Análisis de supervivencia y modelos cure. Aplicación a la COVID-19

Carlos Plou Izquierdo

Trabajo fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: Gerardo Sanz Sáiz
25 de junio de 2021

Prólogo

El trabajo fin de grado con título “Análisis de supervivencia y modelos cure. Aplicación a la COVID-19” que se presenta a continuación, consta de dos partes claramente diferenciadas: una primera que recoge una introducción al análisis de supervivencia, profundizando en los modelos cure, y una segunda en la cual se explica detalladamente el estudio realizado acerca de la hospitalización e ingreso en UCI de los pacientes covid. Animo a realizar una previa y detenida lectura de los dos primeros capítulos para comprender el procedimiento seguido y los resultados obtenidos en el estudio.

Antes de dar paso al trabajo, me gustaría agradecer a Gerardo Sanz Sáiz su implicación, comprensión e ilusión con la que, en todo momento, me ha ayudado a dar mis primeros pasos en el mundo de la investigación. También me gustaría reconocer y agradecer a Miguel Lafuente Blasco tanto el trabajo realizado como los consejos dados.

Por último, mencionar a todas las personas que han formado parte de mi día a día durante estos años y agradecerles todo el cariño y apoyo recibido. Vosotros sois los verdaderos artífices de mis logros.

Disfruten de su lectura, Carlos Plou Izquierdo.

Summary

Survival analysis is the branch of Statistics that deals with the variable "time to occurrence of a given event". Such an event is typically addressed, albeit not exclusively, to health research. However, it may belong to different knowledge areas such as engineering, economy or demography. We will focus our attention on usual survival models and in one that has received a great impulse in recent years: the cure model. This model is characterised by the fact that not every individual considered in the study must experience the event. In this project, we will therefore apply all these concepts to a COVID-19 dataset.

The coronavirus disease-2019 (COVID-19) was declared as a pandemic on 11 March 2020. Since then, multiple studies have been carried out to identify and explore the key features of the disease - for example, the basic reproduction number, the fatality rate or even potential factors and treatments. In particular, we have put the focus on covid patients. By using a COVID-19 database, we have performed this study with the aim of analysing different parameters: the probability of being admitted to the intensive care unit (ICU) upon hospitalisation, time from hospital admission to ICU admission and the ICU length-of-stay according to the information of each individual.

In this way, we begin describing basic concepts of survival analysis together with the usual nonparametric, semiparametric and parametric models (Chapter 1). Secondly, cure models will be explained in detail (Chapter 2) to further implement them for covid patients hospitalisation (Chapter 3).

Survival analysis

In survival analysis, the time-to-event variable is called "lifetime" or "survival time", and the event usually receives the name of "failure" or "death". Besides, the *survival function* at time t, $S(t)$, is defined as the probability of an individual living beyond that time, that is

$$S(t) = P(T > t),$$

where T denotes the lifetime. The main goal of the survival analysis is to estimate this function for any value of t. To get its objective, the survival analysis must address the censoring, which is a set of limitations that may appear in the available data and thus influencing the study.

Another function that characterises the distribution of survival time is the *hazard function*, $h(t)$, which describes the instantaneous failure rate for an individual. Mathematically,

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

Consequently, the *cumulative hazard function*, $H(t)$, is defined as

$$H(t) = \int_0^t h(x)dx.$$

The basis of survival analysis relies on using inference techniques along with some of the wide variety of models that allow not only the estimates of these three functions (displayed above), but also comparing the survival time between different groups of the population and analysing the influence of certain variables on the survival time.

Assuming that the failure and censoring times $0 < t_1 < t_2 < \cdots < t_N$ are known, the *Kaplan-Meier estimator* is defined as

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right),$$

where n_j denotes the number of individuals at risk at t_j and d_j the number of individuals that fail at this time (t_j). In the same way, the *Nelson-Aalen estimator* is given by

$$\hat{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}.$$

Once estimates are computed, in case we want to compare survival curves from different groups A, B of the population, we will use the *Log-Rank test* which is based on the statistical

$$Q = \frac{(O-E)^2}{V},$$

where O is the number of group A individuals that experience the event, E its expected number and V its variance. Under the null hypothesis (equality of both survival functions) the statistical Q follows a χ_1^2 distribution.

Furthermore, it should be noted that if we consider the likely influence of other variables (we will refer to them as covariates and will be denoted as a vector $\mathbf{z}$), then the Cox model is the most accurate. This model is also called *proportional hazards model* and defines the hazard function as

$$h(t, \mathbf{z}, \boldsymbol{\beta}) = h_0(t) e^{\mathbf{z}'\boldsymbol{\beta}},$$

$h_0(t)$ receives the name of baseline hazard function and $\boldsymbol{\beta}$ is a coefficient vector associated with the covariates. Cox proposed a partial likelihood function so as to estimate $\boldsymbol{\beta}$, this estimator will allow us to estimate the Cox survival function.

With a view to evaluate the influence of the covariates, we will compute tests such as *Wald test*, *Score test* or *Likelihood ratio test*. All of them employ similar statistics to evaluate the hypothesis test

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

where $\boldsymbol{\beta}_0$ is specified.

Finally, we consider a class of models called *accelerated failure time models*, in which lifetime distribution is completely specified

$$T = e^{\beta_0 + \mathbf{z}'\boldsymbol{\beta}} \varepsilon,$$

being β_0 an intercept and ε an error component. Depending on ε distribution, we are able to distinguish and discuss two specific examples: the exponential and the Weibull regression models.

Cure model

"Classical" survival analysis assumes that, despite censoring, all subjects under study will experience the failure. Nevertheless, in some contexts, a fraction of the population will never experience it. Cure models handle this situation, and thus becoming an essential branch of the survival analysis. This model separates the population between two different groups:

- *"Cured"* or *non-susceptible*: they will never experience the event, that is, $T = \infty$. Denoted by $B = 0$.
- *"Uncured"* or *susceptible*: they will experience the failure, that is, $T < \infty$. Denoted by $B = 1$.

and it is said that the data contains a *cure fraction* $1 - p$ defined as

$$\lim_{t \to \infty} S_{pop}(t) = 1 - p,$$

being $S_{pop}(t)$ the population survival function. The key challenge of cure models is to distinguish between cured individuals and censored observations which are susceptible.

Cure models are classified into two main groups: *mixture cure models* and *promotion time cure models*.

Promotion time cure models are designed for the analysis of relapse in cancer studies. Its survival function is given by

$$S_{pop}(t) = \exp\left[-\theta F(t)\right],$$

taking $\theta > 0$ and $F(t)$ a proper distribution function.

Mixture cure models are a kind of two-part models whose population survival function is defined as

$$S_{pop}\left(t|\mathbf{x},\mathbf{z}\right) = 1 - p(\mathbf{x}) + p(\mathbf{x}) S_u\left(t|\mathbf{z}\right),$$

where $\mathbf{x}, \mathbf{z}$ are two covariate vectors, $p(\mathbf{x}) = P\left(B = 1|\mathbf{x}\right)$ is called *incidence* and $S_u\left(t|\mathbf{z}\right) = P\left(T > t|\mathbf{z}, B = 1\right)$ takes the name of *latency*. Depending on the assumptions set for the latency and the incidence, there is a wide variety of models based on estimators like Kaplan-Meier or Cox survival function. Since the model contains unobserved information, the *Expectation-Maximisation* (EM) algorithm is considered to maximise the likelihood function and, hence, to estimate the coefficient vectors.

Research

In the last chapter, we apply all the previous concepts to analyse the hospitalisation of covid patients using a dataset. The study consists of two parts: one related to ICU length-of-stay and another one associated with hospital length-of-stay and ICU admission.

On the one hand, in the first part of the study we define the variable of interest as "time from ICU admission to ICU discharge". Since this is a time-to-event variable, survival methods could be applied. Specifically, Kaplan-Meier estimator and LogRank test will be used to illustrate and to prove the influence of covariates such as age or gender. In addition, a Cox model will be suggested upon considering some of the significant covariates.

On the other hand, we consider "time from hospital admission to ICU admission" as the variable of interest through the second section of the study. It is obvious that a large fraction of individuals in dataset will not have been admitted to ICU at the final date of the study. Part of them might be censored observations meaning that admission would occur after the follow-up period. However, most of them will not require ICU stay, being those cured observations. This fact therefore validates the application of the cure model. Particularly, we will apply mixture cure model and then attempt to estimate the incidence and latency.

Índice general

Capítulo 1

Análisis de Supervivencia

En este capítulo explicaremos los pilares en los que se fundamenta el análisis de supervivencia clásico. De este modo, en la sección 1.1 definiremos algunos conceptos básicos así como notaciones que se emplearán a lo largo del trabajo. Seguidamente, en la sección 1.2 presentaremos dos de los estimadores no paramétricos más comunes: Kaplan-Meier y Nelson-Aalen. A continuación, explicaremos el modelo de riesgos proporcionales de Cox en la sección 1.3 para finalizar comentando algún modelo paramétrico en la sección 1.4.

1.1. Conceptos básicos

El **análisis de supervivencia** es la rama de la estadística que estudia el tiempo (años, meses, edad...) que transcurre desde un instante considerado como punto de partida hasta el acontecimiento de un determinado evento (enfermedad, recuperación, muerte...) para cada individuo de una población.

La hipótesis fundamental del análisis de supervivencia es que todos los individuos bajo estudio experimentan el evento de interés. Cabe reseñar que, por norma general, nos centraremos en el estudio de un único evento. Algunos ejemplos pueden ser:

1. Años transcurridos desde que un paciente recibe un trasplante de corazón hasta su muerte. El suceso a estudiar es la muerte.

2. Años de supervivencia para un paciente VIH positivo. El suceso también es el fallecimiento del paciente.

3. Días de ocupación de una cama de hospital por parte de un paciente COVID. El suceso a estudiar es el hecho de dejar una cama libre sin atender el motivo del mismo.

Denotaremos por **T** la variable aleatoria no negativa que describe el tiempo de supervivencia (o de fallo) para un individuo. Siempre que no indiquemos lo contrario, asumiremos que es absolutamente continua. Del mismo modo, usaremos **t** para denotar valores concretos de la anterior variable aleatoria.

Definición 1. Sea T variable aleatoria no negativa con función de distribución $F(t)$ y función de densidad $f(t)$ se define la **función de supervivencia *S(t)*** como

$$S(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f(t).$$

Nota. Esta función representa la probabilidad de que un individuo tenga un tiempo de supervivencia mayor que un valor concreto t. Además, la función es no creciente y verifica que $S(0) = 1$, $S(\infty) = 0$. Luego, acorde con la hipótesis fundamental, todos los individuos experimentarán el suceso de interés.

El principal problema al que se enfrentan este tipo de estudios es la **censura**. La censura tiene lugar cuando tenemos información parcial sobre un individuo. Los principales motivos de la existencia de datos censurados son:

- Fin del estudio: el individuo no ha experimentado el evento en el momento en el que el estudio finaliza. En el ejemplo 3, un enfermo puede estar ingresado en el hospital en el momento que se deja de recoger datos.
- Pérdida de seguimiento o salida del estudio del individuo debido al acontecimiento de un suceso claramente ajeno al de interés. Por ejemplo, muerte por accidente en el ejemplo 1.
- La primera observación del individuo es posterior al verdadero inicio del tiempo de supervivencia. En el ejemplo 2, un individuo entra al estudio tras dar positivo en un test sin conocer cuando se infectó.

Los dos primeros motivos, que son los más comunes y los que consideraremos en este trabajo, reciben el nombre de censura a derecha. En la figura 1.1 podemos ver un caso concreto del ejemplo 1 en un estudio de 40 años de duración. Los pacientes 1 y 5 han fallecido durante el estudio. El resto de pacientes han sido censurados a derecha: el 3 podría ser por pérdida de seguimiento, mientras que los individuos 2 y 4 a causa del fin del estudio.

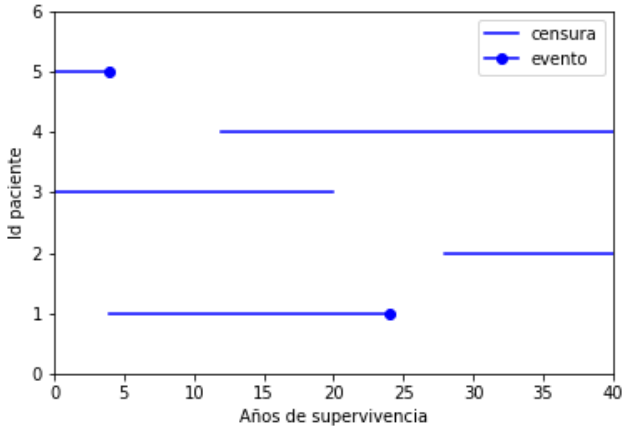


Figura 1.1: Censura en el ejemplo 1.

Asumiremos, entonces, que la censura (a derecha) es una variable aleatoria y la denotaremos por **C** (**c** para valores concretos). De este modo, el tiempo de observación para cada individuo será

$$Y = \text{mín}(T, C), \tag{1.1}$$

denotando por y valores concretos de dicha variable aleatoria. Utilizaremos la variable δ para denotar si un individuo ha experimentado el suceso o ha sido censurado

$$\delta = I(T \leq C),$$

donde $I(\cdot)$ representa la función indicadora. La principal hipótesis que han de cumplir los modelos básicos con datos censurados a derecha es que la distribución de censura y la de la función de supervivencia sean independientes.

Definición 2. En las condiciones de la *Definición* 1, se define la **función de riesgo *h(t)*** como

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t P(T > t)} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}.$$

Nota. Esta función representa la probabilidad de que un individuo experimente el suceso en el instante t teniendo en cuenta que ha "sobrevivido" hasta ese momento.

Definición 3. En las condiciones de la *Definición* 1, se define la **función de riesgo acumulado *H(t)*** como

$$H(t) = \int_0^t h(x)dx = \int_0^t \frac{f(x)}{S(x)}dx = \int_0^t \frac{F'(x)}{1-F(x)}dx = -\log(1-F(x))\Big|_{x=0}^{x=t} = -\log(1-F(t)) = -\log(S(t)).$$

Nota. La función de supervivencia $S(t)$ se puede expresar como

$$S(t) = e^{-H(t)}. \tag{1.2}$$

Supongamos que disponemos de n observaciones independientes (y_i, δ_i) $i = 1, \ldots, n$. En tal caso, la función verosimilitud del conjunto de datos queda determinada por

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

Los principales objetivos que se buscan al realizar un análisis de supervivencia son:

1. Estimar e interpretar las tres funciones mencionadas.
2. Comparar la supervivencia entre dos o más grupos de la población.
3. Analizar la influencia y significación de diferentes factores en la supervivencia.

Los modelos que nos permiten abordar nuestros objetivos se pueden clasificar en paramétricos, semiparamétricos y no paramétricos. Si se asume que la distribución del tiempo de supervivencia sigue una determinada distribución concreta que depende de una serie de parámetros, el modelo será paramétrico. Por otro lado, el modelo será semiparamétrico si se sabe que una parte de dicha distribución viene determinada por una distribución paramétrica, como en el caso anterior, pero otra parte queda sin precisar. Por último, los modelos no paramétricos serán aquellos en los que no se asume que el tiempo de supervivencia siga una determinada distribución. El caso semiparamétrico aparecerá cuando introduzcamos el modelo de regresión de Cox (sección 1.3), el cual asume que la función de riesgo es el producto de dos componentes; una paramétrica y otra que queda sin precisar.

1.2. Estimadores no paramétricos

Los dos estimadores no paramétricos de la función supervivencia más comunes son el estimador de **Kaplan-Meier** y el de **Nelson-Aalen**.

Su principal premisa es que dividen el tiempo en intervalos de acuerdo a si se observa el suceso de interés o se produce la censura, lo que supone en la práctica una discretización del tiempo de supervivencia T. Denotamos por $0 < t_1 < t_2 < \cdots < t_N$ los instantes (ordenados) en los que el evento o la censura tienen lugar para algún individuo.

Definición 4. Dado $t \geq 0$, se llama **conjunto de riesgo en t** al conjunto de individuos que se encuentran en observación en el instante t^-.

Nota. La expresión t^- nos permite incluir aquellos individuos que experimentan el evento o son censurados justo en el instante t. Denotaremos el conjunto de riesgo en t por $R(t)$.

Asimismo, para cada t_j, identificamos n_j con el cardinal de $R(t_j)$ y d_j con el número de eventos observados en t_j. De este modo, acorde con la definición de la función riesgo $h(t)$, tenemos en nuestro caso (tiempo discretizado) que

$$h(t_j) = P(T = t_j | T \geq t_{j-1}) \quad j = 1, \ldots, N.$$

Pudiendo ser estimada sencillamente por

$$\hat{h}(t_j) = \frac{d_j}{n_j} \quad j = 1, \ldots, N. \tag{1.4}$$

1.2.1. Estimador de Kaplan-Meier

El estimador de Kaplan-Meier es el utilizado, por defecto, en la mayoría de paquetes estadísticos. Se fundamenta en el hecho de que la probabilidad de "sobrevivir" en el instante t_j es igual al producto de haber sobrevivido en el instante t_{j-1} y de no experimentar el evento en el instante t_j teniendo en cuenta que "ha sobrevivido" hasta ese momento. En términos matemáticos,

$$\hat{S}(t_j) = \hat{S}(t_{j-1})(1-\hat{h}(t_j)) = \hat{S}(t_{j-1})\left(1-\frac{d_j}{n_j}\right) \quad j = 1,\ldots,N.$$

Aplicando reiteradamente esta fórmula recursiva, se alcanza el **estimador de Kaplan-Meier**

$$\hat{S}(t) = \prod_{j:t_j \leq t}\left(1-\frac{d_j}{n_j}\right). \tag{1.5}$$

Ejemplo: Se realiza un estudio para comparar la eficacia de dos tratamientos frente a una enfermedad. Para ello, se recoge el tiempo (meses) transcurrido desde que se medica al paciente hasta que fallece. Una muestra de los datos de este estudio sería la tabla 1.1 y la gráfica correspondiente a los estimadores de Kaplan-Meier para ambos tratamientos sería la figura 1.2.

Id	Tiempo (meses)	Tratamiento
1	3	A
2	20	B
3	18	A
4	18	A
5	35	B
6	12	B
7	33	B
8	5	A
9	19	B
10	7	A

Tabla 1.1: Datos de los pacientes.

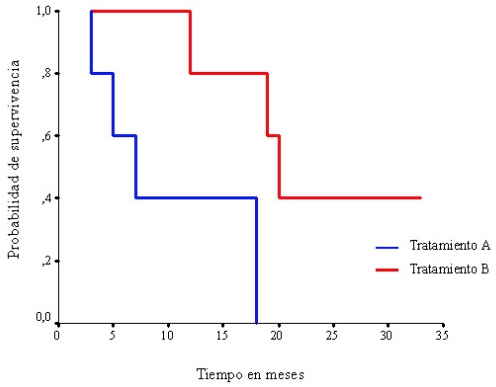


Figura 1.2: Ejemplo del estimador Kaplan-Meier.

En el *Apéndice* A, se calcula un estimador de la varianza de (1.5) por medio del método delta junto con un intervalo de confianza para dicho estimador.

1.2.2. Estimador de Nelson-Aalen

El estimador de Nelson-Aalen aproxima la función de riesgo acumulado *H(t)* a partir de (1.4) para luego estimar la función de supervivencia empleando (1.2). Por consiguiente, definimos el **estimador de Nelson-Aalen** como

$$\hat{H}(t) = \sum_{j:t_j \leq t}\frac{d_j}{n_j}.$$

Se puede ver [1, Sección 3.1] que un estimador de la varianza del estimador de Nelson-Aalen es

$$\mathbb{V}ar\left(\hat{H}(t)\right) = \sum_{j:t_j \leq t}\frac{d_j}{n_j^2},$$

así pues, si denotamos por $\hat{SE}\left(\hat{H}(t)\right)$ a la raíz cuadrada positiva de dicha varianza y $z_{1-\frac{\alpha}{2}}$ al percentil $1-\frac{\alpha}{2}$ de la variable aleatoria normal estándar, concluimos que el intervalo de confianza de significación $0<\alpha<1$ del estimador de Nelson-Aalen es [1, Sección 3.1]

$$\left(\hat{H}(t) - z_{1-\alpha/2}\hat{SE}\left(\hat{H}(t)\right), \hat{H}(t) + z_{1-\alpha/2}\hat{SE}\left(\hat{H}(t)\right)\right).$$

1.2.3. Test Log-Rank

Tal y como hemos comentado previamente, uno de los objetivos es poder determinar si las funciones de supervivencia de dos grupos de individuos son significativamente distintas o no. Ante la falta de precisión a la hora de realizar esta tarea de manera visual, comparando dos gráficas, surge la alternativa de plantear un contraste de hipótesis.

La notación para cada instante t_i viene dada por la tabla 1.2.

Individuos	Grupo A	Grupo B	Total
En riesgo	$n_A(t_i)$	$n_B(t_i)$	$n(t_i)$
Sufren evento	$d_A(t_i)$	$d_B(t_i)$	$d(t_i)$

Tabla 1.2: Notaciones en un instante t_i para el contraste de igualdad de supervivencia entre dos grupos.

Planteamos así, el siguiente contraste de hipótesis

$$H_0 : S_A(t) = S_B(t) \quad \text{vs.} \quad H_1 : S_A(t) \neq S_B(t).$$

El estadístico que nos permite evaluar dicho contraste [10, Sección 1.5], resulta de considerar que el número de individuos de un grupo que experimentan el suceso en el instante t_i ($d_A(t_i)$ o $d_B(t_i)$) sigue una distribución hipergeométrica.

- ☞ Para cada grupo $j = A, B$ el número estimado de acontecimientos en el instante t_i es

$$\hat{e}_j(t_i) = \frac{n_j(t_i)d(t_i)}{n(t_i)}.$$

- ☞ Para cada grupo $j = A, B$ la estimación de la varianza del número de acontecimientos en el instante t_i es

$$\hat{\mathbb{V}ar}(d_j(t_i)) = \frac{n_A(t_i)n_B(t_i)[n(t_i) - d(t_i)]}{n^2(t_i)[n(t_i) - 1]}.$$

Recordamos que N denota el número de instantes de ocurrencia del evento, luego si $O = \sum_{i=1}^{N} d_A(t_i)$, $E = \sum_{i=1}^{N} \hat{e}_A(t_i)$, $V = \sum_{i=1}^{N} \hat{\mathbb{V}ar}(d_A(t_i))$, podemos definir el estadístico de **contraste de Log-Rank** como

$$Q = \frac{(O-E)^2}{V}.$$

Si N es muy grande, entonces Q se puede aproximar asintóticamente (bajo hipótesis nula H_0) mediante una distribución χ^2_1. Y, por lo tanto, si $p_{valor} = P\left(\chi^2_1 \geq Q\right) < 0{,}05$ rechazaremos la hipótesis nula, afirmando que las funciones de supervivencia de A y B son distintas. El test para el caso general de k grupos sigue un razonamiento muy similar [8, Sección 2.4].

1.3. Modelo de riesgos proporcionales de Cox

Hasta este momento, nos hemos limitado a aproximar la función supervivencia de los individuos tratándolos a todos por igual. Sin embargo, resulta obvio pensar que no siempre todos los individuos tienen la misma probabilidad de sufrir un evento concreto. Por ejemplo, si analizamos el evento "padecer cáncer de pulmón" el hecho de ser fumador (o por el contrario, de no serlo) parece, a priori, un factor clave. Del mismo modo, la edad también parece ser un factor a tener en cuenta, pero ¿cómo de influyente es?, ¿hay también diferencias considerables entre ambos sexos? En esta sección, trataremos de responder a este tipo de preguntas.

El modelo de riesgos proporcionales de Cox es el más empleado para analizar la influencia de los diferentes factores que guardan (o no) relación con el acontecimiento de un evento [8, Sección 3.2]. Estos factores reciben el nombre de covariables.

Cox definió la función de riesgo de un individuo con vector de covariables $\mathbf{z} = (z_1, z_2, \ldots, z_m)$, siendo $m \geq 1$, como

$$h(t, \mathbf{z}, \boldsymbol{\beta}) = h_0(t) e^{\mathbf{z}'\boldsymbol{\beta}}, \tag{1.6}$$

recibiendo $h_0(t)$ el nombre de *función de riesgo basal* y siendo $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_m)$ un vector de coeficientes estando cada coeficiente β_j asociado a la covariable z_j.

Destacar que se trata de un modelo semiparamétrico ya que incluye una parte paramétrica $\exp(\mathbf{z}'\boldsymbol{\beta})$, que refleja la influencia de las covariables y una no paramétrica $h_0(t)$, que recoge la influencia del tiempo de supervivencia T.

Definición 5. Bajo el modelo descrito, se denomina **tasa de riesgos** o **hazard ratio (HR)** entre dos individuos con valores $\mathbf{z}_0$ y $\mathbf{z}_1$ en el vector de covariables $\mathbf{z}$ al cociente

$$HR(t, \mathbf{z}_1, \mathbf{z}_0) = \frac{h(t, \mathbf{z}_1, \boldsymbol{\beta})}{h(t, \mathbf{z}_0, \boldsymbol{\beta})} = \frac{h_0(t) e^{\mathbf{z}_1'\boldsymbol{\beta}}}{h_0(t) e^{\mathbf{z}_0'\boldsymbol{\beta}}} = e^{(\mathbf{z}_1 - \mathbf{z}_0)'\boldsymbol{\beta}}.$$

Nota. El HR es independiente del tiempo de supervivencia, es decir, las funciones de riesgo de dos individuos son siempre proporcionales. Esta es la principal hipótesis que ha de cumplir el modelo (1.6) de Cox (hipótesis que deberá ser evaluada [11, Capítulo IV]) y es por ello que recibe el nombre de riesgos proporcionales de Cox.

A partir de (1.6), la función de riesgo acumulado puede ser expresada como

$$H(t, \mathbf{z}, \boldsymbol{\beta}) = \int_0^t h(x, \mathbf{z}, \boldsymbol{\beta}) dx = e^{\mathbf{z}'\boldsymbol{\beta}} \int_0^t h_0(x) dx = e^{\mathbf{z}'\boldsymbol{\beta}} H_0(t),$$

donde $H_0(t)$ se llama *función de riesgo acumulado basal*. Empleando (1.2) definimos la **función de supervivencia en el modelo de Cox** como

$$S(t, \mathbf{z}, \boldsymbol{\beta}) = e^{-H(t, \mathbf{z}, \boldsymbol{\beta})} = [e^{-H_0(t)}]^{\exp(\mathbf{z}'\boldsymbol{\beta})} = [S_0(t)]^{\exp(\mathbf{z}'\boldsymbol{\beta})}, \tag{1.7}$$

siendo $S_0(t)$ la *función de supervivencia basal*.

1.3.1. Ajustes y aproximación del modelo

Suponemos que tenemos n observaciones independientes, cada una de ellas representada (siguiendo la notación de la sección 1.1) mediante la terna $(y_i, \mathbf{z}_i, \delta_i)$ $i = 1, \ldots, n$. Nuestro primer objetivo va a ser buscar un estimador de $\boldsymbol{\beta}$.

Ante las dificultades de obtener el estimador máximo verosímil (EMV) de la función verosimilitud

$$L = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i) = \prod_{i=1}^{n} \left[\frac{h(y_i)}{\sum_{j \in R(y_i)} h_j(y_i)} \right]^{\delta_i} \left[\sum_{j \in R(y_i)} h_j(y_i) \right]^{\delta_i} S(y_i),$$

Cox propuso prescindir de los dos últimos términos puesto que no recogen información de $\boldsymbol{\beta}$ [10, Sección 4.2]. De este modo, formuló la *función de verosimilitud parcial* como

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{h(y_i)}{\sum_{j \in R(y_i)} h_j(y_i)} \right]^{\delta_i} = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{z}_i'\boldsymbol{\beta}}}{\sum_{j \in R(y_i)} e^{\mathbf{z}_j'\boldsymbol{\beta}}} \right]^{\delta_i},$$

y probó que el estimador que se obtiene de esa expresión tienen las mismas propiedades que el EMV [2, Capítulo VII]. Esta fórmula, usualmente, se simplifica prescindiendo de los datos censurados ($\delta_i = 0$) y asumiendo la hipótesis de que todos los tiempos de fallo son distintos. De esta forma, sean $t_1 < t_2 < \cdots < t_N$ los tiempos de fallo ordenados y notando que $y_i = t_i$ por ser $\delta_i = 1\ \forall i = 1,2,\ldots,N$, se alcanza la expresión simplificada

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{e^{\mathbf{z}_i'\boldsymbol{\beta}}}{\sum_{j\in R(t_i)} e^{\mathbf{z}_j'\boldsymbol{\beta}}}.$$

Por lo tanto, la *logverosimilitud parcial* queda caracterizada como

$$l_p(\boldsymbol{\beta}) = \log\left[L_p(\boldsymbol{\beta})\right] = \sum_{i=1}^{N}\left[\mathbf{z}_i'\boldsymbol{\beta} - \log\left(\sum_{j\in R(t_i)} e^{\mathbf{z}_j'\boldsymbol{\beta}}\right)\right].$$

Maximizando esta expresión por medio de métodos numéricos, se obtiene el estimador de máxima verosimilitud parcial $\hat{\boldsymbol{\beta}}_p$ de $\boldsymbol{\beta}$.

El estimador de la matriz de varianzas-covarianzas de $\hat{\boldsymbol{\beta}}_p$ se calcula del mismo modo que si fuera el EMV, es decir, tomando la *matriz de información observada* $I(\boldsymbol{\beta})$

$$I(\boldsymbol{\beta}) = -\frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2},$$

y evaluando su inversa en $\hat{\boldsymbol{\beta}}_p$. Consiguientemente, el estimador de la matriz de varianzas-covarianzas de $\hat{\boldsymbol{\beta}}_p$ es

$$\hat{\mathbb{V}ar}(\hat{\boldsymbol{\beta}}_p) = I(\hat{\boldsymbol{\beta}}_p)^{-1}.$$

En el caso de que haya empates entre los tiempos de fallo, se calculan aproximaciones de la función logverosimilitud parcial [8, Página 85].

Por último, vamos a hallar un estimador de la función supervivencia del modelo de Cox (1.7). Cabe destacar que una vez que se ha conseguido una estimación del vector de coeficientes $\boldsymbol{\beta}$, bastará con estimar la función supervivencia basal para lograr el propósito.

Sean $t_1,\ldots,t_N$ los tiempos de fallo, consideremos, al igual que en el estimador de Kaplan-Meier, que la función supervivencia se puede expresar como el producto de las probabilidades de supervivencia condicionadas

$$\alpha_i = P(T > t_i | T > t_{i-1}) = \frac{P(T > t_i)}{P(T > t_{i-1})} = \frac{S(t_i, \mathbf{z}, \boldsymbol{\beta})}{S(t_{i-1}, \mathbf{z}, \boldsymbol{\beta})}, \quad i = 1, \cdots, N,$$

con $t_0 = 0$. Por ende, se sigue que

$$S(t, \mathbf{z}, \boldsymbol{\beta}) = \prod_{i:t_i<t} \alpha_i = \prod_{i:t_i<t} \frac{S(t_i, \mathbf{z}, \boldsymbol{\beta})}{S(t_{i-1}, \mathbf{z}, \boldsymbol{\beta})} = \prod_{i:t_i<t} \left\{\frac{S_0(t_i)}{S_0(t_{i-1})}\right\}^{\exp(\mathbf{z}'\boldsymbol{\beta})} = \left\{\prod_{i:t_i<t} \frac{S_0(t_i)}{S_0(t_{i-1})}\right\}^{\exp(\mathbf{z}'\boldsymbol{\beta})},$$

y, como consecuencia, deducimos que (bajo las hipótesis del modelo de Cox) podemos plantear la función supervivencia basal como

$$S_0(t) = \prod_{i:t_i<t} \frac{S_0(t_i)}{S_0(t_{i-1})} = \prod_{i:t_i<t} \alpha_{0i},$$

reduciéndose el problema a hallar una estimación de α_{0i}, $i = 1, \cdots, N$. La función verosimilitud puede expresarse en términos de estos coeficientes α_{0i},

$$L = \prod_{i=1}^{N}\left[\prod_{l\in D(t_i)}\left(1 - \alpha_{0i}^{\exp(\mathbf{z}_l'\boldsymbol{\beta})}\right) \prod_{l\in R(t_i)-D(t_i)} \alpha_{0i}^{\exp(\mathbf{z}_l'\boldsymbol{\beta})}\right], \tag{1.8}$$

representando $D(t_i)$ el conjunto de individuos que sufren el fallo en el tiempo t_i.

Ahora, tomamos $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_p$ y maximizamos la verosimilitud con respecto a $\alpha_{01}, \ldots, \alpha_{0N}$. Derivando el logaritmo de (1.8) con respecto a α_{0i}, deducimos que el estimador de máxima verosimilitud ($\hat{\alpha}_{0i}$) de α_{0i} es solución de

$$\sum_{l \in D_i} \frac{\exp(\mathbf{z}_l' \hat{\boldsymbol{\beta}}_p)}{1 - \alpha_{0i}^{\exp(\mathbf{z}_l' \hat{\boldsymbol{\beta}}_p)}} = \sum_{l \in R(t_i)} \exp(\mathbf{z}_l' \hat{\boldsymbol{\beta}}_p). \tag{1.9}$$

Si en cada tiempo t_i se produce un único fallo (no hay empates), entonces (1.9) puede resolverse directamente despejando α_{0i},

$$\hat{\alpha}_{0i} = \left\{ 1 - \frac{\exp(\mathbf{z}_i' \hat{\boldsymbol{\beta}}_p)}{\sum_{l \in R(t_i)} \exp(\mathbf{z}_l' \hat{\boldsymbol{\beta}}_p)} \right\}^{\exp(\mathbf{z}_i' \hat{\boldsymbol{\beta}}_p)}.$$

En caso de que se produzcan empates, será preciso un método iterativo para poder extraer las estimaciones $\hat{\alpha}_{0i}$, $i = 1, \cdots, N$ [10, Sección 4.3]. De esta manera, el estimador de la función supervivencia basal es

$$\hat{S}_0(t) = \prod_{i:t_i < t} \hat{\alpha}_{0i}.$$

Concluimos que un estimador de la función supervivencia de Cox es

$$\hat{S}(t, \mathbf{z}, \boldsymbol{\beta}) = [\hat{S}_0(t)]^{\exp(\mathbf{z}' \hat{\boldsymbol{\beta}}_p)}.$$

1.3.2. Contraste de hipótesis

Una vez que se ha realizado el ajuste del modelo, debemos analizar si las covariables que hemos considerado son significativas o no. Esta cuestión la afrontaremos mediante el contraste de hipótesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

el cual se puede particularizar a casos concretos. Por ejemplo, si queremos analizar la significación conjunta de las covariables escogeremos $\boldsymbol{\beta}_0 = (0, \ldots, 0)$ y, en cambio, si queremos analizar la significación de la covariable z_j tomaremos $\boldsymbol{\beta}_0 = (\hat{\beta}_1, \ldots, \hat{\beta}_{j-1}, 0, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_m)$ con el 0 ocupando la posición j-ésima. Existen tres test distintos que posibilitan evaluar dicho contraste:

- **Test de razón de verosimilitudes (LRT)**: se denota por G y se calcula como el doble de la diferencia de la logverosimilitud parcial cuando el modelo considera todas las covariables y cuando el modelo se encuentra bajo la hipótesis nula H_0. Esto es,

  $$G = 2\left(l_p(\hat{\boldsymbol{\beta}}_p) - l_p(\boldsymbol{\beta}_0)\right).$$

  Bajo hipótesis nula, el estadístico G sigue asintóticamente una distribución χ^2_q siendo q el número de componentes nulas de $\boldsymbol{\beta}_0$. La justificación de este hecho se sigue de aproximar la función verosimilitud parcial mediante un proceso de conteo [2].

- **Test de Wald**: se basa en el hecho de que $\hat{\boldsymbol{\beta}}_p$ sigue asintóticamente una distribución

  $$\hat{\boldsymbol{\beta}}_p \sim N\left(\boldsymbol{\beta}_0, I(\hat{\boldsymbol{\beta}}_p)^{-1}\right).$$

  Por ello, se considera el estadístico

  $$z_W = (\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_0)' I(\hat{\boldsymbol{\beta}}_p)(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_0).$$

  Al igual que antes, bajo hipótesis nula, el estadístico z_W sigue una distribución χ^2_q [12].

■ **Test de Score**: Sea $U(\boldsymbol{\beta}) = \frac{\partial l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, se verifica asintóticamente y bajo la hipótesis nula que

$$U(\boldsymbol{\beta}) \sim N(0, I(\boldsymbol{\beta}))\,.$$

Luego, el estadístico

$$z_{SC} = U(\boldsymbol{\beta}_0)' \left[I(\boldsymbol{\beta}_0)\right]^{-1} U(\boldsymbol{\beta}_0),$$

bajo hipótesis nula, sigue también una distribución χ^2_q [10].

Normalmente, los tres estadísticos (G, z_W, z_{SC}) toman valores muy parecidos, es decir, los tres estadísticos nos suelen conducir a la misma decisión sobre si rechazar o aceptar la hipótesis nula. No obstante, en caso de discrepancia tomaremos el LRT como referente. En caso de aceptar la hipótesis nula, descartaremos del modelo las correspondientes covariables.

1.3.3. Interpretación de parámetros

Una vez que hemos estimado los coeficientes asociados a las covariables, resulta imprescindible conocer el significado que tienen los resultados obtenidos. Vamos a verlo para el caso de una única covariable, el caso general se puede ver en [8, Páginas 108-120].

Para evaluar el cambio producido en una covariable consideramos el logaritmo de la función riesgo

$$g(t,z,\beta) = \log\left[h(t,z,\beta)\right] = \log\left[h_0(t)\right] + z\beta.$$

En particular, la diferencia de esta función entre dos valores de la covariable $z = a$ y $z = b$ es

$$g(t,z=a,\beta) - g(t,z=b,\beta) = \{\log\left[h_0(t)\right] + a\beta\} - \{\log\left[h_0(t)\right] + b\beta\} = (a-b)\beta. \tag{1.10}$$

Por otro lado, la tasa de riesgos para la covariable z es

$$HR(t,z=a,z=b) = e^{(a-b)\beta}, \tag{1.11}$$

siendo esta la expresión que utilizaremos para interpretar el resultado. Contemplamos primero la interpretación para el caso sencillo en el que z es una variable dicotómica y, seguidamente, el caso general.

☞ Si z es una variable dicotómica, es decir $a = 1$ y $b = 0$, la expresión (1.10) se reduce a

$$g(t,z=1,\beta) - g(t,z=0,\beta) = (1-0)\beta = \beta.$$

Por otro lado, la expresión (1.11) es

$$HR(t,z=1,z=0) = e^{(1-0)\beta} = e^{\beta},$$

En efecto, si por ejemplo $\beta = \log 2$ se tiene que el HR es 2, luego los individuos con $z = 1$ tienen el doble de probabilidad de "morir" que los individuos con $z = 0$ en todo momento.

☞ En el caso general, si z es una variable continua (discreta), dados $a = x + c$ y $b = x$ donde $x, c \in \mathbb{R}$ $(x, c \in \mathbb{Z})$, entonces la expresión (1.10) se simplifica

$$g(t,z=a,\beta) - g(t,z=b,\beta) = ((x+c)-x)\beta = c\beta.$$

Por otro lado, la expresión (1.11) es

$$HR(t,z=a,z=b) = e^{((x+c)-x)\beta} = e^{c\beta}.$$

En este caso, si por ejemplo $c = 3$ y $\beta = 0{,}1$ se extrae que $HR = 1{,}35$, luego un incremento de 3 unidades en la covariable z conlleva un aumento del 35 % de la probabilidad de experimentar el evento en cualquier instante del tiempo.

1.4. Estimadores paramétricos

En las secciones anteriores, hemos considerado modelos semiparamétricos y no paramétricos con la ventaja de no tener que especificar la función supervivencia completamente, lo cual nos permite trabajar con dichos modelos sin probar que los datos se ajustan a una determinada distribución. Sin embargo, los modelos paramétricos nos facilitan la estimación de los paramétros, valores ajustados y residuos, entre otras cosas. En esta sección, presentaremos dos de ellos que pertenecen al grupo de **modelos de tiempo de fallo acelerado (AFT)**.

En estos modelos, se considera que la distribución del tiempo de supervivencia T viene dada por

$$T = e^{\beta_0 + \mathbf{z}'\boldsymbol{\beta}}\varepsilon, \tag{1.12}$$

donde β_0 es un intercepto y ε es un término de error que ha de tomar valores positivos. Atendiendo a la distribución de ε, se diferencian dos modelos.

1.4.1. Modelo de regresión Weibull

Si ε sigue una distribución Weibull de parámetro λ, cuya función de densidad recordamos que es

$$f(x) = \lambda x^{\lambda-1} e^{-x^\lambda} \quad x > 0, \tag{1.13}$$

el modelo (1.12) se denomina **modelo de regresión Weibull** y está caracterizado por la siguiente proposición.

Proposición 1.1. *La función de supervivencia del modelo de regresión Weibull es*

$$S(t, z, \boldsymbol{\beta}, \lambda) = \exp\left[-\left(\frac{t}{e^{\beta_0 + z'\boldsymbol{\beta}}}\right)^\lambda\right].$$

Demostración. Se sigue de (1.13) que la función de distribución de una Weibull de parámetro λ es $F(x) = 1 - e^{-x^\lambda}$. Luego, $P(\varepsilon > x) = e^{-x^\lambda}$ y concluimos que

$$S(t, z, \boldsymbol{\beta}, \lambda) = P(T > t) = P\left(e^{\beta_0 + \mathbf{z}'\boldsymbol{\beta}}\varepsilon > t\right) = P\left(\varepsilon > \frac{t}{e^{\beta_0 + \mathbf{z}'\boldsymbol{\beta}}}\right) = \exp\left[-\left(\frac{t}{e^{\beta_0 + \mathbf{z}'\boldsymbol{\beta}}}\right)^\lambda\right].$$

□

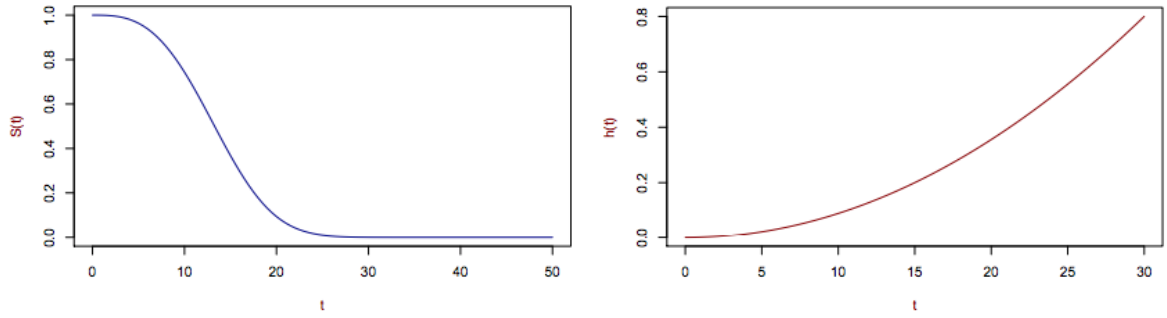
Proposición 1.2. *La función riesgo del modelo de regresión Weibull es*

$$h(t, z, \boldsymbol{\beta}, \lambda) = \frac{\lambda t^{\lambda-1}}{\left(e^{\beta_0 + z'\boldsymbol{\beta}}\right)^\lambda}.$$

Demostración. Se comprueba aplicando el resultado de la *Proposición* 1.1 a la siguiente igualdad

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

□



(a) Función supervivencia. (b) Función riesgo.

Figura 1.3: Un modelo de regresión Weibull.

1.4.2. Modelo de regresión exponencial

Si $\varepsilon \equiv Exp(1)$, entonces el modelo (1.12) es llamado **modelo de regresión exponencial**. Observar que, en tal contexto, la función de densidad del error es

$$f(x) = e^{-x} \quad x > 0, \tag{1.14}$$

coincidiendo con (1.13) para el caso $\lambda = 1$. Por consiguiente, se deducen los siguientes resultados.

Lema 1.3. *La función de supervivencia del modelo de regresión exponencial es*

$$S(t,z,\boldsymbol{\beta}) = \exp\left[-\frac{t}{e^{\beta_0+z'\boldsymbol{\beta}}}\right].$$

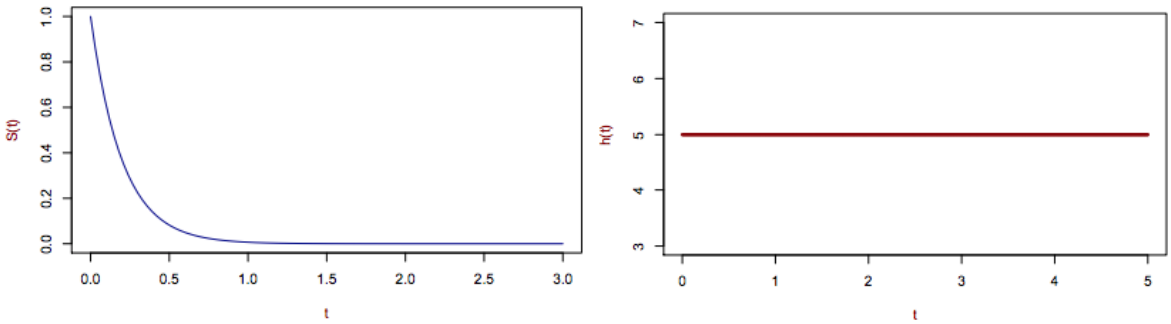
Demostración. Basta con considerar el resultado de la *Proposición* 1.1 en el escenario $\lambda = 1$. □

Lema 1.4. *La función riesgo del modelo de regresión exponencial es*

$$h(t,z,\boldsymbol{\beta}) = \frac{1}{e^{\beta_0+z'\boldsymbol{\beta}}}.$$

Demostración. Tomar meramente la *Proposición* 1.2 con $\lambda = 1$. □

Nota. Subrayar que la función riesgo para un determinado individuo es constante a lo largo del tiempo, es decir, un individuo tiene el mismo riesgo de fallar en cualquier momento del estudio. Asimismo, la probabilidad de fallar en un intervalo $[t, t+\Delta t]$ es independiente de la supervivencia previa a dicho intervalo. Esta característica se debe a la propiedad de *pérdida de memoria* de la distribución exponencial.



(a) Función supervivencia. (b) Función riesgo.

Figura 1.4: Un modelo de regresión exponencial.

Capítulo 2

Cure models

Este capítulo comienza presentando los llamados cure models en la sección 2.1. Seguidamente, en la sección 2.2, nos centraremos en los mixture cure models, el tipo de modelos cure más importante. Finalmente, detallaremos en la sección 2.3 el procedimiento más utilizado para estimar dichos modelos: el algoritmo EM.

2.1. Introducción y conceptos

Tal y como hemos comentado en el Capítulo 1, la hipótesis fundamental del análisis de supervivencia clásico es que todos los individuos bajo consideración, tarde o temprano, experimentarán el suceso de interés. Sin embargo, en determinadas ocasiones, esto no es así ya que existe una parte de los individuos que nunca lo van a experimentar. En términos de supervivencia, se puede expresar diciendo que esos individuos tienen un tiempo de supervivencia infinito.

Para poder considerar esta situación, los modelos de supervivencia clásicos se han extendido a los que ahora se denominan "cure models". Si bien estos modelos habían aparecido en la literatura hace muchos años, no se les había prestado atención hasta tiempos muy recientes y ya, a día de hoy, podemos afirmar que forman una parte muy importante del análisis de supervivencia ([4], [5], [13]). El nombre se debe a que el principal campo en el que se aplican es la medicina. Estos modelos consideran que los individuos pueden ser:

- **"Curados" o no susceptibles**: no experimentarán nunca el evento, es decir, $T = \infty$.
- **"No curados" o susceptibles**: experimentarán el evento en algún momento, es decir, $T < \infty$.

Desde el punto de vista de la medicina o, más ampliamente, de las ciencias de la Salud, el objetivo suele ser estudiar el tiempo de recurrencia de cierta enfermedad. Sin embargo, no todas las aplicaciones se limitan al ámbito de la Salud.

- En Economía, el interés puede estar en el tiempo que transcurre desde que un trabajador pierde su empleo hasta que consigue otro. En este caso, los individuos curados serán aquellos que no vuelven a conseguir ningún otro empleo, de manera que su tiempo de desempleo sería infinito.
- En Ingeniería, el problema puede ser analizar el tiempo transcurrido desde que una máquina o sistema se pone en marcha hasta que se avería. Los individuos curados serían ahora las máquinas o sistemas que nunca se averían.
- En el ámbito estudiantil, se podría considerar el tiempo que transcurre desde que un estudiante se enfrenta a un problema hasta que lo resuelve. Los estudiantes curados serían los que no consiguen resolverlo.

De nuevo, la variable de interés es el tiempo transcurrido T hasta que el suceso de interés ocurre y es claro que cuando t tiende a infinito una parte de los individuos no habrá experimentado el suceso. Denotaremos por $S_{pop}(t)$ a la función supervivencia del conjunto de la población.

Definición 6. Se denomina **cure fraction** a la proporción $1-p$ de individuos curados, esto es

$$\lim_{t\to\infty} S_{pop}(t) = P(T=\infty) = 1-p > 0.$$

Nota. Gráficamente, al analizar la supervivencia, se observará una meseta (zona plana) de altura $1-p$ en la cola derecha. Un claro ejemplo es la figura 2.1 en la cual el valor del cure fraction es 0,267. Además, destacar que si se realiza el análisis de supervivencia exclusivamente para los individuos no curados, entonces la meseta desaparece.

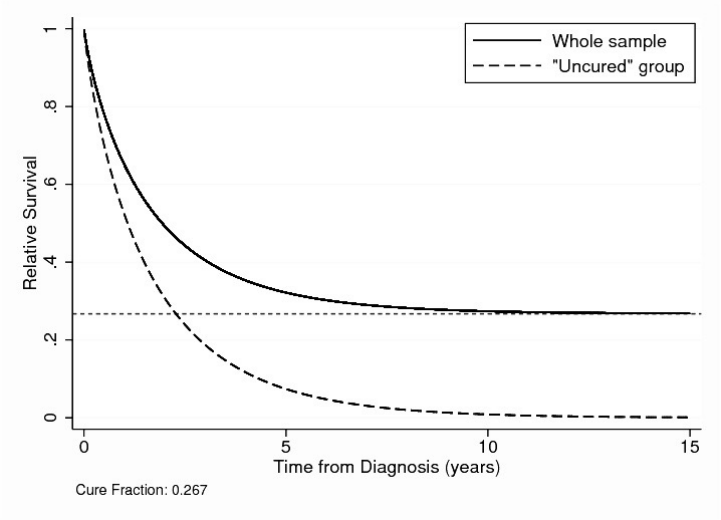


Figura 2.1: Cure rate.

Los modelos cure nos permiten conocer qué covariables están asociadas a una corta o una larga supervivencia. Por ejemplo, podríamos estudiar si una terapia concreta conlleva un incremento o un descenso de la probabilidad de ser un superviviente de larga duración.

Desde sus inicios, se han propuesto varios tipos de modelos cure, que pueden dividirse en dos categorías: **mixture cure models** y **promotion time cure models** (también llamados "non-mixture" cure models).

Los mixture cure models estudian el tiempo de supervivencia considerando que hay dos grupos de individuos que son diferenciados mediante la variable $B = I(T < \infty)$. Distinguiendo así, los individuos susceptibles ($B = 1$) de los curados ($B = 0$). Notar que en caso de que un individuo sea no censurado ($\delta = 1$) se tiene, siguiendo la notación (1.1), que $Y = T$, luego $B = 1$. Sin embargo, si está censurado no podremos determinar B, imposibilitando diferenciar los individuos susceptibles censurados de los curados. Estos modelos estiman la supervivencia de un individuo de la población como la probabilidad de ser curado más la probabilidad de no serlo y no haber sufrido el evento hasta ese momento. Explicaremos en mayor detalle estos modelos en la sección 2.2.

Los promotion time cure models fueron creados con fines biológicos para analizar el tiempo de recidiva en estudios oncológicos, para ello se modelizó el estado de latencia en tumores. En este caso, se asumió que tras un primer diagnótico y eficaz tratamiento de la enfermedad, un número $N \geq 0$ de células cancerosas subsisten en el organismo. Cada una de ellas se encuentra en estado latente durante un tiempo T_k, hasta que desarrolla un nuevo tumor. De esta forma, los individuos susceptibles de recidiva son aquellos con al menos una célula cancerosa en el organismo ($N > 0$), mientras que los curados son los que tienen $N = 0$ células cancerosas.

Estos modelos asumen que N sigue una distribución Poisson de parámetro $\theta > 0$, siendo θ el número medio de células cancerosas que tienen los individuos de una población. Además, se establece que los

tiempos T_k son variables aleatorias independientes con distribución idéntica $F(t)$ e independientes, a su vez, de N.

Proposición 2.1. *La función de supervivencia de los promotion time cure models viene dada por*

$$S_{pop}(t) = \exp[-\theta F(t)].$$

Demostración. Dado $t \geq 0$, se verifica que

$$\begin{aligned} S_{pop}(t) &= P(N=0) + P(T_1 > t, \ldots, T_N > t, N \geq 1) = \exp[-\theta] + \sum_{N=1}^{\infty} (1-F(t))^N \exp[-\theta] \frac{\theta^N}{N!} = \\ &= \exp[-\theta] \left[\sum_{N=0}^{\infty} (1-F(t))^N \frac{\theta^N}{N!} \right] = \exp[-\theta] \exp[\theta(1-F(t))] = \exp[-\theta F(t)]. \end{aligned}$$

□

Nota. Obsérvese que en los promotion time cure models, la proporción de individuos curados es

$$\lim_{t \to \infty} S_{pop}(t) = P(N=0) = \exp[-\theta],$$

y la función supervivencia para los individuos no curados es

$$S_u(t) = P(T > t | N \geq 1) = \frac{P(T > t, N \geq 1)}{P(N \geq 1)} = \frac{\exp[-\theta F(t)] - \exp[-\theta]}{1 - \exp[-\theta]}.$$

Debido a que su aplicación está tan focalizada en estudios oncológicos, no vamos a entrar más en detalle en este tipo de modelos. Algunos de estos modelos se pueden ver en [4, Sección 2.2].

2.2. Mixture cure models

Vamos a expresar, en términos matemáticos, el modelo comentado anteriormente. Con este fin, empezamos definiendo los siguientes términos.

Definición 7. Se denomina **incidencia** p(**x**) a la probabilidad de ser susceptible dado un vector de covariables **x**,

$$p(\mathbf{x}) = P(B = 1 | \mathbf{x}).$$

Definición 8. Se llama **latencia** $S_u(t|\mathbf{z})$ a la probabilidad de tener un tiempo de supervivencia mayor que t condicionado a ser susceptible y dado un vector de covariables **z**, es decir,

$$S_u(t|\mathbf{z}) = P(T > t | \mathbf{z}, B = 1).$$

Nota. Recalcar que la latencia sí que satisface que

$$\lim_{t \to \infty} S_u(t|\mathbf{z}) = 0.$$

En este contexto, la función supervivencia de la población total dados dos vectores de covariables **x**,**z** (que pueden ser iguales o no) puede ser modelada como

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = 1 - p(\mathbf{x}) + p(\mathbf{x}) S_u(t|\mathbf{z}),$$

siendo esta la ecuación que caracteriza a este tipo de modelos. A razón de su definición, podemos obtener una estimación de $S_{pop}(t|\mathbf{x}, \mathbf{z})$ estimando $p(\mathbf{x})$ y $S_u(t|\mathbf{z})$ por separado. Así, dentro del modelo tendremos dos submodelos independientes: uno para estimar la incidencia y otro para estimar la latencia.

2.2.1. Incidencia

Veamos dos posibles estimadores de la incidencia: uno paramétrico a través de la función *logit* y otro no paramétrico basado en el estimador de Kaplan-Meier (sección 1.2.1).

■ **Submodelo paramétrico de la incidencia: Logit**
Relaciona el vector de covariables $\mathbf{x}$ con la incidencia a través de la función *logit*,

$$logit(p) = \log\left(\frac{p}{1-p}\right), \quad p \in (0,1).$$

Por ende, la incidencia viene determinada por

$$logit(p(\mathbf{x})) = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma},$$

donde $\boldsymbol{\gamma}$ es un vector de coeficientes asociado al vector de covariables $\mathbf{x}$ y γ_0 es un intercepto. Se comprueba sencillamente que, en tal caso, el valor de la incidencia es

$$p(\mathbf{x}) = \frac{e^{\gamma_0+\mathbf{x}'\boldsymbol{\gamma}}}{1+e^{\gamma_0+\mathbf{x}'\boldsymbol{\gamma}}}.$$

En particular, si una covariable z_j aumenta una unidad, la probabilidad de sufrir el evento será e^{γ_j} veces mayor/menor (dependiendo del signo de γ_j) que antes. Otros estimadores paramétricos pueden localizarse en [13, Página 9].

■ **Submodelo no paramétrico de la incidencia: Kaplan-Meier**
Vamos a presentar el estimador en el caso de que no haya covariables. El caso general se encuentra en [13, Páginas 55-56]. Denotamos $\hat{S}^{KM}(t)$ al estimador de Kaplan Meier (1.5) y recordamos la notación $0 < t_1 < t_2 < \cdots < t_N$ para los distintos instantes (ordenados) en los que el evento tiene lugar. Habida cuenta de la *Definición* 6, se tiene que

$$1 - \hat{p} = \lim_{t\to\infty} \hat{S}(t) = \hat{S}^{KM}(t_N),$$

de donde se desprende el estimador

$$\hat{p} = 1 - \hat{S}^{KM}(t_N). \tag{2.1}$$

De la estimación de la varianza del estimador de Kaplan-Meier (*Apéndice* A.2), se deduce que la varianza del estimador (2.1) es

$$\hat{p}^2 \sum_{j=1}^{N} \frac{d_j}{n_j(n_j-d_j)}.$$

2.2.2. Latencia

Existe un amplio abanico de estimadores de la latencia. Nosotros vamos a centrarnos en dos de ellos que se sustentan respectivamente en el modelo de riesgos proporcionales de Cox (sección 1.3) y en el modelo de tiempo de fallo acelerado (comentado en la sección 1.4). Otros modelos se hallan en [13, Sección 1. 2].

■ **Submodelo semiparamétrico de la latencia: Cox**
Consideramos el modelo de riesgos proporcionales de Cox para la latencia. En tal caso, extraemos de (1.7) que

$$S_u(t|\mathbf{z}) = [S_0(t)]^{\exp(\mathbf{z}'\boldsymbol{\beta})}.$$

Si especificáramos la función de supervivencia basal $S_0(t)$, nos encontraríamos ante un modelo paramétrico con el cual sería más sencillo obtener las estimaciones de los parámetros. No obstante, vamos a considerar el modelo en su forma semiparamétrica puesto que resulta más atractivo. En concreto, la función de supervivencia basal puede ser expresada como

$$S_0(t) = P(T > t | \mathbf{Z} = 0, B = 1).$$

El modelo cure que asume el submodelo logit en la incidencia y el de Cox en la latencia se denomina **Logistic/Cox (LC) Mixture Cure Model**.

■ **Submodelo semiparamétrico de la latencia: AFT**
Adopta el modelo de tiempo de fallo acelerado para la latencia. Así, a partir de (1.12) se tiene la expresión

$$\log(T) = \beta_0 + \mathbf{z}'\boldsymbol{\beta} + \varepsilon^*.$$

Si especificáramos la distribución del error ε^*, tendríamos modelos como el Weibull o el exponencial (comentados en la sección 1.4). Sin embargo, por la misma razón que antes, no se suele especificar dicha distribución. Obteniéndose así, el modelo semiparámetrico. El modelo cure que asume el submodelo logit en la incidencia y el AFT semiparámetrico en la latencia se denomina **Logistic/Semiparametric AFT Mixture Cure Model**.

2.3. Estimación de parámetros

La estimación de parámetros en los modelos cure se realiza a través del algoritmo EM (*Apéndice* B). Procedemos a explicar los pasos del algoritmo para el LC Mixture Cure Model. De todos modos, el algoritmo para otros modelos como el Logistic/Semiparametric AFT Mixture Cure Model (se puede ver en [4, página 28]) se diferencia solo en el Paso M.

Supongamos que disponemos de la información observada $O_i = (Y_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$. La verosimilitud de cada individuo será una de las tres siguientes expresiones

- Si ha experimentado el evento ($\Delta_i = 1$, $B_i = 1$):

$$p(\mathbf{X}_i)\, h_u(Y_i|\mathbf{Z}_i)\, S_u(Y_i|\mathbf{Z}_i).$$

- Si es susceptible ($B_i = 1$), pero no ha experimentado el evento ($\Delta_i = 0$):

$$p(\mathbf{X}_i)\, S_u(Y_i|\mathbf{Z}_i).$$

- Si es curado ($\Delta_i = 0, B_i = 0$):

$$1 - p(\mathbf{X}_i).$$

donde $h_u(t|\mathbf{z})$ es la función riesgo de los individuos susceptibles. Concluimos que la función verosimilitud de los mixture cure models es

$$L_c(\boldsymbol{\gamma}, \boldsymbol{\beta}, S_0) = \prod_{i=1}^{n} \{p(\mathbf{X}_i)\, h_u(Y_i|\mathbf{Z}_i)\, S_u(Y_i|\mathbf{Z}_i)\}^{\Delta_i B_i} \prod_{i=1}^{n} \{p(\mathbf{X}_i)\, S_u(Y_i|\mathbf{Z}_i)\}^{(1-\Delta_i)B_i} \prod_{i=1}^{n} \{1 - p(\mathbf{X}_i)\}^{(1-\Delta_i)(1-B_i)}. \tag{2.2}$$

Observar que (2.2) puede expresarse como producto de dos factores

$$L_1(\boldsymbol{\gamma}) = \prod_{i=1}^{n} p(\mathbf{X}_i)^{B_i} \{1 - p(\mathbf{X}_i)\}^{1-B_i}, \tag{2.3}$$

$$L_2(\boldsymbol{\beta}, S_0) = \prod_{i=1}^{n} \left[\{h_u(Y_i|\mathbf{Z}_i)\, S_u(Y_i|\mathbf{Z}_i)\}^{\Delta_i B_i} S_u(Y_i|\mathbf{Z}_i)^{(1-\Delta_i)B_i} \right], \tag{2.4}$$

cada una de ellos conteniendo los parámetros respectivos a uno de los submodelos. Luego, podremos estimar por separado la latencia y la incidencia. De este modo, los dos pasos del algoritmo EM en la m-ésima iteración serán:

1. **Paso E:** Computar la esperanza del logaritmo de (2.2) con los valores de los parámetros obtenidos en la iteración anterior $\theta^{(m-1)} = (\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\beta}, S_0)^{(m-1)}$ y con la información observada O_i. La única variable no observada es B_i, para solventar esta cuestión se evalúa su esperanza

$$E\left(B_i|O_i,\theta^{(m-1)}\right) = \Delta_i\left\{1\times P\left(B_i=1|Y=Y_i,\Delta_i=1,\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)}\right)\right\} + (1-\Delta_i)\left\{1\times P\left(B_i=1|Y=Y_i,\Delta_i=0,\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)}\right)\right\}.$$

   Por un lado, recordar que $P(B_i=1|\Delta_i=1)=1$ y por otro, notar que se verifica que

$$P\left(B_i=1|Y=Y_i,\Delta_i=0,\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)}\right) = \frac{P\left(B_i=1,T>Y_i|\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)}\right)}{P\left(T>Y_i|\mathbf{X}=\mathbf{X}_i,\mathbf{Z}=\mathbf{Z}_i,\theta^{(m-1)}\right)}.$$

   En consecuencia, la esperanza de B_i se puede calcular como

$$E\left(B_i|O_i,\theta^{(m-1)}\right) = \Delta_i + (1-\Delta_i)\frac{p^{(m-1)}(\mathbf{X}_i)\,S_u^{(m-1)}(Y_i|\mathbf{Z}_i)}{1-p^{(m-1)}(\mathbf{X}_i)+p^{(m-1)}(\mathbf{X}_i)\,S_u^{(m-1)}(Y_i|\mathbf{Z}_i)} = W_i^{(m)}.$$

   Ahora ya podemos computar la esperanza del logaritmo de (2.2) sustituyendo B_i por $W_i^{(m)}$.

2. **Paso M:** Maximizar la verosimilitud respecto a los parámetros del modelo. Tal y como hemos comentado antes, realizaremos el proceso en dos partes; una maximizando (2.3) respecto a los parámetros correspondientes a la incidencia y otra, maximizando (2.4) respecto a los parámetros correspondientes a la latencia.

   a) Incidencia: nótese que (2.3) es igual a la función verosimilitud de los modelos de regresión logística [7, sección 8.1]. Luego, se usará el método de Newton-Raphson para estimar los parámetros $\gamma_0, \boldsymbol{\gamma}$.

   b) Latencia: se implementa un método propuesto por Sy y Taylor [15] basado en el método utilizado en el modelo de Cox (sección 1.3.1). Primero de todo, estimamos la función de riesgo acumulado basal mediante

$$\hat{H}_0(t) = \sum_{j:Y_j<t}\frac{D(t_j)}{\sum_{k\in R(t_j)} W_k^{(m)}\exp(\mathbf{Z}_k'\boldsymbol{\beta})}.$$

   Seguidamente, sustituimos este estimador en (2.4), obteniendo (asumiendo que no hay empates) la siguiente expresión de la verosimilitud

$$\hat{L}_2(\boldsymbol{\beta}|\mathbf{W}^{(m)}) = \prod_{i=1}^{n}\left\{\frac{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}{\sum_{k\in R(t_j)} W_k^{(m)}\exp(\mathbf{Z}_k'\boldsymbol{\beta})}\right\}^{\Delta_i}, \tag{2.5}$$

   donde $\mathbf{W}^{(m)} = \left(W_1^{(m)},\ldots,W_n^{(m)}\right)$. Así pues, aplicando el método de Newton-Raphson, maximizamos (2.5) con respecto a $\boldsymbol{\beta}$, consiguiendo una estimación de dicho parámetro.

De este modo, se logra una nueva estimación de los parámetros $\theta^{(m)} = (\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\beta}, S_0)^{(m)}$. Se repetirán los dos pasos hasta que se detecte convergencia.

Capítulo 3

Aplicación a la COVID-19

En este capítulo se lleva a acabo el análisis de un problema sanitario ocasionado por la COVID-19 a partir de un conjunto de datos y con el fin de ilustrar el análisis de supervivencia explicado en el Capítulo 1 y los modelos cure descritos en el Capítulo 2. El análisis aquí mostrado se enmarca en un estudio más global en el que se analizan diversos aspectos de los datos, con el objetivo de mejorar la gestión hospitalaria en el contexto de pandemia.

Desde el inicio de la pandemia, numerosas investigaciones de esta índole han sido efectuadas con el objetivo de tratar de dar respuesta a múltiples incógnitas como cuáles eran los factores más influyentes o los tratamientos más efectivos, de ayudar en toma de decisiones críticas o incluso de intentar predecir la evolución de un paciente o del número de casos. Algunas de estas pueden verse en [3], [5] y [9].

El estudio se va a realizar a partir de un conjunto de datos proporcionado por el Gobierno de Aragón a la Universidad de Zaragoza. Dichos datos recogen la información de los individuos de una población que desde el 2 de marzo de 2020 hasta el 23 de marzo de 2021 dieron positivo tras la realización de uno de los posibles test.

La extraordinariedad de los primeros meses de la pandemia conllevó irregularidades en la recogida de datos y, es por ello, que sólo vamos a considerar los tomados a partir del 1 de julio de 2020. A partir de esta fecha, los datos reflejan de manera efectiva lo acontecido, siendo la vacunación (iniciada a finales del año 2020) el principal factor influyente ausente. En particular, la investigación se centra en la hospitalización e ingreso en la Unidad de Cuidados Intensivos (UCI) de los pacientes covid. De este modo, las variables tomadas de la base de datos para cada individuo son:

- *ID* del paciente: único y anónimo.
- *Edad* del individuo en la fecha del diagnóstico.
- *Sexo* (mujer/hombre) del paciente.
- *Grupo del GMA*: es un agrupador de morbilidad que permite estratificar a la población con el fin de identificar a las poblaciones diana. Los grupos son: población sana (0), embarazo y/o parto (10), patología aguda (20), enfermedad crónica de un sistema (31), enfermedad crónica en 2 o 3 sistemas (32), enfermedad crónica en 4 ó más sistemas (33) y neoplasias en el período (40). Cada uno de estos (excepto el de población sana) se fragmenta en 5 subgrupos atendiendo a la complejidad del individuo [6].
- *Fecha de referencia*: fecha de realización del test positivo.
- *Fecha de ingreso hospitalario*, en caso de que la evolución de los síntomas lo haya requerido.
- *Fecha de alta hospitalaria*, en caso de haber sido ingresado.

- *Fecha de ingreso en la UCI*, en caso de que la gravedad de los síntomas haya requerido de cuidados intensivos.

- *Fecha de alta en la UCI*, en caso de haber requerido dicho servicio.

Según criterio médico, todo ingreso hospitalario producido un mes (o más tiempo) después de que el paciente diera positivo, es motivado por causas ajenas a la covid. Así, con el fin de evitar tomar un elevado número de observaciones censuradas, vamos a considerar exclusivamente los datos con *fecha de referencia* anterior al 23 de Febrero de 2021. Dicho de otro modo, descartamos todos los individuos con *fecha de referencia* posterior al 23 de Febrero de 2021 puesto que una parte de ellos habrán sido hospitalizados debido a la covid después del 23 de Marzo de 2021 y no disponemos de esa información.

La manipulación y el análisis de datos ha sido realizado usando el software R [14]. El correspondiente código, los resultados y los gráficos obtenidos se encuentran en el *Apéndice* C, además detallaremos en él lo que representa cada parte del código. No obstante, a lo largo de este capítulo se irán explicando todos los desarrollos.

3.1. Estancia en UCI

Vamos a comenzar analizando la estancia en UCI de los pacientes que han requerido dicho servicio. De este modo, definimos la variable de interés como "tiempo transcurrido desde el ingreso en UCI hasta el alta", siendo el evento a considerar el alta en UCI del paciente ingresado. Resulta importante incidir en que el evento a considerar es el alta, un alta que se puede producir o bien por el fallecimiento del paciente o bien por el traslado a planta del mismo. Este hecho se traduce en que no queremos predecir el pronóstico (recuperación o muerte) de un paciente sino que trataremos de conseguir resultados que puedan resultar muy útiles para determinar e incluso predecir el nivel de ocupación de camas de UCI, siendo este uno de los criterios utilizados durante la pandemia para establecer el nivel de alerta de un territorio.

Para poder realizar el correspondiente análisis de supervivencia hemos seleccionado los individuos que tienen fecha de ingreso y alta de UCI y, seguidamente, hemos creado la variable *evento*: "Recibir alta en UCI". Por defecto, en la base de datos aparecen con fecha de alta "23-03-2021" aquellos individuos censurados, es decir, aquellos que el 23 de Marzo seguían ingresados en UCI sin recibir el alta. Además, hemos calculado la variable *tiempo de estancia* como la diferencia entre la fecha de alta y la de ingreso. Por último, la variable continua *edad* la hemos transformado a una variable de tipo factor (*edadF*) en base a sus cuantiles (57, 65 y 72 años), pero solo hemos considerado tres grupos: menores de 57 años, 58-72 años, mayores de 72 años. El motivo de juntar los grupos de 58-65 años y 66-72 años es que, a priori, no hay mucha diferencia entre ambos.

Eliminando todos los datos que tienen algún valor nulo, obtenemos un conjunto de datos compuesto por 672 individuos. Una muestra de ellos junto con las 5 covariables que consideraremos aparece en la tabla 3.1.

Las observaciones de la variable *tiempo de estancia* pueden ser representadas gráficamente como en la figura C.1 (*Apéndice* C) o incluso si queremos particularizar para un subconjunto concreto de la muestra como en la figura C.2. A vista del primero de estos dos gráficos, podemos observar en qué momentos de la pandemia ha habido mayor ocupación en las Unidades de Cuidados Intensivos, destacando claramente el mes de noviembre. Por contra, el periodo de finales de diciembre e inicio de enero es el de mayor descongestión.

Curva de supervivencia

Representamos gráficamente la función supervivencia mediante el estimador de Kaplan-Meier (1.5) junto con los intervalos de confianza a un nivel del 95 %, es decir, con significación $\alpha = 0{,}05$ para cada

ID	Tiempo de estancia	Evento	EdadF	GrupoGMA	Sexo
1	71	1	58 – 72 años	33	Hombre
2	71	1	58 – 72 años	33	Hombre
3	24	1	Mayor de 72 años	33	Hombre
4	16	0	58 – 72 años	32	Mujer
5	21	1	58 – 72 años	32	Hombre
6	6	1	58 – 72 años	32	Mujer

Tabla 3.1: Muestra del conjunto de datos utilizado para analizar el tiempo de estancia en UCI.

tiempo t a partir de la fórmula de Greenwood (*Apéndice* A.4). Además, añadimos al gráfico una tabla que nos muestra el número de individuos en riesgo para ciertos valores de t (figura C.3).

Observamos que sólo un individuo ha estado más de 120 días en la UCI, de hecho, este individuo está censurado y a día "23-03-2021" llevaba ingresado 138 días. Además, cabe destacar que el 25 % de las personas que ingresan en UCI permanecen más de un mes hasta que son dados de alta. La mediana de la estancia es de 18 días (*página* 33).

La banda de confianza es prácticamente inapreciable, esto se debe a que se ha calculado como la unión de los intervalos de confianza de cada uno de los puntos. Así, el resultado es mucho más preciso, lo cual supone una banda más estrecha.

Test de LogRank

Vamos a aplicar el test de Logrank (sección 1.2.3) con el objetivo de valorar la influencia de las covariables *edadF*, *sexo* y *grupoGMA* sobre la función supervivencia, es decir, sobre el tiempo de estancia de los individuos en la UCI. Además, tomaremos $\alpha = 0{,}05$ como significación del contraste de hipótesis.

- *Sexo*: Se pueden ver las curvas de supervivencia para las mujeres y hombres en la figura C.4. A simple vista no se aprecia gran diferencia en los tiempos de supervivencia entre ambos grupos. No obstante, hay que destacar el hecho de que han ingresado en UCI más del doble de pacientes hombres que de mujeres. En concreto, el test de LogRank nos proporciona la información de la tabla 3.2.

	N	Observed	Expected	$\frac{(O-E)^2}{V}$
sexo=HOMBRE	451	434	440	0,277
sexo=MUJER	221	212	206	0,277

Chisq= 0,3 on 1 degree of freedom, $p = 0{,}6$

Tabla 3.2: Test de LogRank para la covariable *Sexo*.

El p-valor es mayor que 0,05 luego no se rechaza la hipótesis nula de igualdad de las funciones de supervivencia entre ambos grupos, es decir, pese a que entren muchos más hombres que mujeres en UCI, el sexo no influye significativamente en el tiempo de estancia.

- *EdadF*: Se pueden ver las funciones de supervivencia para los tres grupos de edad en la figura C.5. A diferencia de la covariable *sexo*, en este caso sí que se aprecian diferencias entre los grupos. El 50 % de los menores de 58 años que ingresan en UCI están menos de 12 días (los otros dos grupos una semana más en media), no obstante los individuos de ese grupo que permanecen más de 12

días acaban teniendo una estancia media más larga que las de los otros grupos. Por el contrario, el grupo con percentil 75 más bajo es el de mayores de 72 años debido, en gran parte, a los fallecimientos (*página* 35). Veamos si la información de la tabla 3.3 proporcionada por el test de LogRank confirma lo observado.

	N	Observed	Expected	$\frac{(O-E)^2}{E}$
edadF=Menor de 58 años	175	168	157	0,772
edadF=58-72 años	338	326	356	6,01
edadF=Mayor de 72 años	159	152	133	2,788

Chisq= 6,4 on 2 degree of freedom, $p = 0{,}04$

Tabla 3.3: Test de LogRank para la covariable *EdadF*.

El p-valor es menor que 0,05 luego se rechaza la hipótesis nula de igualdad de las funciones de supervivencia entre los tres grupos. Luego, el hecho de pertenecer a un grupo de edad u otro influye significativamente en el tiempo de estancia en UCI.

- *GrupoGMA*: Nos vamos a centrar en analizar si existen diferencias entre los tiempos de supervivencia para los grupos 31, 32 y 33. La elección de estos tres grupos se debe a su significado: enfermedad crónica en un sistema, enfermedad crónica en dos o tres sistemas y enfermedad crónica en 4 o más sistemas. Además, la frecuencia de estos tres grupos es ampliamente superior a la del resto. Las curvas de supervivencia, que se pueden ver en la figura C.6, parecen prácticamente idénticas. Veamos si la información de la tabla 3.4 proporcionada por el test de LogRank ratifica la igualdad.

	N	Observed	Expected	$\frac{(O-E)^2}{E}$
grupoGMA=31	79	74	70,7	0,1577
grupoGMA=32	246	239	236,9	0,0181
grupoGMA=33	289	277	282,4	0,1037

Chisq= 0,3 on 2 degree of freedom, $p = 0{,}9$

Tabla 3.4: Test de LogRank para la covariable *grupoGMA*: 31,32,33.

El p-valor es claramente superior a 0,05 luego no se puede rechazar la hipótesis nula de igualdad de las funciones de supervivencia entre los tres grupos, esto es, no hay diferencias significativas entre los tiempos de supervivencia de los grupos GMA 31,32 y 33.

Regresión de Cox

Vamos a aplicar el modelo de riesgos proporcionales de Cox (sección 1.3) a nuestro caso. Atendiendo a lo razonado en el apartado anterior sobre la significación de las covariables, tomaremos *EdadF* y *grupoGMA*, agrupando los grupos 31, 32 y 33 bajo el nombre "Enfermedad crónica en algún sistema". Un razonamiento similar nos conduce a juntar los grupos 10 y 20 con el nombre "Patología aguda o embarazo". Los otros dos grupos serán "Población sana" y "Neoplasias" en referencia a los grupos 0 y 40 respectivamente. El resumen del modelo se encuentra en la tabla 3.5.

A la vista de los p-valores mostrados en la última columna de la tabla 3.5, no se puede rechazar que los coeficientes de todas las covariables tomen valores nulos. Este razonamiento también se puede deducir de que el 1 se encuentra en todos los intervalos de confianza de los exponentes de los coeficientes. Los resultados de los tres test realizados (sección 1.3.2) arrojan p-valores ligeramente superiores a 0,05.

	coef	exp(coef)	lower .95	upper .95	z	Pr(>|z|)
edadF[T. 58-72 años]	−0,146	0,863	0,715	1,042	−1,531	0,126
edadF[T. Mayor de 72 años]	0,085	1,089	0,870	1,364	0,745	0,456
GMA[T. Embarazo o patología]	0,639	1,895	0,773	4,642	1,398	0,162
GMA[T. Enfermedad crónica]	0,305	1,356	0,606	3,036	0,743	0,458
GMA[T. Neoplasis]	0,407	1,502	0,616	3,660	0,896	0,370

LRT=9.48 on 5 df, $p = 0{,}09$	Wald test=9.7 on 5 df, $p = 0{,}08$	Score test=9.76 on 5 df, $p = 0{,}08$

Tabla 3.5: Resumen del modelo de Cox.

Por todo ello, no se asegura la significación conjunta de las covariables. Aun así, el modelo puede ser de gran utilidad para analizar los tiempos de estancia en UCI por grupos. Si nos fijamos en la segunda columna de la tabla 3.5, las dos primeras filas muestran los *hazard ratios* de los dos grupos de *edadF* respecto al grupo "Menor de 58 años". Por su parte, las tres filas siguientes muestran los *hazard ratios* de los tres grupos de *GMA* respecto al grupo "Población sana". Notar que el hecho de pertenecer a uno de esos tres grupos de GMA supone un incremento de la probabilidad de experimentar el evento.

3.2. Estancia en hospital

Ahora vamos a analizar la estancia en planta y el posible posterior ingreso en UCI de los pacientes que han sido hospitalizados. Así, definimos la variable de interés como "tiempo transcurrido desde la admisión en el hospital hasta el ingreso en UCI", siendo el evento a considerar el ingreso en la Unidad de Cuidados Intensivos. A diferencia del estudio anterior en el que todos los individuos (excepto los censurados) experimentaban el suceso, en este caso gran parte de ellos no lo experimentarán puesto que o bien fallecerán en planta o bien su evolución será favorable y serán dados de alta. De este modo, nos encontramos ante un claro ejemplo en el que puede ser muy útil aplicar los modelos cure (Capítulo 2).

Al igual que antes, para poder realizar el correspondiente análisis de supervivencia hemos seleccionado los individuos con fecha de ingreso en hospital. A partir de aquí, hemos creado la variable *evento*: "Ser ingresado en UCI" y la variable *días* calculada como la diferencia entre la fecha de ingreso en UCI y la de admisión en el hospital. Para todos aquellos individuos que no han experimentado el evento, en la variable *días* aparecerá la diferencia entre el 23 de marzo de 2021 y su correspondiente fecha de hospitalización. Además, la variable continua *edad* la hemos transformado a una variable de tipo factor (*edadF*) con tres grupos: "menor de 58 años", "58-84 años" y "mayor de 84 años" siguiendo un razonamiento idéntico al del estudio anterior.

Con todo ello, si representamos la curva de supervivencia mediante el estimador de Kaplan-Meier (figura C.7) podemos observar una gran meseta en la cola derecha que corresponde al conjunto de individuos que no han sido ingresados en UCI. Este hecho nos conduce a los modelos cure y, en particular, a los mixture cure models. Para poder discernir entre los individuos susceptibles censurados y los curados, calculamos la diferencia máxima entre la fecha de ingreso en UCI y la de hospitalización (39 días). Así, definimos como individuo curado ($B = 0$) toda persona ingresada en hospital antes del 13 de Febrero de 2021 y que no ha precisado UCI. Por contra, los individuos susceptibles ($B = 1$) serán aquellos que o bien han experimentado el suceso o bien han sido hospitalizados con fecha igual o posterior al 13 de Febrero de 2021 y no lo han experimentado.

Considerando las covariables *edadF* y *sexo*, vamos a trabajar con un conjunto de 9823 observaciones. Una muestra de ellas junto con las 5 variables a considerar es la tabla 3.6. Las librerías y paquetes estadísticos de R que permiten hacer uso de los modelos cure han comenzado a ser desarrolladas en

los últimos años y, a día de hoy, todavía resulta bastante complejo utilizarlas. Ante este inconveniente, hemos estimado la latencia e incidencia a partir de la función *SNPCPKtoSurvival* creada por Miguel Lafuente Blasco (Apéndice C.1).

ID	Días	Evento	Curado	EdadF	Sexo
1	112	0	0	Menor de 58 años	Hombre
2	0	1	1	58 – 84 años	Hombre
3	139	0	0	Mayor de 84 años	Mujer
4	32	0	1	58 – 84 años	Mujer
5	189	0	0	Menor de 58 años	Mujer
6	7	1	1	58 – 84 años	Hombre

Tabla 3.6: Muestra del conjunto de datos utilizado para analizar el tiempo de estancia en planta.

Si aplicamos la función mencionada a nuestro conjunto de datos (sin covariables), obtenemos que la incidencia es 0,068, es decir, la probabilidad de ser ingresado en UCI una vez que estás hospitalizado es de 0,068. Así, la *cure fraction* es 0,931 siendo esta la altura de la meseta observada en la figura C.7. Por otra parte, la estimación de la latencia puede verse en la figura C.8, observamos que pasados 40 días todo individuo susceptible es ingresado en UCI y, de hecho, más del 90 % de los susceptibles son trasladados a UCI antes de llevar 10 días hospitalizados. Deducimos que casi todos los traslados de planta a UCI se realizan antes de los 10 días de hospitalización y, por consiguiente, si una persona lleva hospitalizada 10 días o más, la probabilidad de que sea traslada a UCI es prácticamente nula. Otro aspecto reseñable es que la latencia en el momento $t = 0$ toma valor 0,735, en lugar de 1. Esto se debe a que un elevado porcentaje de individuos han sido ingresados en UCI el mismo día que han sido admitidos en el hospital, es decir, la primera observación del individuo coincide con el evento. Este es el caso del individuo 2 de la tabla 3.6.

Con el fin de analizar la influencia de la covariable *sexo*, hemos implementado la función *SNPCPKtoSurvival* a cada uno de los grupos de la covariable. De este modo, la incidencia estimada para las mujeres es 0,047 mientras que la de los hombres es 0,087. Se sigue de estos resultados que la probabilidad de ser susceptible a ser ingresado en UCI una vez que lo estás en planta es casi el doble para los hombres que para las mujeres. Este hecho justifica, en parte, lo visto en el primer estudio de que han ingresado en UCI el doble de pacientes hombres que de mujeres. En cambio, no se pueden observar apenas diferencias en la figura C.9 de las respectivas latencias. En conclusión, la covariable sexo es significativa en la probabilidad de ser susceptible pero no en la función supervivencia de los susceptibles.

Del mismo modo, para la covariable *edadF* hemos obtenido que las incidencias estimadas son 0,066, 0,101 y 0,001 para los grupos “Menor de 58 años”, “58-84 años ” y “Mayor de 84 años” respectivamente. Estos resultados ponen en manifiesto el hecho de que prácticamente ningún paciente mayor de 84 años es ingresado en UCI y que, en cambio, aproximadamente el 10 % de los individuos de entre 58 y 64 años que ingresan en el hospital son susceptibles de ser trasladados a UCI. Esta significativa diferencia se puede observar también en la latencia (figura C.10). Notar que el grupo de “Mayor de 84 años” está condicionado por la falta de observaciones susceptibles. Sin embargo, se observan diferencias entre los otros dos grupos donde la función supervivencia de los susceptibles menores de 58 años está claramente por debajo de la del grupo de “58-84 años ” durante los primeros 17 días.

Los dos estudios llevados a cabo arrojan resultados interesantes sobre la trascendencia de la edad en la probabilidad de entrar en UCI y en la duración de sus estancias en planta y UCI. Por contra, el sexo es influyente en la probabilidad de acabar ingresado en la UCI pero no en la duración de sus correspondientes estancias. La investigación podría ser continuada analizando los factores y la probabilidad de ser hospitalizado, además de considerar la influencia de otras covariables en los eventos estudiados.

Bibliografía

[1] AALEN, O.O., BORGAN, O. AND GJESSING, H.K., *Survival and Event History Analysis*. Springer, New York (2008).

[2] ANDERSEN, P.K., BORGAN, O., GILL, R.D. AND KEIDING, N., *Statistical Models Based on Counting Processes*. Springer, New York (1993).

[3] AGUIAR, M., BIDAURRAZAGA, J., MAR, J., MILLÁN, E., STOLLENWERK, N., *Modelling COVID-19 in the Basque Country from introduction to control measure response*, Scientific Reports,10, 17306 (2020). https://doi.org/10.1038/s41598-020-74386-1

[4] AMICO, M., *Cure Models in Survival Analysis: From Modelling to Prediction Assessment of the Cure Fraction.* PhD thesis, Louvain Catholic University and Leuven Catholic University (2018).

[5] CAO, R., LÓPEZ, C., PEDROSA, M., *Cure models to estimate time until hospitalization due to COVID-19*, Applied Intelligence (2021). https://doi.org/10.1007/s10489-021-02311-8

[6] CLÈRIES, M., MONTERDE, D., VELA, EMILI., *Los grupos de morbilidad ajustados: nuevo agrupador de morbilidad poblacional de utilidad en el ámbito de la atención primaria*, Atención Primaria, 48, 674-682 (2016).

[7] EFRON, B., HASTIE, T., *Computer Age Statistical Inference*. Cambridge University Press, United Kingdom (2016).

[8] HOSMER, D.W., LEMESHOW, S. AND MAY, S., *Applied Survival Analysis*. 2nd Edition, Wiley Series in Probability and Statistics, New York (2008).

[9] HE, W., YI, G., ZHU, Y., *Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis*, Medical Virology, 92, 2543– 2550 (2020). https://doi.org/10.1002/jmv.26041

[10] KALBFLEISCH, J.D. AND PRENTICE, R.L., *The Statistical Analysis of Failure Time Data*. 2nd Edition, Wiley Series in Probability and Statistics, New York (2002).

[11] KLEINBAUM, D.G., *Survival Analysis*. Springer, New York (1996).

[12] LEE, E.T. AND WANG, J.W., *Statistical Methods for Survival Data Analysis* . Wiley Series in Probability and Statistics, New York (2003).

[13] PENG, Y., YU, B., *Cure Models: Methods, Applications and Implementation*. CRC Press (2021).

[14] R CORE TEAM, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (2020).

[15] SY, J AND TAYLOR, J, *Estimation of a Cox proportional hazards cure model*. Biometrics, 56, 227-236 (2000).

[16] USARRALDE, B., *Algoritmo EM Aplicaciones y Extensiones*. Trabajo de Fin de Grado, Universidad de Zaragoza (2017).

Apéndice A

Propiedades del estimador de Kaplan-Meier

En determinadas ocasiones, resulta más interesante conocer el intervalo de confianza o la varianza de un estimador que la propia estimación realizada. Es por ello que detallaremos en este apartado los pasos necesarios para obtener ambas propiedades del estimador de Kaplan-Meier (1.5). Comenzaremos presentando el llamado **método delta** el cual nos permitirá obtener la estimación de la varianza.

A.1. Método delta

Este método es utilizado para estimar la varianza de un estimador que tiene una expresión compleja. Se basa en el siguiente teorema.

Teorema A.1. *Sea X una variable aleatoria tal que $EX = \mu$ y $VarX = \sigma^2$, sea $f : \mathbb{R} \longrightarrow \mathbb{R}$ diferenciable. Entonces, se verifica que*

$$Var[f(X)] \cong \left[f'(\mu)\right]^2 \sigma^2. \tag{A.1}$$

Demostración. Primero, consideramos el desarrollo de Taylor de primer orden de f en torno a μ

$$f(X) \cong f(\mu) + (X-\mu)f'(\mu).$$

De esta igualdad, se sigue que

$$Var[f(X)] \cong Var\left[f(\mu) + (X-\mu)f'(\mu)\right] = Var\left[(X-\mu)f'(\mu)\right] = \left[f'(\mu)\right]^2 Var(X-\mu) = \left[f'(\mu)\right]^2 \sigma^2.$$

□

Nota. En las condiciones del teorema anterior, sean $\hat{\mu}$ y $\hat{\sigma^2}$ estimadores de μ y σ^2 respectivamente, se tiene que

$$Var[f(X)] \cong \left[f'(\hat{\mu})\right]^2 \hat{\sigma^2}.$$

A.2. Varianza del estimador

Procedemos a calcular un estimador de la varianza del estimador de Kaplan-Meier. Para ello, suponemos que la probabilidad de supervivencia en cada instante t_j corresponde a una Bernouilli de parámetro p_j, siendo cada una de las observaciones independientes.

Teorema A.2. *El estimador de la varianza del estimador de Kaplan-Meier es*

$$\hat{\mathbb{V}ar}\left[\hat{S}(t)\right] \cong \left[\hat{S}(t)\right]^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{A.2}$$

Demostración. Consideraremos el logaritmo del estimador de Kaplan-Meier

$$\log\left(\hat{S}(t)\right) = \sum_{j:t_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right) = \sum_{j:t_j \leq t} \log\left(\hat{p}_j\right), \;\; \hat{p}_j = 1 - \frac{d_j}{n_j}$$

ya que la varianza de la suma tiene un cálculo más sencillo que la varianza del producto. Además $\hat{p}_j$ es el estimador media muestral de p_j. Así, por la independencia asumida se sigue que

$$\mathbb{V}ar\left(\log\left(\hat{S}(t)\right)\right) = \mathbb{V}ar\left(\sum_{j:t_j \leq t} \log\left(\hat{p}_j\right)\right) = \sum_{j:t_j \leq t} \mathbb{V}ar\left(\log\left(\hat{p}_j\right)\right).$$

Aplicamos el método delta para estimar $\mathbb{V}ar\left(\log\left(\hat{p}_j\right)\right)$. En nuestro caso, $f(x) = \log(x)$, $X \equiv \hat{p}_j$ cuyo estimador de media es $\hat{\mu}_X = \hat{p}_j$ y de varianza es $\hat{\sigma_X}^2 = \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}$. Luego,

$$\hat{\mathbb{V}ar}\left(\log(\hat{p}_j)\right) \cong \left(\frac{d}{dx}\log(x)\Big|_{x=\hat{\mu}_X}\right)^2 \hat{\sigma_X}^2 \cong \frac{1}{\hat{\mu}_X^2}\hat{\sigma_X}^2 = \frac{1}{\hat{p}_j^2}\frac{\hat{p}_j(1-\hat{p}_j)}{n_j} = \frac{1-\hat{p}_j}{\hat{p}_j n_j} = \frac{d_j}{n_j(n_j-d_j)}.$$

De donde deducimos que

$$\hat{\mathbb{V}ar}\left(\log\left(\hat{S}(t)\right)\right) = \sum_{j:t_j \leq t} \hat{\mathbb{V}ar}\left(\log\left(\hat{p}_j\right)\right) = \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j-d_j)}. \tag{A.3}$$

Para llegar al resultado buscado, debemos volver a aplicar el método delta. En esta ocasión, $g(x) = \exp(x)$, $Y \equiv \log\left(\hat{S}(t)\right)$ cuyo estimador de media es $\hat{\mu}_Y = \log\left(\hat{S}(t)\right)$ y de varianza $\hat{\sigma_Y^2}$ es (A.3). Luego,

$$\hat{\mathbb{V}ar}\left[\hat{S}(t)\right] = \hat{\mathbb{V}ar}\left\{\exp\left[\log\left(\hat{S}(t)\right)\right]\right\} \cong \left(\frac{d}{dx}\exp(x)\Big|_{x=\hat{\mu}_Y}\right)^2 \hat{\sigma_Y}^2 \cong \left[\exp(\hat{\mu}_Y)\right]^2 \hat{\sigma_Y^2} = \left[\hat{S}(t)\right]^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j-d_j)}.$$

□

A.3. Intervalo de confianza del estimador: Fórmula de Greenwood

Para definir un intervalo de confianza del estimador de Kaplan-Meier lo lógico sería utilizar la expresión (A.2). Sin embargo, el uso de esta aproximación puede suponer que el intervalo de confianza se salga del intervalo $[0,1]$ en el que está definida $\hat{S}(t)$. Para solventar este problema, se define la función $\log\left\{-\log[\hat{S}(t)]\right\}$ llamada *log-log función de supervivencia* cuyo rango va desde menos hasta más infinito. Así, el estimador de la varianza de esta función se calcula partiendo de (A.3) y volviendo a aplicar el método delta

$$\hat{\mathbb{V}ar}\left(\log\left\{-\log[\hat{S}(t)]\right\}\right) \cong \left(\frac{d}{dx}\log(x)\Big|_{x=-\hat{\mu}_Y}\right)^2 \hat{\sigma_Y}^2 \cong \frac{1}{\hat{\mu}_Y^2}\hat{\sigma_Y}^2 = \frac{1}{\left[\log\left(\hat{S}(t)\right)\right]^2}\sum_{j:t_j \leq t}\frac{d_j}{n_j(n_j-d_j)}.$$

Denotando $\hat{SE}\left(\log\left\{-\log[\hat{S}(t)]\right\}\right)$ a la raíz cuadrada positiva de la anterior expresión y $z_{1-\frac{\alpha}{2}}$ al percentil $1-\frac{\alpha}{2}$ de la variable aleatoria normal estándar, se tiene que el intervalo de confianza de significación $0 < \alpha < 1$ de la log-log función de supervivencia es

$$\left(\log\left\{-\log[\hat{S}(t)]\right\} - z_{1-\frac{\alpha}{2}}\hat{SE}\left(\log\left\{-\log[\hat{S}(t)]\right\}\right), \log\left\{-\log[\hat{S}(t)]\right\} + z_{1-\frac{\alpha}{2}}\hat{SE}\left(\log\left\{-\log[\hat{S}(t)]\right\}\right)\right)$$

Identificando $\hat{c}_1$ y $\hat{c}_2$ con el extremo superior e inferior de dicho intervalo respectivamente, se deduce que el intervalo de confianza de significación α del estimador de la función supervivencia de Kaplan-Meier es

$$\left(\exp\left[-\exp(\hat{c}_1)\right], \exp\left[-\exp(\hat{c}_2)\right]\right) \tag{A.4}$$

Esta expresión se denomina **fórmula de Greenwood**. Observar que este intervalo de confianza es válido sólo para los tiempos en los que el estimador de Kaplan-Meier está definido, es decir, en los distintos tiempos de fallo o censura.

Apéndice B

Algoritmo EM

El algoritmo EM (Esperanza-Maximización) es un algoritmo propuesto por Dempster, Laird and Rubin que proporciona un método iterativo para estimar parámetros por máxima verosimilitud cuando no todas las variables han sido observadas. Se puede aplicar en multitud de problemas en los que nos enfrentemos a situaciones como la de los modelos cure (sección 2.3).

Denotemos por $\mathbf{Z} = (z_1, \ldots, z_m)$ el conjunto de datos observados en m realizaciones de un cierto experimento y por $\mathbf{X} = (x_1, \ldots, x_m)$ el conjunto de datos no observados en dichas realizaciones, de manera que $\mathbf{Y} = \mathbf{X} \cup \mathbf{Z}$ será el conjunto completo de datos. Los datos no observados $\mathbf{X}$, se consideran una variable aleatoria cuya distribución depende de los parámetros a estimar y de los datos observados. De hecho, la idea básica del algoritmo consiste en asociar un problema de datos incompletos con otro de datos completos relacionando las verosimilitudes de ambos problemas.

Sea $\boldsymbol{\mu}$ un vector de parámetros desconocidos, si $L(\boldsymbol{\mu}, Y) = f(Y \mid \boldsymbol{\mu})$ es la verosimilitud (función de densidad conjunta) de los datos completos, podemos escribir

$$f(\mathbf{Y} \mid \boldsymbol{\mu}) = f(\mathbf{X, Z} \mid \boldsymbol{\mu}) = f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\, f(\mathbf{Z}, \boldsymbol{\mu}),$$

o equivalentemente

$$\log f(\mathbf{Y} \mid \boldsymbol{\mu}) = \log f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}) + \log f(\mathbf{Z}, \boldsymbol{\mu}).$$

Además, puesto que lo que queremos realizar es optimizar el vector de parámetros $\boldsymbol{\mu}$ respecto a los datos observados (no tenemos otros), despejamos y obtenemos

$$\log f(\mathbf{Z}, \boldsymbol{\mu}) = \log f(\mathbf{Y} \mid \boldsymbol{\mu}) - \log f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}) = \log f(\mathbf{X,Z} \mid \boldsymbol{\mu}) - \log f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}).$$

En particular, podemos sustituir $f(\mathbf{Z} \mid \boldsymbol{\mu})$ por $L(\boldsymbol{\mu}, \mathbf{Z})$ y denotar su logaritmo por $l(\boldsymbol{\mu}, \mathbf{Z})$. Así, tenemos que

$$l(\boldsymbol{\mu}, \mathbf{Z}) = \log f(\mathbf{X,Z} \mid \boldsymbol{\mu}) - \log f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}),$$

Ya que consideramos los datos no observados como variables aleatorias, el algoritmo estima dichos datos calculando su media sobre todos sus posibles valores. Dicho de otro modo, para maximizar la expresión anterior, integraremos sobre todos los posibles valores de $\mathbf{X}$,

$$l(\boldsymbol{\mu}, \mathbf{Z}) = \int \log f(\mathbf{X,Z} \mid \boldsymbol{\mu})\, f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\, d\mathbf{X} - \int \log f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\, f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\, d\mathbf{X}$$

A partir de esta expresión, si denotamos por $\boldsymbol{\mu}^{(k)}$ a la estimación de $\boldsymbol{\mu}$ en el paso k-ésimo, podemos definir las siguientes funciones:

$$Q(\boldsymbol{\mu}, \boldsymbol{\mu}^{(k)}) = E\left[\log f(\mathbf{X,Z}|\boldsymbol{\mu}) \mid \mathbf{Z}, \boldsymbol{\mu}^{(k)}\right] = \int \log f(\mathbf{X,Z} \mid \boldsymbol{\mu})\, f(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}^{(k)})\, d\mathbf{X},$$

$$H(\boldsymbol{\mu},\boldsymbol{\mu}^{(k)}) = E\left[\log f(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu})\ |\mathbf{Z},\boldsymbol{\mu}^{(k)}\right] = \int \log f(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu})\ f(\mathbf{X}\mid\mathbf{Z},\boldsymbol{\mu}^{(k)})\ d\mathbf{X}.$$

Dichas funciones verifican que

$$l(\boldsymbol{\mu},\mathbf{Z}) = Q(\boldsymbol{\mu},\boldsymbol{\mu}^{(k)}) - H(\boldsymbol{\mu},\boldsymbol{\mu}^{(k)}).$$

Ahora ya estamos en condiciones de describir el algoritmo.

ALGORITMO

1. Elíjase un valor arbitrario $\boldsymbol{\mu}^{(0)}$ como valor inicial de los parámetros.
2. Repítanse los dos siguientes pasos hasta detectar convergencia mediante un criterio de parada:
   a) **Paso E:** Calcúlese $Q(\boldsymbol{\mu},\boldsymbol{\mu}^{(k)})$.
   b) **Paso M:** Maximícese la función $Q(\boldsymbol{\mu},\boldsymbol{\mu}^{(k)})$ respecto de $\boldsymbol{\mu}$, obteniéndose $\boldsymbol{\mu}^{(k+1)}$.

Otras aplicaciones del algoritmo, así como algunas extensiones del mismo pueden verse en [16].

Apéndice C

Material complementario al estudio

C.1. Código en R

Comenzamos cargando los paquetes estadísticos requeridos y el conjunto de datos.

```
library(ggplot2)
library(survival)
library(survminer)
library(dplyr)
library(ggalt)
library(smcure)
load("~/Desktop/TFG/datos_covid_gerardo.Rdata")
```

Ante la irregularidad de los primeros meses de pandemia, consideraremos los datos tomados a partir del 1 de Julio de 2020. Además, se disponen datos hasta el 23 de Marzo de 2021 y como vamos a analizar ingresos en hospital y UCI, para evitar un elevado número de observaciones censuradas vamos a tomar los datos correspondientes con fecha de test positivo anterior al 23 de Febrero de 2021 puesto que según criterio médico, un ingreso en hospital producido un mes después de dar positivo es por causas ajenas a la covid.

```
datos_todos<-datos_est[as.Date(datos_est$fecha_referencia)>=
as.Date("2020-07-01"),]
datos_todos<-datos_todos[as.Date(datos_todos$fecha_referencia)<=
as.Date("2021-02-23"),]
```

Nuestro primer objetivo es analizar la estancia en UCI de los pacientes covid que han precisado dicho servicio. Tomamos además las variables que vamos a considerar: Id, Sexo, edad, fecha de ingreso, fecha de alta y grupoGMA.

```
datosUCI<-datos_todos[datos_todos$UCI=="UCI_Sí",c(53,70,71,56,
38,33)]
```

Eliminamos los datos que tienen valores nulos en cualquiera de las variables anteriores. Así,

```
datosUCI<-na.omit(datosUCI)
```

Creamos la variable evento: “Recibir alta en UCI” que nos va a permitir realizar el análisis de supervivencia. Por defecto, en la base de datos aparecen con fecha de alta “2021-03-23” aquellos individuos censurados, es decir, aquellos que seguían ingresados en UCI sin recibir el alta el 23 de Marzo.

```
datosUCI$evento<-rep(1, nrow(datosUCI))
datosUCI[datosUCI$dis_ICU=="2021-03-23",]$evento<-rep(0,
nrow(datosUCI[datosUCI$dis_ICU=="2021-03-23",]))
```

La variable edad la vamos a convertir en una variable de tipo factor. Así veamos sus cuantiles para dividirla en grupos.

```
summary(datosUCI$edad)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   57.00   65.00   63.29   72.00   86.00
```

Una posibilidad sería dividir a los individuos en 4 grupos: menores de 57, 58-65 años, 66-72 años mayores de 72 años. No obstante, los dos grupos intermedios tienen, a priori, un análisis similar. Luego, dividimos en tres grupos: menores de 57, 58-72 años y mayores de 72 años.

```
datosUCI<-within(datosUCI, edadF<- factor(edad, levels=seq(7,86,1),
labels=c(rep("Menor de 58 años años",51),rep("58-72 años",15),
rep("Mayor de 72 años",14))))
```

Para poder aplicar los paquetes estadísticos de supervivencia debemos eliminar los datos que tienen misma fecha de entrada y salida de UCI. Además, creamos la variable tiempoestancia que recoge los días que el paciente ha estado ingresado en UCI.

```
datosUCI<-datosUCI[datosUCI$adm_ICU<datosUCI$dis_ICU,]
datosUCI$tiempoestancia<-as.numeric(datosUCI$dis_ICU-datosUCI$
adm_ICU)
```

Con todo ello,

```
head(datosUCI[,c(9,7,8,6,5)])

##      tiempoestancia evento            edadF grupo_GMA   sexo
## 314              71      1       58-72 años        33 HOMBRE
## 676              71      1       58-72 años        33 HOMBRE
## 677              24      1 Mayor de 72 años        33 HOMBRE
## 1130             16      1       58-72 años        32 HOMBRE
## 1246             21      1       58-72 años        32 HOMBRE
## 1317              6      1       58-72 años        32  MUJER
```

Comenzamos viendo un gráfico (figura C.1) sobre los datos que tenemos.

```
gantt_chart_sa <- ggplot(data    = datosUCI ,
                     mapping = aes(x     = adm_ICU,
                                   xend  = dis_ICU,
                                   y     = id_num,
                                   color = factor(evento))) +
  geom_dumbbell(size_x = 1, size_xend = 1)+
  labs(x = "Meses", y = "Número_ID",
       title = " Entrada y salida de la UCI de pacientes covid",
       color = "Evento") +
  theme(axis.text.y = element_blank())+theme_minimal()
```

Para verlo más claro, nos centramos (por ejemplo) en los pacientes de 70 años (figura C.2).

```
gantt_chart_sa <- ggplot(data    = datosUCI %>%
                           filter(edad == 70),
                     mapping = aes(x     = adm_ICU,
                                   xend  = dis_ICU,
                                   y     = id_num,
                                   color = factor(evento)), , break.y.by = 100)
                                    + geom_dumbbell(size_x = 2, size_xend = 2)+
  labs(x = "Meses", y = "Número_ID",
       title = " Entrada y salida de la UCI en los pacientes de 70 años",
       color = "Evento") +
  theme(axis.text.y = element_blank())+
  geom_text(aes(label = adm_ICU), vjust = +1.3, size=2.5) +
  geom_text(aes(x = dis_ICU, label = dis_ICU), vjust = -0.7, size=2.5)
  +theme_minimal()
```

Veamos el estimador Kaplan-Meier del conjunto de los individuos (figura C.3).

```
Survfit <- survfit(Surv(tiempoestancia, evento) ~ 1,  conf.type = "log" ,
conf.int=0.95, type="kaplan-meier", error="greenwood", data=datosUCI)
Survfit

## Call: survfit(formula = Surv(tiempoestancia, evento) ~ 1, data = datosUCI,
##     error = "greenwood", conf.type = "log", conf.int = 0.95,
##     type = "kaplan-meier")
##
##       n  events  median 0.95LCL 0.95UCL
##     672     646      18      16      19

quantile(Survfit, quantiles=c(.25,.5,.75))

## $quantile
## 25 50 75
##  9 18 32
##
## $lower
## 25 50 75
##  8 16 29
##
## $upper
## 25 50 75
## 10 19 36


ggsurvplot(Survfit,risk.table=T,legend.title ='', break.x.by = 20,break.y.by
= 0.2) + labs(title= " Estimación de supervivencia de Kaplan-Meier")
```

Si analizamos por Sexo (figura C.4),

```
Survfit2 <- survfit(Surv(tiempoestancia, evento) ~ sexo,  conf.type = "log" ,
conf.int=0.95, type="kaplan-meier", error="greenwood", data=datosUCI)
Survfit2
```

```
## Call: survfit(formula = Surv(tiempoestancia, evento) ~ sexo, data = datosUCI,
##     error = "greenwood", conf.type = "log", conf.int = 0.95,
##     type = "kaplan-meier")
##
##                n events median 0.95LCL 0.95UCL
## sexo=HOMBRE 451    434     18      16      20
## sexo=MUJER  221    212     17      14      19

quantile(Survfit2, quantiles=c(.25,.5,.75))

## $quantile
##             25 50 75
## sexo=HOMBRE  9 18 32
## sexo=MUJER   9 17 28
##
## $lower
##             25 50 75
## sexo=HOMBRE  8 16 29
## sexo=MUJER   8 14 25
##
## $upper
##             25 50 75
## sexo=HOMBRE 11 20 38
## sexo=MUJER  11 19 37


ggsurvplot(Survfit2, conf.int = T,
           risk.table=T,
           legend.title = "",
           break.x.by = 20,
           legend.labs = c("Hombre",
                           " Mujer"),
           data = datosUCI) +
  labs(title    = " Estimaciones de supervivencia de Kaplan-Meier por Sexo")
```

Veamos el test de Log-Rank.

```
survdiff(Surv(tiempoestancia, evento) ~ sexo, data=datosUCI)

## Call:
## survdiff(formula = Surv(tiempoestancia, evento) ~ sexo, data = datosUCI)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## sexo=HOMBRE 451      434      440    0.0842     0.277
## sexo=MUJER  221      212      206    0.1799     0.277
##
##  Chisq= 0.3  on 1 degrees of freedom, p= 0.6
```

Si analizamos por edadF (figura C.5),

```
Survfit3 <- survfit(Surv(tiempoestancia, evento) ~ edadF,  conf.type = "log" ,
conf.int=0.95, type="kaplan-meier", error="greenwood", data=datosUCI)
Survfit3

## Call: survfit(formula = Surv(tiempoestancia, evento) ~ edadF, data = datosUCI,
##     error = "greenwood", conf.type = "log", conf.int = 0.95,
##     type = "kaplan-meier")
##
##                               n events median 0.95LCL 0.95UCL
## edadF=Menor de 58 años años 175    168     12      11      16
## edadF=58-72 años            338    326     19      17      22
## edadF=Mayor de 72 años      159    152     18      16      22

quantile(Survfit3)

## $quantile
##                               25 50 75
## edadF=Menor de 58 años años  7 12 29
## edadF=58-72 años            11 19 37
## edadF=Mayor de 72 años      10 18 26
##
## $lower
##                               25 50 75
## edadF=Menor de 58 años años  6 11 26
## edadF=58-72 años            10 17 32
## edadF=Mayor de 72 años       8 16 24
##
## $upper
##                               25 50 75
## edadF=Menor de 58 años años  8 16 38
## edadF=58-72 años            12 22 41
## edadF=Mayor de 72 años      12 22 31


ggsurvplot(Survfit3, conf.int = F,
           risk.table=F,
           legend.title = "",
           break.x.by = 20,
           legend.labs = c("Menor de 58",
                           " Entre 58-72", "Mayor de 72"),
           data = datosUCI) +
  labs(title    = " Estimaciones de supervivencia de Kaplan-Meier por Edad")
```

Y el test de Log-Rank

```
survdiff(Surv(tiempoestancia, evento) ~ edadF, data=datosUCI)

## Call:
## survdiff(formula = Surv(tiempoestancia, evento) ~ edadF, data = datosUCI)
##
```

```
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## edadF=Menor de 58 años años 175      168      157     0.772      1.08
## edadF=58-72 años            338      326      356     2.568      6.01
## edadF=Mayor de 72 años      159      152      133     2.788      3.72
##
##  Chisq= 6.4  on 2 degrees of freedom, p= 0.04
```

Veamos los grupos 31, 32 y 33 de la variable grupoGMA (figura C.6),

```
datosUCIGMA<-rbind(datosUCI[datosUCI$grupo_GMA==31,],
datosUCI[datosUCI$grupo_GMA==32,],datosUCI[datosUCI$grupo_GMA==33,])
```

```
Survfit4 <- survfit(Surv(tiempoestancia, evento) ~ grupo_GMA,  conf.type =
"log" ,conf.int=0.95, type="kaplan-meier", error="greenwood", data=datosUCIGMA)


ggsurvplot(Survfit4, conf.int = F,
           risk.table=F,
           legend.title = "",
           break.x.by = 20,
           legend.labs = c("31", "32", "33"),
           data = datosUCI) + labs(title    =
            " Estimaciones de supervivencia de Kaplan-Meier por grupos GMA")
```

Y el test de Logrank es

```
survdiff(Surv(tiempoestancia, evento) ~ grupo_GMA, data=datosUCIGMA)

## Call:
## survdiff(formula = Surv(tiempoestancia, evento) ~ grupo_GMA,
##     data = datosUCIGMA)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## grupo_GMA=31  79       74     70.7    0.1577    0.1880
## grupo_GMA=32 246      239    236.9    0.0181    0.0319
## grupo_GMA=33 289      277    282.4    0.1037    0.2090
##
##  Chisq= 0.3  on 2 degrees of freedom, p= 0.9
```

Vamos a agrupar los grupos de GMA.

```
datosUCI$grupo_GMAF <- factor(x = datosUCI$grupo_GMA, levels = c("0",
"10", "20", "31", "32", "33", "40"), labels = c("Población sana",
"Embarazo o patología","Embarazo o patología", "Enfermedad(es) crónica(s)",
"Enfermedad(es) crónica(s)" , "Enfermedad(es) crónica(s)" ,"Neoplasis" ))
```

Finalmente, el modelo de Cox con las covariables EdadF y GrupoGMAF. Convertimos la variable sexo a tipo factor.

```
datosUCI$sexo <- as.factor(datosUCI$sexo)
CoxModel<-coxph(Surv(tiempoestancia, evento) ~ edadF+grupo_GMAF, method="efron"
,datosUCI)
summary(CoxModel)
```

```
mfrow <- par(mfrow=mfrow(2))
termplot(CoxModel, ask=FALSE)
par(mfrow)
```

Pasamos ahora a la segunda parte del estudio.

```
datosHOSP<-datos_todos[datos_todos$hospitalizado=="Hosp_Sí",c(53,67,69,70,56,38)]
```

Creamos la variable evento. Eliminamos todos los datos que tengan valores nulos en algunas de las variables excepto en la de admisión UCI (puesto que no todos los ingresados han requerido de UCI).

```
datosHOSP<-datosHOSP[!is.na(datosHOSP$UCI)&!is.na(datosHOSP$adm_hos)
&!is.na(datosHOSP$edad)&!is.na(datosHOSP$sexo),]
datosHOSP$evento<-rep(0, nrow(datosHOSP))
datosHOSP[datosHOSP$UCI=="UCI_Sí",]$evento<-rep(1,
nrow(datosHOSP[datosHOSP$UCI=="UCI_Sí",]))
```

La variable edad la vamos a dividir en grupos para convertirla en tipo factor.

```
summary(datosHOSP$edad)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    57.0    72.0    68.8    84.0   106.0
```

Así por la información de los cuantiles y el perfil que hemos visto en el análisis anterior de que los pacientes que ingresan en UCI tienen entre 50 y 84 años, dividimos en tres grupos: menores de 57 años, 58-84 años y mayores de 84 años.

```
datosHOSP<-within(datosHOSP, edadF<- factor(edad, levels=seq(0,106,1),labels=
c(rep("Menor de 58 años",58),rep("58-84 años",27),rep("Mayor de 84 años",22))))
```

Creamos la variable días que recoge los días que pasan desde el ingreso en el hospital hasta que el paciente ingresa en UCI.

```
datosHOSP$dias<-as.numeric(datosHOSP$adm_ICU-datosHOSP$adm_hos)
datosHOSP[is.na(datosHOSP$dias),]$dias<-as.numeric(as.Date("2021-03-23")-
datosHOSP[is.na(datosHOSP$dias),]$adm_hos)
```

Veamos el estimador de Kaplan-Meier (figura C.7),

```
Survfit6 <- survfit(Surv(dias, evento) ~ 1,  conf.type = "log" ,conf.int=0.95,
 type="kaplan-meier", error="greenwood", data=datosHOSP)
ggsurvplot(Survfit6, conf.int = T,
           risk.table=T,
           legend.title = "",
           break.x.by = 20,
           data = datosHOSP) +
  labs(title    = " Estimaciones de supervivencia de Kaplan-Meier")
```

Para definir los individuos curados, veamos cual ha sido la persona con mayor diferencia entre el tiempo de ingreso en hospital y UCI.

```
max(datosHOSP$adm_ICU-datosHOSP$adm_hos, na.rm=T)

## Time difference of 39 days
```

Así, definimos como individuo curado toda persona ingresada en hospital antes del 13 de Febrero de 2020 y que no ha precisado de ingreso en UCI. Dicho de otro modo, todos los hospitalizados con fecha de ingreso posterior al 13 de Febrero no podemos asegurar si son curados o susceptibles censurados.

```
datosHOSP$curado<-rep(1, nrow(datosHOSP))
datosHOSP[datosHOSP$UCI=="UCI_No" & datosHOSP$adm_hos<as.Date("2021-02-13"),]$
curado <-rep(0,nrow(datosHOSP[datosHOSP$UCI=="UCI_No" & datosHOSP$adm_hos<
as.Date("2021-02-13"),]))
```

Los datos con los que vamos a trabajar son:

```
head(datosHOSP[,c(1,9,7,10,8,6)])

##    id_num dias evento curado            edadF   sexo
## 9       9  148      0      0 Mayor de 84 años  MUJER
## 23     23  109      0      0 Menor de 58 años HOMBRE
## 25     25   55      0      0       58-84 años HOMBRE
## 39     39  186      0      0       58-84 años  MUJER
## 46     46  111      0      0 Menor de 58 años  MUJER
## 74     74   42      0      0 Menor de 58 años  MUJER
```

Implementamos la función cure de Miguel Lafuente Blasco,

```
S_NPCPK_to_Survival <- function (data=data) {

      N = nrow(data)

      #Order data
      data.ot <- data[order(data[, 1]), ]
      t = data.ot[,1]
      d = data.ot[,2]
      nu = data.ot[,3]
      cum.nu = cumsum(nu)

      S = rep(1, N)

      #Estimator of survival function
      for (i in 2:N) {
        if(d[i]==0) {S[i] = S[i-1]}
        if(d[i]==1) {S[i]  = S[i-1] * (1 - 1/(N - i + 1 + cum.nu[i-1]))}}

      #Probability of experimenting the event
      p  =  1- min(S)

      #Survival of the time until event
```

```
      survivalF=(S-(1-p))/p
      times=t

      #Extract indexes of the first appeareances of the integers times to build
      the survival function table
      index_for_S=sapply(unique(times), function(x) min(which(times==x)))
      result=cbind(times[index_for_S],survivalF[index_for_S])

      #Discard the out of support points in the survival function
      index_end_support=which(result[,2]==0)
      if(length(index_end_support)==0){
        index_end_support=which(result[,1]==max(t[which(d==1)]))
      } else{
        index_end_support=min(index_end_support)-1
      }
      result=result[1:index_end_support,]

      #To transform P(T>=t) to P(T>t)
      result[,2]=c(result[2:length(result[,2]),2],0)

      # #To delete the no changing probability points (in order to
      interpolate later)
      rr=result[,2]
      indexes_bool=c(F,rr[-1]==rr[-length(rr)])
      result=result[which(indexes_bool==F),]

      return(list(S, p, t,result))
}
```

Así, la incidencia estimada es

```
mcure<-S_NPCPK_to_Survival(datosHOSP[,c(9,7,10)])
mcure[[2]]

## [1] 0.06869019
```

y la latencia puede ser representada (figura C.8),

```
plot(mcure[[4]][,1],mcure[[4]][,2],type = "l", main="Latencia", xlab="Días",
ylab ="Probabilidad")
```

Para comparar entre ambos sexos vamos a aplicar la función a ambos grupos.

```
datosHOSPH<-datosHOSP[datosHOSP$sexo=="HOMBRE",c(9,7,10)]
datosHOSPM<-datosHOSP[datosHOSP$sexo=="MUJER",c(9,7,10)]
```

Las incidencias de ambos grupos

```
mcureH<-S_NPCPK_to_Survival(datosHOSPH)
mcureM<-S_NPCPK_to_Survival(datosHOSPM)
c(mcureH[[2]],mcureM[[2]])

## [1] 0.08708314 0.04742958
```

Veamos las correspondientes latencias (figura C.9),

```
plot(mcureM[[4]][,1],mcureM[[4]][,2],type = "l", main="Latencia según Sexo",
xlab="Días", ylab ="Probabilidad", col="red")
+ lines(mcureH[[4]][,1],mcureH[[4]][,2], col="blue")
legend("right", legend = c("Mujer", "Hombre"),
       lwd = 3, col = c("red", "blue"))
```

Si repetimos el razonamiento con la covariable EdadF.

```
datosHOSPE1<-datosHOSP[datosHOSP$edadF=="Menor de 58 años",c(9,7,10)]
datosHOSPE2<-datosHOSP[datosHOSP$edadF=="58-84 años",c(9,7,10)]
datosHOSPE3<-datosHOSP[datosHOSP$edadF=="Mayor de 84 años",c(9,7,10)]
```

Primero las incidencias.

```
mcureE1<-S_NPCPK_to_Survival(datosHOSPE1)
mcureE2<-S_NPCPK_to_Survival(datosHOSPE2)
mcureE3<-S_NPCPK_to_Survival(datosHOSPE3)
c(mcureE1[[2]],mcureE2[[2]],mcureE3[[2]])

## [1] 0.066940216 0.101008735 0.001248439
```

Veamos las latencias para los tres grupos (figura C.10),

```
plot(mcureE2[[4]][,1],mcureE2[[4]][,2],type = "l", main="Latencia según grupo de
edad", xlab="Días", ylab ="Probabilidad", col="red")
 + lines(mcureE1[[4]][,1],mcureE1[[4]][,2], col="blue")
 + lines(mcureE3[[4]][,1],mcureE3[[4]][,2], col="green")
legend("right", legend = c("Menor de 58 años", "58-84 años", "Mayor de 84 años"),
       lwd = 3, col = c("blue","red", "green"))
```

C.2. Gráficos

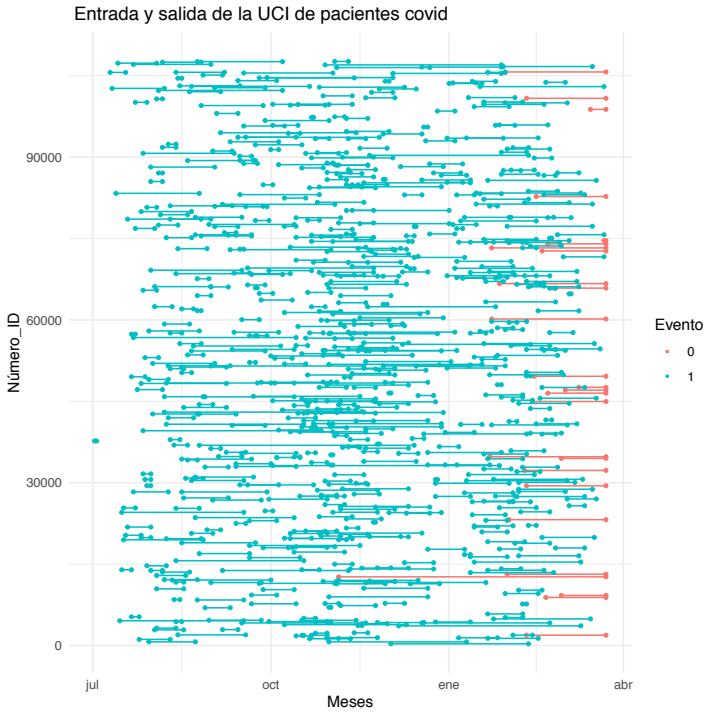


Figura C.1: Observaciones y censura en el evento “dar de alta en UCI”.

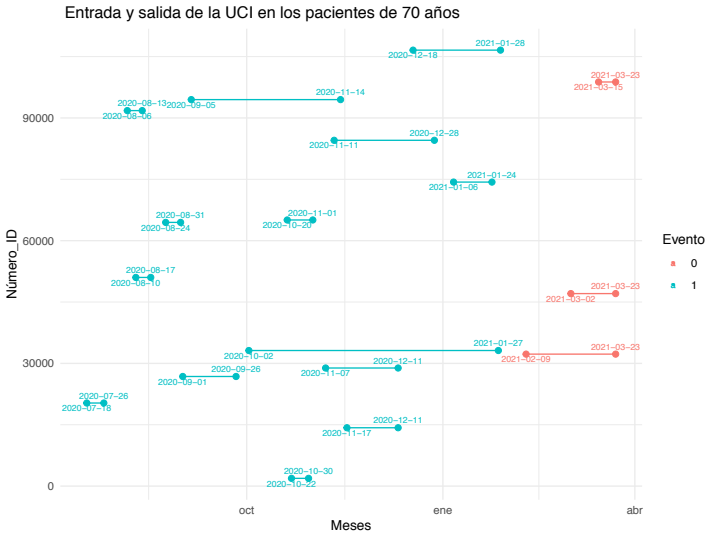


Figura C.2: Observaciones y censura en el evento “dar de alta en UCI” para los pacientes de 70 años.

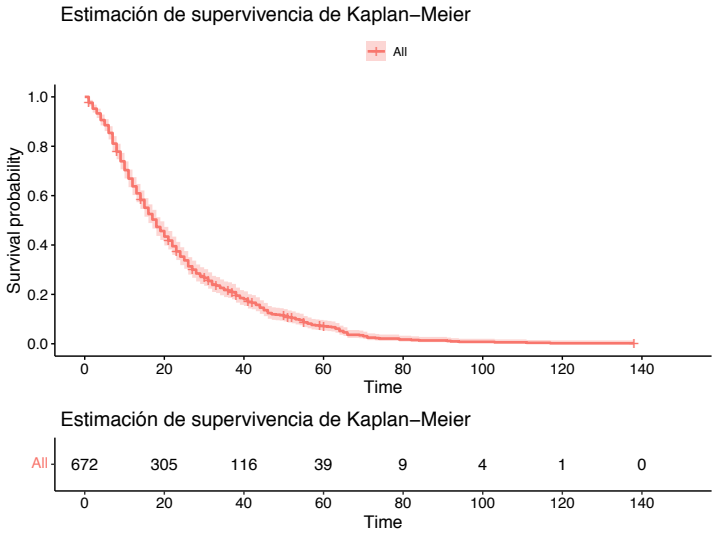


Figura C.3: Estimador de Kaplan-Meier para el evento “dar de alta en UCI”.

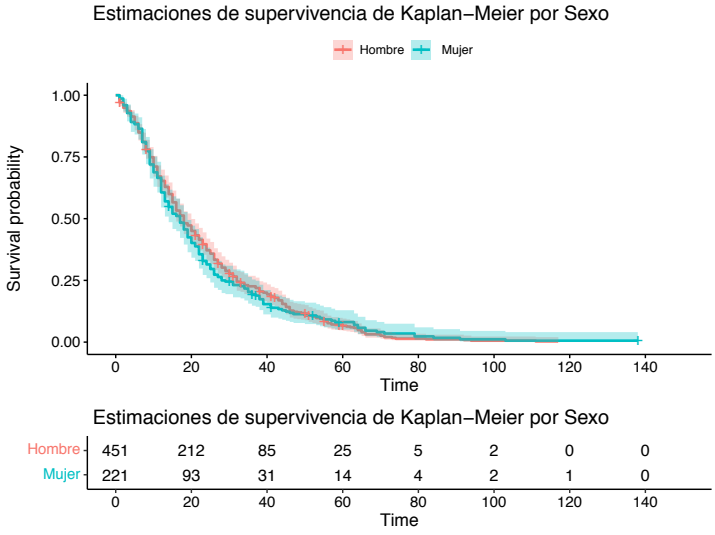


Figura C.4: Estimador de Kaplan-Meier según *sexo* para el evento “dar de alta en UCI”.

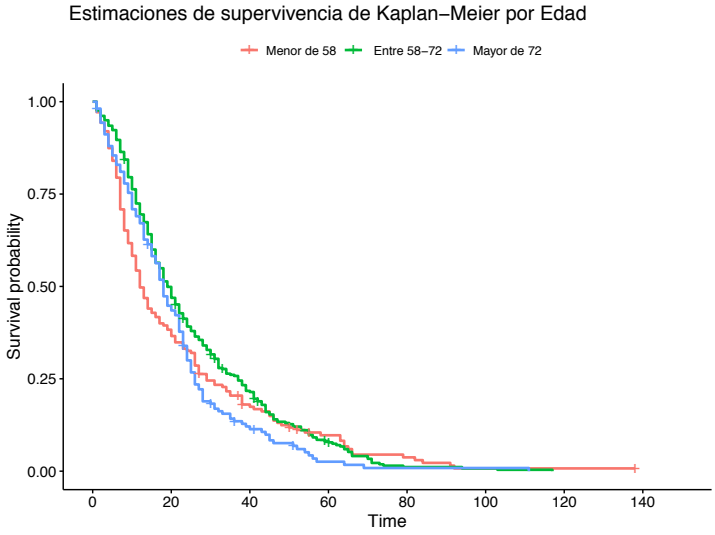


Figura C.5: Estimador de Kaplan-Meier según *edadF* para el evento “dar de alta en UCI”.

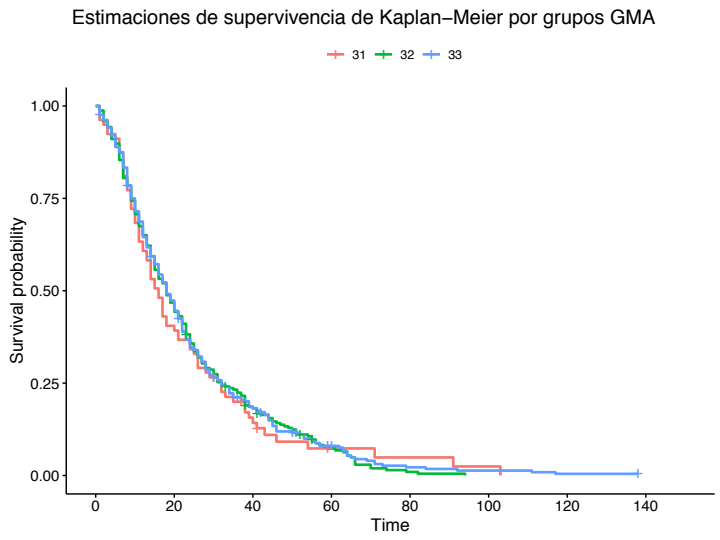


Figura C.6: Estimador de Kaplan-Meier entre los grupos 31, 32 y 33 de la covariable *GMA*.

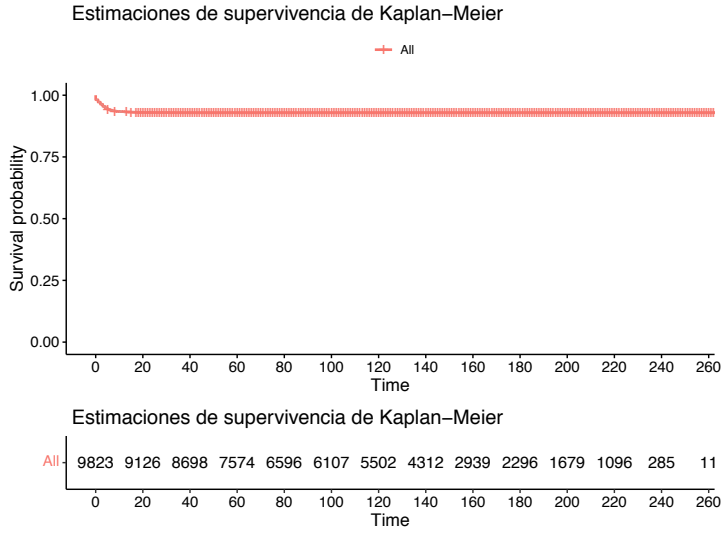


Figura C.7: Estimador de Kaplan-Meier para el evento “ingresar en UCI”.

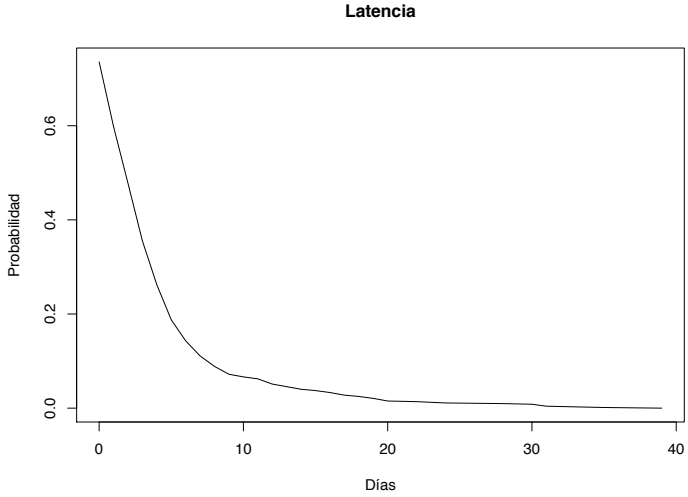


Figura C.8: Estimación de la latencia del conjunto de la población.

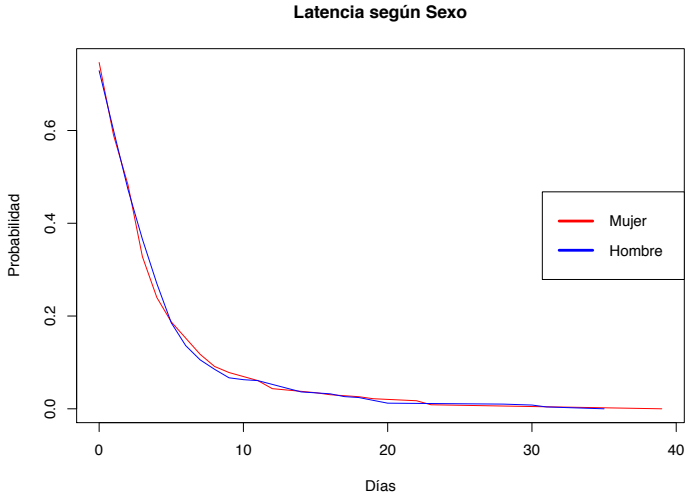


Figura C.9: Estimación de la latencia según *sexo*.

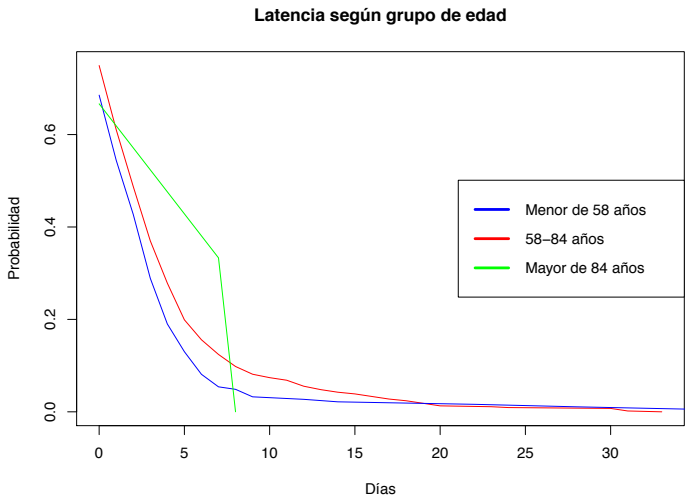


Figura C.10: Estimación de la latencia según *edadF*.