres no estadio IV.

Tabela 4.7: Caraterísticas das mulheres diagnosticadas com cancro da mama entre 1998 e 2005. Proporção dentro do estadiamento.

			Estadiamento									
	Total	%	0 ou I	%	II	%	III	%	IV	%	Desc.	%
Grupo Etário*												
<40 anos	54	6,5	14	12,1	19	7,5	6	7,7	2	3,8	13	3,9
40 a 49 anos	156	18,7	39	33,6	50	19,8	16	20,5	6	11,5	45	13,5
50 a 59 anos	199	23,9	29	25,0	72	28,5	18	23,1	13	25,0	67	20,1
>=60 anos	424	50,9	34	29,3	112	44,3	38	48,7	31	59,6	209	62,6
Cirurgia*												
Fez	745	89,4	116	100,0	253	100,0	70	89,7	17	32,7	289	86,5
Não fez	88	10,6	0	0,0	0	0,0	8	10,3	35	67,3	45	13,5
Radioterapia*												
Fez	307	36,9	69	59,5	113	44,7	34	43,6	6	11,5	85	25,4
Não fez	526	63,1	47	40,5	140	55,3	44	56,4	46	88,5	249	74,6
Quimioterapia*												
Fez	456	54,7	51	44,0	188	74,3	68	87,2	30	57,7	119	35,6
Não fez	377	45,3	65	56,0	65	25,7	10	12,8	22	42,3	215	64,4
Tipo de quimio*												
Adjuvante	292	64,6	45	88,2	156	83,0	28	41,8	8	26,7	55	47,4
Adj. e neoadj.	103	22,8	6	11,8	31	16,5	30	44,8	4	13,3	32	27,6
Neoadjuvante	19	4,2	0	0,0	1	0,5	5	7,5	0	0,0	13	68,4
Só quimio#	38	8,4	0	0,0	0	0,0	4	6,0	18	60,0	16	42,1
Hormonoterapia*												
Fez	361	43,3	64	55,2	134	53,0	37	47,4	5	9,6	121	36,2
Não fez	472	56,7	52	44,8	119	47,0	41	52,6	47	90,4	213	63,8

#Sem cirurgia

Desc. = Desconhecido; Dim. = Dimensão

*p<0,001.

A realização de hormonoterapia é mais frequente nos estadios iniciais. Na Tabela 4.6 observa-se que o estadio 0 ou I e o estadio II têm uma proporção de $17,7\%$ e $37,1\%$, respetivamente, de mulheres que se submetem a esta terapêutica. Dentro de cada um destes estadios a proporção de mulheres que fazem hormonoterapia também é maior do que as que não fazem (Tabela 4.7).

Na Figura 4.3 ilustra-se a relação entre as variáveis Tipo de evento e Estado do indivíduo. Observa-se que $30,4\%$ das mulheres morreram por cancro da mama. Contudo, a maioria dos indivíduos tem o tempo de vida censurado

sendo a causa principal da censura o facto de estarem vivos (55, 6%).

4.2 Estimativas de Kaplan-Meier

O conhecimento da função de sobrevivência é crucial em estudos que envolvam a análise do tempo de vida de indivíduos, pois esta função permite obter a probabilidade de um indivíduo sobreviver para além de determinado instante. O estimador de Kaplan-Meier, que é um estimador não-paramétrico, é uma das formas mais utilizadas para este fim, na presença de dados censurados.

Na Figura 4.4 visualiza-se a estimativa de Kaplan-Meier da função de sobrevivência. Verifica-se que aos 14 anos de *follow-up* a sobrevida estimada encontra-se em 63, 8%.

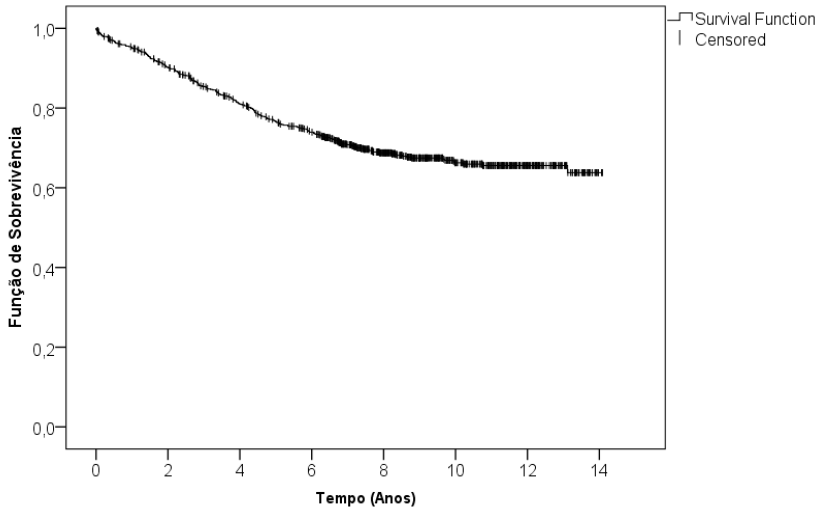

Figura 4.4: Estimativa de Kaplan-Meier da função de sobrevivência das mulheres diagnosticadas com cancro da mama entre 1998 e 2005.

Na Figura 4.5 observamos que, tal como esperado, o prognóstico é tanto melhor quanto menor for o estadio. As mulheres diagnosticadas com os estadios 0 ou I, a partir dos 8 anos após o diagnóstico mantêm uma probabilidade de sobrevivência acima dos 85% . Por seu lado, as mulheres diagnosticadas no último estadio têm uma probabilidade quase nula (6, 9%) de estarem vivas ao fim de 14 anos e é nos dois primeiros anos após o diagnóstico que se verifica um decrescimento acentuado da probabilidade de sobrevivência.

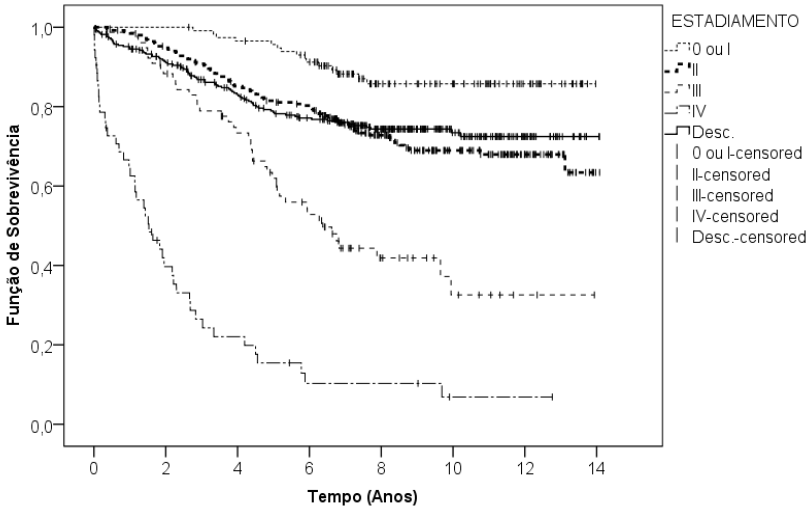

Figura 4.5: Estimativa de Kaplan-Meier das funções de sobrevivência por estadiamento.

O tratamento com cirurgia é claramente vantajoso, embora também esteja subjacente o estadio. Na Figura 4.6 observamos a grande diferença entre a probabilidade de sobreviver para além de 14 anos quando o indivíduo foi submetido a cirurgia $(68,7\%)$ ou quando tal não aconteceu $(17,1\%)$.

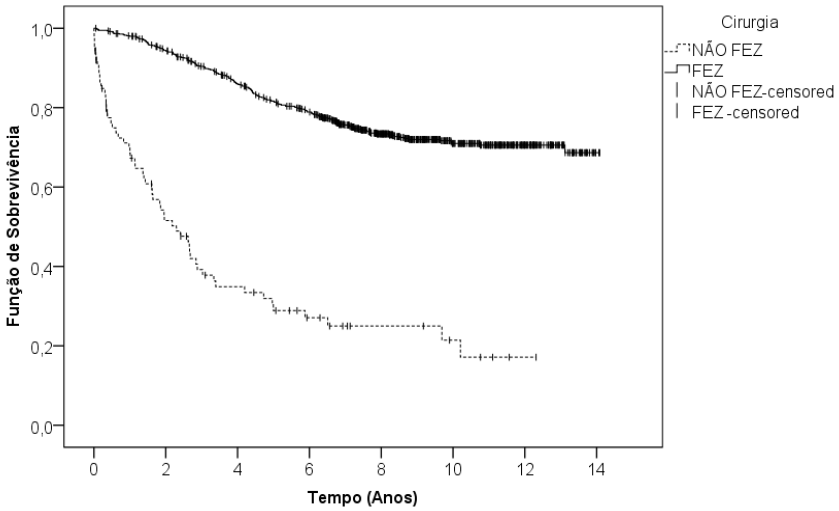

Figura 4.6: Estimativa de Kaplan-Meier das funções de sobrevivência das mulheres que se submeteram ou não a cirurgia.

A probabilidade de sobrevivência diminui de forma semelhante para todas as mulheres quer façam radioterapia ou não (Figura 4.7). Só há uma ligeira diferença no primeiro ano após o diagnóstico, em que é mais benéfico para as mulheres que o fizeram.

Em relação à hormonoterapia, também se observa uma melhor situação para as mulheres submetidas a este tratamento (Figura 4.8).

O comportamento das curvas das estimativas da função de sobrevivência dos indivíduos que fizeram quimioterapia e dos que não fizeram gera alguma surpresa (Figura 4.9). Nos primeiros dois anos, observa-se um melhor prognóstico nas mulheres que se submeteram ao tratamento em questão. Contudo, a partir desse ponto, a situação inverte-se, chegando aos 14 anos com uma diferença de probabilidade de sobrevivência de cerca de 20%, a favor dos indivíduos que não fizeram quimioterapia.

Na Figura 4.10, pode ser observado que o melhor prognóstico está associado aos indivíduos que fizeram quimioterapia adjuvante ou neoadjuvante (esta última, contudo, poderá ser consequência do escasso número de casos). Em seguida surge a quimioterapia neoajuvante e adjuvante e, por fim, a quimioterapia paliativa. De salientar que "a escolha do tipo de quimioterapia está diretamente relacionada com o estadio em que a doente se encontra (ver Tabelas 4.6 e 4.7).

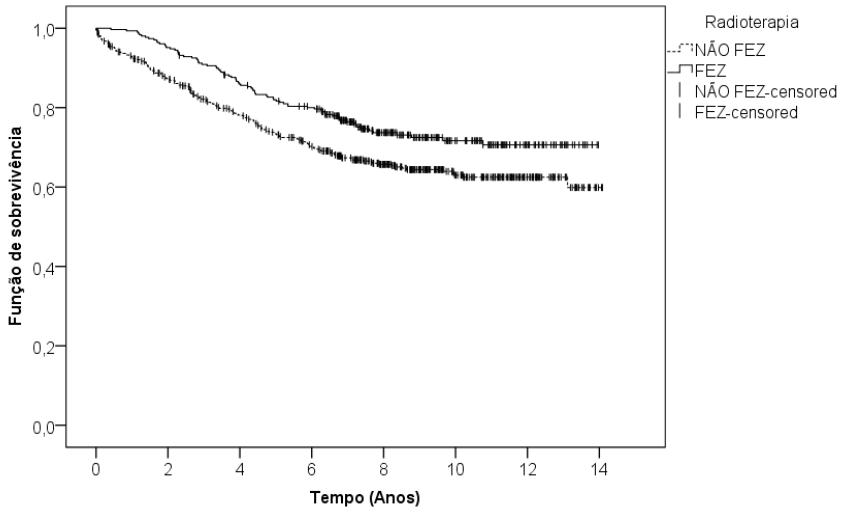

Figura 4.7: Estimativa de Kaplan-Meier das funções de sobrevivência das mulheres que se submeteram ou não a radioterapia.

Para averiguar a razão desta mudança de prognóstico aos 2 anos de *follow-up* foi feita a representação da estimativa de Kaplan-Meier para cada estadio.

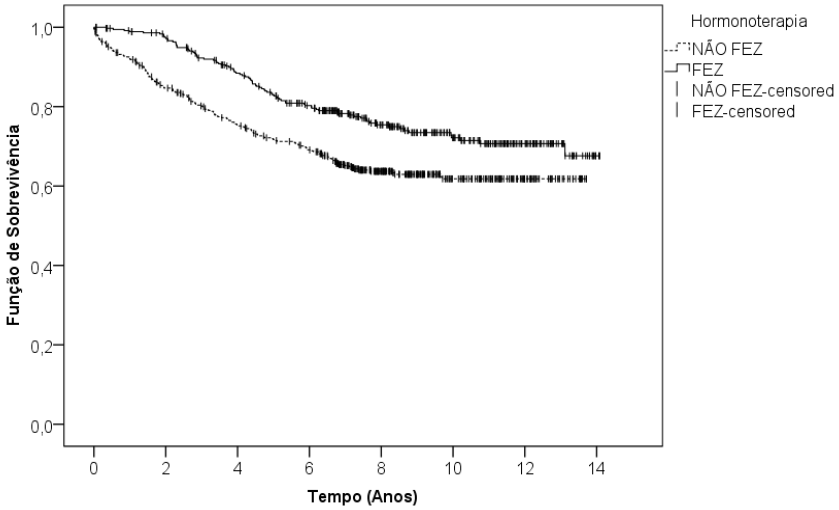

Figura 4.8: Estimativa de Kaplan-Meier das funções de sobrevivência das mulheres que se submeteram ou não a hormonoterapia.

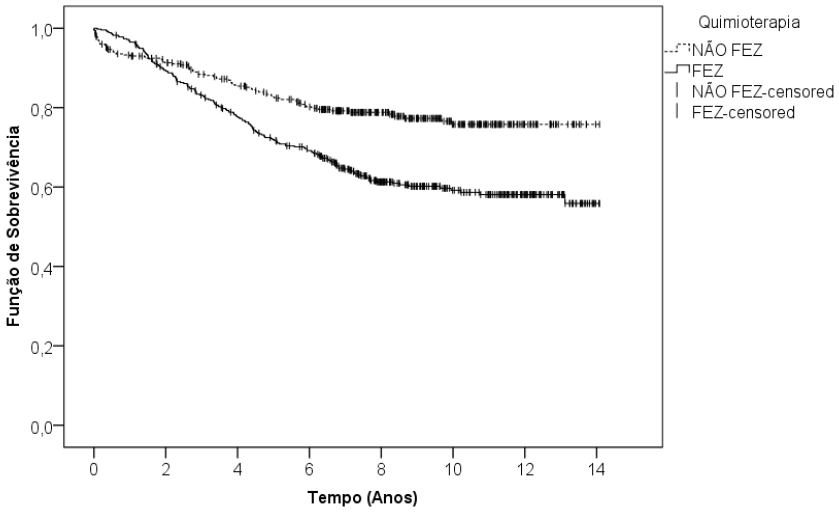

Figura 4.9: Estimativa de Kaplan-Meier para os indivíduos que fizeram ou não quimioterapia.

Nos estadios 0 ou I, os indivíduos que fizeram quimioterapia apresentam uma melhor probabilidade de sobrevivência em relação aos restantes. No estadio III, ocorre o mesmo até aos dez anos e daí em diante a situação inverte-se ligeiramente. Quanto ao último estadio, verifica-se uma diferença

acentuada nos três primeiros anos, favorecendo os que fizeram quimioterapia, mas, a partir daí, esbate-se essa diferença, eventualmente devido ao reduzido número de casos (apenas doze).

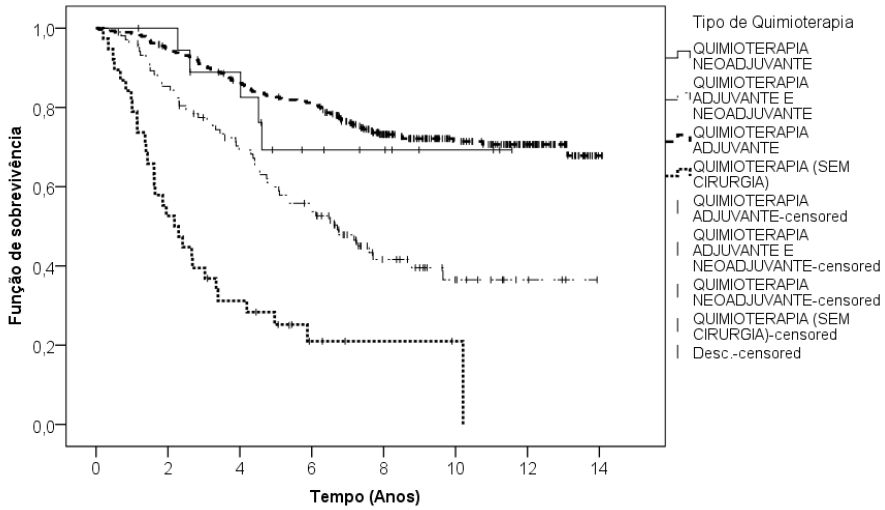

Figura 4.10: Distribuição dos indivíduos submetidos a quimioterapia por tipo.

O gráfico do estadio desconhecido (Figura 4.15) é muito semelhante ao do estadio II (Figura 4.12), o que levanta a suspeita de que a maior parte destes casos estejam nesse estadio.

Ao contrário do que acontece nos outros estadios, o estadio II apresenta um melhor prognóstico para as mulheres que não fizeram quimioterapia. Este facto curioso pode ser uma das razões para a inversão do prognóstico aos dois anos (Figura 4.9) já que este é o grupo com maior representatividade dentro dos estadios conhecidos. Por outro lado, os indivíduos neste estadio são muito heterogéneos, nomeadamente em relação à metastização de gânglios linfáticos (N0 e N1). Por esta razão, optou-se por analisar o prognóstico destes indivíduos, consoante a variável Gânglios (à semelhança do que foi feito por Fields *et al.* (2010), [20]), definida da seguinte forma:

$$\text{Gânglios}=\begin{cases} \text{N0} \\ \text{N1} \end{cases}$$

Note-se que não foram considerados os casos em que os gânglios tinham classificação desconhecida.

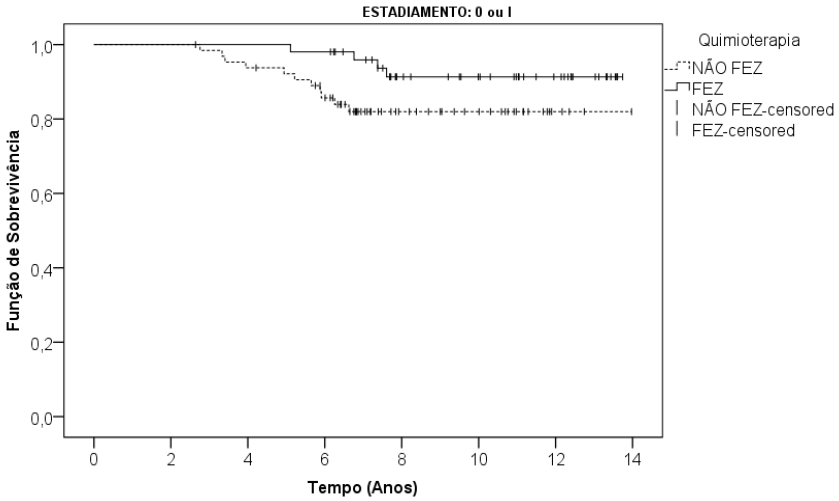

Figura 4.11: Estimativa da função de sobrevivência no estadio 0 ou I consoante tratamento com quimioterapia.

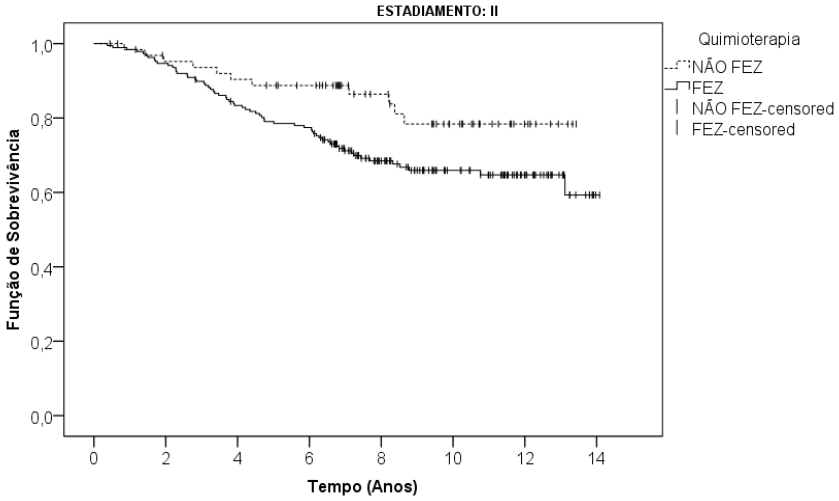

Figura 4.12: Estimativa da função de sobrevivência no estadio II consoante tratamento com quimioterapia.

No estadio II, a distribuição dos indivíduos nas variáveis Quimioterapia e Gânglios é apresentada na Tabela 4.8. Observa-se uma associação estatisticamente significativa ($p < 0,001$) entre as variáveis em questão. Mais de metade das mulheres que não fizeram quimioterapia tinham gânglios negati-

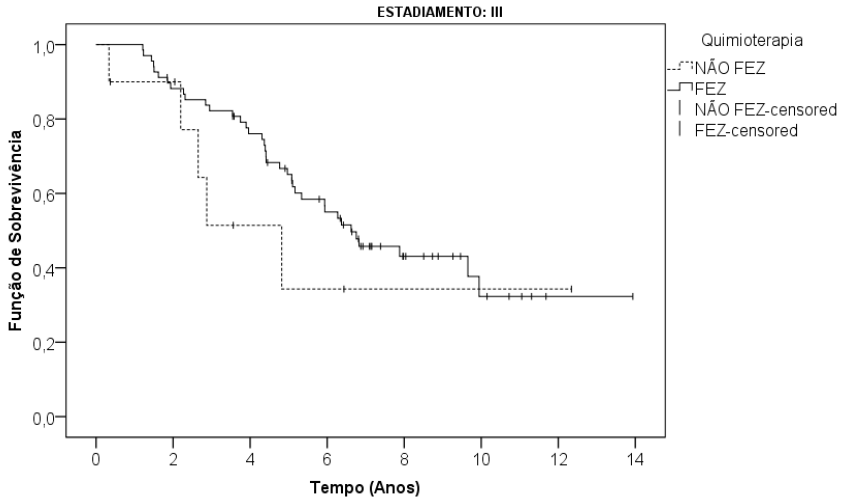

Figura 4.13: Estimativa da função de sobrevivência no estadio III consoante tratamento com quimioterapia.

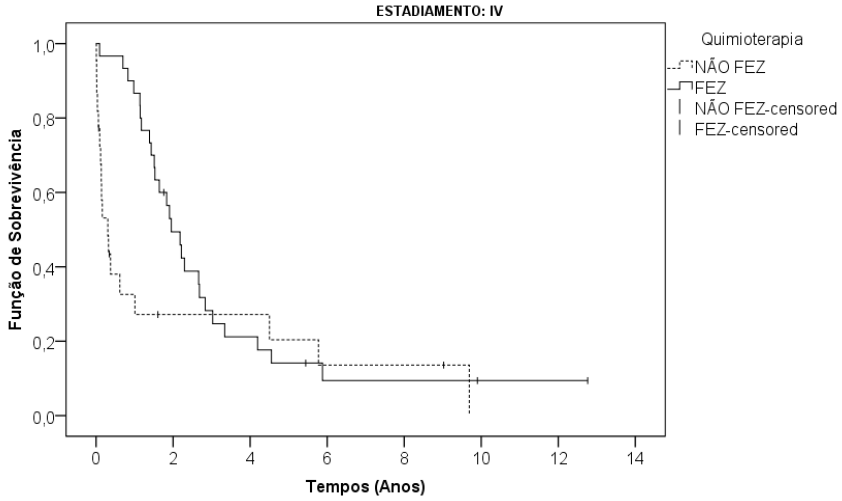

Figura 4.14: Estimativa da função de sobrevivência no estadio IV consoante tratamento com quimioterapia.

vos (55, 9%), enquanto que a maioria das mulheres que foram submetidas ao tratamento tinham gânglios positivos (75, 7%). Uma vez que a presença de metastização nos gânglios indica um maior grau de severidade da doença [28], os dados levam-nos a concluir que a causa do pior prognóstico das mulhe-

res que fazem quimioterapia é a grande proporção de mulheres com gânglios metastizados.

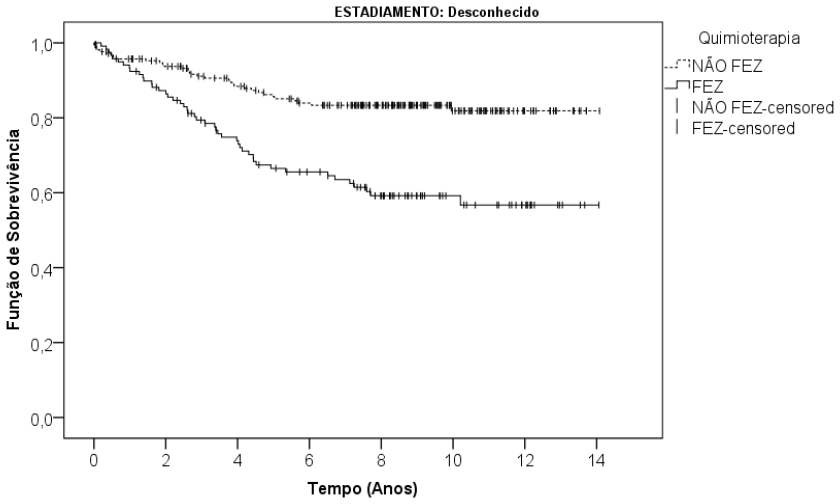

Figura 4.15: Estimativa da função de sobrevivência com estadio desconhecido consoante tratamento com quimioterapia.

Tabela 4.8: Distribuição dos indivíduos por tratamento com quimioterapia e existência de gânglios metastizados.

		N0		N1		Total
Quimioterapia	Não fez	19	55,9%	15	44,1%	34
	Fez	35	24,3%	109	75,7%	144
	Total	54	30,3%	124	69,7%	178

Na Tabela 4.9 observa-se que no caso de ausência de gânglios metastizados apenas morreu um dos 19 indivíduos que não fez quimioterapia. Apesar do número reduzido de indivíduos neste grupo, verifica-se um melhor prognóstico para quem não faz quimioterapia (Figura 4.16). Mesmo assim, no pior dos casos, a probabilidade de sobreviver para além dos 14 anos é de $72,8\%$, enquanto que quando existe presença de metastização dos gânglios, essa probabilidade já baixa para $59,5\%$ e não é relevante o facto de fazer ou não a quimioterapia.

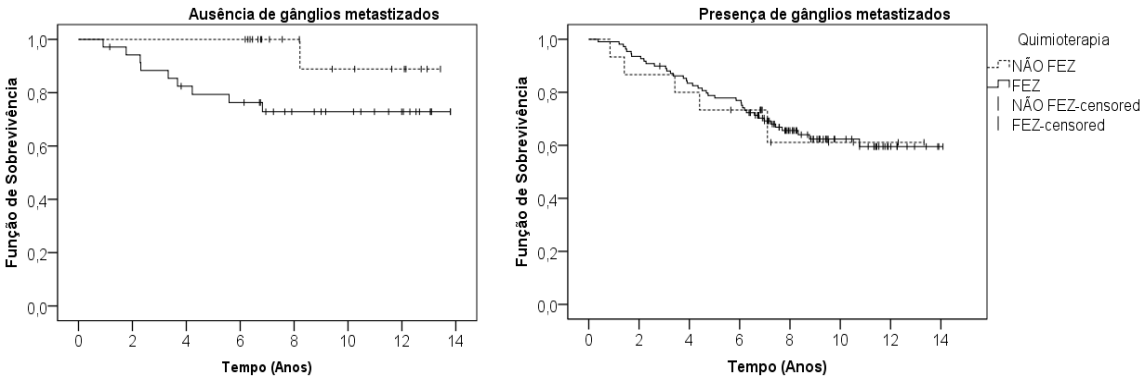

Figura 4.16: Ação dos gânglios linfáticos na sobrevivência dos indivíduos no estadio II com e sem quimioterapia.

No sentido de se perceber melhor as caraterísticas deste estadio, aprofundou-se a análise utilizando a regressão de Cox com a covariável gânglios. Esta covariável é relevante para o prognóstico e a categoria N0 está associada a um tempo de vida mais longo, donde a potencial vantagem de não fazer quimioterapia deve-se na verdade ao favorecimento no que diz respeito à seleção dos indivíduos. Assim, a variável "Gânglios"é uma variável *confounder* pelo que a estimativa adequada para o efeito da quimioterapia no tratamento deve ser ajustada para esta covariável (Figura 4.17).Os indivíduos em N0 têm apenas cerca de metade de risco de morte dos que estão em N1.

Tabela 4.9: Distribuição das mulheres no estadio II, com gânglios negativos por tipo de evento.

		Total N0	N.º de eventos		Censurados	
Quimioterapia	Não fez	19	1	5,3%	18	94,7%
	Fez	35	9	25,7%	26	74,3%
	Total	54	10	18,5%	44	81,5%

4.3 Construção de modelos

Na Análise de Sobrevivência relacionada com a medicina, o registo de informação diversificada para cada indivíduo (traduzida por covariáveis) per-

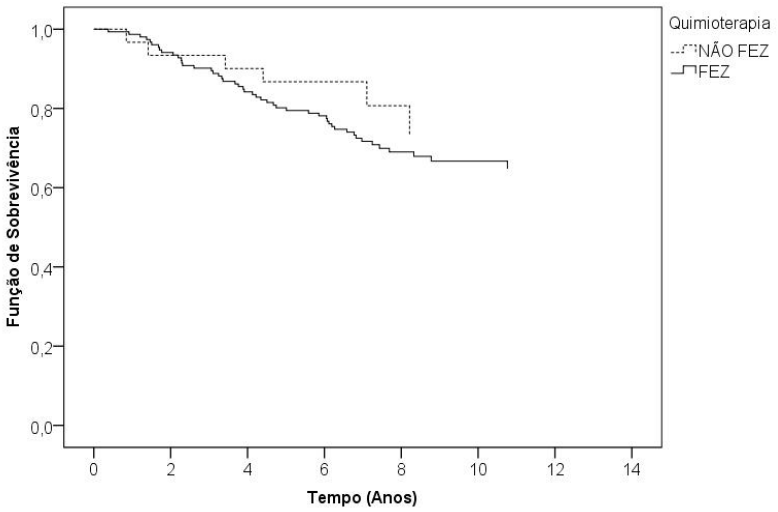

Figura 4.17: Curva de Cox ajustada para a covariável gânglios.

mite estudar de uma forma mais detalhada o tempo de vida dos indivíduos, ou seja, o conhecimento dos valores das covariáveis é um bom auxiliar para o prognóstico.

Para que esta informação possa ser quantificada, são construídos modelos, paramétricos ou não paramétricos. A escolha do modelo mais adequado pode ser feita de várias formas. Uma delas consiste em analisar as variações dos valores de $-2log\widehat{L}$, em que $\widehat{L}$ representa a função de verosimilhança maximizada, e será essa forma que será considerada neste estudo.

Nesta análise, as distribuições paramétricas escolhidas para os tempos de vida dos indivíduos suscetíveis $S_d(t, y)$ foram a distribuição log-logística e a distribuição de Chen.

4.3.1 Modelo de Cox

O modelo de regressão de Cox permite obter uma estimativa do efeito do tratamento, ajustada pelas covariáveis prognóstico. Este ajustamento é útil, especialmente em situações em que o desiquilíbrio entre os grupos relativamente aos fatores prognósticos relevantes é habitual. Neste caso, a estimativa ajustada do efeito do tratamento pode ser consideravelmente diferente da estimativa não ajustada.

Embora aos três anos, sensivelmente, se verifique que a curva da função de sobrevivência dos indivíduos no estadio IV que fizeram quimioterapia se cruza com a correspondente dos que não fizeram esse tratamento (Figura

4.14), não será dada relevância a essa situação na escolha do modelo, uma vez que são poucos os casos (doze) que restam após esse instante.

As curvas das estimativas da função de sobrevivência dos indivíduos no estadio III para cada situação em termos de tratamento com quimioterapia também se cruzam (Figura 4.13). No entanto, a representação das mesmas através do gráfico log-log (Figura 4.18) mostra que as curvas não se cruzam (este gráfico é uma forma empírica de verificar que se pode aplicar o modelo de riscos proporcionais).

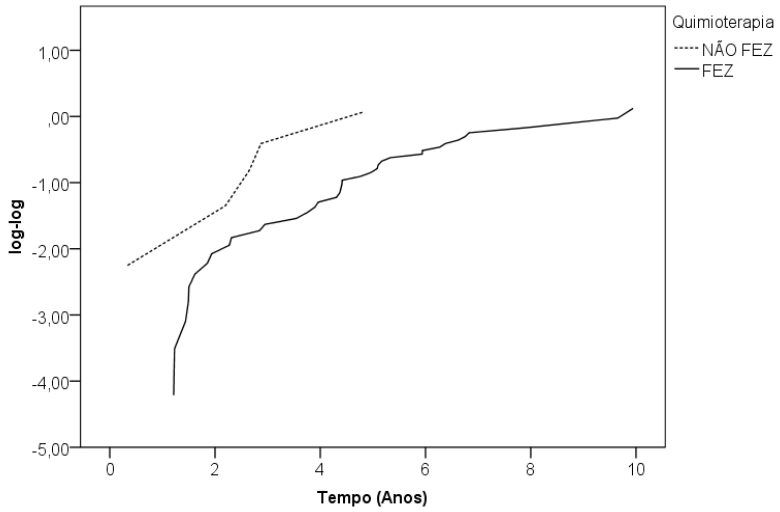

Figura 4.18: Gráfico da função $\log\Big[-\log \widehat{S}(t)\Big]$ dos indivíduos no estadio III de acordo com a variável quimioterapia.

Para as variáveis estadiamento e grupo etário criaram-se variáveis *dummy* como se apresenta nas Tabelas 4.10 e 4.11. Note-se que, apesar de ter sido constatado que a variável "Gânglios"no estadio II era importante para o modelo, esta não será aqui considerada uma vez que em cerca de 30% dos casos esta informação está omissa, o que retiraria representatividade ao estadio II.

Para construir o modelo de regressão de Cox, foi utilizado o método proposto por Collet [13]. Deste modo, para a construção do modelo temos em primeiro lugar a determinação das variáveis que, por si só, são relevantes, ou seja, significativas para $\alpha = 0,10$.

Na Tabela 4.12 observa-se que não existe evidência suficiente para afirmar que a idade seja importante para o modelo. No entanto, o grupo etário dos 50 aos 59 anos levou a uma diminuição relevante do valor de $-2logL$, pelo que a respetiva variável *dummy* será considerada na segunda parte do

Tabela 4.10: Variáveis *dummy* para o estadiamento.

	Variáveis *dummy*			
Estadiamento	E2	E3	E4	Edesc
0 ou I	0	0	0	0
II	1	0	0	0
III	0	1	0	0
IV	0	0	1	0
Desconhecido	0	0	0	1

Tabela 4.11: Variáveis *dummy* para os grupos etários.

	Variáveis *dummy*		
Grupo Etário	GE40a49	GE50a59	GE60emais
$<$ 40 anos	0	0	0
40 a 49 anos	1	0	0
50 a 59 anos	0	1	0
$\geq$ 60 anos	0	0	1

procedimento. A inclusão do estadio II não se mostrou estatisticamente significativo, mas dada a importância desta covariável a nível clínico, esta será tomada em consideração no modelo. As restantes covariáveis revelaram-se importantes. Assim o modelo será construido a partir das variáveis relativas ao estadiamento, cirurgia, radioterapia, quimioterapia, hormonoterapia e grupo etário dos 50 aos 59 anos.

O segundo passso consiste em retirar uma covariável de cada vez ao modelo resultante da escolha anterior e determinar se a exclusão dessa covariável leva ao aumento significativo do valor de $-2logL$, ou seja, se a sua omissão prejudica o modelo. Na Tabela 4.13 está o resumo dos resultados obtidos.

Nesta segunda etapa, todas as covariáveis incluídas no modelo revelaram ser importantes para o mesmo, à exceção da radioterapia, uma vez que a sua exclusão não levou ao aumento significativo do valor de $-2logL$ ($p = 0,104$). Mesmo assim, ainda se construiu o modelo com esta covariável mas como a estimativa apresentava um desvio padrão muito elevado, originou um intervalo de confiança que incluía o valor 1, o que confirmou a decisão de não incluir esta covariável no modelo.

Tabela 4.12: Determinação das variáveis para o modelo - 1^a etapa.

Covariáveis no modelo	-2logL	p-value
Sem covariáveis	3251,017	
Idade	3249,967	0,305
GE40a49	3249,159	0,173
GE50a59	3247,523	0,062
GE60emais	3251,012	0,944
E2	3248,678	0,126
E3	3231,434	<0,001
E4	3139,633	<0,001
Edesc	3243,611	0,006
Cirurgia	3144,863	<0,001
Radioterapia	3242,738	0,004
Quimioterapia	3229,287	<0,001
Hormonoterapia	3237,250	<0,001

Tabela 4.13: Confirmação da importância das covariáveis - 2^a etapa.

Covariáveis no modelo	Covariável excluída	-2logL	p-value
GE50a59; E2; E3; E4; Edesc; Cirurgia; Radioterapia; Quimioterapia; Hormonoterapia	-	3047,531	
E2; E3; E4; Edesc; Cirurgia; Radioterapia; Quimioterapia; Hormonoterapia	GE50a59	3053,276	0,017
GE50a59; E3; E4; Edesc; Cirurgia; Radioterapia; Quimioterapia; Hormonoterapia	E2	3056,063	0,003
GE50a59; E2; E4; Edesc; Cirurgia; Radioterapia; Quimioterapia; Hormonoterapia	E3	3073,120	<0,001
GE50a59; E2; E3; Edesc; Cirurgia; Radioterapia; Quimioterapia; Hormonoterapia	E4	3090,545	<0,001
GE50a59; E2; E3; E4; Cirurgia; Radioterapia; Quimioterapia; Hormonoterapia	Edesc	3052,862	0,021
GE50a59; E2; E3; E4; Edesc; Radioterapia; Quimioterapia; Hormonoterapia	Cirurgia	3077,743	<0,001
GE50a59; E2; E3; E4; Edesc; Cirurgia; Quimioterapia; Hormonoterapia	Radioterapia	3050,181	0,104
GE50a59; E2; E3; E4; Edesc; Cirurgia; Radioterapia; Hormonoterapia	Quimioterapia	3052,964	0,019
GE50a59; E2; E3; E4; Edesc; Cirurgia; Radioterapia; Quimioterapia	Hormonoterapia	3051,830	0,038

A fase seguinte consiste em incluir no modelo obtido na etapa anterior as covariáveis que não foram consideradas como relevantes na etapa inicial

de modo a verificar se a sua inclusão leva a uma diminuição significativa do valor de $-2logL$, i.e., se o modelo melhora.

Tabela 4.14: Inclusão das variáveis rejeitadas inicialmente - 3^a etapa.

Covariáveis no modelo	Covariável incluída	-2logL	p-value
GE50a59; E2; E3; E4; Edesc; Cirurgia; Quimioterapia; Hormonoterapia	-	3050,181	
GE40a49; GE50a59; E2; E3; E4; Edesc; Cirurgia; Quimioterapia; Hormonoterapia	GE40a49	3050,038	0,705
GE50a59; GE60emais; E2; E3; E4; Edesc; Cirurgia; Quimioterapia; Hormonoterapia	GE60emais	3049,732	0,503

Pode-se confirmar na Tabela 4.14 que as variáveis *dummy* dos grupos etários dos 40 a 49 anos e dos 60 e mais anos não são importantes para o modelo. A variável idade, que já tinha sido rejeitada na 1^a etapa, não foi agora considerada porque, obviamente, não é independente do grupo etário. O modelo será composto pelas covariáveis obtidas na segunda etapa, apresentadas na Figura 4.19.

Tendo em conta o procedimento anterior, nenhuma covariável pode ser retirada do modelo sem o prejudicar nem é necessário incluir mais alguma. Temos então o seguinte modelo de Cox

$$\begin{aligned}\widehat{h}(t,\mathbf{z}) = \ & \widehat{h}_0(t)\exp(0,321*\text{GE50a59}+0,818*\text{E2}+1,512*\text{E3}+2,217*\text{E4}\\ & +0,686*\text{Edesc}-1,198*\text{Cirurgia}+0,311*\text{Quimioterapia}\\ & -0,261*\text{Hormonoterapia})\end{aligned} \tag{4.1}$$

A partir do modelo (4.1) temos, por exemplo, para uma mulher com 60 anos, diagnosticada com cancro da mama no estadio III, e tratada com cirurgia e quimioterapia o vetor de covariáveis $z = (0,0,1,0,0,1,1,0)$. Então, pela equação 1.6, a equação 4.2 é a expressão da função de sobrevivência para este indivíduo.

$$\widehat{S}(t,\mathbf{z}) = \widehat{S}_0(t)^{\exp(0,321*0+0,818*0+1,512*1+2,217*0+0,686*0-1,198*1+0,311*1-0,261*0)} \tag{4.2}$$

$$\iff \widehat{S}(t,\mathbf{z}) = \widehat{S}_0(t)^{\exp(1,512-1,198+0,311)}$$

$$\iff \widehat{S}(t, \mathbf{z}) = \widehat{S}_0(t)^{1,868}$$

Verifica-se pela Figura 4.19, no valor de $Exp(B)$ que, dados dois grupos de mulheres na mesma situação em cada covariável, exceto na faixa etária, as mulheres que estão na faixa etária dos 50 aos 59 anos têm pior prognóstico do que as restantes, pois tem mais $37,8\%$ de risco de morte.

Block 0: Beginning Block

Omnibus Tests of Model Coefficients

-2 Log Likelihood
3251,017

Block 1: Method = Enter

Omnibus Tests of Model Coefficients[a]

	Overall (score)			Change From Previous Step			Change From Previous Block		
-2 Log Likelihood	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
3050,181	387,652	8	,000	200,836	8	,000	200,836	8	,000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

							95,0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
GE50a59	,321	,140	5,261	1	,022	1,378	1,048	1,813
E2	,818	,286	8,151	1	,004	2,266	1,292	3,972
E3	1,512	,310	23,790	1	,000	4,537	2,471	8,330
E4	2,217	,345	41,347	1	,000	9,183	4,672	18,052
Edesc	,686	,287	5,721	1	,017	1,985	1,132	3,481
Cirurgia	-1,198	,199	36,385	1	,000	,302	,204	,445
Quimioterapia	,311	,153	4,108	1	,043	1,364	1,010	1,842
Hormonoterapia	-,261	,137	3,645	1	,056	,770	,589	1,007

Figura 4.19: Covariáveis significativas para o modelo.

Comprova-se também a influência do estadio da doença na probabilidade de sobrevivência. Estando em igualdade de circunstâncias em todas as outras covariáveis, e comparando com as mulheres no estadio 0 ou I, as mulheres no estadio II têm uma taxa de morte $126,6\%$ superior; as mulheres no estadio III têm o seu prognóstico agravado, sendo a sua taxa de morte $353,7\%$ superior; e, na pior das situações, no estadio IV, a taxa de morte é extremamente elevada pois é $818,3\%$ superior à dos primeiros estadios (0 ou I). Em relação às mulheres cujo estadiamento é desconhecido, estas têm mais $98,5\%$ de risco de de morte. Este valor é semelhante ao obtido para as mulheres no estadio II, o que leva a crer que, dos estadios desconhecidos, muitos sejam na realidade do estadio II.

Relativamente aos tratamentos, uma mulher que é submetida a cirurgia tem apenas $30,2\%$ do risco de morte de uma mulher que não fez este tratamento. Tal resultado não suscita dúvidas uma vez que a cirurgia só é usada enquanto tem utilidade terapêutica, o que não acontece, por exemplo, na fase terminal da doença.

O tratamento com quimioterapia resulta numa taxa de morte $36,4\%$ superior à das mulheres que não fizeram esta terapêutica. No entanto, como referido anteriormente, esta situação está relacionada com a especificidade do estadio II e eventualmente do estadio desconhecido (já que se suspeita ser semelhante ao estadio II)

Embora o tratamento hormonal pareça indicar um ligeiro favorecimento para as mulheres com os recetores positivos, não é possível obter conclusões acerca do mesmo, dado que o intervalo de confiança associado ao risco contém o valor 1. Note-se que, quando é efetuada a cirurgia, uma das análises anátomo-patológicas realizadas por rotina é a avaliação dos recetores de estrogénio com o objetivo de prever a resposta do doente a um eventual tratamento hormonal. Quando os níveis são baixos, a resposta a um tratamento endócrino é incerto ([16]) ao invés da grande probabilidade de resposta ao tratamento hormonal e consequente aumento da sobrevivência quando os recetores de estrogénio são positivos ([17]).

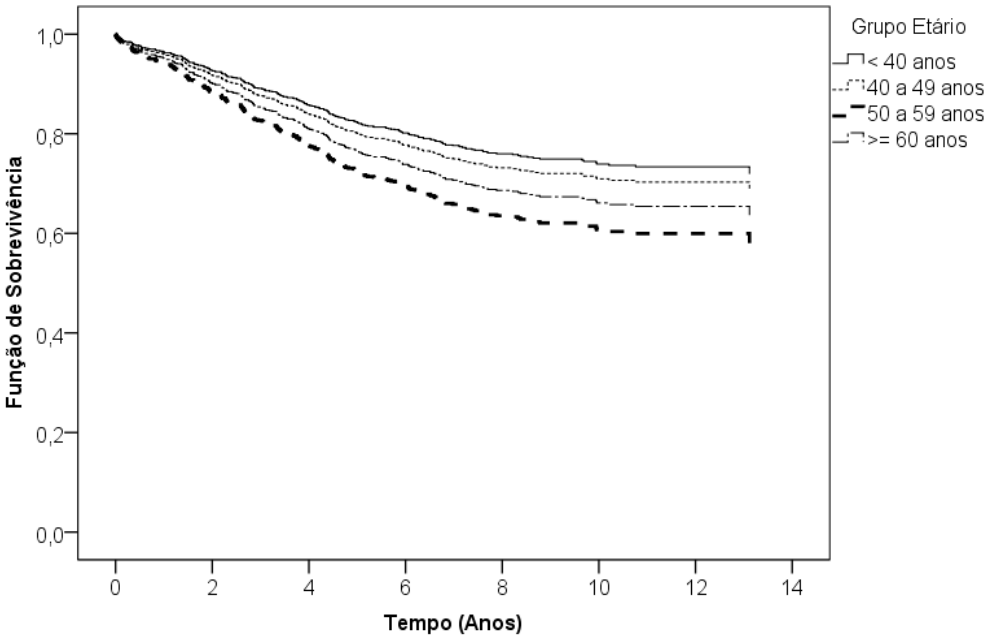

Figura 4.20: Estimativa da função de sobrevivência para cada grupo etário.

O facto de apenas o grupo etário 50 a 59 anos ter sido retido no modelo suscita alguma curiosidade. Visualizando a representação das funções de sobrevivência para cada grupo etário (Figura 4.20), verifica-se que este é o grupo com pior prognóstico, mesmo em relação aos indivíduos mais idosos, o que pode ser a razão da importância deste grupo etário. Poderá existir alguma associação entre a relevância do grupo etário em questão e a implementação do Rastreio do Cancro da Mama na Ilha da Madeira, em 1999, direcionado para as mulheres entre os 45 e os 69 anos, mas a informação disponível não permitiu chegar a alguma conclusão.

Depois de determinado o modelo com o ajustamento para as covariáveis de interesse, a representação das curvas das funções de sobrevivência das mulheres que fizeram quimioterapia e das que não fizeram, inicialmente indicada na Figura 4.9, encontra-se na Figura 4.21. Com este ajustamento, já não existe cruzamento das curvas.

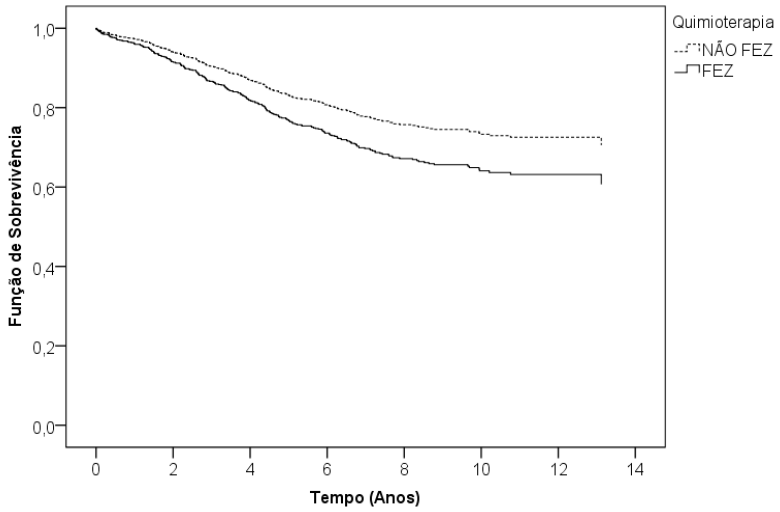

Figura 4.21: Estimativa da função de sobrevivência, usando o modelo de Cox, para os indivíduos que fizeram ou não quimioterapia.

4.3.2 Estimação do modelo de cura baseado na distribuição log-logística

Como já foi referido neste trabalho, o *follow-up* dos indivíduos diagnosticados com cancro da mama necessita de ser longo, para se poder obter conclusões fiáveis sobre a cura desta doença (ver artigo [45]). Por outro lado, Yu *et al.*, [49], referem a necessidade de o *follow-up* ser maior do que o valor da mediana

do tempo de vida dos indivíduos suscetíveis, como forma de assegurar que a maioria dos acontecimentos possam ser observados antes do fim do período de observação. Ora, apesar de o tempo de *follow-up* deste estudo ser de 14 anos, para qualquer indivíduo, só existe a garantia de que ele tenha sido seguido durante, pelo menos, 5 anos. Como o estadio IV é o único em que a mediana do tempo de vida é inferior a 5 anos ($m = 1,5$), o modelo de cura será aplicado apenas para este caso.

Os dados utilizados foram o tempo de vida, em anos, e o indicador de censura, os quais foram compilados em R na base de dados ModCura (ver Anexo A). A programação foi feita em R, baseada em [1], e as bibliotecas específicas utilizadas foram *survival* e *KMsurv*.

Começou-se por obter a estimativa inicial da proporção de curados com base no valor da estimativa de Kaplan-Meier da função de sobrevivência ($\widehat{p}^{(0)} = 0,06863$). De seguida, determinou-se os valores iniciais dos parâmetros da distribuição log-logística utilizando os dados até ao maior tempo de vida observado ($\widehat{\alpha}^{(0)} = 0,97830$ e $\widehat{\lambda}^{(0)} = 0,90397$). Através da utilização do algoritmo EM, foram obtidas as estimativas dos parâmetros do modelo:

$\widehat{p} = 0,00000$
$\widehat{\alpha} = 0,92202$
$\widehat{\lambda} = 0,82329$

Este modelo revela uma taxa de cura nula, o que não está de acordo com a realidade, logo não é um modelo adequado.

4.3.3 Estimação do modelo de cura baseado na distribuição de Chen

Para obter este modelo utilizou-se um procedimento semelhante ao anterior. Apenas se alterou a distribuição para o tempo de vida dos indivíduos suscetíveis pois passou-se a usar a distribuição de Chen e a forma de obter as estimativas iniciais dos parâmetros da distribuição (ver Anexo B).

Assim a estimativa inicial do parâmetro q foi $q^{(0)} = 0,93137$, e as estimativas iniciais de β e λ foram $\beta^{(0)} = 0,5$ e $\lambda^{(0)} = 0,280615$.

Por fim, foram obtidas as seguintes estimativas dos parâmetros do modelo:
$\widehat{q} = 0,9261983$
$\widehat{\beta} = 0,4543725$
$\widehat{\lambda} = 0,3339661$

Com este modelo obteve-se uma taxa de cura de $7,4\%$ o que não só está mais de acordo com a realidade médica, como mais próximo da estimativa

inicial de Kaplan-Meier. Por estas razões, este modelo é preferível ao anterior. A Figura 4.22 reforça esta afirmação.

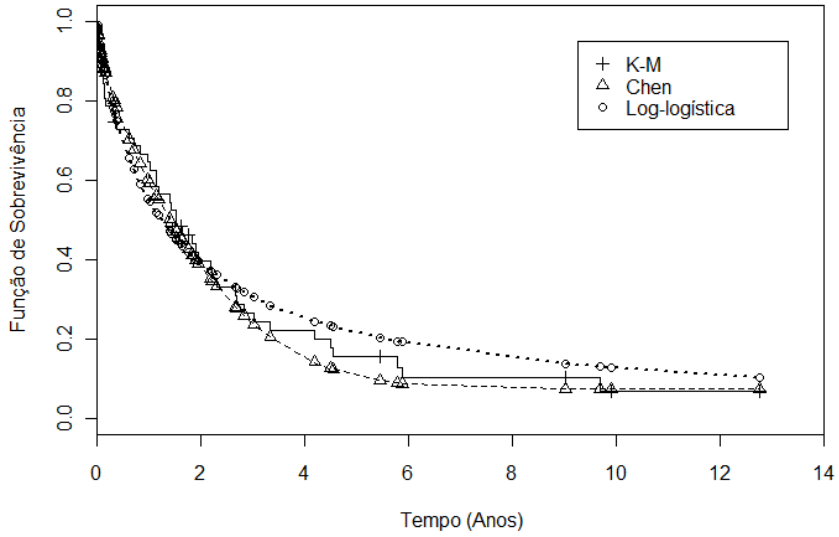

Figura 4.22: Curvas de sobrevivência correspondentes à estimativa de Kaplan-Meier e ao modelo de cura (com a distribuição log-logística e com a distribuição de Chen).

4.4 Comentários finais

Uma das partes mais gratificantes de um trabalho científico é a apresentação dos resultados e das conclusões. Como este é um trabalho com aplicação ao cancro da mama, existiram algumas limitações, pelo que ainda será possível fazer algo mais.

4.4.1 Resultados e conclusões

Neste estudo, as mulheres diagnosticadas com cancro da mama, do tipo não inflamatório, têm idades compreendidas entre os 25 e os 93 anos, sendo a média de 60 ± 14 anos. Em se tratando de estadios conhecidos, a média das idades baixa para 57 ± 13 anos. Categorizando as idades em quatro grupos etários, obteve-se a seguinte distribuição: <40 anos - $6,5\%$; 40 a 49 anos - $18,7\%$; 50 a 59 anos - $23,9\%$; >=60 anos - $50,9\%$ (Tabela 4.7).

Na Tabela 4.6 observa-se que $13,9\%$ das doentes estavam nos estadios 0 ou I, $30,4\%$ estavam no estadio II, $9,4\%$ estavam no estadio III e $6,2\%$ estavam no estadio IV. De salientar a existência de um elevado número de indivíduos com estadio desconhecido $(40,1\%)$ devido, em especial, à falta de informação sobre a existência de metástases à distância.

Verificou-se que para a maioria das mulheres mais jovens (com menos de 40 anos), o diagnóstico ocorreu quando a doença estava principalmente nos estadios iniciais 0 ou I $(25,9\%)$ ou estadio II $(35,2\%)$. Já a situação mais frequente nas mulheres mais idosas, é o desconhecimento do estadio $(49,3\%)$. Nos grupos etários intermédios (40 a 49 anos; 50 a 59 anos), as maiores proporções encontram-se no estadio II $(32,1\%;\ 36,2\%)$ e no estadio desconhecido $(28,8\%;\ 33,7\%)$.

Analisando o estadiamento da doença em relação aos tratamentos, constata-se que todos os indivíduos nos estadios 0 ou I foram submetidos a cirurgia, a maioria fez radioterapia $(59,5\%)$ e hormonoterapia $(55,2\%)$ mas apenas $44,0\%$ foram submetidos a quimioterapia. Desses, a maioria $(88,2\%)$ fez quimioterapia do tipo adjuvante. Todas as mulheres no estadio II realizaram cirurgia, $44,7\%$ fizeram radioterapia, $53,0\%$ fizeram hormonoterapia e $74,3\%$ submeteram-se a quimioterapia, sendo que, destas últimas, $83,0\%$ foram do tipo adjuvante. No caso do estadio III, a percentagem de mulheres submetidas a cirurgia descresce para $89,7\%$, ocorrendo o mesmo em relação à radioterapia e à hormonoterapia (respetivamente, $43,6\%$ e $47,4\%$). No que concerne à quimioterapia, $87,2\%$ das doentes submeteram-se a este tratamento, sendo mais frequente a combinação da adjuvante e da neoadjuvante $(44,8\%)$. No último estadio, há uma grande redução na aplicação dos vários tratamentos, pois apenas $32,7\%$ das mulheres submeteram-se a cirurgia, $11,5\%$ a radioterapia e $9,6\%$ a hormonoterapia. Ainda assim, a maior parte realiza quimioterapia $(57,7\%)$, com $60,0\%$ destas mulheres a não serem submetidas a cirurgia.

Verificou-se que, ao fim de pelo menos 5 anos, a maioria $(55,6\%)$ das mulheres diagnosticadas com cancro da mama estavam vivas e que $7,3\%$ morreram por outra causa (Figura 4.3). Aos 14 anos de *follow-up*, a sobrevida estimada para o total dos casos ainda se encontra acima dos 60% (Figura 4.4). Verificou-se claramente que quanto menor o estadio melhor o prognóstico: a sobrevida estimada para os estadios 0 ou I está acima dos 80%, enquanto no estadio IV esta encontra-se abaixo dos 10% (Figura 4.5).

A probabilidade dos indivíduos estarem vivos ao fim de 14 anos após o diagnóstico é maior se estes foram submetidos a cirurgia - $68,7\%$ de probabilidade de sobrevivência, do que se não forem - $17,1\%$ (Figura 4.6). Embora

com diferenças menos acentuadas, a realização de radioterapia e de hormonoterapia também resulta numa maior taxa de sobrevida (70, 6% e 67, 6%, respetivamente; Figuras 4.7 e 4.8).

O prognóstico das mulheres submetidas a quimioterapia é pior. Embora nos primeiros dois anos, a quimioterapia esteja associada a uma melhor taxa de sobrevivência, esta situação inverte-se a partir desse momento e, ao fim de 14 anos, a sobrevida estimada para os doentes que se submeteram a este tratamento é inferior a 60% (55, 9%), enquanto os que não o realizaram têm uma percentagem de sobrevivência esperada de cerca de 80% (75, 8%). A justificação deste facto envolve diversos fatores. Sabendo que este tratamento pode ser aplicado com diversos objetivos, como sejam a diminuição do tamanho do tumor para preparação do organismo para aplicar outro tratamento (nomeadamente, a cirurgia), a eliminação de algumas potenciais células cancerígenas que tenham restado mesmo após a aplicação de outros tratamentos ou com um fim paliativo (para atenuação dos sintomas na fase terminal da doença), seria de esperar que a curva da função de sobrevivência refletisse estas diferentes realidades. Eventualmente, a aplicação deste tratamento às mulheres no último estadio poderá garantir um prolongamento de vida que durará cerca de dois anos. Esta poderá ser a justificação da maior probabilidade de sobrevivência nos primeiros dois anos nos doentes que se submetem a esta terapêutica. A partir daí, o elevado número de mortes associado principalmente ao estadio IV contribuirá para a acentuada diminuição da probabilidade de sobrevivência. Paralelamente, a heterogeneidade do estadio II, o grupo com mais casos entre aqueles que têm estadio conhecido (50, 7%), no que se refere à metastização dos gânglios, e o peso deste grupo no total dos registos, leva a admitir que o pior prognóstico das mulheres deste grupo que fizeram quimioterapia tenha influenciado o prognóstico geral. A mesma influência suspeita-se ser exercida pelos doentes no estadio desconhecido, uma vez que é o grupo com mais peso no total de casos (40, 1%). Dados os resultados obtidos, o envolvimento do número de gânglios linfáticos metastizados deverá merecer estudo mais aprofundado.

Após a análise individual das variáveis em estudo, foram determinadas as covariáveis significativas através do modelo de regressão de Cox. Tendo como referência os estadios iniciais, 0 ou I, é fator de risco ter a doença diagnosticada em estadios mais avançados: no estadio II a taxa de morte é 126, 6% superior, no estadio III é 353, 7% superior, no estadio IV é 818, 3% superior, e no estadio desconhecido é 98, 5% superior.

Os fatores protetores foram a cirurgia (apenas 30, 2% do risco de morte de quem não fez este tratamento) e a hormonoterapia, embora esta última sem significado estatístico. No sentido contrário, surgem a idade (embora

apenas na faixa etária dos 50 aos 59 anos, com um risco de morte acrescido de $37,8\%$), o estadio (quanto mais elevado pior o prognóstico) e a quimioterapia (mais $36,4\%$ de risco de morte). No entanto, esta última não deve ser analisada por si só, uma vez que, como já foi referido anteriormente, está intimamente ligada com o estadio e, mesmo dentro do estadio, pode ter variações como ocorre no estadio II. A variação de acordo com o estadio também se reflete no tipo de quimioterapia a realizar.

Para as mulheres no estadio IV, estimou-se o modelo de cura (sem covariáveis) baseado em duas distribuições distintas, com o objetivo de obter uma indicação do valor da taxa de cura dos doentes neste estadio. Verificou-se, para este efeito, que o modelo mais adequado foi o que utilizou a distribuição de Chen para os indivíduos suscetíveis e obteve-se assim uma estimativa de taxa de cura de $7,4\%$.

A aplicação do modelo de cura aos dados deste estudo foi restringida ao estadio mais severo devido à limitação do tempo mínimo de *follow-up* e às caraterísticas da doença. Se o tempo de *follow-up* tivesse sido um pouco superior, também seria possível aplicar ao estadio III. Quanto aos restantes estadios, como não é viável calcular a mediana do tempo de vida (a taxa de sobrevivência é sempre superior a 50%), já não é adequado aplicar o modelo. No entanto, sendo o cancro da mama considerado, clinicamente, uma doença de evolução lenta, para esta última situação, não é tão relevante saber se o indivíduo atingiu a cura uma vez que o seu tempo de vida será semelhante ao que teria se não tivesse tido a doença.

4.4.2 Limitações do estudo e trabalho futuro

Ao longo deste estudo surgiram as seguintes limitações:

- Muitos casos com estadio desconhecido;
- Vários casos em que a classificação dos gânglios é desconhecida;
- Ausência de informação sobre os recetores hormonais - importante por estar diretamente relacionada com a decisão de prescrição da terapêutica hormonal;
- Pouco tempo de *follow-up* - apenas o estadio IV tinha os requisitos mínimos para a aplicação do modelo de cura.

Futuramente, seria interessante continuar a seguir estes indivíduos, pois assim poderíamos obter estimativas mais precisas e alargar a aplicação do

modelo de cura a outros estadios, em particular, ao estadio III. Como a mediana do tempo de vida dos doentes no estadio III é de $6,4$ anos, um *follow-up* mínimo de 7 anos já seria suficiente.

Bibliografia

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

[2] Abreu, A.M. (2010/2011) - Apontamentos das aulas de "Complementos de Estatística". Universidade da Madeira.

[3] Abreu, A. M. e Rocha, C. S. (2006). Um novo modelo de cura paramétrico. Em *Ciência Estatística (Luísa Canto e Castro, et al. Editores)*, Edições SPE, pp. 151–162.

[4] Aljawadi, B. A. I., Bakar, M. R. A. e Ibrahim, N. A. (2011). Nonparametric Estimation of Cure Fraction Using Right Censored Data. *American Journal of Scientific Research*, Vol. 14, pp. 79–87.

[5] Aljawadi, B. A. I., Bakar, M. R. A., Ibrahim, N. A. *et al.* (2011). Parametric Estimation of the Cure Fraction Based on BCH Model Using Left-Censored Data with Covariates. *Modern Applied Science*, Vol. 5, 3, pp. 103–110.

[6] American Joint Committe on Cancer (2002). AJCC *Cancer Staging Handbook*. Springer. New York.

[7] Berkson, J. e Gage, R. (1952). Survival cure for cancer patients following treatment. *American Statistical Association Journal.* , Vol. 47, pp. 501–515.

[8] Boag, J. W. (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society.* Series B, Vol. 11, pp. 15–44.

[9] Broët, P., De Rycke, Y., Tubert-Bitter, P. *et al.* (2001). A semiparametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics*, Vol. 57, pp. 844–852.

[10] Chen, M.-H., Ibrahim, J. G. e Sinha, D. (1999). A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of the American Statistical Association.* , Vol. 94, 447, pp. 909–919.

[11] Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function, *Statistics & Probability Letters*, Vol. 49, 2, pp. 155–161.

[12] Cobre, J. (2010). *Modelos de sobrevivência na presença de eventos recorrentes e de longa duração*. Tese de Doutoramento. Universidade Federal de São Carlos. São Carlos, Brasil.

[13] Collet, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

[14] Cooner, F., Banerjee, S., Carlin, B. P. *et al.* (2007). Flexible Cure Rate Modelling Under Latent Activation Schemes, *Journal of the American Statistical Association*, Vol. 102, 478, pp. 560–572.

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

[16] Early Breast Cancer Trialists' Collaborative Group e Clarke (2008). Adjuvant chemotherapy in estrogen-poor breast cancer: patient-level meta-analysis of randomized trials, *Lancet*, Vol. 371, 9606, pp. 29–40.

[17] Early Breast Cancer Trialists' Collaborative Group (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomized trials, *Lancet*, Vol. 365, 9472, pp. 1687-1717.

[18] Farewell, V. T. (1977). A model for a binary variable with time-censored observations. *Biometrika*, Vol. 64, 1, pp. 43–46.

[19] Farewell, V. T. (1982). The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, Vol. 38, pp. 1041–1046.

[20] Fields, R., Jeffe, D. B., Deshpande, A. D. *et al.* (2010). Predictors of axillary lymph node involvement in women with T3 breast cancers: Analysis of 1988-2003 SEER Data. *Journal of Surgical Research*, Vol. 161, 2, pp. 183–189.

[21] Goldman, A. (1984). Survivorship analysis when cure is a possibility: a Monte Carlo study. *Statistics in Medicine*, Vol. 3, pp. 153–163.

[22] Gordon, N.H. (1990). Application of the Theory of Finite Mixtures for the Estimation of 'Cure' Rates of Treated Cancer Patients. *Statistics in Medicine*, Vol. 9, pp. 397–407.

[23] Gouveia, B. P. (2010). *Modelo de Mistura Padrão com Tempos de Vida Exponenciais Ponderados*. Dissertação. Universidade de São Carlos.

[24] Haybittle, J.L. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *British Journal of Radiology*, Vol. 32, pp. 725–733.

[25] Haybittle, J. L. (1983). Is breast cancer ever cured?. *Reviews on Endocrine-Related Cancer*, Vol. 14, pp. 13–18.

[26] Huang, L., Johnson, K. A., Mariotto, A. B. *et al.* (2010). Population-based survival-cure analysis of ER-negative breast cancer, *Breast Cancer Research and Treatment*, Vol. 123, pp. 257–264.

[27] Instituto Nacional de Estatística. Estatísticas no feminino: Ser mulher em Portugal 2001–2011. INE. 2012

[28] Jimenez-Lee, R., Ham, B., Vetto, J. *et al.* (2003). Breast cancer severity score is an innovative system for prognosis. *The American Journal of Surgery*, Vol. 186, pp. 404–408.

[29] Kim, S., Xi, Y. e Chen, M.-H. (2009) A new latent cure rate marker model for survival data. *The Annals of Applied Statistics*. Vol. 3, 3**,** pp. 1124–1146.

[30] Klebanov, L. B. e Yakovlev, A. Y. (2007). A new approach to testing for sufficient follow-up in cure-rate Analysis. *Journal of Statistical Planning and Inference*, Vol. 137, pp. 3557–3569.

[31] Kuk, A. Y. C. e Chen, C.-H. (1992). A Mixture Model Combining Logistic Regression with Proportional Hazards Regression, *Biometrika*, Vol. 79, pp. 531–541.

[32] Langlands, A. O., Pocock, S. J., Kerr, G. R. *et al.* (1979). Long Term Survival of Patients with Breast Cancer: A Study of Curability of the Disease. *Brit. Med. J.*, Vol. 2, pp. 1247–1251.

[33] Laska, E. M. e Meisner, M. J. (1992) Nonparametric Estimation and Testing in Cure Model. *Biometrics*, Vol. 48, pp. 1223–1234.

[34] Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*. Wiley. University of Waterloo, New Jersey, U.S.A.

[35] Li, Y., Tiwari, R. C. e Guha, S. (2005). Mixture Cure Survival Models with Dependent Censoring, *Harvard University Biostatistics Working Paper Series*, working paper 26.

[36] Liga Portuguesa contra o Cancro. Programa de Rastreio de Cancro da Mama da Liga Portuguesa contra o Cancro. http://www.ligacontracancro.pt/gca/index.php?id=42. Consultado em 2012/06/07.

[37] Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, Vol. 79, 4, pp. 731–739.

[38] Maller, R. A. e Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley. New York.

[39] McLachlan, G.J. e Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley, New York.

[40] Ministério da Saúde. http://www.min-saude.pt/portal/ conteudos/ enciclopedia+da+saude/ doencas/ cancro/ cancro+mama.htm. Portal da Saúde. Consultado em 2012/06/07

[41] Paes, A. T. (2007). *Uso de modelos com fração de cura na análise de dados de sobrevivência com omissão nas covariáveis*. Tese de Doutoramento. Universidade de São Paulo.

[42] Rocha, C. e Papoila, A. P. (2009). *Análise de Sobrevivência*. XVII Congresso da Sociedade Portuguesa de Estatística. SPE.

[43] Rodrigues, J., Cancho, V. G., Castro, M. *et al.* (2009). On the unification of the long-term models. *Statistics & Probability Letters*, Vol. 79, 6, pp. 753–759.

[44] Sposto, R., Sather, H.N., Baker, S.A. (1992). A comparison of tests of the difference in the proportion of patients who are cured. *Biometrics.*, Vol. 48, pp. 87–99.

[45] Tai, P., Yu, E., Cserni, G. *et al.* (2005). Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database, *BioMed Central*, 5:48.

[46] Tsodikov, A. (2001). Estimation of Survival Based on Proportional Hazards When Cure is a Possibility. *Mathematical and Computer Modelling*, Vol. 33, pp. 1227–1236.

[47] Yakovlev A. Y., Asselain, B., Bardou, V. J., *et al.* (1993) . A Simple Stochastic Model of Tumor Recurrence and Its Applications to Data on pre-menopausal Breast Cancer. Em Biometrie et Analyse de Dormees Spatio – Temporelles (Editores: Asselain, B., Boniface, M., Duby, C., *et al.*). Societé Française de Biométrie, ENSA Rennes, France, Vol. 12, pp. 66–82

[48] Yakovlev, A. e Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and TheirBiostatistical Applications*. World Scientific. Singapore.

[49] Yu, B., Tiwari, R. C., Cronin, K. A. *et al.* (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, Vol. 23, pp. 1733–1747.

[50] Yu, B. e Peng. Y. (2008). Mixture cure models for multivariate survival data. *Computational Statistics & Data Analysis*, Vol. 52, pp. 1524–1532.

[51] Zhang, J e Peng, Y. (2009). Accelerated Hazards Mixture Cure Model. *Lifetime Data Analysis*, Vol. 15, 4, pp. 455–467

[52] Zhao, X., Wu, X. e Zhou, X. (2009). A change-point model for survival data with long-term survivors. *Statistica Sinica*, Vol. 19, pp. 377–390.

Anexos

Anexo A

Procedimento para obtenção do modelo de cura baseado na distribuição log-logística

Os comandos utilizados foram os que se seguem.

```
ModCura<-data.frame(TempoAnos=c(0.008, 0.014, 0.016, 0.033, 0.060,
0.068, 0.090, 0.093, 0.118, 0.134, 0.142, 0.162, 0.307, 0.323, 0.342, 0.375,
0.616, 0.699, 0.827, 0.978, 1.008, 1.140, 1.145, 1.178, 1.389, 1.430, 1.512,
1.529, 1.605, 1.641, 1.759, 1.833, 1.907, 1.953, 2.181, 2.214, 2.296, 2.663,
2.682, 2.833, 3.027, 3.340, 4.195, 4.501, 4.553, 5.444, 5.775, 5.879, 9.025,
9.690, 9.901, 12.764), TipoEvento=c(1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
0, 1, 1, 0, 1, 0, 0))
```

No início da análise, calcula-se a estimativa inicial da proporção de suscetíveis ($1 - p = q$) através dos seguintes comandos.

```
survCMama<-Surv(ModCura$TempoAnos,ModCura$TipoEvento)
fCMama<-survfit(survCMama~1)
edit(fCMama)
```

Após estas instruções, o programa apresenta uma série de resultados, tais como a estimativa de Kaplan-Meier da função de sobrevivência, que se apresenta de seguida:

```
surv = c(0.980769230769231, 0.961538461538461, 0.942307692307692,
0.923076923076923, 0.903846153846154, 0.903846153846154,
0.884197324414716, 0.864548494983277, 0.84489966555184,
0.825250836120401, 0.805602006688963, 0.785953177257525,
```

0.766304347826087, 0.746655518394649, 0.746655518394649,
0.726475639519118, 0.706295760643587, 0.686115881768055,
0.665936002892524, 0.645756124016993, 0.625576245141462,
0.605396366265931, 0.5852164873904, 0.565036608514869,
0.544856729639338, 0.524676850763807, 0.504496971888276,
0.484317093012745, 0.484317093012745, 0.463259828099148,
0.463259828099148, 0.441199836284902, 0.419139844470657,
0.397079852656412, 0.375019860842167, 0.352959869027922,
0.330899877213677, 0.308839885399432, 0.286779893585187,
0.264719901770941, 0.242659909956696, 0.220599918142451,
0.198539926328206, 0.176479934513961, 0.154419942699716,
0.154419942699716, 0.128683285583096, 0.102946628466477,
0.102946628466477, 0.0686310856443181, 0.0686310856443181,
0.0686310856443181)

Como o maior tempo de vida observado é 9.690, temos que
fCMama(9.690)=0.0686310856443181, logo
$q^{(0)} = 1 - 0.0686310856443181 = 0.9313689143556819$

Em relação aos valores iniciais dos parâmetros da distribuição log-logística, fez-se o ajustamento mas com os dados apenas até $t_{(r)}$, isto é, as observações são

tCMamat<-c(0.008, 0.014, 0.016, 0.033, 0.060, 0.068, 0.090, 0.093, 0.118, 0.134, 0.142, 0.162, 0.307, 0.323, 0.342, 0.375, 0.616, 0.699, 0.827, 0.978, 1.008, 1.140, 1.145, 1.178, 1.389, 1.430, 1.512, 1.529, 1.605, 1.641, 1.759, 1.833, 1.907, 1.953, 2.181, 2.214, 2.296, 2.663, 2.682, 2.833, 3.027, 3.340, 4.195, 4.501, 4.553, 5.444, 5.775, 5.879, 9.025, 9.690)
e

deltaCMamat<-c(1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1)

Assim, tem-se:
survreg(Surv(tCMamat,deltaCMamat)˜1,dist="loglogistic")
O programa R devolve o seguinte resultado
Call:
survreg(formula = Surv(tCMamat, deltaCMamat) ˜1, dist = "loglogistic")
Coefficients:
(Intercept)
0.1031932
Scale= 1.022179

Loglik(model)= -79.1 Loglik(intercept only)= -79.1
n= 50

Note-se que *alpha*=1/Scale e que *lambda*=exp(-Intercept/Scale), donde $alpha^{(0)}$=0.9783022 e $lambda^{(0)}$ =0.9039745

No que se segue, optou-se por dividir os dados iniciais em dois grupos, visto que as expressões das derivadas parciais necessárias ao algoritmo a desenvolver incluem somas que envolvem separadamente"os dados observados ($i = 1, ..., m$) e os dados censurados ($1 = m + 1, ..., n$). Assim, os dados observados são

tCMo<-c(0.008, 0.014, 0.016, 0.033, 0.060, 0.090, 0.093, 0.118, 0.134, 0.142, 0.162, 0.307, 0.323, 0.375, 0.616, 0.699, 0.827, 0.978, 1.008, 1.140, 1.145, 1.178, 1.389, 1.430, 1.512, 1.529, 1.641, 1.833, 1.907, 1.953, 2.181, 2.214, 2.296, 2.663, 2.682, 2.833, 3.027, 3.340, 4.195, 4.501, 4.553, 5.775, 5.879, 9.690)

e os dados censurados são

tCMc<-c(0.068, 0.342, 1.605, 1.759, 5.444, 9.025, 9.901, 12.764)

Algoritmo EM

Para estimar q:

```
emq<-function(tCMc,lambda,alpha,q)
{
q<-(1/52)*(44+sum(q/(1+(1-q)*lambda*tCMcˆalpha)))
return(q)
}
emq(tCMc,0.9039745,0.9783022,0.9313689143556819)
[1] 0.9607197
```

Para estimar *lambda*:

```
eml<-function(tCMo,tCMc,lambda,alpha,q,n)
{
for (i in 1:n)
{
veroslI<-44*(log(alpha)+log(lambda))+(alpha-1)*sum(log(tCMo))-2*
sum(log(1+lambda*tCMcˆalpha))-sum(q*log(1+lambda*tCMcˆalpha)/
(1+ (1-q)*lambda*tCMcˆalpha))
d1lamb<-(44/lambda)-2*sum((tCMoˆalpha)/(1+lambda*tCMoˆalpha))-
```

```
q*sum((tCMcˆalpha)/((1+(1-q)*lambda*tCMcˆalpha)*(1+lambda*
tCMcˆalpha)))
d2lamb<- -(44/(lambdaˆ2))+2*sum((tCMoˆ(2*alpha))/(1+lambda*
tCMoˆalpha)ˆ2)+q*sum((tCMcˆ(2*alpha))/((1+(1-q)*lambda*tCMcˆalpha)*
(1+lambda*tCMcˆalpha)ˆ2))
lambda<-lambda-d1lamb/d2lamb
veroslF<-44*(log(alpha)+log(lambda))+(alpha-1)*sum(log(tCMo))-2*
sum(log(1+lambda*tCMcˆalpha))-sum(q*log(1+lambda*tCMcˆalpha)/
(1+(1-q)*lambda*tCMcˆalpha))
Difl<–2*veroslF+2*veroslI
#Critério de paragem
if(Difl<=0.00001&Difl>=-0.00001) return(lambda)
}
#Para saber quando não há convergência
return(-1)
}
eml(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,10)
[1] 0.8918337
eml(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,25)
[1] 0.8918337
eml(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,5)
[1] 0.8918337
eml(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,4)
[1] -1
```

Para estimar alpha:

```
ema<-function(tCMo,tCMc,lambda,alpha,q,n)
{
for(i in 1:n)
{
verosaI<-44*(log(alpha)+log(lambda))+(alpha-1)*sum(log(tCMo)) -2*
sum(log(1+lambda*tCMcˆalpha))-sum(q*log(1+lambda*tCMcˆalpha)/
(1+(1-q)*lambda*tCMcˆalpha))
d1alpha<-(44/alpha)+sum(log(tCMo))-2*lambda*
sum((tCMoˆalpha)*log(tCMo)/(1+lambda*tCMoˆalpha))-
lambda*q*sum((tCMcˆalpha)*log(tCMc)/((1+ (1-q)*lambda*
tCMcˆalpha)*(1+lambda*tCMcˆalpha)))
d2alpha<–(44/(alphaˆ2))-2*lambda*sum(((tCMoˆalpha)*
(log(tCMo))ˆ2)/(1+lambda*tCMoˆalpha)ˆ2)-lambda*q*
sum(((tCMcˆalpha)*(log(tCMc))ˆ2)/((1+(1-q)*lambda*tCMcˆalpha)*
```

```
(1+lambda*tCMcˆalpha)ˆ2))
alpha<-alpha-d1alpha/d2alpha
verosaF<-44*(log(alpha)+log(lambda))+(alpha-1)*sum(log(tCMo))-2*
sum(log(1+lambda*tCMcˆalpha))- sum(q*log(1+lambda*
tCMcˆalpha)/(1+(1-q)*lambda*tCMcˆalpha))
Difa<–2*verosaF+2*verosaI
#Critério de paragem
if(Difa<=0.00001&Difa>=-0.00001) return (alpha)
}
#Para saber quando ainda não há divergência
return (-2)
}

ema(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,5)
[1] 0.9593713
ema(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,4)
[1] -2
```

As três funções anteriores (emq, eml, ema) são usadas no algoritmo final, que envolve todas as iterações até a convergência.

```
emlogis<-function(tCMo,tCMc,lambda, alpha,q,n)
{
for (i in 1:n)
{
verostI<-44*log(q/(1-q))+52*log(1-q)+(log(q/(1-q)))*q*sum(1/(1+
(1-q)*lambda*tCMcˆalpha))+44*(log(alpha)+log(lambda))+(alpha-1)*
sum(log(tCMo))-2*sum(log(1+lambda*tCMcˆalpha))-sum(q*log(1+
lambda*tCMcˆalpha)/ (1+(1-q)*lambda*tCMcˆalpha))
q1<-emq(tCMc,lambda,alpha,q)
lambda1<-eml(tCMo,tCMc,lambda,alpha,q,n)
alpha1<-ema(tCMo,tCMc,lambda,alpha,q,n)
q<-q1
lambda<-lambda1
alpha<-alpha1
verostF<-44*log(q/(1-q))+52*log(1-q)+(log(q/(1-q)))*q*sum(1/(1+
(1-q)*lambda*tCMcˆalpha))+44*(log(alpha)+log(lambda))+(alpha-1)*
sum(log(tCMo))-2*sum(log(1+lambda*tCMcˆalpha)) -sum(q*log(1+
lambda*tCMcˆalpha)/(1+(1-q)*lambda*tCMcˆalpha))
Dif<- -2*verostF+2*verostI
final<-c(i,q1,lambda1,alpha1)
```

```
if(Dif<=0.00001&Dif>=-0.00001) return(final)
}
return (-3)
}

mlogis(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,50)
[1] 42.0000000 1.0000000 0.8232959 0.9220229
mlogis(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,42)
[1] 42.0000000 1.0000000 0.8232959 0.9220229
mlogis(tCMo,tCMc,0.9039745,0.9783022,0.9313689143556819,41)
[1] -3
```

Observação

Como se obteve uma taxa de cura nula, o modelo obtido corresponde ao modelo habitual (sem ser de cura). Tendo esta informação, os parâmetros podem ser estimados da forma habitual, ou seja, usando a função *survreg*, da biblioteca *survival.*

Anexo B

Procedimento para obtenção do modelo de cura baseado na distribuição de Chen

Para determinar as estimativas iniciais de β e λ, realizou-se o seguinte procedimento.

Tendo em conta a Equação (3.2), vem que

$$\begin{cases} S(0.008) = 0.06863 + 0.93137\exp[\lambda(1-\exp(0.008^{\beta}))] \\ S(5.879) = 0.06863 + 0.93137\exp[\lambda(1-\exp(5.879^{\beta}))] \end{cases} \Longrightarrow$$

$$\begin{cases} \exp[\lambda(1-\exp(0.008^{\beta}))] = (0.98 - 0.06863)/(0.93137) \\ \exp[\lambda(1-\exp(5.879^{\beta}))] = (0.10 - 0.06863)/(0.93137) \end{cases}$$

Se $\beta = 1$,

$$\begin{cases} \exp[\lambda(1-\exp(0.008))] = 0.97853 \\ \exp[\lambda(1-\exp(5.879))] = 0.03368 \end{cases} \Longrightarrow$$

$$\begin{cases} \lambda = (\log(0.97853))/(1-\exp(0.008)) \\ \lambda = (\log(0.03368))/(1-\exp(5.879)) \end{cases} \Longrightarrow \begin{cases} \lambda = 2.70214 \\ \lambda = 0.00951 \end{cases}$$

Se $\beta = 0.5$,

$$\begin{cases} \exp[\lambda(1-\exp(0.008^{0.5}))] = 0.97853 \\ \exp[\lambda(1-\exp(5.879^{0.5}))] = 0.03368 \end{cases} \Longrightarrow$$

$$\begin{cases} \lambda = (\log(0.97853))/(1-\exp(0.008^{0.5})) \\ \lambda = (\log(0.03368))/(1-\exp(5.879^{0.5})) \end{cases} \Longrightarrow \begin{cases} \lambda = 0.23197 \\ \lambda = 0.32926 \end{cases}$$

Então, $\beta^{(0)} = 0.5$ e $\lambda^{(0)} = (0.23197 + 0.32926)/2 = 0.280615$.

Observação:
Ao iniciar o algoritmo com outros valores próximos de β, não se verificaram diferenças significativas nos valores das estimativas obtidas.

Algoritmo EM

Cada iteração é atualizada à custa dos valores da iteração anterior.

Para estimar q:

```
emqc<-function(tCMc,lambda,beta,q)
{
q<-(1/52)*(44+q*sum(exp(lambda*(1-exp(tCMcˆbeta)))/((1-q)+q*
exp(lambda*(1-exp(tCMcˆbeta))))))
return(q)
}

emqc(tCMc, 0.280615, 0.5, 0.9313689143556819)
[1] 0.926088
```

Para estimar lambda:

```
emlc<-function(tCMo,tCMc,lambda,beta,q)
{
lambda<-44/(q*sum(((exp(lambda*(1-exp(tCMcˆbeta))))/(1-q+q*
exp(lambda*(1 -exp(tCMcˆbeta)))))*(exp(tCMcˆbeta)-1))+
sum(exp(tCMoˆbeta)-1))
return(lambda)
}

emlc(tCMo,tCMc,0.280615, 0.5, 0.9313689143556819)
[1] 0.2971614
```

Para estimar beta:

```
emb<-function(tCMo,tCMc,lambda,beta,q,n)
{
for(i in 1:n)
{verosaI<-lambda*q*sum(((exp(lambda*(1-exp(tCMcˆbeta))))/(1-q+
q*exp(lambda*(1- exp(tCMcˆbeta)))))*(1-exp(tCMcˆbeta)))+44*
```

```
log(lambda*beta)+(beta-1)*sum(log(tCMo))+ sum(tCMoˆbeta)-
lambda*sum(1-exp(tCMoˆbeta))
d1beta<- -lambda*q*sum(((exp(lambda*(1-exp(tCMcˆbeta))))/(1-q+
q*exp(lambda*(1-exp(tCMcˆbeta)))))*(log(tCMc))*(tCMcˆbeta)*
exp(tCMcˆbeta))+44/beta+sum((log(tCMo))*(1+tCMoˆbeta))-
lambda*sum((log(tCMo))*(tCMoˆbeta)*exp(tCMoˆbeta))
d2beta<- -(lambda*q*sum(((exp(lambda*(1-exp(tCMcˆbeta))))/(1-q+
q*exp(lambda*(1-exp(tCMcˆbeta)))))*((log(tCMc))ˆ2)*(tCMcˆbeta)*
exp(tCMcˆbeta)*(1+tCMcˆbeta))+44/(betaˆ2)-sum(((log(tCMo))ˆ2)*
tCMoˆbeta)+lambda*sum(((log(tCMo))ˆ2)*(tCMoˆbeta)*
exp(tCMoˆbeta)*(1+tCMoˆbeta)))
beta<-beta-d1beta/d2beta
verosaF<-lambda*q*sum(((exp(lambda*(1-exp(tCMcˆbeta))))/(1-q+
q*exp(lambda*(1-exp(tCMcˆbeta)))))*(1-exp(tCMcˆbeta)))+44*
log(lambda*beta)+(beta-1)* sum(log(tCMo))+sum(tCMoˆbeta)-
lambda*sum(1-exp(tCMoˆbeta))
Difa<- -2*verosaF+2*verosaI
#Critério de paragem
if(Difa<=0.00001&Difa>=-0.00001) return(beta)
}
#Para saber quando ainda não há convergência
return(-2)
}

emb(tCMo,tCMc,0.280615, 0.5, 0.9313689143556819,13)
[1] 0.4756724
emb(tCMo,tCMc,0.280615, 0.5, 0.9313689143556819,12)

[1] -2
```

As três funções anteriores são usadas no algoritmo final, que envolve todas as iterações até à convergência.

```
emchen<-function(tCMo,tCMc,lambda,beta,q,n)
{
for (i in 1:n)
{
verostI<-8*log(1-q)+44*log(q)+(log(q)-log(1-q))*q*sum((exp(lambda*
(1-exp(tCMcˆbeta))))/(1-q+q*exp(lambda*(1-exp(tCMcˆbeta)))))+
lambda*q*sum(((exp(lambda *(1-exp(tCMcˆbeta))))/(1-q+q*
exp(lambda*(1-exp(tCMcˆbeta)))))*(1-exp(tCMcˆbeta)))+44*
log(lambda*beta)+(beta-1)*sum(log(tCMo))+sum(tCMoˆbeta)+
```

```
lambda*sum(exp(tCMoˆbeta))
q1<-emqc(tCMc,lambda,beta,q)
lambda1<-emlc(tCMo,tCMc,lambda,beta,q)
beta1<-emb(tCMo,tCMc,lambda,beta,q,n)
q<-q1
lambda<-lambda1
beta<-beta1
verostF<-8*log(1-q)+44*log(q)+(log(q)-log(1-q))*q*sum((exp(lambda*
(1-exp(tCMcˆbeta))))/(1-q+q*exp(lambda*(1-exp(tCMcˆbeta)))))+
lambda*q*sum(((exp(lambda*(1-exp(tCMcˆbeta))))/(1-q+
q*exp(lambda*(1-exp(tCMcˆbeta)))))*(1-exp(tCMcˆbeta)))+
44*log(lambda*beta)+(beta-1)*sum(log(tCMo))+sum(tCMoˆbeta)+
lambda*sum(exp(tCMoˆbeta))
Dif<- -2*verostF+2*verostI
final<-c(i,q1,lambda1,beta1)
if(Dif<=0.00001&Dif>=-0.00001) return (final)
}
return(-3)
}

emchen(tCMo,tCMc,0.280615, 0.5, 0.9313689143556819,100)
[1] 25.0000000 0.9261983 0.3339661 0.4543725
```

Anexo C

Criação do gráfico em R para comparação das curvas de sobrevivência

O gráfico apresentado na Figura 4.22 pode ser obtido através dos seguintes comandos.

#Para obter a curva da estimativa de Kaplan-Meier da função de sobrevivência.
plot(fCMama,conf.int=FALSE,xlim=c(0,14),ylim=c(0,1),xlab="Tempo (Anos)", ylab="Função de Sobrevivência", lwd= 1.5)

#Para definir a função de sobrevivência através da distribuição de Chen.
surchen<-1- 0.9261983 + 0.9261983 * exp(0.3339661 * (1-exp (ModCura$TempoAnosˆ0.543725)))

#Para adicionar ao gráfico a curva da função definida anteriormente.
lines(ModCura$TempoAnos, surchen,type="l", lty = "dashed", lwd= 1.5)
points(ModCura$TempoAnos, surchen, pch=2)

#Para definir a função de sobrevivência através da distribuição log-logística.
survloglogis<-1-1+1*(1/(1+0.82329* ModCura$TempoAnosˆ0.92202))

#Para adicionar ao gráfico a curva da função definida anteriormente.
lines(ModCura$TempoAnos, survloglogis, type="l", lty=9, lwd= 2)
points(ModCura$TempoAnos, survloglogis,pch=1)

#Para atribuir a legenda.

legend(locator(n=1),legend=c("K-M","Chen","Log-logística"), pch=c(3,2,1))

Observação

Como a taxa de cura é nula com a distribuição log-logística, o modelo obtido corresponde ao modelo habitual (sem ser de cura). Assim sendo, para definir a função de sobrevivência poderia ter sido usada a função *pllogis* da biblioteca *actuar*, com a devida adaptação, uma vez que a função densidade não tem a mesma expressão.