UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE CIÊNCIAS MATEMÁTICAS

Modelos de sobrevivência paramétricos para estudo da mortalidade associada à insuficiência cardíaca

Ana Isabel Nunes dos Santos

Mestrado em Estatística e Investigação Operacional

Especialização em Estatística

Dissertação orientada por:

Marília Cristina de Sousa Antunes

2025

Aos meus avós

Agradecimentos

Ao longo deste percurso, várias pessoas foram fundamentais para a realização desta tese e do mestrado.

Um enorme agradecimento à professora Marília Antunes, orientadora desta tese de mestrado, pela sua disponibilidade, atenção, ajuda e contribuições para a realização deste trabalho.

Um especial agradecimento aos meus pais pelo apoio e por me terem proporcionado a oportunidade de realizar este mestrado, sem eles não seria possível.

Um sentido agradecimento ao meu irmão e à minha cunhada pela ajuda preciosa e atenção dedicada.

Às minhas amigas, o meu obrigada pelo apoio, ajuda e incentivo, e pelos momentos de descontracção que aliviaram o stress nesta jornada e ao longo de todo o percurso académico.

Um agradecimento a todos os familiares, colegas e professores que de algum modo me ajudaram ao longo deste percurso.

Resumo

A insuficiência cardíaca é a principal causa de internamentos a nível mundial. Trata-se de uma doença cardiovascular, em que o coração não é capaz de bombear a quantidade de sangue necessária para o corpo, podendo este deixar de funcionar normalmente. Apesar de não ter cura, o diagnóstico precoce desta doença e o tratamento adequado permitem que os pacientes tenham uma melhor qualidade de vida. Para tal, torna-se essencial conhecer os factores de risco que contribuem para o aparecimento desta doença.

No Paquistão, a incidência das doenças cardiovasculares tem vindo a aumentar. O difícil acesso a cuidados de saúde, escolhas alimentares pouco saudáveis, conflitos regionais e a ocorrência de catástrofes naturais são alguns exemplos que contribuem para este aumento.

No presente trabalho, tem-se como objectivo identificar os factores que influenciam o tempo de vida de doentes com insuficiência cardíaca, seguidos no Instituto de Cardiologia do Hospital *Allied*, na cidade de *Faisalabad*, no Paquistão, entre os meses de Abril e Dezembro, do ano de 2015. Estes dados provêm de um artigo científico, realizado no Paquistão, que teve como objectivos calcular as taxas de morte devido à insuficiência cardíaca e conhecer os principais factores de risco, recorrendo ao modelo de Cox.

Neste estudo, pretende-se analisar os dados, com recurso ao *package survival* do *software* estatístico R, utilizando modelos de regressão paramétricos de Análise de Sobrevivência, com o intuito de identificar os factores que influenciam o tempo de vida dos indivíduos. Na análise preliminar, utilizou-se a estimativa de Kaplan-Meier e o teste log-rank.

Idade, fracção de ejecção, creatinina, pressão arterial, anemia, CPK e sódio foram identificados como factores de risco significativos para a mortalidade por insuficiência cardíaca.

Palavras-chave: insuficiência cardíaca, modelo de Cox, modelos paramétricos, Análise de Sobrevivência

Abstract

Heart failure is the leading cause of hospitalizations worldwide. It is a cardiovascular disease in which the heart is unable to pump the necessary amount of blood throughout the body, and the body may cease to function normally. Although there is no cure, early diagnosis and appropriate treatment allow patients to have a better quality of life. Therefore, it is essential to understand the risk factors that contribute to the development of this disease.

In Pakistan, the incidence of cardiovascular disease has been increasing. Poor access to healthcare, unhealthy dietary choices, regional conflicts, and the occurrence of natural disasters are some examples of factors contributing to this increase.

This study aims to identify the factors that influence the survival time of patients with heart failure treated at the Institute of Cardiology at Allied Hospital in Faisalabad, Pakistan, between April and December 2015. This data comes from a research article published in Pakistan, which aimed to calculate death rates due to heart failure and identify the main risk factors using the Cox model.

This study aims to analyze the data using the R statistical computing software with the survival package, using parametric regression models from survival analysis to identify the factors that influence survival time. The preliminary analysis used the Kaplan-Meier estimate and the log-rank test.

Age, ejection fraction, creatinine, blood pressure, anemia, CPK and sodium were found as significant risk factors for mortality among heart failure patients.

Key-words: heart failure, Cox model, parametric models, Survival analysis

Índice

1 Introdução 1
2 Análise de Sobrevivência 3
2.1 Introdução 3
2.2 Conceitos Básicos 3
2.2.1 Censura 3
2.2.2 Truncatura 4
2.2.3 Função de Sobrevivência 5
2.2.4 Função de Risco 5
2.2.5 Função de Risco Cumulativa 5
2.3 Estimação não paramétrica 6
2.3.1 Estimativa de Kaplan-Meier 7
2.3.2 Teste log-rank 8
2.4 Distribuições de Probabilidade 9
2.4.1 Distribuição Exponencial 9
2.4.2 Distribuição de Weibull 9
2.4.3 Distribuição Gama 10
2.4.4 Distribuição log-normal 10
2.4.5 Distribuição log-logística 11
2.4.6 Distribuição de Gompertz 11
2.5 Covariáveis 11
2.6 Modelos de Regressão 12
2.6.1 Introdução 12
2.6.2 Modelo de Cox 14
2.6.3 Modelos Paramétricos 17
2.7 Análise de Resíduos 18
2.7.1 Resíduos de Cox-Snell 19
2.7.2 Resíduos Padronizados 19
2.7.3 Resíduos de Schoenfeld 19
2.7.4 Resíduos martingala 20
2.7.5 Resíduos deviance 20
3 Análise Estatística 21
3.1 Descrição dos dados 21
3.2 Estimação não paramétrica 24
3.3 Modelo de Cox 28
3.4 Abordagem paramétrica 31
4 Discussão 37

5 Conclusão 39
6 Referências bibliográficas 40

Lista de Figuras

Figura 3.1 Estimativas de Kaplan-Meier por (a) género, (b) fumador/não fumador e (c) diabético/não diabético 24
Figura 3.2 Estimativas de Kaplan-Meier por (a) não ter/ter pressão arterial elevada e (b) não anémico/anémico 25
Figura 3.3 Estimativas de Kaplan-Meier por nível de (a) fracção de ejecção, (b) sódio e (c) plaquetas 26
Figura 3.4 Estimativas de Kaplan-Meier por nível de creatinina e por género 27
Figura 3.5 Estimativas de Kaplan-Meier por nível de CPK e por género 27
Figura 3.6 *Output* do modelo de Cox ajustado com todas as variáveis em estudo 28
Figura 3.7 *Output* do modelo de Cox ajustado com todas as variáveis significativas 29
Figura 3.8 Teste da proporcionalidade dos riscos 29
Figura 3.9 Resíduos de Schoenfeld do modelo de Cox ajustado 30
Figura 3.10 Resíduos deviance do modelo de Cox ajustado 31
Figura 3.11 Adequação dos modelos paramétricos 31
Figura 3.12 Modelo ajustado utilizando a distribuição Exponencial 32
Figura 3.13 Modelo ajustado utilizando a distribuição de Weibull 33
Figura 3.14 Modelo ajustado utilizando a distribuição log-normal 33
Figura 3.15 Modelo ajustado utilizando a distribuição log-logística 34
Figura 3.16 Resíduos de Cox-Snell dos quatro modelos ajustados 35
Figura 3.17 Resíduos deviance dos quatros modelos ajustados 36

Lista de Tabelas

Tabela 2.1 Tabela de Contingência 8
Tabela 3.1 Frequências absolutas de cada variável numérica de acordo com o seu nível.....22
Tabela 3.2 Frequência absoluta e respectiva percentagem de cada variável categórica de censurados versus não censurados.....23
Tabela 3.3 Média das variáveis numéricas de censurados versus não censurados 23

1 Introdução

De acordo com a Federação Mundial do Coração (2021), a insuficiência cardíaca é a principal causa de internamentos hospitalares a nível mundial, afectando mais de 64 milhões de pessoas em todo o mundo.

A insuficiência cardíaca é uma doença grave e crónica, que ocorre quando o coração não é capaz de bombear a quantidade de sangue necessária para o corpo, nem de relaxar e receber novamente o sangue de forma normal. Isto acontece porque as paredes do coração espessam, impedindo, assim, a passagem do sangue, o que significa que o organismo pode ficar sem receber o oxigénio e os nutrientes suficientes, deixando de funcionar normalmente.

Existem diversos factores que podem originar a insuficiência cardíaca, tais como problemas de saúde relacionados com o coração (como por exemplo: elevada pressão arterial, ataque cardíaco ou doença das artérias coronárias), tabagismo, diabetes, colesterol elevado, histórico familiar de doença cardíaca… Apesar da insuficiência cardíaca poder aparecer em qualquer idade, é uma doença muito associada ao envelhecimento da população, sendo a maior causa de internamentos hospitalares para indivíduos acima dos 65 anos (SNS24, 2025).

Dificuldade em respirar, taquicardia constante, inchaço nas pernas e cansaço extremo são alguns dos sintomas que devem alertar os doentes a procurarem ajuda médica.

Apesar de não existir cura para a insuficiência cardíaca, o tratamento certo pode reduzir os sintomas, melhorar a qualidade de vida do doente e ajudar a que este viva por mais anos.

No presente trabalho, serão analisados dados de pacientes com insuficiência cardíaca que foram acompanhados no Instituto de Cardiologia do Hospital *Allied*, no Paquistão, entre os meses de Abril e Dezembro do ano de 2015.

O Paquistão situa-se no sul do continente asiático e é o quinto país mais populoso do mundo. De acordo com um artigo publicado, em Abril de 2023, pela Associação Americana do Coração, a incidência das doenças cardiovasculares neste país está a aumentar.

Um artigo publicado, em Setembro de 2024, no jornal *The Express Tribune*, menciona que estudos já realizados demonstraram que os paquistaneses sofrem, frequentemente, de ataques cardíacos e outras doenças cardiovasculares, até uma década mais cedo do que as populações dos países ocidentais.

Escolhas alimentares menos saudáveis, o acesso dificultado a cuidados de saúde, conflitos regionais e catástrofes naturais são algumas das causas que contribuem para o aumento da incidência das doenças cardiovasculares no Paquistão.

Perante esta situação, o estudo da saúde cardiovascular torna-se crucial. Identificar e estudar os factores de risco pode ser uma forma de prevenção da doença e, por consequência, diminuir o número de casos da mesma.

Os dados provêm de um artigo científico, “Survival analysis of heart failure patients: A case study”, realizado no Paquistão, no ano de 2017, onde foram analisados com o objectivo de estimar as taxas de mortalidade por insuficiência cardíaca e identificar os principais factores de risco, recorrendo ao Modelo de Cox.

Neste estudo, pretende-se fazer uma análise estatística aos dados, com recurso ao *package survival* do *software* R, utilizando modelos de regressão paramétricos de Análise de Sobrevivência, de forma a identificar os factores que influenciam o tempo de vida. Os modelos de regressão paramétricos, se os pressupostos forem verificados, são mais precisos, e os estimadores de parâmetros obtidos são mais eficientes do que os obtidos nos modelos não paramétricos. De forma a comparar os resultados obtidos, pretende-se fazer, também, uma análise não paramétrica aos dados, utilizando estimativas não paramétricas e o modelo semi-paramétrico de Cox.

Este trabalho está estruturado da seguinte forma:

No segundo capítulo, é feita uma introdução à Análise de Sobrevivência, com a definição dos principais conceitos e funções utilizados, assim como às distribuições de probabilidade e aos métodos de estimação não paramétricos mais comuns. São, ainda, abordados os modelos de regressão paramétricos e o modelo semi-paramétrico de Cox, incluindo o método de selecção de variáveis e a análise de resíduos, de modo a fornecer o enquadramento necessário para a aplicação prática dos mesmos.

No terceiro capítulo, são descritos os dados em estudo e introduzidas as variáveis em análise, com uma breve descrição das variáveis clínicas, de forma a contextualizar o seu significado clínico. Segue-se uma análise descritiva e a apresentação dos resultados obtidos nos modelos aplicados.

No capítulo quatro, são discutidos os resultados obtidos nos diferentes modelos ajustados e é feita uma comparação com os resultados do artigo de onde provêm os dados.

Por último, no quinto capítulo, são enumerados os factores que se revelaram significativos no tempo de vida dos pacientes com insuficiência cardíaca.

2 Análise de Sobrevivência

2.1 Introdução

A Análise de Sobrevivência é a área da Estatística que se ocupa do estudo de dados de sobrevivência. Estes dados representam tempos de vida de indivíduos pertencentes a uma determinada população. Designa-se por tempo de vida ou tempo de sobrevivência o tempo decorrido desde um instante inicial previamente definido até à ocorrência de um acontecimento de interesse. Os métodos estatísticos da Análise de Sobrevivência podem ser aplicados em variadíssimas áreas e, como tal, o acontecimento de interesse pode assumir diversas formas. Recaída após um estado de remissão, morte de um indivíduo, fim de um período de desemprego, falha de componentes electrónicas são alguns exemplos do que poderá ser um acontecimento de interesse.

O que distingue a Análise de Sobrevivência das demais áreas da Estatística é a possibilidade da existência de dados censurados, isto é, quando não se observa o acontecimento de interesse durante o período em que os indivíduos estão sob observação. Noutras áreas, estes dados não são tidos em conta, logo há perda de informação, mas tal não acontece na Análise de Sobrevivência, onde existem métodos que permitem que este tipo de dados seja incluído no estudo. Para além da censura, existe também a truncatura, que ocorre aquando da exclusão de indivíduos do estudo.

O tempo de vida de um indivíduo pode ser influenciado por diversos factores, designados por factores de risco ou de prognóstico. Através de modelos de regressão, é possível estudar o efeito que estes factores têm no tempo de vida, sendo estes considerados como variáveis independentes que poderão influenciar a variável dependente, o tempo de sobrevivência.

2.2 Conceitos Básicos

Nesta secção, são apresentados alguns conceitos e funções bastante utilizados na Análise de Sobrevivência.

Para as funções que serão definidas, considere-se a variável aleatória T, contínua e não negativa, que representa o tempo de vida de um indivíduo pertencente a uma dada população homogénea, ou seja, não são considerados factores que representam a heterogeneidade da população.

2.2.1 Censura

Diz-que um dado é censurado quando o acontecimento de interesse não foi observado durante o período em que o indivíduo esteve em estudo, o que significa que não se sabe o valor exacto do tempo de vida desse indivíduo. Existem diferentes tipos de censura:

Censura à direita

Designa-se por censura à direita quando o tempo de vida é superior ao valor registado. Por exemplo, num estudo clínico, com data final previamente definida, se a morte por uma determinada doença for o acontecimento de interesse, serão censurados à direita os tempos de vida dos indivíduos que sobreviveram até à data final do estudo (também designada por censura administrativa), dos indivíduos que morreram por outra causa que não a doença em

estudo e dos indivíduos que são perdidos para o *follow-up* (quando se perde contacto com o indivíduo). A censura à direita pode ser ainda dividida em três tipos:

Tipo I: onde os períodos de observação são previamente definidos pelo investigador.

Tipo II: o estudo termina após ter sido observado o acontecimento de interesse pela n-ésima vez, sendo n um número pré-determinado pelo investigador.

Censura aleatória: o estudo tem uma data final previamente definida pelo investigador, mas a entrada dos indivíduos no estudo é feita de forma aleatória.

Censura à esquerda

Designa-se por censura à esquerda quando o tempo de vida é inferior ao valor registado, ou seja, não se tem conhecimento do momento exacto em que ocorreu o acontecimento de interesse, mas sabe-se que aconteceu antes do valor registado. Por exemplo, se o acontecimento de interesse for a idade com que uma criança realizou uma tarefa pela primeira vez e, à data de entrada no estudo, a criança já realizou essa tarefa, a observação correspondente será censurada à esquerda, pois sabe-se que o acontecimento de interesse ocorreu, mas não se sabe a idade precisa em que tal sucedeu.

Censura intervalar

Designa-se por censura intervalar quando não se sabe o momento exacto em que ocorreu o acontecimento de interesse, mas sabe-se que se deu num determinado intervalo de tempo. Trata-se de censura intervalar – caso I - se a única informação que se dispõe, num determinado instante de monitorização, é se o acontecimento de interesse ocorreu ou não. Este tipo de dados designa-se por dados do estado actual. Trata-se de censura intervalar – caso II - quando se sabe o intervalo de tempo em que ocorreu o acontecimento de interesse.

2.2.2 Truncatura

Diz-se que existe truncatura quando, por um processo de selecção inerente ao planeamento do estudo, são apenas estudados indivíduos a quem tenha ocorrido o mesmo acontecimento. Tal como na censura, existe truncatura à esquerda e truncatura à direita.

Truncatura à esquerda

Existe truncatura à esquerda quando são incluídos no estudo apenas os indivíduos que satisfazem uma determinada condição antes da ocorrência do acontecimento de interesse, ou seja, são observados os indivíduos cujo tempo de vida é superior ao instante em que ocorreu a condição necessária. Este tipo de truncatura favorece o aparecimento de tempos de sobrevivência mais longos pelo facto de indivíduos terem de sobreviver até um determinado instante. Por exemplo, se se pretender estudar o tempo de vida até à ocorrência de um acidente de viação e considerar apenas os indivíduos que possuem carta de condução, os que não possuem são excluídos do estudo.

Truncatura à direita

Existe truncatura à direita se o acontecimento de interesse ocorreu antes de uma certa data, ou seja, o tempo de vida é inferior a um determinado valor conhecido. Por exemplo, se se pretender estudar o tempo de vida até à ocorrência de um acidente de viação e se se considerar apenas os indivíduos que possuem carta de condução há, no máximo, 5 anos, os que possuem carta há mais de 5 anos são excluídos do estudo.

2.2.3 Função de Sobrevivência

Define-se função de sobrevivência como sendo a probabilidade de um indivíduo sobreviver para além do instante *t*. Denota-se por $S(t)$ e é dada pela expressão:

$$S(t) = P(T \geq t)\,, t \geq 0 \qquad (2.1)$$

Trata-se de uma função monótona decrescente e contínua. Quando $t = 0$, a função toma o valor um, ou seja, no instante inicial, todos os indivíduos estão vivos. Quando $t \to +\infty$, a função aproxima-se do valor zero, ou seja, a probabilidade de o indivíduo sobreviver decresce ao longo do tempo.

A função densidade de probabilidade no instante t representa a taxa instantânea de morte nesse instante.

$$f(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t+dt)}{dt} \qquad (2.2)$$

2.2.4 Função de Risco

A função de risco (*hazard function*), também conhecida por função intensidade, taxa de falha ou força de mortalidade, é a taxa instantânea de morte no instante t, sabendo que o indivíduo sobreviveu até esse instante e é dada pela expressão:

$$h(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t+dt \mid T \geq t)}{dt} \qquad (2.3)$$

E satisfaz as seguintes propriedades:

$$h(t) \geq 0$$

$$\int_0^\infty h(t)dt = \infty$$

2.2.5 Função de Risco Cumulativa

Pode-se ainda definir a função de risco cumulativa, que representa o risco de ocorrência do acontecimento de interesse até ao instante t. Trata-se de uma função não negativa e monótona crescente e é definida pela expressão seguinte:

$$H(t) = \int_0^t h(u)du \ , t \geq 0 \quad (2.4)$$

Como a função de risco descreve a probabilidade instantânea de morte de um indivíduo ao longo do tempo, a forma desta função corresponderá às alterações do risco de morte ao longo do tempo.

As formas mais comuns da função de risco são as seguintes:

Monótona crescente: situação mais comum, onde a probabilidade de sobreviver diminui ao longo do tempo. Isto acontece, por exemplo, quando os indivíduos são seguidos durante um período da sua vida, no qual ocorre o envelhecimento gradual.

Monótona decrescente: situação menos comum, em que a probabilidade de sobreviver aumenta ao longo do tempo. Pode acontecer em casos como o de bebés que têm de ser submetidos a uma intervenção cirúrgica à nascença.

Constante: a probabilidade de sobrevivência não se altera ao longo do tempo. Pode acontecer se o tempo de vida representar o tempo até à ocorrência de um acidente ou de uma doença rara.

***Bathub-shaped*:** ao início, a função de risco decresce; durante um período, mantém-se constante; e, a partir de um determinado instante, passa a ser crescente, fazendo lembrar a forma de uma banheira, daí a sua designação. Tal acontece se o indivíduo for seguido desde o nascimento até à morte real.

***Hump-shaped*:** a função de risco assume a forma de uma bossa, ou seja, é crescente inicialmente e, ao fim de algum tempo, torna-se decrescente. Tal pode acontecer em doentes sujeitos a uma cirurgia, onde, no pós-operatório, o seu estado de saúde tenha piorado e, à medida que vão recuperando, o risco de morte vai diminuindo.

2.3 Estimação não paramétrica

Uma análise de um conjunto de dados inicia-se, por norma, por uma estatística sumária das variáveis. No caso dos dados de sobrevivência, utiliza-se, também, um estimador da função de sobrevivência. O mais utilizado na área da Análise de Sobrevivência é a estimativa de Kaplan-Meier. A sua representação gráfica permite ter uma ideia do comportamento da curva de sobrevivência, ao longo do tempo. Se se pretender comparar a distribuição de vida de diversos grupos, a representação gráfica, para cada grupo, da estimativa de Kaplan-Meier, permite avaliar, de um modo informal, se existem diferenças entre os grupos comparados, relativamente à sua curva de sobrevivência. Mas, para uma análise mais rigorosa, é necessário recorrer a um teste de hipóteses. O mais usual e mais potente é o teste log-rank.

Estes métodos são designados por não paramétricos por não assumirem uma distribuição subjacente aos tempos de vida.

2.3.1 Estimativa de Kaplan-Meier

Se não existirem dados censurados, a função de sobrevivência é dada pela proporção de indivíduos que sobreviveram para além do instante t.

$$\hat{S}(t) = \frac{nº\ de\ indivíduos\ vivos\ no\ instante\ t}{nº\ de\ indivíduos\ no\ início\ do\ estudo}, t \geq 0 \quad (2.5)$$

Na presença de dados censurados, a função de sobrevivência é dada pela estimativa de Kaplan-Meier. Trata-se de uma função em escada, que decresce a cada instante de morte distinto.

Considere-se uma amostra de dimensão n, onde $t_{(1)}, \dots, t_{(r)}$ correspondem aos instantes de morte ordenados e distintos e $r \leq n$. O número de indivíduos em risco imediatamente antes de $t_{(i)}, i = 1, \dots, r$ denota-se por n_i e o número de mortes ocorridas nesse mesmo instante denota-se por d_i. A probabilidade de um indivíduo sobreviver para além do instante $t_{(1)}$, dado que sobreviveu até esse instante, é a razão entre o número de indivíduos em risco nesse instante sobre o número de indivíduos em risco no início do estudo. A probabilidade de um indivíduo sobreviver para além do instante $t_{(2)}$ será o produto da probabilidade de ter sobrevivido até $t_{(1)}$ pela probabilidade de ter sobrevivido até $t_{(2)}$, dado que sobreviveu até $t_{(1)}$. A estimativa de Kaplan-Meier é, então, dada pela expressão

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (2.6)$$

Até ao primeiro instante de morte, $\hat{S}(t)$ toma o valor um. Se a maior observação registada não for censurada, $\hat{S}(t)$ toma o valor zero. Se a maior observação registada for censurada, $\hat{S}(t)$ nunca toma o valor zero e não está definida para além desse instante, está definida apenas até ao último instante de morte.

A estimativa da variância de $\hat{S}(t)$ é dada pela expressão

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

denominada por fórmula de Greenwood.

Pode-se construir um intervalo de confiança para o verdadeiro valor de $\hat{S}(t)$ no instante t_0, dado que $\hat{S}(t)$ tem uma distribuição assintótica normal de valor médio $S(t)$ e variância dada pela fórmula de Greenwood. Um intervalo de $100(1-\alpha)\%$ de confiança para $\hat{S}(t_0)$ é dado por

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

Apesar de ser o intervalo mais utilizado, não está isento de problemas, principalmente por ser simétrico. Se a estimativa de $\hat{S}(t_0)$ estiver próxima de zero ou de um, os limites do intervalo podem estar fora do intervalo (0,1). Uma solução passaria por substituir o limite inferior do intervalo por zero ou o limite superior por um. Uma alternativa seria obter um intervalo de confiança para uma transformação de $\hat{S}(t_0)$, como por exemplo $\log[-\log \hat{S}(t_0)]$.

2.3.2 Teste log-rank

O teste log-rank permite testar se existem diferenças significativas entre as funções de sobrevivência de diferentes grupos.

Considerem-se dois grupos de indivíduos. Sejam $t_{(1)} < \cdots < t_{(r)}$ os instantes de morte distintos, d_j o número de indivíduos que morreram no instante t_j, d_{ij} o número de indivíduos do grupo $i, i = 1{,}2$ que morreram no instante t_j, n_j o número de indivíduos em risco imediatamente antes do instante t_j e n_{ij} o número de indivíduos do grupo $i, i = 1{,}2$ em risco imediatamente antes do instante t_j. Esta informação está resumida na tabela seguinte:

Tabela 2.1 Tabela de Contingência

Grupo	*Nº de mortes em* t_j	*Nº de sobreviventes para além de* t_j	*Nº de indivíduos em risco no instante* t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
	d_j	$n_j - d_j$	n_j

As hipóteses a testar são:

$$H_0: S_1(t) = S_2(t) \; vs \; H_1: S_1(t) \neq S_2(t)$$

Mantel e Haenszel (1959) consideraram a distribuição condicional das frequências observadas em cada célula, dados os totais marginais, sob a validade da hipótese nula. Supondo H_0 verdadeira, a distribuição de d_{1j}, condicional aos valores das marginais é hipergeométrica e é dada pela expressão

$$p\left(d_{1j} | d_j, n_j, n_{1j}\right) = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}$$

O valor médio e a variância condicionais de d_{1j}, sob a hipótese nula, são:

$$e_{1j} = \frac{n_{1j} d_j}{n_j}$$

$$v_{1j} = \frac{n_{1j} n_{2j} d_j \left(n_j - d_j\right)}{n_j^2 \left(n_j - 1\right)}$$

respectivamente.

Para obter uma medida global do desvio dos valores observados de d_{1j}, relativamente aos valores esperados, considere-se

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

com valor médio 0 e variância $\sum_{j=1}^{k}(v_{1j})$.

A estatística de teste é dada por

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

que, sob H_0, tem distribuição assintótica χ_1^2.

2.4 Distribuições de Probabilidade

Nesta secção, são apresentadas as distribuições de probabilidade mais comuns na área da Análise de Sobrevivência.

2.4.1 Distribuição Exponencial

Seja T uma variável aleatória que segue uma distribuição exponencial de parâmetro λ ($\lambda > 0$), que tem como função densidade de probabilidade, para $t \geq 0$,

$$f(t) = \lambda e^{-\lambda t} \qquad (2.7)$$

As funções de sobrevivência e de risco são dadas por

$$S(t) = e^{-\lambda t} \qquad (2.8)$$

$$h(t) = \lambda \qquad (2.9)$$

respectivamente, para $t \geq 0$.

Esta distribuição caracteriza-se por ter uma função de risco constante, ou seja, o risco de morte não se altera ao longo do tempo. Uma das características da distribuição exponencial é a "falta de memória", o que justifica a função de risco ser constante.

2.4.2 Distribuição de Weibull

Seja T uma variável aleatória que segue uma distribuição de Weibull com parâmetro de escala λ ($\lambda > 0$) e parâmetro de forma γ ($\gamma > 0$), tem-se, para $t \geq 0$, a função densidade de probabilidade

$$f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}} \qquad (2.10)$$

As funções de sobrevivência e de risco são dadas por

$$S(t) = e^{-\lambda t^{\gamma}} \qquad (2.11)$$

$$h(t) = \lambda\gamma t^{\gamma-1} \qquad (2.12)$$

respectivamente, para $t \geq 0$.

A função de risco pode tomar três formas dependendo do valor de γ:

- Se $\gamma > 1$, a função de risco será monótona crescente;
- Se $0 < \gamma < 1$, a função de risco será monótona decrescente;

- Se $\gamma = 1$, a função de risco será constante (corresponde à distribuição exponencial).

Pela função de risco ser tão flexível, a distribuição de Weibull é provavelmente o modelo mais utilizado na Análise de Sobrevivência.

2.4.3 Distribuição Gama

Seja T uma variável aleatória que segue uma distribuição gama com parâmetro de escala λ ($\lambda > 0$) e parâmetro de forma α ($\alpha > 0$), a função densidade de probabilidade é dada por, para $t \geq 0$,

$$f(t) = \frac{\lambda(\lambda t)^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} \qquad (2.13)$$

A função de sobrevivência é dada pela expressão

$$S(t) = 1 - I(\alpha, \lambda t) \qquad (2.14)$$

sendo $\mathrm{I}(\alpha, x)$ a função gama incompleta, que é definida por

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

Dependendo do valor de α, a função de risco pode tomar as seguintes formas:

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

2.4.4 Distribuição log-normal

Seja T uma variável aleatória que segue uma distribuição log-normal, se $\log T$ tem distribuição normal com valor médio μ e variância σ^2, a função densidade de probabilidade de T é dada por

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} exp\left[-\frac{1}{2}\left(\frac{log\, t - \mu}{\sigma}\right)^2\right] \qquad (2.15)$$

A função de sobrevivência é dada por

$$S(t) = 1 - \Phi\left(\frac{log\, t - \mu}{\sigma}\right) \qquad (2.16)$$

onde $\Phi(.)$ é a função de distribuição da Normal (0,1).

A função de risco é unimodal. Quando $t = 0$, $h(t) = 0$. A função de risco é crescente desde zero até um valor máximo, que depende do valor de σ, e passa a ser decrescente com $\lim_{t\to\infty} h(t) = 0$.

2.4.5 Distribuição log-logística

Seja T uma variável aleatória que segue uma distribuição log-logística com parâmetro de escala λ $(\lambda > 0)$ e parâmetro de forma α $(\alpha > 0)$, a função densidade de probabilidade, para $t \geq 0$, é dada por

$$f(t) = \frac{\alpha\lambda t^{\alpha-1}}{(1+\lambda t^{\alpha})^2} \quad (2.17)$$

As funções de sobrevivência e de risco são dadas por

$$S(t) = \frac{1}{1+\lambda t^{\alpha}} \quad (2.18)$$

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}} \quad (2.19)$$

respectivamente, para $t \geq 0$.

A função de risco pode tomar as seguintes formas:

- Se $\alpha > 1$, a função de risco é crescente desde zero até um valor máximo, dado pela expressão $t = \left(\frac{\alpha-1}{\lambda}\right)^{\frac{1}{\alpha}}$, e passa a ser decrescente com $\lim_{t\to\infty} h(t) = 0$;
- Se $0 < \alpha < 1$, a função de risco é monótona decrescente.

2.4.6 Distribuição de Gompertz

Seja T uma variável aleatória que segue uma distribuição Gompertz, a função densidade de probabilidade, para $t \geq 0$ e $\theta > 0$, é dada por

$$f(t) = \theta exp(\alpha t) exp\left\{\frac{\theta}{\alpha}[1 - exp(\alpha t)]\right\} \quad (2.20)$$

As funções de sobrevivência e de risco são dadas por

$$S(t) = exp\left\{\frac{\theta}{\alpha}[1 - exp(\alpha t)]\right\} \quad (2.21)$$

$$h(t) = \theta exp(\alpha t) \quad (2.22)$$

respectivamente, para $t \geq 0$ e $\theta > 0$.

A função de risco é monótona crescente quando $\alpha > 0$ e monótona decrescente quando $\alpha < 0$.

2.5 Covariáveis

Como já foi referido anteriormente, o tempo de vida pode ser influenciado por diversos factores, designados por factores de risco ou de prognóstico. Estes factores podem ser propriedades intrínsecas dos indivíduos, tratamentos ou variáveis exógenas. Para estudar o efeito dos mesmos, sempre que possível, regista-se, para cada indivíduo, os valores dessas variáveis, denominadas covariáveis ou variáveis explicativas, que representam os factores que se acredita que afectem o tempo de vida e que fornecem informação acerca da heterogeneidade existente na população.

As covariáveis podem ser constantes, se o valor permanece inalterado ao longo do período em que o indivíduo se encontra sob observação (como por exemplo, o género do indivíduo ou uma variável que indica o tratamento a que o indivíduo está sujeito), ou dependentes do tempo, se o valor se altera ao longo do período de observação (como por exemplo, a pressão arterial, se for medida em momentos diferentes ao longo do estudo). As covariáveis dependentes do tempo podem ainda ser divididas em externas ou internas. Uma covariável classifica-se como externa, se não está directamente relacionada com o mecanismo que regula a morte dos indivíduos, ou interna, se resulta de uma medição feita ao indivíduo durante o estudo, ou seja, é apenas observada num indivíduo que está vivo e que não foi censurado, fornecendo assim informação acerca do seu tempo de sobrevivência. As covariáveis externas podem ainda ser divididas em fixas, caso o seu valor não se altere ao longo do estudo; definidas, se o seu valor é pré-determinado (por exemplo, um factor sob controlo do investigador que o faz variar de forma pré-determinada, ao longo do estudo); ou ancilárias, quando é o resultado de um processo que é exterior ao indivíduo (por exemplo, o nível de poluição atmosférica num estudo de ocorrência de ataques de asma).

2.6 Modelos de Regressão

2.6.1 Introdução

Para estudar o efeito dos possíveis factores de risco no tempo de vida, utiliza-se um modelo de regressão, onde a variável dependente é o tempo de vida e as variáveis independentes correspondem às covariáveis. Para tal, é necessário especificar para T um modelo de distribuição dado o vector de covariáveis $\boldsymbol{z} = (z_1, z_2, \dots, z_p)'$ associado a um determinado indivíduo, o que pode ser feito utilizando distribuições de famílias paramétricas ou semi-paramétricas. Devido à sua flexibilidade, o modelo de Cox, um modelo semi-paramétrico, é o mais utilizado na análise do tempo de vida. No entanto, por vezes, é possível admitir um modelo paramétrico para o tempo de vida.

Na Análise de Sobrevivência, os modelos podem ser divididos em três classes, descritas abaixo:

Modelos com funções de risco proporcionais
Apesar dos indivíduos apresentarem valores diferentes nas covariáveis, as funções de risco correspondentes são proporcionais.
Para dois indivíduos com vector de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, a razão das funções de risco não depende de t, logo não varia ao longo do tempo.

$$\frac{h(t;\boldsymbol{z}_1)}{h(t;\boldsymbol{z}_2)}$$

A função de risco de T, dado $\boldsymbol{z}$, pode ser escrita da seguinte forma

$$h(t;\boldsymbol{z}) = h_0(t)\gamma(\boldsymbol{z})$$

onde $\gamma(\mathbf{0}) = 1$ e $h_0(t)$ representa a função de risco de um indivíduo a que está associado o vector $\boldsymbol{z} = \mathbf{0}$ (designa-se por indivíduo padrão).

O factor de proporcionalidade $\gamma(\boldsymbol{z})$, designado por risco relativo, representa a razão entre o risco de morte de um indivíduo a que esteja associado o vector de covariáveis $\boldsymbol{z}$ e o risco de morte de um indivíduo a que esteja associado o vector $\boldsymbol{z} = \mathbf{0}$.

$$\gamma(\boldsymbol{z}) = \frac{h(t;\boldsymbol{z})}{h_0(t)}$$

A função de sobrevivência de T, dado $\boldsymbol{z}$, é dada pela seguinte expressão

$$S(t;\boldsymbol{z}) = S_0(t)^{\gamma(\boldsymbol{z})}$$

Neste tipo de modelos, as covariáveis têm um efeito multiplicativo na função de risco.

Modelos de tempo de vida acelerado

Também chamados de modelos log-lineares para T. Neste modelo, a variável T é dada por

$$T = \frac{T_0}{\psi(\boldsymbol{z})}$$

onde a T_0 corresponde a função de sobrevivência a que está associado o vector de covariáveis $\mathbf{0}$. Para um indivíduo com vector de covariáveis $\boldsymbol{z}$, as funções de risco e de sobrevivência são dadas, respectivamente, pelas expressões seguintes:

$$h(t;\boldsymbol{z}) = h_0\big(t\psi(\boldsymbol{z})\big)\psi(\boldsymbol{z})$$

$$S(t;\boldsymbol{z}) = S_0\big(t\psi(\boldsymbol{z})\big)$$

onde $h_0(t)$ e $S_0(t)$ são, respectivamente, as funções de risco e de sobrevivência do indivíduo padrão.

Neste modelo, as covariáveis têm um efeito multiplicativo em t, ou seja, a sua função é acelerar (ou travar) o tempo até à morte do indivíduo, em relação ao indivíduo padrão.

Na forma logarítmica, o modelo é dado por

$$\log T = \mu + \boldsymbol{\alpha}'\boldsymbol{z} + \sigma\varepsilon$$

onde μ é o termo independente, $\boldsymbol{\alpha}$ é um vector de parâmetros de regressão, σ é um parâmetro de escala e ε é uma variável aleatória que representa o erro e cuja distribuição não depende de $\boldsymbol{z}$.

A variável T_0 é dada por $e^{\mu+\sigma\varepsilon}$ com função de sobrevivência S_0.

$$\log T_0 = \mu + \boldsymbol{\alpha}'\mathbf{0} + \sigma\varepsilon \Leftrightarrow T_0 = e^{\mu+\sigma\varepsilon}$$

Para um indivíduo com vector de covariáveis $\boldsymbol{z}$, a função de sobrevivência é dada por

$$S(t;\boldsymbol{z}) = S_0\big(te^{-\boldsymbol{\alpha}'\boldsymbol{z}}\big)$$

Neste caso, o efeito das covariáveis consiste numa modificação da escala do tempo através do factor de aceleração $\exp(-\boldsymbol{\alpha}'\boldsymbol{z})$. Se este factor for inferior a um, o tempo até à ocorrência do acontecimento de interesse é travado pelas covariáveis. Se for superior a um, o tempo até à ocorrência do acontecimento de interesse é acelerado pelas covariáveis.

Modelos de possibilidades proporcionais

A possibilidade (odds) de sobrevivência para além do instante t é dada pela razão

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

Neste tipo de modelos, a possibilidade de um indivíduo com vector de covariáveis $\boldsymbol{z}$ sobreviver para além do instante t é proporcional à possibilidade do indivíduo padrão sobreviver para além do instante t.

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

onde $\eta = \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p$ em que z_j representa o valor da j-ésima covariável, com $j = 1,2,\ldots,p$.
Neste caso, as covariáveis têm um efeito multiplicativo na possibilidade de um indivíduo sobreviver para além do instante t.
Aplicando o logaritmo à expressão anterior, conclui-se que η é o logaritmo da razão entre a possibilidade de sobrevivência de um indivíduo com vector de covariáveis $\boldsymbol{z}$ e a possibilidade de sobrevivência do indivíduo padrão. Trata-se, então, de um modelo linear.

$$\eta = \log\left[\frac{S(t;\boldsymbol{z})}{1-S(t;\boldsymbol{z})} \Big/ \frac{S_0(t)}{1-S_0(t)}\right]$$

2.6.2 Modelo de Cox

Em 1972, Cox propôs um modelo semi-paramétrico que, devido à sua versatilidade e flexibilidade, se tornou no modelo de regressão mais utilizado na análise de tempos de vida.

Considere-se uma variável aleatória contínua T, que representa o tempo de vida. Para um indivíduo a que esteja associado o vector de covariáveis $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)'$, no instante t, a função de risco é dada por

$$\begin{aligned} h(t;\boldsymbol{z}) &= h_0(t)exp(\boldsymbol{\beta}'\boldsymbol{z}) \\ &= h_0(t)exp(\beta_1 z_1 + \cdots + \beta_p z_p) \end{aligned} \qquad (2.23)$$

onde $\beta_1, \ldots, \beta_p$ são os coeficientes de regressão desconhecidos que representam o efeito das covariáveis na sobrevivência e $h_0(t)$ representa a função de risco de um indivíduo a que está associado o vector $\boldsymbol{z} = \boldsymbol{0}$.

O efeito das covariáveis é modelado parametricamente, mas a função de risco subjacente $h_0(t)$ não é especificada, daí ser um modelo semi-paramétrico.

Trata-se de um modelo de riscos proporcionais, dado que, para dois indivíduos com vector de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, a razão das funções de risco não depende do tempo.

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

As covariáveis têm um efeito multiplicativo na função de risco, de acordo com o factor $exp(\boldsymbol{\beta}'\boldsymbol{z})$, designado por risco relativo.

O modelo de Cox baseia-se no princípio de que a influência das covariáveis na função de risco não se altera ao longo do período em que os indivíduos estão sob observação.

Interpretação dos coeficientes

Considerem-se dois indivíduos aos quais estão associados os vectores de covariáveis $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, respectivamente, que apenas diferem nos valores da covariável z_j. Tem-se que

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

$$= exp\left(\beta_j\left(z_{1j} - z_{2j}\right)\right)$$

O factor $exp\left(\beta_j\right)$ representa o risco relativo de ocorrência de um acontecimento para dois indivíduos que diferem uma unidade nos valores da covariável z_j, mantendo os valores das restantes variáveis constantes.

Função de verosimilhança

Para a inferência sobre β, Cox baseou-se na função de verosimilhança dada por

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(\beta' \boldsymbol{z}_{(i)})}{\sum_{l \epsilon R_i} exp(\beta' \boldsymbol{z}_l)} \qquad (2.24)$$

onde $R_i = R\left(t_{(i)}\right) = \left\{j: t_j \geq t_{(i)}\right\}$ é o conjunto de indivíduos em risco no instante $t_{(i)}$ e $t_{(1)} < \cdots < t_{(k)}$, com $k < n$, os k tempos de vida distintos.

A função $L(\beta)$, considerada por Cox, não é a verosimilhança habitual. Para o modelo de Cox, a função de verosimilhança é da forma

$$L[\beta, h_0(t)] = \prod_{i=1}^{n} \left[h_0(t_i) exp(\beta' \boldsymbol{z}_i) S_0(t_i)^{exp(\beta' \boldsymbol{z}_i)}\right]^{\delta_i} \left[S_0(t_i)^{exp(\beta' \boldsymbol{z}_i)}\right]^{1-\delta_i}$$

$$= \prod_{i \epsilon D} \frac{exp(\beta' \boldsymbol{z}_i)}{\sum_{l \epsilon R_i} exp(\beta' \boldsymbol{z}_l)} \prod_{i \epsilon D} \left(h_0(t_i) \sum_{l \epsilon R_i} exp(\beta' \boldsymbol{z}_l) \right) \prod_{i=1}^{n} S_0(t_i)^{exp(\beta' \boldsymbol{z}_i)}$$

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

Como o primeiro factor de $L[\beta, h_0(t)]$ coincide com $L(\beta)$, esta pode ser considerada como verosimilhança parcial. Como $L(\beta)$ não depende de $h_0(t)$, é possível fazer inferência sobre β sem especificar $h_0(t)$.

Diversos autores consideraram que, sob condições de regularidade bastante gerais, o estimador de máxima verosimilhança parcial de β é consistente e assintoticamente normal com valor médio β e matriz de covariância $I(\beta)^{-1}$, onde

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

Existência de observações censuradas

Na expressão $L(\beta)$ apenas são considerados tempos de vida distintos. Na presença de observações empatadas, pode-se utilizar uma aproximação da função de verosimilhança, proposta por Peto (1972) e Breslow (1974), dada pela expressão

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(\boldsymbol{\beta}' \boldsymbol{s}_i)}{\left[\sum_{l \epsilon R_i} exp(\boldsymbol{\beta}' \boldsymbol{z}_l)\right]^{d_i}} \qquad (2.25)$$

onde $\boldsymbol{s}_i = \sum_{j=1}^{d_i} \boldsymbol{z}_{ij}$ para $i = 1, \ldots, k$.

Estimação da função de sobrevivência

Tendo obtido $\hat{\beta}$ a partir da verosimilhança parcial, Kalbfleisch e Prentice (1973) determinaram um estimador de máxima verosimilhança não paramétrico de $S_0(t)$, que é dado por

$$\hat{S}_0(t) = \prod_{i:t_{(i)} \leq t} \hat{\alpha}_i \qquad (2.26)$$

onde

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

Trata-se de uma função em escada com descontinuidades a cada instante de morte distinto.

Método de selecção de variáveis

Collett (2003) propôs um método, descrito abaixo, para seleccionar o modelo que melhor se ajusta aos dados:

1. Ajustar modelos com apenas uma covariável. Calcular, para cada um dos modelos ajustados, o valor da estatística $-2 \log \hat{L}$ e comparar com o valor da estatística para o modelo nulo. Assim, determinam-se as covariáveis que levam a uma redução significativa da estatística $-2 \log \hat{L}$.
2. Ajustar um modelo com as covariáveis que, no passo anterior, se revelaram significativas, e calcular o valor da estatística $-2 \log \hat{L}$. Retirar uma covariável de cada vez e reter, apenas, aquelas que levam a um aumento significativo do valor da estatística.
3. As variáveis que não foram consideradas no primeiro passo podem revelar-se importantes na presença de outras. Estas variáveis são, então, adicionadas, uma de cada vez, ao modelo obtido no segundo passo. Aquela que reduzir significativamente o valor da estatística $-2 \log \hat{L}$ deverá ser retida.
4. É feita uma verificação final para garantir que nenhuma variável significativa é deixada fora do modelo e que nenhuma variável incluída possa ser omitida.

Collett recomenda, ainda, que se utilize o nível de significância de 10% na decisão de incluir ou omitir covariáveis no modelo.

2.6.3 Modelos Paramétricos

Trata-se de um modelo paramétrico se os tempos de sobrevivência assumirem uma distribuição de probabilidade. Se os pressupostos forem verificados, é possível obter estimadores dos parâmetros mais precisos, comparativamente aos obtidos no Modelo de Cox.

Para escolher um modelo paramétrico, poder-se-ia estimar a função de risco, representá-la graficamente e tomar em conta o seu comportamento, mas não seria o suficiente para verificar a adequação do modelo. Uma forma mais apropriada e informativa será comparar graficamente a estimativa não paramétrica de Kaplan-Meier com a estimativa paramétrica pelo modelo considerado da função de sobrevivência.

De seguida, são apresentados os modelos de regressão paramétricos mais comuns na Análise de Sobrevivência.

Modelo de regressão Weibull

É o único modelo que tanto pode ser considerado como modelo de riscos proporcionais ou como modelo de tempo de vida acelerado.

Formulando o modelo de Weibull como modelo de riscos proporcionais, tem-se, para um indivíduo com vector de covariáveis $\boldsymbol{z}$, a seguinte função de risco

$$h(t;\boldsymbol{z}) = h_0(t)\gamma(\boldsymbol{z}) = \lambda\gamma t^{\gamma-1}exp(\beta'\boldsymbol{z}) \qquad (2.27)$$

onde o tempo de vida tem distribuição de Weibull com parâmetro de escala $\lambda exp(\beta'\boldsymbol{z})$ e parâmetro de forma γ. O efeito das covariáveis é apenas no parâmetro de escala. O parâmetro de forma mantém-se inalterado.

A função de sobrevivência é dada pela expressão

$$S(t;\boldsymbol{z}) = exp(-(\lambda t)^{\gamma}) \qquad (2.28)$$

Modelo de regressão log-logístico

Há situações em que o modelo de Weibull não é o indicado para modelar o tempo de vida. Uma das alternativas é o modelo log-logístico, o único modelo que pode ser formulado como um modelo de possibilidades proporcionais. Para um indivíduo com vector de covariáveis $\boldsymbol{z}$, tem-se a função de sobrevivência

$$S(t;\boldsymbol{z}) = \frac{1}{1+\lambda exp(\beta'\boldsymbol{z})t^{\alpha}} \qquad (2.29)$$

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

Método de selecção de variáveis

Na análise de regressão, pretende-se identificar as variáveis que influenciam de forma significativa o tempo de sobrevivência dos indivíduos. Para tal, um dos métodos mais comuns é a utilização do teste de Wald, onde se avalia se o coeficiente β_j, que representa o efeito da covariável z_j no tempo de sobrevivência do indivíduo, influencia significativamente o tempo de vida. As hipóteses testadas são

$$H_0: \beta_j = 0 \; vs \; H_1: \beta_j \neq 0$$

e a estatística de teste é dada por

$$\frac{\hat{\beta}_j^2}{var(\hat{\beta}_j)}$$

que tem, sob H_0, distribuição assintótica χ_1^2.

Mas as estimativas de $\hat{\beta}_j$ nem sempre são independentes umas das outras, pelo que é útil recorrer a métodos que permitam comparar modelos. Um dos métodos utilizados é o *stepwise*, que consiste na introdução e remoção de covariáveis uma a uma. No primeiro caso, designado por *Forward Selection*, adiciona-se cada covariável uma a uma e testa-se qual delas tem um efeito mais significativo, para que seja, posteriormente, adicionada ao modelo. O processo repete-se até que nenhuma das covariáveis que não foram incluídas tenha um efeito significativo no modelo. No segundo caso, designado por *Backward Selection*, o modelo inicial contém todas as covariáveis e vão-se retirando uma a uma, testando qual delas afecta menos o modelo, sendo esta a que é retirada. O processo repete-se até que qualquer uma das covariáveis que se retire tenha um efeito significativo no modelo. Há, ainda, um terceiro caso, designado por *Both Selection*, que combina as duas primeiras abordagens, permitindo a introdução e exclusão de covariáveis.

Critério de Informação de Akaike

A escolha do modelo mais apropriado, de entre os vários modelos possíveis, não necessariamente aninhados, pode ser feita com base no Critério de Informação de Akaike.

$$AIC = -2 \log \hat{L} + 2(p + 1 + k)$$

onde $\hat{L}$ representa a máxima verosimilhança, p é o número de parâmetros do modelo ajustado, $k = 0$ para o modelo exponencial e $k = 1$ para os modelos Weibull, log-logístico e log-normal.

O melhor modelo será aquele que tiver um menor valor de AIC.

2.7 Análise de Resíduos

Para averiguar a adequabilidade do modelo ajustado, é fundamental fazer-se uma análise de resíduos. Um resíduo consiste na diferença entre o valor observado da variável dependente e o valor predito pelo modelo. Na presença de observações censuradas, a definição não é análoga.

2.7.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell permitem verificar o ajustamento global do modelo e são definidos por

$$e_i = \widehat{H}(t_i; \boldsymbol{z}_i) \qquad (2.30)$$

onde $\widehat{H}$ é a função de risco cumulativa estimada do modelo ajustado.

Os resíduos de Cox-Snell seguem uma distribuição exponencial se o modelo ajustado aos dados for adequado. Na representação gráfica dos pontos $\left(\widehat{e_i}, \widehat{H}_{e_i}(\widehat{e_i})\right)$, se estes estiverem sobre a recta $y = x$, conclui-se que o modelo é adequado.

2.7.2 Resíduos Padronizados

Outra alternativa para avaliar a adequação do modelo remete para os resíduos padronizados, baseados na representação log-linear do modelo, que são definidos por

$$s_i = \frac{log\ t_i - \widehat{\mu} - \widehat{\boldsymbol{\alpha}}\boldsymbol{z}_i}{\widehat{\sigma}} \qquad (2.31)$$

Tal como nos resíduos de Cox-Snell, se os pontos $\left(\widehat{s_i}, \widehat{H}_{s_i}(\widehat{s_i})\right)$ estiverem sobre a bissetriz dos quadrantes ímpares, conclui-se que o modelo é adequado.

2.7.3 Resíduos de Schoenfeld

Para o i-ésimo indivíduo a que corresponde a covariável z_j, o resíduo de Schoenfeld é dado pela expressão

$$r_{ji} = \delta_i\{z_{ji} - a_{ji}\} \qquad (2.32)$$

onde

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

e

$$a_{ji} = \frac{\sum_{l \epsilon R_i} z_{jl} exp\left(\widehat{\boldsymbol{\beta}}' z_l\right)}{\sum_{l \epsilon R_i} exp\left(\widehat{\boldsymbol{\beta}}' z_l\right)}$$

com $j = 1, \ldots, p$ e R_i o conjunto de indivíduos em risco no instante t_i.

Para um indivíduo cuja observação for censurada, o resíduo de Schoenfeld toma o valor zero. Para distinguir estas observações daquelas em que o valor observado é igual ao valor previsto, os resíduos das observações censuradas são tomados como valores omissos.

No caso dos indivíduos cuja morte foi observada, o resíduo consiste na diferença entre o valor da covariável z_j e a média ponderada dos valores dessa covariável para todos os indivíduos em risco no instante t_i. O peso associado a cada indivíduo é $exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_l\right)$, onde l pertence ao conjunto de indivíduos em risco em t_i.

No gráfico dos resíduos de Schoenfeld *versus* tempo de vida, se se observar que os dados se dispõem numa nuvem aleatória de pontos, centrada em zero, o modelo ajustado aos dados é adequado.

2.7.4 Resíduos martingala

Os resíduos martingala permitem verificar a adequabilidade da forma de regressão e são definidos por

$$\hat{m}_i = \delta_i - \hat{e}_i \qquad (2.33)$$

onde δ_i é uma variável indicatriz, que toma o valor um se a observação for não censurada e que toma o valor zero se a observação for censurada, e $\hat{e}_i$ é o resíduo de Cox-Snell dado pela equação (2.30).

No gráfico dos resíduos martingala, se a nuvem de pontos $(x_{ij}, \hat{m}_i)$ tiver uma tendência linear, a variável X_j está bem representada no modelo.

2.7.5 Resíduos deviance

Os resíduos deviance permitem detectar a presença de *outliers* no modelo e são definidos por

$$\hat{d}_i = sgn(\hat{m}_i)\sqrt{-2(\hat{m}_i + \delta_i\, ln(\delta_i - \hat{m}_i))} \qquad (2.34)$$

onde sng é a função sinal e $\hat{m}_i$ é o resíduo martingala do i-ésimo indivíduo. Estes resíduos centram os resíduos martingala em torno de zero, o que torna mais fácil a identificação de *outliers*. Os pontos $(t_i, \hat{d}_i)$ indicam a presença de *outliers* no modelo.

3 Análise Estatística

3.1 Descrição dos dados

Em estudo, estão 299 pacientes, 194 homens e 105 mulheres, com insuficiência cardíaca, que foram seguidos no Instituto de Cardiologia do Hospital *Allied*, na cidade de *Faisalabad*, no Paquistão, entre os meses de Abril e Dezembro, do ano de 2015. O acontecimento de interesse é a morte devido à insuficiência cardíaca, tendo esta sido observada em 96 indivíduos. As observações censuradas correspondem aos indivíduos em que não foi observado o acontecimento de interesse até à data final do período em que estiveram em estudo. Foram também registados os valores das variáveis idade, género, fumador, diabetes, pressão arterial, anemia, fracção de ejecção, sódio, creatinina, plaquetas e CPK, que representam potenciais factores de risco.

Após uma análise descritiva inicial, verificou-se que o período de observação dos pacientes variou entre os 4 e os 285 dias, tendo a sua mediana sido de 115 dias. Em estudo, estão indivíduos com idades compreendidas entre os 40 e os 95 anos, inclusive, sendo a sua mediana igual a 60 anos. Dos 299 indivíduos, 96 (32.1%) são fumadores, 125 (41.8%) são diabéticos e 105 (35.1%) têm pressão arterial elevada. De realçar que há uma discrepância, em relação ao artigo de onde os dados provêm, na frequência absoluta dos pacientes que têm pressão arterial elevada. No artigo, é referido que corresponde a 106 o número de indivíduos com pressão arterial elevada. Optou-se por considerar o valor presente no conjunto de dados utilizado neste estudo.

A anemia é uma condição clínica na qual o número de glóbulos vermelhos e a quantidade de hemoglobina dentro das hemácias no sangue é inferior aos valores considerados normais. Na recolha de dados, a anemia foi avaliada pelo seu nível de hematócrito. O hematócrito é um exame que mede a quantidade, em percentagem, de glóbulos vermelhos no sangue. Um indivíduo é considerado anémico se o seu nível de hematócrito for inferior a 36, que é considerado o valor mínimo normal. Dos 299 pacientes, 129 (43.1%) têm anemia. Tal como na variável da pressão arterial, existe uma discrepância na frequência absoluta, relativamente ao artigo, e o valor a considerar será, mais uma vez, o da base de dados utilizada neste estudo. De acordo com o artigo, há 137 indivíduos anémicos.

A fracção de ejecção (EF) é a quantidade de sangue, em percentagem, que é bombeada pelos ventrículos em cada batimento. De acordo com a Associação Americana do Coração (2025), quando esta percentagem se encontra entre os 50% e os 70%, inclusive, o corpo recebe a quantidade de sangue necessária para funcionar normalmente. Entre os 41% e os 49%, é considerada no limite, e inferior ou igual a 40% é considerada bastante reduzida, o que geralmente indica insuficiência cardíaca. Um valor superior a 70% é considerado elevado e pode ocorrer em pessoas com cardiomiopatia hipertrófica. Dividindo a variável da fracção de ejecção de acordo com as percentagens descritas, constata-se que 219 (73.2%) indivíduos têm uma fracção de ejecção baixa, 20 (6.7%) indivíduos têm a percentagem entre os 41% e os 49%, 59 (19.7%) indivíduos têm a percentagem de fracção de ejecção no intervalo considerado normal e, por fim, há apenas um indivíduo com uma fracção de ejecção superior a 70%, com cerca de 80%.

O sódio é um mineral fundamental na regulação da pressão arterial, pois ajuda a manter a quantidade e a distribuição de água no corpo equilibradas. Os valores ideais de sal no sangue variam entre os 135 e os 145 mEq/L (Tua Saúde, 2025). A ingestão deficiente de sódio pode levar a problemas de saúde, tal como a insuficiência cardíaca. Dividindo a variável sódio, de acordo com os valores referidos, obtém-se que 83 (27.8%) indivíduos têm sódio baixo, 214 (71.6%) indivíduos têm o valor de sódio no intervalo ideal e apenas 2 (0.7%) indivíduos têm sódio alto.

A creatinina é uma substância, presente no sangue, produzida pelos músculos, que serve para produzir energia para a contracção muscular. Os valores de referência da creatinina variam de acordo com a idade e o género. Nos homens, os valores ideais estão entre os 0.7 e os 1.3 mg/dL. Nas mulheres, os valores ideais são entre os 0.6 e os 1.2 mg/dL (Tua Saúde, 2025). Uma das consequências da creatinina alta é a hipertensão arterial, o que pode levar a problemas de insuficiência cardíaca. Formando grupos por género de acordo com os valores referidos, obtém-se que, no grupo dos homens, 3 (1.5%) têm creatinina baixa, 137 (70.6%) têm a creatinina no intervalo dos valores normais e 54 (27.8%) têm creatinina alta; no grupo das mulheres, apenas uma tem creatinina baixa, 70 (66.7%) têm o valor da creatinina no intervalo dos 0.6 aos 1.2 mg/dL e 34 (32.4%) têm creatinina alta.

As plaquetas são as células mais pequenas do sangue e são responsáveis pela coagulação sanguínea. Os valores de referência são 150 000 e 450 000 plaquetas por microlitro de sangue (Tua Saúde, 2024). De acordo com os valores de referência, obtém-se que 27 (9%) indivíduos têm plaquetas baixas, 259 (86.6%) indivíduos têm o número de plaquetas dentro do intervalo de referência e 13 (4.3%) indivíduos têm plaquetas altas. Um elevado número de plaquetas pode ter como consequência problemas cardíacos.

A CPK, sigla de creatinofosfoquinase, é uma enzima presente no cérebro, no coração e nos tecidos musculares. Se ocorrer uma lesão num destes órgãos, esta enzima é liberada no sangue, havendo um aumento da sua concentração. Tal como na creatinina, os valores de referência da CPK variam de acordo com o género. No caso dos homens, os valores de referência são 32 e 294 unidades por litro; no caso das mulheres, os valores de referência são 33 e 211 unidades por litro (Tua Saúde, 2024). Tendo em conta estes valores, no grupo dos homens, obtém-se que apenas 2 homens têm CPK baixa, 101 (52.1%) homens têm a CPK no intervalo de referência e 91 (46.9%) homens têm CPK alta; no grupo das mulheres, nenhuma paciente tem CPK baixa, 47 (44.8%) têm o valor da CPK dentro do intervalo de referência e 58 (55.2%) mulheres têm CPK alta. Um elevado valor de CPK pode levar a problemas cardíacos.

Na tabela seguinte, estão sumarizados os grupos descritos acima com a respectiva frequência absoluta.

Tabela 3.1 Frequências absolutas de cada variável numérica de acordo com o seu nível

		Baixo	No limite	Normal	Elevado
EF		219	20	59	1
Sódio		83		214	2
Creatinina	**Homens**	3		137	54
	Mulheres	1		70	34
Plaquetas		27		259	13
CPK	**Homens**	2		101	91
	Mulheres	0		47	58

Na tabela 3.2, é apresentada, para cada variável categórica, a frequência absoluta e respectiva percentagem do grupo de censurados e do grupo de não censurados.

Tabela 3.2 Frequência absoluta e respectiva percentagem de cada variável categórica de censurados versus não censurados

		Censurados (203)	**Não Censurados (96)**
Género	**Masculino**	132 (65%)	62 (64.6%)
	Feminino	71 (35%)	34 (35.4%)
Fumador	**Sim**	66 (32.5%)	30 (31.2%)
	Não	137 (67.5%)	66 (68.8%)
Diabético	**Sim**	85 (41.9%)	40 (41.7%)
	Não	118 (58.1%)	56 (58.3%)
Pressão Arterial Elevada	**Sim**	66 (32.5%)	39 (40.6%)
	Não	137 (67.5%)	57 (59.4%)
Anemia	**Sim**	83 (40.9%)	46 (47.9%)
	Não	120 (59.1%)	50 (52.1%)

Pode-se constatar que o acontecimento de interesse teve uma maior ocorrência no que respeita ao género, em pacientes do sexo masculino, e na categoria "não", no que respeita às restantes variáveis categóricas.

Na tabela 3.3, para cada variável contínua, é apresentada a média do grupo de indivíduos censurados e a média do grupo de indivíduos não censurados.

Tabela 3.3 Média das variáveis numéricas de censurados versus não censurados

	Censurados (203)	**Não Censurados (96)**
Idade	58.76 anos	65.22 anos
EF	40.27%	33.47%
Sódio	137.22 mEq/L	135.38 mEq/L
Creatinina	1.18 mg/dL	1.84 mg/dL
Plaquetas	266 657.5 plaquetas/µl	256 381plaquetas/µl
CPK	540.05 unidades/L	670.2 unidades/L

A média das idades é superior no grupo dos não censurados, o que não surpreende, dado que a insuficiência cardíaca está associada ao envelhecimento da população. Também os valores de creatinina e de CPK são superiores no grupo de indivíduos que morreu devido à doença em estudo. Nas restantes variáveis numéricas, fracção de ejecção, sódio e plaquetas, os valores são superiores no grupo dos censurados.

3.2 Estimação não paramétrica

Para uma abordagem não paramétrica, foi utilizada a estimativa de Kaplan-Meier, de modo a comparar as funções de sobrevivência de dois ou mais grupos, através da sua representação gráfica, que mostra a estimativa da probabilidade de sobrevivência em diferentes intervalos de tempo e, assim, tornando possível visualizar e aferir se existem diferenças, relativamente à sobrevivência dos indivíduos, entre os grupos comparados. Em caso afirmativo, a variável poderá afectar o tempo de vida. A estimativa de Kaplan-Meier e a respectiva representação gráfica foram, então, construídas para as variáveis categóricas e para as variáveis numéricas agrupadas, tendo em conta os seus valores de referência. A par da estimativa de Kaplan-Meier, foi realizado o teste não paramétrico log-rank. Testou-se a hipótese nula de não existirem diferenças entre os grupos comparados contra a hipótese alternativa de existirem diferenças entre os grupos comparados. Os resultados obtidos são apresentados abaixo.

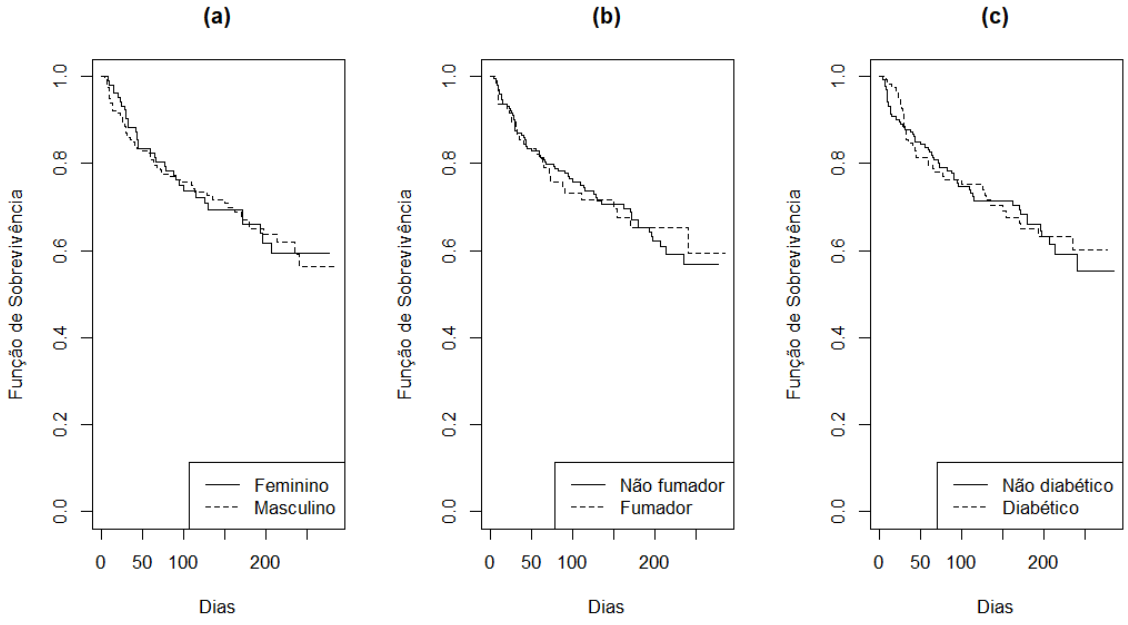

Figura 3.1 Estimativas de Kaplan-Meier por (a) género, (b) fumador/não fumador e (c) diabético/não diabético

Nos três gráficos apresentados na Figura 3.1, as curvas de sobrevivência mantêm-se próximas e cruzam-se diversas vezes, não aparentando haver diferenças significativas. Os valores-p obtidos foram 0.9, 1 e 0.8 para as variáveis género, fumador e diabético, respectivamente, o que leva à não rejeição da hipótese de não existirem diferenças significativas.

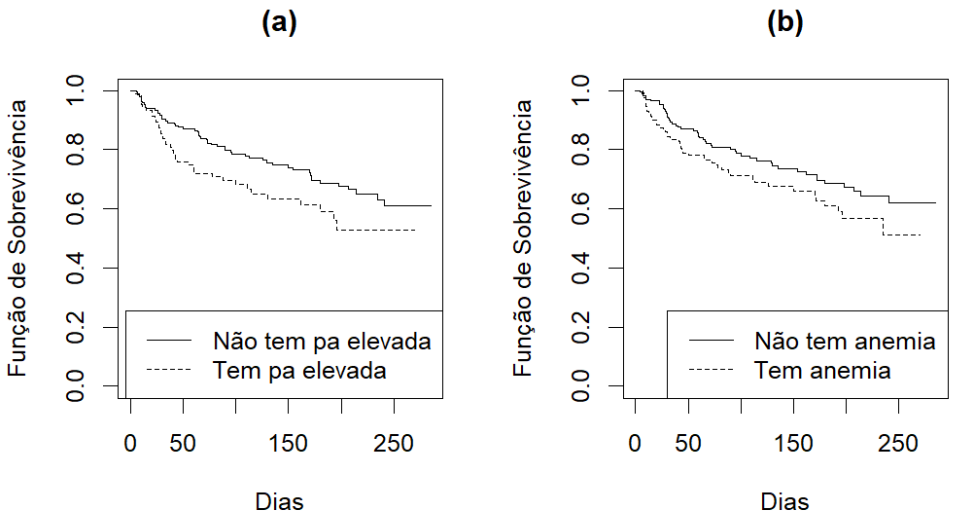

Figura 3.2 Estimativas de Kaplan-Meier por (a) não ter/ter pressão arterial elevada e (b) não anémico/anémico

No gráfico (a) da Figura 3.2, a curva de sobrevivência do grupo de indivíduos que tem pressão arterial elevada é inferior. O valor-p obtido foi 0.04, logo, aos níveis de significância de 5% e 10%, há evidência de que existem diferenças significativas. Na Figura 3.2(b), a curva de sobrevivência dos indivíduos anémicos é inferior. O valor-p 0.1 leva à rejeição da hipótese nula, apenas ao nível de significância de 10%.

De recordar que, aquando do agrupamento dos valores das variáveis numéricas, de acordo com os valores de referência, houve grupos com, no máximo, três elementos. Uma estimativa de Kaplan-Meier com um baixo volume de dados leva a uma estimativa menos confiável. Para combater este problema, os indivíduos pertencentes a estes grupos foram colocados noutro grupo. No caso da variável da fracção de ejecção, como uma percentagem mais baixa pode ter como consequência uma doença cardíaca, o indivíduo com alta percentagem foi colocado no grupo com valores dentro do intervalo normal. O mesmo procedimento foi aplicado na variável sódio. Nas variáveis creatinina e CPK, um valor elevado pode levar a problemas cardíacos, logo os indivíduos pertencentes ao menor grupo foram colocados no grupo que tem os valores dentro do intervalo de referência. De seguida, são apresentados os gráficos obtidos, tendo em conta as alterações feitas.

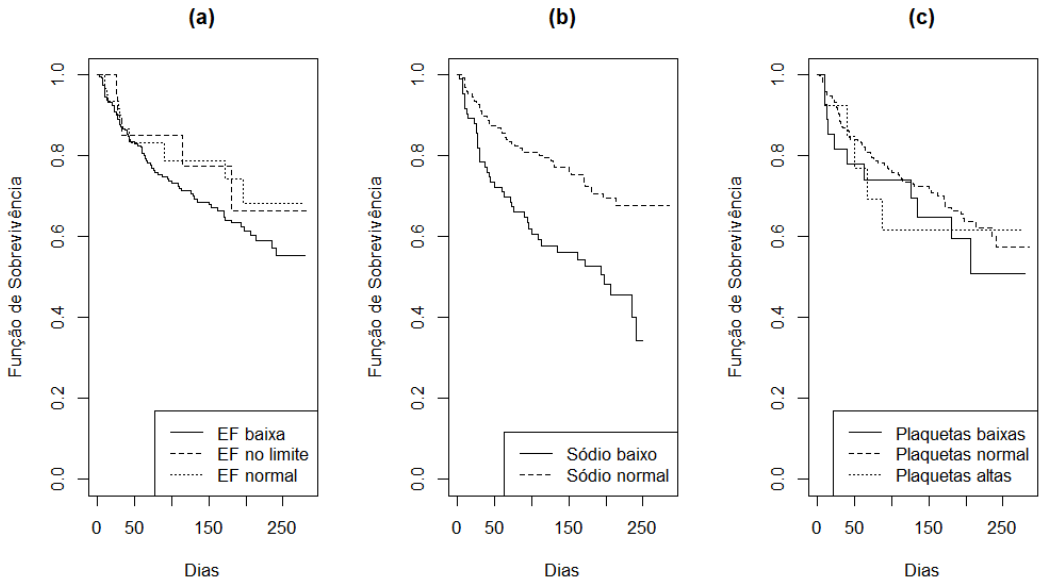

Figura 3.3 Estimativas de Kaplan-Meier por nível de (a) fracção de ejecção, (b) sódio e (c) plaquetas

No gráfico relativo à fracção de ejeccção, a curva de sobrevivência do grupo com as percentagens menores é inferior, comparativamente às restantes curvas. No teste log-rank, obteve-se um valor-p de 0.4, que aos níveis usuais de significância, leva à não rejeição da hipótese nula, não havendo, assim, evidência de que existam diferenças significativas entre as curvas de sobrevivência. O gráfico ao centro, referente ao sódio, mostra que a curva de sobrevivência dos indivíduos que têm o nível de sódio no intervalo de referência é superior. O valor-p obtido foi de 0.00008, havendo, assim, evidência de que existem diferenças significativas entre as curvas de sobrevivência. No gráfico (c), relativo às plaquetas, as curvas de sobrevivência mantêm-se próximas, cruzando-se por vezes. O valor-p 0.7 leva a crer que esta variável não influencie o tempo de vida.

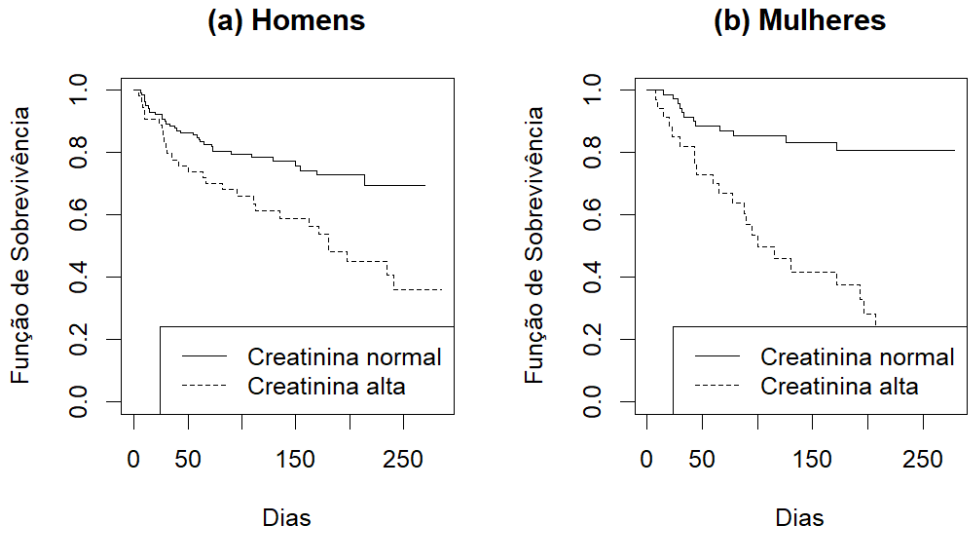

Figura 3.4 Estimativas de Kaplan-Meier por nível de creatinina e por género

Em ambos os géneros, a curva correspondente aos indivíduos com creatinina no valor normal é superior, comparativamente à dos indivíduos com creatinina elevada. No teste log-rank, o valor-p obtido, nos dois casos, foi próximo de zero, havendo evidência de que existam diferenças significativas entre as curvas de sobrevivência.

Por último, no gráfico da variável CPK, representado na Figura 3.5(a), no género masculino, no início, as curvas estão bastante próximas e, posteriormente, afastam-se, sendo a curva da CPK alta inferior. No género feminino, as curvas mantêm-se próximas. Obteve-se um valor-p de 0.3 para os homens e 0.5 para as mulheres, não havendo, em ambos os casos, evidência de existirem diferenças significativas.

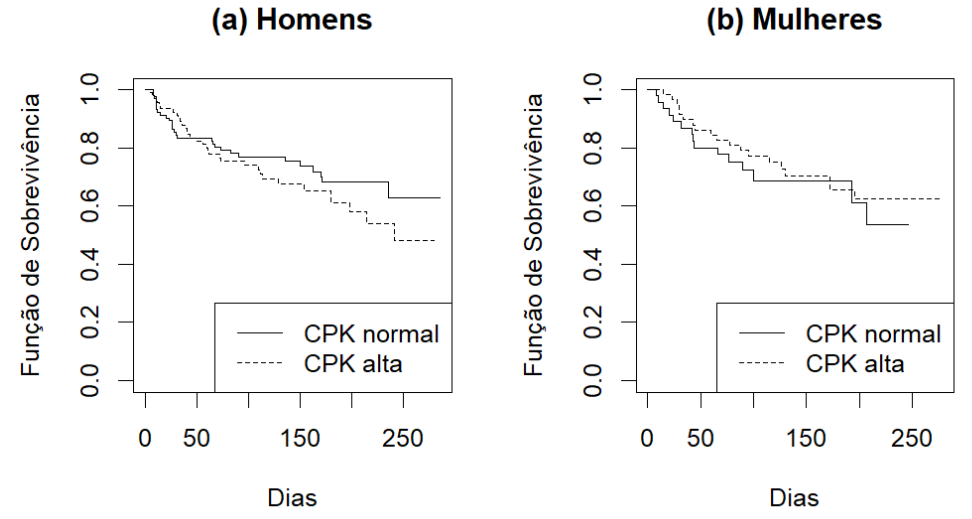

Figura 3.5 Estimativas de Kaplan-Meier por nível de CPK e por género

3.3 Modelo de Cox

Para identificar os factores que afectam o tempo de sobrevivência dos indivíduos, foi ajustado o modelo de regressão de Cox. No modelo inicial, incluíram-se todas as variáveis em estudo. O *output* obtido está representado na Figura 3.6.

```
> modelo_cox_total
Call:
coxph(formula = Surv(TIME, Event) ~ Gender + Smoking + Diabetes +
    BP + Anaemia + Age + Ejection.Fraction + Sodium + Creatinine +
    Pletelets + CPK, data = dados, ties = "breslow")

                       coef  exp(coef)   se(coef)      z        p
Gender1          -2.399e-01  7.867e-01  2.516e-01 -0.953   0.3404
Smoking1          1.186e-01  1.126e+00  2.513e-01  0.472   0.6370
Diabetes1         1.373e-01  1.147e+00  2.231e-01  0.615   0.5384
BP1               4.719e-01  1.603e+00  2.164e-01  2.181   0.0292
Anaemia1          4.553e-01  1.577e+00  2.169e-01  2.099   0.0358
Age               4.619e-02  1.047e+00  9.327e-03  4.952 7.36e-07
Ejection.Fraction -4.884e-02  9.523e-01  1.051e-02 -4.647 3.37e-06
Sodium           -4.441e-02  9.566e-01  2.324e-02 -1.911   0.0560
Creatinine        3.115e-01  1.366e+00  6.962e-02  4.475 7.64e-06
Pletelets        -5.160e-07  1.000e+00  1.128e-06 -0.458   0.6473
CPK               2.209e-04  1.000e+00  9.910e-05  2.229   0.0258

Likelihood ratio test=81.17  on 11 df, p=8.766e-13
n= 299, number of events= 96
```

Figura 3.6 Output *do modelo de Cox ajustado com todas as variáveis em estudo*

Na primeira coluna, tem-se as estimativas dos betas e, na segunda coluna, o risco relativo que se obtém calculando a exponencial de $\hat{\beta}$. Um risco relativo superior a um indica um aumento do risco de morte, enquanto um valor inferior a um é indicador de uma diminuição do risco de morte. Dividindo o valor estimado pelo *standard error* (se), obtém-se o valor observado da estatística de teste do teste de Wald, apresentado na quarta coluna. E, na última coluna, tem-se o valor-p do respectivo teste. As variáveis que se revelaram significativas, ao nível de significância de 10%, foram pressão arterial, anemia, idade, fracção de ejecção, sódio, creatinina e CPK.

Para o método de selecção de variáveis, retiraram-se, uma a uma, as variáveis que não demonstraram ser significativas. Compararam-se os modelos, recorrendo-se ao teste da razão de verosimilhanças, cujo valor-p é apresentado na última coluna. O modelo obtido, com todas as variáveis significativas, está representado na figura seguinte.

```
> summary(modelo_cox_final)
Call:
coxph(formula = Surv(TIME, Event) ~ BP + Anaemia + Age + Ejection.Fraction +
    Sodium + Creatinine + CPK, data = dados, ties = "breslow")

  n= 299, number of events= 96

                      coef  exp(coef)   se(coef)      z Pr(>|z|)
BP1                4.922e-01  1.636e+00  2.138e-01  2.302   0.0213 *
Anaemia1           4.418e-01  1.556e+00  2.151e-01  2.054   0.0400 *
Age                4.335e-02  1.044e+00  8.842e-03  4.903 9.44e-07 ***
Ejection.Fraction -4.732e-02  9.538e-01  1.031e-02 -4.589 4.46e-06 ***
Sodium            -4.600e-02  9.550e-01  2.331e-02 -1.973   0.0485 *
Creatinine         3.045e-01  1.356e+00  6.849e-02  4.445 8.77e-06 ***
CPK                2.097e-04  1.000e+00  9.811e-05  2.137   0.0326 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
BP1                  1.6359     0.6113    1.0758    2.4875
Anaemia1             1.5555     0.6429    1.0204    2.3713
Age                  1.0443     0.9576    1.0264    1.0626
Ejection.Fraction    0.9538     1.0485    0.9347    0.9733
Sodium               0.9550     1.0471    0.9124    0.9997
Creatinine           1.3559     0.7375    1.1856    1.5507
CPK                  1.0002     0.9998    1.0000    1.0004

Concordance= 0.738  (se = 0.027 )
Likelihood ratio test= 79.78  on 7 df,   p=2e-14
Wald test            = 87.51  on 7 df,   p=4e-16
Score (logrank) test = 86.96  on 7 df,   p=5e-16
```

Figura 3.7 Output *do modelo de Cox ajustado com todas as variáveis significativas*

As variáveis que não contribuíram significativamente para o modelo foram removidas, ficando apenas as variáveis pressão arterial, anemia, idade, fracção de ejecção, sódio, creatinina e CPK.

O passo seguinte foi testar se o pressuposto da proporcionalidade dos riscos é violado. Ao nível de significância de 10%, todas as variáveis verificam o pressuposto da proporcionalidade dos riscos, excepto a variável da fracção de ejecção que apenas verifica o pressuposto ao nível de significância de 1%.

```
> (teste_cox<-cox.zph(modelo_cox_final))
                     chisq df     p
BP                 0.00489  1 0.944
Anaemia            0.00235  1 0.961
Age                0.04753  1 0.827
Ejection.Fraction  4.74001  1 0.029
Sodium             0.07984  1 0.778
Creatinine         2.01569  1 0.156
CPK                0.92968  1 0.335
GLOBAL            10.79930  7 0.148
```

Figura 3.8 Teste da proporcionalidade dos riscos

Testada a hipótese da proporcionalidade dos riscos, realizou-se uma análise de resíduos. Para verificar se o modelo ajustado é adequado, representou-se graficamente os resíduos de Schoenfeld. Na generalidade, as nuvens de pontos centram-se em zero, salvo uns valores que se afastam, reforçando, assim, a adequação do modelo de Cox ajustado.

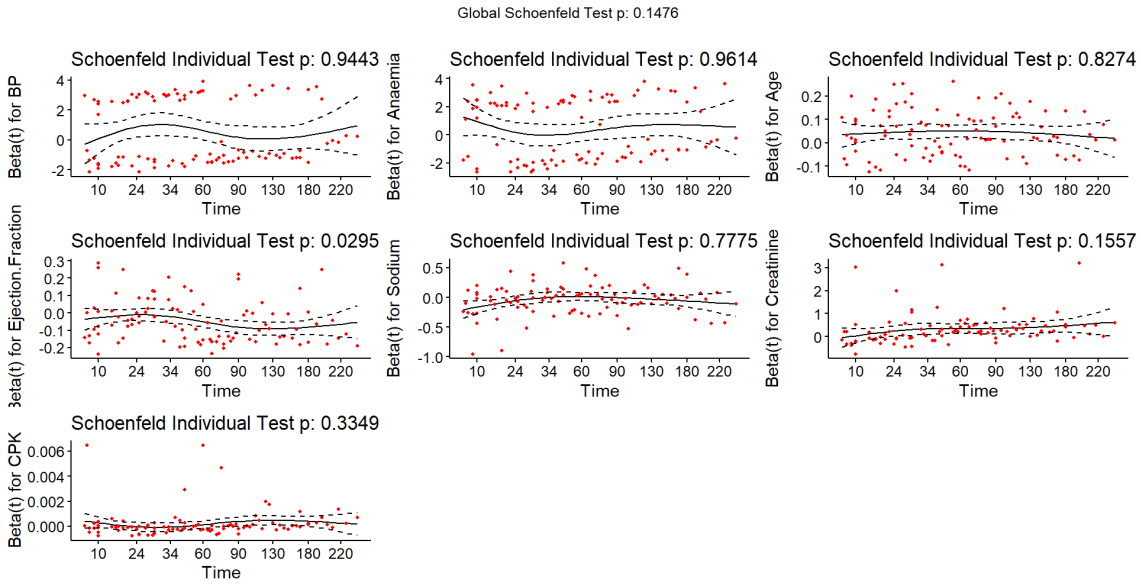

Figura 3.9 Resíduos de Schoenfeld do modelo de Cox ajustado

Para avaliar a adequabilidade da forma das variáveis no modelo, representou-se cada variável em função dos resíduos martingala. Verificou-se uma tendência linear, assegurando que as variáveis estão bem representadas no modelo. Utilizaram-se, ainda, os resíduos deviance, de modo a detectar a presença de *outliers*. Não se verificaram valores extremos que pusessem em causa o ajustamento do modelo.

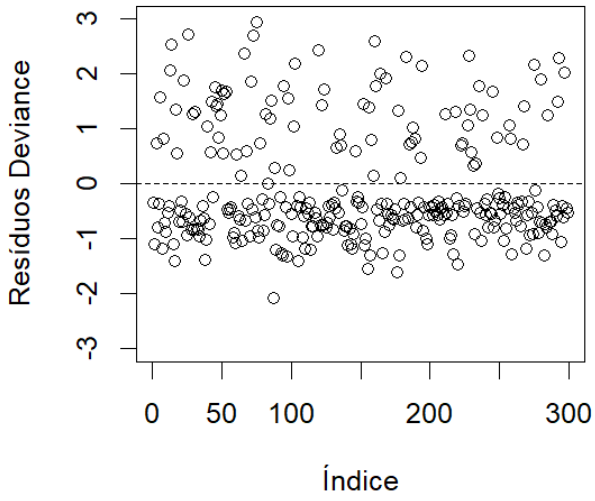

Figura 3.10 Resíduos deviance do modelo de Cox ajustado

3.4 Abordagem paramétrica

Com o intuito de averiguar os modelos paramétricos que se adequam aos dados em estudo, comparou-se graficamente a estimativa de Kaplan-Meier com a estimativa paramétrica pelo modelo considerado para a função de sobrevivência. As distribuições consideradas foram a exponencial, Weibull, log-normal e log-logística. Os gráficos obtidos são apresentados na Figura 3.11.

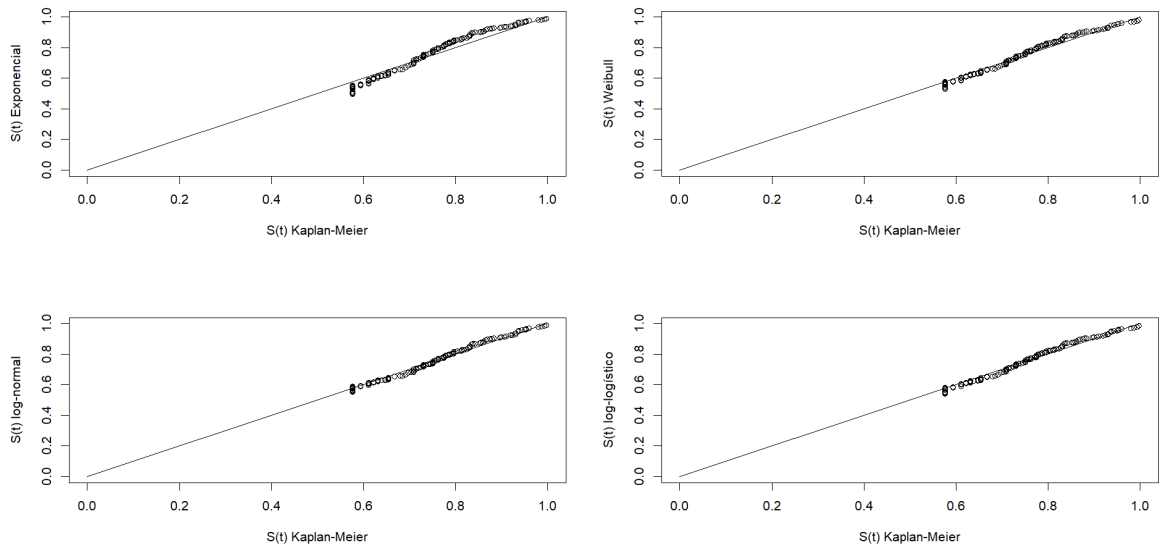

Figura 3.11 Adequação dos modelos paramétricos

Todos os modelos têm uma boa adequação aos dados, sendo o modelo exponencial o que mais se afasta. O modelo log-normal aparenta ser o que se adequa melhor, embora seja semelhante aos restantes modelos. Utilizou-se, também, o critério de informação de Akaike para comparar os modelos. O modelo log-normal foi o que apresentou o menor valor de AIC, sugerindo o melhor ajustamento aos dados. Por sua vez, o modelo exponencial foi o que apresentou o maior valor de AIC, corroborando, assim, a análise feita graficamente.

Encontrados os modelos que se adequam aos dados, o próximo passo será criar um modelo de regressão para avaliar o efeito das covariáveis no tempo de vida dos indivíduos. Para cada distribuição, para a selecção de variáveis, recorreu-se ao método *stepwise*, com base no valor de AIC, tendo sido considerados os três casos: *forward*, *backward* e *both*. Os *outputs* obtidos são apresentados nas figuras seguintes.

```
Call:
survreg(formula = Surv(TIME, Event) ~ Age + Ejection.Fraction +
    Creatinine + BP + Anaemia + CPK + Sodium, data = dados, dist = "exponential")
                     Value Std. Error     z       p
(Intercept)       1.90e+00   3.13e+00  0.61   0.542
Age              -4.57e-02   8.80e-03 -5.20 2.0e-07
Ejection.Fraction 4.95e-02   1.03e-02  4.78 1.8e-06
Creatinine       -3.18e-01   6.68e-02 -4.76 1.9e-06
BP1              -5.26e-01   2.11e-01 -2.49   0.013
Anaemia1         -4.79e-01   2.13e-01 -2.25   0.024
CPK              -2.26e-04   9.84e-05 -2.30   0.021
Sodium            4.54e-02   2.31e-02  1.96   0.050

Scale fixed at 1

Exponential distribution
Loglik(model)= -628.9   Loglik(intercept only)= -672.5
        Chisq= 87.29 on 7 degrees of freedom, p= 4.5e-16
Number of Newton-Raphson Iterations: 5
n= 299
```

Figura 3.12 Modelo ajustado utilizando a distribuição Exponencial

```
Call:
survreg(formula = Surv(TIME, Event) ~ Age + Ejection.Fraction +
    Creatinine + BP + Anaemia + CPK + Sodium, data = dados, dist = "weibull")
                    Value Std. Error     z       p
(Intercept)      1.795142   3.252849  0.55   0.581
Age             -0.046745   0.009462 -4.94 7.8e-07
Ejection.Fraction  0.050860   0.011278  4.51 6.5e-06
Creatinine      -0.325122   0.071385 -4.55 5.3e-06
BP1             -0.533074   0.219493 -2.43   0.015
Anaemia1        -0.486299   0.221263 -2.20   0.028
CPK             -0.000231   0.000102 -2.26   0.024
Sodium           0.046749   0.024193  1.93   0.053
Log(scale)       0.035998   0.088754  0.41   0.685

Scale= 1.04

Weibull distribution
Loglik(model)= -628.8   Loglik(intercept only)= -670.4
        Chisq= 83.25 on 7 degrees of freedom, p= 3e-15
Number of Newton-Raphson Iterations: 6
n= 299
```

Figura 3.13 Modelo ajustado utilizando a distribuição de Weibull

```
Call:
survreg(formula = Surv(TIME, Event) ~ Age + Ejection.Fraction +
    Creatinine + Sodium + BP + CPK + Anaemia, data = dados, dist = "lognormal")
                    Value Std. Error     z       p
(Intercept)      -0.336105   3.634636 -0.09 0.92632
Age              -0.046723   0.010409 -4.49 7.2e-06
Ejection.Fraction  0.043154   0.011233  3.84 0.00012
Creatinine       -0.357001   0.103687 -3.44 0.00058
Sodium            0.062446   0.026753  2.33 0.01958
BP1              -0.510678   0.251867 -2.03 0.04260
CPK              -0.000244   0.000115 -2.11 0.03449
Anaemia1         -0.527003   0.251187 -2.10 0.03590
Log(scale)        0.489546   0.080044  6.12 9.6e-10

Scale= 1.63

Log Normal distribution
Loglik(model)= -631.1   Loglik(intercept only)= -666.3
        Chisq= 70.44 on 7 degrees of freedom, p= 1.2e-12
Number of Newton-Raphson Iterations: 4
n= 299
```

Figura 3.14 Modelo ajustado utilizando a distribuição log-normal

```
Call:
survreg(formula = Surv(TIME, Event) ~ Age + Ejection.Fraction +
    Creatinine + BP + Sodium + Anaemia + CPK, data = dados, dist = "loglogistic")
                    Value Std. Error     z       p
(Intercept)        0.663936   3.692788  0.18 0.85732
Age               -0.048405   0.010139 -4.77 1.8e-06
Ejection.Fraction  0.049385   0.011972  4.13 3.7e-05
Creatinine        -0.354196   0.098154 -3.61 0.00031
BP1               -0.544761   0.242267 -2.25 0.02454
Sodium             0.053844   0.027211  1.98 0.04784
Anaemia1          -0.490213   0.245452 -2.00 0.04581
CPK               -0.000224   0.000115 -1.95 0.05078
Log(scale)        -0.111833   0.087473 -1.28 0.20108

Scale= 0.894

Log logistic distribution
Loglik(model)= -630.4   Loglik(intercept only)= -669.2
        Chisq= 77.51 on 7 degrees of freedom, p= 4.4e-14
Number of Newton-Raphson Iterations: 4
n= 299
```

Figura 3.15 Modelo ajustado utilizando a distribuição log-logística

Como se pode constatar, nos quatro modelos, o conjunto de covariáveis que se revelaram significativas foi semelhante. Idade, fracção de ejecção, creatinina, pressão arterial, sódio, anemia e CPK foram as covariáveis que mostraram ter um efeito significativo no tempo de vida dos pacientes. As variáveis idade, creatinina, pressão arterial, anemia e CPK apresentam um sinal negativo, indicando que o aumento de uma unidade destas variáveis implica uma diminuição do tempo de sobrevivência. Tome-se como exemplo o modelo exponencial, o aumento de uma unidade, na variável idade, mantendo as restantes constantes, leva a uma diminuição do tempo de vida, calculado por exp(0.0457)=1.0468, ou seja, cerca de 4.68%. Por outro lado, um aumento de uma unidade, na variável da fracção de ejecção, mantendo as restantes constantes, leva ao aumento do tempo de vida, calculado por exp(0.0495)=1.0507, ou seja, cerca de 5.07%.

A comparação dos valores de AIC indicou que o modelo exponencial foi o que apresentou um melhor ajustamento, seguido do modelo de Weibull, enquanto o modelo log-normal foi o que apresentou um valor de AIC superior.

Por fim, para averiguar a adequabilidade dos modelos ajustados, realizou-se uma análise de resíduos. Para avaliar o ajustamento global do modelo, utilizaram-se os resíduos de Cox-Snell, onde foi feita a representação gráfica destes resíduos *versus* a função de risco cumulativa dos mesmos. Se estes estiverem sobre a recta da bissetriz dos quadrantes ímpares, conclui-se que o modelo é adequado.

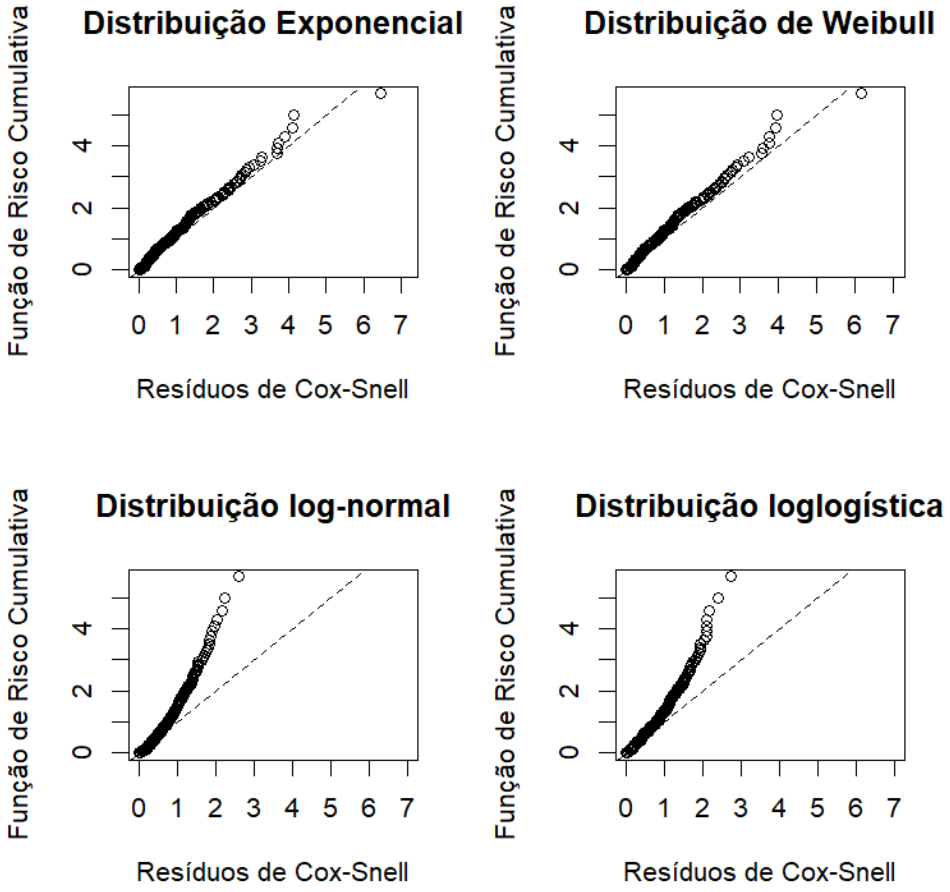

Figura 3.16 Resíduos de Cox-Snell dos quatro modelos ajustados

Nos gráficos dos modelos exponencial e Weibull apresentados na Figura 3.16, os pontos situam-se próximos da recta de referência. Já nos modelos log-normal e log-logístico, os pontos afastam-se da recta da bissetriz dos quadrantes ímpares, indicando um pior ajustamento.

Para avaliar a forma de regressão de cada uma das covariáveis, foram calculados os resíduos martingala. Se, na representação gráfica, estes resíduos formarem uma nuvem de pontos com uma tendência linear, significa que a covariável está bem representada no modelo. Na generalidade, nos gráficos obtidos, a nuvem de pontos revelou uma tendência linear, ainda que com alguns valores mais afastados.

Por último, para detectar a existência de *outliers*, foram utilizados os resíduos deviance. Se, na representação gráfica, os pontos não tiverem em torno de zero, significa que se está na presença de *outliers*. Nos quatros modelos, como se pode observar nos gráficos da Figura 3.17, a maior parte dos resíduos deviance situam-se entre -3 e 3, havendo apenas uns pontos dispersos.

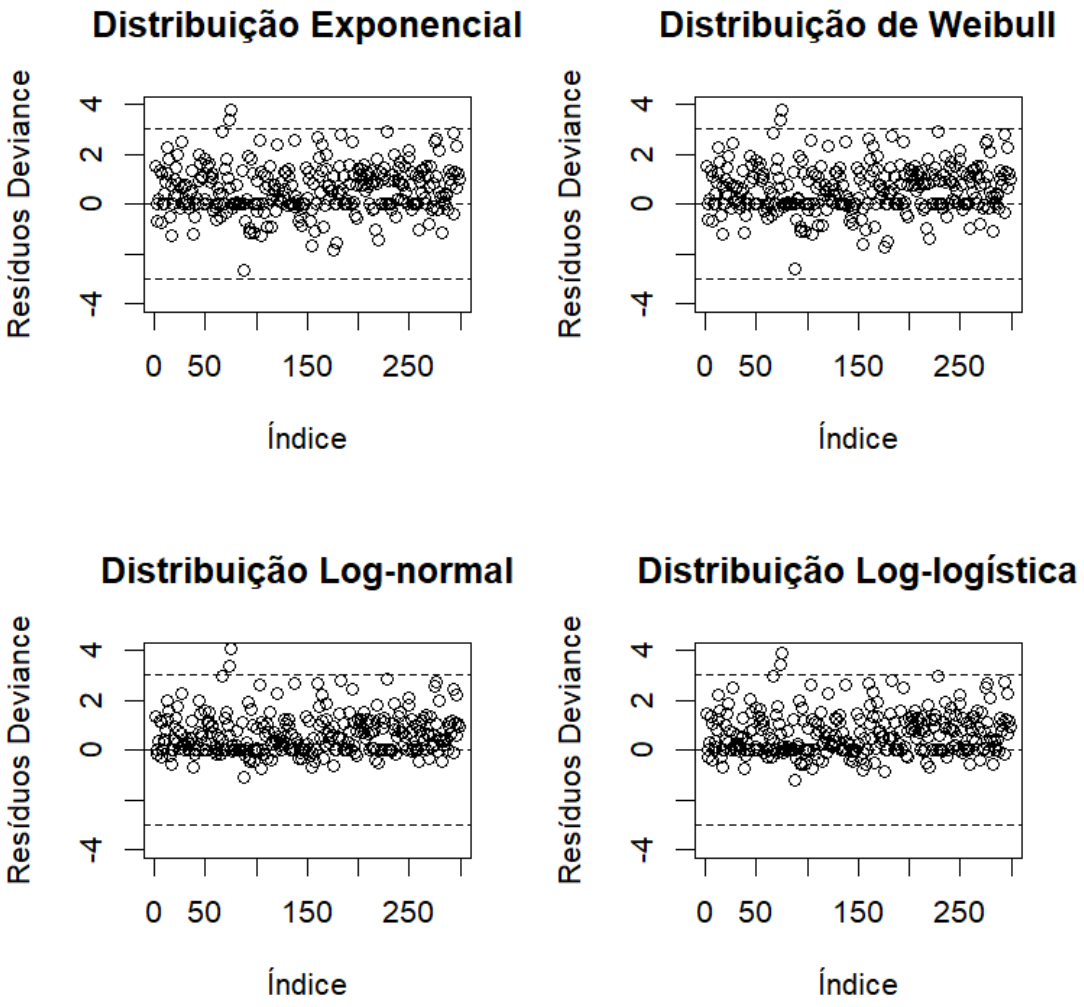

Figura 3.17 Resíduos deviance dos quatros modelos ajustados

4 Discussão

No presente estudo, pretendeu-se identificar os factores que influenciavam o tempo de vida de pacientes com insuficiência cardíaca, aplicando modelos de regressão paramétricos.

Na análise não paramétrica realizada, os gráficos da estimativa de Kaplan-Meier e os resultados obtidos no teste log-rank evidenciaram diferenças significativas entre os grupos comparados nas variáveis pressão arterial, sódio e creatinina. Na variável que indica se o indivíduo tem anemia, apesar de, na representação gráfica, as curvas de sobrevivência se afastarem, o teste de log-rank não corrobora este resultado ao nível de significância de 5%.

Do ajustamento do modelo semi-paramétrico de Cox, onde foram consideradas todas as variáveis em estudo, as que se revelaram significativas foram pressão arterial, anemia, idade, fracção de ejecção, sódio, creatinina e CPK. Os resultados que podem surpreender remetem para o facto de as variáveis fumar e ter diabetes não apresentarem um risco associado à insuficiência cardíaca. A verificação do pressuposto da proporcionalidade dos riscos e a análise de resíduos realizada indicaram um bom ajustamento do modelo aos dados.

Os resultados obtidos no modelo de Cox mostraram que há um maior risco de morte nos indivíduos que têm pressão arterial elevada, relativamente aos que não têm, assim como, nos indivíduos anémicos, há um risco superior, relativamente aos indivíduos não anémicos. O aumento de uma unidade na idade ou na creatinina também leva a um aumento do risco de morte. Por outro lado, o aumento da quantidade da percentagem de fracção de ejecção ou da quantidade de sódio, presente no sangue, reduz o risco de morte. Apesar de CPK se ter revelado significativa, o risco relativo é igual a um, pelo que não existe um aumento ou uma diminuição do risco de morte, caso o valor de CPK se altere.

No que respeita aos modelos paramétricos, foram considerados os modelos exponencial, Weibull, log-normal e log-logístico. No ajustamento sem a inclusão de covariáveis e com base no critério de informação de Aikake, o modelo log-normal foi o que se revelou mais indicado, com um menor valor de AIC. O modelo exponencial foi o que apresentou um pior desempenho. No ajustamento com covariáveis, a comparação dos resultados, bem como a análise de resíduos, permitiu concluir que os modelos exponencial e de Weibull apresentaram um ajustamento global mais adequado. Dada a sua flexibilidade, não é de admirar que o modelo de Weibull se tenha verificado adequado aos dados. Tanto o modelo log-normal quanto o modelo log-logístico, na análise gráfica dos resíduos de Cox-Snell, se afastaram da recta de referência, revelando um mau ajustamento. Da comparação dos valores de AIC, o modelo log-normal foi o que apresentou o maior valor, contrastando com os resultados obtidos no ajustamento sem covariáveis. A análise dos resíduos martingala confirmou a adequação das variáveis no modelo, embora se tenham observado alguns pontos mais afastados. Também nos resíduos deviance, verificaram-se uns valores mais afastados, representando possíveis *outliers*.

Os resultados dos modelos ajustados mostraram que as variáveis idade, fracção de ejecção, creatinina, pressão arterial, anemia, CPK e sódio são factores que contribuem para a alteração do tempo de sobrevivência. O aumento da idade, valores elevados de creatinina e de CPK estão associados a um aumento do risco, assim como ter pressão arterial elevada e anemia. Valores mais baixos de fracção de ejecção e de sódio também constituem um factor de risco acrescido - resultados que vão ao encontro da literatura, que identifica estes factores como relevantes em doentes com insuficiência cardíaca. As variáveis que indicam se o indivíduo fuma e se o indivíduo tem diabetes não se revelaram significativas, indo de encontro ao que diz a literatura, que diz que estes factores estão relacionados com doenças cardíacas.

Através do critério de informação de Akaike, foram comparados os modelos paramétricos ajustados com o modelo de Cox, tendo sido este a apresentar o menor valor de AIC, evidenciando um melhor ajustamento aos dados.

No artigo que serviu de base para este estudo, os factores identificados como contribuidores para o risco de mortalidade devido à insuficiência cardíaca foram idade, creatinina, pressão arterial, anemia, fracção de ejecção e sódio, recorrendo ao ajustamento do modelo de Cox, o que difere apenas numa variável, comparativamente ao presente trabalho, onde também a CPK foi identificada como um factor de risco. Apesar das discrepâncias notadas nas variáveis categóricas indicativas de pressão arterial e anemia, tanto no artigo como neste estudo, estas revelaram-se significativas no tempo de vida dos indivíduos com insuficiência cardíaca.

5 Conclusão

Em suma, as variáveis idade, fracção de ejecção, creatinina, pressão arterial, anemia, CPK e sódio revelaram ter um efeito significativo no tempo de vida, enquanto as variáveis género, fumador, diabetes e plaquetas não o demonstraram. O aumento da idade, um valor elevado de creatinina, ter pressão arterial elevada, ter anemia e um elevado valor de CPK contribuem para o aumento do risco de mortalidade entre pacientes com insuficiência cardíaca. Por outro lado, o aumento da fracção de ejecção e da quantidade de sódio no sangue podem reduzir a taxa de mortalidade.

O aumento da idade é inevitável, mas a monitorização de parâmetros como a creatinina, fracção de ejecção e sódio, que demonstraram estar associados ao risco de morte, podem levar ao diagnóstico precoce de alterações cardíacas, permitindo uma intervenção mais eficaz.

Em estudos futuros, poder-se-á explorar se existe interacção entre as covariáveis e fazer-se uma análise estratificada nas covariáveis, cujos valores de referência variam de acordo com o género.

6 Referências bibliográficas

Ahamad T., Munir A., Bhatti S., Aftab M., Raza M. (2017). *Survival analysis of heart failure patients: A case study*.

American Heart Association. (2025). *Ejection Fraction Heart Failure Measurement*. [online]. [Acedido a 2 de Maio de 2025]. Disponível em: https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement

Bokhari, S. (2024). *Cardiovascular Health in Pakistan: A Growing Concern*. The Aga Khan University Hospitl. [online]. [Acedido a 22 de Julho de 2025]. Disponível em: https://hospitals.aku.edu/pakistan/AboutUs/News/Pages/cardiovascular-health.aspx

Borges, A. (2014). *Análise de Sobrevivência com o R*. Dissertação de Mestrado, Universidade da Madeira.

Cleveland Clinic. (2022). *Ejection Fraction*. [online]. [Acedido a 2 de Maio de 2025]. Disponível em: https://my.clevelandclinic.org/health/articles/16950-ejection-fraction

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Fourth Edition. Chapman & Hall/CRC.

CUF. *Anemia*. [online]. [Acedido a 2 de Maio de 2025]. Disponível em: https://www.cuf.pt/saude-a-z/anemia

CUF. *Diabetes*. [online]. [Acedido a 19 de Setembro de 2025]. Disponível em: https://www.cuf.pt/saude-a-z/diabetes

CUF. *Tabagismo*. [online]. [Acedido a 19 de Setembro de 2025]. Disponível em: https://www.cuf.pt/saude-a-z/tabagismo

Hayes, A. (2025). *Stepwise Regression Explained: Uses, Benefits, and Drawbacks*. [online]. [Acedido a 12 de Setembro de 2025]. Disponível em: https://www.investopedia.com/terms/s/stepwise-regression.asp

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

Samad, Z., Hanif, B. (2023). *Cardiovascular Diseases in Pakistan: Imagining a Postpandemic, Postconflict Future*. AHA|ASA Journals. [online]. [Acedido a 22 de Julho de 2025]. Disponível em: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.122.059122

Serranho, P. (2015). *Notas de Análise de Sobrevivência.* [PDF]. Universidade Aberta. [Acedido a 14 de Novembro de 2024]. Disponível em: https://repositorioaberto.uab.pt/bitstreams/1179869b-f452-48c0-bebf-3903b452ecf5/download

SNS 24. (2024). *Insuficiência cardíaca*. [online]. [Acedido a 24 de Novembro de 2024]. Disponível em: https://www.sns24.gov.pt/tema/doencas-do-coracao/insuficiencia-cardiaca/

Therneau, T. (2024). *Package "suvival"*. [PDF]. [Acedido a 18 de Setembro de 2025]. Disponível em: https://cran.r-project.org/web/packages/survival/survival.pdf

Tua Saúde. (2024). *Anemia: o que é, sintomas, tipos, causas e tratamento*. [online]. [Acedido a 2 de Maio de 2025]. Disponível em: https://www.tuasaude.com/anemia/

Tua Saúde. (2024). *CPK (creatinofosfoquinase): o que significa e porque está alto ou baixo*. [online]. [Acedido a 4 de Setembro de 2025]. Disponível em: https://www.tuasaude.com/exame-cpk/

Tua Saúde. (2025). *Creatinina: o que é, quando fazer o exame (e valores normais)*. [online]. [Acedido a 4 de Setembro de 2025]. Disponível em: https://www.tuasaude.com/creatinina/

Tua Saúde. (2023). *Hematócrito (Hct): o que é e porque está alto ou baixo*. [online]. [Acedido a 2 de Maio de 2025]. Disponível em: https://www.tuasaude.com/hematocrito-hct/

Tua Saúde. (2024). *Plaquetas: o que são, funções e valores de referência*. [online]. [Acedido a 2 de Maio de 2025]. Disponível em: https://www.tuasaude.com/funcao-das-plaquetas/

Tua Saúde. (2025). *Sódio: o que é, funções (e porque pode estar baixo ou alto)*. [online]. [Acedido a 4 de Setembro de 2025]. Disponível em: https://www.tuasaude.com/sodio-na/

World Heart Federation. (2021). *Heart Failure*. [online]. [Acedido a 2 de Janeiro de 2025]. Disponível em: https://world-heart-federation.org/what-we-do/heart-failure/