Universidade do Minho
Escola de Ciências

Leandro dos Santos Xavier Duarte

Aplicação de técnicas de análise de sobrevivência numa coorte de COVID-19

dezembro de 2023

Aplicação de técnicas de análise de sobrevivência numa coorte de COVID-19
Leandro Duarte
UMinho | 2023

Universidade do Minho
Escola de Ciências

Leandro dos Santos Xavier Duarte

Aplicação de técnicas de análise de sobrevivência numa coorte de COVID-19.

Dissertação de Mestrado
Mestrado em
Estatística para
Ciência de Dados

Trabalho efetuado sob a orientação do
Prof. Doutor Luís Filipe Meira Machado e da
Doutora Carla Maria Gonçalves de Macedo Moreira

dezembro de 2023

Agradecimentos

Gostaria de agradecer à professora doutora Carla Moreira por todo o apoio, orientação e por permitir que eu fizesse parte do projeto SUMcohort. Este projeto foi extremamente importante para mim porque me permitiu trabalhar com uma equipe maravilhosa e extremamente qualificada, abordou um tema de meu interesse e ainda permitiu que eu pudesse me dedicar exclusivamente ao estudo através da bolsa de projeto EXPL/MAT-STA/0956/2021 patrocinada pela FCT, a qual também agradeço.

Serei eternamente grato ao professor doutor Luís Meira Machado pela calma, paciência e por ser sempre solícito e muito didático. À professora doutora Ana Paula Amorim por ser sempre muito atenciosa e à mestra Inês Gonçalves pelas incansáveis horas de trabalho presencial na UMinho.

Agradeço à Paula Meireles e à Joana Pinto Costa pelo suporte no entendimento da base de dados, nos conceitos e particularidades do COVID-19 e pela contribuição direta na elaboração do artigo.

Agradeço à professora doutora Arminda Manuela Gonçalves pela incomensurável dedicação nas aulas e nas tarefas de diretora do curso e à professora doutora Susana Margarida Ferreira Sá Faria por possuir uma didática que supera quaisquer expectativas.

Aos futuros doutores Jhonathan Barrios e Vitor Mattos pela amizade e por servirem de inspiração nesta jornada acadêmica. Ao mestre (no xadrez) Elias por ter sempre me incentivado a seguir com os estudos.

Agradeço aos meus pais, familiares e amigos que sempre me apoiaram nesta jornada. São tantas as pessoas que me ajudaram que, se eu fosse nomeá-las, a lista ficaria demasiadamente extensa, felizmente.

Por fim, dedico este trabalho ao meu tio Roberto e a todos que faleceram em decorrência do COVID-19 e do descaso do governo da época.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledge the Code of Ethical Conduct of the University of Minho.

Universidade do Minho, 27 de dezembro de 2023,

Leandro dos S. X. Duarte

(Leandro dos Santos Xavier Duarte)

Resumo

O COVID-19 é causado pelo vírus SARS-COV-2, que é parte da família coronavírus e que pode causar desde doenças mais leves até as mais graves. Além disso, ele possui grande capacidade de espalhar-se rapidamente e de apresentar muitas mutações genéticas, fazendo com que surgissem (e ainda possam surgir) diversas variantes [1].

Outro facto que devemos nos atentar é que mesmo após a cura do vírus, ainda há a possibilidade de sofrermos pelas sequelas causadas no processo de recuperação, as quais podem perdurar por semanas ou até serem irreversíveis, o que é chamado de *Long-COVID* (ou COVID prolongada, em português), *Post-Covid*, *Post-Acute Sequelae* entre outros termos [1].

Os estudos publicados sobre o tema compreendem questões acerca da mortalidade ou do tempo de internamentos, pouco se sabendo sobre o tempo de resolução de todos os sintomas e quais são os fatores que influenciam a sua resolução, os quais são importantes para prevermos como será a progressão da doença num indivíduo ao longo do tempo, determinando os possíveis fatores de risco (ou os protetores) até a completa cura para que as instituições de saúde possam agir de forma rápida e precisa.

No intuito de alcançarmos esse objetivo, analisamos a base de dados fornecida pelo laboratório do Centro Hospitalar Universitário de São João (CHUSJ), que possui informações dos pacientes diagnosticados com COVID-19 e que responderam ao questionário sobre quais sintomas tiveram e por quantos dias os sofreram. À época do estudo, ainda não havia disponibilidade de vacina para o COVID-19, razão pela qual isto não pôde ser contemplado no estudo.

Considerando que estamos interessados no tempo até a resolução dos sintomas e que há observações censuradas, as técnicas mais adequadas são as de análise de sobrevivência, pois ignorar a presença de censura pode fazer com que os riscos sejam subestimados (ou superestimados, dependendo do tipo de censura existente). Mais precisamente, aplicamos técnicas de modelos de cura de mistura (MCM) porque a função de sobrevivência estava em um platô, apresentando mais

de um terço dos participantes sem a resolução de todos os sintomas durante o período observado.

Como conclusão, vimos que os fatores em estudo, idade, sexo ao nascimento, presença de comorbidades, perceção de renda, uso regular de medicamentos e hospitalização foram significativamente associadas ao tempo de resolução de todos os sintomas, apesar de alguns não entrarem nos modelos finais devido ao critério de parcimónia.

Para o modelo de regressão de Cox múltiplo, os homens possuíram, em média, aproximadamente 62% mais risco de terem o evento de interesse (resolução de todos os sintomas) quando comparado às mulheres (HR: 1,619; 95% do IC: [1,469 , 1,785]); participantes com comorbidades apresentaram, em média, aproximadamente 12% a menos de risco de terem o evento de interesse quando comparados com os que não possuíam comorbidades (HR: 0,879; 95% do IC: [0,797 , 0,970]); à medida que piora a perceção da renda menor é o risco de ter o evento de interesse; indivíduos que foram hospitalizados durante a fase aguda da COVID-19 apresentaram, em média, aproximadamente 24% a menos de risco quando comparados aos que não foram hospitalizados (HR: 0,761; 95% CI: [0,660 , 0,878]). Estas interpretações também foram obtidas pelo modelo de regressão paramétrica com a distribuição de Weibull, o qual também pode ser considerado com o pressuposto de riscos proporcionais. Também foram apresentados modelos paramétricos sob as distribuições lognormal e loglogística, os quais contiveram os mesmos fatores dos modelos anteriores.

Para o Modelo de Cura de Mistura (MCM), a variável sexo ao nascimento deve estar presente na incidência (efeito de longo prazo) e na latência (efeito de curto prazo), enquanto comorbidades deve estar presente na incidência e hospitalização na latência. Utilizando este modelo, a estimativa para a proporção de indivíduos que não obteriam a resolução de todos os sintomas foi de 41,8% para mulheres com comorbidades, 35,2% para mulheres sem comorbidades, 29,0% para homens com comorbidades e de 23,7% para homens sem comorbidades, enquanto o fator hospitalização não interfere na estimativa de curados.

Palavras-chaves: Análise de Sobrevivência; Modelos de Cura de Mistura; COVID-19; Estimador de Kaplan-Meier; Modelos de Regressão de Cox

Abstract

COVID-19 is caused by the SARS-COV-2 virus, which is part of the coronavirus family and can cause anything from mildest to most severe illnesses. In addition, it has a great capacity to spread rapidly and to present many genetic mutations, causing several variants to emerge (and still may arise) [1].

Another fact that we must pay attention to is that even after the cure of the virus, there is still the possibility of suffering from the sequelae caused in the recovery process, which can last for weeks or until they are irreversible, which is called Long-COVID, Post-COVID, Post-Acute Sequelae among other terms [1].

The studies published on the subject include questions about mortality or hospitalisations, little is known about the time to resolution of all symptoms and what factors influence it, so that we can predict how the progression of the disease will be in an individual over time, determining the possible improvements (or complications) until the complete cure so that health institutions can act quickly and accurately.

To achieve this goal, we analysed the database provided by the central laboratory of the Centro Hospitalar Universitário de São João (CHUSJ), which has information on patients diagnosed with SARS-CoV-2 infection and who answered the questionnaire about what symptoms they had and for how many days they suffered them. At the end of the study, there was still no vaccine available for COVID-19, which is why it could not be included in the study.

Considering that we are interested in the time to resolution of all symptoms and that there are censored observations, the most appropriate techniques are those of survival analysis, because ignoring the presence of censoring can underestimate (or overestimate, depending on the type of censoring) the risks involved. More precisely, we applied techniques from mixture cure models (MCM) because the survival curve was in a plateau, with more than a third of the participants not having resolution of all symptoms during the observed period.

In conclusion, we saw that the factors age, sex at birth, presence of comorbidities, income perception, regular use of medications and hospitalisation are significantly associated with the time of resolution of all symptoms, although some variables were not included in the final models due to the parsimony criterion.

For the multiple Cox regression model, men had, on average, approximately 62% more risk of having the event of interest (resolution of all symptoms) when compared to women (HR: 1.619; 95% of CI: [1.469, 1.785]); Participants with comorbidities had, on average, approximately 12% less risk of having the event of interest when compared with those who did not have comorbidities (HR: 0.879; 95% of CI: [0.797, 0.970]); as the perception of income worsens, the lower the risk of having the event of interest becomes; individuals who were hospitalised during the acute phase of COVID-19 had, on average, approximately 24% lower risk when compared to those who were not hospitalized (HR: 0.761; 95% CI: [0.660, 0.878]). These interpretations were also obtained by the parametric regression model with the Weibull distribution, which can also be considered with the assumption of proportional risks. Parametric models were also presented under the lognormal and loglogistic distributions, which contained the same factors as the previous models.

For the MCM, the variable sex at birth was present in incidence (long-term effect) and in latency (short-term effect), while comorbidities were present in incidence and hospitalisation in latency. Using this model, the estimate for the proportion of individuals who would not achieve resolution of all symptoms was 41.8% for women with comorbidities, 35.2% for women without comorbidities, 29.0% for men with comorbidities and 23.7% for men without comorbidities, while the hospitalisation factor does not interfere in the estimation of long-term cured patients.

Keywords: Survival Analysis; Mixture Cure Models; COVID-19; Kaplan-Meier estimator; Cox regression

"A disciplina é a mãe do êxito"

Ésquilo

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de acrónimos

AFT	*Accelerated Failure Time*
CHUSJ	Centro Hospitalar Universitário de São João
HR	*Hazard Ratio*
KM	Kaplan-Meier
MCM	Modelo de Cura de Mistura
MV	Máxima Verosimilhança
OR	*Odds Ratio*
PH	*Proportional Hazard*

Capítulo 1

Introdução

1.1 COVID-19 e COVID prolongada

O COVID-19 é causado pela infeção do vírus pertencente à família SARS-CoV-2 e, como outras doenças respiratórias, pode causar tosse (seca ou com expetoração), cefaleia, congestão nasal, fadiga, inflamações na garganta, perda ou alteração no olfato e paladar, dispneia, nódulos pulmonares, linfopenia e doenças renais, podendo causar hospitalização ou até mesmo a morte [2]. A gravidade da infeção pode ser designada como leve, moderada, severa ou crítica, sendo que 81% dos casos observados na China foram classificados como leve ou moderado, 14% como severo e 5% como crítico [3].

A duração dos sintomas pode variar de duas semanas (maioria dos casos) até mesmo meses, o que neste caso pode designar a condição chamada COVID prolongada ou pós COVID [1]. Segundo a Organização Mundial da Saúde (OMS), a condição é designada por COVID prolongada quando os sintomas permanecem por 3 meses após a infeção por SARS-CoV-2 [1] [4].

Os sintomas persistentes da COVID prolongada mais comuns foram fadiga ou fraqueza muscular, ansiedade ou depressão [5,6], comprometimento da difusão pulmonar [7], anosmia, disgeusia, comprometimento da concentração e memória, dispneia, amigdalite estreptocócica, tontura [8] e neurocognitivos [9].

[1] os sintomas podem ser intermitentes, tendo uma duração mínima de 2 meses

1.2 Motivação e objetivo

Os diferentes estudos acerca da resolução de todos os sintomas consideraram diversos intervalos temporais porque, há pouco tempo, ainda estava em debate a definição de COVID prolongada. Por exemplo, em [10], o tempo observado foi de 7 meses; em [8, 11], consideraram 12 meses; em [12], observaram aos 6, 12 e 18 meses.

Outro ponto controverso é em relação ao percentual de indivíduos que não experimentaram a resolução de todos os sintomas, pois, em [10], 93,2% dos participantes alegaram ainda possuir sintomas; em [11], 80% dos respondentes ainda os possuíram; em [12], os percentuais de indivíduos sem resolução de todos os sintomas foram de 18,4%, 10,1% e 7,8% de resolução, nos períodos de 6, 12 e 18 meses, respetivamente; em [8], a conclusão a que chegaram é que os indivíduos com gravidade leve retornaram à normalidade no primeiro ano.

Além da disparidade encontrada nos percentuais de resolução de todos os sintomas, há grande variedade de técnicas estatísticas, com indivíduos proveninentes de diferentes regiões. Por exemplo, análise de sobrevivência clássica na China [13], modelos de cura na Índia [14] e Espanha [15], estatística espacial no Brasil [16], amostras pareadas nos EUA [17], e riscos competitivos em Portugal [18]. Ademais, também há estudos que acompanharam a resolução de um sintoma em específico como, por exemplo, diabetes em [19], perda de paladar e olfato em [20], saúde mental em [21–23] e alopécia em [24]. Considerando que cada método tem as suas vantagens e desvantagens, em [25] há a referida discussão.

No que diz respeito aos fatores que dificultam a recuperação dos sintomas, idade avançada, sexo feminino, histórico de câncer, histórico de consumo de tabaco, alto índice de massa corporal (30 ou mais) e possuir mais de 4 sintomas foram apontados por [12]. Diversos estudos indicaram que o sexo feminino pode ser considerado um fator de risco para o processo de recuperação [20, 22, 26–28] o que é surpreendente, pois a taxa de mortalidade é maior entre os homens [18]. O sexo e género foram estudados especificamente em [29, 30].

Com base nessas considerações, o propósito deste estudo é investigar o padrão de evolução do tempo necessário para a resolução completa dos sintomas resultantes da infeção pelo SARS-CoV-2, identificando os fatores que exercem influência nesse processo.

1.3 Estrutura da dissertação

O Capítulo 1 tem o objetivo de explicar o que é a COVID prolongada e trazer as motivações que culminaram na elaboração deste trabalho. No Capítulo 2, apresentamos as técnicas de análise de sobrevivência que foram utilizadas no tratamento da base fornecida pelo laboratório do Centro Hospitalar Universitário de São João (CHUSJ). No Capítulo 3, abordamos as técnicas de modelos de cura de mistura, dada a necessidade de utilizarmos esses conceitos no Capítulo 4.

No Capítulo 4, apresentamos, resumidamente, diveros estudos realizados ao redor do mundo sobre o tempo de resolução de todos os sintomas causados pela infeção de SARS-CoV-2. No Capítulo 5, comparamos os nossos resultados com os obtidos nos estudos apresentados. Por fim, no Apêndice, apresentamos todos os comandos utilizados em R para a confeção deste trabalho, desde a manipulação da base de dados até a construção dos modelos utilizados, explicando-os detalhadamente.

Portanto, este trabalho pode ser considerado tanto como um guia para a utilização das técnicas estatísticas de análise de sobrevivência e de modelos de cura quanto um estudo focado em explanar sobre a existência e gravidade da COVID prolongada, cujo sintomas podem ser perenes.

Capítulo 2

Principais conceitos de análise de sobrevivência

Este capítulo tem como objetivo apresentar os conceitos essenciais de análise de sobrevivência para proporcionar melhor entendimento das técnicas utilizadas no Capítulo 4, levando em consideração as particularidades do cenário lá verificado. As referências utilizadas para a sua elaboração foram [31–35].

As funções apresentadas são válidas quando a variável aleatória T for contínua e não negativa, pois é dessa forma que, frequentemente, são aplicadas, mesmo quando, na teoria, a variável resposta deveria ser considerada discreta.

2.1 Função de sobrevivência

A função de sobrevivência, denotada por $S(t)$, representa a probabilidade de ocorrência do evento ser posterior ao tempo t:

$$S(t) = P(T > t), \;\; t \geq 0 \tag{2.1}$$

De realçar que a função de sobrevivência é contínua e monótona não crescente, apresentando $S(t) = 1$ para o tempo de origem ($t = 0$) e se assume que $S(t) = 0$ quando $t \to \infty$.

Caso $S(t) > 0$ quando $t \to \infty$, dizemos que a função de sobrevivência é imprópria. Nesses casos, pode ser preferível utilizar modelos de cura, os quais são explicados no Capítulo 3.

É possível examinar a função de sobrevivência para toda a amostra ou realizar uma segmentação em grupos de participantes com base em alguma variável explicativa (ou covariável). Por exemplo, é viável comparar as curvas de sobrevivência entre mulheres e homens para investigar se existe alguma indicação de que essa variável possa ser significativa para compreender a variável de resposta. O termo "curvas de sobrevivência" é utilizado quando se realiza uma análise gráfica desse fenómeno.

2.2 Função de risco

A função de risco, denotada por $h(t)$, representa a taxa de ocorrência do evento de interesse no instante t, supondo que ainda não ocorreu referido evento até esse momento:

$$h(t) = \lim_{dt \to 0^+} \frac{P(t \geq T < t + dt | T \geq t)}{dt} \tag{2.2}$$

Embora seja informalmente vista como uma "probabilidade condicional", esta medida não é uma probabilidade porque ela não está limitada a 1 como ocorre com probabilidades, embora seja não negativa. Além disso, verifica a propriedade $\int_0^\infty h(t) = \infty$.

A relevância da função de risco reside na capacidade de compreender as variações da variável resposta ao longo do tempo. Essa variação pode manifestar-se como crescimento, decrescimento, constante ou adotar formas não monótonas, tais como em formato de banheira (*bathtube-shaped*), em formato de elevação (*hump-shaped*), entre outras possibilidades.

2.3 Função de risco acumulado

A função de risco acumulado, denotada por $H(t)$, corresponde à acumulada da função de risco da origem até o tempo t:

$$H(t) = \int_0^t h(u)du, \;\; t \geq 0 \tag{2.3}$$

A função de risco acumulado pode ser escrita com relação à função de sobrevivência $S(t)$, pois $H(t) = -log(S(t))$.

2.4 Função de distribuição

A função de distribuição, denotada por $F(t)$, representa a probabilidade de o evento de interesse ocorrer até o evento t, ou seja:

$$F(t) = P(T \leq t), \;\; t \geq 0 \tag{2.4}$$

A função de distribuição é monótona e não decrescente, verificando-se $F(t) = 0$ para $t = 0$ e $F(t) = 1$ para $t \to \infty$. Além disso, destaca-se que $F(t)$ pode ser escrita em função de $S(t)$, pois $F(t) = 1 - S(t)$.

Tendo em vista que ela é complementar à função de sobrevivência, não costuma ser utilizada em análise de sobrevivência. Contudo, se soubermos qual é a distribuição da variável resposta, seria mais fácil para construir modelos paramétricos.

2.5 Função densidade de probabilidade

A função densidade de probabilidade é uma taxa instantânea de ocorrência do evento no instante t:

$$f(t) = \lim_{dt \to 0^+} \frac{P(t \geq T < t + dt)}{dt} \tag{2.5}$$

2.6 Relação entre as funções de análise de sobrevivência

As funções de sobrevivência, risco, risco acumulado, distribuição e densidade de probabilidade estão relacionadas, ou seja, se conhecermos uma é possível determinar todas as outras, inequivocamente.

$$\begin{aligned} S(t) &= 1 - F(t) = \exp(-H(t)) \\ h(t) &= -\frac{dlog(S(t))}{dt} = \frac{f(t)}{S(t)} \\ H(t) &= -log(S(t)) \\ f(t) &= h(t) \exp(-H(t)) \end{aligned} \tag{2.6}$$

2.7 Censuras

Uma característica distintiva da análise de sobrevivência é a capacidade de lidar com informações censuradas (ou parciais), que se manifestam pela ausência de informações completas sobre a variável resposta. As censuras podem ocorrer à direita (sendo a mais comum), à esquerda e de forma intervalar.

Para que os métodos de análise de sobrevivência sejam válidos, as censuras precisam ser não informativas, ou seja, é necessário que os tempos de censura sejam independentes (entre si e dos verdadeiros tempos – não observados) e identicamente distribuídos.

A variável de interesse é o tempo decorrido desde a origem até a ocorrência do evento de interesse, representada por T. Seja C uma variável aleatória que denota o tempo de censura. O tempo observado, t, é idêntico ao tempo verdadeiro apenas quando não há censura. Denotamos por Y a variável aleatória composta por T e C (considerando que são independentes).

Para distinguir se o valor representado corresponde ao tempo até o evento de interesse ou ao tempo de censura, introduzimos uma variável indicadora δ. Esta é igual a 1 na presença do evento e igual a 0 na ocorrência da censura.

2.7.1 Censura à direita

A censura à direita em análise de sobrevivência ocorre quando não se observa o evento de interesse para alguns dos participantes durante o período de estudo, mas a informação disponível indica que esses participantes ainda estão em risco de experimentar o evento no futuro. Em outras palavras, a observação do tempo até o evento para esses participantes é truncada ou incompleta devido a circunstâncias além do controle do estudo e que o tempo do evento será posterior ao tempo t, representado como $t+$.

Utilizando a notação apresentada, temos $Y_i = min(T_i, C_i)$, para cada indivíduo i da amostra de tamanho n.

2.7.2 Censura à esquerda

Define-se censura à esquerda quando há a ocorrência do evento de interesse em um período anterior ao observado, podendo ocorrer, inclusive, antes do início do estudo, razão pela qual utili-

zamos a notação $t-$. Por exemplo, na aplicação do Capítulo 4, 69 participantes tiveram o evento de interesse (resolução de todos os sintomas), mas não sabemos quando isso ocorreu. Nesse caso, o tempo observado será da data de início dos sintomas até a data da entrevista telefónica, configurando-se censura à esquerda.

Utilizando a notação apresentada, temos $Y_i = max(T_i, C_i)$, para cada indivíduo i da amostra de tamanho n.

2.7.3 Censura intervalar

A censura intervalar ocorre quando não se pode registar com exatidão o tempo até a ocorrência do evento de interesse, registando-o num intervalo $(t_e, t_d]$. Trata-se, portanto, de uma generalização dos outros tipos de censura, pois a censura à direita pode ser escrita em intervalos do tipo $(t_c, +\infty]$, sendo t_c o tempo de censura e à esquerda pode ser escrita em intervalos de $(t_0, t_c]$, sendo t_0 o tempo de início dos estudos (ou pode escolher um tempo anterior esse). Contudo, o mais comum é utilizar o termo censura intervalar quando há mais tempos de inspeções, por exemplo, em consultas médicas, mesmo que não haja regularidade nas visitas.

2.8 Truncatura

Vimos, na seção anterior, que a censura ocorre quando a informação sobre a observação é parcialmente obtida, sendo anterior (à esquerda), posterior (à direita) ou entre (intervalar) os tempos observados. A truncatura também se refere à informação incompleta, mas nesse caso o indivíduo deixa de entrar no estudo devido ao critério de seleção inerente ao planejamento da pesquisa.

2.8.1 Truncatura à esquerda

Denomina-se truncatura à esquerda quando o indivíduo não pode ser observado devido à ocorrência do evento de interesse antes do período de observação. Isto pode fazer com que haja uma superestimação do evento de interesse, pois provavelmente os casos de maior risco foram desconsiderados. Esse viés de seleção é denominado de uso de dados prevalentes.

A diferença entre censura à esquerda e truncatura à esquerda consiste no fato do indivíduo ser observado no primeiro caso, mas no segundo não. Para elucidar a diferença entre eles, retomemos

a aplicação feita no Capítulo 4, onde há 69 casos de censura à esquerda, pois observamos os pacientes e sabemos que eles experenciaram o evento de interesse, embora não se lembravam à época do final do estudo (data da entrevista telefónica) quando tiveram a resolução de todos os sintomas. Contudo, os indivíduos que faleceram antes da entrevista telefónica (devido ao COVID-19 ou por outro motivo) não foram observados e, consequentemente, não puderam entrar no estudo e, por isso, são considerados truncados à esquerda.

2.8.2 Truncatura à direita

A truncatura à direita ocorre quando se selecionam apenas os indivíduos que tiveram o evento de interesse que, neste caso, costuma ser a morte. Por exemplo, analisar uma base com registos de óbito decorrentes de COVID-19 e, por ela, determinar quais são os maiores fatores de risco. Todos os pacientes com COVID-19 que estavam vivos à época do início do estudo são considerados truncados à direita.

2.9 Estimador de Kaplan-Meier

Tratando-se de estimação não paramétrica, o estimador de Kaplan-Meier (KM) é utilizado para obter estimativas para a função de sobrevivência, levando em consideração a presença de censura [36]. Ao considerar que a sobrevivência de cada indivíduo é independente de outro, considera o produto das probabilidades e, por isso, também é conhecido como estimador produto-limite.

Sejam $t_1 < t_2 < \cdots < t_k$ os k tempos ordenados e não censurados, d_j o número de eventos ocorridos no tempo t_j e n_j o número de indivíduos em risco, ou seja, os que ainda não tiveram o evento de interesse. O estimador de KM para a sobrevivência é dado pela seguinte expressão:

$$\widehat{S}_{KM}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.7}$$

O estimador de KM é consistente, tem normalidade assintótica, possui a forma de uma função escada, cujos saltos ocorrem apenas na presença de eventos, está definido apenas até o t_{max} observado e assume o valor zero caso essa observação não seja censurada. Caso contrário, o estimador de KM será uma função imprópria, ou seja, maior que zero para o t_{max}. Contudo, é

bom salientar que isso, por si, não indica que a sobrevivência também seja imprópria, pois o tempo observado pode apenas não ter sido suficiente.

Podemos escrever de outra forma o estimador de KM. Considere $Y_i = min(T_i, C_i)$ os tempos observados e ordenados[1], a variável indicadora $\delta_i = I(T_i \leq C_i)$ e o peso

$$W_i = \frac{\Delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[j]}}{n-j+1}\right].$$

Então, a forma alternativa para 2.7 é:

$$\widehat{S}_{KM}(t) = 1 - \sum_{i=1}^{n} W_i I\left(Y_i \leq t\right) \tag{2.8}$$

É fácil verificar que a estimativa de 2.8 é igual à empírica

$$\widehat{S}_{KM}(t) = \frac{1}{n} \sum_{t_j > t} I\left(t_j > t\right)$$

quando não há a presença de censura.

Podemos estimar a variância do estimador de KM através da fórmula de Greenwood, dada por:

$$\widehat{var}(\widehat{S}_{KM}(t)) = \left(\widehat{S}_{KM}(t)\right)^2 \sum_{i:t_j \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{2.9}$$

2.9.1 Intervalo de confiança para a sobrevivência

Considerando que o estimador de KM é consistente, tem normalidade assintótica e temos a estimativa da sua variância, podemos construir intervalos de confiança da forma $[LI, LS]$ para $\hat{S}(t)$ ao grau de confiança de $100(1-\alpha)$%, da seguinte forma:

$$\begin{aligned} LI &= S(t) - Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\widehat{S}_{KM}(t))} \\ LS &= S(t) + Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}(\widehat{S}_{KM}(t))} \end{aligned} \tag{2.10}$$

onde LI e LS são o limite inferior e superior, respetivamente, e $Z_{1-\frac{\alpha}{2}}$ é o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição normal reduzida N(0,1).

[1] Em caso de empates, os tempos não censurados vêm primeiro.

Percebe-se que o intervalo de 2.10 é simétrico. Contudo, em alguns casos, ele pode apresentar valores fora do intervalo [0,1] da função de sobrevivência. Para contornar esse problema, o melhor é aplicar alguma transformação como, por exemplo, logística ou a complementar log-log [31].

2.9.2 Estimador de Beran

O estimador proposto por Beran [37] pode ser visto como uma generalização do estimador de KM, que permite considerar o efeito da covariável X na estimativa da sobrevevência. É definido por:

$$\hat{S}_h(t|x) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{\delta_{[i]} B_{h(i)}(x)}{\sum_{r=1}^{n} B_{h(r)}(x)}\right), \tag{2.11}$$

onde

$$B_{h(i)}(x) = \frac{K_h(x - X_{[i]})}{\sum_{j=1}^{n} \left(K_h(x - X_{[j]})\right)} \tag{2.12}$$

e K_h é o função kernel reescalada por um parâmetro de suavização h e $X_{[i]}$ é a covariável da observação i no tempo $Y_i = min(T_i, C_i)$.

2.10 Nelson-Aalen

O estimador de Nelson-Aalen estima a função de risco acumulado e é dado por:

$$\widehat{H}_{NA}(t) = \sum_{i=1}^{n} \frac{I\left(Y_i \leq t, \delta = 1\right))}{\sum_{j=1}^{n} I\left(Y_j \geq Y_i\right))} = \sum_{t_j \leq t} \frac{d_j}{n_j} \tag{2.13}$$

E a variância estimada é

$$\widehat{var}\left(\widehat{H}_t\right) = \sum_{t_j \leq t} \frac{d_j}{n_j^2}.$$

Considerando a associação entre as funções de risco acumulado e de sobrevivência (2.6), podemos estimar a sobrevivência pelo estimador de Nelson-Aalen, sendo, portanto, uma alternativa ao estimador de KM. Embora apresentem valores próximos, observa-se, nas aplicações práticas, que o estimador de Nelson-Aalen tem melhor comportamento quando as amostras são pequenas.

2.11 Testes não paramétricos para comparar sobrevivências

A hipótese nula (H_0) dos testes não paramétricos considera que as curvas de sobrevivência da variável explicativa analisada são iguais. Sob H_0, a distribuição assintótica é $\chi^2_{(k-1)}$, onde k é o número de categorias da variável explicativa.

Para a i-ésima categoria das k existentes, no instante fixado t, denotando por $W_i(t)$ um fator (ou peso), $O_i(t)$ o número de ocorrências observadas, $E_i(t)$ o número de ocorrências esperadas e $R_i(t)$ o número de indivíduos em risco.

Para determinar os testes para $k = 2$, basta considerar apenas uma das categorias, pois será simétrica para a outra. Por conveniência, escolhemos $i = 1$, conforme se verifica em 2.14.

$$\begin{aligned} O_1 - E_1 &= \sum_t W_i(t) \left(O_1(t) - E_1(t)\right) \\ var\left(O_1 - E_1\right) &= \sum_t W_i(t)^2 \frac{R_1(t) R_2(t) O_1(t) \left(R(t) - O(t)\right)}{R(t)^2 \left(R(t) - 1\right)} \\ Log - rank &= \frac{\left(O_1 - E_1\right)^2}{var\left(O_1 - E_1\right)} \end{aligned} \tag{2.14}$$

Para k > 2, é necessário utilizar operações matriciais, não obstante o raciocínio seja o mesmo de 2.14, bastando obter o vetor das diferenças das $(k-1)$ categorias e a matriz de variâncias-covariâncias de dimensão $(k-1)$ x $(k-1)$.

Quando $W_i(t) = 1$, temos o teste de log-rank; o teste de Peto-Peto apresenta $W_i(t) = \hat{S}_i(t)$. O de Gehan considera que $W_i(t) = R_i(t)$, o de Tarone-Ware utiliza $W_i(t) = \sqrt{R_i(t)}$ e o Fleming-Harrington possui $W_i(t) = \hat{S}_i(t)^p \left(1 - \hat{S}_i(t)^q\right)$. O teste de Fleming-Harrington é uma generalização do teste de log-rank $(q = p = 0)$ e de Peto-Peto $(p = 1$ e $q = 0)$, sendo que os parâmetros p e q servem para dar maior peso nos momentos iniciais $(p > 0$ e $q = 0)$ ou maior peso nos momentos tardios $(p = 0$ e $q > 0)$.

2.12 Estimação paramétrica

A estimação paramétrica depende da distribuição, que geralmente é desconhecida. O procedimento usual é considerar mais de uma distribuição e ver qual delas melhor se ajusta à amostra. O

método de estimação mais utilizado é o de máxima verosimilhança (Máxima Verosimilhança (MV)), pois considera as observações censuradas, é consistente, eficiente e ainda possui normalidade assintótica.

Para análise de sobrevivência, as distribuições mais comuns são a exponencial, Weibull, gama, lognormal, loglogística, Rayleigh, Pareto e Gompertz. Apresentaremos apenas as que foram utilizadas no Capítulo 4 (Weibull, lognormal e loglogística) e também a exponencial, pois ela se trata, na verdade, de um caso particular da Weibull.

2.12.1 Exponencial

De entre as distribuições mais utilizadas para modelar o tempo, é a mais simples, pois possui um único parâmetro positivo θ. A função de densidade de probabilidade, de distribuição, de sobrevivência, de risco e de risco acumulado são dadas, respetivamente, por:

$$f(t) = \theta \exp(-\theta t) \tag{2.15}$$

$$F(t) = 1 - \exp(-\theta t) \tag{2.16}$$

$$S(t) = \exp(-\theta t) \tag{2.17}$$

$$h(t) = \theta \tag{2.18}$$

$$H(t) = \theta t \tag{2.19}$$

Considerando que a função de risco desta distribuição não depende do tempo 2.18, ou seja, é constante, dizemos que sofre da propriedade de "falta de memória". Além disso, temos que o valor médio de vida desta distribuição é $\frac{1}{\theta}$.

2.12.2 Weibull

A função de densidade de probabilidade, de distribuição, de sobrevivência, de risco e de risco acumulado da distribuição de Weibull(θ, k) são dadas, respetivamente, por:

$$f(t) = k\theta^k t^{k-1} \exp(-\theta t)^k \tag{2.20}$$

$$F(t) = 1 - \exp(-\theta t)^k \tag{2.21}$$

$$S(t) = \exp(-\theta t)^k \tag{2.22}$$

$$h(t) = k\theta^k t^{k-1} \tag{2.23}$$

$$H(t) = (\theta t)^k \tag{2.24}$$

Percebe-se que, quando $k = 1$, a distribuição Weibull se torna uma exponencial de parâmetro θ. Além disso, observando-se a função de risco 2.23, tem-se que será monótona e crescente para $k > 1$ e decrescente para $0 < k < 1$.

2.12.3 Lognormal

Considere uma variável Z normal padrão, ou seja, $Z \sim N(0, 1)$. A função exponencial de Z terá distribuição lognormal Y, ou seja, $Y = \exp(Z) \sim LN(0, 1)$.

Generalizando para o caso em que T seja lognormal com valor esperado μ e variância σ^2, suas funções de densidade de probabilidade, acumulada, sobrevivência, risco e risco acumulado são dadas, respetivamente, por:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right) \tag{2.25}$$

$$F(t) = \phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{2.26}$$

$$S(t) = 1 - \phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{2.27}$$

$$h(t) = \frac{f(t)}{S(t)} \tag{2.28}$$

$$H(t) = -\log\left(S(t)\right), \tag{2.29}$$

onde Φ é a distribuição da normal padrão acumulada.

A distribuição lognormal tem a função de risco crescente até atingir um ponto máximo, depois fica decrescente tendendo a 0 quando $t \to \infty$. Embora o risco diminuir com o aumento do tempo não costuma ocorrer em situações práticas, esse cenário foi justamente o encontrado no exemplo do Capítulo 4, pois o risco de resolução de todos os sintomas parece diminuir para valores elevados de t, dando a impressão de que não ocorrerão mais eventos de interesse quando $t \to \infty$.

2.12.4 Loglogística

A distribuição loglogística possui os parâmetros λ e α. Suas funções de densidade de probabilidade, sobrevivência, risco e risco acumulado são dadas, respetivamente, por:

$$f(t) = \frac{\alpha\lambda t^{\alpha-1}}{\left(1+\lambda t^{\alpha}\right)^2} \tag{2.30}$$

$$S(t) = \frac{1}{1+\lambda t^{\alpha}} \tag{2.31}$$

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}} \tag{2.32}$$

$$H(t) = log\left(1+\lambda t^{\alpha}\right) \tag{2.33}$$

Para $0 < \alpha \leq 1$, a função de risco é monótona decrescente. Para $\alpha > 1$, a função de risco da distribuição loglogística possui o formato de banheira (*bathtub-shaped*), pois é crescente até atingir o valor máximo em $t = \left(\frac{\alpha-1}{\lambda}\right)^{\frac{1}{\alpha}}$, decrescendo para 0 quando $t \to \infty$. Por exemplo, segundo o modelo de regressão paramétrica com distribuição loglogística apresentado no Capítulo 4, temos $\alpha = 1,225$, confirmando que no cenário lá verificado o risco tem formato de banheira.

2.12.5 Interpretação dos modelos de regressão paramétrica

Os modelos de regressão paramétrica são modelos *Accelerated Failure Time* (AFT), no qual cada fator aumentará (ou diminuirá), em média, a sobrevivência. Para obtermos a interpretação

em AFT, aplicamos a exponencial dos coeficientes β do modelo.

Contudo, quando a distribuição adotada é a Weibull ou a exponencial, também podemos interpretar os modelos em termos de *Proportional Hazard* (PH), utilizando a fórmula

$$\exp(-\frac{\beta z}{\alpha}) \tag{2.34}$$

onde β são os coeficientes do modelo, z é o valor da covariável e α é a escala da distribuição.

2.12.6 Estimador de máxima verosimilhança

Seja T a variável aleatória que represente o tempo até o evento de interesse, $f(t)$ a função densidade da distribuição e $S(t)$ a sobrevivência.

Sem a presença de censura, a função de verosimilhança é o produtório dos tempos observados no conjunto de tempos observados O, ou seja,

$$L = \prod_{i \in O} f(t_i).$$

Com a presença de censura, lembrando que os tempos observados são independentes dos tempos censurados, podemos incluir a sobrevivência dos indivíduos censurados no produtório acima. Assim, denotando por D o conjunto dos indivíduos censurados à direita, teremos que

$$L = \prod_{i \in O} f(t_i) \times \prod_{i \in D} S(t_i+) = \prod_{i}^{n} f(t_i)^{\delta_i} S(t_i+)^{1-\delta_i}$$

,

onde δ_i é a função indicadora de evento ou não (censura) e t_i+ é para indicar que a censura é à direita ($t_i+ > t_i$).

Podemos considerar todos os tipos de censura censura na mesma expressão:

$$L = \prod_{i \in O} f(t_i) \times \prod_{i \in D} S(t_i) \times \prod_{i \in E} (1 - S(t_i-)) \times \prod_{i \in I} (S(t_d) - S(t_e)) \tag{2.35}$$

onde t_i- é para indicar que a censura é à esquerda, pertencente ao conjunto E, d e e são utilizados para representar um intervalo do tipo $(t_d, t_e]$ do conjunto I.

Portanto, o estimador de máxima verosimilhança permite a presença de censura e a estimativa para o parâmetro θ será o valor que maximiza a função L descrita em 2.35.

2.13 Modelos de riscos proporcionais de Cox

2.13.1 Estimação semi paramétrica e hipótese de riscos proporcionais

Proposto por Cox [38], o modelo de regressão de riscos proporcionais de Cox também considera a presença de censura (apenas à direita), com a vantagem de ser de fácil interpretação, configurando-se em uma das técnicas mais utilizadas em análise de sobrevivência.

Seja T o tempo e X um vetor de p covariáveis, ou seja, $X = (X_1, \ldots, X_p)^T$. A função de risco é definida por:

$$h(t|x) = h_0(t)\,\exp(\beta x), \tag{2.36}$$

onde $h_0(t)$ é a função de risco suporte, β é o vetor de coeficientes de regressão que representa o efeito de cada covariável em T.

Denominam-se de riscos proporcionais porque a razão de riscos para dois indivíduos X_1 e X_2 é independente de t. Com efeito:

$$\frac{h(t|x_1)}{h(t|x_2)} = \exp\left(\beta\left(x_1 - x_2\right)\right) \tag{2.37}$$

Para verificarmos se o pressuposto de riscos proporcionais é válido, podemos recorrer à função *cox.zph()* da biblioteca *survival* do R, que considera os resíduos de Schoenfeld [39], que permitem analisar esse pressuposto globalmente e por cada covariável isoladamente.

Uma forma intuitiva de verificar graficamente o pressuposto de riscos proporcionais é construindo um gráfico com os resíduos de Schoenfeld pelo tempo. Caso a reta seja horizontal e, por isso, não depender do tempo, então temos a sugestão de que os riscos serão proporcionais.

2.13.2 Análise de resíduos

Em análise de sobrevivência, foram definidos diversos tipos de resíduos por causa da presença de censura. Utilizamos os de Cox-Snell, que permitem analisar também o ajustamento global do modelo (análise de diagnóstico), e os de Schoenfeld, que permitem analisar a hipótese de riscos proporcionais.

2.13.3 Resíduos de Cox-Snell

De acordo com o teorema da transformação uniformizante, se T for uma variável aleatória contínua, então tanto a função de distribuição $F(t)$ quanto a sobrevivência $S(t)$ terão distribuição uniforme $U(0,1)$. Por conseguinte, $H(t) = -log(S(t))$ tem distribuição exponencial Exp(1).

A função de risco acumulada para o modelo de Cox será:

$$H(t|\mathbf{x}) = \int_0^t h(u|\mathbf{x})du = \int_0^t h_0(u) \exp(\beta\mathbf{x})\, du = \exp(\beta\mathbf{x})\, H_0(t) \tag{2.38}$$

Assim, o resíduo para cada indivíduo i é dado pela estimativa $\hat{H}(t;\mathbf{x})$, ou seja,

$$r_i = \hat{H}(t_i) = \exp\left(\hat{\beta}x\right) \hat{H}_0(t_i) \tag{2.39}$$

onde $\hat{\beta}$ e $\hat{H}_0(t)$ são obtidos por MV.

Contudo, considerando que $t_i^+ < t_i$ e que, por conseguinte, $\hat{H}(t_i^+) < \hat{H}(t_i)$, o resíduo de Cox-Snell $r_i^{'}$ seria subestimado em presença de censura. Para contornar isso, geralmente adota-se um fator $c = 1$, resultando em $r_i^{'} = r_i + 1 - \delta_i$.

2.13.4 Resíduos de Schoenfeld

Para cada tempo t_i, o resíduo de Schoenfeld [39] é definido por

$$\mathbf{r}_i(\beta) = \mathbf{x}_i - \frac{\sum_{j \in R(t_i)} \mathbf{x}_j \exp\left(\beta^T \mathbf{x}_j\right)}{\sum_{j \in R(t_i)} \exp\left(\beta^T \mathbf{x}_j\right)} \tag{2.40}$$

sendo $R(t_i)$ a média ponderada de todos os indivíduos em risco nesse tempo.

Para padronizar os resíduos de Schoenfeld, divide-se pela variância do estimador β_k obtido no modelo de Cox, ou seja,

$$r_i^*(\beta_k) = \frac{r_i(\beta_k)}{var(\beta_k)}.$$

Intuitivamente, para que o pressuposto de riscos proporcionais seja válido, não pode haver tendência no gráfico de $r_i^*(\beta_k)$ pelo tempo.

2.13.5 Técnicas alternativas

O que fazer quando a hipótese de riscos proporcionais é rejeitada? Quando a base de dados é "grande", apesar de ser um termo subjetivo, podemos desconsiderar o fato dos riscos não serem proporcionais. Alternativamente, pode-se aplicar outro tipo de modelo, como os de regressão paramétrica e os de tempo de vida acelerado, ou aplicar o modelo de Cox com estratificação ou, ainda, considerar variáveis dependentes do tempo.

2.14 Método de seleção de modelos e de variáveis

2.14.1 Wald

O princípio da parcimónia é utilizado na construção de modelos estatísticos para enfatizar os fatores mais importantes, reduzir a multicolinearidade entre as covariáveis, além de apresentar um modelo mais simples, o que falicitaria a interpretação. Nesse sentido, o intuito é determinar quais são as variáveis imprescindíveis no modelo, ou seja, as que influenciam significativamente a variável resposta ou, em caso contrário, que tenha alguma relevância subjetiva para a área, por exemplo, algum fator de interesse clínico.

Seja β um vetor de parâmetros com dimensão p que contenha o efeito das p covariáveis. Para cada componente β_j, $j = 1, \ldots, p$, verificamos se é significativamente diferente de 0 através do seguinte teste de hipótese:

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

O teste de Wald pode ser utilizado quando o estimador é proveniente do método da MV [35], baseado na distribuição assintótica de $\hat{\beta}_j$:

$$Z = \frac{\hat{\beta}_j}{\sqrt{var(\hat{\beta}_j)}} \tag{2.41}$$

Sob H_0, Z tem distribuição assintótica qui-quadrado com 1 grau de liberdade (χ^2_1).

O problema do teste de Wald é que ele considera que cada compenente β_j seja independente entre si, o que não necessariamente ocorre na prática.

No intuito de minimizar os efeitos decorrentes da multicolinearidade, aplicamos o método *stepwise* que combina técnicas de remover uma covariável (*backward*) ou adicioná-la (*forward*),

uma a uma, por um processo iterativo, até obter um modelo adequado.

2.14.2 AIC

O critério de Akaike (AIC) também pode ser considerado no processo de seleção dos modelos estatísticos, pois ele penaliza o acréscimo de covariáveis no modelo, contribuindo para que o modelo final seja parcimonioso.

Suponha que um modelo foi construído com as p covariáveis disponíveis e que $\hat{L}$ seja o estimador de máxima verosimilhança obtido pelo modelo. O AIC é definido por:

$$AIC = 2p - 2\log(\hat{L}) \tag{2.42}$$

Pode-se replicar a fórmula para todos os modelos candidatos e escolher o que proporcionar o menor valor de AIC.

Capítulo 3

Modelos de cura

A função de sobrevivência é monótona e não crescente, tendendo a 0 para $t \to +\infty$. Contudo, em muitas situações práticas, pode existir uma parcela $1 - p$ da população que seja imune ao evento de interesse, razão pela qual se aplica técnicas de modelos de cura.

Há, essencialmente, duas vertentes para os modelos de cura, *promotion time cure models* (PTCM) [40, 41] e o modelo de cura de mistura (MCM) [42, 43], o qual foi utilizado no Capítulo 4.

O PTCM é um processo estocástico e foi designado resolver problema biológicos que analisava o número N de células cancerígenas, cuja distribuição era de Poisson com parâmetro positivo θ. Assumindo que a distribuição $F(t)$ de T era independente de N, a sobrevivência era definida por:

$$S(t) = \exp\left(-\theta F(t)\right).$$

O termo Mistura de MCM é utilizado porque considera que existem na população dois tipos de indivíduos misturados: aqueles que já experimentaram (ou experimentarão) o evento de interesse e os imunes, que nunca experimentarão o evento de interesse, mesmo que o tempo de observação seja ampliado indefinidamente.

Contudo, não é possível segregar inequivocamente os indivíduos suscetíveis ao evento de interesse e os imunes, pois as observações censuradas podem ser provenientes de indivíduos suscetíveis ($\upsilon = 0$) ou imunes ($\upsilon = 1$). Considerando que o tempo observacional seja suficientemente grande ($\tau_0 < \tau_C$), podemos admitir que os indivíduos que não tiveram o evento de interesse são provavelmente imunes.

Os modelos de cura de mistura podem ser paramétricos, semi-paramétricos e não paramétricos.

3.1 Modelos de cura de mistura

A notação utilizada será a mesma adotada no Capítulo 2, com a inclusão de:

- υ: variável indicadora que assume o valor 1 se a observação corresponder à população imune (ou curada) e 0 se pertencer à suscetível (ou não curada);
- Z: vetor de covariáveis utilizado na incidência;
- X: vetor de covariáveis utilizado na latência, que pode ser igual a Z;
- $p(z)$: incidência (efeito de longo prazo);
- $(1–p(z)) = \lim_{t\to+\infty} S(t|z) > 0$: a taxa de cura;
- $S_0(t|x) = P(Y > t|X = x, \upsilon = 0)$: função de sobrevivência dos suscetíveis, ou seja, a latência (efeito de curto prazo).

3.1.1 Modelo de cura de mistura paramétrico

O MCM paramétrico, proposto inicialmente por [42], considerou a incidência como uma constante e a latência como possuidora da distribuição lognormal.

Com a generalização de [44], embora ainda tenha considerado a incidência como uma constante, permitiu utilizar covariáveis para a latência com efeitos fixos e multiplicativos, além de generalizar a distribuição para qualquer uma que seja proveniente de um modelo AFT, ou seja, da seguinte forma:

$$\log(Y) = \beta_0 + \beta X + \sigma\epsilon \tag{3.1}$$

onde Y corresponde ao tempo observado, cada β é um parâmetro do modelo, σ é o parâmetro da distribuição e ϵ é o erro. Para a estimação dos parâmetros, utiliza-se o método da máxima verosimilhança, assim como no caso da regressão paramétrica.

Modelos paramétricos AFT podem ser construídos utilizando a biblioteca *gfcure* do R [45], cuja função tem o mesmo nome.

3.1.2 Modelo de cura de mistura semi-paramétrico

Os modelos de cura semi-paramétricos permitem maior flexibilidade que os modelos paramétricos, pois a incidência não precisa ser considerada constante e, para a latência, podemos construir um modelo de riscos proporcionais de Cox ou um modelo AFT.

Na prática, consideram que a incidência tenha um modelo de regressão logística:

$$p(z) = \frac{\exp(\alpha_0 + \alpha^T Z)}{1 + \exp(\alpha_0 + \alpha^T Z)} \tag{3.2}$$

Para a latência, quando utilizado um modelo AFT, as distribuições utilizadas são exponencial, Weibull, lognormal, gamma ou F. Caso a distribuição da função suporte seja Weibull [43], temos:

$$S_0(t|x) = \exp\left(-\lambda \exp(\beta^t x) t^\rho\right) \tag{3.3}$$

onde λ é um parâmetro de forma e ρ é um parâmetro de escala.

Caso seja escolhida a função suporte de riscos proporcionais (PH) de Cox, temos:

$$S_0(t|x) = S_{0,b}(t)^{\exp(\beta^T x)} \tag{3.4}$$

onde $S_{0,b}(t)$ é o termo não paramétrico do modelo.

O modelo com distribuição Weibull, assim como o de Cox, também pode ser interpretado como um modelo de riscos proporcionais (PH), razão pela qual podemos considerar a função de risco como sendo:

$$h(t|x) = h_0(t) c\left(\beta^T x\right) \tag{3.5}$$

onde $c(\cdot)$ é uma função *link* que pode ser paramétrica ou não paramétrica. É comum utilizar a função exponencial como uma função *link*.

Considerando a relação entre a função de risco e a de sobrevivência, podemos aplicar a seguinte transformação:

$$S(t|x) = S_0(t)^{c(\beta^T x)} \tag{3.6}$$

A estimação dos parâmetros não pode ser feita por máxima verosimilhança por causa do parâmetro υ que é desconhecido na ocorrência de censura, pois o indivíduo pode ser suscetível

($\upsilon = 0$) ou imune ($\upsilon = 1$). O método utilizado para a maximização da função de verosimilhança é o algoritmo *expectation-maximization* (EM) [46].

Para os modelos semi-paramétricos, pode-se aplicar a biblioteca *smcure* do R [47], cuja função tem o mesmo nome. Este modelo foi utilizado no Capítulo 4 e o código foi disponibilizado no Apêndice A.5.

3.1.3 Modelo de cura de mistura não paramétrico

O MCM não paramétrico possui mais flexibilidade que os anteriores, pois pode conter *splines* [48], estruturas *single-index* [49], métodos computacionais como o *bootstrap* [50] ou ainda ter estrutura completamente não paramétrica como proposta por [51], ao utilizar o estimador de Beran ($S_h(\cdot)$) [37] com os pesos $B_h(\cdot)$ e $K_h(\cdot)$ de Nadaraya-Watson . Ao considerar o maior tempo observado (T^1_{max}) e escolher os parâmetros de suavização h e b, temos:

$$\begin{aligned} \hat{S}_h(T^1_{max}|z) = 1 - \hat{p}_h(z) = \prod_{i=1}^{n} \left(1 - \frac{\delta_i B_{h(i)}(z)}{\sum_{r=1}^{n} B_{h(r)}(z)}\right) \\ \hat{S}_{0,b}(t|z) = \frac{\hat{S}_b(t|z) - (1 - \hat{p}_b(z))}{\hat{p}_b(z)} \end{aligned} \tag{3.7}$$

sendo $B_h(i) = \frac{K_h(z - Z_{(i)})}{\sum_{j=1}^{n} K_h(z - Z_j)}$ e $K_h(u) = \frac{1}{h} K(\frac{u}{h})$.

Os MCM modelos não paramétricos podem ser obtidos na biblioteca *npcure* do R [52].

3.2 Estimação dos parâmetros em MCM

3.2.1 Médoto MV

Dada a amostra aleatória de tamanho n, temos $(Z_i, X_i, Y_i, \delta_i)$, para $i = 1, \ldots, n$. Na presença de evento, temos que $\delta_i = 1$ e, consequentemente, $\upsilon_i = 0$, a densidade de T é dada por:

$$f(t|z, x) = p(z) f_0(t|x) \tag{3.8}$$

onde $f_0(t|x)$ é a densidade condicional dos indivíduos suscetíveis ($\upsilon_i = 0$). A sobrevivência condicional é dada por:

$$S(t|z,x) = 1 - p_z + p_z S_0(t|x) \tag{3.9}$$

Portanto, a função de verosimilhança é dada por:

$$L = \prod_{i=1}^{n} f\left(Y_i|X_i, Z_i\right)^{\delta_i} S\left(Y_i|X_i, Z_i\right)^{1-\delta_i} \tag{3.10}$$

Para obtermos o valor máximo da função 3.10, aplicamos ao logaritmo l da verosimilhança a otimização de Newton-Raphson:

$$l = \sum_{i=1}^{n} \delta_i [logp(Z_i) + logf\left(Y_i|X_i\right)] + \sum_{i=1}^{n} \left(1-\delta_i\right) log[1 - p(Z_i) + p(Z_i) S_0\left(Y_i|X_i\right)] \tag{3.11}$$

3.2.2 Algoritmo expectation-maximization (EM)

O algoritmo EM consiste em maximizar a função de verosimilhança L. Contudo, a função de verosimilhança escrita em 3.10 considera apenas os indivíduos com $\upsilon_i = 0$. A função de verosimilhança para υ desconhecido tem a seguinte forma:

$$L = \prod_{i=1}^{n} [p(Z_i) f_0(Y_i|X_i)^{\delta_i} S_0(Y_i|X_i)]^{1-\delta_i}]^{1-\upsilon_i} [1 - p(Z_i)]^{\upsilon_i} \tag{3.12}$$

Podemos decompor 3.12 em termos de incidência e latência ao logaritmo da verosimilhança (l) da seguinte forma:

$$l = l_1 + l_2$$

, onde:

$$\begin{aligned} l_1(p|\vec{\upsilon}) &= \sum_{i=1}^{n} \left(1 - \upsilon_i\right) log(p(Z_i) + \upsilon_i log[1 - p(Z_i)] \\ l_2(S_0|\vec{\upsilon}) &= \sum_{i=1}^{n} \delta_i log f_0(Y_i|X_i) + \left(1 - \upsilon_i - \delta_i\right) log S_0(Y_i|X_i) \end{aligned} \tag{3.13}$$

Considerando que o vetor composto pelos indicadores de cura $\vec{\upsilon} = (\upsilon_1 \ldots \upsilon_n)$ é desconhecido, ele será substiuído pela esperança condicional da incidência (l_1) ou da latência (l_2) no seguinte processo iterativo:

parte E de EM:

Para a r-ésima iteração da incidência e da latência, $p^{(r)}$ e $S_0^{(r)}$, respetivamente, de cada indivíduo i, calculamos a esperança condicional

$$w_i^{(r)} = E(\upsilon_i|p^{(r)}, S_0^{(r)}, X_i, Z_i, \delta_i) = (1-\delta_i)\,\frac{1-p^{(r)}(Z_i)}{1-p^{(r)}(Z_i)+p^{(r)}(Z_i)S_0^{(r)}(Y_i, X_i)} \quad (3.14)$$

A equação 3.14 proporciona o valor zero na presença de evento. Na presença de censura, a esperança condicional de 3.14 é, na verdade, apenas a seguinte probabilidade condicional:

$$P(\upsilon_i|X_i, Z_i, Y_i, \delta_i = 0) = \frac{1-p(Z_i)}{1-p(Z_i)+p(Z_i)S_0(Y_i|X_i)} \quad (3.15)$$

parte M de EM:

Esta etapa consiste em maximizar $l_1(\cdot)$ e $l_2(\cdot)$ para obter a $r+1$-ésima iteração, retornando à parte E do processo.

3.3 Qualidade do ajuste

No intuito de verificar se o modelo escolhido está bem ajustado, temos testes para a incidência ($p(z)$) e para a latência ($S_0(t)$), os quais dependerão do tipo de modelo escolhido.

3.3.1 Qualidade do ajuste na incidência

É comum considerar a incidência como tendo a distribuição logística e as hipóteses a serem testadas são [53]:

$H_0 : p = p_\theta$ para algum $\theta \in \Theta$ vs $H_1 : p \neq p_\theta$ para todo $\theta \in \Theta$,

onde Θ é o espaço finito de parâmetros e p_θ é uma função conhecida que possui o parâmetro $\theta \in \Theta$.

A estatística do teste é [54]:

$$\hat{\tau_n} = nh^{\frac{1}{2}}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}(Z_i) - p_{\hat{\theta}}(Z_i)\right)^2 \quad (3.16)$$

onde h é o parâmetro de suavização, $\hat{p}$ é um estimador não paramétrico de p e $\hat{\theta}$ é o estimador MV para θ sob H_0.

3.3.2 Qualidade do ajuste para a latência

Maller e Zhou [55] analisaram a correlação entre o ajuste do modelo e o estimador de KM e, caso fosse próximo a 1, indicaria, informalmente, que o modelo está bem ajustado.

Formalmente, as hipóteses a serem testadas são:

$H_0 : S_0 = S_{0\theta}$ para algum $\theta \in \Theta$ vs $H_1 : S_0 \neq S_{0\theta}$ para todo $\theta \in \Theta$.

3.4 Testes de hipóteses

Para que os modelos de cura sejam válidos, precisamos que o período observacional seja suficientemente grande [56] e que exista uma parcela $p < 1$ de indivíduos imunes [57].

3.4.1 Teste para a suficiência do tamanho do período observacional

Para que o período observacional seja suficientemente grande, precisamos testar se o ponto final à direita τ_0, proveniente da sobrevivência $S_0(t)$, é significativamente menor que τ_C, proveniente da distribuição da variável aleatória censura $C(t)$.

Então, as hipóteses a serem testadas são:

$H_0 : \tau_0 \geq \tau_C$ vs $H_1 : \tau_0 < \tau_C$,

com $\tau_0 = \sup\{t : S_0(t) > 0\}$ e $\tau_C = \sup\{t : C(t) > 0\}$.

A estatística do teste é $a_n = \left(1 - \frac{N_n}{n}\right)^n$, onde N_n é o número de observações não censuradas no intervalo $(2T_n^1 - T_n; T_n^1]$, composto por $T_n^1 = \max_{i:\delta_i=1}\{T_i\}$ e $T_n = \max_{i=1,\dots,n}\{T_i\}$.

Portanto, caso a_n seja menor que um valor α fixado (por exemplo, $\alpha = 0,05$), rejeitamos H_0 ao nível de significância α.

Contudo, aconselha-se também a construir um gráfico com a estimativa de KM porque o teste utiliza os períodos finais de observação, que costumam ser instáveis por possuírem menos indivíduos, dos quais muitos são informações censuradas.

3.4.2 Teste para a existência de imunes

Considerando que o tempo observacional tenha sido suficientemente grande, precisamos verificar se há indivíduos imunes. Esse teste, também proposto por Maller-Zhou, possui as seguintes hipóteses:

$H_0 : p = 1$ (sem imunes) vs $H_1 : p < 1$,

sendo p a incidência e 1–p é a taxa de cura.

Pela razão de verosimilhanças, temos:

$d_n = -2[l(\hat{\theta_0} - l(\hat{\theta_1}]$

onde $\hat{\theta_0} =$ arg max$_{\theta:p=1} l(\theta)$ e $\hat{\theta_1} =$ arg max$_\theta l(\theta)$

Sob H_0, tem-se:

$P(d_n < t) \to \frac{1}{2} + \frac{1}{2} P\left(\chi_1^2 < t\right).$

Contudo, mesmo quando rejeitamos H_0, é aconselhável observar um gráfico com a estimativa de KM para reforçar a necessidade da aplicação de modelos de cura.

Esse teste foi disponibilizado na biblioteca *npcure* do R [52] sob a função *testmz*. Utilizamos esta função no Capítulo 4 e o código foi disponibilizado em A.5.

3.5 Inferência quando a cura é parcialmente observada

Nas seções anteriores, apresentamos os modelos de cura clássicos, os quais consideram que apenas os indivíduos que experimentaram o evento de interesse ($\delta = 1$) possuem o indicador de cura υ conhecido (e igual a zero), não sendo possível distinguir os indivíduos imunes ($\upsilon = 1$) e os suscetíveis ($\upsilon = 0$). Esse cenário é verificado na maioria das aplicações de modelos de cura, inclusive no exemplo de aplicação do Capítulo 4.

Contudo, há contextos em que alguns dos indivíduos podem ser considerados imunes devido a algum critério preestabelecido como possuir tempo de sobrevivência superior a um limite, ou que exista algum diagnóstico médico. Por exemplo, considerar que se não houver remissão de cancro num paciente durante 5 anos, ele será considerado curado (ou imune).

Há ainda a mescla das técnicas de riscos competitivos com as de cura, pois na ocorrência de algum evento competitivo, podemos considerar que houve cura observada, pois não ocorrerá mais o evento de interesse.

Ignorar a existência de cura parcialmente observada ou considerar o problema apenas sob a perspetiva de análise de riscos competitivos (por exemplo, considerando a cura como um evento competitivo) pode fazer com que as estimativas fiquem enviesadas. Por exemplo, em [58], através de simulações, mostraram o viés de estimadores que negligenciaram a existência de fração de indivíduos já curados.

Para considerar a cura parcialmente observada, utiliza-se a variável indicadora ξ que será igual a 1 quando υ for conhecido. As situações englobadas serão as seguintes:

- **Ocorrência de evento**: $Y_i = T_i, \delta = 1, \upsilon = 0$ e $\xi = 1$. Assim, temos que $\xi\upsilon = 0$;
- **Evento competitivo ou indivíduo considerado curado**: $Y_i = C_i, \delta = 0, \upsilon = 1$ e $\xi = 1$. Assim, temos que $\xi\upsilon = 1$;
- **Observação censurada**: $Y_i = C_i, \delta = 0, \upsilon$ é desconhecido e, por isso, $\xi = 0$. Assim, temos que $\xi\upsilon = 0$, independentemente se $\upsilon = 0$ ou $\upsilon = 1$.

Para mais informações sobre cura parcialmente observada, sugerimos a consulta a [52, 59]. Para bibliotecas no R sobre o tema, tem-se *npcure* e *npcurePK*.

Capítulo 4

Exemplo de aplicação

4.1 Identificação da base de dados

Conduzimos um estudo coorte prospetivo cuja base de dados foi disponibilizada pelo laboratório do Centro Hospitalar Universitário de São João (CHUSJ), localizado em Porto, composta por indivíduos infetados por SARS-CoV-2 entre março de 2020 e dezembro de 2020, os quais foram acompanhados ambulatoriamente. Os pacientes decidiram voluntariamente participar do estudo, caracterizado pela aplicação de questionário telefónico aplicado por profissionais treinados. O protocolo do estudo foi aprovado pela Comissão de Ética do Centro Hospitalar de São João e da Faculdade de Medicina da Universidade do Porto em 16 de abril de 2020 [60].

A base foi inicialmente composta por 7539 indivíduos. Após excluirmos 460 que faleceram e 1796 que não conseguimos contacto telefónico, os 5283 pacientes restantes foram convidados a participar do estudo. Excluindo 1802 que se recusaram a participar do estudo, 271 menores de idade e 433 assintomáticos, a base final foi resultou em 2777 participantes (Figura 4.1).

4.1.1 Variáveis respostas

As variáveis respostas utilizadas são parecidas e, por isso, estão associadas e também pode provocar confusão ao leitor, razão pela qual as separamos nas subsecções posteriores. Além disso, explicamos o porquê da preferência da variável tempo até a resolução de todos os sintomas para a construção de modelos estatísticos.

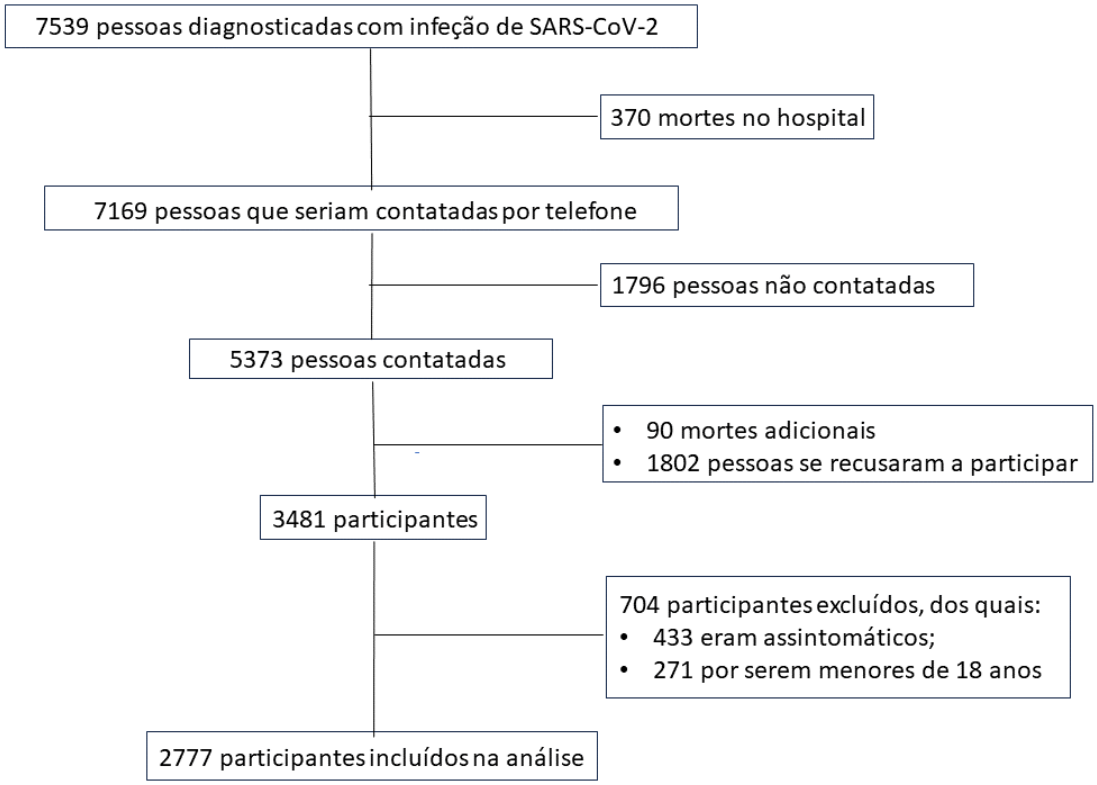

Figura 4.1: Fluxograma para a seleção dos participantes elegíveis.

4.1.1.1 Todos os sintomas resolvidos

Na data da aplicação do questionário telefónico (considerada final do estudo), os participantes foram inquiridos se todos os sintomas foram resolvidos até aquela data, cuja resposta poderia ser "sim" ou "não".

Agregando essa informação pela base total ou por categorias das variáveis explicativas, calculamos a proporção da resolução de todos os sintomas, que foi utilizada nas análises descritivas e na construção de intervalos de confiança.

De salientar que não utilizamos técnicas de modelos lineares generalizados porque temos observações censuradas. Há algoritmos de machine learning que até consideram a presença de censura, mas a variável "houve a resolução de todos os sintomas" refere-se, exclusivamente, à época do final do estudo.

4.1.1.2 Tempo até a resolução de todos os sintomas

Outra variável resposta proveniente do questionário telefónico era quantos dias, a partir da data do início dos sintomas, decorreram até a sua completa resolução.

A observação pode ser censurada à direita, pois o fim do estudo pode ser inferior ao tempo até a ocorrência do evento de interesse, ou à esquerda, caracterizada pelo fato de ter ocorrido o evento de interesse, mas não sabemos precisamente quando ocorreu a resolução de todos os sintomas.

Considerando que os modelos de cura de mistura (MCM) e os de Cox não permitem a utilização de censura à esquerda, excluímos as 69 observações censuradas em todos os processos para conseguirmos comparar melhor os resultados das diversas técnicas utilizadas.

Por fim, não obstante essa variável seja discreta (número de dias), consideramos que não houve prejuízo ao considerar o tempo até a resolução dos sintomas como se fosse contínuo, prática esta comum na aplicação de técnicas de análise de sobrevivência.

4.1.2 Variáveis explicativas

As variáveis explicativas utilizadas na nossa análise foram:

- **Idade**: idade, em anos, agrupada nos intervalos [18;30), [30;40), [40-50), [50-60), [60-70), [70-80) e [80; $+\infty$);
- **Sexo**: sexo registado ao nascimento, podendo assumir os valores "1" para masculino e "0" para feminino;
- **Escolaridade**: nível educacional agrupado em "secundário ou inferior", "bacharelado ou técnico" e "mestrado ou superior";
- **Renda**: renda mensal do agregado familiar, em euros, categorizada em [0;600), [600;1000], (1001;1500], (1500;2000], (2000;2500], (2500;3000], (3000, $+\infty$) e "não sabe ou não quis informar";
- **Perceção da renda**: perceção da renda do agregado familiar, categorizada em "insuficiente", "precisa ter cuidados com os gastos", "suficiente" e "confortável";
- **Estatuto fumador**: relativo ao hábito de fumar antes da infeção por SARS-CoV-2, podendo ser caracterizada em "fumador", "ex-fumador" e "não fumador";

- **Álcool**: relativo ao consumo (e frequência) de bebidas alcoólicas antes da infeção por SARS-CoV-2, agrupadas em "nunca", "até uma vez por semana", "menos que diariamente" (e mais que uma vez por semana) e "diariamente";
- **Atividade cotidiana**: relativo a como o participante passava a maior parte do dia antes da infeção por SARS-CoV-2, podendo assumir os valores "sentado", "em pé, mas sem correr" e "muito ativo";
- **Desporto**: relativo à prática desportiva antes da infeção por SARS-CoV-2, podendo assumir os valores "sim" e "não";
- **IMC**: utilizando o peso (em kg) e a altura (em cm) do participante antes da infeção por SARS-CoV-2, calculamos o índice de massa corporal e o categorizamos em "abaixo do peso ou ideal", "sobrepeso" e "obesidade";
- **Comorbidades**: presença de comorbidades caracterizada pelo fato do participante necessitar de cuidados prévios, podendo assumir os valores "sim" e "não";
- **Medicamentos**: uso regular de medicamentos, podendo assumir os valores "sim" e "não";
- **Doença respiratória do sono**: histórico de doenças respiratórias do sono, podendo assumir os valores "sim" e "não";
- **Hospitalização**: hospitalização no período agudo de SARS-CoV-2, podendo assumir os valores "sim" e "não".

4.2 Resumo de estudos relacionados à resolução de todos os sintomas

Antes de procedermos com as análises estatísticas da base de dados introduzida, é de bom alvitre apresentarmos resumidamente alguns estudos sobre a resolução dos sintomas decorrentes da infeção por SARS-CoV-2, efetuados ao redor do mundo, os quais utilizaram as mais variadas técnicas, consideraram diferentes períodos da infeção, analisaram diferentes quantidade de sintomas e apresentaram distintos tempos observacionais.

Barak Mizrahi e outros [8]

Publicado em novembro de 2022, em Israel, teve base de dados eletrónica composta por 1.913.234 de pacientes de todas as idades, possuídores de sintomas leves e que testaram positivo para a infeção de SARS-CoV-2 entre março de 2020 a outubro de 2021. Cada um dos 70 sintomas listados foi estudado isoladamente através de pareamentos em dois momentos da infeção (de 30 a 180 dias e de 180 a 360 dias) por idade, estatuto de vacinação, variantes de SARS-CoV-2 e sexo, e os riscos foram comparados através da razão de riscos.

Principais conclusões:

- a maioria dos sintomas desaparecem ao final de um ano após o diagnóstico;
- o fator sexo apresentou pouco efeito no risco de resolução dos sintomas;
- a vacinação reduz o risco de dispneia e outras doenças similares.

Olivier Robineau e outros [12]

Publicado em novembro de 2022, teve base de dados composta por 3972 participantes diagnosticados com infeção de SARS-CoV-2 entre abril e junho de 2020, que preencheram um questionário acerca de sintomas persistentes entre junho e setembro de 2021. Foram considerados como sintomas persistentes quando eles perduraram por, pelo menos, 2 meses. A técnica utilizada foi análise de sobrevivência e os fatores analisados foram idade, sexo, presença de comorbidades antes do diagnóstico e outros fatores socioeconômicos.

Principais conclusões:

- Após um ano, a proporção de indivíduos que tiveram resolução de todos os sintomas foi de 89,9%;
- Idade avançada (acima de 60 anos), sexo feminino, histórico de câncer, histórico de consumo de tabaco, obesidade e possuir mais de 4 sintomas foram fatores que diminuíram o risco de resolução de todos os sintomas.

Elisa Gentilotti e outros [22]

Publicado em julho de 2023, teve base de dados composta por 1796 participantes diagnosticados com infeção de SARS-CoV-2 entre fevereiro de 2020 e junho de 2021 acompanhados por até um ano. A técnica utilizada foi análise de componentes principais (PCA, em inglês) para fazer

agrupamentos dos sintomas. Além de verificar a persistência dos sintomas, o estudo levou em consideração o efeito dos sintomas persistentes na qualidade de vida.

Principais conclusões:

- 57% dos participantes não tiveram a resolução de todos os sintomas em um ano;
- sexo feminino, sintomas gastrointestinais e complicações renais durante a fase aguda da infeção foram fatores de risco para desenvolver sintomas persistentes.

B. Blomberg e outros [5]

Publicado em junho de 2021, teve base de dados composta por 357 participantes diagnosticados com infeção de SARS-CoV-2 em dois hospitais alemães entre fevereiro e abril de 2020. Foram considerados como sintomas persistentes quando eles perduraram por, pelo menos, 6 meses. A técnica utilizada para a determinação dos fatores associados aos sintomas persistentes foi regressão logística.

Principais conclusões:

- 61% dos participantes possuíram sintomas persistentes;
- sexo feminino com maior risco de possuir fadiga;
- 13% dos participantes menores de 16 anos possuíram sintomas persistentes.

Lixue Huang e outros [6]

Publicado em augusto de 2021, na China, teve base de dados composta por 1276 participantes diagnosticados com infeção de SARS-CoV-2 entre janeiro e maio de 2020, os quais responderam questionário. As técnicas utilizadas foram regressão logística e também comparações com grupos controles, caracterizados por pacientes que não foram infetados. Os períodos considerados foram 6 e 12 meses depois da infeção.

Principais conclusões:

- 68% dos participantes possuíam ao menos um sintoma aos 6 meses e 49% aos 12 meses;
- mulheres foram mais propensas a ter fadiga, fraqueza muscular, ansiedade ou depressão, quando comparadas aos homens.

Xiaojun Wu e outros [7]

Publicado em julho de 2021, o estudo teve o objetivo de acompanhar a permanência dos problemas respiratórios decorrentes da infeção de SARS-CoV-2 aos 3, 6, 9 e 12 meses após a infeção em 135 pacientes que foram hospitalizados por pneumonia severa e submetidos a exames médicos em um estudo longitudinal. A técnica estatística utilizada para a obtenção dos resultados foi regressão logística.

Principais conclusões:

- quase um terço dos pacientes possuíam algum problema respiratório após os 12 meses de acompanhamento;
- mulheres tiveram maior risco de comprometimento persistenta da difusão pulmonar.

Bishal Gyawali e Christopher Booth [61]

Publicado em março de 2021, o estudo é composto por análises e explicações técnicas de diversos tipos de sintomas persistentes (o critério foi de 4 semanas ou mais) decorrentes da infeção de SARS-CoV-2, além de análise detalhada de estudos realizados por outros pesquisadores, no intuito de que discussões possam ser multidisciplinares.

Max Augustin e outros [62]

Publicado em julho de 2021, teve base de dados composta por 958 participantes diagnosticados com infeção de SARS-CoV-2 entre abril e dezembro de 2020, com o objetivo de identificar sintomas persistentes aos 4 e 7 meses. Os sintomas considerados foram: ageusia (perda do paladar), asnomia (perda de olfato), fadiga e falta de ar. A técnica utilizada foi regressão logística.

Principal conclusão: 34,8% tiveram sintomas que perduraram por 7 meses.

Knut Stavem e outros [26]

Publicado em dezembro de 2020, teve base de dados composta por 451 participantes diagnosticados com infeção de SARS-CoV-2 até junho de 2020, os quais foram acompanhados por 6 meses. O objetivo do estudo era determinar a persistência dos sintomas innformados pelos participantes de uma lista com 23 sintomas. O número de sintomas foi utilizado para determinar a gravidade da infeção e para determinar a persistência dos sintomas. A técnica utilizada foi análise multivariada de dados.

Principais conclusões:

- 33% dos homens e 47% das mulheres não tiveram a resolução de todos os sintomas, no período de 6 meses;

- o número de sintomas e a presença de comorbidades estão associados à persistência dos sintomas.

Carole Sudre e outros [63]

Publicado em março de 2021, teve base de dados composta por 4182 indivíduos que informaram seus sintomas através de um aplicativo, os quais foram observados por até 12 semanas. O critério para a condição COVID prolongada foi a permanência dos sintomas por 28 dias ou mais.

Principais conclusões:

- aumento na idade e no IMC aumentaram o risco de ter a condição COVID prolongada;

- número de sintomas e sexo feminino são fatores de risco para COVID prolongada;

- 13,3% dos participantes apresentaram COVID prolongada e 2,3% do total de participantes não tiveram a resolução de todos os sintomas ao término das 12 semanas.

Jade Ghosn e outros [64]

Publicado em maio de 2021, teve base de dados composta por 1137 pacientes que foram hospitalizados durante a fase aguda da infeção. O objetivo do estudo foi averiguar a frequência de sintomas persistentes em 3 e 6 meses após a infeção, através da visita e avaliação médica, o que pode contribuir para que apenas sintomas relevantes sejam reportados.

Principais conclusões:

- 68% e 60% dos pacientes possuíam ao menos um sintoma persistente nas visitas M3 e M6, respetivamente;

- género feminino esteve associado a possuir 3 ou mais sintomas no M6;

- 25% dos participantes possuíram 3 ou mais sintomas persistentes no M6;

- houve pouca resolução dos sintomas no período decorrente entre o M3 e o M6;

Hannah Davis e outros [65]

Publicado em julho de 2021, teve base de dados composta por 3762 participantes de 56 países: 1020 com infeção confirmada de SARS-CoV-2 e 2742 com suspeita de infeção (resultado negativo ou que não foram testados) durante setembro e novembro de 2020. 66 sintomas foram rastreados no período de 7 meses da coorte. Foram medidos o impacto na vida, trabalho e no retorno à saúde. A técnica utilizada foi análise de sobrevivência, com adição de técnicas para agrupar (*cluster*) os diversos sintomas analisados.

Principais conclusões:

- 91% dos participantes tiveram os sintomas por mais de 35 semanas;
- excetuando a perda de paladar e olfato, as trajetórias nas curvas de sobrevivência não se alteraram para os grupos com infeção confirmada e para o com suspeita;
- fadiga foi o sintoma com maior prevalência após 6 meses de infeção.

Viet-Thi Tran e outros [11]

Publicado em abril de 2022, teve base de dados composta por 968 pacientes com infeção de SARS-CoV-2 confirmada, os quais responderam um questionário que continha 53 sintomas. A coorte francesa possui dados recolhidos entre dezembro de 2020 agosto de 2021. Outro critério de seleção foi que os participantes precisavam padecer dos sintomas por mais de 2 meses. Os fatores analisados foram idade (em agrupamentos), género, nível educacional e hospitalização durante a fase aguda da infeção. A técnica utilizada foi análise de sobrevivência.

Principais conclusões:

- género feminino está associado ao aumento de sintomas persistentes;
- 85% dos participantes não tiveram a resolução de todos os sintomas em 12 meses;
- os sintomas persistentes são contínuos ou podem ser intermitentes.

Jessica S. e outros [9]

Publicado em julho de 2021, teve base de dados composta por 96 pacientes maiores de 18 anos infetados entre fevereiro e abril de 2020 e que possuíam sintomas por pelo menos 5 meses. O objetivo do estudo foi analisar o desenvolvimento dos sintomas ao longo do tempo e os fatores considerados foram idade, género, gravidade da doença e presença de ANA (*antinuclear antibody*).

Principais conclusões:

- mulheres apresentaram significativamente mais sintomas neurocognitivos que os homens;
- um ano após a infeção, apenas 22,9% dos participantes tiveram a resolução de todos os sintomas;
- sintomas neurocognitivos podem persistir por ao menos um ano após a infeção, reduzindo a qualidade de vida.

Xiaoyu Fang e outros [66]

Publicado em dezembro de 2021, teve base de dados composta por 1233 participantes maiores de 60 anos. A técnica utilizada foi regressão logística.

Principal conclusão: 51,1% dos pacientes possuíam ao menos um sintoma após um ano.

Benjamin K. J. Tan e outros [20]

Publicado em junho de 2022, teve base de dados composta por 3699 pacientes provenientes de outras base de dados. O objetivo foi analisar a persistência da perda do paladar e do olfato decorrentes da infeção de SARS-CoV-2. A técnica utilizada foi modelos de cura de mistura paramétricos.

Principais conclusões:

- A permanência dos sintomas foi de 74,1%, 85,8%, 90,0% e 95,7% até 30, 60, 90 e 180 dias, respetivamente;
- mulheres apresentaram menos chance de recuperação do paladar e olfato;
- 5% dos indivíduos apresentaram sintomas persistentes ao término do estudo.

4.3 Análise de dados

As Tabelas 4.1 e 4.2 mostram as variáveis explicativas, suas categorias, frequências absolutas e relativas das variáveis explicativas. Além disso, apresentamos a estimativa pontual da proporção de resolução de todos os sintomas até a data final do estudo (entrevista telefónica) e um intervalo de confiança com grau de $(1-\alpha) = 95\%$, os quais foram construídos utilizando o teorema do limite central para proporções:

$$IC = [LI, LS] = \left[\hat{p}_k - z_{1-\frac{\alpha}{2}}\sqrt{\left(\hat{p}_k\frac{(1-\hat{p}_k)}{n_k}\right)}, \hat{p}_k + z_{1-\frac{\alpha}{2}}\sqrt{\left(\hat{p}_k\frac{(1-\hat{p}_k)}{n_k}\right)}\right] \quad (4.1)$$

onde $\hat{p}_k$ é a estimativa da proporção de resolução de todos os sintomas na categoria k, n_k será o número de participantes nessa categoria e $z_{1-\frac{\alpha}{2}}$ é o quantil de percentagem $\left(1-\frac{\alpha}{2}\right)$%, obtido pela distribuição normal padrão.

Aparentemente, a idade apresentou um comportamento em formato banheira (*bathtub-shaped*), pois as primeiras faixas de idade apresentaram uma maior proporção de resolução de todos os sintomas, decaindo até a faixa [50,59) anos e, após isso, voltou a aumentar. Quanto ao sexo ao nascimento, os homens apresentaram maior resolução de todos os sintomas que as mulheres (72,2% contra 57,9%). Para o nível de escolaridade, os que possuíram mestrado ou superior apresentaram maior proporção na resolução de sintomas que os participantes pertencentes às outras duas categorias, que apresentaram praticamente a mesma proporção. A variável renda não parece apresentar uma associação linear com a variável resposta e, ademais, quase 20% dos participantes não quiseram informar a renda. Em contrapartida, a perceção da renda parece estar linearmente associada à resolução de todos os sintomas, pois à medida que se teve uma melhor perceção da renda, maior foi a proporção da resolução de todos os sintomas. Para estatuto fumador, consumo de bebidas alcoólicas, atividade física cotidiana, prática de desporto, IMC, doença respiratória do sono e hospitalização não houve indícios de que estejam associadas à variável resposta. Participantes com comorbidades apresentaram menor proporção de resolução que os que não possuíam (66,6% contra 61,0%) e os que não necessitavam tomar medicamentos regularmente apresentaram maior proporção de resolução de todos os sintomas (66,6% contra 61,3%).

4.3.1 Atribuição de valores faltantes

Atribuímos o valor da mediana ou da moda para o preenchimento das informações faltantes (*missings*) nas variáveis explicativas. As explicativas que tiveram mais valores ausentes foram perceção de renda (7,3%), estatuto fumador (7,0%) e consumo de bebidas alcoólicas (7,0%). Quanto à data de início dos sintomas, utilizada apenas para calcular o tempo de censura para os que não tiveram a resolução de todos os sintomas até o final do estudo, tivemos 141 (5,1%) observações preenchidas como "não sabe". Em vez de excluirmos tais observações, utilizamos a data do diag-

Tabela 4.1: Variáveis explicativas, frequências absoluta e relativa, proporção e IC com 95%

Variável		Frequências	Proporção e IC com 95%
Idade	[18,30[	454 (16,3)	68,3 [64,0 , 72,6]
	[30,40[	488 (17,6)	64,1 [59,9 , 68,4]
	[40,50[	566 (20,4)	59,2 [55,1 , 63,2]
	[50,60[	510 (18,4)	58,0 [53,8 , 62,3]
	[60,70[	360 (12,9)	65,0 [60,1 , 69,9]
	[70,80[	263 (9,5)	68,4 [62,8 , 74,1]
	[80,+inf[	136 (4,9)	70,6 [62,9 , 78,3]
Sexo	Feminino	1680 (60,5)	57,9 [55,5 , 60,2]
	Masculino	1097 (39,5)	72,2 [69,6 , 74,9]
Escolaridade	Secundário ou inferior	1951 (70,3)	62,5 [60,3 , 64,6]
	Bacharelado ou técnico	626 (22,5)	63,6 [59,8 , 67,4]
	Mestrado ou superior	200 (7,2)	73,5 [67,4 , 79,6]
Renda	[0,600]	193 (6,9)	64,8 [58,0 , 71,5]
	]600,1000]	579 (20,8)	61,5 [57,5 , 65,4]
	]1000,1500]	581 (20,9)	61,6 [57,7 , 65,6]
	]1500,2000]	349 (12,6)	59,6 [54,5 , 64,7]
	]2000,2500]	214 (7,7)	71,0 [65,0 , 77,1]
	]2500,3000]	130 (4,7)	70,8 [63,0 , 78,6]
	]3000, +inf[	201 (7,2)	67,2 [60,7 , 73,7]
	Não informada	530 (19,1)	63,8 [59,7 , 67,9]
Perceção da renda	Insuficiente	268 (9,6)	51,9 [45,9 , 57,9]
	Precisa ter cuidado	703 (25,3)	61,3 [57,7 , 64,9]
	Suficiente	1196 (43,1)	65,2 [62,5 , 67,9]
	Confortável	610 (22,0)	67,9 [64,2 , 71,6]
Estatuto fumador	Fumador	259 (9,3)	62,5 [56,6 , 68,4]
	Ex-fumador	575 (20,7)	64,0 [60,1 , 67,9]
	Não fumador	1943 (70,0)	63,5 [61,4 , 65,6]

Tabela 4.2: Variáveis explicativas, frequências absoluta e relativa, proporção e IC com 95% (cont.)

Variável		Frequências	Proporção e IC com 95%
Álcool	Nunca	941 (33,9)	61,7 [58,6-64,8]
	Até uma vez	746 (26,9)	61,5 [58,0-65,0]
	Menos diariamente	567 (20,4)	64,9 [61,0-68,8]
	Diariamente	523 (18,8)	68,1 [64,1-72,1]
Ativ. cotidiana	Sentado	703 (25,3)	66,9 [63,4-70,4]
	Em pé, sem correr	1144 (41,2)	61,8 [59,0-64,6]
	Muito ativo	930 (33,5)	63,1 [60,0-66,2]
Desporto	Não	1392 (50,1)	63,5 [61,0-66,0]
	Sim	1385 (49,9)	63,5 [61,0-66,0]
IMC	Abaixo ou ideal	1114 (40,1)	64,5 [61,7 , 67,3]
	Sobrepeso	1046 (37,7)	64,6 [61,7 , 67,5]
	Obesidade	617 (22,2)	60,0 [56,1 , 63,9]
Comorbidades	Não	1241 (44.7)	66,6 [64,0 , 69,3]
	Sim	1536 (55.3)	61,0 [58,6 , 63,4]
Medicamentos	Não	1171 (42.2)	66,6 [63,9 , 69,3]
	Sim	1606 (57.8)	61,3 [58,9 , 63,7]
Doença resp. sono	Não	2659 (95.8)	63,7 [61,9 , 65,5]
	Sim	118 (4.2)	58,5 [49,6 , 67,4]
Hospitalização	Não	2336 (84.1)	63,9 [61,9 , 65,8]
	Sim	441 (15.9)	61,7 [57,1 , 66,2]

nóstico mais 7 dias, sendo que este valor foi a diferença média entre o tempo de diagnóstico e de início dos sintomas para as outras observações.

4.3.2 Número de sintomas

No questionário telefónico, estavam incluídos os 20 sintomas mais comuns causados pela infeção de SARS-CoV-2, além de um campo para o preenchimento de outros sintomas. Os sintomas reportados pelos participantes, em ordem decrescente de frequência relativa, foram: astenia (69,03%), mialgia (64,03%), perda ou alteração no paladar (58,62%), cefaleia (58,37%), perda ou alteração no olfato (56,21%), febre (52,07%), tosse seca (47,68%), arrepios (35,94%), diarreia (34,10%), dispneia (33,96%), rinorreia (33,06%), lombalgia (31,62%), dor torácica (25,50%), náusea ou vómitos (24,45%), cervicalgia (25,59%), dores abdominais (19,05%), odinofagia (18,69%), tosse com expetoração (14,98%), exantema (5,80%) e hemoptise (2,59%). O número de sintomas por participantes variou entre 1 até 19, com a mediana (mais precisamente, 50,95%) relatando possuir 7 sintomas.

A Tabela 4.3 e a Figura 4.2 mostram a associação entre o número de sintomas e a resolução de todos os sintomas.

Vemos que, à medida que aumenta o número de sintomas, diminui a proporção de resolução de todos os sintomas. Contudo, preferimos não utilizar essa variável na elaboração dos modelos, para dar mais ênfase às outras variáveis explicativas.

4.4 Estimativas de KM e testes não paramétricos

Na Figura 4.3, apresentamos a estimativa de Kaplan-Meier (KM) e as bandas com 95% de confiança para a variável tempo até a resolução de todos os sintomas. Ressaltamos que existiam 69 (2,5%) observações censuradas à esquerda, mas as excluímos porque não podemos utilizá-las nos modelos de Cox e nos de Cura. Essas curvas sugerem que o decréscimo é mais acentuado nos primeiros 50 dias após a data de início dos sintomas, com menor resolução até aproximadamente os 100 dias e, após isso, parece que a curva se estabilizou atingindo um platô. Por outras palavras, a taxa de resolução de todos os sintomas é alta nos primeiros dias, desacelerando até o ponto de parecer que, mesmo se aumentássemos o período observacional, não haveria mais o evento de

Tabela 4.3: Frequência absoluta e proporção de resolução dos sintomas por número de sintomas

Número de sintomas	**Frequência absoluta**	**Proporção de resolução**
1 ou outro	166	86,1%
2	211	87,2%
3	228	75,9%
4	239	72,8%
5	255	72,2%
6	263	67,3%
7	238	61,8%
8	206	59,2%
9	204	55,9%
10	178	47,8%
11	134	51,5%
12	138	52,9%
13	99	46,5%
14	92	34,8%
15 ou mais	126	32,5%

Figura 4.2: Frequência absoluta e proporção de resolução de todos os sintomas segundo o número de sintomas.

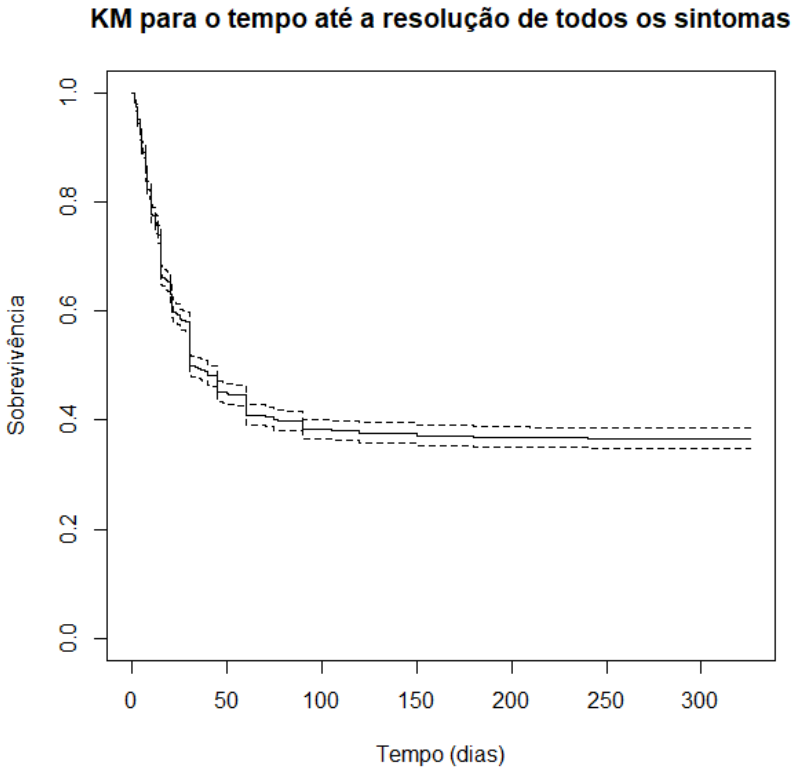

Figura 4.3: Estimativa de KM e bandas de confiança a 95%.

interesse. As medidas-resumo para o tempo até a resolução de todos os sintomas é (em dias): mínimo = 1, primeiro quartil = 14, mediana = 40, terceiro quartil = 126 e máximo = 326. Em termos absolutos, dos 2777 participantes, 1764 (63,5%) experienciaram a resolução de todos os sintomas até o final do estudo (329 dias após a data de início dos sintomas), dos quais 1138 (64,5%) tiveram resolução de todos os sintomas até o dia 29, 363 (20,6%) de 30 até 59 dias, 129 (7,3%) do dia 60 ao 89, 52 (3,0%) do dia 90 ao 119 e 82 (4,7%) necessitaram de 120 dias ou mais para terem o evento de interesse. É importante esclarecer que, neste contexto, quando a função de sobrevivência atinge 0%, isso indicaria que todos os participantes experimentaram a resolução de todos os sintomas (evento de interesse). Portanto, não se trata de uma interpretação literal de sobrevivência, como comumente observado em outras aplicações de análise de sobrevivência.

Convém ressaltar que, dos 1147 participantes que tiveram COVID prolongada (sintomas por 90 dias ou mais), apenas 134 tiveram resolução de todos os sintomas, ou seja, menos de 11,7% dos que padeceram desta condição.

A Tabela 4.4 contém os valores p para os testes log-rank e Peto-Peto. O nível de significância adotado nos testes foi de 5%.

Percebe-se, pela observação da Tabela 4.4, que ambos os testes culminaram nas mesmas conclusões. Assim, as variáveis que apresentaram diferenças significativas nas curvas de sobrevivência foram idade, sexo, nível de escolaridade, perceção da renda, consumo de bebidas alcoólicas, IMC, comorbidades, medicamentos e hospitalização.

Além dos testes mencionados, é aconselhável observar os gráficos com as curvas de sobrevivência, pois, às vezes, apenas uma categoria destoa das outras, fazendo com que a hipótese nula seja rejeitada, sem que a variável tenha boa capacidade de explicar devidamente o comportamento da variável resposta.

No que diz respeito à idade (Figura 4.4), parece que as curvas são semelhantes nos primeiros dias. Com o passar dos dias, as categorias [50, 60[e [60,70[têm maior desaceleração na resolução de todos os sintomas, apresentando menor resolução de todos os sintomas em relação às outras categorias.

O gráfico da Figura 4.5 sugere que, desde os primeiros dias, as curvas de sobrevivência descolam-se, fazendo com que a taxa de resolução dos todos os sintomas seja menor para as participantes do sexo feminino quando comparadas aos do sexo masculino.

Aparentemente, pelo gráfico da Figura 4.6, as três categorias da escolaridade possuem curvas

Tabela 4.4: Testes de log-rank e de Peto-Peto

Variável	Graus de liberdade	p valor (log-rank)	p valor (Peto-Peto)
Idade	6	0,002	0,004
Sexo	1	<0,001	<0,001
Escolaridade	2	0,022	0,013
Renda	7	0,096	0,174
Perceção da renda	3	<0,001	0,011
Estatuto fumador	2	0,752	0,378
Álcool	3	0,007	0,001
Atividade cotidiana	2	0,143	0,113
Desporto	1	0,687	0,682
IMC	2	0,045	0,036
Comorbidades	1	<0,001	<0,001
Medicamentos	1	<0,001	<0,001
Doença respiratória do sono	1	0,237	0,376
Hospitalização	1	0,003	0,001

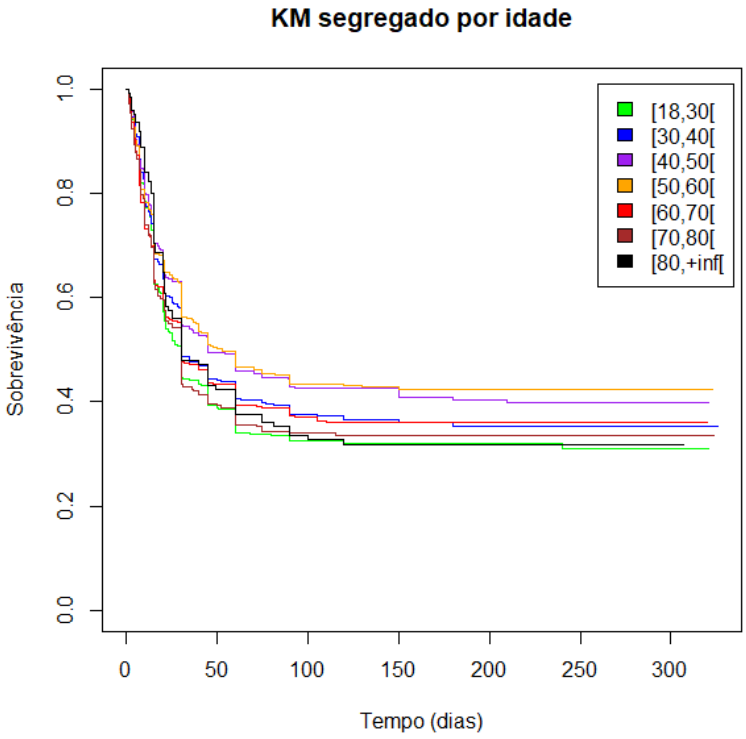

Figura 4.4: Estimativa de KM segregado por idade.

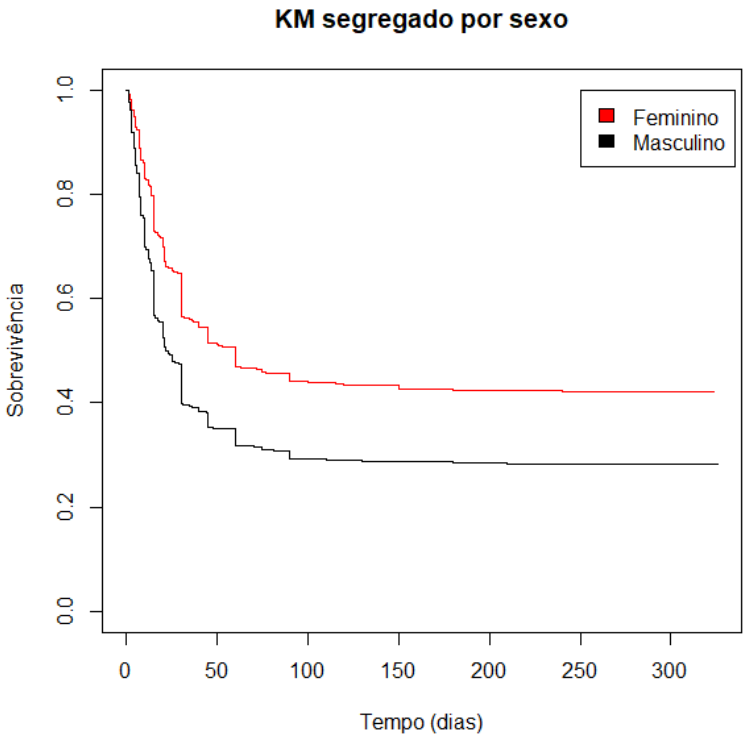

Figura 4.5: Estimativa de KM segregado por sexo.

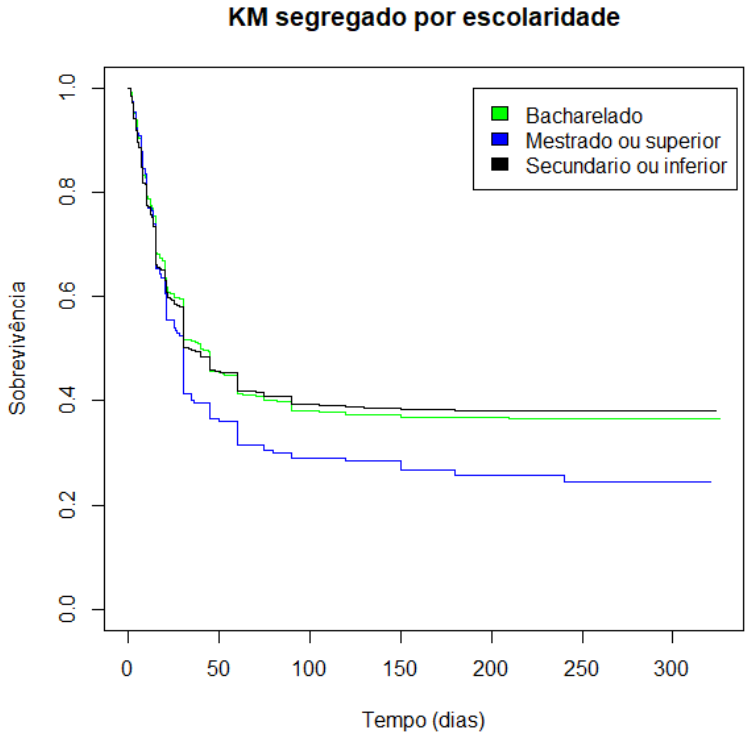

Figura 4.6: Estimativa de KM segregado por escolaridade.

de sobrevivência próximas nos primeiros dias e depois a resolução de todos os sintomas é mais acentuada para os que possuem mestrado ou superior.

Confirmando o que fora encontrado nos testes, a visualização do gráfico da Figura 4.7 sugere que as curvas segregadas por renda não são muito distintas, embora haja um descolamento delas um pouco maior nos últimos dias.

A interpretação da Figura 4.8 é que, após aproximadamente após o vigésimo quinto dia, à medida que melhora a perceção da renda, melhor é a taxa de resolução de todos os sintomas.

Confirmando o que fora encontrado nos testes, parece que as três curvas do estatuto fumador (Figura 4.9) estão sobrepostas.

Apesar da hipótese nula ser rejeitada em ambos os testes, não conseguimos observar diferenças entre as curvas de sobrevivência nas categorias da variável álcool (Figura 4.10). Além disso, caso realmente houvesse alguma diferença entre as curvas de sobrevivência, não é intuitivo que indivíduos que o aumento no consumo de bebidas alcoólicas acarrete em uma acelerção na resolução de todos os sintomas.

Além dos resultados dos testes, percebe-se, pela observação do gráfico da Figura 4.11, que

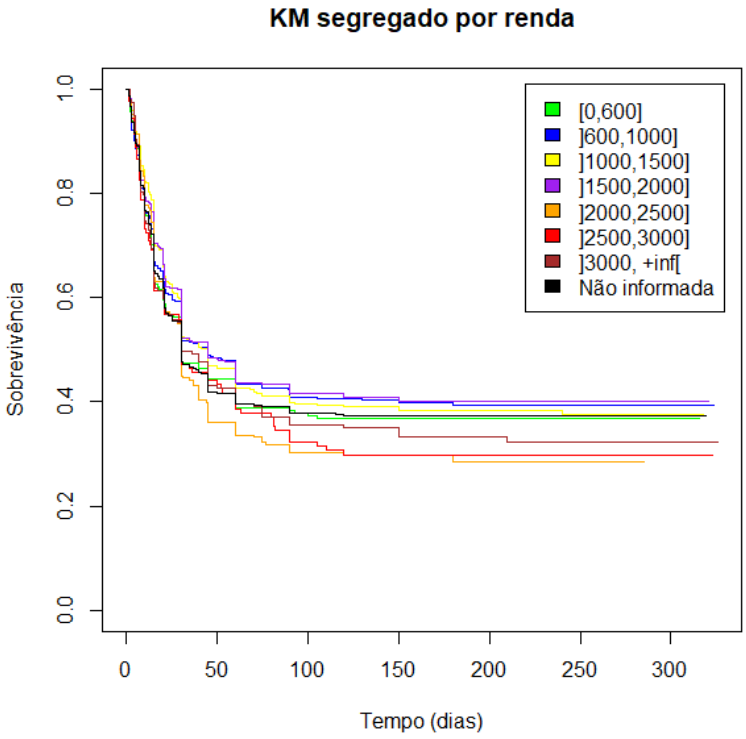

Figura 4.7: Estimativa de KM segregado por renda.

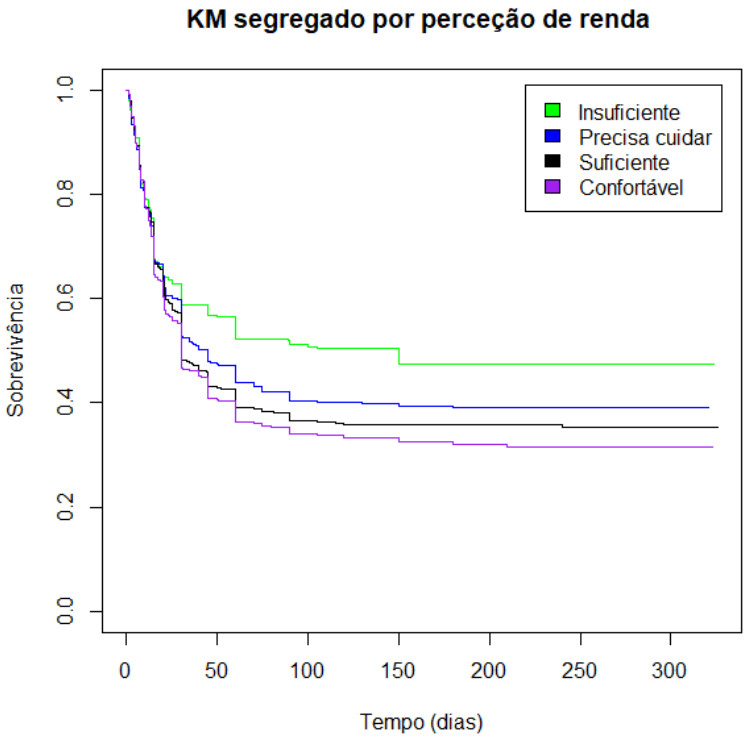

Figura 4.8: Estimativa de KM segregado por perceção da renda.

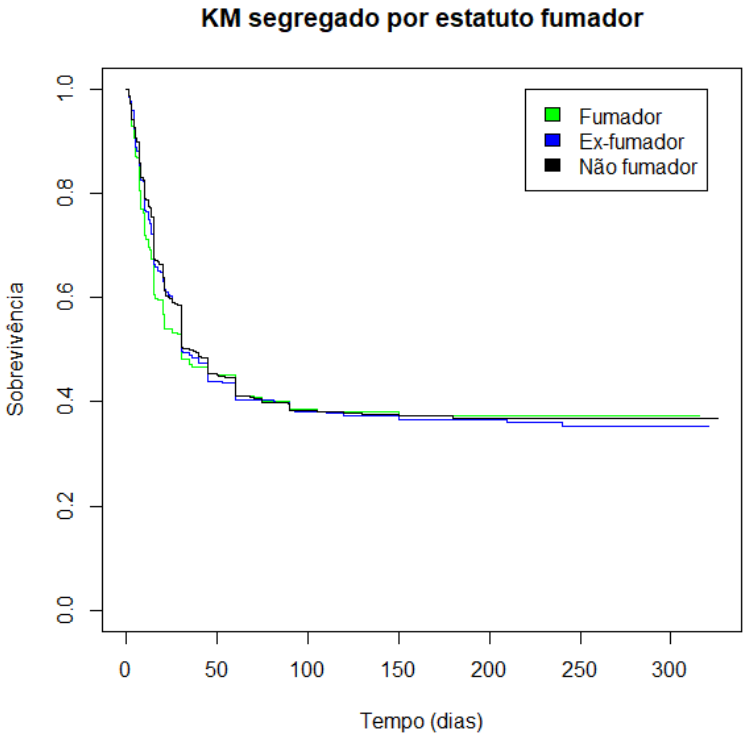

Figura 4.9: Estimativa de KM segregado por estatuto fumador.

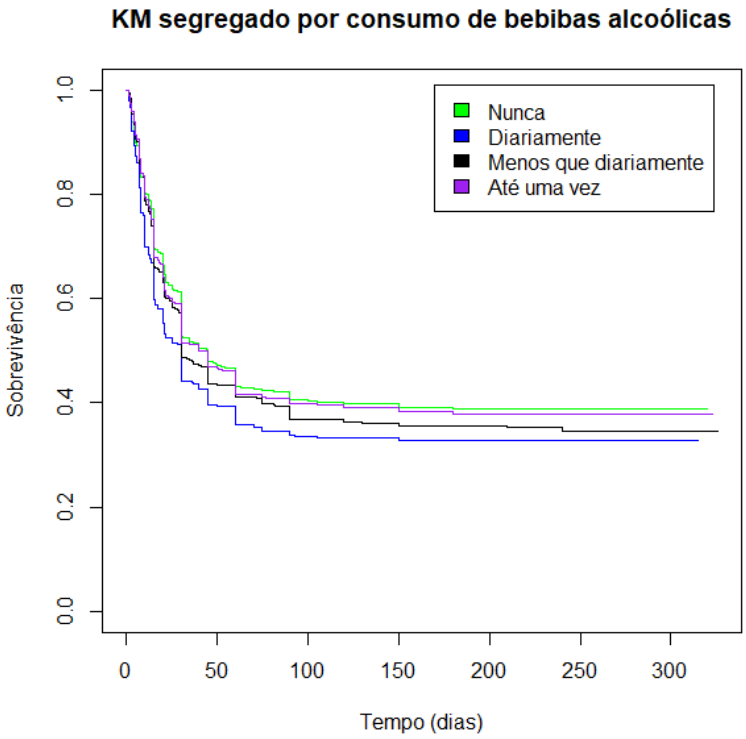

Figura 4.10: Estimativa de KM segregado por consumo de bebidas alcoólicas.

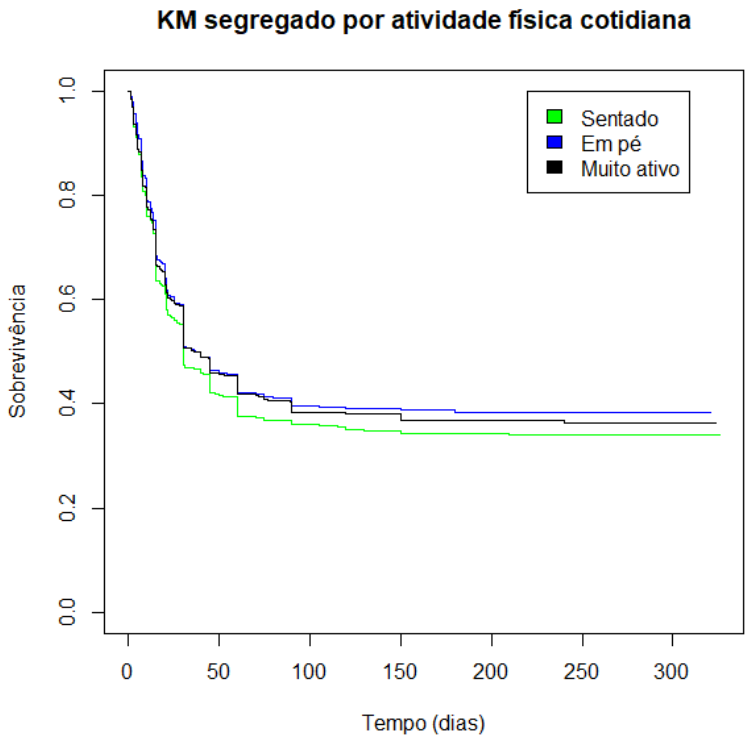

Figura 4.11: Estimativa de KM segregado por atividade física cotidiana.

as curvas de sobrevivências da atividade física cotidiana estão praticamente sobrepostas, ou seja, esta variável não parece estar associada ao tempo de resolução de todos os sintomas.

Confirmando o que fora encontrado nos testes, parece que as curvas da Figura 4.12 estão sobrepostas, indicando que a prática de desporto não está associada ao tempo de resolução de todos os sintomas.

Ao que indica na Figura 4.13, após aproximadamente ao quinquagésimo dia, os indivíduos obesos apresentaram menor taxa de resolução de todos os sintomas em relação às outras categorias.

A Figura 4.14 parece indicar que, nos primeiros dias, as curvas de sobrevivência da variável comorbidades se cruzam, e, após isso, a taxa de resolução de todos os sintomas é menos desacelerada para os participantes que não possuem comorbidades.

A Figura 4.15 sugere que as curvas se cruzam nos primeiros dias, com menor desaceleração na sobrevivência dos indivíduos que não necessitam tomar regularmente alguma medicação.

Observando-se o gráfico $a)$ da Figura 4.16, parece que há um descolamento entre as curvas de sobrevivência quando segregadas pela variável doença respiratória do sono. Contudo, como temos

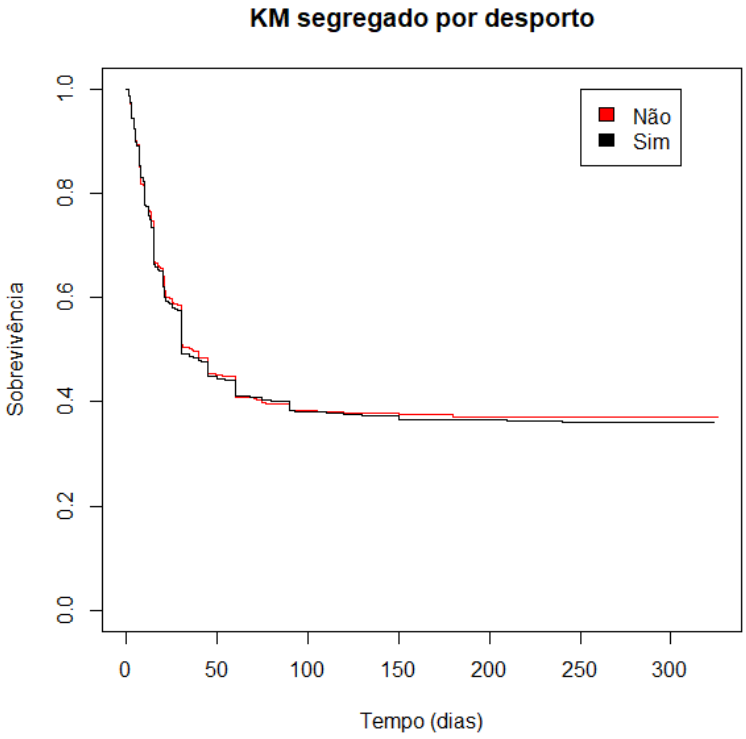

Figura 4.12: Estimativa de KM segregado por desporto.

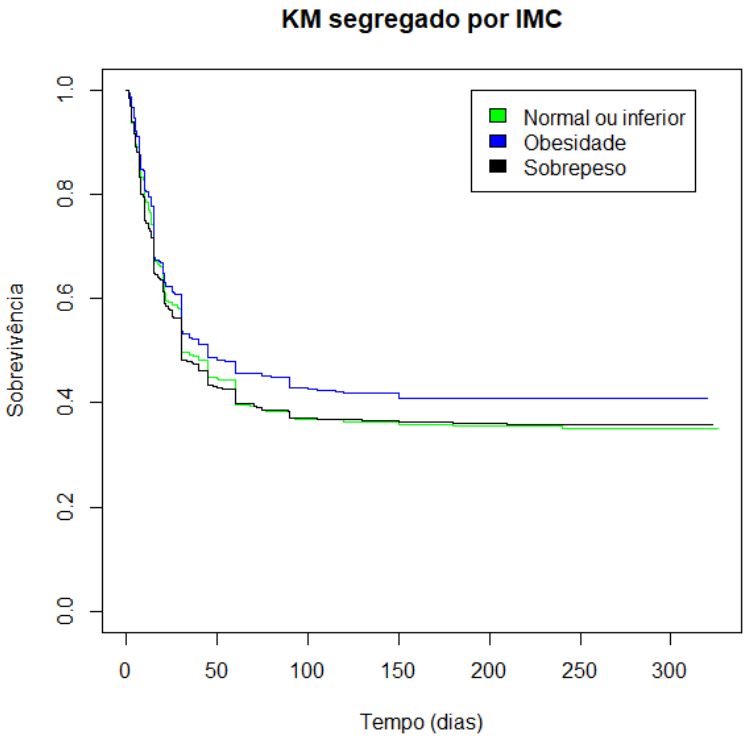

Figura 4.13: Estimativa de KM segregado por IMC.

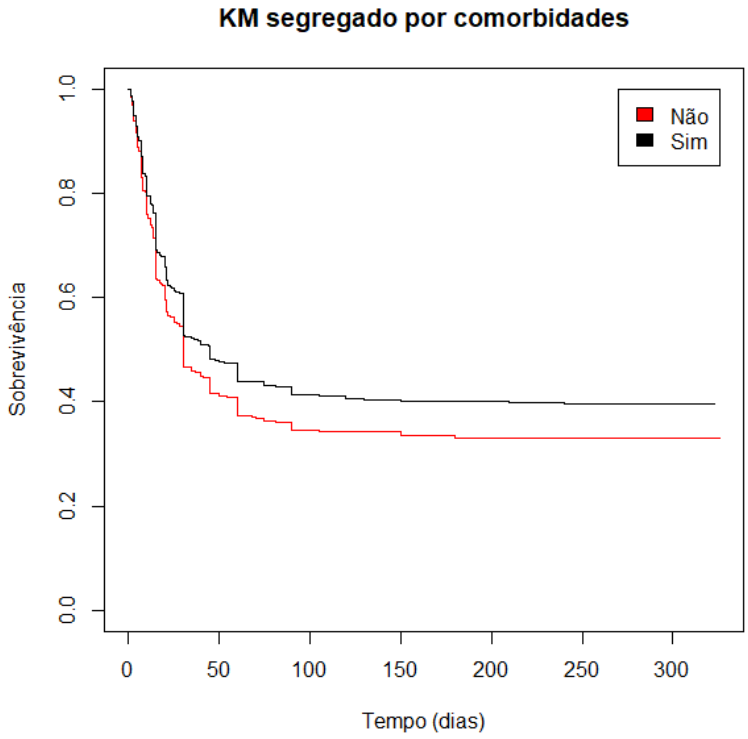

Figura 4.14: Estimativa de KM segregado por comorbidades.

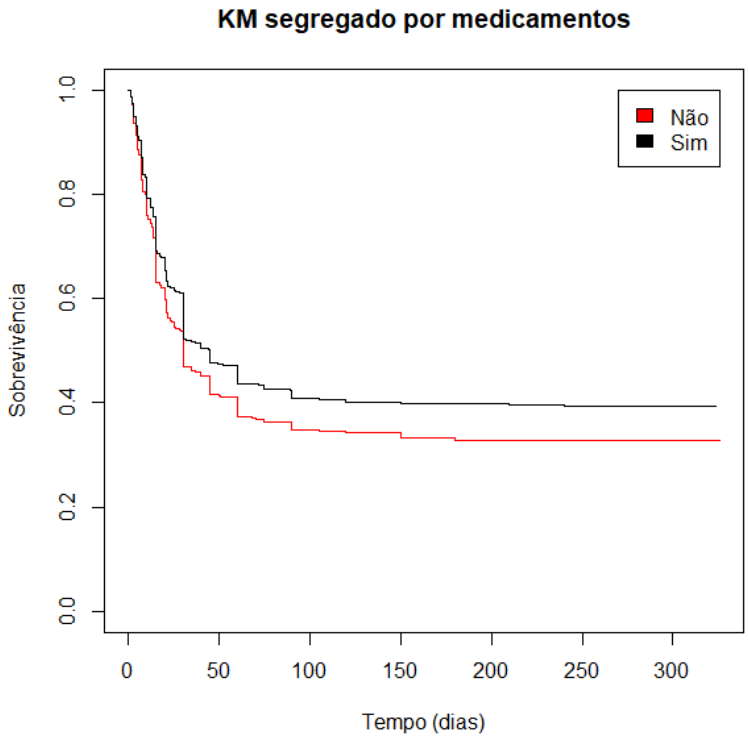

Figura 4.15: Estimativa de KM segregado por uso regular de medicação.

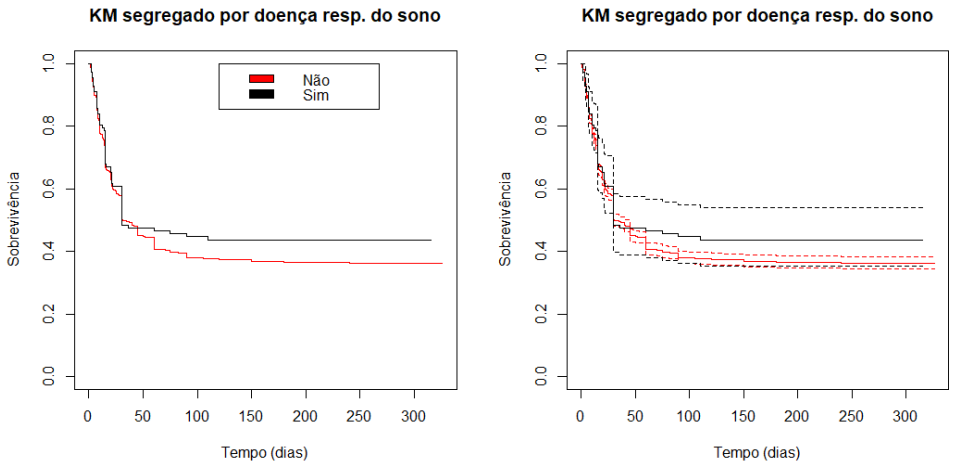

Figura 4.16: Estimativa de KM segregado por doença respiratória do sono: a) Sem bandas de confiança; b) Com bandas de confiança a 95%

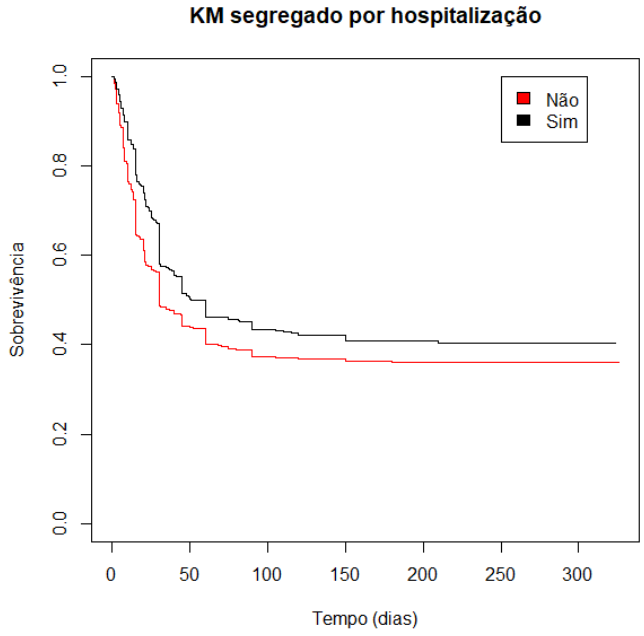

Figura 4.17: Estimativa de KM segregado por hospitalização.

poucos indivíduos que possuem doença respiratória do sono, não houve evidências suficientes para rejeitarmos a hipótese nula. O gráfico $b)$ da Figura 4.16 mostra o quão amplas são as bandas de confiança para a categoria "sim" dessa variável explicativa.

A Figura 4.17 mostra que, desde os primeiros dias, os participantes que não foram hospitalizados apresentaram maior taxa de resolução de todos os sintomas quando comparados aos que foram hospitalizados na fase aguda de SARS-CoV-2.

4.5 Modelo de regressão de Cox

4.5.1 Construção do modelo

Para a construção do modelo de regressão de riscos proporcionais de Cox, partimos do modelo completo, composto por todas as variáveis explicativas que apresentaram alguma relevância nos testes de log-rank e Peto-Peto, excluindo as menos significativas (ao nível de 5%) como em um processo *stepwise*.

Tabela 4.5: Coeficientes do modelo de regressão de Cox

Variável	**Categoria**	**HR**	**IC de 95%**	**p valor**
Sexo	Feminino	referência		
	Masculino	1,619	[1,469 , 1,785]	<0,001
Comorbidades	Sim	0,879	[0,797 , 0,970]	0,011
	Não	referência		
Perceção da renda	Insuficiente	0,721	[0,600 , 0,868]	<0,001
	Precisa ter cuidado	0,930	[0,824 , 1,049]	0,235
	Suficiente	referência		
	Confortável	1,039	[0,920 , 1,174]	0,538
Hospitalização	Sim	0,761	[0,660 , 0,878]	<0,001
	Não	referência		

Na Tabela 4.5, vemos que o modelo final que obtivemos foi composto por sexo, comorbidades, perceção da renda e hospitalização. O *Hazard Ratio* (HR) obtido sugere que homens tiveram, aproximadamente, 62% mais risco de ter o evento de interesse; participantes com comorbidades

apresentaram quase 12% menos risco ([1,000 , 0,879]) quando comparados aos que não possuíam comorbidades; à medida que melhorou a perceção da renda, aumentou o risco de resolução de todos os sintomas; participantes que foram hospitalizados na fase aguda da infeção por SARS-CoV-2 tiveram risco de, aproximadamente, 24% menor quando comparados aos que não foram hospitalizados.

4.5.2 Análise de Resíduos

Na Figura 4.18, temos o gráfico dos resíduos de Cox-Snell, para verificarmos se o modelo final ficou bem ajustado.

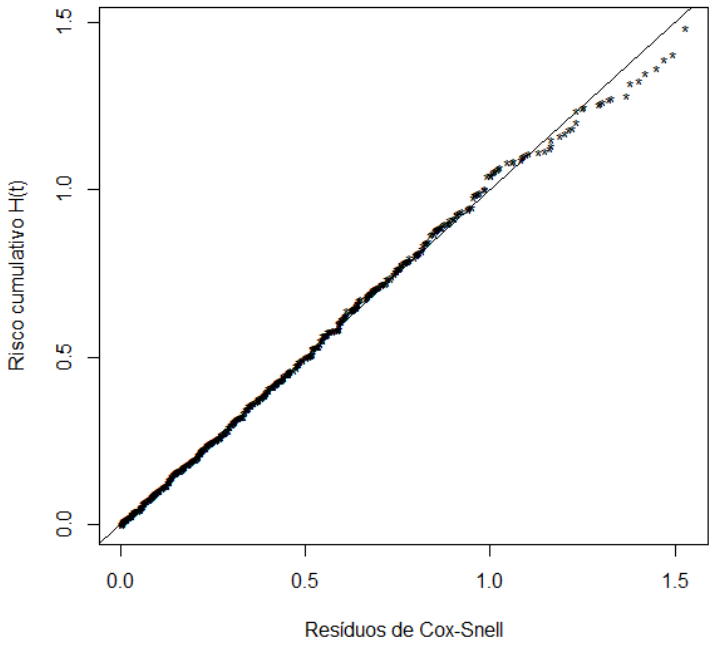

Figura 4.18: Resíduos de Cox-Snell.

Portanto, o modelo final parece se ajustar bem aos dados, pois a maioria dos pontos está sobre a reta, não obstante haver um descolamento para os riscos cumulativos dos tempos mais altos.

A Tabela 4.6 evidencia que a hipótese de riscos proporcionais foi rejeitada globalmente e pelas variáveis sexo, perceção de renda e hospitalização. Acreditamos que tal fato não inviabiliza o modelo obtido porque a base de dados pode ser considerada "grande".

Tabela 4.6: Teste para a hipótese de riscos proporcionais

Variável	Graus de liberdade	p valor
Sexo	1	<0,001
Comorbidades	1	0,795
Perceção da renda	3	0,009
Hospitalização	1	<0,001
GLOBAL	**6**	**<0,001**

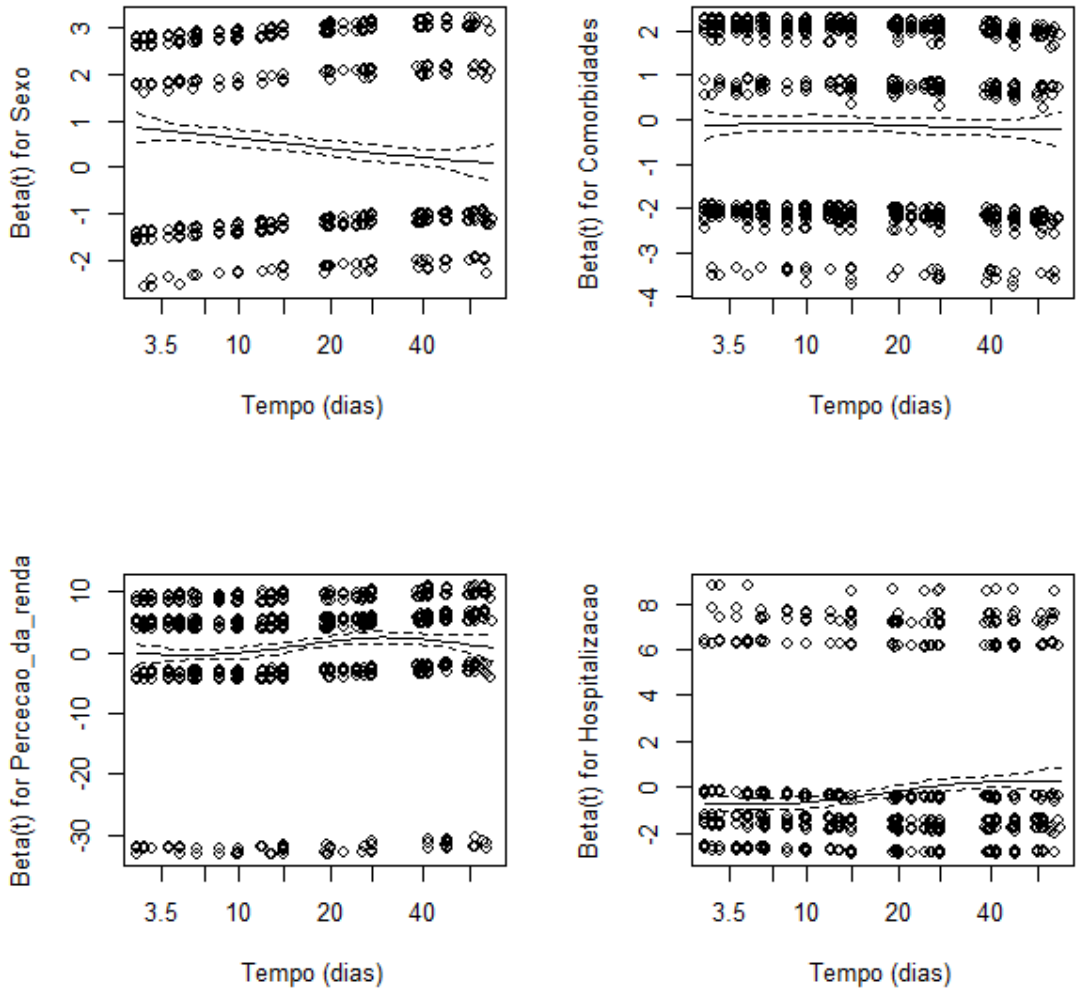

Figura 4.19: Resíduos de Schoenfeld.

Os gráficos da Figura 4.19 mostram os resíduos de Schoenfeld para que, intuitivamente, analisemos o pressuposto de riscos proporcionais. As retas parecem horizontais, mas isso ocorreu porque a presença de valores extremos (*outliers*) distorceu os gráficos, os quais são perceptíveis quando atentamos aos eixos verticais (betas).

Contudo, devemos lembrar que podemos ignorar esse pressuposto quando utilizamos grandes amostras (a base em questão possui 2708 participantes). Além disso, construímos modelos de regressão paramétrica para compararmos os resultados obtidos.

4.6 Modelos de regressão paramétrica

4.6.1 Weibull

Tabela 4.7: Coeficientes do modelo paramétrico com distribuição Weibull

Variável	**Categoria**	**Coeficiente**	**Erro padrão**	**p valor**
Intercepto		4,975	0,089	<0,001
Log(escala)		0,555	0,021	<0,001
Sexo	Feminino	referência		
	Masculino	-0,844	0,087	<0,001
Comorbidades	Sim	0,256	0,088	0,003
	Não	referência		
Perceção da renda	Insuficiente	0,647	0,164	<0,001
	Precisa ter cuidado	0,125	0,107	0,244
	Suficiente	referência		
	Confortável	-0,097	0,108	0,371
Hospitalização	Sim	0,448	0,127	<0,001
	Não	referência		

A Tabela 4.7 apresenta o modelo de regressão paramétrica sob a distribuição Weibull. Para a interpretação em termos de HR por categorias, aplicamos a fórmula de 2.34. Então, temos que o risco do evento de interesse para os participantes do sexo masculino foi de, aproximadamente, 62% maior que o do feminino ($\exp(-\frac{(-0,844)}{\exp(0,555)} = 1,623$); participantes com comorbidades tiveram

quase 14% a menos de risco de ter o evento de interesse quando comparados aos participantes sem comorbidades; à medida que piorou a perceção da renda, menor foi o risco de ter resolução de todos os sintomas; os indivíduos hospitalizados tiveram quase 23% a menos de risco de resolução de todos os sintomas quando comparados aos participantes que não hospitalizados.

4.6.2 Lognormal

Tabela 4.8: Coeficientes do modelo paramétrico com distribuição lognormal

Variável	Categoria	Coeficiente	Erro padrão	p valor
Intercepto		4,194	0,087	<0,001
Log(escala)		0,703	0,019	<0,001
Sexo	Feminino	referência		
	Masculino	-0,834	0,084	<0,001
Comorbidades	Sim	0,204	0,085	0,017
	Não	referência		
Perceção da renda	Insuficiente	0,459	0,150	0,002
	Precisa ter cuidado	0,079	0,103	0,443
	Suficiente	referência		
	Confortável	-0,030	0,108	0,7784
Hospitalização	Sim	0,524	0,019	<0,001
	Não	referência		

A Tabela 4.8 apresenta o modelo de regressão paramétrica sob a distribuição lognormal. Para a interpretação em termos de AFT por categorias, aplicamos a função exponencial aos coeficientes do modelo, cuja interpretação é em relação à sobrevivência, que, em nosso contexto, representa ter a resolução de todos os sintomas mais aceleradamente (exp$(\beta) < 1$) ou menos aceleradamente (exp$(\beta) > 1$). Assim, temos que o tempo de resolução de todos os sintomas para os participantes do sexo masculino é, em média, aproximadamente, 0,434 (exp$(-0,834)$) vezes (ou 56,6% menor que) o tempo das participantes do sexo feminino; para os participantes com comorbidades, o tempo de resolução de todos os sintomas é, em média, aproximadamente, 1,227 (exp$(0,204)$) vezes (ou 22,7% maior que) o tempo dos participantes sem comorbidades; à medida que piorou a

perceção da renda, maior foi o tempo até a resolução de todos os sintomas; para os participantes hospitalizados, o tempo até a resolução de todos os sintomas foi, em média, aproximadamente 1,689 (exp$(0,524)$) vezes (ou 68,9% maior que) o tempo dos indivíduos não hospitalizados.

4.6.3 Loglogística

Tabela 4.9: Coeficientes do modelo paramétrico com distribuição loglogística

Variável	Categoria	Coeficiente	Erro padrão	p valor
Intercepto		4,089	0,089	<0,001
Log(escala)		0,203	0,021	<0,001
Sexo	Feminino	referência	0,087	
	Masculino	-0,890		<0,001
Comorbidades	Sim	0,216	0,088	0,014
	Não	referência		
Perceção da renda	Insuficiente	0,564	0,161	<0,001
	Precisa ter cuidado	0,109	0,107	0,309
	Suficiente	referência		
	Confortável	-0,020	0,109	0,085
Hospitalização	Sim	0,546	0,124	<0,001
	Não	referência		

A Tabela 4.9 apresenta o modelo de regressão paramétrica sob a distribuição loglogística. Para a interpretação em termos de AFT, temos que o tempo de resolução de todos os sintomas para os participantes do sexo masculino é, em média, aproximadamente, 0,411 (exp$(-0,890)$) vezes (ou 58,9% menor que) o das participantes do sexo feminino; para os participantes com comorbidades, o tempo de resolução de todos os sintomas foi, em média, aproximadamente, 1,241 vezes (exp$(0,216)$) (ou 24,1% maior que) o tempo dos participantes sem comorbidades; à medida que piorou a perceção da renda, maior foi o tempo até a resolução de todos os sintomas; para os participantes hospitalizados, o tempo até a resolução de todos os sintomas foi, em média, aproximadamente 1,727 (exp$(0,546)$) vezes (ou 72,7% maior que) o tempo dos indivíduos não hospitalizados.

4.7 Comparação entre os modelos de Cox e paramétrico com distribuição Weibull

Considerando que o modelo paramétrico com distribuição Weibull pode ser interpretado em termos de riscos proporcionais e que as variáveis utilizadas em 4.7 foram as mesmas que em Cox (4.5), apresentamos a Tabela 4.10 comparando os HR deles.

Tabela 4.10: Comparação dos modelos de Cox e Weibull em termos de HR

Variável	Categoria	HR (Cox)	HR (Weibull)
Sexo	Feminino		
	Masculino	1,619	1,623
Comorbidades	Sim	0,879	0,863
	Não		
Perceção da renda	Insuficiente	0,721	0,690
	Precisa ter cuidado	0,930	0,931
	Suficiente		
	Confortável	1,039	1,057
Hospitalização	Sim	0,761	0,773
	Não		

Portanto, pela Tabela 4.10, percebe-se que os HR dos modelos de Weibull e de Cox ficaram bem próximos, o que reforça nossas conclusões sobre a influência dos fatores utilizados.

4.8 Comparação entre os modelos de regressão paramétrica

Observando os valores da Tabela 4.11, percebe-se que o modelo de Weibull atribuiu maior peso ao fator perceção da renda (variou de 1,909 a 0,988) e menor peso ao fator hospitalização, enquanto as interpretações dos modelos lognormal e loglogístico ficaram relativamente próximos.

Observando os valores da Tabela 4.12, conclui-se que o modelo de regressão paramétrica com distribuição lognormal é o melhor, seguido pelo modelo com distribuição loglogística. Isto é

Tabela 4.11: Comparação entre os modelos de regressão paramétrica em termos de AFT

Variável	Categoria	Weibull	Lognormal	Loglogística
Sexo	Feminino	referência	referência	referência
	Masculino	0,430	0,434	0,411
Comorbidades	Sim	1,291	1,227	1,241
	Não	referência	referência	referência
Perceção da renda	Insuficiente	1,909	1,582	1,758
	Precisa ter cuidado	1,133	1,082	1,115
	Suficiente	referência	referência	referência
	Confortável	0,988	0,97	0,980
Hospitalização	Sim	1,565	1,689	1,727
	Não	referência	referência	referência

Tabela 4.12: Comparação dos modelos de regressão paramétrica pelo critério AIC

Modelo	AIC
Weibull	18676,43
Lognormal	18213,21
Loglogística	18310,41

justificado porque ambos as distribuições assumem que o risco diminui com o decorrer do tempo, o que não acontece com a distribuição de Weibull, pois seu modelo assume que os riscos são proporcionais.

4.9 Modelo de cura de mistura

No nosso contexto, a sobrevivência representa a percentagem de participantes que ainda não tiveram a resolução de todos os sintomas até o instante t, não tendo, por conseguinte, o sentido literal. Da mesma forma, o sentido de "cura" e de "imunes" são contrários aos comumente utilizados, pois "cura" representa os sintomas que serão perenes e "imunes" representam os pacientes considerados que não terão a resolução de todos os sintomas.

Utilizamos técnicas de modelos de cura porque a estimativa obtida pelo estimador de Kaplan-Meier indicou a presença de um platô. Além disso, aplicamos o teste de Maller-Zhou e, ao nível de significância de 5%, rejeitamos a hipótese nula de que não haveria indivíduos imunes.

Para a construção do modelo de cura, utilizamos as variáveis sexo, comorbidades, perceção da renda e hospitalização, ou seja, todas as variáveis existentes nos modelos paramétricos e semi paramétricos, o que definimos como ponto de partida. Após isso, fomos excluindo e incluindo variáveis uma a uma, manualmente, até obter o modelo final, apresentado em 4.13. Esse modelo foi confirmado pela função *penPHcure* da biblioteca com o mesmo nome, que seleciona automáticamente as covariáveis presentes na incidência e na latência, conforme apresentado no apêndice A.5.

O modelo apresentado na Tabela 4.13 foi composto pelas variáveis sexo e comorbidades na incidência (efeito de longo prazo) e por sexo e hospitalização na latência (efeito de curto prazo). A interpretação dos coeficientes da incidência é em termos de *Odds Ratio* (OR), pois foi determinada a partir de um modelo logístico. Para a latência, proveniente de um modelo semi paramétrico de Cox, a interpretação dos coeficientes é em termos de HR. Assim, temos que a OR dos homens foi 1,755 (exp$(0,562)$) vezes a OR das mulheres e a OR dos participantes com comorbidades foi 0,757 (exp$(-0,279)$) vezes a OR dos participantes sem comorbidades; a HR dos homens foi 1,459 (exp$(0,562)$) vezes a HR das mulheres e a HR dos participantes com comorbidades foi 0,650 (exp$(-0,279)$) a HR dos participantes sem comorbidades.

Para estimarmos a proporção de imunes segundo o modelo de cura de mistura da Tabela

Tabela 4.13: Coeficientes do modelo de cura de mistura (MCM)

Variável	Categoria	Coeficiente	Erro padrão	p valor
Intercepto		0,609	0,069	<0,001
Variáveis para a incidência (efeito de longo prazo)				
Sexo	Feminino	referência		
	Masculino	0,562	0,098	<0,001
Comorbidades	Sim	-0,279	0,084	<0,001
	Não	referência		
Variáveis para a latência (efeito de curto prazo)				
Sexo	Feminino	referência		
	Masculino	0,378	0,051	<0,001
Hospitalização	Sim	-0,430	0,076	<0,001
	Não	referência		

4.13, apresentamos na Tabela 4.14 as 8 possíveis combinações obtidas por sexo, comorbidades e hospitalização (2 para sexo (masculino ou feminino) x 2 para comorbidades (sim ou não) x 2 para hospitalização (sim ou não)). Contudo, apenas as variáveis presentes na incidência alteram a estimativa da percentagem de imunes, pois a variável hospitalização está presente apenas no efeito de curto prazo (latência).

Portanto, pela Tabela 4.14, vemos que, aproximadamente, 41,8% das participantes do sexo feminino com comorbidades não experimentarão a resolução de todos os sintomas, independentemente se foram ou não hospitalizadas; para os homens sem comorbidades, a estimativa foi que, aproximadamente, 23,7% deles não terão o evento de interesse.

Podemos construir gráficos com as curvas de sobrevivência ajustadas pelo MCM. Nesse sentido, o gráfico da Figura 4.20 apresenta a previsão com a maior proporção de imunes (mulheres com comorbidades) e a menor proporção (homens sem comorbidades), evidenciando a maior diferença na incidência e o da Figura 4.21 evidencia uma diferença para a latência, ao variar apenas a variável hospitalização.

A curva tracejada da Figura 4.20 representa os homens sem comorbidades, que possuem a menor percentagem de imunes (23,7%) e a contínua representa as mulheres com comorbidades

Tabela 4.14: Estimativa de curados por

Sexo	Comorbidades	Hospitalização	1 - $\hat{p}$(z)
Feminino	Sim	Sim	41,8%
Feminino	Sim	Não	41,8%
Feminino	Não	Sim	35,2%
Feminino	Não	Não	35,2%
Masculino	Sim	Sim	29,0%
Masculino	Sim	Não	29,0%
Masculino	Não	Sim	23,7%
Masculino	Não	Não	23,7%

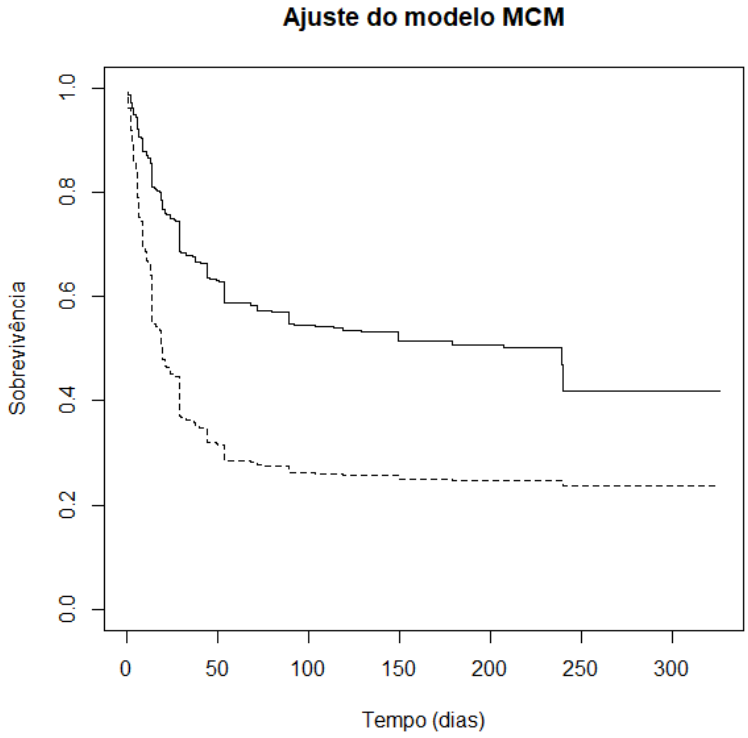

Figura 4.20: MCM para previsão de sobrevivência - enfoque na incidência.

Legenda: curva contínua: mulheres com comorbidades e que foram hospitalizadas; **curva tracejada:** homens sem comorbidade e que não foram hospitalizados.

(41,8%). A variável hospitalização não interfere no platô, sendo utilizada apenas para que o gráfico fosse construído.

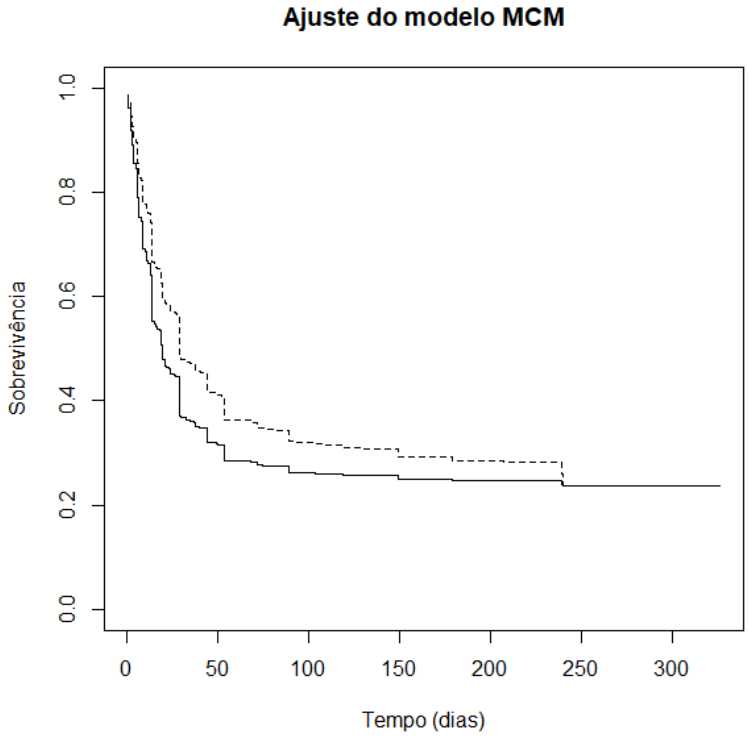

Figura 4.21: MCM para previsão de sobrevivência - enfoque na latência.

Legenda: curva contínua: homens sem comorbidades e que não foram hospitalizadas; **curva tracejada:** homens sem comorbidade e que foram hospitalizados.

As curvas da Figura 4.21 representam homens sem comorbidades, e a diferença entre as curvas consiste na hospitalização, pois na contínua os indivíduos não necessitam ser hospitalizados enquanto na tracejada necessitam de hospitalização. Observe que os pacientes se recuperam mais rapidamente na curva contínua, mas estão no mesmo platô porque possuem as mesmas variáveis da incidência (sexo masculino e sem comorbidades).

Capítulo 5

Discussão e Conclusões

5.1 Principais resultados deste trabalho

Na Figura 4.3 do Capítulo 4, apresentamos a estimativa de KM para a variável tempo até a resolução de todos os sintomas, que sugeria que mesmo se aumentássemos o tempo observável, não teríamos muitos eventos de interesse, pois parecia que a curva tinha atingido um platô. Isso foi evidenciado pela aplicação do teste de Maller e Zhou [57], pois rejeitamos a hipótese nula de que a curva de sobrevivência decairia para o valor zero ($p < 0,001$). É importante salientar que, entre os 1147 indivíduos (41,3% do total) que desenvolveram COVID prolongada (sintomas por mais de 90 dias), apenas 134 (11,7%) apresentaram resolução de todos os sintomas até o final do estudo.

Levando essas informações em consideração, construímos um MCM que teve sexo ao nascimento e hospitalização na incidência (efeito de longo prazo) e sexo ao nascimento e presença de comorbidades na latência (efeito de curto prazo). Além disso, estimamos a parcela de indivíduos "imunes" (no nosso contexto, imunes são os participantes que não experimentarão a resolução de todos os sintomas) segundo essas variáveis. Então, para os homens não hospitalizados e sem comorbidades, a proporção estimada de imunes é de 23,7% (mais baixa) e para as mulheres hospitalizadas e com comorbidades, a proporção estimada foi de 41,8% (a mais alta).

Ademais, apresentamos um modelo de riscos proporcionais de Cox, no qual as variáveis sexo ao nascimento, presença de comorbidades, perceção da renda e hospitalização foram significativas. Ademais, os modelos de regressão paramétrica com distribuições Weibull, lognormal e loglogística utilizaram os mesmos fatores. É importante recordar que, mesmo não sendo incorporadas a esses modelos, as variáveis idade, uso regular de medicamentos e nível educacional também

estão relacionadas ao tempo de resolução de todos os sintomas.

5.1.1 Limitações

Alguns fatores podem inviabilizar os resultados obtidos nesta dissertação. Primeiramente, não conseguimos isolar o tempo de resolução por sintoma porque esta informação foi única por participante.

Em segundo lugar, não foi possível mensurar a gravidade da doença, que deveria ser feita medindo a saturação de oxigenação e o nível de respiração [3]. Para contornar essa situação, utilizamos a variável hospitalização. O número de sintomas também poderia indicar a gravidade da doença ou ser utilizado como variável explicativa, como feito em outros estudos, mas preferimos desconsiderar essa informação porque minimizaria a influência das outras variáveis.

Em terceiro lugar, o fato do participante informar a data de resolução de todos os sintomas e a data do início dos sintomas pode acarretar em arredondamentos e em possíveis esquecimentos. Por exemplo, 69 participantes (2,48% do total) afirmaram experienciar o evento de interesse, mas não souberam indicar quando isso ocorreu (censura à esquerda). Essas observações foram consideradas apenas para a variável proporção de resolução de todos os sintomas (Tabelas 4.1 e 4.2). Caso isso ocorresse em maior quantidade, poderia sobrestimar a taxa de imunes, pois participantes que tiveram a resolução de todos os sintomas (evento de interesse) foram excluídos da amostra.

Por fim, apenas os pacientes que sobreviveram à época do questionário telefónico foram considerados elegíveis, o que se afigura como truncatura à esquerda. Esta exclusão pode superestimar a resolução de todos os sintomas, uma vez que casos mais graves podem ter sido omitidos.

5.2 Discussão

Os diversos estudos sobre a resolução de todos os sintomas, apresentados resumidamente no Capítulo 4, utilizaram as mais diversas técnicas e obtiveram as mais variadas conclusões, especialmente no que tange a porcentagens de indivíduos com sintomas persistentes. Com efeito, em [12], cerca de 10% dos indivíduos não tiveram a resolução de todos os sintomas em um ano, enquanto que em [9], quase 80% dos indivíduos não tiveram a resolução de todos os sintomas.

Alguns fatores que podem ter contribuído para essa amplitude no percentual de resolução de todos os sintomas:

- **época da infeção:** dependendo da época da infeção e da região de estudo, pode haver outras variantes do vírus ou os indivíduos podem estar vacinados;
- **gravidade da doença:** alguns estudos consideraram apenas sintomas leves (*mild*), que é a gravidade mais frequente.
- **período de observação:** alguns estudos consideraram apenas os primeiros meses, enquanto outros tiveram a duração de um ano;
- **quantidade de sintomas:** alguns estudos consideraram alguns sintomas específicos, o que pode acarretar na seleção de indivíduos com doenças mais graves, ou menos graves, dependendo dos sintomas investigados;
- **tamanho amostral:** variando de 96 indivíduos a 1.913.234, isso pode afetar especialmente as conclusões nos testes de hipóteses;
- **forma de mensuração:** houve estudos que um professional aferiu os sintomas, enquanto outros utilizaram valores informados pelos participantes;
- **aplicação de filtros:** alguns fatores podem ter sido utilizados para critério de seleção da amostra, por exemplo, considerar apenas maiores de 60 anos, ou apenas pacientes que foram hospitalizados, ou apenas pacientes que tiveram sintomas por pelo menos 5 meses;
- **influência externa:** pessoas próximas, mídias e outras fontes pode aumentar (ou diminuir) o ”medo”coletivo, que pode influenciar o indivíduo infetado. Aliás, a própria crença de estar infetado pode causar sintomas persistentes [10].

Contudo, mesmo com divergências nas porcentagens dos sintomas persistentes, a maioria dos estudos indicou a existência de COVID prolongada, o que, por si, já seria um motivo suficiente para não negligenciarmos o impacto dessa condição na qualidade de vida das pessoas. Casos de pacientes que precisaram de terapia de substituição renal [67] e de acarretamento de diabetes devido à infeção de SARS-CoV-2 [19,68] são exemplos de que os sintomas podem causar sequelas duradouras.

Muitos estudos confirmaram que o sexo feminino é um fator de risco para a persistência dos sintomas [5–7, 9, 11, 12, 20, 22, 27, 63]. Isso pode ser surpreendente porque a mortalidade é maior entre os homens [14, 18, 69]. Embora isso possa ter ocorrido devido à truncatura à esquerda, há discussões de natureza genética para entender a relação entre o sexo atribuído ao nascimento e também do género no desenvolvimento do COVID-19 [30], e em relação a outras doenças infeciosas [29].

O estudo de Robineau e outros [12] se assemelha ao nosso em tamanho amostral, variáveis utilizadas e período de infeção. Ademais, confirma a existência de sintomas persistentes após um ano e que as variáveis idade, sexo, obesidade e número de sintomas estão associados ao tempo até a resolução de todos os sintomas.

5.3 Conclusão

Ao término do estudo, cujo acompanhamento foi de aproximadamente 10 meses, 36,48% dos participantes não tiveram a resolução de todos os sintomas. Entre os 1147 (41,3%) pacientes que atingiram a condição COVID prolongada, apenas 11,7% experienciaram a resolução de todos os sintomas.

Os fatores associados ao tempo até a resolução de todos os sintomas mais significativos, segundo os modelos de regressão paramétrica e de riscos proporcionais de Cox, foram sexo, perceção da renda, presença de comorbidades e hospitalização durante a fase aguda da doença. Quando comparados às categorias de referência, em média, temos as seguintes interpretações para o risco de resolução de todos os sintomas: homens possuíram 61,9% mais risco; participantes com comorbidades tiveram 12,1% a menos de risco; à medida que piorou a perceção da renda, menor foi o risco; participantes que foram hospitalizados apresentaram 24% a menos de risco.

Pelo modelo de cura de mistura semi-paramétrico construído, o fator sexo esteve presente na incidência (efeito de longo prazo) e na latência (efeito de curto prazo), enquanto comorbidades esteve presente apenas na incidência e hospitalização apenas na latência. A estimativa para a proporção de indivíduos que não obteriam a resolução de todos os sintomas foi de 41,8% para mulheres com comorbidades, 35,2% para mulheres sem comorbidades, 29,0% para homens com comorbidades e de 23,7% para homens sem comorbidades, enquanto o fator hospitalização não interfere na estimativa de curados.

A maioria dos estudos analisados sugeriram a existência de indivíduos que não obtiveram a resolução de todos os sintomas, embora houvesse divergências entre as porcentagens de resolução. Além disso, muitos estudos concluíram que o sexo feminino, presença de comorbidades e hospitalização durante a fase aguda da doença são fatores de risco aos sintomas persistentes.

Bibliografia

[1] CDC, “Long COVID or post-COVID conditions,” Dec. 2023. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html

[2] J. Baj, H. Karakuła-Juchnowicz, G. Teresiński, G. Buszewicz, M. Ciesielka, R. Sitarz, A. Forma, K. Karakuła, W. Flieger, P. Portincasa, and R. Maciejewski, “COVID-19: Specific and non-specific clinical manifestations and symptoms: The current state of knowledge,” *J. Clin. Med.*, vol. 9, no. 6, Jun. 2020.

[3] R. T. Gandhi, J. B. Lynch, and C. Del Rio, “Mild or moderate covid-19,” *N. Engl. J. Med.*, vol. 383, no. 18, pp. 1757–1766, Oct. 2020.

[4] World Health Organization (WHO), “Post COVID-19 condition (long COVID),” Dec. 2023. [Online]. Available: https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition

[5] B. Blomberg, K. G.-I. Mohn, K. A. Brokstad, F. Zhou, D. W. Linchausen, B.-A. Hansen, S. Lartey, T. B. Onyango, K. Kuwelker, M. Sævik, H. Bartsch, C. Tøndel, B. R. Kittang, Bergen COVID-19 Research Group, R. J. Cox, and N. Langeland, “Long COVID in a prospective cohort of home-isolated patients,” *Nat. Med.*, vol. 27, no. 9, pp. 1607–1613, Sep. 2021.

[6] L. Huang, Q. Yao, X. Gu, Q. Wang, L. Ren, Y. Wang, P. Hu, L. Guo, M. Liu, J. Xu, X. Zhang, Y. Qu, Y. Fan, X. Li, C. Li, T. Yu, J. Xia, M. Wei, L. Chen, Y. Li, F. Xiao, D. Liu, J. Wang, X. Wang, and B. Cao, “1-year outcomes in hospital survivors with COVID-19: a longitudinal cohort study,” *Lancet*, vol. 398, no. 10302, pp. 747–758, Aug. 2021.

[7] X. Wu, X. Liu, Y. Zhou, H. Yu, R. Li, Q. Zhan, F. Ni, S. Fang, Y. Lu, X. Ding, H. Liu, R. M. Ewing, M. G. Jones, Y. Hu, H. Nie, and Y. Wang, “3-month, 6-month, 9-month, and 12-month respi-

ratory outcomes in patients following COVID-19-related hospitalisation: a prospective study," *Lancet Respir. Med.*, vol. 9, no. 7, pp. 747–754, Jul. 2021.

[8] B. Mizrahi, T. Sudry, N. Flaks-Manov, Y. Yehezkelli, N. Kalkstein, P. Akiva, A. Ekka-Zohar, S. S. B. David, U. Lerner, M. Bivas-Benita, and S. Greenfeld, "Long covid outcomes at one year after mild sars-cov-2 infection: nationwide cohort study," *BMJ*, vol. 380, 2023.

[9] J. Seeßle, T. Waterboer, T. Hippchen, J. Simon, M. Kirchner, A. Lim, B. Müller, and U. Merle, "Persistent symptoms in adult patients 1 year after coronavirus disease 2019 (COVID-19): A prospective cohort study," *Clin. Infect. Dis.*, vol. 74, no. 7, pp. 1191–1198, Apr. 2022.

[10] J. Matta, E. Wiernik, O. Robineau, F. Carrat, M. Touvier, G. Severi, X. de Lamballerie, H. Blanché, J.-F. Deleuze, C. Gouraud, N. Hoertel, B. Ranque, M. Goldberg, M. Zins, C. Lemogne, and Santé, Pratiques, Relations et Inégalités Sociales en Population Générale Pendant la Crise COVID-19-Sérologie (SAPRIS-SERO) Study Group, "Association of self-reported COVID-19 infection and SARS-CoV-2 serology test results with persistent physical symptoms among french adults during the COVID-19 pandemic," *JAMA Intern. Med.*, vol. 182, no. 1, pp. 19–25, Jan. 2022.

[11] V.-T. Tran, R. Porcher, I. Pane, and P. Ravaud, "Course of post COVID-19 disease symptoms over time in the ComPaRe long COVID prospective e-cohort," *Nat. Commun.*, vol. 13, no. 1, p. 1812, Apr. 2022.

[12] O. Robineau, M. Zins, M. Touvier, E. Wiernik, C. Lemogne, X. de Lamballerie, H. Blanché, J.-F. Deleuze, P. M. Saba Villarroel, C. Dorival, J. Nicol, R. Gomes-Rima, E. Correia, M. Coeuret-Pellicer, N. Druesne-Pecollo, Y. Esseddik, C. Ribet, M. Goldberg, G. Severi, F. Carrat, and Santé, Pratiques, Relations et Inégalités Sociales en Population Générale Pendant la Crise COVID-19-Sérologie (SAPRIS-SERO) Study Group, "Long-lasting symptoms after an acute COVID-19 infection and factors associated with their resolution," *JAMA Netw. Open*, vol. 5, no. 11, p. e2240985, Nov. 2022.

[13] Z. Wang, Y. Liu, L. Wei, J. S. Ji, Y. Liu, R. Liu, Y. Zha, X. Chang, L. Zhang, Q. Liu, Y. Zhang, J. Zeng, T. Dong, X. Xu, L. Zhou, J. He, Y. Deng, B. Zhong, and X. Wu, "What are the risk factors of hospital length of stay in the novel coronavirus pneumonia (COVID-19) patients? a survival analysis in southwest china," *PLoS One*, vol. 17, no. 1, p. e0261216, Jan. 2022.

[14] E. P. Sreedevi and P. G. Sankaran, “Statistical methods for estimating cure fraction of COVID-19 patients in india,” Jun. 2020, unpublished.

[15] M. Pedrosa-Laza, A. López-Cheda, and R. Cao, “Cure models to estimate time until hospitalization due to COVID-19: A case study in galicia (NW spain),” *Appl. Intell.*, vol. 52, no. 1, pp. 794–807, 2022.

[16] R. R. Castro, R. S. C. Santos, G. J. B. Sousa, Y. T. Pinheiro, R. R. I. M. Martins, M. L. D. Pereira, and R. A. R. Silva, “Spatial dynamics of the COVID-19 pandemic in brazil,” *Epidemiol. Infect.*, vol. 149, no. e60, p. e60, Feb. 2021.

[17] A. C. Hernandez-Romieu, S. Leung, A. Mbanya, B. R. Jackson, J. R. Cope, D. Bushman, M. Dixon, J. Brown, T. McLeod, S. Saydah, D. Datta, K. Koplan, and F. Lobelo, “Health care utilization and clinical characteristics of nonhospitalized adults in an integrated health care system 28-180 days after COVID-19 diagnosis - georgia, may 2020-march 2021,” *MMWR Morb. Mortal. Wkly. Rep.*, vol. 70, no. 17, pp. 644–650, Apr. 2021.

[18] M. T. Ribeiro, “A COVID-19 na população portuguesa : uma análise de riscos competitivos,” Ph.D. dissertation, Faculdade de Ciências da Universidade de Lisboa, 2022. [Online]. Available: http://hdl.handle.net/10451/53755

[19] C. E. Barrett, A. K. Koyama, P. Alvarez, W. Chow, E. A. Lundeen, C. G. Perrine, M. E. Pavkov, D. B. Rolka, J. L. Wiltz, L. Bull-Otterson, S. Gray, T. K. Boehmer, A. V. Gundlapalli, D. A. Siegel, L. Kompaniyets, A. B. Goodman, B. E. Mahon, R. V. Tauxe, K. Remley, and S. Saydah, “Risk for newly diagnosed diabetes >30 days after SARS-CoV-2 infection among persons aged <18 years - united states, march 1, 2020-june 28, 2021,” *MMWR Morb. Mortal. Wkly. Rep.*, vol. 71, no. 2, pp. 59–65, Jan. 2022.

[20] B. K. J. Tan, R. Han, J. J. Zhao, N. K. W. Tan, E. S. H. Quah, C. J.-W. Tan, Y. H. Chan, N. W. Y. Teo, T. C. Charn, A. See, S. Xu, N. Chapurin, R. K. Chandra, N. Chowdhury, R. Butowt, C. S. von Bartheld, B. N. Kumar, C. Hopkins, and S. T. Toh, “Prognosis and persistence of smell and taste dysfunction in patients with covid-19: meta-analysis with parametric cure modelling of recovery curves,” *BMJ*, vol. 378, p. e069503, Jul. 2022.

[21] J. H. Han, K. N. Womack, M. W. Tenforde, D. C. Files, K. W. Gibbs, N. I. Shapiro, M. E. Prekker, H. L. Erickson, J. S. Steingrub, N. Qadir, A. Khan, C. L. Hough, N. J. Johnson, E. W. Ely, T. W. Rice, J. D. Casey, C. J. Lindsell, M. N. Gong, V. Srinivasan, N. M. Lewis, M. M. Patel, W. H. Self, and Influenza and Other Viruses in the Acutely Ill (IVY) Network, "Associations between persistent symptoms after mild COVID-19 and long-term health status, quality of life, and psychological distress," *Influenza Other Respi. Viruses*, vol. 16, no. 4, pp. 680–689, Jul. 2022.

[22] E. Gentilotti, A. Górska, A. Tami, R. Gusinow, M. Mirandola, J. Rodríguez Baño, Z. R. Palacios Baena, E. Rossi, J. Hasenauer, I. Lopes-Rafegas, E. Righi, N. Caroccia, S. Cataudella, Z. Pasquini, T. Osmo, L. Del Piccolo, A. Savoldi, S. Kumar-Singh, F. Mazzaferri, M. G. Caponcello, G. de Boer, G. L. Hara, ORCHESTRA Study Group, P. De Nardo, S. Malhotra, L. M. Canziani, J. Ghosn, A.-M. Florence, N. Lafhej, B. T. F. van der Gun, M. Giannella, C. Laouénan, and E. Tacconelli, "Clinical phenotypes and quality of life to define post-COVID-19 syndrome: a cluster analysis of the multinational, prospective ORCHESTRA cohort," *EClinicalMedicine*, vol. 62, p. 102107, Aug. 2023.

[23] J. S. Rogers-Brown, V. Wanga, C. Okoro, D. Brozowsky, A. Evans, D. Hopwood, J. R. Cope, B. R. Jackson, D. Bushman, A. C. Hernandez-Romieu, R. A. Bonacci, T. McLeod, J. R. Chevinsky, A. B. Goodman, M. G. Dixon, C. Lutfy, J. Rushmore, E. Koumans, S. B. Morris, and W. Thompson, "Outcomes among patients referred to outpatient rehabilitation clinics after COVID-19 diagnosis - united states, january 2020-march 2021," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 70, no. 27, pp. 967–971, Jul. 2021.

[24] B. Nguyen and A. Tosti, "Alopecia in patients with COVID-19: A systematic review and meta-analysis," *JAAD Int.*, vol. 7, pp. 67–77, Jun. 2022.

[25] S. H. Saydah, J. T. Brooks, and B. R. Jackson, "Surveillance for post-COVID conditions is necessary: Addressing the challenges with multiple approaches," *J. Gen. Intern. Med.*, vol. 37, no. 7, pp. 1786–1788, May 2022.

[26] K. Stavem, W. Ghanima, M. K. Olsen, H. M. Gilboe, and G. Einvik, "Persistent symptoms 1.5-6 months after COVID-19 in non-hospitalised subjects: a population-based cohort study," *Thorax*, vol. 76, no. 4, pp. 405–407, Apr. 2021.

[27] J. Ghosn, L. Piroth, O. Epaulard, P. Le Turnier, F. Mentré, D. Bachelet, C. Laouénan, and French COVID cohort study and investigators groups, "Persistent COVID-19 symptoms are highly prevalent 6 months after hospitalization: results from a large prospective cohort," *Clin. Microbiol. Infect.*, vol. 27, no. 7, pp. 1041.e1–1041.e4, Jul. 2021.

[28] M. Taquet, Q. Dercon, S. Luciano, J. R. Geddes, M. Husain, and P. J. Harrison, "Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19," *PLoS Med.*, vol. 18, no. 9, Sep. 2021.

[29] M. Lipoldová and P. Demant, "Gene-specific sex effects on susceptibility to infectious diseases," *Front. Immunol.*, vol. 12, p. 712688, Oct. 2021.

[30] C. Gebhard, V. Regitz-Zagrosek, H. K. Neuhauser, R. Morgan, and S. L. Klein, "Impact of sex and gender on COVID-19 outcomes in europe," *Biol. Sex Differ.*, vol. 11, no. 1, p. 29, May 2020.

[31] J. P. Klein and M. L. Moeschberger, *Survival analysis*, 2nd ed., ser. Statistics for Biology and Health. New York, NY: Springer, Mar. 2005.

[32] P. Hougaard, *Analysis of multivariate survival data*, 2000th ed., ser. Statistics for Biology and Health. Springer, Dec. 2012.

[33] M. Tableman and J. S. Kim, *Survival analysis using S*, ser. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, Jul. 2003.

[34] S. Selvin, *Practical guides to biostatistics and epidemiology: Survival analysis for epidemiologic and medical research*. Cambridge University Press, Feb. 2010.

[35] M. S. Carvalho, V. L. Andreozzi, C. T. Codeço, D. P. Campos, M. T. S. Barbosa, and S. E. Shimakura, *Análise de sobrevivência: teoria e aplicações em saúde*. Editora FIOCRUZ, 2011.

[36] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Am. Stat. Assoc.*, vol. 53, no. 282, p. 457, Jun. 1958.

[37] R. Beran, "Nonparametric regression with randomly censored survival data," Jan. 1981.

[38] D. R. Cox, “Regression models and life-tables,” *J. R. Stat. Soc.*, vol. 34, no. 2, pp. 187–202, Jan. 1972.

[39] D. Schoenfeld, “Partial residuals for the proportional hazards regression model,” *Biometrika*, vol. 69, no. 1, pp. 239–241, 1982.

[40] A. Yakovlev, A. D. Tsodikov, and B. Asselain, “Stochastic models of tumor latency and their biostatistical applications,” 1996.

[41] A. D. Tsodikov, “A proportional hazards model taking account of long-term survivors.” *Biometrics*, vol. 54 4, pp. 1508–16, 1998.

[42] J. W. Boag, “Maximum likelihood estimates of the proportion of patients cured by cancer therapy,” *J. R. Stat. Soc.*, vol. 11, no. 1, pp. 15–44, Jan. 1949.

[43] V. T. Farewell, “The use of mixture models for the analysis of survival data with long-term survivors,” *Biometrics*, vol. 38, no. 4, pp. 1041–1046, Dec. 1982.

[44] R. Saikia and M. P. Barman, “A review on accelerated failure time models,” 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:207824546

[45] Y. Peng, K. B. Dear, and J. W. Denham, “A generalized F mixture model for cure rate estimation,” *Stat. Med.*, vol. 17, no. 8, pp. 813–830, Apr. 1998.

[46] M. Amico and I. Van Keilegom, “Cure models in survival analysis,” *Annu. Rev. Stat. Appl.*, vol. 5, no. 1, pp. 311–342, Mar. 2018.

[47] C. Cai, Y. Zou, Y. Peng, and J. Zhang, “smcure: an r-package for estimating semiparametric mixture cure models,” *Comput. Methods Programs Biomed.*, vol. 108, no. 3, pp. 1255–1260, Dec. 2012.

[48] L. Wang, P. Du, and H. Liang, “Two-component mixture cure rate model with spline estimated nonparametric components,” *Biometrics*, vol. 68, no. 3, pp. 726–735, Sep. 2012.

[49] M. Amico, I. Van Keilegom, and C. Legrand, “The single-index/cox mixture cure model,” *Biometrics*, vol. 75, no. 2, pp. 452–462, Jun. 2019.

[50] A. López-Cheda, M. A. Jácome, and R. Cao, "Nonparametric latency estimation for mixture cure models," *Test (Madr.)*, vol. 26, no. 2, pp. 353–376, Jun. 2017.

[51] A. López-Cheda, R. Cao, M. A. Jácome, and I. Van Keilegom, "Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models," *Comput. Stat. Data Anal.*, vol. 105, pp. 144–165, Jan. 2017.

[52] W. C. Safari, I. López-de Ullibarri, and M. A. Jácome, "Nonparametric inference for mixture cure model when cure information is partially available," *Eng. Proc.*, vol. 7, no. 1, p. 17, Oct. 2021.

[53] U. U. Müller and I. Van Keilegom, "Goodness-of-fit tests for the cure rate in a mixture cure model," *Biometrika*, vol. 106, no. 1, pp. 211–227, Dec. 2018.

[54] J. Xu and Y. Peng, "Nonparametric cure rate estimation with covariates," *Can. J. Stat.*, vol. 42, no. 1, pp. 1–17, Mar. 2014.

[55] R. A. Maller and X. Zhou, "Survival analysis with long-term survivors," 1996.

[56] R. A. Maller and S. Zhou, "Testing for sufficient follow-up and outliers in survival data," *J. Am. Stat. Assoc.*, vol. 89, no. 428, p. 1499, Dec. 1994.

[57] ——, "Testing for the presence of immune or cured individuals in censored survival data," *Biometrics*, vol. 51, no. 4, pp. 1197–1205, Dec. 1995.

[58] M. A. Nicolaie, J. M. G. Taylor, and C. Legrand, "Vertical modeling: analysis of competing risks data with a cure fraction," *Lifetime Data Anal.*, vol. 25, no. 1, pp. 1–25, Jan. 2019.

[59] P. W. Bernhardt, "A flexible cure rate model with dependent censoring and a known cure threshold," *Stat. Med.*, vol. 35, no. 25, pp. 4607–4623, Nov. 2016.

[60] ISPUP, "Ispup website. from sars-cov-2 infection to covid-19. study of viral kinetics and immune response to understand contagion and clinical evolution." [Online]. Available: https://ispup.up.pt/en/

[61] B. Gyawali and C. M. Booth, "Cancer treatments should benefit patients: a common-sense revolution in oncology," *Nat. Med.*, vol. 28, no. 4, pp. 617–620, Apr. 2022.

[62] M. Augustin, P. Schommers, M. Stecher, F. Dewald, L. Gieselmann, H. Gruell, C. Horn, K. Vanshylla, V. D. Cristanziano, L. Osebold, M. Roventa, T. Riaz, N. Tschernoster, J. Altmueller, L. Rose, S. Salomon, V. Priesner, J. C. Luers, C. Albus, S. Rosenkranz, B. Gathof, G. Fätkenheuer, M. Hallek, F. Klein, I. Suárez, and C. Lehmann, "Post-COVID syndrome in non-hospitalised patients with COVID-19: a longitudinal prospective cohort study," *Lancet Reg. Health Eur.*, vol. 6, Jul. 2021.

[63] C. H. Sudre, B. Murray, T. Varsavsky, M. S. Graham, R. S. Penfold, R. C. Bowyer, J. C. Pujol, K. Klaser, M. Antonelli, L. S. Canas, E. Molteni, M. Modat, M. Jorge Cardoso, A. May, S. Ganesh, R. Davies, L. H. Nguyen, D. A. Drew, C. M. Astley, A. D. Joshi, J. Merino, N. Tsereteli, T. Fall, M. F. Gomez, E. L. Duncan, C. Menni, F. M. K. Williams, P. W. Franks, A. T. Chan, J. Wolf, S. Ourselin, T. Spector, and C. J. Steves, "Attributes and predictors of long COVID," *Nat. Med.*, vol. 27, no. 4, pp. 626–631, Apr. 2021.

[64] J. Ghosn, L. Piroth, O. Epaulard, P. Le Turnier, F. Mentré, D. Bachelet, C. Laouénan, and French COVID cohort study and investigators groups, "Persistent COVID-19 symptoms are highly prevalent 6 months after hospitalization: results from a large prospective cohort," *Clin. Microbiol. Infect.*, vol. 27, no. 7, pp. 1041.e1–1041.e4, Jul. 2021.

[65] H. E. Davis, G. S. Assaf, L. McCorkell, H. Wei, R. J. Low, Y. Re'em, S. Redfield, J. P. Austin, and A. Akrami, "Characterizing long COVID in an international cohort: 7 months of symptoms and their impact," *EClinicalMedicine*, vol. 38, Aug. 2021.

[66] X. Fang, C. Ming, Y. Cen, H. Lin, K. Zhan, S. Yang, L. Li, G. Cao, Q. Li, and X. Ma, "Post-sequelae one year after hospital discharge among older COVID-19 patients: A multi-center prospective cohort study," *J. Infect.*, vol. 84, no. 2, pp. 179–186, Feb. 2022.

[67] S. Yende and C. R. Parikh, "Long COVID and kidney disease," *Nat. Rev. Nephrol.*, vol. 17, no. 12, pp. 792–793, Dec. 2021.

[68] A. V. Raveendran, R. Jayadevan, and S. Sashidharan, "Long COVID: An overview," *Diabetes Metab. Syndr.*, vol. 15, no. 3, pp. 869–875, May 2021.

[69] G. Salinas-Escudero, M. F. Carrillo-Vega, V. Granados-García, S. Martínez-Valverde, F. Toledano-Toledano, and J. Garduño-Espinosa, “A survival analysis of COVID-19 in the mexican population,” *BMC Public Health*, vol. 20, no. 1, p. 1616, Oct. 2020.

Apêndice A

Código R comentado

A.1 Manipulação da base de dados

```
#Define o diretorio
setwd("C:/colocar_nome_do_caminho")

#Selecao de variaveis/colunas da base de dados
raw<-read.csv("colocar_nome_da_base.csv")
var=c(1,2,7,8,...)

#Criacao do Data Frame
df<-subset(raw, select = var)

#Exclusao de assintomaticos
df<-df[df$G3Q00039==1&!is.na(df$G3Q00039),]

#Construcao de variaveis####
#Idade
calc_Idade<-function(d,n){return(pmax(floor((d-n)/365.25),0))}
df["Idade"]<-mapply(calc_Idade,df$dtaDiag.dtDiag,  df$G1Q00002)

calc_Idade_AGG<-function(age){if(age<18) return("[0,18[")
```

```
  else if(age<30) return("[18,30[")
  else if(age<40) return("[30,40[")
  else if(age<50) return("[40,50[")
  else if(age<60) return("[50,60[")
  else if(age<70) return("[60,70[")
  else if(age<80) return("[70,80[")
  else return("[80,+inf[")
}

df["Idade_AGG"]<-sapply(df$Idade,FUN=calc_Idade_AGG)
#Utilizei mapply e sapply para as outras funcoes
#Exclui os censurados a esquerda e os menores de idade
```

O código $A.1$ apresenta os comandos básicos de manipulação da base de dados, incluindo a construção de variáveis não existentes na base inicial (e.g., Idade, variável contínua), além de eventuais agrupamento de categorias (e.g., Idade_AGG, variável categorizada nas faixas de idade).

A.2 Gráficos, testes de log-rank e de Peto-Peto

```
library(survival)

surv<-Surv(df_event$Time,df_event$Status)
#Log-rank test, rho=0 (default)
a<-survdiff(surv~factor(Renda),data=df_event)
round(1-pchisq(a$chisq,1),3) #apenas para aumentar a precisao

#Peto-Peto test, rho=1
survdiff(surv~factor(Escolaridade_AGG),rho=1,data=df_event)
```

```
#Graficos
km0<-survfit(surv~1,data=df_event) #KM para o tempo
plot(km0, main = "Título",xlab = "Tempo", ylab = "Sobrevivência")

km1<-survfit(surv~factor(Idade_AGG),data=df_event)
plot(km1, main = "Título",col=c("green","blue","purple", "orange",
"red", "brown","black"), xlab = "Tempo (dias)", ylab = "Sobrevivência")

legend(260, 1.01, legend=c("[18,30[","[30,40[","[40,50[","[50,60[",
"[60,70[","[70,80[","[80,+inf["), fill = c("green","blue","purple",
"orange","red","brown","black"))
```

Para a aplicação dos testes de log-rank e Peto-Peto, basta aplicar a função *survdiff* e alterar o valor do *rho* para cada variável que será testada. A função *Surv* serve para indicarmos quais obverservações são censuradas ou se experenciaram o evento de interesse.

A.3 Modelos de regressão paramétrica

```
#Escolhendo categorias de referencia das variaveis que entraram nos modelos
df_event["Sexo"]<-factor(df_event$G1Q00001, levels=c(0,1), labels=c(0,1))

#Weibull
fit_weibull<-survreg(surv~Sexo+Comorbidades+
                       Percecao_da_renda+Hospitalizacao,
                       data=df_event,dist="weibull")
summary(fit_weibull)
#dist="lognormal" (e "loglogistic")

#Para interpretar em termos de AFT
```

```
exp(coef(fit_weibull))
```

A função *survreg* constrói um modelo de regressão paramétrica com as variáveis indicadas. Para a escolha da distribuição adotada, alteramos o parâmetro *dist*. A função *summary* serve para vermos os coeficientes do modelo e os p valores dos testes.

O modelo com distribuição Weibull pode ser interpretado em termos de riscos proporcionais (utilizando a fórmula de 2.34) ou em termos de AFT (exp(coeficientes)). Contudo, para as distribuições lognormal e loglogística, temos apenas a interpretação em termos de AFT.

A.4 Modelos de Cox, análises de resíduos

```
fit_cox <- coxph(surv ~ Sexo + Comorbidades + Percecao_da_renda +
Hospitalizacao,  data = df_event)
summary(fit_cox)

#Analises de diagnostico e de residuos
ph.test<-cox.zph(fit_cox) #H0: Os riscos sao proporcionais

#Residuos de Cox-Snell
model <- fit_cox
rc<-abs(df_event$Status-model$residuals)
km.rc<-survfit(Surv(rc,df_event$Status)~1)
summary.km.rc<-summary(km.rc)
rcu<-summary.km.rc$time
surv.rc<-summary.km.rc$surv

plot(rcu,-log(surv.rc),type="p",pch="*",xlab="Cox.Snell residual",
ylab="Cumulative hazard");abline(a=0,b=1)

#Residuos de Schoenfeld
```

```
detail<-coxph.detail(model)
time<-detail$y[,2]
status<-detail$y[,3]
sch<-resid(model,type="schoenfeld")
plot(time[status==1],sch[,1],xlab="Tempo de sobrevivência ordenado",
     ylab="Resíduos de Schoenfeld para o sexo")
```

O modelo de regressão de riscos proporcionais de Cox foi construído utilizando a função *coxph*. A função *cox.zph* foi utilizada para a aplicação do teste de proporcionalidade dos riscos, sendo a hipótese nula: os riscos são proporcionais verificada em cada variável e globalmente.

Nesta etapa também foram construídos os resíduos de Cox-Snell e os de Schoenfeld.

A.5 Modelos de cura de mistura

```
#Reduzir o numero de variaveis
df_cure<-df_event[,c("Time","Status","Sexo", "Comorbidades",
                     "Percecao_da_renda","Hospitalizacao")]

#Para deixar variavel numerica
calc_Sex2<-function(t){ return(as.numeric(t)-1)}

df_cure["Sexo_num"]<-sapply(df_cure$Sex,FUN=calc_Sex2)

calc_Ill2<-function(t){return(as.numeric(t)-1)}

df_cure["Comorbidades_num"]<-sapply(df_cure$Comorbidades,FUN=calc_Ill2)

#Parametric AFT Cure Models ####
library(gfcure)
```

```
mcm.ph.weibull <- gfcure(Surv(df_cure$Time,df_cure$Status) ~ Sexo, ~ Sexo,
                         data = df_cure, dist = "weibull")
```

Para a preparação dos modelos de cura, precisamos criar o data frame reduzido *df_cure* porque se a base de dados possuir muitas variáveis, as funções não serão executadas. Além disso, precisamos transformar as variáveis categóricas para numéricas.

O modelo AFT não foi utilizado nesta dissertação. Entretando, apresentamos o código para a apreciação. Detalhes sobre o *gfcure* pode ser consultados em [45].

```
#Parametric PH Cure Models ####
library(smcure)
sm.ph <- smcure(Surv(Time, Status) ~ Sexo_num + Hospitalizacao,
cureform = ~ Sexo_num + Comorbidades_num, data=df_cure, model="ph", Var=TRUE)
```

O modelo de cura utilizado nesta dissertação foi o semi-paramétrico dos riscos proporcionais através da função *smcure* da biblioteca de mesmo nome. O objeto *Surv* também é utilizado como nos modelos paramétricos e semi-paramétricos do Capítulo 4. As primeiras variáveis são para a latência e as de *cureform* são para a incidência.

```
# Construindo duas curvas
newinc <- cbind(c(0,1),c(1,0)) #(Sexo),(Comorbidades) =>
(feminino + sim, masculino + nao) - incidencia
newlat <- cbind(c(0,1),c(1,0)) #(Sexo),(Hosp) =>
(feminino + sim, masculino + nao) - latencia

pred.sm.ph <- predictsmcure(sm.ph, newX=, newlat, newZ = newinc, model = "ph")

plotpredictsmcure(pred.sm.ph,model="ph", ylab = "Sobrevivência",
xlab = "Tempo (dias")
title(main="Ajuste do modelo MCM")
```

Como as variáveis para a incidência e as para a latência podem ser diferentes, no momento da previsão, utilizamos os objetos *newinc*, que representa as variáveis da incidência (sexo e comorbidades), e *newlat*, que representa as variáveis da latência (sexo e hospitalização). Em cada curva será composta pelas características assumidas em cada variável. Por exemplo, suponha que queremos criar 3 curvas: 1) mulheres com comorbidades = cbind(c(0),c(1)); 2) mulheres sem comorbidades = cbind(c(0),c(0)); 3) homens sem comorbidades = cbind(c(1),c(0)). Juntando essas 3 curvas no objeto *newinc*, teremos o vetor cbind(c(0,0,1),c(1,0,0)).

As funções *predictsmcure* e *plotpredictsmcure* pertencem à biblioteca *smcure* e servem para fazer as previsões e construir o gráfico de sobrevivência com as curvas escolhidas. Contudo, não é possível alterar a escala do gráfico, razão pela qual modificamos a função *plotpredictsmcure*, conforme se verifica no próximo trecho do script.

```
plotpredictsmcure <-
  function(object, type="S", xlab="Time",ylab="Predicted Survival Probability",
  model=c("ph","aft"), ylim=c(0,1),...)
  {
    pred <- object$prediction
    if(model=="ph"){
      pdsort <- pred[order(pred[,"Time"]),]
      if(length(object$newuncureprob)==1) plot(pdsort[,"Time"],pdsort[,1],
      type="S")
      else
        matplot(pdsort[,"Time"],pdsort[,1:(ncol(pred)-1)],col=1,type="S",
        lty=1:(ncol(pred)-1),xlab=xlab,ylab=ylab,ylim=ylim)
    }
    if(model=="aft"){
      nplot=ncol(pred)/2
      pdsort <- pred[order(pred[,1+nplot]),c(1,1+nplot)]
      plot(pdsort[,2],pdsort[,1],xlab=xlab,ylab=ylab,col=1,type="S",ylim=ylim)
      if(nplot>1){
        for(i in 2:nplot){
```

```
          pdsort<- pred[order(pred[,i+nplot]),c(i,i+nplot)]
          lines(pdsort[,2],pdsort[,1],lty=i,type="S")
        }
      }
    }
  }
```

Considerando que a função *plotpredictsmcure* utiliza a função *matplot*, basta utilizar seu parâmetro *ylim*. Embora tenhamos escolhido o modelo "ph"da função, alteramos também a parte "aft"da função..

```
# Parcela de imunes
cure.ph<-1-pred.sm.ph$newuncureprob

# Testar a adequabilidade de parcela de curados
library(npcure)
testmz(Time, Status, df_cure)

#Para selecionar as variaveis automaticamente
library(penPHcure)

set.seed(123) # caso queira reproduzir os resultados
df_cure$start <- 0

# Modelo de cura PH padrao.
# O argumento which.X = "mean" serve para considerar o peso
# do tempo no historico.
# Para conduzir a inferencia por bootstrapping (default tem 100 reamostragens),
#usamos o argumento inference = TRUE,

fit <- penPHcure(Surv(start, Time, Status) ~ Hospitalizacao +
```

```
Sexo_num + Comorbidades_num,
        cureform = ~  Hospitalizacao + Sexo_num + Comorbidades_num,
        data = df_cure, which.X = "mean", inference = TRUE, nboot =  50)
summary(fit) # Foram escolhidas as mesmas variaveis que o feito manualmente
```

A primeira parte do código é para a determinação da percentagem de "imunes", ou seja, de indivíduos que não terão a resolução de todos os sintomas, que é o complementar da incidência, por isso a fórmula $1 - p(z)$.

O teste de Maller-Zhou é *testmz* da biblioteca *npcure*, o qual contém modelos não paramétricos.

Por fim, a função *penPHcure* da biblioteca de mesmo nome serve para a criação do modelo PH, com a vantagem de selecionar automaticamente as variáveis da incidência e da latência. O modelo escolhido por esta ferramente coincidiu com o que fizemos manualmente pela função *smcure*.