DM

Análise de Sobrevivência Aplicada à Educação

DISSERTAÇÃO DE MESTRADO

Susana Maria Pereira da Silva

MESTRADO EM MATEMÁTICA, ESTATÍSTICA E APLICAÇÕES

UNIVERSIDADE da MADEIRA

A Nossa Universidade

www.uma.pt

setembro | 2021

Análise de Sobrevivência Aplicada à Educação

DISSERTAÇÃO DE MESTRADO

Susana Maria Pereira da Silva

MESTRADO EM MATEMÁTICA, ESTATÍSTICA E APLICAÇÕES

ORIENTAÇÃO

Ana Maria Cortesão Pais Figueira da Silva Abreu

Análise de Sobrevivência Aplicada à Educação

DISSERTAÇÃO DE MESTRADO

Susana Maria Pereira da Silva

MESTRADO EM MATEMÁTICA, ESTATÍSTICA E APLICAÇÕES

JÚRI

Maria Teresa Alves Homem de Gouveia

Rita Maria César e Sá Fernandes

Ana Maria Cortesão Pais Figueira da Silva Abreu

“O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo.
Lute. Acredite. Conquiste. Perca. Deseje. Espere. Alcance. Invada. Caia.
A vitalidade é demonstrada não apenas pela persistência, mas pela capacidade de começar de novo.
Seja tudo o quiser ser, mas acima de tudo, seja sempre você”....

Hudson Alves

Agradecimentos

Foi com um enorme sacrifício que finalizo mais uma fase da minha vida. Tudo o que foi conseguido não teria sido possível sem o apoio e a força de Deus, pois existiram muitos momentos difíceis, principalmente nesta fase pandémica da Covid-19 que estamos todos a viver.

Estou muito grata à minha família, ao meu marido e às minhas duas filhas, por todos os momentos que não passei com eles e que, apesar disso, sempre me compreenderam e me apoiaram.

Foram dispensadas muitas horas de trabalho, de luta, de sacrifício, mas graças ao apoio e a disponibilidade da minha orientadora, Professora Doutora Ana Maria Cortesão Pais Figueira da Silva Abreu, muitos desses constrangimentos foram ultrapassados.

Agradeço também ao Professor Doutor João Filipe Pereira Nunes Prudente, Presidente da Escola Superior de Tecnologias e Gestão (ESTG), da Universidade da Madeira (UMa), à D. Sílvia Gomes, do Secretariado da ESTG, dos Cursos Técnicos Superiores Profissionais, à Dra. Margarida Santana, administradora dos Serviços de Ação Social da Universidade da Madeira (SASUMa) e à D. Paula Vasconcelos, da Unidade de Assuntos Académicos (UAA), pois sem os seus contributos, esta investigação não teria sido possível de concretizar.

Não podia também deixar de dar uma palavra especial aos professores e colegas do Mestrado em Matemática, Estatística e Aplicações, pelos conhecimentos transmitidos, pelo carinho e apoio ao longo do curso. Agradeço pelas amizades criadas e por terem feito parte desta minha jornada. Por último, agradeço a todos aqueles que, de uma forma direta ou indireta, contribuíram para a realização deste trabalho.

Resumo

Esta dissertação assentou em investigar o percurso escolar dos alunos dos quatro Cursos Técnicos Superiores Profissionais (CTeSP) (Agricultura Biológica, Contabilidade e Fiscalidade, Redes e Sistemas Informáticos e Guias da Natureza) da Escola Superior de Tecnologias e Gestão da UMa, entre os anos letivos 2015/2016 e 2019/2020.

Numa primeira vertente foi feita uma análise descritiva dos dados com base em indicadores de três dimensões: social, pessoal e escolar. Em seguida, aplicou-se a teoria da Análise de Sobrevivência ao tempo que os alunos levam até concluir o curso.

Da análise univariada efetuada às 15 variáveis presentes do estudo, verificámos que as variáveis CTeSP, ano letivo de início, ex-aluno, apoio social e trabalhador-estudante se mostraram significativas ao nível de significância 5%, em relação ao tempo até à conclusão do curso. No que se refere ao modelo de Cox, verificou-se que os antigos alunos da UMa apresentam um menor risco em terminar o curso ($36,10\%$ inferior) comparativamente aos alunos que nunca o foram. Os alunos com apoio social apresentam uma tendência positiva no seu desempenho comparativamente aos que não o têm e, por fim, comparando o CTeSP em Guias da Natureza com os restantes, verificou-se que os alunos que ingressaram nos CTeSP em Contabilidade e Fiscalidade e em Redes e Sistemas Informáticos, apresentam um risco de terminar o curso de $68,32\%$ e $126,27\%$, respetivamente, superior. No caso do CTeSP em Agricultura Biológica, o risco é o mesmo. Assim, estas três covariáveis demonstraram ter um forte impacto no tempo de duração de um curso.

Palavras-Chave: Análise de Sobrevivência; Educação; Cursos Técnicos Superiores Profissionais; Percurso escolar; Linguagem R; Tempo de permanência.

Abstract

This dissertation was based on investigating the academic trajectory of students from the four Higher Technical Professional Courses (CTeSP) (Biological Agriculture, Accounting and Taxation, Networks and Computer Systems and Nature Guides) of the UMa School of Technology and Management, between academic years 2015/2016 and 2019/2020.

In the first aspect, a descriptive analysis of the data was carried out based on indicators of three dimensions: social, personal and school. Then, the Survival Analysis theory was applied to the time it takes students to complete the course.

From the univariate analysis performed on the 15 variables present in the study, we found that the variables CTeSP, starting school year, ex-student, social support and student-worker were significant at the 5% significance level, in relation to the time until completion of the course. With regard to the Cox model, it was found that former UMa students have a lower risk of finishing the course ($36,10\%$ lower) compared to students who have never been. Students with social support show a positive trend in their performance compared to those who do not and, finally, comparing the CTeSP in Nature Guides with the others, it was found that students who joined the CTeSP in Accounting and Taxation and Computer Networks and Systems, present a higher risk of finishing the course of $68,32\%$ and $126,27\%$, respectively, higher. In the case of CTeSP in Organic Agriculture, the risk is the same. Thus, these three covariates have been shown to have a strong impact on the duration of a course.

Keywords: Survival Analysis; Education; Professional Higher Technical Courses; School route; R language; Lenght of stay.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Análise de Sobrevivência

Neste capítulo são descritos alguns dos principais conceitos da Análise de Sobrevivência, modelos fundamentais, assim como métodos não paramétricos de estimação (Rocha e Papoila (2009) [1]).

1.1 Introdução

A Análise de Sobrevivência é uma área Estatística que permite estudar o tempo de vida ou de sobrevivência desde um instante inicial bem definido até à ocorrência de um dado acontecimento de interesse. Assim sendo, este acontecimento de interesse, denominado também tempo de falha, pode ter vários sentidos, como a morte, o divórcio, a venda de um imóvel, a conclusão de uma licenciatura, entre outros. O tempo pode ser medido em diferentes unidades, como sejam dias, semanas, meses ou anos.

A medicina é uma das áreas onde este tipo de análise tem um forte impacto, mas também pode ser usada em outras áreas como, por exemplo, no desporto, na política, na psicologia, na informática e na educação. Assim, a Análise de Sobrevivência poderá ser aplicada para estudar o tempo que um aluno demora a concluir um curso ou o tempo que um casal leva até se divorciar ou até mesmo, o tempo que um ex-recluso permanece em liberdade até reincidir novamente no crime.

O que é caraterístico deste tipo de análise de dados é a eventual existência de observações censuradas. Este tipo de dados surge quando, durante o período de tempo em que os indivíduos estão a ser observados, para alguns, não ocorre o acontecimento de interesse.

Na Análise de Sobrevivência, para não existir perda de informação, os tempos censurados são também incluídos no estudo, conjuntamente com os tempos de vida efetivamente observados. Assim, a censura origina dados incompletos, mas não é a única forma de isso acontecer. A truncatura também origina este tipo de dados e ocorre quando apenas são incluídos no estudo os indivíduos que satisfazem determinada condição.

O tempo de vida dos indivíduos também pode sofrer influência de vários fatores, nomeadamente caraterísticas próprias do indivíduo observado como, por exemplo, o sexo e a idade, e características externas ao indivíduo como, por exemplo, o tratamento administrado a um dado indivíduo ou o estabelecimento de ensino que frequentou. Estes fatores, que poderão interferir no tempo de sobrevivência de um indivíduo, são denominados por variáveis explanatórias ou covariáveis.

Na Análise de Sobrevivência, os modelos de regressão são instrumentos de grande importância e utilidade, sendo que serão abordados neste documento modelos adaptados aos aspetos concretos das observações recolhidas.

1.2 Conceitos fundamentais

Seja T uma variável aleatória absolutamente contínua, não negativa, que representa o tempo de vida de um indivíduo de uma dada população homogénea, em que a sua distribuição pode ser caraterizada através de uma das seguintes funções: $S(t)$, a função de sobrevivência; $f(t)$, a função densidade de probabilidade; e $h(t)$, a função de risco. Em relação à função de sobrevivência, esta é definida como sendo a probabilidade de um indivíduo sobreviver para além do instante t e representa-se da seguinte forma:

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

em que $F(t)$ representa a função de distribuição de T.

A função de sobrevivência verifica as seguintes propriedades:

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

Quanto à função densidade de probabilidade num instante t, esta é definida por:

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

No que concerne à função de risco (*hazard function*), esta também é designada por função intensidade, taxa de falha ou força de mortalidade e descreve a evolução ao longo do tempo da probabilidade instantânea de morte de um indivíduo, sendo definida da seguinte forma:

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

A função de risco verifica as seguintes propriedades:

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

Como consequência das três definições anteriormente referidas é possível estabelecer as seguintes relações entre a função de sobrevivência, a função de densidade de probabilidade e a função de risco:

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

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

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

$$f(t) = -S'(t). \tag{1.2}$$

Define-se ainda a função de risco cumulativa, $H(t)$, que é uma função não negativa e monótona crescente, da seguinte forma:

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

Assim sendo, de (1.1), obtemos:

$$S(t) = \exp[-H(t)] \Leftrightarrow H(t) = -\log S(t).$$

1.3 Função de risco

A função de risco é a função mais adequada para estudar o modo como o risco de morte se altera ao longo do tempo. De facto, consoante a forma desta função, assim o modo como o risco de morte evolui, tal como se descreve a seguir. Assim, considerando as formas mais comuns, a função de risco pode ser:

- **monótona crescente:** é a forma mais comum e os modelos mais aplicados em Análise de Sobrevivência apresentam este tipo de função de risco. Ocorre quando os indivíduos são observados num período da sua vida durante o qual acontece um envelhecimento progressivo. Nesta situação, de entre os sobreviventes num dado instante, a proporção de indivíduos que morrem nesse mesmo instante, aumenta com o tempo.

- **monótona decrescente:** é a menos comum, pois representa uma situação em que, quanto mais tempo o indivíduo sobrevive, menor é a probabilidade de morte no instante seguinte.

- **constante:** surge quando o risco de morte não se modifica. Por exemplo, quando o estudo é efetuado durante um período de tempo razoavelmente curto para se poder considerar que o risco de morte é inalterado, isto é, constante;

- **bathtub-shaped:** surge em populações em que os indivíduos são acompanhados desde o seu nascimento até à sua morte. Neste caso, a função de risco, no início, é decrescente, durante um largo período de tempo é constante e no final da vida, devido ao envelhecimento da população, é crescente;

- **hump-shaped ou unimodal:** no início, o risco de morte é crescente, depois, ao fim de algum tempo, passa a ser decrescente.

Em suma, a função de risco pode ser monótona (crescente, decrescente ou constante) ou não monótona, sendo que, a todas estas formas correspondem sempre uma função de sobrevivência decrescente.

1.4 Mecanismos de censura

Quando falamos de censura, referimos aqueles indivíduos que, durante o período em que foram observados, não foi exequível verificar o acontecimento de interesse, o que faz com que haja perda de informação.

Portanto, uma observação é censurada quando não é possível determinar o seu valor exato, tendo-se conseguido apenas obter um limite inferior desse valor (censura à direita), ou um limite superior (censura à esquerda), ou até mesmo, ambos os casos (censura intervalar).

Existem vários tipos de censura, sendo que a mais usual é a censura à direita. Assim, num determinado estudo, pode existir:

- Censura à direita;

- Censura à esquerda;

- Censura intervalar;

- Censura dupla;

- Censura independente e não informativa.

Censura à direita: quando apenas se sabe que o tempo de vida excede um dado valor, visto que, a observação do indivíduo termina antes da ocorrência do acontecimento de interesse. Se o indivíduo inicia o estudo no instante t_0 e morre no instante $t_0 + t$, sendo o valor t desconhecido, pode ser devido a duas situações. O indivíduo permanece vivo até ao final do estudo, ou, por outro lado, o indivíduo é perdido para o *follow-up*, isto é, durante o acompanhamento. Se existir informação de que o indivíduo se encontra vivo no instante $t_0 + c$, em que $c < t$, o tempo $c - t_0$ é designado por tempo de sobrevivência censurado e estamos perante a presença de um caso de censura à direita.

Existem três tipos de censura à direita: censura de tipo I, censura de tipo II e censura aleatória.

- **Censura de tipo I:** seja uma amostra de n indivíduos e T_i a variável aleatória que representa o tempo de vida do i-ésimo indivíduo. Consideremos que a cada indivíduo corresponde um período de observação c_i designado por tempo de censura potencial, tal que a morte desse indivíduo apenas será observada se ocorrer durante esse período. Assim sendo, as observações são da forma (t_i, δ_i), $i = 1, ..., n$, em que $t_i = min(T_i, c_i)$ e

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

  A censura de tipo I ocorre quando os períodos de observação $c_1, ..., c_n$ são fixados previamente pelo investigador. Neste caso, o número de mortes observadas é uma variável aleatória. Assim sendo, só podemos ter conhecimento do tempo de vida de um indivíduo se a morte ocorrer antes do instante pré-definido.

- **Censura de tipo II:** o estudo finaliza no instante em que é observada a r-ésima morte, sendo que, r representa um número pré-determinado $(1 \leqslant r \leqslant n)$. Neste caso, o tempo de duração do estudo é uma variável aleatória.

- **Censura aleatória:** este é o tipo de censura mais geral, em que a cada indivíduo está associado um tempo de vida T_i e um tempo de censura (potencial) C_i, onde os tempos de censura e os tempos de vida são variáveis aleatórias mutuamente independentes e T_i e C_i são igualmente variáveis aleatórias independentes. Logo, as observações consistirão nos pares de variáveis aleatórias (γ_i, δ_i), $i = 1, ..., n$, em que $\gamma_i = min(T_i, C_i)$ e

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

Censura à esquerda: este tipo de censura é menos comum e ocorre quando o tempo de vida é menor que o tempo C_e observado. Nesta situação, pode ocorrer que o acontecimento de interesse tenha acontecido antes do indivíduo entrar em estudo.

Assim sendo, para uma amostra de n indivíduos, as observações corresponderão a pares de variáveis aleatórias (φ_i, ω_i), $i = 1, ..., n$, em que $\varphi_i = max(T_i, C_e)$ e

$$\omega_i = \begin{cases} 1 & se \quad \varphi_i \geq C_e \\ 0 & se \quad \varphi_i < C_e. \end{cases}$$

Censura intervalar: quando não se sabe o instante exato que em acontece o acontecimento de interesse, mas sabemos ter acontecido num determinado intervalo aleatório de tempo. Por outras palavras, este tipo de censura ocorre quando não temos conhecimento dos tempos de sobrevivência precisos, mas temos conhecimento apenas que eles ocorrem dentro de um intervalo.

Por exemplo, consideremos um estudo do qual interessa-nos saber o tempo de recorrência de um determinado cancro após uma cirurgia para remoção do tumor. Vamos supor que, três meses depois da cirurgia o paciente é observado e é constatado que este está livre da doença, mas quando este é examinado seis meses depois da cirurgia, é verificado a recorrência da mesma doença.

Neste caso, em particular, o tempo de recorrência real do paciente é desconhecido, apenas temos conhecimento que este está entre três e seis meses, logo, estamos perante um tempo de censura intervalar [27].

Existem dois tipos de censura intervalar: caso I (quando apenas temos conhecimento que, num determinado instante de acompanhamento, o acontecimento de interesse já aconteceu ou ainda não) e caso II (quando temos conhecimento do intervalo durante o qual se concretizou o acontecimento de interesse).

Censura dupla: existem duas situações distintas em que esta designação é usada por diferentes autores. A primeira situação está relacionada com estudos de sobrevivência em que fazem parte dados censurados à direita, outros à esquerda e as restantes observações são exatas. A segunda situação está relacionada quando tanto a origem como o tempo que decorre até ocorrer o acontecimento de interesse são censurados.

Censura independente e não informativa: é necessário que, para os métodos usuais de Análise de Sobrevivência sejam válidos, a censura seja independente. Estamos perante um caso de censura independente quando a razão para ocorrer censura é independente da razão que leva, por exemplo, à morte, isto é, quando a censura é independente do mecanismo responsável pela morte de um indivíduo [3]. A censura não informativa determina que a distribuição do tempo de censura não depende do parâmetro θ, que é o parâmetro de interesse que carateriza a distribuição do tempo de vida.

1.5 Variáveis explanatórias ou covariáveis

Tal como já havíamos referido anteriormente, o propósito fundamental da Análise de Sobrevivência é o estudo do tempo de vida dos indivíduos. Todavia, a sua sobrevivência poderá ser afetada por vários fatores de risco ou de prognóstico, como tratamentos, propriedades intrínsecas do indivíduo ou até mesmo, variáveis exógenas. Estes fatores são designados por variáveis explanatórias ou covariáveis. Os valores individuais destas variáveis devem ser registados sempre que possível, uma vez que fornecem informações sobre a heterogeneidade presente numa determinada população.

As variáveis explanatórias ou covariáveis podem ser classificadas em:

- **constante** se o seu valor não sofre qualquer alteração durante todo o tempo em que o indivíduo está a ser observado.
- **dependente do tempo** se o seu valor se modifica ao longo do período de observação. Estas covariáveis dependentes do tempo podem ser classificadas em:
  - **externa** se não está diretamente relacionada com o mecanismo que regula a morte dos indivíduos.
  - **interna** se advém de uma análise realizada sobre um indivíduo enquanto está vivo e não censurado, ao longo do tempo que foi observado, facultando informações sobre o seu tempo de vida.

1.6 Função de verosimilhança

Vamos supor que a distribuição do tempo de vida segue um determinado modelo paramétrico, indexado por um vetor de parâmetros $\boldsymbol{\theta}$ que se pretende estimar. Para construir a função de verosimilhança é necessário considerar o contributo dado por cada uma das observações, que varia consoante o tipo de censura, na eventualidade de não se tratar de uma observação exata.

Assim sendo, essa contribuição dada por cada indivíduo, dependente do tempo de vida observado ou de acordo com o mecanismo de censura a que está sujeito, será dada pela:

- Função densidade, $f(t_i)$, para tempos de vida exatos;
- Função de sobrevivência, $S(t_i)$, para observações censuradas à direita;
- Função de distribuição, $1 - S(t_i)$, para observações censuradas à esquerda;
- Diferença entre as funções de sobrevivência nos extremos do intervalo l_i e r_i, $S(l_i) - S(r_i)$, para observações sujeitas a censura intervalar – caso II.

Consideremos então que pretendemos construir a função de verosimilhança quando temos observações não censuradas e observações censuradas à direita.

No caso de uma observação corresponder a um tempo de vida exato, isto é, uma observação não censurada, esta facultará informação sobre a probabilidade de ocorrer um acontecimento de interesse nesse instante, que será aproximadamente igual à função densidade da variável tempo de vida $f(t)$, nesse mesmo instante.

Em relação a uma observação censurada à direita, o verdadeiro tempo de vida é superior ao tempo de vida observado, sendo a informação facultada através da função de sobrevivência $S(t)$, nesse mesmo instante.

Assim, dada uma amostra de dimensão n, $((t_1, \delta_1), ..., (t_n, \delta_n))$, e admitindo que a distribuição do tempo de censura não depende do vetor de parâmetros de interesse $\boldsymbol{\theta}$, podemos representar toda a inferência sobre este vetor na verosimilhança seguinte

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

ou ainda em

$$L = \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i). \tag{1.4}$$

De referir que, os resultados assintóticos usuais sobre a teoria da máxima verosimilhança permanecem válidos, em condições de regularidade bastante gerais nos processos de morte e censura. Com efeito, o estimador de máxima verosimilhança $\widehat{\boldsymbol{\theta}}$ tem distribuição assintótica normal multivariada com valor médio $\boldsymbol{\theta}$ e matriz de covariância $I(\boldsymbol{\theta})^{-1}$, em que $I(\boldsymbol{\theta})$ é a matriz de informação de Fisher.

1.7 Estimação não paramétrica da função de sobrevivência através do estimador de Kaplan-Meier

Se não existir censura, a função de sobrevivência num dado instante t é estimada pela proporção de tempos de vida observados de valor superior a t, função esta que se designa por função de sobrevivência empírica.

Por outro lado, quando existem observações censuradas, nem todos os tempos de vida serão observados. Assim sendo, Kaplan e Meier (1958) [22] para superar esta situação, propuseram um estimador não paramétrico da função de sobrevivência, denominado estimador de Kaplan-Meier ou estimador "produto-limite".

Consideremos $t_{(1)}, ..., t_{(r)}$ os instantes de morte diferentes numa amostra de dimensão n, em que $r \leq n$, d_i o número de mortes ocorridas no instante $t_{(i)}$ e n_i o número de indivíduos em risco no instante $t_{(i)}$. Então, o estimador de Kaplan-Meier da função de sobrevivência define-se como sendo:

$$\widehat{S}(t) = \prod_{i:t_{(i)} \leq t} \frac{n_i - di}{n_i} = \prod_{i:t_{(i)} \leq t} (1 - \frac{d_i}{n_i}), \tag{1.5}$$

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

O cálculo da estimativa da variância de $\widehat{S}(t)$ pode ser efetuado através da seguinte expressão, denominada por fórmula de Greenwood:

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

Seguidamente são apresentadas algumas observações sobre o estimador de Kaplan-Meier:

- quando não existe censura, o estimador de Kaplan-Meier coincide com a função de sobrevivência empírica;
- se a maior observação registada t^* for censurada, $\widehat{S}(t)$ nunca toma o valor zero. Neste caso, considera-se que a estimativa está definida somente até esse instante, sendo $\widehat{S}(t) = \widehat{S}(t_{(r)})$ para $t_{(r)} \leq t \leq t^*$;
- $\widehat{S}(t) = 0$ para $t \geq t_{(r)}$, se $t_{(r)}$ for a maior observação registada, ou seja, se a maior observação for não censurada;
- $\widehat{S}(t)$ é um estimador consistente de $S(t)$ e pode ser considerado como um estimador de máxima verosimilhança não paramétrico de $S(t)$, em certas condições de regularidade;
- O estimador de Kaplan-Meier é autoconsistente, segundo a definição de Efron (1967) [9].

Dado que $\widehat{S}(t)$ tem uma distribuição assintótica normal de valor médio $S(t)$ e variância estimada através da fórmula de Greenwood, um intervalo de $100(1-\alpha)\%$ de confiança para a função de sobrevivência no instante t_0 é obtido por

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

em que $z_{1-\frac{\alpha}{2}}$ representa o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição $N(0,1)$.

Apesar deste intervalo de confiança ser o mais usual e o que se obtém, por norma, através de *software* estatístico, pode originar alguns problemas, como sejam, os seus limites se situarem fora do intervalo $(0,1)$.

Por esta razão, há outras alternativas, entre elas, começar por obter um intervalo de confiança para uma transformação de $\widehat{S}(t_0)$ (por exemplo, $\log(\widehat{S}(t_0))$ ou $\log[-\log(\widehat{S}(t_0))]$ e, depois de aplicar a transformação inversa, efetuar o cálculo do intervalo de confiança para $\widehat{S}(t_0)$.

Uma vez que, em geral, a distribuição do tempo de vida é assimétrica positiva, é preferível recorrer ao uso da mediana para caracterizar a localização do centro da distribuição.

Assim, dado que $\widehat{S}(t)$ é a estimativa de Kaplan-Meier da função de sobrevivência, então a estimativa da mediana do tempo de vida será definida como sendo

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

em que $t_{(i)}$ é o i-ésimo instante de morte, $i = 1, ..., r$.

Como a estimativa da função de sobrevivência nem sempre atinge o valor zero (basta que a última observação corresponda a um tempo de vida censurado), em particular, pode nem atingir o valor $0,5$. Neste último caso, não será possível obter uma estimativa da mediana, mas apenas de algum quantil inferior, cuja expressão geral é dada por

$$\widehat{X_p} = min\{t_{(i)} : \widehat{S}(t_{(i)}) \leq 1-p\}.$$

1.8 Estimação não paramétrica da função de risco cumulativa

Tendo em conta a relação entre $H(t)$ e $S(t)$, um estimador natural de $H(t)$ é

$$\widehat{H}(t) = -\log \widehat{S}(t).$$

O estimador de Nelson-Aalen, também designado por função de risco cumulativa empírica, é um estimador alternativo e mais habitual, com melhor comportamento para amostras pequenas. Este estimador é definido por

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

e a sua estimativa da variância é dada por

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

Tal como a função de risco é importante para estudar o modo como o risco de morte se altera ao longo do tempo, também a função de risco cumulativa assume um papel relevante na identificação de modelos para o tempo de vida. De facto, como $H(t)$ é uma função não decrescente, ela será convexa se $h(t)$ for crescente, côncava se $h(t)$ for decrescente e linear se $h(t)$ for constante.

Novamente devido à relação entre $S(t)$ e $H(t)$, através do estimador de Nelson Aalen, obtém-se um outro estimador para $S(t)$, designado por estimador de Breslow, definido por

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

1.9 Testes não paramétricos

A obtenção da estimativa de Kaplan-Meier da função de sobrevivência para cada um dos grupos e a sua representação gráfica, além da representação gráfica das estimativas de Nelson-Aalen da função de risco cumulativa, permite-nos ter informações úteis sobre o comportamento de curvas de sobrevivência para cada grupo e, consequentemente, investigar potenciais diferenças significativas nas mesmas. No entanto, para haver uma avaliação mais rigorosa acerca da existência de eventuais diferenças significativas entre as várias curvas de sobrevivência é necessário efetuar alguns testes de hipóteses.

Existem diversos testes não paramétricos que permitem comparar as curvas de sobrevivência de diferentes grupos de indivíduos. No entanto, a seleção do teste mais apropriado depende de vários fatores, nomeadamente dos padrões de morte e da censura nos diversos grupos e da relação entre as funções de risco correspondentes, como também da hipótese alternativa de interesse.

Assim sendo, considerando dois grupos de indivíduos, em que S_i representa a função de sobrevivência de um individuo no i-ésimo grupo $(i = 1,2)$, para qualquer teste de hipóteses, as hipóteses a testar serão sempre:

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

Nesta secção vamos abordar apenas os testes log-rank e Gehan.

1.9.1 Teste log-rank

Sejam m e n as dimensões de duas amostras de dois grupos (grupo 1 e grupo 2), respetivamente. Representamos por:

- $t_1 < \cdots < t_k$ os instantes de morte distintos correspondentes aos $m+n$ indivíduos;
- d_j o número de mortes ocorridas em t_j, em que $j = 1, \ldots, k$;
- d_{ij} o número de mortes ocorridas em t_j no grupo i, em que $i = 1, 2$;
- n_j o número de indivíduos em risco logo antes de t_j, em que $j = 1, \ldots, k$;
- n_{ij} o número de indivíduos em risco logo antes de t_j no grupo i, em que $i = 1, 2$.

Tendo em conta que, sob H_0 a distribuição de d_{1j} condicional a d_j e a n_j é hipergeométrica, o valor médio e a variância de d_{1j} são, respetivamente,

$$e_{1j} = \frac{n_{1j} d_j}{n_j} \quad e \quad \vartheta_{1j} = \frac{n_j n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)},$$

sendo que e_{1j} significa o número esperado de mortes no instante t_j no grupo 1.

Com o objetivo de obter uma medida global do desvio dos valores observados de d_{1j} em relação aos valores esperados, supomos a estatística seguinte

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

sendo que $\sum d_{1j} - \sum e_{1j}$ representa a diferença entre o número total de mortes observadas e esperadas no instante t_j no grupo 1.

Assim sendo, $E(U) = 0$, dado que $E(d_{1j}) = e_{1j}$ e $var(U) = \sum_{j=1}^{k} \vartheta_{ij}$, dado que é assumida a independência entre as mortes que ocorreram nos k instantes.

A estatística de teste apresentada por Mantel e Haenszel (1959) [33] é

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

que, sob H_0 apresenta uma distribuição assintótica Qui-quadrado χ_1^2, com 1 grau de liberdade.

Este teste é o mais adequado para detetar afastamentos da hipótese de igualdade das distribuições dos tempos de vida de dois grupos quando, as respetivas funções de risco são proporcionais. O teste log-rank também é potente para situações em que as funções de risco não são proporcionais, mas não se cruzam.

Uma das formas de podermos avaliar a validade da hipótese de riscos proporcionais, passa pela representação gráfica das estimativas das funções de sobrevivência. Logo, se as estimativas das funções de sobrevivência não se intersetam, então não há indicação de que a hipótese de riscos proporcionais não seja válida.

Uma outra alternativa gráfica para avaliar a hipótese de riscos proporcionais, é a seguinte:

Se $h_1(t)$ e $h_2(t)$ forem funções de risco proporcionais então $h_1(t) = e^{\beta} h_2(t)$, donde

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

e ainda

$$\log[-\log S_1(t)] = \beta + \log[-\log S_2(t)].$$

Deste modo averiguamos se os gráficos do logaritmo das funções de risco cumulativas correspondentes aos dois grupos são equidistantes ao longo do tempo. Logo, considerando as estimativas de $\widehat{S}_1(t)$ e $\widehat{S}_2(t)$ não fundamentadas na hipótese de riscos proporcionais, o gráfico de $\log[-\log[\widehat{S}_1(t)]$ *versus* t será aproximadamente paralelo ao gráfico $\log[-\log[\widehat{S}_2(t)]$ *versus* t, no caso de $h_1(t)$ e $h_2(t)$ serem proporcionais.

1.9.2 Teste de Gehan

A generalização do teste de Mann-Whitney-Wilcoxon para dados censurados ou teste de Wilcoxon generalizado foi proposto por Gehan (1965) [12].

A estatística de teste do teste de Gehan pode ser escrita na forma

$$Q_G = \frac{U_G^2}{var(U_G)}$$

em que,

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

onde $e_{1j} = n_{1j}d_j/n_j$. A variância da estatística U_G é

$$V_G = \sum_{j=1}^{k} n_j^2 \vartheta_{1j}$$

e a estatística de teste do teste de Gehan será

$$W = U_G^2/V_G.$$

Sob a validade de H_0, W tem distribuição assintótica de Qui-quadrado χ_1^2, com 1 grau de liberdade.

Verificamos que cada diferença $(d_{1j} - e_{1j})$ é ponderada por n_j, isto é, pelo número de indivíduos em risco no instante t_j. Então, de acordo com a expressão U_G constata-se que é atribuído maior peso às diferenças $(d_{1j} - e_{1j})$ correspondentes aos instantes onde o número total de indivíduos em risco é superior, isto é, aos instantes na fase inicial do estudo.

Assim sendo, por esta razão este teste é considerado menos sensível que o teste long-rank pois dá menos importância a diferenças entre o número observado e o número esperado de mortes que se observem na cauda direita da distribuição do tempo de vida. É o teste mais potente para detetar os efeitos a curto prazo, devido ao maior peso que é facultado nas observações mais pequenas.

É importante salientar que estes dois testes (log-rank e Gehan) podem ser generalizados de modo a permitir a comparação de r grupos, sendo $r \geq 3$. Sob a validade de $H_0 : S_1 = ... = S_r$, a estatística de teste tem distribuição assintótica Qui-quadrado com $r-1$ graus de liberdade. Assim, desta forma, o teste log-rank continua a ter a mesma designação e a generalização do teste de Gehan (teste utilizado para $r = 2$), que na verdade é uma generalização do teste de Kruskal-Wallis, designa-se por teste de Breslow.

Neste estudo para comparar as curvas de sobrevivência dos vários grupos da mesma variável em estudo, foram utilizados estes três testes: log-rank ou Mantel-Haenzel, Gehan e Breslow. Todos os testes estatísticos realizados foram bilaterais ou bicaudais e adotou-se um nível de significância de 5%.

1.10 Modelo de regressão de Cox

1.10.1 Introdução

Devido à sua versatilidade e flexibilidade, o modelo proposto por Cox (1972) [7] tornou-se no modelo de regressão mais utilizado na análise de tempos de vida. A sua aplicação abrange um grande número de situações práticas, nomeadamente em áreas como a medicina, a engenharia e a sociologia. O facto de o modelo de Cox ser formulado com base na relação entre a função de risco e as covariáveis é um dos aspetos inovadores deste modelo. Este modelo pressupõe que o efeito das covariáveis se mantém constante ao longo do tempo. Seguidamente será apresentada a descrição do modelo de Cox, bem como, testes de hipóteses para a sua validação e as suas representações gráficas.

1.10.2 Definição do modelo de Cox

Seja T uma variável aleatória contínua, que representa o tempo de vida de um indivíduo e $\boldsymbol{z} = (z_1, \ldots, z_p)'$ o vetor de covariáveis associado a cada indivíduo. Seja ainda $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ o vetor dos coeficientes de regressão (desconhecidos) que representam o efeito das covariáveis na sobrevivência e $h_0(t)$ a função de risco subjacente (função arbitrária não negativa) que é a que corresponde a um indivíduo associado o vetor de covariáveis nulo, $\boldsymbol{z} = \boldsymbol{0}$. Então, o modelo de regressão proposto por Cox (1972), definido à custa das funções de risco, tem a seguinte expressão:

$$h(t;\boldsymbol{z}) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{z}) = h_0(t)\exp(\beta_1 z_1 + \ldots + \beta_p z_p). \tag{1.6}$$

Assim, trata-se de um modelo de regressão semi-paramétrico, uma vez que, apesar do efeito das covariáveis ser modelado parametricamente, não é especificada a função de risco subjacente $h_0(t)$. Além disso, as covariáveis têm um efeito multiplicativo na função de risco, sendo que o fator $\exp(\boldsymbol{\beta}'\boldsymbol{z})$, designado por risco relativo, representa o risco de morte de um indivíduo com vetor de covariáveis $\boldsymbol{z}$ relativamente a um indivíduo padrão (aquele em que $\boldsymbol{z} = \boldsymbol{0}$).

A razão das funções de risco de dois indivíduos com vetores de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$ não depende de t

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

e as funções de risco são proporcionais.

De acordo com o fator $\exp(\boldsymbol{\beta}')$, denominado por risco relativo, isto é, o risco de morte de um dado indivíduo em relação ao indivíduo padrão, as covariáveis têm um efeito multiplicativo. Logo, de acordo com o modelo de Cox (1972), a influência das covariáveis na função de risco não se altera durante o período de observação dos indivíduos.

A quantidade $\boldsymbol{\beta}'\boldsymbol{z}_i = \sum_{j=1}^{p} \beta_j z_{ij}$. representa o índice de prognóstico (*risk score*) para o i-ésimo indivíduo.

De acordo com a relação entre a função de risco, a função de sobrevivência e a igualdade (1.6), o modelo de regressão de Cox, pode ser apresentado com base na função de sobrevivência, isto é,

$$S(t;\boldsymbol{z}) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\boldsymbol{z})} \tag{1.7}$$

em que $S_0(t)$ é a função de sobrevivência de um dado indivíduo, com vetor de covariáveis $\boldsymbol{z} = \boldsymbol{0}$, também denominada por função de sobrevivência subjacente.

1.10.3 Modelo de Cox estratificado

Suponhamos que as funções de risco correspondentes às diversas categorias de uma dada variável qualitativa são evidentemente não proporcionais. Pode acontecer que existam subgrupos de indivíduos em cada categoria que apresentam funções de risco proporcionais, portanto uma extensão do modelo de Cox que verifica esta situação é dada pelo modelo

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

para $j = 1,..,m$, onde m representa o número de categorias da covariável em questão e $\boldsymbol{z}$ o vetor das restantes covariáveis. Este modelo baseia-se no facto que para os indivíduos na categoria j, a que estão associados os vetores de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, as funções de risco são proporcionais se:

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

Embora os indivíduos em categorias diferentes possam apresentar funções de risco não proporcionais, dado que as funções $h_{0_1}(t),...,h_{0_m}(t)$ são arbitrárias e não relacionadas, os coeficientes de regressão β não dependem da categoria e, por essa razão, o efeito das covariáveis em todas categorias será o mesmo.

Para cada categoria j é obtida uma função de verosimilhança parcial $L_j(\beta)$ e a estimação dos parâmetros β é feita através da soma dos logaritmos das funções de verosimilhança parciais, sendo $L(\beta)$ maximizada em relação a β.

No entanto, importa referir que a estratificação relativa a uma dada covariável também apresenta as suas desvantagens, uma vez que impede a estimação do efeito dessa covariável no tempo de vida, isto é, não permite avaliar o efeito da covariável responsável pela estratificação, para além do facto de que se a covariável em questão é contínua será necessário categorizá-la, o que por si só também acarreta problemas.

1.10.4 Coeficientes de regressão

A interpretação dos coeficientes de regressão é fundamental para a compreensão da relação entre as variáveis explanatórias e a variável resposta no modelo de Cox.

Habitualmente a sua interpretação é feita através de $\exp(\beta_j)$, pois esta quantidade tem um significado mais direto no que diz respeito ao risco de morte.

Se considerarmos dois indivíduos associados aos respetivos vetores de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, que diferem nos valores da covariável z_j. Através da função de risco, temos

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

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

1.10.5 Função de verosimilhança

Vamos considerar um estudo com n indivíduos e que foram observados k tempos de vida distintos, tal que $t_{(1)} < ... < t_{(k)}$, com $k < n$. Então, o conjunto de risco no instante $t_{(i)}$ é definido da seguinte forma

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

e representa o conjunto de índices associados aos indivíduos em observação imediatamente antes do instante $t_{(i)}$.

A função de verosimilhança baseada para realizar inferência sobre $\boldsymbol{\beta}$, Cox (1972), é definida da seguinte forma

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{z}_{(i)})}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{z}_l)} \tag{1.8}$$

em que $\boldsymbol{z}_{(i)}$ é o vetor de covariáveis associado ao indivíduo que morre no instante $t_{(i)}$.

Dado que $L(\boldsymbol{\beta})$ não depende de $h_0(t)$, então é possível realizar a inferência sobre o vetor de parâmetros $\boldsymbol{\beta}$, ignorando $h_0(t)$. Note-se que a função de verosimilhança $L(\boldsymbol{\beta})$, considerada por Cox é uma verosimilhança parcial. A verosimilhança completa é representada da seguinte forma

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

em que D é o conjunto de indivíduos nos quais a morte foi observada.

Em condições de regularidade bastante gerais (Andersen e Gill, 1982 [24]), o estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$ é consistente, assintoticamente normal com valor médio $\boldsymbol{\beta}$ e matriz de covariância $I(\boldsymbol{\beta})^{-1}$, em que $I(\boldsymbol{\beta})$ representa a matriz de informação de Fisher,

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

Note-se que a função de verosimilhança definida em (1.8) é válida no caso de as observações serem todas distintas, como é o caso se for considerado um modelo contínuo. No entanto, devido a limitações na medição e registo dos dados, podem ocorrer observações com o mesmo valor. Nestas situações, é necessário fazer uma pequena alteração na expressão da função de verosimilhança, alteração esta integrada no *software* estatístico.

1.10.6 Estimação da função de sobrevivência

Tendo em conta (1.7), para poder obter estimativas de $S(t;\boldsymbol{z})$ é necessário estimar $S_0(t)$. Serão apresentadas em seguida duas expressões dos estimadores de $S_0(t)$: a proposta por Kalbfleisch e Prentice (1973) [25] e a proposta por Breslow (1974) [4].

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

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

Admitindo que $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ (onde $\widehat{\boldsymbol{\beta}}$ corresponde ao estimador obtido a partir da verosimilhança parcial) e maximizando a função de verosimilhança em relação a $\alpha_1, \ldots, \alpha_k$, para $i = 1, \ldots, k$, temos as seguintes equações de máxima verosimilhança

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

Se $d_i = 1$, em que $i = 1, ..., k$, obtemos a seguinte equação

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

Caso contrário, seria necessário resolver a equação através de um método iterativo.

Assim, o estimador de máxima verosimilhança de $S_0(t)$, proposto por Kalbfleisch e Prentice (1973), é representado do seguinte modo

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

que é uma função em escada, com descontinuidades em cada instante de morte observado $t_{(i)}$. Deste modo, através da equação (1.7), a função de sobrevivência estimada associada ao vetor de covariáveis $\boldsymbol{z}$ para um dado indivíduos será então

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

Prova-se a consistência e a normalidade assintótica de $S(t; \boldsymbol{z})$, se as hipóteses adequadas no mecanismo de censura forem verificadas.

Vamos considerar que a distribuição do tempo de vida possui uma função de risco constante entre os instantes de morte sucessivos e que as observações censuradas que surgem entre os instantes $t_{(i)}$ e $t_{(i+1)}$ são censuradas em $t_{(i)}$. Então, o estimador da função de risco subjacente proposto por Breslow (1974)[4], no intervalo $]t_{(i-1)}, t_{(i)}]$ é descrito da seguinte forma

$$\widehat{h}_i = \frac{d_i}{[t_{(i)} - t_{(i-1)}] \sum_{l \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}.$$

O estimador da função de risco cumulativa, no instante t proposto por Breslow (1974) não necessita da utilização de métodos iterativos quando $d_i > 1$, para algum i é então

$$\widetilde{H}_0(t) = -\log \widetilde{S}_0(t) = \sum_{i: t_{(i)} \leq t} \frac{d_i}{\sum_{l \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}.$$

Logo, a correspondente função de sobrevivência será

$$\widetilde{S}_0(t) = \prod_{i: t_{(i)} \leq t} \left[\exp \left(\frac{-d_i}{\sum_{l \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)} \right) \right].$$

Então, $\widetilde{H}_0(t)$ representa o estimador de Nelson-Aalen e $\widehat{S}_0(t)$ o estimador de Kaplan-Meier.

1.10.7 Comparação de distribuições do tempo de vida

O modelo de Cox pode ser aplicado em situações em que se pretende testar a hipótese de igualdade das distribuições do tempo de vida para dois grupos de indivíduos contra a hipótese alternativa de que as distribuições são distintas (sendo as funções de risco proporcionais).

Vamos definir z como sendo uma covariável binária indicatriz, isto é,

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

Através de (1.7), as funções de sobrevivência S_1 e S_2 correspondentes aos dois grupos estão relacionadas através de

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

donde testar $H_0 : S_1(t) = S_2(t)$ é o mesmo que testar $H_0 : \beta = 0$.

Sejam $t_{(1)} < ... < t_{(k)}$ os instantes de morte distintos relativos aos $m+n$ indivíduos dos dois grupos e, para $j = 1,..,k$ e $i = 1,2$, seja ainda:

- d_j: número de mortes ocorridas em $t_{(j)}$
- d_{ij}: número de mortes ocorridas em $t_{(j)}$ no grupo i
- n_j: número de indivíduos em risco em $t_{(j)}$
- n_{ij}: número de indivíduos em risco em $t_{(j)}$ no grupo i

Sob o modelo de Cox, admitindo que existem poucas observações empatadas, e considerando $r_2 = \sum_{j=1}^{k} d_{2j}$, temos que

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

Logo,

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

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

Para testar $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ recorre-se a um teste bastante simples denominado teste score, que não requer o cálculo de $\widehat{\beta}$, em que a estatística de teste é obtida através

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

a qual, sob H_0, tem distribuição assintótica $N(0,1)$.

No caso de existir um número considerável de observações empatadas, deve-se usar um teste que tenha em conta a natureza discreta dos dados.

Esse teste é semelhante ao atrás descrito mas onde é necessário fazer uma correção à expressão de $I(0)$.

Uma vez que, sob H_0, Z^2 tem distribuição assintótica χ^2_1, este teste é equivalente ao teste log-rank, por vezes, denominado por teste de Cox-Mantel.

1.10.8 Método de seleção de variáveis

Numa análise de regressão pretende-se identificar quais as variáveis independentes que são importantes para prever a variável dependente (ou resposta). Com o modelo de Cox a situação é semelhante, pois pretende-se saber quais as covariáveis que são relevantes para prever o tempo de vida dos indivíduos.

Como o coeficiente β_j representa o efeito da covariável z_j na sobrevivência de um indivíduo, para podermos avaliar a existência de evidências de que essa covariável afeta de forma significativa o tempo de vida, é necessário testar as hipóteses

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

utilizando o teste de Wald, cuja estatística de teste $\widehat{\beta}_j^2 / var(\widehat{\beta}_j)$ apresenta, sob H_0, distribuição assintótica χ_1^2.

De referir, que neste caso, estamos a testar a hipótese de que a covariável z_j, na presença das restantes covariáveis, não tem influência significativa na sobrevivência do indivíduo. De uma forma geral, as estimativas $\widehat{\beta}$ não são independentes umas das outras, o que faz com que seja mais difícil a interpretação dos resultados de testes sobre os coeficientes associados a covariáveis que fazem parte de um modelo. Assim, é recomendado recorrer a métodos que admitam a comparação de modelos alternativos.

Seja um modelo de Cox com p covariáveis (modelo 1) e um modelo de Cox em que estão incluídas q covariáveis adicionais (modelo 2), isto é,

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

Pretende-se saber se os q termos adicionais incluídos no modelo 2 fazem com que este modelo seja significativamente melhor do que o modelo 1 (caso isso não se verifique, os q termos podem ser omitidos e o modelo 1 é considerado adequado).

Para tal, testam-se as hipóteses

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

A estatística de teste associada é $-2\log(\widehat{L}_1/\widehat{L}_2)$, onde $\widehat{L}_1$ e $\widehat{L}_2$ representam as estimativas de máxima verosimilhança para o grupo 1 e grupo 2, respetivamente, a qual tem distribuição assintótica χ_q^2. Este teste é designado por teste da razão das verosimilhanças dos modelos 1 e 2.

De modo a acautelar para os possíveis inconvenientes dos métodos de seleção automática de covariáveis realizados pelo *software* estatístico, Collett (2003) [6] propôs os seguintes passos para a seleção do modelo que melhor se ajusta aos dados:

1. Ajustamos todos os modelos que contêm apenas uma covariável. Calcula-se os valores da estatística $-2\log\widehat{L}$ para cada um dos modelos e compara-se com o valor da estatística para o modelo nulo (sem covariáveis). Desta forma, verificamos quais as covariáveis que, por si só, contribuíram para uma redução significativa do valor da estatística de teste e que podem influenciar o tempo de sobrevivência dos indivíduos.

2. Incluímos num só modelo as covariáveis que se revelaram ser potencialmente importantes na fase anterior (passo 1) e calcula-se o valor da estatística de teste $-2\log\widehat{L}$. Retira-se deste modelo uma covariável de cada vez, de forma a verificar se alguma deixou de ser relevante na presença das restantes. Retemos apenas no modelo as covariáveis que contribuíram para um aumento significativo do valor da estatística.

3. As variáveis que não foram consideradas no passo 2, pois quando consideradas isoladamente, não eram importantes, podem nesta fase revelar-se importantes na presença de outras. Assim, estas covariáveis passam então a ser incluídas, uma de cada vez, no modelo resultante do passo 2 e ficam apenas retidas no modelo as covariáveis que levaram a uma redução significativa do valor de $-2\log\widehat{L}$.

4. Deve-se fazer uma verificação final, de forma a garantir que nenhuma covariável pode ser omitida sem levar a um aumento significativo do valor de $-2\log\widehat{L}$ e que, nenhuma covariável não incluída leva a uma diminuição significativa do valor desta estatística.

Quanto ao nível de significância considerado para a inclusão ou omissão de covariáveis, este não deve ser muito pequeno, pelo que Collett (2003) sugere que se utilize o valor $\alpha \approx 0.1$.

1.10.9 Resíduos

Os resíduos são fundamentais para verificarmos se um modelo de regressão é adequado ou não. São definidos, no caso da regressão linear, como sendo a diferença entre o valor observado da variável resposta e o valor predito pelo modelo.

A existência de observações censuradas e a própria forma do modelo de Cox, leva a que não se possa fazer uma definição análoga para este modelo.

Por essa razão, a definição de resíduo é mais complexa e menos direta na modelação do tempo de vida, em relação a outros modelos de regressão.

Assim sendo, para o modelo de Cox foram definidos vários tipos de resíduos.

Nesta secção vamos abordar apenas dois: os resíduos de Cox-Snell e de Schoenfeld.

Resíduos de Cox-Snell

Para Cox e Snell (1968) [7], se o modelo de Cox é correto, então os resíduos devem comportar-se como uma amostra que advém de uma determinada distribuição conhecida. Este tipo de resíduos foram os primeiros a serem propostos para o modelo de Cox e são muito úteis na avaliação do ajustamento global do modelo final.

Pelo teorema da transformação uniformizante e tendo em conta a relação entre $S(t)$ e $H(t)$, vem que $H(t)$ tem distribuição exponencial de valor médio 1.

Por outro lado, como no modelo de Cox temos

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

então, o resíduo para o i-ésimo indivíduo, $i = 1,...,n$ será definido como sendo

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

onde $\widehat{\beta}$ e $\widehat{H}_0(t)$ são as estimativas de máxima verosimilhança parcial. Os valores estimados $\widehat{H}(t_i)$ terão propriedades semelhantes aos verdadeiros valores $H(t_i)$, apenas se o modelo que foi ajustado aos dados é satisfatório. Assim, os resíduos r_i devem comportar-se, aproximadamente, como uma amostra aleatória resultante de uma população com distribuição exponencial de valor médio 1.

Se uma determinada observação é censurada, então também o é o resíduo correspondente. Observemos, então, como se obtém os resíduos de Cox-Snell modificados.

Seja t_i^* uma observação censurada à direita e t_i o verdadeiro, embora desconhecido, tempo de vida desse indivíduo. O valor correto do resíduo para esse indivíduo seria $\widehat{H}(t_i)$, no entanto apenas podemos calcular $\widehat{H}(t_i^*)$. Uma vez que $t_i > t_i^*$, vem que $\widehat{H}(t_i) > \widehat{H}(t_i^*)$. Então, os resíduos de Cox-Snell modificados são obtidos através da adição de uma constante positiva, denominada por excesso residual:

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

O mais comum é considerar $\Delta = 1$.

Os resíduos de Cox-Snell diferem dos resíduos usados em regressão linear, uma vez que têm propriedades bastante distintas, nomeadamente não se distribuem de forma simétrica em torno do zero, nem podem tomar valores negativos.

Para averiguar a adequabilidade do modelo, verificamos se a amostra dos resíduos pode ser considerada como proveniente de uma população exponencial de valor médio um, isto é, de uma distribuição exponencial de valor médio. Uma das possibilidades possíveis consiste em fazer a representação gráfica dos pontos $\left(r'_i, \widetilde{H}(r'_i)\right)$, sendo $\widetilde{H}(r'_i)$ a estimativa de Nelson-Aalen da função de risco cumulativa dos resíduos. O modelo será adequado se a representação gráfica obtida for aproximadamente uma reta de declive um e ordenada na origem nula.

Resíduos de Schoenfeld

Os resíduos propostos por Schoenfeld (1982) [26], são úteis para a avaliação do pressuposto de proporcionalidade, após o ajustamento dos dados do modelo de Cox.

Estes resíduos diferem dos resíduos de Cox-Snell em dois aspetos:

- Não é necessário obter uma estimativa da função de risco cumulativa;
- A cada indivíduo não corresponde apenas um resíduo, mas sim um conjunto de valores, um por cada covariável que foi incluída no modelo de regressão de Cox.

No caso do *i*-ésimo indivíduo em estudo, o resíduo de Schoenfeld correspondente à covariável z_j, $j = 1, ..., p$ é

$$r_{ji} = \delta_i \left\{ z_{ji} - a_{ji} \right\}, \quad i = 1, ..., n$$

onde

$$a_{ji} = \frac{\sum\limits_{l \in R_i} z_{jl} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)}{\sum\limits_{l \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l)},$$

sendo o indicador de censura habitual definido por

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

Estes resíduos são sempre nulos, para indivíduos com tempo de vida censurado. Assim, os resíduos que correspondem a observações censuradas são indicados como valores omissos, de modo a poderem ser distinguidos dos resíduos que são verdadeiramente iguais a zero.

No caso de um indivíduo em que a morte foi observada em t_i, o resíduo será definido como sendo a diferença entre o valor da covariável z_j correspondente a esse indivíduo e uma média ponderada dos valores dessa covariável para todos os indivíduos em risco no instante t_i.

Caso o modelo que foi ajustado aos dados seja adequado, então o respetivo gráfico dos resíduos de Schoenfeld *versus* os tempos de vida ou *versus* as ordens dos tempos de vida, deverá ser uma nuvem aleatória de pontos, centrada em zero.

Alguns anos mais tarde, Grambsch e Therneau (1994) [29] propuseram uma nova versão destes resíduos, designados por resíduos de Schoenfeld padronizados, que afirmam ser mais eficaz em detetar afastamentos no modelo assumido.

1.10.10 Verificar a hipótese de riscos proporcionais

Existem vários métodos para verificar a validade da proporcionalidade das funções de risco, hipótese que é fundamental no modelo de Cox.

Método gráfico

Numa primeira fase começa-se por usar o método gráfico, antes do ajustamento do modelo. Através de (1.7), o modelo de Cox verifica a seguinte igualdade

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

logo

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

donde, para dois vetores de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, a distância entre $\log[-\log S(t;\boldsymbol{z}_1)]$ e $\log[-\log S(t;\boldsymbol{z}_2)]$ será constante e igual a $\boldsymbol{\beta}'\boldsymbol{z}$, o que quer dizer que as curvas do gráfico serão paralelas.

Consideremos os M subgrupos definidos por todas as combinações dos valores que as variáveis assumem, para p covariáveis fixas que assumem um pequeno número de valores. Por conseguinte, para efetuar uma validação gráfica da hipótese de riscos proporcionais deve ser obtida, para cada um das covariáveis, a estimativa de Kaplan-Meier da função de sobrevivência para cada um desses m grupos homogéneos de indivíduos, para $m = 1,...,M$ e fazer a representação gráfica de $\log[-\log \widehat{S}_m(t)]$ *versus* t num mesmo gráfico.

O modelo de Cox com as p covariáveis apenas poderá ser ajustado aos dados, se os gráficos forem razoavelmente paralelos. Geralmente, este gráfico é criado colocando $\log t$ no eixo das abcissas.

Podem surgir alguns problemas no método gráfico, nomeadamente se o número de indivíduos em cada subgrupo for muito pequeno. Além de que, quanto maior for o número de subgrupos mais complicado será a interpretação dos gráficos. Por essa razão, é frequente que a validade da hipótese de riscos proporcionais seja analisada, de forma separada, para cada covariável.

Aplicação dos resíduos de Schoenfeld

Um outro método utilizado, diz respeito aos resíduos de Schoenfeld padronizados, que são fundamentais para efetuar a avaliação do pressuposto de riscos proporcionais depois do ajustamento do modelo de Cox.

De acordo com Grambsch e Therneau (1994) [29], o valor médio do resíduo padronizado de Schoenfeld no instante t_i, para a covariável z_j, é

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

sendo $\beta_j(t_i)$ o coeficiente de z_j, que varia ao longo do tempo, calculado no instante de morte t_i e $\widehat{\beta}_j$ a estimativa de β_j no modelo de Cox já ajustado aos dados.

Será através de um gráfico dos valores de $r^*_{ji} + \widehat{\beta}_j$ *versus* tempo t, ou alguma função do tempo $g(t)$, que teremos informações sobre a forma do coeficiente $\beta_j(t)$, e, consequentemente, da forma como o efeito da covariável poderá depender o tempo.

No gráfico, se surgir uma linha horizontal significa que o coeficiente de z_j é constante e que é válida a hipótese de riscos proporcionais.

Assim, de forma a facilitar a interpretação é usual integrar no gráfico uma curva de suavização LOWESS com o respetivo intervalo de confiança, bem como, é recomendado a marcação de linhas horizontais de referência em zero e em $\widehat{\beta}_j$, de modo a permitir uma análise mais completa do gráfico.

Para complementar a observação do gráfico deve-se realizar um teste formal, baseado nos resíduos de Schoenfeld padronizados, proposto por Grambsch e Therneau (1994). O teste apresenta uma versão global e uma específica para cada uma das covariáveis.

Em relação à versão específica para cada covariável, notemos que se o modelo não for de riscos proporcionais, então o efeito das covariáveis não será constante ao longo do tempo. Logo pode ser escrito na forma

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

em que se escreve $\beta(t)$ como sendo uma regressão em $g(t)$, onde $\overline{g}_j$ representa a média dos $g_j(t_i)$ e $g_j(t_i)$ é uma função que é conhecida como, por exemplo, $\log t$. Então, testar a hipótese de riscos proporcionais corresponde a testar, para cada j, as hipóteses $H_0 : \theta_j = 0$ *versus* $H_1 : \theta_j \neq 0$.

No caso da versão global, testa-se simultaneamente que todos os θ_j são nulos.

A estatística de teste (Grambsch e Therneau, 1994) pode ser entendida como uma medida da correlação entre os resíduos associados a cada covariável e os tempos de vida, sendo que, sob H_0, apresenta uma distribuição χ^2_1.

Se a hipótese nula for rejeitada, então concluímos que existe evidência de que a correlação é não nula e, consequentemente, que existe evidência de não proporcionalidade.

1.11 Programa R e alguns pacotes de Análise de Sobrevivência

O R [30] é um sistema de computação científica e estatística, programável, ideal para análises estatísticas. Foi criado em 1993 por Ross Ihaka e por Robert Gentleman e trata-se de uma ferramenta *Open Source*, de acesso livre, estando disponível gratuitamente para todos os utilizadores. Existe uma comunidade ativa de investigadores que desenvolvem novas funcionalidades, de forma a atualizar e a ampliar as funcionalidades do sistema, tornando o R num ambiente mais desenvolvido, genérico e multifacetado ([20] e [23]).

Assim sendo, o R é um excelente aliado para extrair, transformar, analisar e apresentar dados, devido a sua vasta diversidade de técnicas gráficas e estatísticas, testes, gráficos com qualidade e precisão, entre outras funcionalidades, por isso, no caso específico da Análise de Sobrevivência, naturalmente que é uma mais-valia na obtenção de resultados, explorando todos os cenários estatísticos possíveis, com rigor e com qualidade. Existem outros *packages* que igualmente podem ser utilizados na Análise de Sobrevivência.

O R é compatível com diversas versões da *Microsoft Windows*, *Mac Os X*, *Linux*, entre outros e tem a possibilidade de instalar sempre que for necessário, novos pacotes (*packages*), alargando assim o seu leque de funções. Tem também a particularidade de estabelecer ligações com algumas interfaces, nomeadamente, com o *MS Excel*, o *Minitab*, o *SPSS* (*Statistical Package for the Social Sciences*), entre outros.

Apesar do R dispor de vários menus que possibilitam a realização de algumas operações, está orientado para a utilização de uma linha de comandos. Todavia, também oferece ao utilizador uma série de programas integrados de desenvolvimento que apresentam uma interface mais versátil e amigável, como por exemplo, o *R Commander* [15], facilitando assim a sua utilização.

Na Análise de Sobrevivência existe um *package* que é considerado o mais importante, o *survival* [31]. Este tipo de *package* permite efetuar análises descritivas, testes para duas amostras, aplicar o modelo de Cox, os modelos de tempo de vida acelerado (técnicas paramétricas) e *Case-cohort* (estudo de coorte). Se o *package survival* for utilizado com o *R Commander*, o acesso às funções dos menus torna-se muito mais fácil.

Também no âmbito da Análise de Sobrevivência existem três packages que são igualmente fundamentais e que também são *plug-ins* do *R Commander*, como o caso do *RcmdrPlugin.EZR (Easy R)* [14], o *RcmdrPlugin.KMggplot2* [21] e o *RcmdrPlugin.survival* [10].

Em relação ao *plug-in RcmdrPlugin.EZR*, este permite adicionar uma diversidade de funções estatísticas, das quais fazem parte a Análise de Sobrevivência, sobretudo no que diz respeito às meta-análises, ao cálculo da dimensão da amostra e às análises de curvas ROC. Para poder utilizar este*plug-in* é necessário efetuar a sua instalação no*R Commander*.

No que concerne ao *plug-in RcmdrPlugin.KMggplot2*, este é basicamente gráfico e por isso muito útil na construção de gráficos da estimativa de Kaplan-Meier da função de sobrevivência. No caso de construir gráficos mais elaborados, estes podem ser construídos através das funcionalidades do *package ggplot2*.

Por fim, no que diz respeito ao *RcmdrPlugin.survival*, este *plug-in* é uma extensão do que já tem disponível no *R Commander*, adicionando somente alguns novos itens aos menus que já existiam. Tem disponível itens próprios para o modelo de Cox, para o modelo de regressão paramétrico, para a estimação de curvas de sobrevivência. Permite um melhor manuseamento dos dados e testar as diferenças entre as curvas de sobrevivência, para além de efetuar outros testes, diagnósticos e construção de gráficos.

Anteriormente foram referidos alguns *packages* que podem ser utilizados na Análise de Sobrevivência, mas existe uma variedade de outros*packages* como por exemplo: *eha*; *KMsurv*; *relsurv*; *riskRegression*; *rms*; *smcure*; *survcomp*; entre outros.

Para esta dissertação o programa R, versão 64 bits 4.0.4. de 15 de fevereiro de 2021, será o instrumento principal que usarei, para poder efetuar todos os cálculos estatísticos necessários, com recurso ao programa integrado *R Commander*, versão 2.7-1. e aos *packages*: *RcmdrPlugin.survival* e *RcmdrPlugin.KMggplot2*.

Capítulo 2

Estudo de alguns CTeSP da UMa

Neste capítulo será feita uma introdução aos Cursos Técnicos e Superiores Profissionais (CTeSP), indicada a forma de recolha dos dados e a confidencialidade dos mesmos (ver **Anexo** A), a metodologia adotada e o tratamento e análise dos dados. Numa primeira fase será realizada a análise descritiva e, posteriormente, aplicada a Análise de Sobrevivência no estudo do tempo até à conclusão do CTeSP.

2.1 Enquadramento

A Universidade da Madeira é uma instituição que está atenta a eventuais problemáticas e na imagem que projeta para o exterior, tendo sempre em conta como preocupação central os estudantes.

Este estudo será mais um contributo nesta procura constante de melhoria contínua. Pensando em todos estes aspetos, este trabalho irá abordar os CTeSP que tiveram a sua primeira edição no ano letivo 2015/16, nomeadamente os cursos de: Agricultura Biológica (AB); Contabilidade e Fiscalidade (CF); Guias da Natureza (GN); e Redes e Sistemas Informáticos (R&S I).

Os CTeSP são formações de nível superior que concedem um diploma de técnico superior profissional que conferem qualificação profissional de nível 5 do Quadro Nacional de Qualificação (QNQ) e nível 5 do Quadro Europeu de Qualificações (QEQ). Têm duração de 2 anos letivos e um total de 120 ECTS (European Credit Transfer and Accumulation System, em português, Sistema Europeu de Transferência e Acumulação de Créditos).

Como os CTeSP estão associados às empresas e às necessidades do mercado de trabalho, o número de vagas depende da quantidade de instituições e empresas que acolhem os estagiários. Da mesma forma, a abertura dos CTeSP está condicionada ao número mínimo de candidatos selecionados, bem como, ao financiamento pelo Fundo Social Europeu, por intermédio do Programa Operacional Madeira 14-20 [16].

Os alunos que pretendem ingressar num CTeSP são dispensados da realização de exames nacionais e é-lhes garantido um estágio em contexto real de trabalho no último semestre, além de que, são cursos que conferem uma maior empregabilidade e possibilidade de prossecução noutras licenciatura (com creditação, na licenciatura em que ficar admitido, da formação realizada e nos termos da legislação em vigor). Os estudantes dos CTeSP também têm a possibilidade de candidatar-se para a aquisição de uma bolsa de estudo, bem como aos restantes apoios sociais previstos para todos os estudantes do ensino superior.

Podem candidatar-se a estes cursos titulares de um curso do ensino secundário ou de habilitação legitimamente equivalente; alunos maiores de 23 anos, após realização de provas que avaliam as capacidades do indivíduo para o curso que pretende frequentar, nos termos do Decreto-Lei n.º 64/2006, de 21 de março, alterado pelo Decreto-Lei n.º 113/2014, de 16 de julho [8] e o Decreto-Lei n.º 63/2016, de 13 de setembro [5]; titulares de um diploma de Curso de Especialização Tecnológica (CET), de um diploma de técnico superior profissional ou ainda, de um grau de ensino superior e por fim, os

estudantes que finalizam os cursos de formação profissional de nível secundário ou equivalente, sejam em escolas, sejam em outras entidades da rede do Politécnico da Universidade da Madeira [18] e [17].

O objetivo principal deste trabalho é o estudo do sucesso dos alunos que iniciaram algum dos CTeSP já referidos, entre os anos letivos 2015/2016 e 2018/2019, através da aplicação da Análise de Sobrevivência. Assim, a variável relevante é o tempo que os alunos precisam para concluir o seu curso.

2.2 Caracterização da amostra

Nesta secção, pretende-se evidenciar quais os procedimentos adotados para a seleção da amostra, bem como a sua caracterização, descrever os instrumentos e procedimentos utilizados para a recolha, aplicação, análise e tratamento de dados.

2.2.1 Metodologia

Este é um estudo prospetivo cujo objetivo é investigar o percurso escolar dos alunos que iniciaram algum dos CTeSP já referidos, entre os anos de 2015/2016 e 2018/2019. A base de dados contém 248 registos e os alunos foram seguidos até 28 de fevereiro de 2021, de modo que mesmo os alunos que iniciaram o curso em 2018/2019 fossem seguidos o tempo suficiente de modo a poderem terminar o curso no tempo regulamentar. Assim sendo, o período de *follow-up* foi de 30 de setembro de 2015 a 28 de fevereiro de 2021.

Pretende-se, acima de tudo, com este projeto de investigação, identificar os fatores que influenciam o sucesso dos alunos que frequentam os CTeSP.

2.2.2 Recolha de dados

De forma a assegurar a qualidade da informação e a coerência da mesma, os dados foram facultados por vários serviços da UMa, em especial pela Escola Superior de Tecnologia e Gestão.

A primeira inscrição nestes cursos ocorreu a 30 de setembro de 2015. Por esta razão, foi considerada esta data para o início do estudo. De referir que em 2016/17 a UMa decidiu não abrir o CTeSP em Agricultura Biológica, pelo que este curso tem apenas 3 edições no período considerado, ao contrário dos restantes que têm 4 (**Tabela 2.1**).

Tabela 2.1: Cronologia do estudo.

Cursos	**Ano letivo 2015/2016**	**Ano letivo 2016/2017**	**Ano letivo 2017/2018**	**Ano letivo 2018/2019**
CTeSP	AB CF GN R&S I	- CF GN R&S I	AB CF GN R&S I	AB CF GN R&S I

Os estudantes inseridos neste estudo foram anonimizados e integrados num estudo global pelo que a obtenção de um consentimento individualizado não foi considerada necessária. A informação recolhida destinou-se exclusivamente ao tratamento estatístico no âmbito deste estudo, respeitando o caráter confidencial dos dados.

Assim, foi elaborado um documento com o propósito de obter o consentimento do titular dos dados, com a informação do tipo de dados a aceder (ver **Anexo** A).

Desta forma, ficam salvaguardadas as regras relativas à proteção, tratamento e livre circulação dos dados pessoais, aplicáveis desde 25 de maio de 2018 e que revoga a Diretiva 95/46/CE (Regulamento

Geral sobre a Proteção de Dados - RGPD), no Regulamento (União Europeia - UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016 [32].

2.2.3 Variáveis recolhidas e variáveis utilizadas no estudo

Importa começar por distinguir os conceitos entre variáveis recolhidas e variáveis utilizadas no estudo. Variáveis recolhidas são aquelas que contem informações de interesse que foram recolhidas da população e com a possibilidade de sofrerem alterações, não sendo consideradas as variáveis finais da investigação, enquanto as variáveis utilizadas no estudo são as que contem informações de interesse que foram utilizadas efetivamente no estudo, sendo consideradas as variáveis finais da investigação.

Existem várias variáveis que direta ou indiretamente podem influenciar o percurso e o desempenho escolar dos alunos nos CTeSP, no entanto, de modo a contextualizar os indicadores mais pertinentes foram apenas consideradas três dimensões essenciais, com as respetivas variáveis (ver **Tabela B.1**):

- **Dimensão pessoal:** concelho de residência, data de nascimento, sexo, naturalidade e nacionalidade;
- **Dimensão social:** dados relativos aos pais, nomeadamente a habilitação literária (grau de escolaridade) e a profissão;
- **Dimensão escolar:** habilitação de acesso ao curso; cursos CTeSP; ano letivo de início e de fim do curso; data de início e de fim do curso; tempo para finalizar o curso (dias); o estado do aluno (se terminou ou não o curso); a situação do aluno (se desistiu/abandonou o curso, se mudou de curso, se ainda está frequentar o curso ou se já terminou o curso); fase de candidatura; se já foi aluno da UMa; o n.º de ECTS; se é trabalhador-estudante e se recebeu algum apoio social.

Tendo em conta os princípios de proteção, recolha e gestão de dados de acordo com o Regulamento Geral sobre a Proteção de Dados (RGPD) e de modo a proteger as informações pessoais dos alunos que de alguma forma possam identificar (direta ou indiretamente) o aluno, foi gerada uma amostra aleatória de valores inteiros de 1 a 248 (variável identificação na **Tabela C.1**) usando a função *sample* do programa R. Os alunos são identificados somente a partir destes valores.

Em relação à variável data de nascimento, no instrumento criado para registar os dados recolhidos nesta investigação, este tipo de informação não consta, nem mesmo nos resultados do estudo, sendo esta variável substituída pela variável idade (ver **Tabela C.1**). Este procedimento é mais uma salvaguarda para o anonimato dos dados recolhidos.

Vamos agora abordar numa primeira análise as variáveis recolhidas. Nos **Anexos** é apresentado uma tabela de síntese com a descrição de cada uma das 22 variáveis recolhidas (**Anexo**, **Tabela B.1**).

A partir destas variáveis, foram construídas as variáveis utilizadas no estudo. Das variáveis recolhidas foram retiradas quatro variáveis: a nacionalidade, tendo em conta que a larga maioria dos alunos apresentava nacionalidade portuguesa (95,56%), a data de início e de data de fim do curso, uma vez que estas variáveis já foram contempladas no cálculo da variável tempo para finalizar o curso (dias) e, por fim, a variável razões para ainda não ter terminado o curso, uma vez que não foi possível recolher este tipo de informação. De referir que, no caso da variável ano letivo de fim do curso, apenas foi retirada a categoria 2020/2021, uma vez que devido à situação pandémica da COVID-19, foi decidido prolongar os prazos de aulas/avaliações, fazendo com que os três alunos que estavam contemplados nestas categorias, fossem transferidos para a categoria 2019/2020. Com efeito, a categoria 2020/2021 deixa de ser contemplada nesta variável.

Das variáveis recolhidas, foram modificadas oito variáveis do estudo: o concelho de residência, a data de nascimento, a naturalidade, a habilitação literária do pai e da mãe, a profissão do pai e da mãe e a fase de candidatura.

Em relação à primeira variável, decidiu-se reduzir o número de categorias, uma vez que existiam vários concelhos de residência com um número muito reduzido de alunos ou até mesmo sem alunos.

No caso da segunda variável, data de nascimento, optou-se por efetuar este procedimento não apenas para facilitar a própria análise estatística, mas também para salvaguardar a divulgação de um dado considerado sensível. Assim sendo, a data de nascimento dos alunos passa a ser substituída pela variável idade do aluno no momento do ato da matrícula do curso.

No que diz respeito à terceira variável, após a análise dos dados recolhidos verificou-se que existiam demasiadas categorias, dispersando assim muita informação, pelo que se optou por reduzir esse número, de modo a recolher com uma maior precisão as informações de interesse. Assim, a naturalidade passa apenas a ter cinco categorias. Esta decisão foi igualmente aplicada no caso das variáveis habilitação literária do pai e da mãe (sete categorias) e no caso também da variável profissão do pai e da mãe (seis categorias).

No que concerne à variável fase de candidatura, foi necessário reduzir o número de categorias que inicialmente haviam sido propostas. Assim, em vez de três categorias (1.ª fase, 2.ª fase e 3.ª fase), esta variável passou a ter duas (1.ª fase e 2.ª e 3.ª fases).

Por outro lado, numa segunda análise e, após observar os dados recolhidos, foram determinadas quais as variáveis que serão efetivamente utilizadas no estudo. Seguidamente serão descritas cada uma dessas variáveis:

- **Identificação** - gerada no programa R a partir do comando sample de modo a não poder identificar concretamente cada dos alunos alvo do estudo, obtendo-se desta forma uma amostra aleatória de valores inteiros de 1 a 248.

- **Concelho de residência** - permite recolher informações acerca do concelho de residência do aluno. Inicialmente esta variável apresentava 12 categorias, tendo sido reduzida para apenas quatro:

    1. Funchal, Câmara de Lobos e Santa Cruz mantiveram-se inalteradas;
    2. Calheta, Machico, Ponta do Sol, Porto Moniz, Porto Santo, Ribeira Brava, Santana, São Vicente e Fora da RAM formaram a categoria Restantes concelhos.

- **Sexo** - variável dicotómica que classifica o sexo do aluno, em masculino ou feminino.

- **Idade (anos)** - definida em anos e obtida através da data de inscrição do aluno e da sua data de nascimento.

- **Naturalidade** - identifica o local de nascimento do aluno. Inicialmente apresentava doze categorias, sendo restringida posteriormente para apenas cinco:

    1. Calheta (Arco da Calheta e Ponta do Pargo), Câmara de Lobos (Estreito de Câmara de Lobos e Jardim da Serra e Quinta Grande), Funchal (Sé, São Martinho, Santo António, Monte, São Pedro, São Gonçalo, Imaculado Coração de Maria, Santa Maria Maior, Santa Luzia e São Roque), Machico (Porto da Cruz, Caniçal, Santo António da Serra e Água de Pena), Ponta do Sol (Canhas), Porto Moniz, Porto Santo, Ribeira Brava (Campanário), Santa Cruz (Gaula, Camacha e Caniço), Santana (Faial e São Jorge) e São Vicente formaram a categoria RAM;
    2. Lisboa, Porto, Viana do Castelo, Vila Real, Açores formaram a categoria Portugal (exceto RAM);
    3. Moçambique, África do Sul e Guiné-Bissau formaram a categoria África;
    4. Brasil e Venezuela formaram a categoria América do Sul;
    5. Espanha, Reino Unido, Itália e Alemanha formaram a categoria Europa (exceto Portugal).

- **Habilitação literária dos pais** - permite recolher informações acerca das habilitações literárias dos pais do aluno. Inicialmente esta variável apresentava quinze categorias, sendo posteriormente restringida para sete, tendo em conta que algumas apresentavam um número reduzido de observações. Assim, congregamos os dados num menor número de categorias:

    1. Não sabe ler nem escrever/sem escolaridade e Sabe ler e escrever (sem ter o 4.º ano) formaram a categoria Sem escolaridade;
    2. 4.º ano de escolaridade (antiga 4.ª classe), 6.º ano de escolaridade, 9.º ano de escolaridade e 12.º ano de escolaridade mantiveram-se inalteradas;
    3. Ensino Médio (Antigo Magistério), Curso sem Grau, Curso de Especialização Tecnológica, Bacharelato, Licenciatura, Mestrado e Doutoramento formaram a categoria Ensino pós-secundário (superior e não superior);
    4. Desconhecido e Outra formaram a categoria Outra.

- **Profissão dos pais** - permite recolher informações acerca da profissão dos pais do aluno. Inicialmente esta variável apresentava dez categorias, sendo posteriormente reduzida a seis, pelos mesmos motivos que nas habilitações literárias dos pais:

    1. Desempregado(a), Doméstico(a), Aposentado(a)/Reformado(a), Trabalhador por Conta de Outrem mantiveram-se inalteradas;
    2. Trabalhador por Conta Própria como Empregador e Trabalhador por Conta Própria como Isolado formaram a categoria Trabalhador por Conta Própria;
    3. Estudante, Serviço Militar, Trabalhador Familiar Não Renumerado e Outra situação, formaram a categoria Outra situação.

- **Habilitação de acesso ao curso** - habilitação de acesso dos alunos aos CTeSP's. Das categorias apresentadas (Cursos Científico-Humanísticos, Cursos de Educação e Formação Profissional, Cursos de Especialização Tecnológica (CET), Formação superior e Outra (Ensino Recorrente, etc.), apenas a categoria Outra foi substituída para Informação Omissa.

- **Cursos CTeSP** - identifica os quatro cursos alvo deste estudo. Está categorizada da seguinte forma: Agricultura Biológica (AB), Contabilidade e Fiscalidade (CF), Guias da Natureza (GN) e Redes e Sistemas Informáticos (R&S I).

- **Ano letivo de início do curso** - anos letivos de ingresso dos alunos: 2015/2016, 2016/2017, 2017/2018 e 2018/2019.

- **Ano letivo de fim do curso** - anos letivos de finalização do curso pelos alunos: 2016/2017, 2017/2018, 2018/2019, 2019/2020 e Ainda não terminou. Foi retirada a categoria 2020/2021, passando a fazer parte desta variável apenas cinco categorias.

- **Tempo para finalizar o curso (dias)** - obtido com base na diferença entre a data de início do curso (coincide com a data da ficha de inscrição) e a data de fim do curso (coincide com a data da nota do estágio).

- **Estado do aluno** - variável dicotómica que indica se o aluno terminou (=1) ou não (=0) o curso.

- **Situação do aluno** - representa a situação ou estado atual do aluno no fim do período de observação e apresenta como categorias: Desistiu/Abandonou, Mudou de curso, Em curso e Terminou. Nesta variável, consideramos que o aluno desistiu/abandonou desde que tenha estado um ano matriculado e não dois.

- **Fase de candidatura** - corresponde às diferentes fases de candidatura aos cursos CTeSP. Inicialmente estava categorizada em três fases: 1.ª fase, 2.ª fase e 3.ª fase, passando apenas a apresentar duas categorias: 1.ª fase e 2.ª e 3.ª fases.

- **Aluno da UMa (ex-aluno)** - variável dicotómica que avalia se o aluno já foi (=1) ou não (=0) aluno da UMa.

- **№ de ECTS** - créditos obtidos pelo estudante na aprovação de uma disciplina (neste caso de um curso) nos termos do Decreto-Lei n.º 42/2005 [19]. O mínimo é 0 ECTS que significa que o aluno desistiu, abandonou ou transferiu de curso antes de ter aproveitamento em alguma disciplina e o máximo será 120 ECTS, que significa que o aluno finalizou o curso.

- **Trabalhador-estudante** - variável dicotómica que avalia se o aluno possui (=1) ou não (=0) o estatuto de trabalhador-estudante.

- **Apoio social** - variável dicotómica que avalia se o aluno recebeu algum apoio social (=1) (bolsa de estudo, ...) ou não (=0). Caso os alunos obtenham bolsa nos dois anos letivos do CTeSP, essa bolsa é contabilizada apenas no 1.º ano, isto é, no ano em que entraram para o curso.

Nos **Anexos** é apresentado uma tabela de síntese com a descrição de cada uma das 19 variáveis utilizadas neste estudo (**Tabela C.1**).

2.2.4 Tratamento e análise de dados

Após a obtenção da informação solicitada, esta foi registada com auxílio do *software Microsoft Excel* para *Microsoft* 365 MSO 64 bits. Através do programa R versão 64 bits 4.0.4. de 15 de fevereiro de 2021 [30] foi elaborado toda a análise e tratamento estatístico. Após tratamento dos dados, foi feita uma compilação de todos os dados recorrendo ao *software TeXstudio* 2.12.18 (gif 2.12.18) ([28] e [13]).

2.3 Análise descritiva dos dados

A amostra em estudo é constituída pelos 248 alunos que realizaram a sua primeira inscrição nos CTeSP da Universidade da Madeira, nomeadamente nos quatro cursos: Agricultura Biológica, Contabilidade e Fiscalidade, Guias da Natureza e Redes e Sistemas Informáticos, entre o ano letivo 2015/2016 (ano de abertura dos CTeSP na UMa) e o ano 2018/2019. O fim do estudo ocorreu a 28 de fevereiro de 2021. Todos os alunos foram seguidos por um período mínimo de 2 anos letivos de modo que tivessem, no mínimo, o período regulamentar para realizar o seu curso.

Antes de falarmos de candidatos, alunos inscritos/matriculados e alunos diplomados, convém definir os conceitos que os distinguem. Assim:

- Candidatos: Indivíduos que preencheram a ficha de candidatura de acesso aos CTeSP.

- Alunos inscritos/matriculados: Alunos que têm pelo menos um ano de inscrição.

- Alunos diplomados: Alunos que finalizaram o seu curso.

Durante o período de tempo do ano letivo 2015/2016 ao ano letivo 2018/2019 candidataram-se aos CTeSP da UMa 391 indivíduos. Na **Tabela 2.2** verifica-se que os candidatos do sexo Masculino (M) (66,50%) predominam em relação aos candidatos do sexo Feminino (F) (33,50%). O ano letivo 2018/2019 foi o que registou mais candidatos (30,95%). Por outro lado, o ano letivo 2015/2016 foi onde se verificou, no Total (T), menos candidatos (19,69%).

Tabela 2.2: Distribuição dos alunos candidatos aos CTeSP, por sexo e ano letivo (% linha).

CTeSP	2015/2016			2016/2017			2017/2018		
	F	M	T	F	M	T	F	M	T
AB	4	15	19	9	9	18	-	-	-
%	6,35	23,81	30,16	14,29	14,29	28,57	-	-	-
CF	13	10	23	10	11	21	14	15	29
%	12,62	9,71	22,33	9,71	10,68	20,39	13,59	14,56	28,16
GN	5	10	15	20	12	32	17	15	32
%	4,63	9,26	13,89	18,52	11,11	29,63	15,74	13,89	29,63
R&S I	1	19	20	3	36	39	1	21	22
%	0,85	16,24	17,09	2,56	30,77	33,33	0,85	17,95	18,80
Total	23	54	77	42	68	110	32	51	83
%	5,88	13,81	19,69	10,74	17,39	28,13	8,18	13,04	21,23

CTeSP	2018/2019			Total		
	F	M	T	F	M	T
AB	6	20	26	19	44	63
%	9,52	31,75	41,27	30,16	69,84	100
CF	20	10	30	57	46	103
%	19,42	9,71	29,13	55,34	44,66	100
GN	7	22	29	49	59	108
%	6,48	20,37	26,85	45,37	54,63	100
R&S I	1	35	36	6	111	117
%	0,85	29,91	30,77	5,13	94,87	100
Total	34	87	1211	131	260	391
%	8,70	22,25	30,95	33,50	66,50	100

Numa outra perspetiva, na **Tabela 2.3** observou-se que, no conjunto dos quatro anos letivos, existem mais indivíduos candidatos ao CTeSP em Redes e Sistemas Informáticos (29,92%) do que nos restantes. Os candidatos do sexo feminino escolheram preferencialmente o CTeSP em Contabilidade e Fiscalidade (43,51%) e o que escolheram menos foi o CTeSP em Redes e Sistemas Informáticos (4,58%). Relativamente aos candidatos do sexo masculino, escolheram preferencialmente o CTeSP em Redes e Sistemas Informáticos (42,69%) e menos o CTeSP em Agricultura Biológica (16,92%).

Na **Tabela 2.4** verifica-se que o número total de alunos inscritos/matriculados do sexo masculino (66,53%) predomina em relação ao número total de alunos inscritos/matriculados do sexo feminino (33,47%), com percentagens muito próximas das verificadas para os candidatos. O ano letivo 2015/2016 foi o que registou mais alunos inscritos/matriculados (28,63%), apesar de não ter sido o ano com mais candidatos. Por outro lado, o ano letivo 2017/2018 foi o que registou menos alunos inscritos/matriculados (18,95%).

De referir que embora tenham sido contabilizados 9 alunos inscritos/matriculados no CTeSP em Agricultura Biológica, no ano letivo 2016/2017, inicialmente foram 15 alunos (mínimo estipulado para cada curso), no entanto, este valor foi reduzido devido a alunos que decidiram abandonar o curso.

Tabela 2.3: Distribuição dos candidatos aos CTeSP, por sexo e ano letivo (% coluna).

CTeSP	2015/2016			2016/2017			2017/2018		
	F	M	T	F	M	T	F	M	T
	%	%	%	%	%	%	%	%	%
AB	4	15	19	9	9	18	-	-	-
	17,39	27,78	24,68	21,43	13,24	16,36	-	-	-
CF	13	10	23	10	11	21	14	15	29
	56,52	18,52	29,87	23,81	16,18	19,09	43,75	29,41	34,94
GN	5	10	15	20	12	32	17	15	32
	21,74	18,52	19,48	47,62	17,65	29,09	53,13	29,41	38,55
R&S I	1	19	20	3	36	39	1	21	22
	4,35	35,19	25,97	7,14	52,94	35,45	3,13	41,18	26,51
Total	23	54	77	42	68	110	32	51	83
%	100	100	100	100	100	100	100	100	100

CTeSP	2018/2019			Total		
	F	M	T	F	M	T
	%	%	%	%	%	%
AB	6	20	26	19	44	63
	17,65	22,99	21,49	14,50	16,92	16,11
CF	20	10	30	57	46	103
	58,82	11,49	24,79	43,51	17,69	26,34
GN	7	22	29	49	59	108
	20,59	25,29	23,97	37,40	22,69	27,62
R&S I	1	35	36	6	111	117
	2,94	40,23	29,75	4,58	42,69	29,92
Total	34	87	121	131	260	391
%	100	100	100	100	100	100

Tabela 2.4: Distribuição de alunos inscritos/matriculados nos CTeSP, por sexo e ano letivo (% linha).

CTeSP	2015/2016			2016/2017			2017/2018		
	F	M	T	F	M	T	F	M	T
AB	2	14	16	4	5	9	-	-	-
%	5,26	87,5	42,11	10,53	13,16	23,68	-	-	-
CF	12	9	21	10	9	19	9	6	15
%	15,38	42,9	26,92	12,82	11,54	24,36	11,54	7,69	19,23
GN	4	10	14	9	5	14	8	6	14
%	7,02	71,4	24,56	15,79	8,77	24,56	14,04	10,53	24,56
R&S I	1	19	20	2	18	20	1	17	18
%	1,33	25,33	26,67	2,67	24	26,67	1,33	22,67	24
Total	19	52	71	25	37	62	18	29	47
%	7,66	20,97	28,63	10,08	14,92	25	7,26	11,69	18,95

CTeSP	2018/2019			Total		
	F	M	T	F	M	T
AB	1	12	13	7	31	38
%	2,63	31,58	34,21	18,42	81,58	100
CF	15	8	23	46	32	78
%	19,23	10,26	29,49	58,97	41,03	100
GN	5	10	15	26	31	57
%	8,77	17,54	26,32	45,61	54,39	100
R&S I	0	17	17	4	71	75
%	0	22,67	22,67	5,33	94,67	100
Total	21	47	68	83	165	248
%	8,47	18,95	27,42	33,47	66,53	100

Analisando os dados numa outra vertente, podemos verificar através da **Tabela 2.5** que são contabilizados mais alunos inscritos/matriculados no CTeSP em Contabilidade e Fiscalidade (31,45%) e menos no CTeSP em Agricultura Biológica (15,32%). Contudo, esta última observação seria expectável, uma vez que este curso não abriu no ano letivo 2017/2018. Também se constatou que os alunos inscritos/matriculados do sexo feminino escolheram preferencialmente o CTeSP em Contabilidade e Fiscalidade (55,42%) e o que escolheram menos foi o CTeSP em Redes e Sistemas Informáticos (4,82%). Relativamente aos alunos inscritos/matriculados do sexo masculino, escolheram preferencialmente o CTeSP em Redes e Sistemas Informáticos (43,03%) e menos o CTeSP em Agricultura Biológica e o CTeSP em Guias da Natureza (18,79%, respetivamente).

Tabela 2.5: Distribuição de alunos inscritos/matriculados nos CTeSP, por sexo e ano letivo (% coluna).

CTeSP	2015/2016			2016/2017			2017/2018		
	F	M	T	F	M	T	F	M	T
	%	%	%	%	%	%	%	%	%
AB	2	14	16	4	5	9	-	-	-
	10,53	26,92	22,54	16	13,51	14,52	-	-	-
CF	12	9	21	10	9	19	9	6	15
	63,16	17,31	29,58	40	24,32	30,65	50	20,69	31,91
GN	4	10	14	9	5	14	8	6	14
	21,05	19,23	19,72	36	13,51	22,58	44,44	20,69	29,79
R&S I	1	19	20	2	18	20	1	17	18
	5,26	36,54	28,17	8	48,65	32,26	5,56	58,62	38,30
Total	19	52	71	25	37	62	18	29	47
%	100	100	100	100	100	100	100	100	100

CTeSP	2018/2019			Total		
	F	M	T	F	M	T
	%	%	%	%	%	%
AB	1	12	13	7	31	38
	4,76	25,53	19,12	8,43	18,79	15,32
CF	15	8	23	46	32	78
	71,43	17,02	33,82	55,42	19,39	31,45
GN	5	10	15	26	31	57
	23,81	21,28	22,06	31,33	18,79	22,98
R&S I	0	17	17	4	71	75
	0	36,17	25	4,82	43,03	30,24
Total	21	47	68	83	165	248
%	100	100	100	100	100	100

Observando o número de vagas disponíveis (V) em relação ao número de candidatos (C) verificamos através da **Tabela 2.6** que desde o ano letivo 2015/2016 até ao ano letivo 2018/2019, tem-se verificado uma tendência evolutiva muito positiva. O CTeSP em Guias da Natureza é o curso que apresenta uma maior afluência de candidatos face ao número de vagas disponíveis (180%). Por outro lado, o CTeSP em Agricultura Biológica é o que apresenta uma menor adesão de candidatos (105%).

Tabela 2.6: Distribuição comparativa do número de vagas disponíveis com o número de candidatos aos CTeSP.

CTeSP	2015/2016			2016/2017			2017/2018		
	V	C	%	V	C	%	V	C	%
AB	20	19	95	20	18	90	-	-	-
CF	21	23	109,52	21	21	100	21	29	138,10
GN	15	15	100	15	32	213,33	15	32	213,33
R&S I	20	20	100	20	39	195	20	22	110
Total	76	77	101,32	76	110	144,74	56	83	148,21

CTeSP	2018/2019			Total		
	V	C	%	V	C	%
AB	20	26	130	60	63	105
CF	24	30	125	84	103	122,62
GN	18	29	161,11	60	108	180
R&S I	20	36	180	80	117	146,25
Total	76	121	159,21	284	391	137,68

Numa outra perspetiva (ver **Tabela** 2.7), no que concerne ao número de vagas (V) e os alunos inscritos/matriculados (AI), podemos verificar que durante o período de observação, a nível geral, $87,32\%$ das vagas disponíveis foram preenchidas. O ano letivo 2015/2016 foi o que verificou uma maior taxa de ocupação das vagas ($93,42\%$) e com menor o ano letivo 2016/2017 ($81,58\%$). Ao nível dos cursos, o CTeSP em Redes e Sistemas Informáticos foi o que apresentou uma maior taxa de ocupação das vagas ($93,75\%$) e com menor o CTeSP em Agricultura Biológica ($63,33\%$).

Tabela 2.7: Distribuição comparativa do número de vagas disponíveis com o número de alunos inscritos/matriculados nos CTeSP.

CTeSP	2015/2016			2016/2017			2017/2018		
	V	AI	%	V	AI	%	V	AI	%
AB	20	16	80	20	9	45	-	-	-
CF	21	21	100	21	19	90,48	21	15	71,43
GN	15	14	93,33	15	14	93,33	15	14	93,83
R&S I	20	20	100	20	20	100	20	18	90
Total	76	71	93,42	76	62	81,58	56	47	83,93

CTeSP	2018/2019			Total		
	V	AI	%	V	AI	%
AB	20	13	65	60	38	63,33
CF	24	23	95,83	87	78	89,66
GN	18	15	83,33	63	57	90,48
R&S I	20	17	85	80	75	93,75
Total	76	68	89,47	284	248	87,32

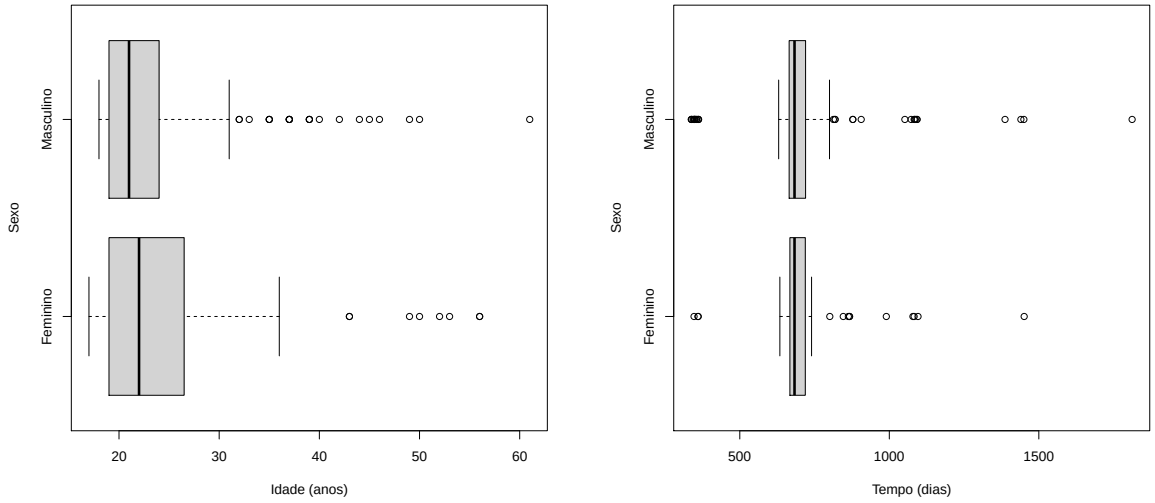
Tendo em conta a **Tabela** D.1 disponível nos **Anexos**, verifica-se que, em relação ao concelho de residência dos alunos, a maior percentagem de alunos inscritos reside no Funchal ($46,77\%$) e a menor nos restantes concelhos (exceto Câmara de Lobos e Santa Cruz) ($24,19\%$).

Analisando a variável idade, constata-se que a média de idades dos alunos é 24,05 anos, com um desvio padrão de 8,18 anos, amplitude interquartil de 6 anos (IQR - *Interquatile Range*), idade mínima de 17 anos, mediana de 21 anos e a idade máxima de 61 anos.

Note-se que 25% dos alunos apresenta idade inferior a 19 anos e 75% apresenta idade inferior a 25 anos.

Cruzando esta informação com a variável sexo verificamos que a idade média dos alunos do sexo feminino é 24,99 anos, com um desvio padrão 9,58 anos. A idade mínima dos alunos do sexo feminino é 17 anos, a idade mediana 22 anos e a idade máxima 56 anos.

De referir que 25% dos alunos do sexo feminino apresenta idade inferior a 19 anos e 75% idade inferior a 26,5 anos. No que diz respeito à idade média dos alunos do sexo masculino, esta é de 23,58 anos, com um desvio padrão de 7,37 anos. A idade mínima dos alunos do sexo masculino é 18 anos, a idade mediana 21 anos e a idade máxima 61 anos. Verifica-se que 25% dos alunos do sexo masculino apresenta idade inferior a 19 anos e 75% idade inferior a 24 anos (**Figura 2.1a**).



(a) **(b)**

Figura 2.1: Distribuição da idade dos alunos inscritos/matriculados por sexo (a) e distribuição do tempo dispensado para finalizar o curso por sexo (b).

Ao analisar a variável tempo despendido para finalizar o curso constata-se que o tempo médio despendido para o efeito é 707,27 dias, com um desvio padrão de 173,52 dias e amplitude interquartil de 55 dias. O tempo mínimo despendido é 338 dias, a mediana 683 dias e o tempo máximo despendido é 1812 dias. Note-se que 25% dos alunos finalizaram o curso em menos de 665 dias e 75% em menos de 720 dias.

Cruzando esta informação com a variável sexo verificamos que o tempo médio despendido pelos alunos do sexo feminino para finalizar o curso é de 712,7952 dias, com um desvio padrão aproximadamente igual a 139,95 dias e amplitude interquartil de 51,5 dias. O tempo mínimo despendido é 347 dias, a mediana é de 683 dias e o tempo máximo é de 1451 dias.

Note-se que 25% dos alunos do sexo feminino finalizam o curso em menos de 667,5 dias e 75% em menos de 719 dias. No que diz respeito ao tempo médio despendido pelos alunos do sexo masculino para finalizar o curso é de 704,49 dias, com um desvio padrão igual a 188,5 dias e amplitude interquartil de 55 dias. O tempo mínimo é 338 dias, a a mediana 683 dias e o tempo máximo 1812 dias.

Verifica-se que 25% dos alunos do sexo masculino finalizam o curso em menos de 665 dias e 75% em menos de 720 dias (**Figura 2.1b**).

Quanto aos valores mencionados anteriormente para o tempo mínimo despendido pelos alunos, estes valores são explicados pela presença de *outliers* inferiores correspondentes a alunos que terminaram o curso antes do tempo, dado virem de CET's e solicitarem a creditação de unidades curriculares

desses cursos. Como tal, fizeram o curso não só em menos tempo que os restantes alunos, como também em tempo inferior à duração normal do curso (2 anos).

Da representação gráfica anterior verifica-se que, em ambos os sexos, os dados distribuem-se de forma assimétrica (à direita), revelando a existência de alguns valores discrepantes/atípicos (*outliers*). Estes valores correspondem a alguns alunos com idades bastante superiores aos restantes. Os *outliers* acima do limite superior significam a existência de alunos que demoraram mais tempo a finalizar o curso, num período de tempo superior a 2 anos letivos.

Também foi possível apurar na **Tabela D.1** disponível nos **Anexos**, que relativamente à variável naturalidade a grande maioria dos alunos é natural da Região Autónoma da Madeira (81,05%). No entanto, observaram-se várias origens, nomeadamente, África (Moçambique (3), África do Sul (2) e Guiné-Bissau (1)) e Europa (Itália (2), Reino Unido (2), Alémanha (1) e Espanha (1)).

Em relação à variável habilitação literária do pai e da mãe dos alunos verifica-se que prevalece em ambos o 4.º ano de escolaridade (antiga 4.ª classe), sendo que este valor é um pouco superior no caso do pai (33,47%) relativamente à mãe (25,40%).

Quanto às variáveis Profissão dos pais, de acordo com os resultados obtidos ainda na **Tabela D.1** nos **Anexos**, tanto o pai como a mãe dos alunos inscritos/matriculados apresentam uma maior percentagem no cargo de Trabalhador por Conta de Outrem comparativamente às restantes categorias destas variáveis, observando-se uma maior expressividade no que se refere às mães dos alunos (43,55%).

Observou-se, da mesma forma que, comparando as percentagens dos pais com as mães dos alunos para o mesmo cargo, os pais são os que apresentam maior percentagem de Desempregados (18,15%), Aposentados/Reformados (10,89%) e Trabalhadores por Conta Própria (7,66%), sendo que, para o cargo de Doméstico, quase um sexto das mães são domésticas, não sendo detetados casos em que o pai seja Doméstico.

Relativamente à habilitação de acesso ao curso, foi possível apurar que cerca de dois terços dos alunos finalizaram o ensino secundário através da modalidade Educação e Formação Profissional (66,53%), 26,21% através dos Cursos Científico-Humanísticos, 4,44% dos alunos já apresentava uma formação superior e, por fim, 2,42% através dos Cursos de Especialização Tecnológica (CET). De referir que esta informação está omissa num aluno (0,4%) pois, apesar de se saber que detém o 12.º ano, não foi possível identificar se era proveniente de um Curso Científico-Humanístico ou de um Curso de Educação e Formação Profissional.

De acordo com a **Tabela D.1** disponível nos **Anexos**, verificou-se que o ano letivo 2016/2017, de todos os anos letivos de fim de curso, foi o ano que se detetou mais alunos diplomados (22,58%), sendo que este valor tem vindo a diminuir nos anos letivos seguintes. Todavia, esta tendência poderá ser modificada/alterada se alguns dos 19,76% dos alunos que ainda não terminaram o curso, finalizarem ainda no final deste ano letivo 2020/2021. Os três alunos que se diplomaram, no ano letivo 2020/2021, isto é, que conseguiram finalizar o curso até dia 28/02/2021, foram contemplados na categoria 2019/2020, uma vez que devido à situação epidemiológica do novo Coronavírus – COVID 19, os prazos das avaliações foram alargados.

Relativamente à variável estado do aluno, até à data final da recolha dos dados (fim do *follow-up*), verificou-se que a grande maioria dos alunos terminou o curso (80,24%).

Um outro aspeto que também importa referir é que, em termos gerais, verificou-se através da **Tabela 2.8** que os alunos inscritos/matriculados no ano letivo 2015/2016 são os que apresentam mais alunos diplomados (85,92%) comparativamente aos alunos inscritos/matriculados no ano letivo 2018/2019 (72,06%). No entanto, é importante salientar que os alunos do ano letivo 2018/2019 estiveram a ser seguidos apenas em dois anos, ao contrário dos restantes que foram seguidos durante, pelo menos, 3 anos.

Tabela 2.8: Relação do número de alunos inscritos/matriculados com o número de alunos diplomados.

	2015/2016		2016/2017		2017/2018		2018/2019		Total	
	№	%	№	%	№	%	№	%	№	%
Alunos inscritos/ matriculados	71	100	62	100	47	100	68	100	248	100
Alunos diplomados	61	85,92	52	83,87	37	78,72	49	72,06	199	80,24

Numa outra perspetiva podemos verificar através da **Tabela 2.9** que o CTeSP em Redes e Sistemas Informáticos é o curso que mais contribui para a percentagem de alunos diplomados (32,66%) e o que menos contribui é o CTeSP em Agricultura Biológica (13,57%). No entanto, uma vez que este curso não abriu no ano letivo 2017/2018 é compreensível existir menos alunos inscritos/matriculados e esse facto influenciar o resultado anterior.

Tabela 2.9: Distribuição dos alunos diplomados (S-Sim, N-Não e T-Total), por curso e por ano letivo (% coluna).

CTeSP	2015/2016			2016/2017			2017/2018		
	S	N	T	S	N	T	S	N	T
	%	%	%	%	%	%	%	%	%
AB	12	4	16	8	1	9	-	-	-
	19,67	40	22,54	15,38	10	14,52	-	-	-
CF	20	1	21	14	5	19	10	5	15
	32,79	10	29,58	26,92	50	30,65	27,03	50	31,91
GN	11	3	14	12	2	14	11	3	14
	18,03	30	19,72	23,08	20	22,58	29,73	30	29,79
R&S I	18	2	20	18	2	20	16	2	18
	29,51	20	28,17	34,62	20	32,26	43,24	20	38,30
Total	61	10	71	52	10	62	37	10	47
%	100	100	100	100	100	100	100	100	100

CTeSP	2018/2019			Total		
	S	N	T	S	N	T
	%	%	%	%	%	%
AB	7	6	13	27	11	38
	14,29	31,58	19,12	13,57	22,45	15,32
CF	20	3	23	64	14	78
	40,82	15,79	33,82	32,16	28,57	31,45
GN	5	10	15	26	31	57
	18,37	31,58	22,06	21,61	28,57	22,98
R&S I	13	4	17	65	10	75
	26,53	21,05	25	32,66	20,41	30,24
Total	49	19	68	199	49	248
%	100	100	100	100	100	100

Em relação ao sexo observou-se algumas diferenças que podem ser verificadas na **Tabela 2.10**, porém os alunos do sexo feminino são os que revelaram, com mais expressividade, melhores resultados relativamente aos alunos do sexo masculino, finalizando com maior frequência os cursos a que se candidataram (86,7%).

Tabela 2.10: Relação do número alunos inscritos/matriculados com o sexo e a finalização do curso.

Sexo	Ingresso ao curso	Finalização do curso			
		Sim		Não	
		№	%	№	%
Feminino	83	72	86,7	11	13,3
Masculino	165	127	77	38	23
Total	248	199	80,24	49	19,76

No que se refere à variável situação do aluno, tal como já foi referido anteriormente, a maioria dos alunos finalizaram o curso (80,24%). No entanto, verifica-se que quase um sexto dos alunos desistiu/abandonou do curso que se havia matriculado (16,53%).

A fase de candidatura claramente predominante é a 1.ª fase, com 94,76% dos alunos, sendo que, apenas 5,24% dos alunos candidataram-se na 2.ª e 3.ª fases.

Quanto à variável ex-aluno da UMa, a larga maioria dos alunos desta amostra nunca tinham sido alunos da UMa (85,05%).

Em relação à variável número de ECTS verifica-se que o número médio de ECTS que os alunos possuem é 107,38 ECTS, com um desvio padrão de 29,68 ECTS. O valor mínimo é 0 ECTS e, como 80,24% dos alunos já terminaram o curso, o 1.º quartil, a mediana, o 3.º quartil e o máximo é 120 ECTS.

Também foi possível observar que a maioria dos alunos não obtiveram o estatuto de trabalhador-estudante (87,9%) e que 50% dos alunos receberam apoio social.

A **Tabela D.1** disponível nos **Anexos** mostra uma análise geral de todas as informações estatísticas apresentadas anteriormente.

2.4 Análise de Sobrevivência aplicada aos CTeSP

Embora a área de aplicação mais frequente da Análise de Sobrevivência seja a medicina, com um forte impacto na sua terminologia, existem muitas mais, como por exemplo a educação. No entanto, a sua aplicação na educação não é muito comum, pelo que constitui um desafio aplicar esta teoria ao percurso escolar dos alunos de alguns CTeSP da Escola Superior de Tecnologias e Gestão da UMa.

Desta forma, pretende-se difundir a utilização da Análise de Sobrevivência na vertente educativa, contribuindo com mais uma ferramenta para a descrição e compreensão do percurso escolar dos alunos.

Neste estudo, o acontecimento de interesse irá corresponder à conclusão do respetivo CTeSP e, para cada aluno, o seu tempo de vida será o tempo decorrido desde que iniciou o curso até à conclusão do mesmo, sendo a censura dos dados à direita e do tipo I.

Para esta análise foi utilizada a mesma base de dados apresentada anteriormente, constituída por 248 alunos e a variável dicotómica indicadora da ocorrência do acontecimento de interesse (terminar o curso) é a variável estado do aluno.

Quanto às variáveis utilizadas, algumas das 19 variáveis apresentadas na **Tabela C.1** nos **Anexos** sofreram alterações. Assim sendo, a variável idade (anos) foi convertida numa variável com categorias, com vista a facilitar a interpretação e visualização dos dados, através da função *quantile* gerada no programa R, sendo assim formada a variável IdadeC. Com este procedimento foram geradas quatro categorias: [17,19],]19,21],]21,25] e]25,61].

Relativamente a variável situação do aluno, ano letivo de fim do curso e número de ECTS, optou-se por retirar estas três variáveis desta parte da análise, uma vez que acabariam por dar as mesmas informações que a variável estado do aluno, sendo assim variáveis redundantes. Para a variável estado

do aluno foi codificada o valor 0, se o aluno ainda não terminou o curso e, por outro lado, o valor 1, se o aluno já concluiu/terminou o curso.

Portanto, na modelação da variável tempo até à conclusão do CTeSP, consideramos como possíveis variáveis explicativas: idade de ingresso dos alunos em anos (categorizada); concelho de residência; sexo; naturalidade; habilitações literárias dos pais; profissão dos pais; habilitação de acesso ao curso; CTeSP; ano letivo de início; estado do aluno no fim do período observado; fase de candidatura; se já tinha sido aluno da UMa; se possui estatuto de trabalhador-estudante e se tem apoio social.

Nos **Anexos** é apresentado a **Tabela E.1** com a descrição de cada uma destas variáveis utilizadas nesta secção. Para as variáveis que apresentam categorias com informação muito extensa, foram criadas abreviaturas de modo a facilitar a visualização e interpretação da informação.

Numa primeira fase procedeu-se a uma análise univariada onde serão obtidas as estimativas de Kaplan-Meier (KM) da função de sobrevivência e apartir delas a mediana do tempo de permanência no curso até à conclusão do mesmo. Depois, serão comparadas estimativas de KM da função de sobrevivência relativas às das variáveis em estudo, quer graficamente, quer através da realização dos testes de hipóteses.

Nestes testes, a hipótese nula H_0 será para verificar se os grupos em análise apresentam uma função de sobrevivência igual e a hipótese alternativa H_1 será a situação contrária, isto é, se os grupos em análise apresentarem diferentes funções de sobrevivência.

Por norma é usual e conveniente os investigadores assumirem 5% como um nível de significância padrão (ou nível α), sendo que será igualmente assumido nesta análise esse valor.

Por fim, será apresentado o modelo de Cox para modelar o tempo até à conclusão do curso, assim como as principais conclusões que podemos retirar a partir das estimativas dos coeficientes de regressão.

2.4.1 Análise univariada

Para estudar a distribuição do tempo até finalizar o curso (em dias), isto é, o tempo que o aluno demorou para terminar o CTeSP, será feita uma primeira abordagem que consiste na obtenção da estimativa de Kaplan-Meier da função de sobrevivência, tendo em conta cada uma das variáveis, e a respetiva representação gráfica, de modo a ter uma noção do comportamento das curvas de sobrevivência para os vários grupos [11].

Na **Tabela 2.11** encontram-se as estimativas da mediana (*me*) do tempo de vida, obtidas com base na estimativa de Kaplan-Meier da função de sobrevivência, respetivos intervalos de confiança a 95% (*IC*95%) e percentis 25 e 75.

Tabela 2.11: Mediana (*me*), respetivos intervalos de confiança a 95% (*IC*95%), percentis 25 e 75.

Variáveis	*me* (IC95%)	Percentins	
		25	75
Concelho de residência:			
Funchal	696 (680; 715)	667	729
Camara de Lobos	683 (682; 712)	672	721
Santa Cruz	695 (671; 729)	667	740
Restantes Concelhos	682 (678; 708)	667	718
Sexo:			
Masculino	688 (681; 711)	667	728
Feminino	683 (682; 708)	669,5	728

Variáveis	me (IC95%)	Percentins	
		25	75
IdadeC:			
[17,19]	682 (673; 696)	667	718
]19,21]	689 (676; 715)	667	728
]21,25]	711 (682; 728)	668	738
]25,61]	711 (688; 723)	679	792
Naturalidade:			
África	676 (667; NA)	667	681
América do Sul	714 (694; 797)	683	797
Europa (exceto Portugal)	720 (710; NA)	710	728
Portugal (exceto RAM)	668 (665; NA)	665	800
RAM	683 (681; 701)	667	727
Habilitação literária do pai:			
Sem escolaridade	681 (672; NA)	672	693
4.º ano de escolaridade	682 (673; 692)	665	720
6.º ano de escolaridade	683 (673; 707)	665	728
9.º ano de escolaridade	683 (677; 729)	672	729
12.º ano de escolaridade	712 (682; 792)	671	792
Ensino pós-secundário	712 (697; 726)	689	733
Outra	727 (710; NA)	696	863
Habilitação literária da mãe:			
Sem escolaridade	688 (682; NA)	681,5	948
4.º ano de escolaridade	682 (672; 711)	665	724
6.º ano de escolaridade	681 (676; 696)	666	711
9.º ano de escolaridade	683 (673; 712)	667	723
12.º ano de escolaridade	711 (682; 729)	667	729
Ensino pós-secundário	721 (696; 792)	682	818
Outra	718 (682; NA)	682	727
Profissão do pai:			
Trabalhador por Conta de Outrem	684 (682; 712)	667	728
Desempregado	676 (671; 711)	664	719
Aposentado/Reformado	711 (682; 792)	677	792
Trabalhador por Conta Própria	711 (682; NA)	677	846
Outra situação	693 (681; 711)	669	725
Profissão do mãe:			
Trabalhador por Conta de Outrem	692 (682; 711)	665	728
Desempregada	681 (673; 719)	667	720
Doméstica	682 (673; 701)	672	712
Aposentada/Reformada	712 (682; NA)	680	797
Trabalhador por Conta Própria	725 (682; NA)	682	729
Outra situação	697 (673; 702)	667	867
Habilitação de acesso ao curso:			
Cursos Científico-Humanísticos	708 (683; 725)	673,5	729
Educação e Formação Profissional	682 (676; 696)	666	728
Cursos de Especialização Tecnológica	711 (680; NA)	680	711
Formação Superior	711 (708; NA)	708	740
Informação Omissa	718 (NA; NA)	718	718

Variáveis	me (IC95%)	Percentins	
		25	75
CTeSP:			
Agricultura Biológica	712 (711; 727)	708	728
Contabilidade e Fiscalidade	683 (682; 707)	668	724
Redes e Sistemas Informáticos	673 (672; 676)	671	867
Guias da Natureza	712 (696; 801)	664	720
Ano letivo de início do curso:			
2015/2016	683 (678; 708)	672	712
2016/2017	667 (665; 697)	662	712
2017/2018	676 (671; 719)	666	727
2018/2019	728 (724; 729)	684,5	800
Fase de candidatura:			
1.ª fase	684 (682; 708)	668	728
2.ª e 3.ª fases	695 (653; NA)	653	797
Aluno da UMa (ex-aluno):			
Sim	711 (697; 729)	683	740
Não	683 (679; 697)	667	727
Trabalhador-estudante:			
Sim	712 (683; 728))	682	728
Não	683 (680; 701)	667	728
Apoio social:			
Sim	680 (673; 685)	665	718
Não	711 (693; 723)	675	738

Nota: NA=Não aplicável.

Seguidamente realiza-se uma análise gráfica prévia das curvas de sobrevivência para todas as variáveis do estudo, de modo a verificar se estas se cruzam ou não, sendo escolhidos, desta forma, os testes adequados para cada variável (log-rank, Gehan e Breslow) e os resultados de cada teste são apresentados na **Tabela 2.12.**

Tabela 2.12: Resultados dos testes de log-rank, Gehan e Breslow para as comparações das curvas de sobrevivência de todas as variáveis do estudo.

Variáveis	Teste aplicado	χ^2	gl	Valor-p
Concelho de residência	Breslow	1,4	3	0,7
Sexo	Gehan	0	1	0,9
IdadeC	Breslow	5,8	3	0,1
Naturalidade	Breslow	8,8	4	0,07
Habilitação literária do pai	Breslow	10,9	6	0,09
Habilitação literária da mãe	Breslow	10,5	6	0,1
Profissão do pai	Breslow	5,8	4	0,2
Profissão do mãe	Breslow	3,3	5	0,6
Habilitações de acesso ao curso	Breslow	6	4	0,2
CTeSP	Breslow	20,4	3	0,0001
Ano letivo de início do curso	Breslow	36,5	3	0,00000006
Fase de candidatura	Gehan	0,9	1	0,3
Aluno da UMa (ex-aluno)	log-rank	5,7	1	0,02
Trabalhador-estudante	Gehan	4,5	1	0,03
Apoio social	log-rank	9,9	1	0,002

Com base na **Tabela 2.11**, podemos averiguar que, em termos gerais, a mediana dos vários grupos da variável concelho de residência varia entre 682 e 696 dias, não havendo diferenças significativas entre estes, como se pode comprovar pelo valor-p obtido no teste de Breslow (valor-p = 0,7), teste mais adequado, uma vez que as curvas de sobrevivência das 4 categorias desta variável cruzam-se em vários locais ao longo do tempo (ver **Figura E.1** nos **Anexos**).

De acordo com as **Figuras E.2** e **E.3** nos **Anexos**, com base na estimativa da mediana do tempo de duração do curso tendo em conta as variáveis IdadeC e sexo, podemos verificar que os alunos com faixas etárias entre os 21 e 25 anos e entre os 25 e 61 anos são os que demoram mais tempo a terminar o curso comparativamente aos restantes grupos e que os alunos do sexo masculino (M) tendem a demorar mais tempo a concluir o curso do que os alunos do sexo feminino (F). No entanto, as diferenças entre os grupos destas duas variáveis não são significativas, uma vez que os valores-p obtidos no teste de Breslow (valor-p = 0,1) e Gehan (valor-p = 0,9) são superiores a 5%.

No que diz respeito à naturalidade, embora se possa verificar que os alunos que nasceram na Europa exceto Portugal (E (exceto PT)) são os que demoram mais tempo a concluir o curso (**Figura E.4** nos **Anexos**), em termos gerais, a mediana varia entre 668 e 720 dias, não havendo diferenças significativas entre os grupos, uma vez que o valor-p obtido no teste de Breslow foi de 0,07 (ver **Tabela 2.12**).

No que concerne à profissão do pai e da mãe do aluno, em ambos os casos, se verifica que, os alunos cujo pai e mãe estão desempregados (De), são os que demoram menos tempo a realizar o curso, com 676 dias e 681 dias respetivamente, e os alunos cujo pai e mãe são trabalhadores por conta própria (TCP) ou estão aposentados/reformados (A/R) são os que demoram mais tempo a finalizar o curso (ver nos **Anexos** as **Figuras E.5** e **E.6**, respetivamente). No entanto, estas diferenças não se mostraram significativas dado os valores-p obtidos no teste de Breslow serem superiores a 5% (ver **Tabela 2.12**).

Quanto às habilitações literárias dos pais dos alunos, verificamos que, para o pai do aluno, em termos gerais, a mediana varia entre 681 e 727 dias (ver **Tabela 2.11**) e quanto à mãe, em termos gerais, a mediana varia entre 681 e 721 dias. No entanto, observando as **Figuras E.7** e **E.8** nos **Anexos**, podemos verificar que não existem diferenças consideráveis entre os grupos destas duas variáveis, portanto, como forma de comprovar esta análise, foi aplicado o teste de Breslow, uma vez que as curvas de sobrevivência dos grupos de cada variável se cruzam ao longo do tempo.

Desta aplicação, foi obtido o valor-p para cada variável (valor-p = 0,09 e valor-p = 0,1, respetivamente), sendo, através deles, comprovado que o teste não é significativo para as duas variáveis.

Relativamente às habilitações de acesso ao curso obtidas pelos alunos, verificamos, com base na **Figura E.9** nos **Anexos** e a **Tabela 2.11**, que os alunos que tinham como habilitações de acesso ao Curso a Educação e Formação Profissional terminaram o curso mais cedo do que os alunos que tinham outras habilitações de acesso, embora não haja diferenças significativas entre os grupos desta variável, dado o valor-p obtido (0,2) no teste de Breslow.

Quanto à fase de candidatura de ingresso, averiguamos que, com base na **Figura** (ver **Anexos**) **E.10** e na **Tabela 2.11**, os alunos que entraram na UMa na 1.ª fase de candidatura demoram menos tempo a realizar o curso comparativamente aos alunos que ingressaram na 2.ª e 3.ª fases. No entanto, de acordo com o valor-p obtido no teste de Gehan (valor-p = 0,3) concluímos que as curvas de sobrevivência dos dois grupos desta variável para além de se cruzarem ao longo do tempo, não são estatisticamente diferentes, sendo o teste não significativo.

Embora, até agora, todas os testes aplicados às 10 variáveis apresentadas anteriormente não se tenham revelado significativos, isso já não se verifica para a variável CTeSP, uma vez que o valor-p obtido no teste de Breslow foi de 0,0001. Dado isto, concluímos que, pelo menos uma das funções de sobrevivência dos quatro grupos da variável CTeSP é estatisticamente diferente das restantes, verificando que a variável CTeSP tem um impacto estatisticamente significativo ao nível de 5% no tempo até a conclusão do curso.

Além disso, com base na **Figura 2.2a**, podemos averiguar que os alunos que escolheram o CTeSP em Guias da Natureza (GN) demoram mais tempo para terminarem o curso comparativamente aos alunos que escolhem outros cursos. Os alunos que ingressam no CTeSP em Redes e Sistemas Informáticos (R&S I) são os que demoram menos tempo a realizar o curso, apresentando uma mediana de 683 dias, seguindo-se os alunos do CTeSP em Contabilidade e Fiscalidade (CF) com 673 dias e terminado com os alunos que ingressaram no CTeSP em Guias da Natureza ou no CTeSP em Agricultura Biológica (AB) com a mesma mediana de duração do curso (712 dias).

Relativamente à variável ano letivo de início do curso, com base na **Tabela 2.11**, verificamos que os alunos que ingressaram no ano letivo 2016/2017 demoraram menos tempo a realizar o curso comparativamente aos alunos que ingressam nos restantes anos letivos. No entanto, observando a **Figura 2.2b**, podemos verificar que as curvas de sobrevivência dos vários grupos desta variável estão muito próximas umas das outras, não evidenciando muitas diferenças. Contudo, aplicando o teste de Breslow, concluímos, com base no valor-p obtido (0,00000006 < 0,05), que o tempo até a conclusão do curso depende, estatisticamente ao nível de 5%, do ano letivo de início do curso.

Quanto à variável ex-aluno, verificamos que a mediana até à conclusão do curso para os alunos que já foram alunos da UMa é 29 dias a mais em relação aos alunos que nunca o foram (**Tabela 2.11**) e, com base na **Figura 2.2c**, podemos concluir que os alunos que já foram alunos da UMa apresentam, a partir dos 711 dias (mediana para os ex-alunos), um menor risco em terminar o curso, ou seja, demoram mais tempo ou nem terminam. Como as curvas de sobrevivência dos dois grupos desta varável nunca se cruzam ao longo do tempo, o teste mais adequado a aplicar seria o teste log-rank. Assim, desta aplicação, concluímos que as funções de sobrevivência dos dois grupos da variável ex-aluno são estatisticamente diferentes, tendo em conta o valor-p obtido (0,02 < 0,05).

Por fim, para as variáveis trabalhador-estudante e apoio social, concluímos, com base na **Tabela 2.11**, que os alunos que não são trabalhadores-estudantes, em termos medianos, terminam o seu curso ao fim de 683 dias, enquanto os alunos que o são o fazem ao fim de 712 dias, e que os alunos que não têm apoio social demoram mais tempo a terminar o curso (711 dias).

Além disso, pela análise da **Figura 2.2e**, podemos concluir que as curvas de sobrevivência dos dois grupos da variável apoio social estão praticamente sobrepostas até aos 711 dias, mas após esse dia, a curva de sobrevivência dos alunos com apoio social está abaixo da curva de sobrevivência dos alunos sem apoio social. Assim sendo, podemos constatar que os alunos sem apoio social têm um menor risco de terminarem o curso, refletindo-se com mais tempo para o fazer.

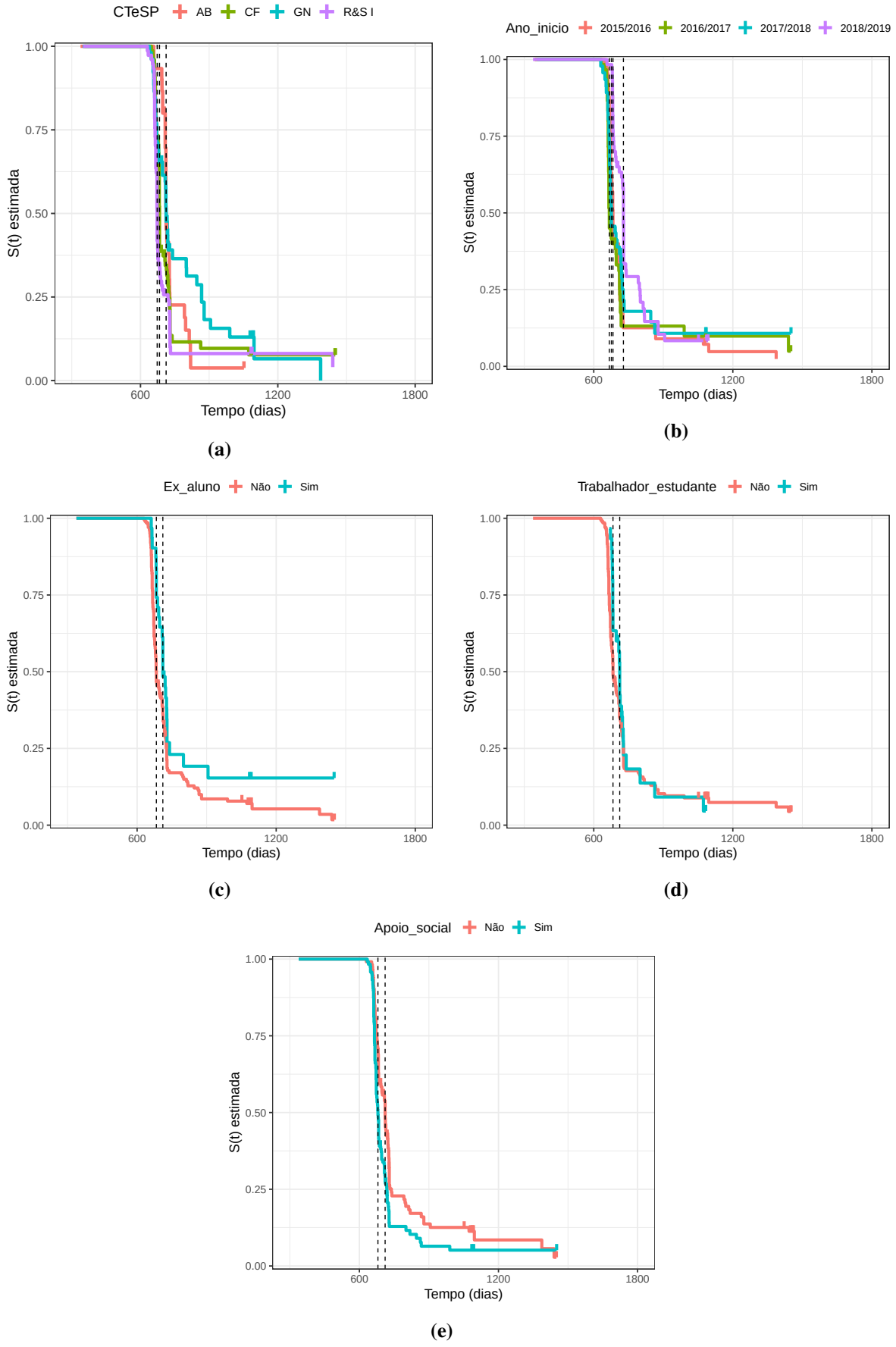


Figura 2.2: Estimativas de KM da função de sobrevivência (em dias) segundo as variáveis CTeSP (a), ano letivo de início do curso (b), ex-aluno (c), trabalhador-estudante (d) e apoio social (e) onde as linhas tracejadas representam a mediana.

Por outro lado, para verificar se as curvas de sobrevivência dos dois grupos destas variáveis, representadas na **Figura 2.2d** e na **Figura 2.2e** respetivamente, são estatisticamente diferentes, foi aplicado o teste de Gehan e log-rank, respetivamente, uma vez que as curvas de sobrevivência dos grupos da variável trabalhador-estudante cruzam-se ao longo do tempo, mas no caso da variável apoio social esse cruzamento não é evidente. Com base nos valores-p obtidos (0,03 e 0,002, respetivamente), concluímos que, com 95% de confiança, o apoio social e o facto do aluno ser trabalhador-estudante têm um impacto estatisticamente significativo ao nível de 5% no tempo até à conclusão do curso.

Assim, em suma, verificamos que, aplicando o teste de hipóteses adequado a cada variável, as variáveis CTeSP, ano letivo de início do curso, ex-aluno da UMa, trabalhador-estudante e apoio social foram as únicas cujas funções de sobrevivência dos seus respetivos grupos são estatisticamente diferentes ao nível de significância de 5%, portanto têm um impacto estatisticamente significativo no tempo até a conclusão do curso CTeSP feito pelo aluno.

Em contrapartida, podemos concluir que as funções de sobrevivência dos respetivos grupos das variáveis: sexo; IdadeC; concelho de residência; naturalidade; habilitações de acesso ao curso; habilitações literárias do pai e da mãe; profissão do pai e da mãe e fase de candidatura não são estatisticamente diferentes, tendo em conta o valor-p obtido no respetivo teste de hipóteses, concluindo desta forma que o tempo até à conclusão do curso não depende destas variáveis.

2.4.2 Modelo de regressão de Cox

Para aplicar este método ao estudo, começamos por ajustar o modelo de Cox utilizando o método de seleção de covariáveis proposto por Collett (2003). Para aplicar este método, foi incluído no modelo inicial todas as covariáveis que se revelaram significativas (valor-$p < 0{,}05$) na análise univariada e as covariáveis IdadeC e sexo, pois apesar de na análise univariada não terem sido estatisticamente significativas ao nível de 5%, consideramos ser importante a sua inclusão nesta fase inicial [2].

Começamos então por ajustar os modelos contendo apenas uma covariável de cada vez, calculamos os valores da estatística $-2\log\widehat{L}$ para cada um dos modelos univariados e comparamos com o valor obtido para o modelo nulo (sem covariáveis). Seguidamente, determinamos quais as covariáveis que, por si só, levam a uma redução significativa ao nível de 10% do valor da estatística de teste e que são potencialmente importantes, fazendo assim parte do modelo inicial na fase seguinte.

Tal como se pode verificar na **Tabela 2.13**, apenas as covariáveis trabalhador-estudante e sexo não levaram a uma redução significativa do valor da estatística de teste, obtendo um valor de valor-p de 0,3071 e 0,5917, respetivamente. Quanto às restantes covariáveis, verificamos que todas revelaram uma influência significativa no tempo de duração do curso CTeSP feito pelos alunos, pois, para todas estas, o valor-p obtido foi inferior a 0,1.

Tabela 2.13: Valores de $\log\widehat{L}$, $-2\log\widehat{L}$, χ^2 e valor-p para o modelo nulo e modelos univariados.

	$\log\widehat{L}$	$-2\log\widehat{L}$	χ^2	Valor-p
Modelo nulo	-920,21	1840,42	-	-
CTeSP	-914,61	1829,22	11,206	0,01066
Ano_inicio	-910,24	1820,48	19,932	0,0001753
Ex_aluno	-917,16	1834,32	6,0931	0,01357
Trabalhador_estudante	-919,69	1839,38	1,0433	0,3071
Apoio_social	-915,40	1830,8	9,6277	0,001917
IdadeC	-916,58	1833,16	7,253	0,06426
Sexo	-920,07	1840,14	0,2878	0,5917

Na segunda fase, incluímos as covariáveis potencialmente importantes obtidas na fase anterior num único modelo (Modelo completo: CTeSP; ano letivo de início do curso; ex-aluno; apoio social e IdadeC), calculamos o valor da estatística $-2\log\widehat{L}$, omitimos uma covariável de cada vez e retemos no modelo apenas aquelas que levaram a um aumento significativo ao nível de 10% do valor da estatística de teste. Como as covariáveis trabalhador-estudante e sexo não foram significativas na fase anterior já não serão usadas nesta fase.

Nesta segunda fase, como se pode apurar pela **Tabela 2.14**, todas as covariáveis incluídas no modelo resultante da primeira fase revelaram-se significativas ao nível de 10%, à exceção da covariável IdadeC, uma vez que a sua exclusão não levou a um aumento significativo da estatística de teste $-2\log\widehat{L}$ (valor-p = 0,2377 > 0,1). Portanto, na fase seguinte o modelo apenas irá incluir as covariáveis CTeSP, ano letivo de início do curso, ex-aluno e apoio social.

Tabela 2.14: Valores de $\log\widehat{L}$, $-2\log\widehat{L}$, χ^2 e valor-p para os vários modelos da segunda fase.

	$\log\widehat{L}$	$-2\log\widehat{L}$	χ^2	Valor-p
Modelo completo	-896,65	1793,3	-	-
Modelo sem CTeSP	-903,36	1806,72	13,411	0,003827
Modelo sem Ano_inicio	-906,51	1813,02	19,71	0,0001949
Modelo sem Ex_aluno	-898,57	1797,14	3,8329	0,05026
Modelo sem Apoio_social	-899,12	1798,24	4,9392	0,02625
Modelo sem IdadeC	-898,76	1797,52	4,2298	0,2377

Na terceira fase, voltamos a incluir, no modelo obtido na segunda fase, as variáveis que, quando consideradas isoladamente, não eram importantes e que, portanto, não foram consideradas na segunda fase, mais propriamente as covariáveis trabalhador-estudante e sexo, pois estas podem revelar-se significativas na presença de outras. O objetivo desta inclusão é averiguar se ao serem incluídas, uma de cada vez, levam a uma redução significativa ao nível de 10% do valor da estatística de teste $-2\log\widehat{L}$, sendo assim consideradas importantes para o modelo resultante da segunda fase.

Como se pode verificar pela **Tabela 2.15**, estas covariáveis não levaram a uma redução significativa ao nível de 10% da estatística de teste (valor-p = 0,4851 e valor-p = 0,6792), logo serão omitidas do modelo final.

Tabela 2.15: Valores de $\log\widehat{L}$, $-2\log\widehat{L}$, χ^2 e valor-p para os modelos da terceira fase.

	$\log\widehat{L}$	$-2\log\widehat{L}$	χ^2	Valor-p
Modelo com CTeSP, Ano_inicio, Apoio_social, Ex_aluno	-898,76	1797,52	-	-
Modelo com CTeSP, Ano_inicio, Apoio_social, Ex_aluno, Trabalhador_estudante	-898,52	1797,04	0,4875	0,4851
Modelo com CTeSP, Ano_inicio, Apoio_social, Ex_aluno, Sexo	-898,68	1797,36	0,171	0,6792

Na última fase, fazemos uma verificação final, para garantir que nenhuma covariável pode ser retirada sem levar a um aumento significativo da estatística de teste $-2\log\widehat{L}$ e que nenhuma covariável excluída leva a uma redução significativa do valor da estatística de teste. No nosso caso, como pela terceira fase já fizemos a verificação da covariáveis excluídas na primeira fase, no modelo obtido no final da segunda fase, iremos apenas fazer a verificação da significância da inclusão da covariável excluída na segunda fase (IdadeC).

Com base na **Tabela 2.16**, podemos concluir que a covariável IdadeC não levou a uma redução significativa ao nível de 10% da estatística $-2\log\widehat{L}$, portanto será omitida do modelo final. Quanto às restantes covariáveis, não é necessário efetuar uma segunda verificação, pois estão todas incluídas no modelo final.

Tabela 2.16: Valores de $\log\widehat{L}$, $-2\log\widehat{L}$, χ^2 e valor-p para os modelos da quarta fase.

	$\log\widehat{L}$	$-2\log\widehat{L}$	χ^2	Valor-p
Modelo com CTeSP, Ano_inicio, Apoio_social, Ex_aluno	-898,76	1797,52	-	-
Modelo com CTeSP, Ano_inicio, Apoio_social, Ex_aluno, IdadeC	-896,65	1793,3	4,2298	0,2377

Agora que o modelo de Cox foi obtido através do método de seleção de Collett (2003), precisamos de verificar se o pressuposto de riscos proporcionais é válido para as 4 covariáveis do modelo obtido. Para tal, começamos por averiguar se as curvas das estimativas de Kaplan-Meier da função sobrevivência para cada uma destas covariáveis não se cruzam.

Como podemos observar, com base nas **Figuras 2.2a**, **2.2b**, **2.2c** e **2.2e**, o pressuposto de proporcionalidade dos riscos parece ter sido violado pelas covariáveis CTeSP e ano letivo de início do curso, uma vez que as curvas definidas para cada categoria destas covariáveis se cruzam ao longo do período em estudo. Já para as covariáveis ex-aluno e apoio social, este pressuposto não parece ser violado, uma vez que as curvas das duas categorias destas covariáveis não se cruzam durante o período do estudo.

Outra forma empírica de confirmar se a hipótese de riscos proporcionais é válida para todas as covariáveis obtidas no final do método de seleção de Collett (2003), é recorrer ao método gráfico através da representação dos gráficos $\log(-\log(\widehat{S}(t)))$ *versus* t. Pela **Figura 2.3** podemos observar que a hipótese de riscos proporcionais é violada para as covariáveis ano letivo de início do curso e CTeSP, pois as curvas da função $\log(-\log(\widehat{S}(t)))$ *versus* t das várias categorias de cada covariável acabam-se cruzando em algumas partes do gráfico ao longo do tempo, enquanto que para as covariáveis ex-aluno e apoio social, as curvas desta função são razoavelmente paralelas ao longo do tempo, não violando assim o pressuposto de riscos proporcionais.

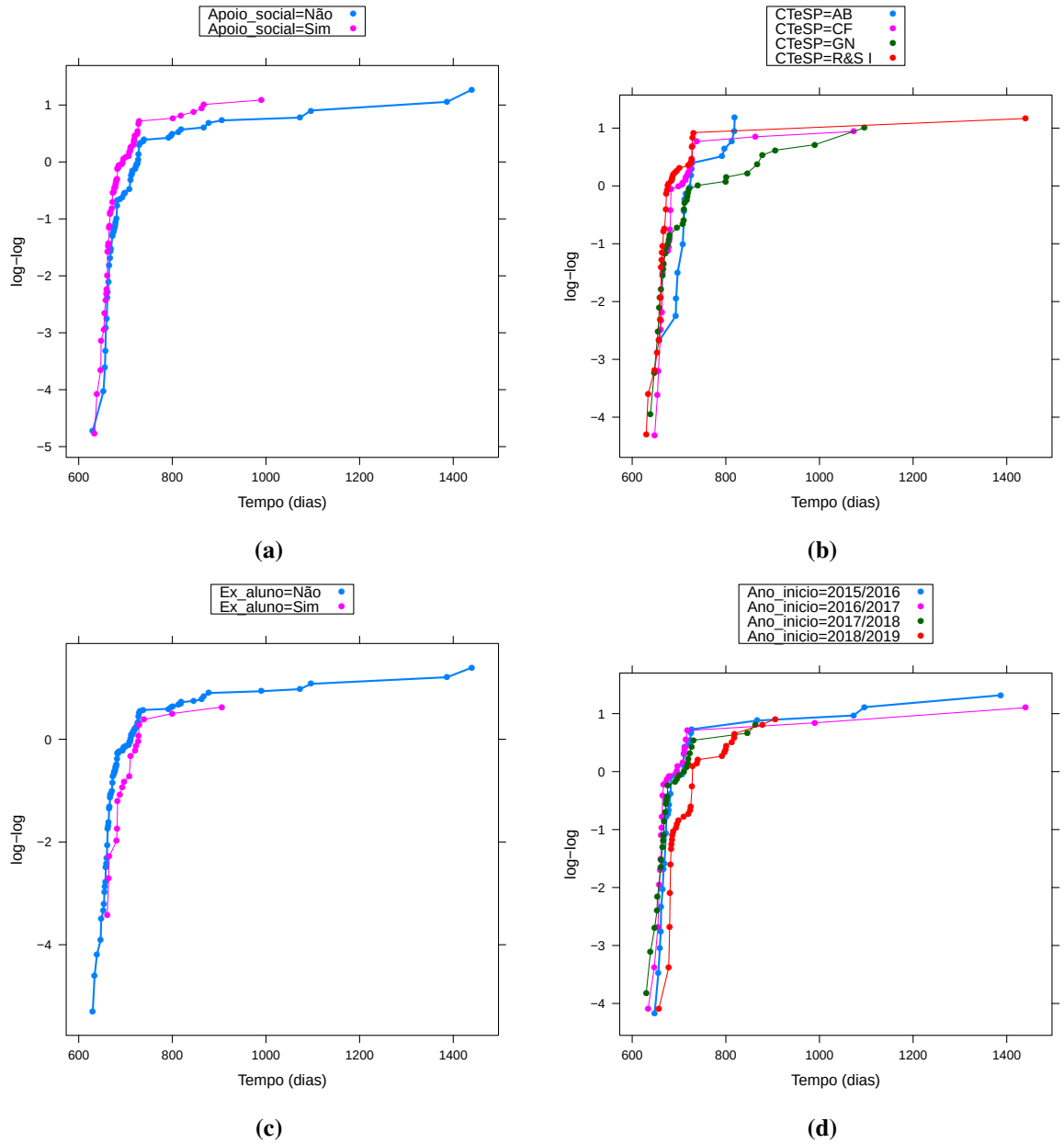


Figura 2.3: Gráficos da função $\log(-\log(\widehat{S}(t)))$ *versus* t segundo o apoio social (a), CTeSP (b), ex-aluno (c) e ano letivo de início do curso (d).

No entanto, embora o cruzamento das estimativas de Kaplan-Meier não implique inevitavelmente que o mesmo ocorre com os verdadeiros valores da função de sobrevivência, leva-nos a suspeitar que possa não existir proporcionalidade dos riscos das categorias destas covariáveis. Assim sendo, para podermos assegurar se a hipótese de riscos proporcionais é válida para todas as covariáveis optamos por recorrer à análise dos resíduos de Schoenfeld.

Na **Tabela 2.17 - Modelo de Cox 1** verificamos que, ao nível de significância de 5%, o pressuposto de proporcionalidade das funções de risco é violado nas covariáveis CTeSP e ano letivo de início do curso (valor-p = 0,000028610 e valor-p = 0,000000037, respetivamente), enquanto nas covariáveis apoio social e ex-aluno, este pressuposto é válido, dado o valor-p obtido para cada covariável ser superior a 5%. No entanto, olhando para o valor-p global chegamos a conclusão de que este modelo não é o melhor, pois o valor-p global é muito pequeno (valor-p = 0,000000014 < 0,05).

Tabela 2.17: Teste à proporcionalidade das funções de risco com base nos resíduos de Schoenfeld.

Modelo de Cox 1			
Covariável	χ^2	gl	Valor-p
CTeSP	23,718	3	0,000028610
Apoio_social	1,109	1	0,29
Ano_inicio	37,437	3	0,000000037
Ex_aluno	0,904	1	0,34
Global	52,342	8	0,000000014

Assim, com base na representação gráfica dos resíduos de Schoenfeld, podemos comprovar a conclusão obtida anteriormente, uma vez que, como estes resíduos são independentes do tempo, um gráfico que mostra um padrão não aleatório em relação ao tempo é evidência de violação do pressuposto de riscos proporcionais. Portanto, como se pode verificar pela **Figura 2.4**, a proporcionalidade das funções de risco é violada nas covariáveis CTeSP e ano letivo de início, uma vez que os resíduos correspondentes não apresentam um padrão aleatório em torno de zero, enquanto que nas covariáveis apoio social e ex-aluno, isso já não se verifica.

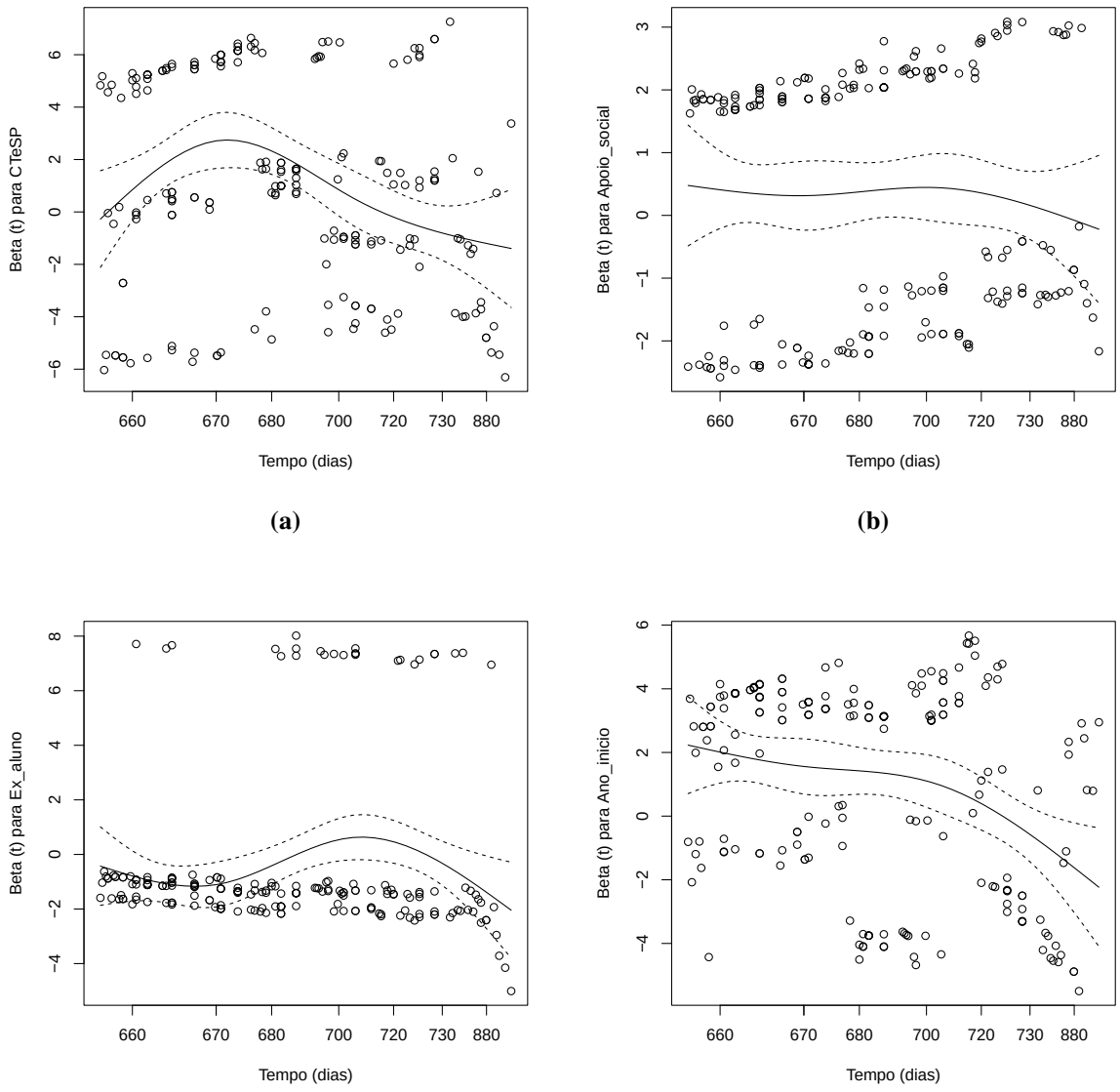


Figura 2.4: Resíduos de Schoenfeld para as covariáveis CTeSP (a), apoio social (b), ex-aluno (c) e ano letivo de início do curso (d).

Dado isto, decidimos estratificar o modelo pela covariável com o valor-p mais pequeno, obtido no modelo anterior no teste à proporcionalidade das funções de risco, ou seja, a covariável ano letivo de início do curso.

Contudo, antes de efetuar esta estratificação, apuramos que, como o programa R, por norma, assume como categoria de referência a comparar, a primeira categoria, por ordem alfabética, apresentada pela covariável CTeSP, que neste caso é o CTeSP em Agricultura Biológica (AB), e, visto esta categoria não ser a mais adequada a ser comparável por haver cruzamento da curva da categoria AB ao longo do período de tempo com as restantes categorias apresentadas, foi decidido alterar a categoria de referência AB para a categoria GN (Guias da Natureza).

Para tal, esta covariável foi codificada com o valor 1 se o aluno ingressou no CTeSP em Guias da Natureza, o valor 2 se o aluno ingressou no CTeSP em Agricultura Biológica, o valor 3 se ingressou no CTeSP em Contabilidade e Fiscalidade e o valor 4 se ingressou no CTeSP em Redes e Sistemas Informáticos, sendo assim designada de CTeSP1.

Com base na informação descrita na **Tabela 2.18** - **Modelo de Cox 2**, podemos verificar que as covariáveis apoio social e ex-aluno não violam o pressuposto da proporcionalidade das funções de risco, em termos globais, ao nível global, com nível de significância de 5% (valor-p = 0,969 e valor-p = 0,463, respetivamente) e que, embora a covariável CTeSP1 viole o pressuposto ao nível de 5% (mas não a 1%), verificamos que houve um aumento do valor-p global comparativamente ao **Modelo de Cox 1** (passou de 0,000000014 para 0,062) que, por sua vez, indica-nos que existe proporcionalidade das funções de risco ao nível de significância de 5% (valor-p = 0,062 > 0,05), portanto concluímos que este será o modelo mais adequado aos nossos dados, ficando assim determinado.

Tabela 2.18: Teste à proporcionalidade das funções de risco com base nos resíduos de Schoenfeld.

Modelo de Cox 2 (Estratificado por Ano_inicio)			
Covariável	χ^2	gl	Valor-p
CTeSP1	9,99527	3	0,019
Apoio_social	0,00155	1	0,969
Ex_aluno	0,53826	1	0,463
Global	10,50191	5	0,062

No que diz respeito à expressão matemática do modelo de Cox estratificado pela covariável ano letivo de início (**Modelo de Cox 2**), composto pelas covariáveis CTeSP1, apoio social e ex-aluno, escrita com base nas funções de risco, é a seguinte:

$$\begin{aligned} h(t;\mathbf{z}) =& h_0(t)\exp(\beta_1 Apoio\_social[Sim] + \beta_2 Ex\_aluno[Sim] \\ &+ \beta_3 CTeSP1[AB] + \beta_4 CTeSP1[CF] + \beta_5 CTeSP1[R\&SI]) \end{aligned}$$

ou

$$\begin{aligned} \frac{h(t;\mathbf{z})}{h_0(t)} =& \exp(\beta_1 Apoio\_social[Sim] + \beta_2 Ex\_aluno[Sim] \\ &+ \beta_3 CTeSP1[AB] + \beta_4 CTeSP1[CF] + \beta_5 CTeSP1[R\&SI]) \end{aligned}$$

Os valores das estimativas dos parâmetros deste modelo encontram-se na **Tabela 2.19**. Para verificar se as estimativas dos parâmetros são significativas, observa-se se o valor 1 não pertence ao intervalo de confiança estimado ou se o valor-p correspondente é inferior ao nível de significância de 5%.

Tabela 2.19: Estimativas dos parâmetros do modelo de Cox estratificado pela covariável ano letivo de início.

	$\widehat{\beta}_i$	$exp(\widehat{\beta}_i)(IC95\%)$	$SE(\widehat{\beta}_i)$	Teste Wald	Valor-p
CTeSP1 [AB]	-0,009011	0,991029 (0,6021; 1,6312)	0,254249	-0,035	0,9717
CTeSP1 [CF]	0,520672	1,683158 (1,1179; 2,5343)	0,208810	2,494	0,0126
CTeSP1 [R&S I]	0,816547	2,262674 (1,5060; 3,3994)	0,207690	3,932	0,0000844
Apoio_social [Sim]	0,266867	1,305866 (0,9674; 1,7628)	0,153078	1,743	0,0813
Ex_aluno [Sim]	-0,447926	0,638952 (0,4150; 0,9837)	0,220128	-2,035	0,0419

A partir deste modelo, verificamos que, por exemplo, para um aluno sem apoio social, sendo ex-aluno da UMa e pertencente ao CTeSP em Contabilidade e Fiscalidade, o vetor de covariáveis que lhe corresponde é dado por $\boldsymbol{z} = (0,1,0,1,0)$. Portanto, através da equação (1.7), a expressão da função de sobrevivência para este aluno é a seguinte:

$$\widehat{S}(t;\boldsymbol{z}) = \widehat{S}_0(t)^{\exp(0{,}266867\times 0 - 0{,}447926\times 1 - 0{,}009011\times 0 + 0{,}520672\times 1 + 0{,}816547\times 0)} \Leftrightarrow$$
$$\widehat{S}(t;\boldsymbol{z}) = \widehat{S}_0(t)^{\exp(0{,}072746)}.$$

Com base na informação descrita na **Tabela 2.19**, verificamos que, estando os alunos em igualdade de circunstâncias nas covariáveis apoio social e ex-aluno e, comparando com os alunos que ingressaram no CTeSP em Guias da Natureza, os alunos que ingressaram no CTeSP em Contabilidade e Fiscalidade apresentam um risco de terminar o curso de $68,32\%$ superior e que, os alunos que ingressaram no CTeSP em Redes e Sistemas Informáticos apresentam risco acrescido de $126,27\%$ superior ao dos alunos que ingressaram no CTeSP em Guias da Natureza. Quanto aos alunos que ingressaram no CTeSP em Agricultura Biológica, concluímos que estes apresentam o mesmo risco de terminar o curso em relação aos alunos que ingressaram no CTeSP em Guias da Natureza, uma vez que não existem diferenças significativas entre os grupos de alunos destas duas categorias, dado o valor-p não ter sido significativo ao nível de 5% (valor-p = 0,9717 > 0,05).

No que diz respeito à covariável ex-aluno, quer os alunos tenham ou não apoio social e independente do curso que ingressaram, averiguamos que os alunos que já foram alunos da UMa apresentam um menor risco em terminar o curso ($36,10\%$ inferior) comparativamente aos alunos que nunca o foram.

Em relação à covariável apoio social, concluímos que os alunos que têm apoio social apresentam o mesmo risco de terminar o curso comparativamente aos alunos que não o têm, uma vez que não existem diferenças significativas entre os grupos de alunos desta covariável, pelo facto do valor-p obtido ser superior ao nível de significância de 5% (valor-p = 0,0813 > 0,05).

No entanto, pelo facto desta covariável pertencer ao nosso modelo (tendo em conta que o seu valor-p está entre 0,05 e 0,1), pode-se verificar uma aparente tendência para que o facto de um aluno ter apoio social influencie positivamente o seu desempenho.

Capítulo 3

Conclusão

Este trabalho teve como base uma aplicação prática, mais propriamente o estudo de quatro Cursos Técnicos Superiores Profissionais (Agricultura Biológica, Contabilidade e Fiscalidade, Guias da Natureza e Redes e Sistemas Informáticos), da Escola Superior de Tecnologias e Gestão, da Universidade da Madeira, com o objetivo, de investigar o percurso escolar destes alunos, desde o ano letivo 2015/2016 até ao ano letivo 2019/2020.

Embora a área de aplicação mais frequente da Análise de Sobrevivência seja a medicina, com um forte impacto na sua terminologia, existem muitas mais, como o desporto, a política, a economia, a educação, entre outras. No entanto, a sua aplicação nesta última área não é muito comum, pelo que foi um desafio aplicar esta teoria ao percurso escolar dos alunos dos CTeSP da UMa, durante o período alvo referido anteriormente.

Neste estudo, verificou-se que, entre o ano letivo 2015/2016 e o ano letivo 2018/2019, candidataram-se aos CTeSP 391 alunos, entre os quais $66,50\%$ do sexo masculino e $33,50\%$ do sexo feminino. Durante este período de tempo, foi registado um maior número de candidaturas ($30,95\%$) no ano letivo 2018/2019 e um menor número ($19,69\%$) no ano letivo 2015/2016. Relativamente aos CTeSP, concluímos que, no conjunto dos quatro anos letivos, o curso com mais alunos candidatos foi o CTeSP em Redes e Sistemas Informáticos ($29,92\%$).

Os 391 candidatos disponham de 284 vagas, mas matricularam-se nos quatro CTeSP 248 alunos, constituindo-se, desta forma, a nossa amostra. Destes 248 alunos inscritos, $66,53\%$ eram do sexo masculino e $33,47\%$ do sexo feminino. Ao analisar os quatro anos letivos de acesso ao curso, durante o nosso período de estudo, verificamos que o ano letivo 2015/2016 foi o que revelou ter um maior número de alunos matriculados pela primeira vez ($28,63\%$) contrapondo o ano letivo 2017/2018, com apenas $18,95\%$. Dentro destas inscrições, $31,45\%$ foram para o CTeSP em Contabilidade e Fiscalidade e apenas $15,32\%$ para o CTeSP em Agricultura Biológica.

Além disso, concluímos que os alunos inscritos no ano letivo 2015/2016 foram os que apresentam uma maior taxa de sucesso em terminar o curso ($85,92\%$), tendo-se verificado a situação contrária nos alunos inscritos no ano letivo 2018/2019 ($72,06\%$).

Observando o número de vagas disponíveis em relação ao número de candidatos, verificamos que, durante o período de estudo, existe uma tendência evolutiva positiva, sendo que o CTeSP em Guias da Natureza o curso com maior afluência de candidatos (180%) face ao número de vagas disponíveis e o CTeSP em Agricultura Biológica, o curso com menor adesão (105%).

De um nível geral, $87,32\%$ das vagas disponíveis foram preenchidas durante todo o período de observação, verificando-se uma maior taxa de ocupação das vagas ($93,42\%$) no ano letivo 2015/2016 e menor taxa de ocupação ($81,52\%$) no ano letivo 2016/2017. Ao nível dos cursos, o CTeSP em Redes e Sistemas Informáticos foi o que apresentou uma maior taxa de ocupação das vagas ($93,75\%$) e o CTeSP em Agricultura Biológica, o curso com menor taxa ($63,33\%$).

Fazendo uma análise introspetiva na nossa amostra, este estudo evidenciou que, a maior preponderância de alunos residia no concelho do Funchal ($46,77\%$), naturais da Região Autónoma da

Madeira $(81,05\%)$ e apresentam uma média de idades de 24,05 anos. No que concerne às habilitações de acesso ao curso, dois terços dos alunos finalizaram o ensino secundário através da modalidade dos Cursos de Educação e Formação Profissional $(66,53\%)$, $26,21\%$ através dos Cursos Científico-Humanísticos, $4,44\%$ através de uma Formação Superior e $2,42\%$ através dos Cursos de Especialização Tecnológica.

Com esta análise também foi possível verificar que, até à data final da recolha dos dados (fim do *follow-up*), a grande maioria dos alunos havia terminado o curso $(80,24\%)$.

De um modo geral, podemos averiguar que a maioria dos alunos se candidatou ao CTeSP da UMa na 1.ª fase de candidatura $(94,76\%)$, nunca haviam sido alunos da UMa $(85,05\%)$, não apresentavam o estatuto de trabalhador-estudante $(87,9\%)$ e, curiosamente, exatamente 50% haviam recebido apoio social.

Em relação às variáveis habilitação literária do pai e da mãe verificou-se que, em ambos os casos, o 4.º ano de escolaridade foi a habilitação com maior número de observações e, quanto às respetivas profissões, tanto o pai como a mãe dos alunos, apresentam um maior número de registos desempenhando o cargo de Trabalhador por Conta de Outrem.

No que concerne aos CTeSP propriamente ditos, o CTeSP em Redes e Sistemas Informáticos foi o curso que mais contribuiu para a percentagem de alunos diplomados $(32,66\%)$ e o CTeSP em Agricultura Biológica $(13,57\%)$ o curso que menos contribuiu.

Após efetuar uma análise univariada às 15 variáveis utilizadas na Análise de Sobrevivência, este estudo divulgou que os alunos que já haviam sido alunos da UMa e que não recebem qualquer tipo de apoio social, foram os que demoram mais tempo a finalizar o seu respetivo CTeSP. Esta observação foi possível averiguar através da representação das estimativas de Kaplan-Meier da função de sobrevivência destas duas variáveis.

De modo a efetuar uma avaliação mais rigorosa acerca da existência de eventuais diferenças significativas entre as várias curvas de sobrevivência de todas as 15 variáveis mencionadas anteriormente, foi aplicado três testes de hipóteses: log-rank ou Mantel-Haenzel, Gehan e a sua generalização, teste de Breslow.

Com a realização destes testes estatísticos adequados a cada a variável, concluímos que as variáveis CTeSP, ano letivo de início do curso, ex-aluno da UMa, apoio social e trabalhador-estudante apresentaram um impacto estatisticamente significativo ao nível de 5% no tempo até a conclusão do curso CTeSP do aluno.

Em contrapartida, pudemos averiguar que as funções de sobrevivência dos respetivos grupos das variáveis: a fase de candidatura; as habilitações literárias do pai e da mãe; a profissão do pai e da mãe; a naturalidade; o concelho de residência; o sexo; as habilitações de acesso ao curso e a idade do aluno não foram estatisticamente diferentes, concluindo, desta forma, que o tempo até a conclusão do curso não dependeu destas variáveis.

Após efetuar uma análise univariada aos nossos dados, decidimos aplicar um modelo semi-paramétrico, mais propriamente o modelo de Cox. Para aplicar este método ao nosso estudo, começamos por ajustar o modelo de Cox utilizando o método de seleção de covariáveis proposto por Collett (2003). Deste método, obtivemos um modelo com as covariáveis CTeSP, ano letivo de início do curso, apoio social e ex-aluno, pois foram as únicas que se revelaram sempre significativas ao nível de 10%, verificando-se assim uma influência significativa no tempo de duração do CTeSP feito pelos alunos.

Dado isto, decidimos verificar se o pressuposto de riscos proporcionais foi válido para as quatro covariáveis do modelo obtido. Para tal, começamos por averiguar se as curvas das estimativas de Kaplan-Meier da função sobrevivência se cruzam, violando assim o pressuposto. Desta análise, concluímos que o pressuposto de proporcionalidade dos riscos parecia ter sido violado pelas covariáveis ano letivo de início do curso e CTeSP, enquanto, para as covariáveis ex-aluno e apoio social, isso já não se verificava.

Seguidamente, foi feito uma segunda verificação deste pressuposto recorrendo ao método gráfico através da representação dos gráficos $\log(-\log(\widehat{S}(t)))$ *versus* t. Desta representação, reforçamos a

nossa conclusão obtida na análise anterior, pois as covariáveis ex-aluno e apoio social foram as únicas cujas curvas foram razoavelmente paralelas ao longo do tempo, não violando assim o pressuposto de riscos proporcionais.

Por fim, para efetuar uma última verificação da validação do pressuposto de riscos proporcionais para as quatro covariáveis obtidas no final do método de seleção de Collet (2003) recorremos à análise dos resíduos de Schoenfeld. Desta análise, foi obtido, primeiramente, um modelo de Cox onde este pressuposto era de facto violado nas covariáveis CTeSP e ano letivo de início do curso, enquanto que, nas covariáveis apoio social e ex-aluno, este pressuposto era válido ao nível de significância de 5%. Portanto, optou-se por estratificar o modelo pela covariável ano letivo de início do curso, uma vez que esta teria sido a covariável que apresentava o valor-p mais pequeno nesta primeira análise. No entanto, há que salientar que, antes de efetuar esta estratificação, a covariável CTeSP foi recodificada (CTeSP1), tendo sido alterada na sua categoria de referência.

Assim sendo, desta estratificação, foi obtido um modelo estratificado por ano letivo de início do curso composto pelas covariáveis CTeSP1, apoio social e ex-aluno. Após uma análise com base nos valores-p obtidos para cada covariável, chegamos à conclusão que, embora a covariável CTeSP1 tenha violado o pressuposto de riscos proporcionais ao nível de significância de 5% (mas não de 10%), verificou-se um aumento do valor-p global comparativamente ao modelo de Cox anterior, indicando que existe proporcionalidade das funções de risco, e que, por sua vez, as covariáveis apoio social e ex-aluno não violavam este pressuposto ao nível de 5%. Portanto, este seria o modelo mais adequado aos nossos dados, ficando assim determinado.

Em termos dos valores obtidos para as estimativas dos parâmetros para o modelo de Cox determinado, verificamos que não existiam diferenças significativas entre os grupos de alunos que ingressaram no CTeSP em Agricultura Biológica e no CTeSP em Guias da Natureza e, que os alunos que tinham apoio social apresentavam o mesmo risco de terminar o curso que os alunos que não o tinham, uma vez que o valor-p respetivo não foi significativo ao nível de significância de 5%. Contudo, dado o valor-p obtido estar compreendido entre 0,05 e 0,1, verificou-se uma tendência positiva no desempenho do aluno pelo facto de o aluno ter apoio social.

Em relação à covariável ex-aluno e as restantes categorias da covariável CTeSP1, averiguamos que os alunos que já foram alunos da UMa apresentam um menor risco em terminar o curso ($36,10\%$ inferior) comparativamente aos alunos que nunca o foram e que, os alunos que ingressaram no CTeSP em Contabilidade e Fiscalidade e no CTeSP em Redes e Sistemas Informáticos apresentam, respetivamente, um risco de terminar o curso de $68,32\%$ e $126,27\%$ superior em relação aos alunos que ingressaram no CTeSP em Guias da Natureza.

Após todas estas conclusões, consideramos que os objetivos propostos foram cumpridos. Contudo, dado a situação pandémica em que nos encontramos, devido à COVID-19, debatemo-nos com esta limitação, uma vez que dificultou a abertura de espaços e o contacto direto com as pessoas, bem como a outras restrições. Para além deste constrangimento, alia-se as minhas funções profissionais, o meu papel de mãe de duas filhas e o facto de ter estado muitos anos sem rever os conhecimentos que haviam sido adquiridos na altura da minha licenciatura na UMa. O fator tempo também foi sem dúvida uma limitação muito difícil de gerir, devido a todos os aspetos já referidos anteriormente.

Para investigações futuras certamente seria uma mais valia refletir e compreender toda a dinâmica envolvente do aluno desde a sua entrada nos CTeSP, a sua adaptação, retenção, mudança de curso, transferência ou desistência, de forma a combater estes flagelos e compreender, de alguma forma, se estes advêm de desmotivação; integração académica e social; razões financeiras; razões de saúde; razões familiares; razões profissionais; o facto de não gostar do curso que escolheu; entre outras razões.

Um outro aspeto que também teria interesse investigar está relacionado com a situação pós-curso dos alunos que finalizaram os CTeSP, de modo a tentar compreender se estes alunos estão a desempenhar funções na área que se licenciaram; se estão a trabalhar, mas não na área que se licenciaram; se continuam a estudar; se estão desempregados; se emigraram; entre outras possíveis situações.

Acreditamos que esta dissertação possa contribuir para um melhor entendimento do objeto de estudo exposto, na tentativa de melhor compreender o percurso escolar dos alunos, neste caso em concreto, dos alunos dos CTeSP da UMa.

De referir que toda a teoria e resultados narrados nesta tese foram retirados de livros, teses, artigos e fontes eletrónicas apresentados na bibliografia, pelo que a sua consulta certamente será uma mais-valia.

Anexos

Anexo A

Pedido de autorização para recolha de dados

UNIVERSIDADE da MADEIRA

Escola Superior de Tecnologias e Gestão

Autoriza-se a aluna Susana Maria Pereira da Silva, mestranda da Universidade da Madeira, que está a realizar a sua dissertação de mestrado sob a supervisão da Professora Doutora Ana Abreu, a recolher os dados necessários ao seu estudo sobre os alunos dos CTeSP da Escola Superior de Tecnologias e Gestão.

Mais concretamente recolher informação sobre os alunos que se matricularam pela primeira vez nestes cursos nos anos letivos 2015/16, 2016/17 e 2017/18.

A informação a recolher é a que consta do questionário em anexo e que deve ser disponibilizada pelo Secretariado da ESTG, quando for solicitada.

Funchal, 2 de dezembro de 2019

O Presidente da Escola Superior de Tecnologias e Gestão

Prof Doutor João Filipe Pereira Nunes Prudente

QUESTIONÁRIO

Percurso Escolar dos Alunos dos Cursos Técnicos Superiores Profissionais (CTeSP) da Universidade da Madeira (UMa)

Dimensão Pessoal

1. Concelho onde reside: ☐ Calheta ☐ Câmara de Lobos ☐ Funchal ☐ Machico ☐ Ponta do Sol ☐ Porto Moniz ☐ Porto Santo ☐ Ribeira Brava ☐ Santa Cruz ☐ Santana ☐ São Vicente
2. Idade: _____anos
3. Género: ☐ Feminino ☐ Masculino
4. Naturalidade: _____
5. Nacionalidade: ☐ Portuguesa ☐ Outra. Qual?_____

Dimensão Social

1. Profissão dos pais:

	Pai	Mãe
Quadros superiores de Administração Pública, dirigentes e quadros superiores de empresas	☐	☐
Especialistas das profissões intelectuais e científicas	☐	☐
Técnicos e profissionais de nível intermédio	☐	☐
Pessoal administrativo e similares	☐	☐
Pessoal dos serviços e vendedores	☐	☐
Agricultores e trabalhadores qualificados da agricultura e pesca	☐	☐
Operários, artífices e trabalhadores similares	☐	☐
Operadores de instalações e máquinas e trabalhadores da montagem	☐	☐
Trabalhadores não qualificados	☐	☐
Pessoal das Forças Armadas	☐	☐
Doméstico(a)	☐	☐
Aposentado(a)/reformado(a)	☐	☐
Desempregado(a)	☐	☐
Outra	☐	☐

2. Habilitações literárias dos pais:

	Pai	Mãe
Não sabe ler nem escrever/sem escolaridade	☐	☐
Sabe ler ou escrever sem possuir o 4.º ano de escolaridade	☐	☐
4.º ano de escolaridade (antiga 4.ª classe)	☐	☐
6.º ano de escolaridade	☐	☐
9.º ano de escolaridade	☐	☐
Ensino secundário complementar ou equivalente	☐	☐
Ensino médio/cursos técnicos profissionais	☐	☐
Ensino superior (bacharelato, licenciatura, mestrado, doutoramento)	☐	☐
Desconheço	☐	☐

Dimensão Escolar

1. Curso CTeSP:

   ☐ Agricultura Biológica ☐ Contabilidade e Fiscalidade ☐ Guias da Natureza ☐ Redes e Sistemas Informáticos ☐ Sistemas Eletrónicos e Instalações Elétricas ☐ Tecnologias e Programação de Sistemas de Informação

2. Área de formação:

   2.1. ☐ Ensino secundário: ☐ Ciências e Tecnologias ☐ Ciências Socioeconómicas ☐ Línguas e Humanidades ☐ Artes Visuais

   2.2. ☐ Ensino Profissional: Nome do curso_____

3. Ano letivo que iniciou o CTeSP: ☐ 2015/2016 ☐ 2016/2017 ☐ 2017/2018
4. Fase de candidatura: ☐ 1.ª Fase ☐ 2.ª Fase ☐ 3.ª Fase
5. Número de ECTS creditados: _____
6. Tipo de regime adotado: ☐ Diurno ☐ Pós-laboral ☐ Misto
7. Estatuto de trabalhador estudante: ☐ Sim ☐ Não
8. Apoio social (bolsa de estudo, ...): ☐ Sim ☐ Não
9. Ano letivo que finalizou o CTeSP: ☐ 2016/2017 ☐ 2017/2018 ☐ 2018/2019 ☐ Ainda não terminou
10. Razões para ainda não ter terminado o curso: ☐ Não obteve aproveitamento (classificação; faltas superiores a 25% das aulas previstas;...) ☐ Mudou de curso ☐ Desistiu ☐ Já terminou
11. Razões para ter abandonado o curso: ☐ Mudou de curso ☐ Emigrou ☐ Razões familiares ☐ Razões profissionais ☐ Razões de saúde ☐ Razões financeiras ☐ Outra razão. Qual?_____
12. Situação pós-curso: ☐ Trabalha na área ☐ Trabalha mas não na área ☐ Continua a estudar ☐ Desempregado(a) ☐ Emigrou ☐ Outra razão. Qual?_____

Anexo B

Descrição das variáveis recolhidas

Tabela B.1: Descrição das variáveis recolhidas.

Variáveis	Código	Descrição	Tipo de variável	Categorização
Identificação	ID_R	Amostra (sample) de 1 a 248 valores inteiros gerados aleatoriamente no programa R.	Quantitativa - Ordinal	-
Concelho de residência	creside	Concelho atual de residência do aluno.	Qualitativa - Nominal	• Calheta • Câmara de Lobos • Funchal • Machico • Ponto do Sol • Porto Moniz • Porto Santo • Ribeira Brava • Santa Cruz • Santana • São Vicente • Fora da RAM
Sexo	sexo	Sexo do aluno.	Qualitativa - Nominal	• Masculino • Feminino
Data de nascimento	datanasc	Data de nascimento do aluno.	Qualitativa - Ordinal	-
Naturalidade	naturalidade	Local de nascimento do aluno.	Qualitativa - Nominal	• Calheta (Arco da Calheta e Ponta do Pargo) • Câmara de Lobos (Estreito de Câmara de Lobos e Jardim da Serra e Quinta Grande) • Funchal (Sé, S. Martinho, Sto. António, Monte, S. Pedro, S. Gonçalo, Imaculado Coração de Maria, Sta. Maria Maior, Sta. Luzia e S. Roque) • Machico (Porto da Cruz, Caniçal, Santo António da Serra e Água de Pena) • Ponta do Sol (Canhas) • Porto Moniz • Porto Santo • Ribeira Brava (Campanário) • Santa Cruz (Gaula, Camacha e Caniço) • Santana (Faial e S. Jorge) • São Vicente • Fora da RAM (Lisboa, Porto, Viana do Castelo, Vila Real, Açores, Espanha, Brasil, Venezuela, Reino Unido, Itália, África e Alemanha)
Nacionalidade	nacionalidade	País de nascimento do aluno ou onde adquiriu a naturalização.	Qualitativa - Nominal	• Portuguesa • Outra
Habilitação literária dos pais	hablitpai e hablitmae	Habilitação literária do pai e da mãe do aluno.	Qualitativa - Nominal	• Não sabe ler nem escrever/sem escolaridade • Sabe ler e escrever (sem ter o 4.º ano) • 4.º ano de escolaridade (antiga 4.ª classe) • 6.º ano de escolaridade • 9.º ano de escolaridade • 12.º ano de escolaridade • Ensino Médio (Antigo Magistério) • Curso sem Grau • Curso de Especialização Tecnológica • Bacharelato • Licenciatura • Mestrado • Doutoramento • Outra • Desconhecido
Profissão dos pais	profpai e profmae	Profissão do pai e da mãe do aluno.	Qualitativa - Nominal	• Desempregado(a) • Doméstico(a) • Estudante • Aposentado(a)\ Reformado(a) • Serviço Militar • Trabalhador Familiar Não Renumerado • Trabalhador por Conta de Outrem • Trabalhador por Conta Própria como empregador • Trabalhador por Conta Própria como isolado • Outra situação
Habilitação de acesso ao curso	habacescurso	Habilitação de acesso dos alunos aos CTeSP's.	Qualitativa - Nominal	• Cursos Científico-Humanísticos • Educação e Formação Profissional • Cursos de Especialização Tecnológica • Formação Superior • Outra

Variáveis	Código	Descrição	Tipo de variável	Categorização
Cursos CTeSP	ctesp	Nome dos CTeSP.	Qualitativa - Nominal	• Agricultura Biológica • Contabilidade e Fiscalidade • Guias da Natureza • Formação Superior • Redes e Sistemas Informáticos
Ano letivo de início do curso	perioletinicio	Ano letivo de ingresso do curso.	Qualitativa - Ordinal	• 2015/2016 • 2016/2017 • 2017/2018 • 2018/2019
Ano letivo de fim do curso	perioletfim	Ano letivo de finalização do curso. No caso do ano letivo 2020/2021 será até o dia 21/02/2021.	Qualitativa - Ordinal	• 2016/2017 • 2017/2018 • 2018/2019 • 2019/2020 • 2020/2021 • Ainda não terminou
Data de início do curso	datainicio	Data de início do curso.	Qualitativa - Ordinal	-
Data de fim do curso	datafim	Data de conclusão do curso.	Qualitativa - Ordinal	-
Tempo para finalizar o curso	tempofimcurso	Tempo (em dias) que o aluno levou até terminar o curso. Calculado com base na diferença entre a data de início do curso e a data de fim do curso.	Quantitativa - Contínua	-
Situação do aluno	situacaoaluno	Situação ou estado atual do aluno no fim do período de observação ou avaliação.	Qualitativa - Nominal	• Desistiu/abandonou • Mudou de curso • A decorrer (em curso) • Já terminou
Fase de candidatura	fasecand	Fase de candidatura do CTeSP.	Qualitativa - Ordinal	• 1.ª fase • 2.ª fase • 3.ª fase
Aluno da UMa (ex-aluno)	jalunouma	Se o aluno já foi aluno da UMa.	Qualitativa - Nominal	• Não • Sim
N.º de ECTS	numects	Tempo de trabalho despendido pelo estudante na obtenção da aprovação de uma disciplina (neste caso de um curso) nos termos do Decreto-Lei n.º 42/2005	Quantitativa - Contínua	Mínimo: 0 ECTS e Máximo: 120 ECTS
Trabalhador-estudante	trabestud	Se o aluno possui estatuto de trabalhador-estudante.	Qualitativa - Nominal	• Não • Sim
Apoio social	apoiosoc	Se o aluno tem recebeu algum apoio social (bolsa de estudo, . . .).	Qualitativa - Nominal	• Não • Sim
Razões para ainda não ter terminado o curso	naoterminar	Razões por ainda não ter terminado o curso.	Qualitativa - Nominal	• Não obteve aproveitamento (classificação; faltas superiores a 25% das aulas previstas; . . .) • Mudou de curso • Desistiu • Já terminou

Anexo C

Tabelas da análise descritiva dos dados

Tabela C.1: Descrição das variáveis utilizadas na análise descritiva dos dados.

Variáveis	Código	Descrição	Tipo de variável	Categorização
Identificação	ID_R	Amostra (sample) de 1 a 248 valores inteiros gerados aleatoriamente no programa R.	Quantitativa - Ordinal	-
Concelho de residência	Concelho	Concelho atual de residência do aluno.	Qualitativa - Nominal	• Funchal (Fu) • Câmara de Lobos (CL) • Santa Cruz (SC) • Restantes Concelhos (RC)
Sexo	Sexo	Sexo do aluno.	Qualitativa - Nominal	• Masculino • Feminino
Idade	Idade	Idade atual do aluno (anos) a partir da data de início do curso.	Quantitativa - Contínua	-
Naturalidade	Naturalidade	Local de nascimento do aluno.	Qualitativa - Nominal	• África (A) • América do Sul (AS) • Europa (exceto Portugal) (E (exceto PT)) • Portugal (exceto RAM) (PT (exceto RAM)) • RAM
Habilitação literária dos pais	Habilitacoes_pai e Habilitacoes_mae	Habilitação literária do pai e da mãe do aluno.	Qualitativa - Nominal	• Sem Escolaridade (SE) • 4.º ano de escolaridade (antiga 4.ª classe) • 6.º ano de escolaridade • 9.º ano de escolaridade • 12.º ano de escolaridade • Ensino Pós-Secundário (superior e não superior) (EPS) • Outra
Profissão dos pais	Profissao_pai e Profissao_mae	Profissão do pai e da mãe do aluno.	Qualitativa - Nominal	• Desempregado(a) (De) • Doméstico(a) (Do) • Aposentado(a)/Reformado(a) (A/R) • Trabalhador por Conta de Outrem (TCO) • Trabalhador por Conta Própria (TCP) • Outra Situação (OS)
Habilitação de acesso ao curso	Habilitacoes_entrada	Habilitação de acesso dos alunos aos CTeSP's.	Qualitativa - Nominal	• Cursos Científico-Humanísticos (CCH) • Educação e Formação Profissional (EFP) • Cursos de Especialização Tecnológica (CET) • Formação Superior (FS) • Informação Omissa (IO)
Cursos CTeSP	CTeSP	Nome dos CTeSP.	Qualitativa - Nominal	• Agricultura Biológica (AB) • Contabilidade e Fiscalidade (CF) • Guias da Natureza (GN) • Redes e Sistemas Informáticos (R&S I)
Ano letivo de início do curso	Ano_inicio	Ano letivo de ingresso do curso.	Qualitativa - Ordinal	• 2015/2016 • 2016/2017 • 2017/2018 • 2018/2019
Ano letivo de fim do curso	Ano_fim	Ano letivo de finalização do curso.	Qualitativa - Ordinal	• 2016/2017 • 2017/2018 • 2018/2019 • 2019/2020 • Ainda não terminou
Tempo para finalizar o curso	Tempo_dias	Tempo (em dias) que o aluno levou até terminar o curso.	Quantitativa - Contínua	-
Estado do aluno	Estado_aluno	Verificação se o aluno terminou ou não o curso.	Qualitativa - Nominal	• Não • Sim
Situação do aluno	Situacao_aluno	Situação ou estado atual do aluno no fim do período de observação ou avaliação.	Qualitativa - Nominal	• Desistiu/Abandonou (D/A) • Mudou de Curso (MC) • Em Curso (EC) • Terminou (T)
Fase de candidatura	Fase	Fase de candidatura do CTeSP.	Qualitativa - Ordinal	• 1.ª fase • 2.ª e 3.ª fases
Aluno da UMa (ex-aluno)	Ex_aluno	Se o aluno já foi aluno da UMa (ex-aluno da UMa).	Qualitativa - Nominal	• Não • Sim
N.º de ECTS	ECTS	Tempo de trabalho despendido pelo estudante na obtenção da aprovação de uma disciplina (neste caso de um curso).	Quantitativa - Contínua	Mínimo: 0 ECTS e Máximo: 120 ECTS
Trabalhador-estudante	Trabalhador_estudante	Se o aluno possui estatuto de trabalhador-estudante.	Qualitativa - Nominal	• Não • Sim

Variáveis	Código	Descrição	Tipo de variável	Categorização
Apoio social	Apoio_social	Se o aluno tem recebeu algum apoio social (bolsa de estudo,...).	Qualitativa - Nominal	• Não • Sim

Anexo D

Síntese da informação estatística

Tabela D.1: Síntese da informação estatística.

Variável	№ de registos					
Candidaturas	391					
Vagas	284					
Inscrições	248					
Data do *follow-up*	30/9/2015 a 28/02/2021					
		№	%			
Concelho de residência	Funchal	116	46,77			
	Câmara de Lobos	32	12,90			
	Santa Cruz	40	16,13			
	Restantes Concelhos	60	24,19			
		№	%			
Sexo	Masculino	165	66,53			
	Feminino	83	33,47			
		№				
Idade (anos)	Mínimo	17				
	1.º Quartil	19				
	Mediana	21				
	Média	24,05				
	3.º Quartil	25				
	Máximo	61				
	Desvio-padrão	8,18				
		№	%			
Naturalidade	RAM	201	81,05			
	América do Sul	28	11,29			
	Portugal (exceto RAM)	7	2,82			
	África	6	2,42			
	Europa (exceto Portugal)	6	2,42			
		№	%		№	%
Habilitação literária dos pais	Pai:			Mãe:		
	Sem escolaridade	6	2,42	Sem escolaridade	8	3,23
	4.º ano de escolaridade	83	33,47	4.º ano de escolaridade	63	25,40
	6.º ano de escolaridade	62	25	6.º ano de escolaridade	52	20,97
	9.º ano de escolaridade	26	10,48	9.º ano de escolaridade	40	16,13
	12.º ano de escolaridade	34	13,71	12.º ano de escolaridade	43	17,34
	Ensino pós-secundário	24	9,68	Ensino pós-secundário	36	14,52
	Outra	13	5,24	Outra	6	2,42
		№	%		№	%
Profissão dos pais	Pai:			Mãe:		
	Desempregado	45	18,15	Desempregada	39	15,73
	Doméstico	0	0	Doméstica	40	16,13
	Trabalhador por Conta Própria	19	7,66	Trabalhador por Conta Própria	11	4,44
	Trabalhador por Conta de Outrem	97	39,11	Trabalhador por Conta de Outrem	108	43,55
	Aposentado/Reformado	27	10,89	Aposentada/Reformada	17	6,85
	Outra Situação	60	24,19	Outra Situação	33	13,31
		№	%			
Habilitação de acesso ao curso	Formação Superior	11	4,44			
	Educação e Formação Profissional	165	66,53			
	Cursos Científico-Humanísticos	65	26,21			
	Cursos de Especialização Tecnológica	6	2,42			
	Informação Omissa	1	0,40			
		№	%			
CTeSP	Contabilidade e Fiscalidade	78	31,45			
	Redes e Sistemas Informáticos	75	30,24			
	Guias da Natureza	57	22,98			
	Agricultura Biológica	38	15,32			
		№	%			
Ano letivo de início do curso	2015/2016	71	28,63			
	2016/2017	62	25			
	2017/2018	47	18,95			
	2018/2019	68	27,42			
		№	%			
Ano letivo de fim do curso	2016/2017	56	22,58			
	2017/2018	54	21,77			
	2018/2019	37	14,92			
	2019/2020	52	20,97			
	Ainda não terminou	49	19,76			

		№	
Tempo para finalizar o curso (dias)	Mínimo	338	
	1.º Quartil	665	
	Mediana	683	
	Média	707,27	
	3.º Quartil	720	
	Máximo	1812	
	Desvio-padrão	173,52	
		№	%
Estado do aluno	Sim	199	80,24
	Não	49	19,76
		№	%
Situação do aluno	Já terminou	199	80,24
	Desistiu/abandonou	41	16,53
	A decorrer (em curso)	4	1,61
	Mudou de curso	4	1,61
		№	%
Fase de candidatura	1.ª fase	235	94,76
	2.ª e 3.ª fases	13	5,24
		№	%
Aluno da UMa (ex-aluno)	Não	211	85,05
	Sim	37	14,92
		№	
N.º de ECTS	Mínimo	0	
	1.º Quartil	120	
	Mediana	120	
	Média	107,38	
	3.º Quartil	120	
	Máximo	120	
	Desvio-padrão	29,68	
		№	%
Trabalhador-estudante	Não	218	87,9
	Sim	30	12,1
		№	%
Apoio social	Não	124	50
	Sim	124	50

Anexo E

Tabela e figuras da Análise de Sobrevivência

Tabela E.1: Variáveis utilizadas na Análise de Sobrevivência.

Variáveis	Código	Categorização	Abreviatura
Identificação	ID_R	Valores de 1 a 248, gerados aleatoriamente no programa R.	-
Concelho de residência	Concelho	• Funchal • Câmara de Lobos • Santa Cruz • Restantes Concelhos	• Fu • CL • SC • RC
Sexo	Sexo	• Masculino • Feminino	• M • F
Idade	IdadeC	• [17,19] •]19,21] •]21,25] •]25,61]	• [17,19] • (19,21] • (21,25] • (25,61]
Naturalidade	Naturalidade	• África • América do Sul • Europa (exceto Portugal) • Portugal (exceto RAM) • RAM	• A • AS • E (exceto PT) • PT (exceto RAM) • RAM
Habilitação literária dos pais	Habilitacoes_pai e Habilitacoes_mae	• Sem Escolaridade • 4.º ano de escolaridade • 6.º ano de escolaridade • 9.º ano de escolaridade • 12.º ano de escolaridade • Ensino Pós-Secundário • Outra	• SE • 4.º ano • 6.º ano • 9.º ano • 12.º ano • EPS • O
Profissão dos pais	Profissao_pai e Profissao_mae	• Desempregado(a) • Doméstico(a) • Aposentado(a)/Reformado(a) • Trabalhador por Conta de Outrem • Trabalhador por Conta Própria • Outra Situação	• De • Do • A/R • TCO • TCP • OS
Habilitação de acesso ao curso	Habilitacoes_entrada	• Cursos Científico-Humanísticos • Educação e Formação Profissional • Cursos de Especialização Tecnológica • Formação Superior • Informação Omissa	• CCH • EFP • CET • FS • IO
Cursos CTeSP	CTeSP	• Agricultura Biológica • Contabilidade e Fiscalidade • Guias da Natureza • Formação Superior • Redes e Sistemas Informáticos	• AB • CF • GN • FS • R&S I
Ano letivo de início do curso	Ano_inicio	• 2015/2016 • 2016/2017 • 2017/2018 • 2018/2019	• 2015/2016 • 2016/2017 • 2017/2018 • 2018/2019
Tempo para finalizar o curso	Tempo_dias	-	-
Estado do aluno	Estado_aluno1	• Não • Sim	• 0 • 1
Fase de candidatura	Fase	• 1.ª fase • 2.ª e 3.ª fases	• 1.ª fase • 2.ª e 3.ª fases
Aluno da UMa (ex-aluno)	Ex_aluno	• Não • Sim	-
Apoio social	Apoio_social	• Não • Sim	-

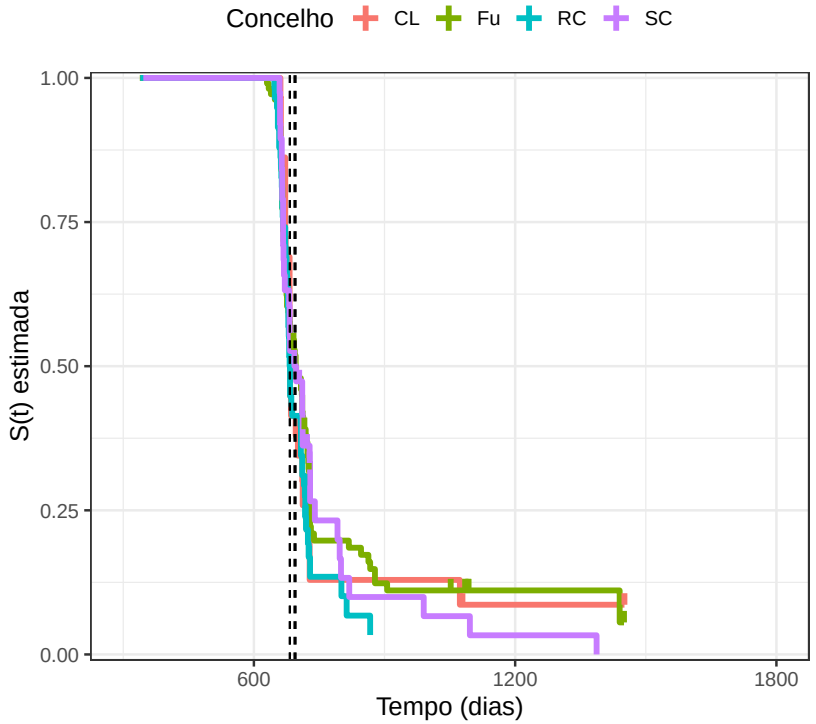


Figura E.1: Estimativa de KM da função de sobrevivência (em dias) segundo o concelho de residência, onde as linhas tracejadas representam a mediana.

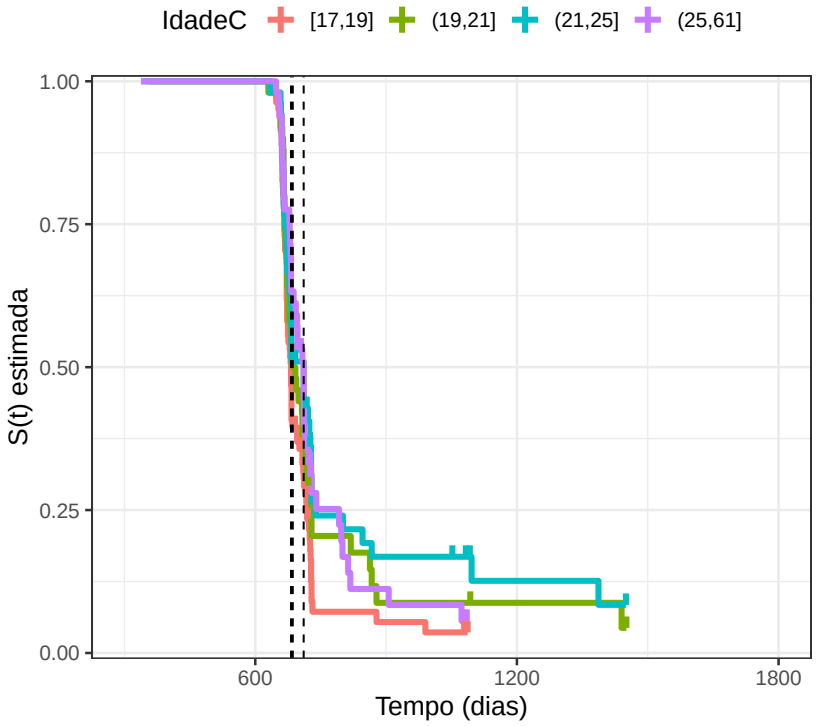


Figura E.2: Estimativa de KM da função de sobrevivência (em dias) segundo a IdadeC, onde as linhas tracejadas representam a mediana.

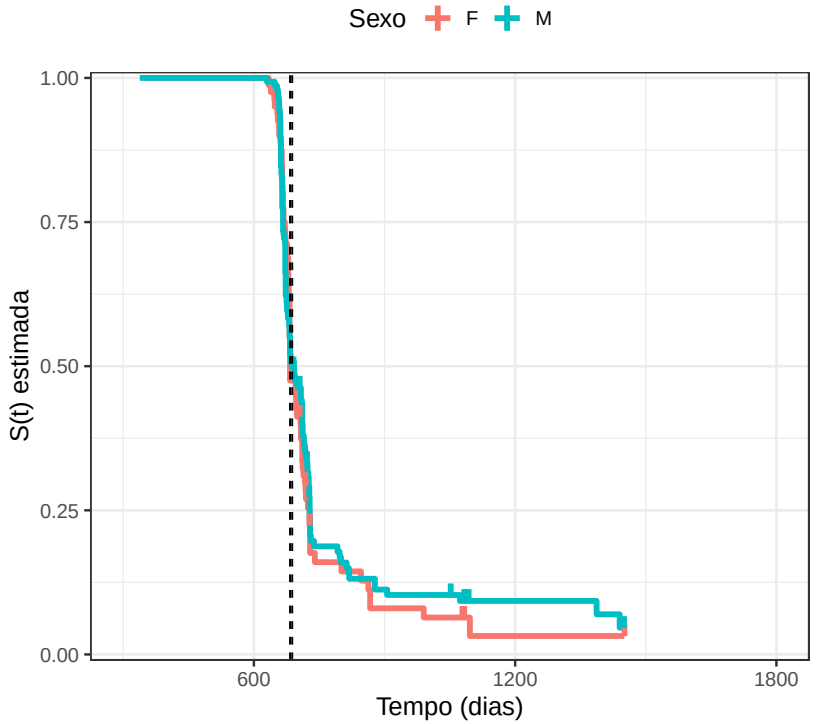


Figura E.3: Estimativa de KM da função de sobrevivência (em dias) segundo o sexo, onde as linhas tracejadas representam a mediana.

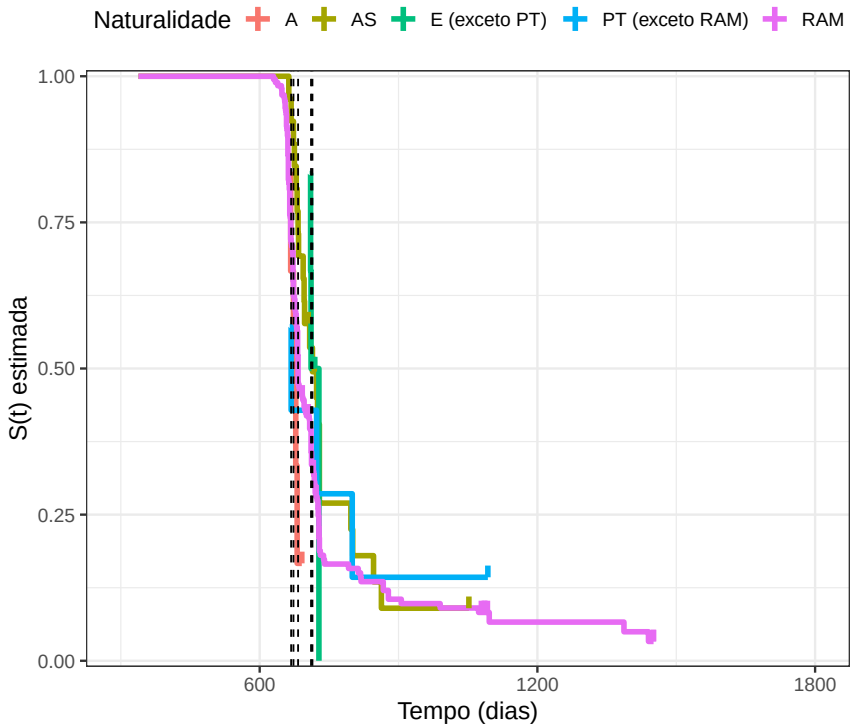


Figura E.4: Estimativa de KM da função de sobrevivência (em dias) segundo a naturalidade, onde as linhas tracejadas representam a mediana.

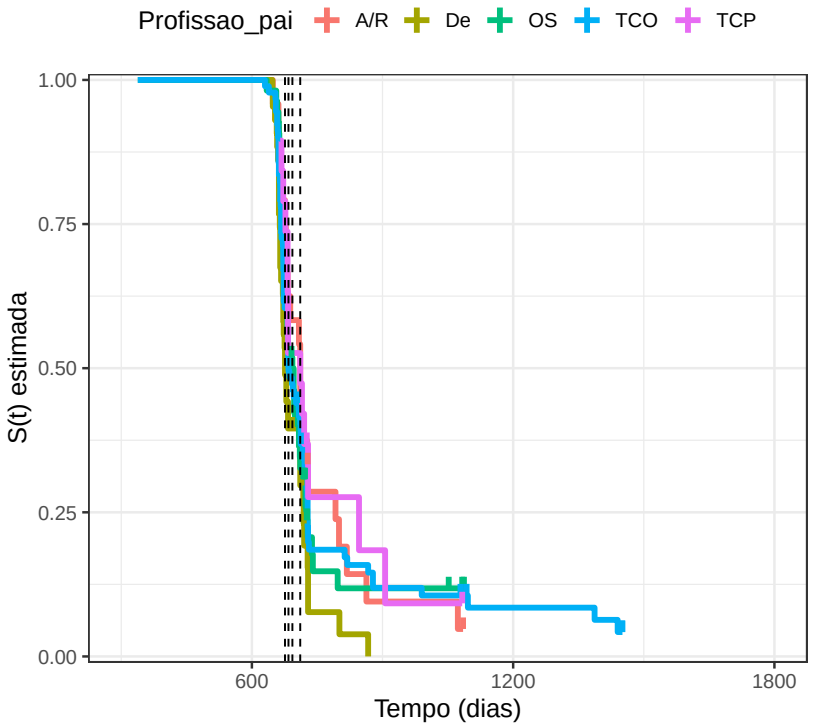


Figura E.5: Estimativa de KM da função de sobrevivência (em dias) segundo a profissão do pai, onde as linhas tracejadas representam a mediana.

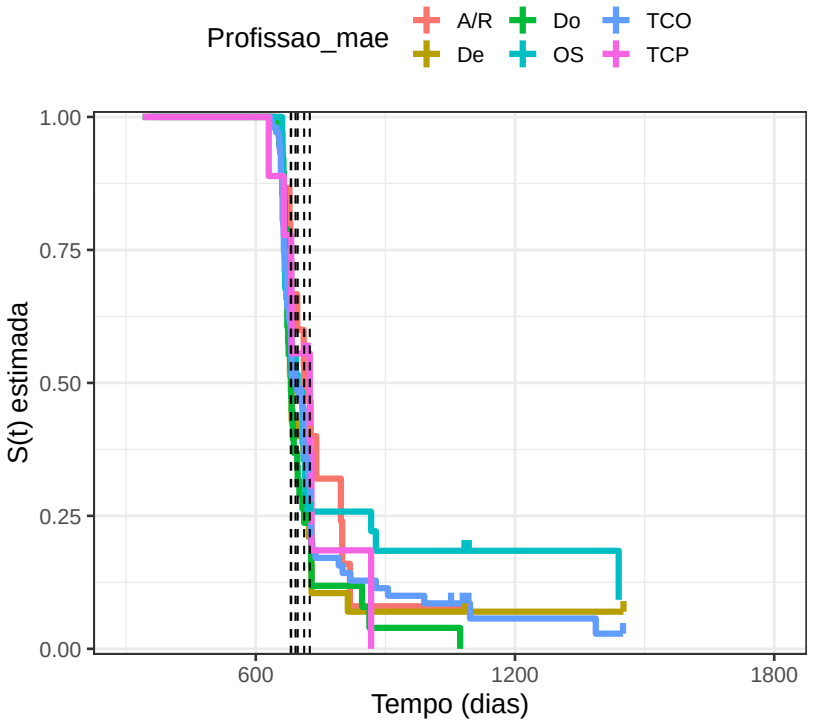


Figura E.6: Estimativa de KM da função de sobrevivência (em dias) segundo a profissão da mãe, onde as linhas tracejadas representam a mediana.

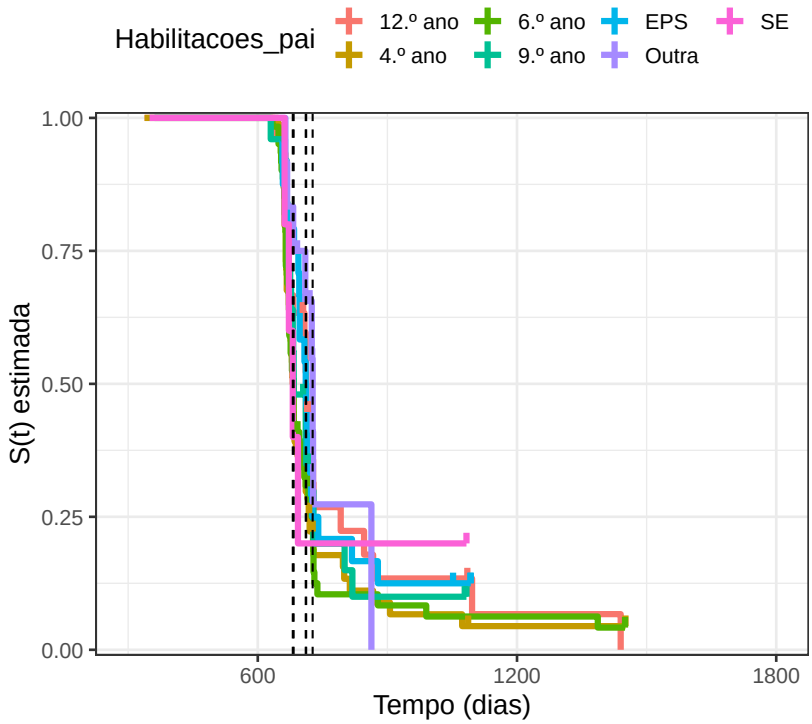


Figura E.7: Estimativa de KM da função de sobrevivência (em dias) segundo a habilitação literária do pai, onde as linhas tracejadas representam a mediana.

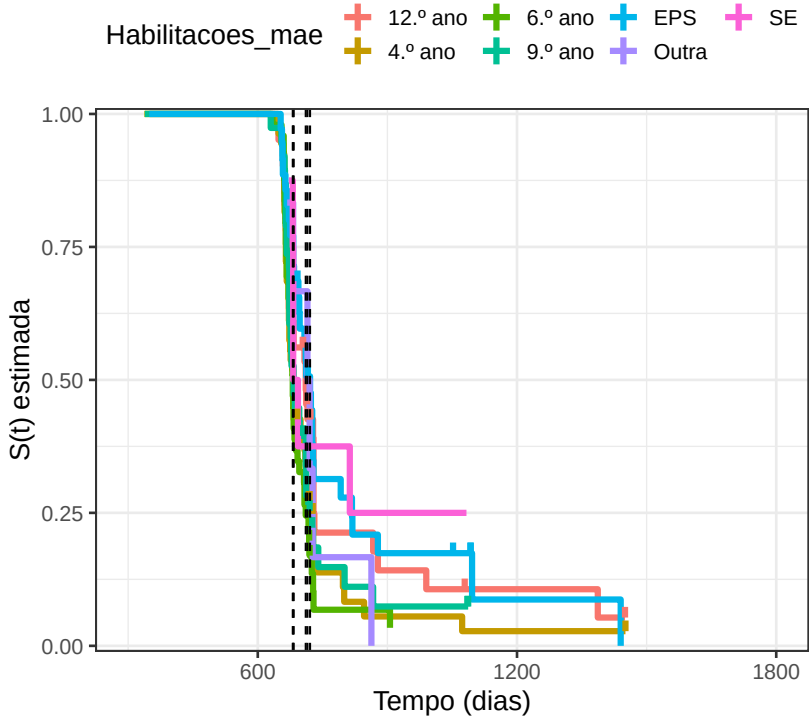


Figura E.8: Estimativa de KM da função de sobrevivência (em dias) segundo a habilitação literária da mãe, onde as linhas tracejadas representam a mediana.

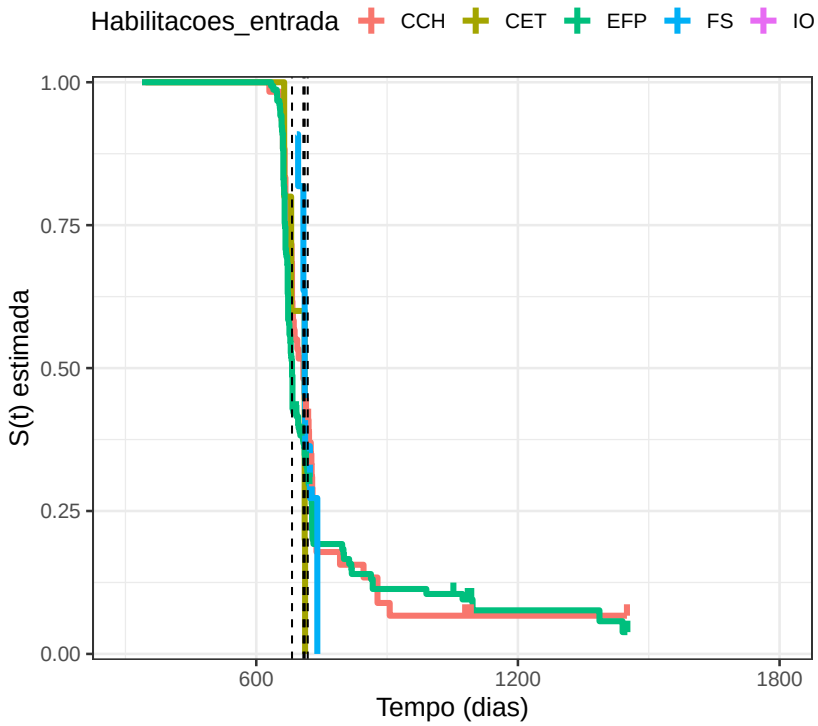


Figura E.9: Estimativa de KM da função de sobrevivência (em dias) segundo a habilitações de acesso ao curso, onde as linhas tracejadas representam a mediana.

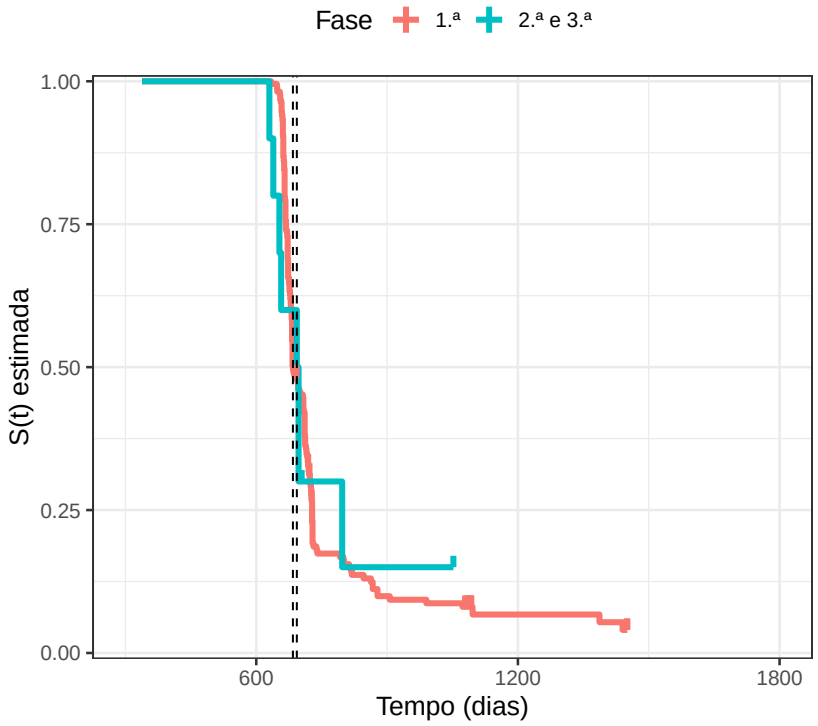


Figura E.10: Estimativa de KM da função de sobrevivência (em dias) segundo a fase de candidatura, onde as linhas tracejadas representam a mediana

Bibliografia

[1] C. Rocha e A.L. Papoila. *Análise de Sobrevivência*. Ed. por XVII Congresso da Sociedade Portuguesa de Estatística. 2009. ISBN: 978-972-8890-22-3.

[2] A.C.F. Alves. «Aplicação ao Cancro da Mama Feminino». Dissertação de mestrado. Faculdade de Ciências Exatas e da Engenharia. Universidade da Madeira, 2012.

[3] A.I.M. Borges. «Análise de Sobrevivência com o R». Dissertação de mestrado. Faculdade de Ciências Exatas e da Engenharia. Universidade da Madeira, 2014.

[4] N.E. Breslow. «Covariance analysis of censored survival data». Em: *Biometrics* 30 (1974), pp. 89–99.

[5] Tecnologia e Ensino Superior Ciência. *Decreto-Lei n. º 63/2016, de 13 de setembro*. Set. de 2016. URL: https://dre.pt/web/guest/pesquisa/-/search/75319452/details/normal?l=1 (acedido em 20/02/2021).

[6] D. Collett. *Modelling Survival Data im Medical Research*. Ed. por Boca Raton Chapman & Hall/CRC. 2.ª Edição. 2003.

[7] D.R. Cox e E.J. Snell. «A general definition of residuals (with discution)». Em: *Journal of the Royal Statistical Society*. A 30 (1968), pp. 248–275.

[8] Ministério da Educação e Ciência. *Decreto-Lei n. º 113/2014, de 16 de julho*. Jul. de 2014. URL: https://dre.pt/web/guest/pesquisa/-/search/55021010/details/normal?q=Decreto-Lei+n.+%C2%BA%20113%2F2014 (acedido em 20/02/2021).

[9] B. Efron. «The two sample problem with censored data». Em: *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4 (1967). Ed. por New York: Prentice-Hall, pp. 831–853.

[10] J. Fox. *RcmdrPlugin.survival: R Commander Plug-in for the 'survival' Package. Package do R versão 1.2-1*. 2019. URL: https://cran.r-project.org/web/packages/RcmdrPlugin.survival/index.html (acedido em 26/02/2021).

[11] R.B. Freitas. «Modelação estatística do tempo de duração de procedimentos nos serviços académicos de uma Universidade». Dissertação de mestrado. Escola de Ciências e Tecnologia. Universidade de Évora, 2015.

[12] E.A. Gehan. «A generalized Wilcoxon test for comparing arbitrarily singly censored samples». Em: *Biometrika* 52 (1965), pp. 203–223.

[13] M. Mendes e J.A. Almeida. *Preparação de Textos Científicos Usando LATEX*. Ed. por Lda Lisboa: Edições Sílabo. 1.ª Edição. 2005. ISBN: 972-618-361-8.

[14] Y. Kanda. *RcmdrPlugin.EZR: Plug-in do R Commander para o pacote EZR (Easy R)*. Versão 1.54. 2020. URL: https://cran.r-project.org/web/packages/RcmdrPlugin.EZR/index.html (acedido em 21/02/2021).

[15] J. Fox e M. Bouchet-Valat. *Rcmdr: R Commander. Package do R versão 2.7-1*. 2019. URL: http://CRAN.R-project.org/package=Rcmdr (acedido em 26/02/2021).

[16] Universidade da Madeira. *Cursos Técnicos Superiores Profissionais*. URL: https://cursotecnicosuperuma.pt/.

[17] Universidade da Madeira. *Documentos ESTG - Regulamento dos CTeSP*. 2017. URL: http://estg.uma.pt/index.php?option=com_docman&task=cat_view&gid=44&Itemid=69&lang=pt (acedido em 20/02/2021).

[18] Universidade da Madeira. *Documentos ESTG-Edital das Candidaturas aos CTeSP*. 2019. URL: http://estg.uma.pt/index.php?option=com_docman&task=cat_view&gid=44&Itemid=69&lang=pt. (acedido em 17/01/2021).

[19] Inovação e Ensino Superior Ministério da Ciência. *Decreto-Lei n. º 42/2005, de 22 de fevereiro*. URL: https://dre.pt/home/-/dre/606304/details/maximized (acedido em 20/02/2021).

[20] D.F. Moore. *Applied Survival Analysis Using R*. Ed. por Springer International Publishing AG Switzerland. 2016940055. 2016. ISBN: 978-3-319-31245-3.

[21] K. Nagashima. *RcmdrPlugin.KMggplot2: R Commander Plug-in for Data Visualization with 'ggplot2'. Package do R versão 0.2-6*. 2019. URL: https://cran.r-project.org/web/packages/RcmdrPlugin.KMggplot2/index.html (acedido em 20/02/2021).

[22] E.L. Kaplan e P. Meier. «Non-parametric estimation from incomplete observations». Em: *Journal of the American Statistical Association* 53 (1958), pp. 457–481.

[23] M. Rocha e P.G. Ferreira. *Análise e Exploração de Dados com R*. Ed. por Lda Lisboa: FCA - Editora de Informática. 2017.

[24] P.K. Anderson e R.D. Gil. «Cox's regression model for counting processes: A large sample study». Em: *Annals of Statistics* 10 (1982), pp. 1100–1120.

[25] J.D. Kalbfeisch e R.L. Prentice. «Marginal likelihoods based on Cox's regression and life model». Em: *Biometrika* 60 (1973), pp. 267–278.

[26] D.A. Schoenfeld. «Partial rsiduals for the proportional hazards regression model». Em: *Biometrika* 69 (1982), pp. 239–241.

[27] E. Strapasson. «Comparação de modelos com censura intervalar em análise de sobrevivência». Tese de mestrado. Escola Superior de Agricultura "Luiz de Queiroz". Universidade de São Paulo, 2007.

[28] B.V.D. Zander; J. Sundermeyer; D. Braun e T. Hoffmann. *TeXstudio: LaTeX made comfortable. Versão 2.12.18*. 2009. URL: https://texstudio.org. (acedido em 22/01/2021).

[29] P.M. Grambsch e T.M. Therneau. «Proportional hazards tests and diagnostics based on weighted residuals». Em: *Biometrika* 81 (1994), pp. 515–526.

[30] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2016. URL: http://www.R-project.org/ (acedido em 20/02/2021).

[31] T. Therneau T.M. e Lumley. *Survival Package*. Versão 3,2-13. 2021. URL: https://cran.r-project.org/web/packages/survival/index.html (acedido em 20/02/2021).

[32] Jornal Oficial da União Europeia. *Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho*. Abr. de 2016. URL: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679&from=DA (acedido em 20/02/2021).

[33] N. Mantel e W. Haenszel. «Statistical aspects of the analysis of data from retrospective studies of disease». Em: *Journal of the National Cancer Institute* 22 (1959), pp. 719–748.