Análisis de Supervivencia.
Modelos de riesgos competitivos.

Mª Elisa Lafuerza Coronas

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: Francisco Javier López Lorente
28 de Junio de 2021

Prólogo

El análisis de supervivencia es una herramienta fundamental en estudios de seguimiento de individuos o sistemas mecánicos. En las ciencias de la salud, en la búsqueda del tratamiento o estudio de determinadas enfermedades o procesos clínicos, resulta de utilidad el estudio de los datos obtenidos. Existen técnicas y programas estadísticos avanzados que permiten realizar este tipo de estudios.

Desde que David Cox propuso en 1972 el modelo de regresión para riesgos proporcionales, el análisis de supervivencia ha ido incorporando nuevas técnicas para su estudio. Hay diferentes métodos tradicionales propuestos para este tipo de análisis estadístico, en el que únicamente se tiene en cuenta la posibilidad de un suceso de interés, tales como el método de Kaplan-Meier.
Sin embargo, en la realidad normalmente cabe la posibilidad de que más de un suceso de interés tenga lugar, hablamos entonces de riesgos competitivos. En cuyo caso, no podemos usar este tipo de métodos. Proporcionaremos entonces una extensión del modelo de riesgos proporcionales de Cox.

El objetivo de este TFG es el desarrollo de la teoría del análisis de supervivencia para así llegar a un modelo que estime la probabilidad de supervivencia en presencia de riesgos competitivos. Aplicaremos los resultados obtenidos a un conjunto de datos reales, donde estudiaremos el tiempo en UCI de pacientes diagnosticados por COVID-19. Además, veremos el efecto que tienen sobre la evolución de los pacientes variables tales como el sexo, la edad y la hipertensión arterial.
Los principales libros en los que está basado este trabajo y de los cuales he recabado gran parte de la información han sido: *Statistical Models and Methods for Lifetime Data* [7] y *Survival Analysis: A Self-Learning Text* [6]. Además de algunos artículos tales como *Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risks analysis* [10] y *Survival analysis in the presence of competing risks* [15] para la aplicación a un conjunto de datos reales con el entorno de programación R.

Abstract

Survival analysis consists of a set of statistical tools and methods that study the time until a certain event takes place, and whose objective is to estimate the survival function. It was originated in medicine with the study of mortality tables. Later, it was used to study the duration and reliability of different materials or machine elements. Other examples of its application are the departure of employees from a company or the end of the commercial relationship established between a company and a customer.

One of the drawbacks of survival analysis is that it has to deal with incomplete data. Once the study is completed, some individuals have experienced the event of interest and others have not, so that the survival time of those who have not experienced the event of interest is unknown, which is when we speak of censored data. The inclusion of the information provided by these data characterizes the survival analysis. In addition, it is common that the different observations do not have the same starting time, or the loss of track of some of them throughout the study.

The concept of event is fundamental in survival analysis. In fact, we are only interested in two data in this type of study; the follow-up time of the individual and a variable that indicates whether the time is a lifetime or censored time. Although the studies have a specific start and end date, we are interested in knowing the time at which the individual joins the study. In addition, the last observation is important; the time at which the individual leaves or is lost to follow-up, or experiences the event of interest. We have a censored data if the study ends without the individual having experienced the event of interest. We can call the period marked by these two times, in which follow-up begins and ends for each individual, the risk period. During this period the individual is at risk of experiencing the event.

Given a random variable T>0 that measures the time until the occurrence of the event, some relevant functions are worth noting. One is the survival function which provides the probability that an individual survives until time t given that he has not yet experienced the event. And the other is the hazard function which provides the probability that an individual will experience the event at time t at which he is being observed and given that he has survived until t.

The probability of survival can be estimated nonparametrically by the Kaplan-Meier method, taking into account both censored and uncensored observation times. It assumes for the censored ones that they would have behaved in the same way as those followed until the event occurred.

Sometimes it is interesting to include information about each individual, which can affect his or her survival time. In survival analysis, the Cox regression model or proportional hazards model is worth mentioning. This is a set of models used to estimate the relationship between a set of explanatory variables and survival time or the hazard function. It is based on the main hypothesis that risks are proportional.

In clinical or biomedical research it is very common to find lifetime data associated with multiple events or causes of failure. The events that make difficult or modify the probability of experiencing the cause of the event are called competing risks. In this type of study, we are interested in both the time, T>0, and the cause of failure, $V \in [1, ..., k]$.

We can specify the joint distribution of the pair (T,V) through two functions. The hazard function for a given cause, which provides the instantaneous risk of experiencing the event of interest for that cause for an individual taking into account the set of individuals at risk at that given time. On the other hand, we have the cumulative incidence function, CIF, which provides the probability of experiencing the event of interest before a certain time and before any other competing risk takes place.

Thus, survival analysis is a very appropriate technique for follow-up studies characterized by a variable duration of follow-up of individuals in the presence of censored data.

Notación

- T variable de tiempo de supervivencia.
- t valor concreto de la variable T.
- f función de masa de probabilidad de la variable T.
- F función de distribución de la variable T.
- S función de supervivencia de la variable T.
- h función de riesgo de la variable T.
- H función de riesgo acumulada de la variable T.
- $j = 0, 1, .., \infty$ tiempos discretos.
- i=1,..,n individuos.
- T_i tiempo de supervivencia para el individuo i.
- C censura.
- C_i tiempo de censura para el individuo i.
- $Z_i = min(T_i, C_i)$ tiempo observado para el individuo i.
- δ_i variable indicadora de censura para el individuo i.
- L función de verosimilitud de la muestra.
- f_i función de masa de probabilidad para el individuo i.
- S_i función de supervivencia para el individuo i.
- h_i función de riesgo para el individuo i.
- $Y_i(t) = I(T_i \geq t, \quad C_i \geq t)$ el individuo i no ha fallado ni ha sido censurado antes del tiempo t.
- $N_i(t) = Y_i(t)I(T_i = t)$ el individuo i falla en el tiempo t sin haber sido censurado antes.
- $\hat{S}$ estimador de la función de supervivencia.
- n(t) número de individuos que están en riesgo en el tiempo t.
- d(t) número de sucesos que tienen lugar en t.
- $\hat{h}$ estimador de la función de riesgo.
- $\tau = max(j, \quad n(j) > 0)$.
- $\hat{H}$ estimador de la función de riesgo acumulada.

- $\mathbf{x}$ vector de dimensión px1 de variables explicativas o covariables, fijas.
- h_0 función de riesgo basal en el modelo de riesgos proporcionales.
- β vector de dimensión px1 formado por los coeficientes de regresión.
- $h(t|\mathbf{x}) = h_0(t)exp(\beta'\mathbf{x})$ función de riesgo para T dado $\mathbf{x}$ en el modelo de reisgos proporcionales.
- S_0 función de supervivencia basal en el modelo de reisgos proporcionales.
- $S(t|\mathbf{x}) = [S_0(t)]^{exp(\beta'\mathbf{x})}$ función de supervivencia para T dado $\mathbf{x}$ en el modelo de reisgos proporcionales.
- $R_i = R(t_i)$ el conjunto de individuos en riesgo justo antes de t_i.
- $\mathbf{x}_i$ covariable asociada al individuo i.
- $\hat{\beta}$ estimador de β.
- $\hat{H}_0$ estimador para H_0.
- $\hat{h}_0$ estimador para h_0.
- HR razón de riesgos.
- U vector Score.
- I matrix de información de Fisher.
- W estadístico Wald.
- $V \in \{1,...,m\}$ variable que indica el modo de fallo o causa.
- h_k función de riesgo específico para la k-ésima causa.
- H_k función de riesgo acumulada para la k-ésima causa.
- S_k función de supervivencia para la k-ésima causa.
- F_k función de incidencia acumulada, CIF.
- f_k subfunción de masa de probabilidad para la k-ésima causa.
- π_k distribución de probabilidad de la k-ésima causa.
- V_i modo de fallo del individuo i.
- $W_i = \delta_i V_i$ variable que proporciona el modo de fallo para aquellos individuos que no son censurados y 0 para los censurados.
- $\delta_{ik} = I(V_i = k)$.
- d_k número de individuos que falla en un tiempo por la causa k.
- h_{0k} función de riesgo basal asociada a la k-ésima causa en el modelo de riesgos proporcionales.
- $d_k(t)$ número de sucesos que tienen lugar en t por la causa k.

Índice general

Capítulo 1

Introducción al análisis de supervivencia.

Llamamos análisis de supervivencia al conjunto de técnicas estadísticas que permiten estudiar la variable "tiempo hasta que ocurre un determinado suceso". Normalmente, llamamos a este tiempo, tiempo de supervivencia, aunque no siempre tiene por qué ser la muerte tal suceso de interés.

En ciertas ocasiones resulta difícil determinar cuándo tiene lugar el suceso. Por ejemplo, el caso de la aparición de un tumor. El principal problema es especificar modelos que representen las distribuciones de tiempo de vida y poder hacer inferencias basadas en estos modelos.

1.1. Tiempo de supervivencia.

En el análisis de supervivencia, la variable temporal que consideramos es el tiempo de supervivencia, que es el tiempo que un individuo ha sobrevivido durante un periodo de seguimiento. Vamos a denotar esta variable como T, donde T toma valor no negativo. Puede medirse en años, meses, días... , y puede modelarse como una variable continua o discreta.

1.2. Distribuciones de tiempo de vida.

Como acabamos de mencionar, T denota la variable aleatoria no negativa que representa los tiempos de vida de los individuos de una población. En ese caso, t será cualquier valor concreto de esa variable. Si T es una variable aleatoria absolutamente continua, denotamos por f(t) a la función de densidad de T, de manera que, $F(t) = P(T \leq t) = \int_0^t f(x)\,dx$ será su función de distribución.

La **función de supervivencia, S(t)**, es la probabilidad de que un individuo tenga un tiempo de supervivencia, T, mayor o igual a un determinado tiempo t: $S(t) = P(T \geq t) = \int_t^\infty f(x)\,dx$. Notar que t toma valores en [0, ∞). La función de supervivencia es una función continua monótona no creciente. Toma valor 1 para t=0, ya que el tiempo de supervivencia es no negativo. Por otra parte, $\lim_{t\to+\infty} S(t) = 0$, es decir, si el estudio perdurase en el tiempo ningún individuo sobreviviría para un tiempo suficientemente grande, y en tal caso, la probabilidad de supervivencia sería 0.

Se define **la función de riesgo, h(t)**, como $h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t}$. La función de riesgo especifica la tasa instantánea de ocurrencia del evento en el momento t, dado que el individuo ha sobrevivido hasta t. Además, notar que $\Delta t h(t)$ representa la probabilidad de que un individuo falle en el intervalo $[t, t+\Delta t)$ dado que ha sobrevivido hasta t. Es más bien una tasa, ya que representa la probabilidad por unidad de tiempo y no toma valores entre 0 y 1 sino entre 0 e ∞.

Alguna característica a destacar de h(t) es que puede empezar en cualquier punto y crecer o decrecer a lo largo del tiempo. Es siempre no negativa y no tiene porque estar acotada superiormente. Además, $\int_0^\infty h(t)dt = +\infty$. Otra interpretación es que indica el modo en el que el riesgo de fallo varía con el tiempo.

Proposición 1.1. *Sea T una variable aleatoria no negativa con función de densidad f. Sean S(t) y h(t) sus respectivas funciones de supervivencia y riesgo. Se verifica la siguiente relación:* $h(t) = \frac{f(t)}{S(t)}$.

Dem. Aplicando la propiedad de probabilidad condicional en el intervalo $(t, t+\Delta t)$ dado que el individuo sobrevive hasta t, se sigue que:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t \cap T \geq t) \setminus P(T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t) \setminus P(T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \frac{1}{P(T \geq t)} \\ &= \frac{F'(t)}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{1.1}$$

□

Es interesante definir la **función de riesgo acumulada**; $H(t) = \int_0^t h(x)\,dx$.

Proposición 1.2. *Existe una relación entre la función de supervivencia y la función de riesgo acumulada que viene dada por:* $S(t) = exp[-H(t)] = exp(-\int_0^t h(x)\,dx)$.

Dem. Sabemos que $f(t) = F'(t) = -S'(t)$ y en consecuencia, $h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t)$, con S(0)=1. De donde se sigue el resultado. □

De manera inmediata se tiene también que, $f(t) = h(t)S(t) = h(t)exp(-\int_0^t h(x)\,dx)$.

1.3. Caso discreto.

A veces, cuando los tiempos de vida están agrupados o medidos de una determinada manera, por ejemplo, en días, podemos tratar T como una variable aleatoria discreta que puede tomar valores 0,1,2.. Sea la función de masa de probabilidad $f(j) = P(T = j)$, para j=0,1,2.. La función de supervivencia será en tal caso:

$$S(t) = P(T \geq t) = \sum_{j \geq t} f(j)$$

Es una función continua a izquierda no creciente que toma valor 1 en t=0 y 0 cuando $t \to \infty$.
La función de riesgo discreta se define como:

$$h(j) = P(T = j | T \geq j) = \frac{f(j)}{S(j)} \quad j = 0, 1, 2..$$

Como $f(j) = S(j) - S(j+1)$ se sigue que $h(j) = 1 - \frac{S(j+1)}{S(j)}$ y entonces,

$$S(j+1) = S(j)[1-h(j)] = S(j-1)[1-h(j-1)][1-h(j)] = ...$$

$$S(t) = \prod_{j<t}[1-h(j)]$$

1.4. Censura.

Un aspecto importante en el estudio de tiempos de vida y que distingue su análisis estadístico del de otras variables es la existencia de censura. La censura ocurre cuando por diversos motivos, no se conoce el tiempo T sino solo alguna cota suya. Existen diferentes tipos de censura a tener en cuenta, entre ellos:

- Censura a derecha. Ocurre cuando un estudio finaliza sin que el suceso de interés haya tenido lugar, solo se sabe que el suceso será posterior a un momento dado, el cual determina una cota inferior para T. Por ejemplo, sea T la variable de tiempo desde la detección de un tumor hasta la muerte. El estudio puede acabar antes de que todos los individuos hayan fallecido y por tanto, no conocemos con exactitud T para estos individuos sino únicamente que será posterior a la finalización del estudio.

- Censura por la izquierda. Ocurre cuando se desconoce el momento en el que tuvo lugar el suceso, solo se sabe que ocurrió antes de un momento dado. En tal caso, tenemos una cota superior para T. Por ejemplo, queremos saber en que momento los niños de una determinada clase aprenden a leer cuando están en el curso en el que se les enseña. Puede ocurrir que algún niño haya aprendido a leer antes en su casa. Entonces, desconocemos el momento en que aprendió a leer, solo sabemos que fue antes del comienzo del curso, el cual determina una cota superior.

- Censura por intervalos. Ocurre cuando lo que sabemos es que el suceso ha tenido lugar en un intervalo de tiempo (t_1, t_2), pero se desconoce el momento exacto. Estamos ante censura a derecha si $t_1 = 0$ y censura a izquierda si $t_2 = \infty$. Por ejemplo, en el seguimiento de una enfermedad cuando un paciente desarrolla determinados síntomas entre dos revisiones periódicas.

En nuestro estudio vamos a trabajar con datos censurados a derecha. Algunos mecanismos de censura a derecha son; la censura de tipo I, la censura de tipo II y la censura aleatoria.

- Censura de tipo I. Se da cuando cada individuo tiene un tiempo de censura conocido, $C_i > 0$, de manera que $T_i > 0$ se observa si $T_i \leq C_i$ y en caso contrario, solo sabemos que $T_i > C_i$. Los tiempos de censura pueden variar de individuo a individuo. Tiene lugar cuando el estudio se realiza a lo largo de un determinado periodo de tiempo. Por ejemplo, en ensayos clínicos, la entrada en el estudio suele ser escalonada junto con una fecha final donde termina.

- Censura de tipo II. Tiene lugar cuando el estudio empieza con n individuos y termina una vez r sucesos han tenido lugar, con r previamente fijado. En este caso, los r primeros tiempos son observados, mientras que los n-r restantes quedan censurados por la derecha. Por ejemplo, cuando ponemos a funcionar n aparatos al mismo tiempo y terminamos el estudio cuando r de ellos han fallado.

- Censura aleatoria independiente. Tiene lugar cuando cada individuo tiene un tiempo de vida T y de censura C, que son variables aleatorias mutuamente independientes con funciones de supervivencia S(t) y G(t), respectivamente. Por ejemplo, cuando en un ensayo clínico, los pacientes entran al estudio en momentos distintos y pueden recibir tratamientos distintos. De esta forma. la censura de cada individuo puede darse por distintas motivos.

Concretamente, en este TFG, vamos a tratar con datos censurados a derecha de tipo I. En primer lugar, vamos a establecer la notación. Sean n individuos con tiempos de vida $T_1, ..., T_n$, variables aleatorias, las cuales vamos a suponer independientes e idénticamente distribuidas. Definimos como C_i el tiempo de censura para el individuo i. Tomamos los C_i como constantes fijas. En general, pueden ser aleatorios, y en el caso de que sean independientes de los T_i, los resultados proporcionados en este TFG pueden darse por válidos también. Notar que cuando observamos los datos no tenemos el valor de $(T_1, .., T_n)$ sino que observamos las variables $Z_i = min(T_i, C_i)$. Definimos una variable indicadora de censura, $\delta_i = I(T_i = Z_i)$, de manera que toma valor 1 cuando $Z_i = T_i$ y 0 cuando $Z_i = C_i$. De esta forma, los datos que observamos son de la forma (Z_i, δ_i) para i=1,..,n.

Capítulo 2

Estimación del modelo.

En este capítulo, nuestro objetivo va a ser encontrar estimadores para la función de supervivencia. Con la presencia de datos censurados la función de distribución empírica es un estimador sesgado de la función de distribución poblacional, luego buscamos obtener estimadores para esta, que tengan en cuenta la censura. Esto puede llevarse a cabo por el método de Kaplan-Meier, que más adelante mostraremos. También daremos un estimador para la función de riesgo acumulada, el estimador de Nelson-Aalen.

Finalizaremos el capítulo con el estudio del modelo de regresión de Cox o modelo de riesgos proporcionales. El objetivo en este caso es plantear un modelo de regresión para el riesgo o para la supervivencia en función de ciertas variables explicativas. Normalmente, trabajamos con la función de riesgo y buscamos la existencia de relaciones entre alguna de las variables explicativas y el riesgo de los individuos. Teniendo en cuenta la relación entre la función de riesgo y la de supervivencia, si encontramos variables significativas para el riesgo también lo serán para la supervivencia.

2.1. Función de verosimilitud.

Veamos en primer lugar cómo construir la función de verosimilitud de la muestra para el caso general. Los distintos tipos de observaciones contribuyen de manera distinta a la verosimilitud.

Proposición 2.1. *La verosimilitud de la muestra en el caso general teniendo en cuenta la censura viene dada de la siguiente forma:*

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

donde $S(Z_i+)$ representa la desigualdad estricta, es decir, $P(T_i > Z_i)$. También podemos expresarlo como $S(Z_i+1)$ para el caso discreto y $S(Z_i)$ para el caso continuo.

Dem. Demostremos este resultado en el caso discreto.
Sea una muestra de n individuos con $T_1, ..., T_n$ tiempos de vida discretos e independientes. Recordar que $Z_i = min(T_i, C_i)$. Así pues, T_i es observado si $T_i \leq C_i$. Por otro lado, $\delta_i = I(T_i = Z_i) = I(T_i \leq C_i)$. Luego, los datos observados para los n individuos son de la forma (Z_i, δ_i)
Notar que dado que observamos Z_i, estamos ante dos posibilidades. Por un lado, tenemos $P(Z_i = C_i, \quad \delta_i = 0) = P(T_i > C_i)$ y por el otro, $P(Z_i = T_i, \quad \delta_i = 1) = f(Z_i)$, que es la función de masa de probabilidad. Así, la distribución conjunta queda:

$$f(Z_i)^{\delta_i} P(T_i > C_i)^{1-\delta_i}$$

Y la función de verosimilitud se sigue de la independencia de los T_i. □

Esta expresión se puede generalizar al caso en que la función de supervivencia de cada individuo sea distinta, dependiendo de sus características (covariables). En ese caso, y denotando por f_i y S_i sus

respectivas funciones de masa de probabilidad y de supervivencia, tenemos:

$$L = \prod_{i=1}^{n} f_i(Z_i)^{\delta_i} S_i(Z_i+)^{1-\delta_i}$$

Por otra parte, sea:

- $Y_i(t) = I(T_i \geq t, \quad C_i \geq t)$. El individuo i no ha fallado ni ha sido censurado antes del tiempo t.
- $N_i(t) = Y_i(t)I(T_i = t)$. El individuo i falla en el tiempo t sin haber sido censurado antes.

Teniendo en cuenta esta notación podemos escribir la función de verosimilitud para el caso general como:

$$\begin{aligned} L &= \prod_{i=1}^{n} f(Z_i)^{\delta_i} S(Z_i+)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \prod_{j=0}^{\infty} h_i(j)^{N_i(t)} [1-h_i(j)]^{Y_i(j)(1-N_i(j))} \end{aligned} \tag{2.1}$$

dado que; $f_i(Z_i)^{\delta_i} S_i(Z_i+)^{1-\delta_i} = \prod_{j=0}^{Z_i} h_i(j)^{N_i(j)} [1-h_i(j)]^{Y_i(j)(1-N_i(j))} = \prod_{j=0}^{\infty} h_i(j)^{N_i(j)} [1-h_i(j)]^{Y_i(j)(1-N_i(j))}$, con el convenio $0^0 = 1$.

2.2. Estimadores de supervivencia (Métodos no paramétricos).

En la práctica, normalmente obtenemos estimaciones de funciones de supervivencia que son funciones constantes a trozos, aún en el caso de que el tiempo de supervivencia se considere continuo.

2.2.1. El estimador de Kaplan-Meier.

Sea una muestra aleatoria simple de tamaño n. Es decir, suponemos que los T_i son variables aleatorias independientes e idénticamente distribuidas. Para el caso en el que no hubiera censura la función de supervivencia empírica sería $\hat{S}(t) = \frac{(n\,obs \geq t)}{n}$, función constante a trozos que decrece $\frac{b}{n}$ después de cada instante de tiempo, donde b es el número de tiempos de vida que toman ese valor.
En presencia de censura, no conocemos exactamente qué ocurre con aquellos individuos con tiempo de vida superior al tiempo de censura. Usamos en ese caso el estimador de Kaplan-Meier, también llamado estimador del límite producto. Es un método no paramétrico (no supone ninguna función de probabilidad) y por máxima verosimilitud, es decir, se basa en maximizar la función de verosimilitud de la muestra.
En primer lugar, ordenamos los tiempos en que tienen lugar los distintos sucesos: $t_1 < t_2 < .. < t_a$, $a \leq n$, de manera que en un mismo tiempo pueden tener lugar varios. El estimador de Kaplan-Meier viene dado por:

$$\hat{S}(t) = \prod_{j<t} \frac{n(j) - d(j)}{n(j)} \tag{2.2}$$

donde $n(j)$ representa el número de individuos que están en riesgo en tiempo j y $d(j)$ el número de sucesos que tienen lugar en tiempo j, es decir, $n(j) = \sum_{i=1}^{n} Y_i(j)$ y $d(j) = \sum_{i=1}^{n} N_i(j)$. En caso de que el tiempo de censura y el de vida coincidan, por convenio, tomamos ese tiempo como tiempo de vida esto es, no se considera el individuo como censurado.
El estimador de Kaplan-Meier es una función constante a trozos, continua a izquierda, que toma valor 1 en t=0 y disminuye en un factor $\frac{n(j)-d(j)}{n(j)}$ inmediatamente después de cada tiempo de vida t_i. La estimación no cambia en los tiempos de censura. Sin embargo, su efecto si se nota en los valores de $n(j)$ y, por tanto, en el tamaño de los saltos en $\hat{S}(t)$.

Proposición 2.2. *El estimador de Kaplan-Meier es un estimador de máxima verosimilitud no paramétrico para la función de supervivencia.*

Dem. Lo hacemos para el caso discreto.
Supongamos $T_1,...,T_n$ tiempos de vida independientes e idénticamente distribuidos que tienen distribución discreta y sean S(j) y h(j) funciones de supervivencia y riesgo para j=0,1.., respectivamente. Consideramos la distribución de T a través de su función de riesgo, h(t), tratándola como el parámetro. Como hemos visto antes, $L=\prod_{i=1}^{n}\prod_{j=0}^{\infty}h(j)^{N_i(j)}[1-h(j)]^{Y_i(j)(1-N_i(j))}$. Teniendo en cuenta que $n(j)=\sum_{i=1}^{n}Y_i(j)$ representa el número de individuos en riesgo en j y que $d(j)=\sum_{i=1}^{n}N_i(j)$ representa el número de sucesos que han ocurrido en j podemos escribir la verosimilitud anterior como:

$$L=\prod_{j=0}^{\infty}h(j)^{d(j)}[1-h(j)]^{n(j)-d(j)}$$

Dado que h(t) es el parámetro, como la función de verosimilitud es producto de funciones de h(t), para maximizarla basta maximizar cada factor por separado. Notar que si j es un tiempo en el que no tiene lugar ningún fallo $h(j)^{d(j)}[1-h(j)]^{n(j)-d(j)}=h(j)^0[1-h(j)]^{n(j)}$ que se maximizará cuando h(j)=0, por el convenio $0^0=1$.
Para el caso en el que en j tienen lugar uno o más fallos procedemos tomando el logaritmo:

$$\log\left(h(j)^{d(j)}[1-h(j)]^{n(j)-d(j)}\right)=d(j)\log h(j)+(n(j)-d(j))\log(1-h(j))$$

y derivamos respecto del parámetro e igualando a cero:

$$\frac{d(j)}{h(j)}-\frac{n(j)-d(j)}{1-h(j)}=0$$

obtenemos que $\hat{h}(j)=\frac{d(j)}{n(j)}$, para $j=0,1,..,\tau$, donde $\tau=max(j,\quad n(j)>0)$. Como $S(t)=\prod_{j<t}[1-h(j)]$, sustituimos y tenemos:

$$\hat{S}(t)=\prod_{j<t}[1-\hat{h}(j)]=\prod_{j<t}\left[1-\frac{d(j)}{n(j)}\right]$$

□

En cuanto al valor de h(j) para $j>\tau$ notemos que:

- Si todos los individuos que están en riesgo en τ presentan el suceso en τ, $h(\tau)^{d(\tau)}[1-h(\tau)]^{n(\tau)-d(\tau)}=h(\tau)^{d(\tau)}[1-h(\tau)]^0$ expresión que se maximiza para $h(\tau)=1$. $\hat{S}(\tau+)=0$. Como S(t) es no creciente $\hat{S}(j)=0,\forall j>\tau$.
- Si alguno de los individuos que está en riesgo en τ no presenta el suceso en τ, $d(\tau)<n(\tau)$. En ese caso, la verosimilitud será máxima cuando tome valor 1, luego cuando $d(\tau+1)=0$ y $n(\tau+1)-d(\tau+1)=0$, independientemente del valor de $h(\tau+1)$. Por lo tanto, el estimador de h no está defindio para $t>\tau$, y tampoco lo está $\hat{S}(t)$, para $t>\tau$. Ocurre cuando el último tiempo es un tiempo censurado.

2.2.2. Estimador de la varianza para el estimador de Kaplan-Meier.

Para el caso en el que no hubiera censura, un estimador para la varianza del estimador de la función de supervivencia empírica sería: $\widehat{Var}(\hat{S}(t))=\frac{\hat{S}(t)(1-\hat{S}(t))}{n}$. En presencia de datos con censura existen varios estimadores de la varianza del estimador de Kaplan-Meier, uno de los más conocidos es el siguiente.

Proposición 2.3. *La fórmula de Greenwood proporciona un estimador de la varianza del estimador de Kaplan-Meier. Este viene dada por:*

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j<t} \frac{d(j)}{(n(j)-d(j))n(j)}$$

Esta estimación de la varianza se puede obtener de la teoría estándar de máxima verosimilitud en muestras grandes.

Dem. Usaremos el método Delta, que afirma que si una sucesión Y_n cumple $\sqrt{n}(Y_n - a) \to N(0,\sigma)$ en distribución, y g es una función con derivada continua tal que $g'(a) \neq 0$, entonces Var $g(Y_n) \approx \frac{(g'(a))^2\sigma^2}{n}$. Dado el estimador de Kaplan-Meier (2.2): $\hat{S}(t) = \prod_{j<t} \frac{n(j)-d(j)}{n(j)}$, si tomamos como función su logaritmo,

$$\log \hat{S}(t) = \sum_{j<t} \log\left(1 - \frac{d(j)}{n(j)}\right) = \sum_{j<t} \log(1-\hat{h}(j))$$

Obviamente $\hat{S}(t) = exp[log(\hat{S}(t))]$. Podemos utilizar el método Delta considerando $g(y) = e^y$, de donde se sigue fácilmente que:

$$Var(\hat{S}(t)) \approx (S(t))^2 Var(\log(\hat{S}(t)))$$

y, por tanto, $Var(\hat{S}(t))$ se puede estimar por:

$$(\hat{S}(t))^2 Var(\log(\hat{S}(t))) = (\hat{S}(t))^2 Var\left(\sum_{j<t} \log(1-\hat{h}(j))\right) \tag{2.3}$$

Para analizar este valor, estudiaremos el comportamiento asintótico de $\hat{h}(j)$. Tomaremos solo los valores de j=0,1,2.. en los que d(j)>0, ya que en el resto $\hat{h}(j) = 0$.
La matriz de derivadas segundas de log L es diagonal. En efecto,

$$\log L = \sum_{j=0}^{\infty} d(j)\log h(j) + (n(j)-d(j))\log(1-h(j))$$

$$\frac{\partial \log L}{\partial h(j)} = \frac{d(j)}{h(j)} - \frac{n(j)-d(j)}{1-h(j)}$$

por lo que, $\frac{\partial^2 \log L}{\partial h(j)\partial h(u)} = 0$ y, por tanto, los estimadores máximo verosímiles de h(u) y h(j), $\hat{h}(u)$ y $\hat{h}(j)$, son asintóticamente incorrelados. Además las entradas diagonales son:

$$\frac{\partial^2 \log L}{\partial h(j)^2} = -\frac{d(j)}{h(j)^2} - \frac{n(j)-d(j)}{(1-h(j))^2} \approx -\frac{d(j)}{\left(\frac{d(j)}{n(j)}\right)^2} - \frac{n(j)-d(j)}{\left(1-\frac{d(j)}{n(j)}\right)^2} = -\frac{n(j)^3}{d(j)(n(j)-d(j))} = -\frac{n(j)}{\hat{h}(j)(1-\hat{h}(j))}$$

Por tanto, la varianza del estimador máximo verosímil de h(j), $\hat{h}(j)$ será, asintóticamente, $\frac{\hat{h}(j)(1-\hat{h}(j))}{n(j)}$.
Volviendo a (2.3), debido a la incorrelación asintótica tenemos que:

$$Var\left(\sum_{j<t} \log(1-\hat{h}(j))\right) \approx \sum_{j<t} Var(\log(1-\hat{h}(j)))$$

Aplicamos el método Delta con g(y)=log(1-y) y tenemos:

$$Var(\log(1-\hat{h}(j))) \approx Var(\hat{h}(j))\frac{1}{(1-h(j))^2} \approx \frac{\hat{h}(j)(1-\hat{h}(j))}{n(j)}\frac{1}{(1-\hat{h}(j))^2} = \frac{d(j)}{(n(j)-d(j))n(j)}$$

de donde se sigue el resultado. □

Teniendo en cuenta la normalidad asintótica de los estimadores máximo verosímiles, como $\sqrt{n}(\hat{S}(t) - S(t))$ es asintóticamente normal, podemos construir intervalos de confianza para la función de supervivencia a nivel $1-\alpha$.

$$\hat{S}(t) - z_{\frac{(1-\alpha)}{2}}\widehat{Var}[\hat{S}(t)] \leq S(t) \leq \hat{S}(t) + z_{\frac{(1-\alpha)}{2}}\widehat{Var}[\hat{S}(t)]$$

Sin embargo, cuando el número de tiempos de vida no censurados es pequeño o cuando S(t) es próximo a 0 o a 1 la variable no está bien aproximada por una N(0,1).

2.2.3. Estimador de Nelson Aalen y de su varianza.

Otro estimador también usado es el estimador de Nelson-Aalen o función de riesgo empírica acumulativa. El estimador de Nelson-Aalen es un estimador para la función de riesgo acumulado, es decir, para $H(t) = \int_0^t h(s)\,ds$. Notar que H(t) es lineal si h(t) es constante y es convexa si h(t) es monótona. Para el caso discreto, un estimador para la función de riesgo acumulada sería el estimador de Nelson-Aalen:

$$\hat{H}(t) = \sum_{j \leq t} \frac{d(j)}{n(j)}$$

Un estimador asintótico para la varianza de $\hat{H}(t)$, siguiendo el procedimiento anterior para la de $\hat{S}(t)$, es $\widehat{Var}[\hat{H}(t)] = \sum_{j \leq t} \frac{d(j)(n(j)-d(j))}{n(j)^3}$.
Alternativamente, podríamos tomar también el siguiente estimador para su varianza: $\widehat{Var}[\hat{H}(t)] = \sum_{j \leq t} \frac{d(j)}{n(j)^2}$.
En muestras grandes, ambos dan resultados parecidos.
Otra opción sería tomar una estimación a partir de la función de supervivencia teniendo en cuenta que $\hat{S}(t) = exp[-\hat{H}(t)]$, y en consecuencia, $\hat{H}(t) = -log\hat{S}(t)$. Ambos estimadores, $\hat{S}(t)$ y $\hat{H}(t)$ son consistentes y asintóticamente normales.

2.3. Modelo de regresión de Cox.

En muchas ocasiones, el tiempo de supervivencia de un individuo depende de un conjunto de variables explicativas. Ante esta situación, el modelo de riesgos proporcionales de Cox es el más usado para representar el efecto que estas tienen. Ejemplos de covariables pueden ser, por ejemplo; el sexo, la edad...
Sea T la variable de tiempo de vida continua y sea $\mathbf{x}$ un vector de dimensión px1 de variables explicativas, covariables, fijas. Las funciones de riesgo para T dado $\mathbf{x}$ son de la forma $h(t|\mathbf{x}) = h_0(t)r(\mathbf{x})$, con h_0 y $r(\mathbf{x})$ funciones que toman valores positivos. Este modelo proporciona una expresión de la función de riesgo para un individuo en un tiempo t dado un conjunto de variables explicativas. Llamamos a $h_0(t)$ función de riesgo basal, y se corresponde con el riesgo de un individuo cuando todas las variables explicativas toman valor 0. Normalmente r(x) es de la forma $exp(\beta'\mathbf{x})$ con β un vector de dimensión px1 formado por los coeficientes de regresión. De esta forma, la función de riesgo para T dado $\mathbf{x}$ toma la forma,

$$h(t|\mathbf{x}) = h_0(t)exp(\beta'\mathbf{x})$$

Notar que la fórmula tiene dos partes, la de la función de riesgo basal que no depende del vector de las $\mathbf{x}$'s, por lo que podemos decir que esta parte es no paramétrica, y la de la exponencial, que no depende de t pero sí de los parámetros. De ahí que se diga que este modelo es semi-paramétrico.
Como conocemos la relación entre S(t) y H(t) podemos expresar la correspondiente función de supervivencia como $S(t|\mathbf{x}) = S_0(t)^{r(\mathbf{x})}$ con $S_0(t) = exp[-H_0(t)]$ función de supervivencia basal. En este caso,

la función de supervivencia para T dado **x** queda:

$$\begin{aligned} S(t|\mathbf{x}) &= exp\left(-\int_0^t h(u|\mathbf{x})\,du\right) \\ &= exp\left(-exp(\beta'\mathbf{x})\int_0^t h_0(u)\,du\right) \\ &= \left(exp\left(-\int_0^t h_0(u)\,du\right)\right)^{exp(\beta'\mathbf{x})} \\ &= [S_0(t)]^{exp(\beta'\mathbf{x})} \end{aligned} \tag{2.4}$$

donde $S_0(t)$ es la función de supervivencia basal.

2.3.1. Estimación de los parámetros y de la función basal.

Dada una muestra aleatoria de tiempos de vida, donde las observaciones vienen dadas en la forma (Z_i, δ_i) y dado un vector de covariables **x**, el objetivo es estimar los parámetros y la función de riesgo basal para así poder obtener una estimación para $h(t|\mathbf{x}) = h_0(t)exp(\beta'\mathbf{x})$.
Los parámetros del modelo no pueden estimarse por el método de máxima verosimilitud al ser desconocida la forma específica de $h_0(t)$. Existen varios métodos que se pueden usar para ello. Vamos a explicar el que propuso Cox. Se basa en el método de verosimilitud parcial. La diferencia con respecto al método de la verosimilitud está en que el método de verosimilitud se basa en el producto de todas las verosimilitudes para todos los individuos de la muestra, mientras que el método de verosimilitud parcial se basa en el producto de las verosimilitudes de todos los sucesos ocurridos, es decir, el producto de la probabilidad de que haya sido el individuo i el que ha fallado en t, dado el conjunto de individuos en riesgo en tiempo t.
La función de verosimilitud propuesta por Cox considera una muestra aleatoria censurada de tiempos de vida, (Z_i, δ_i), de n individuos, donde hay a tiempos de vida distintos $(t_1 < ... < t_a)$ y n-a tiempos censurados. Sea $R_i = R(t_i)$ el conjunto de individuos vivos y no censurados justo antes de t_i, es decir, el conjunto de individuos en riesgo para t_i. Notar que $card(R_i) = n(t_i)$.
La verosimilitud parcial se basa en el producto de las verosimilitudes de todos los sucesos ocurridos, $L = L_1 x ... x L_a = \prod_{i=1}^{a} L_i$. Sean $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_a$ las covariables asociadas a los individuos que han fallado en $t_1, .., t_a$. Notar que las observaciones censuradas no aportan información sobre el tiempo de fallo en el cálculo de la verosimilitud parcial. La verosimilitud parcial del fallo del individuo i se calcula como la probabilidad condicional de que haya sido ese individuo el que ha fallado dado el conjunto de individuos en riesgo, es decir,

$$\frac{h(t_i|\mathbf{x}_i)}{\sum_{l\in R_i} h(t_i|\mathbf{x}_l)} = \frac{h_0(t_i)exp(\beta'\mathbf{x}_i)}{\sum_{l\in R_i} h_0(t_i)exp(\beta'\mathbf{x}_l)} = \frac{exp(\beta'\mathbf{x}_i)}{\sum_{l\in R_i} exp(\beta'\mathbf{x}_l)}$$

Notar que esta fracción representa la 'verosimilitud parcial' para el individuo que falla en t_i.
La verosimilitud parcial conjunta de toda la muestra queda entonces:

$$L(\beta) = \prod_{i=1}^{a} L_i = \prod_{i=1}^{a} \frac{exp(\beta'\mathbf{x}_i)}{\sum_{l\in R_i} exp(\beta'\mathbf{x}_l)} \tag{2.5}$$

Notar que a pesar de que $L(\beta)$ no es realmente una expresión de verosimilitud, como hemos dicho antes, ya que no puede expresarse como la probabilidad de ninguna muestra observable, si puede tratarse como una función de verosimilitud a la hora de hacer la estimación de β [11].
Para la estimación de β maximizamos el logaritmo de la función de verosimilitud parcial, $L(\beta)$, que no depende de la función de riesgo basal, como si esta fuera una función de verosimilitud ordinaria y derivamos respecto de los parámetros e igualamos a 0. Obtenemos un estimador, $\hat{\beta}$, consistente y asintóticamente normal. Sin embargo, el sistema no es fácil de resolver y es necesario recurrir a métodos

numéricos como el de Newton-Raphson.

La fórmula (2.5) es válida cuando no hay empates, es decir, los individuos mueren en tiempos distintos. En el caso discreto sin embargo, puede ocurrir que varios individuos mueran en el mismo tiempo. En ese caso, existen varias correciones de la expresión anterior, como la de Efron [9], aunque en caso de que el número de empates sea pequeño se suele usar la verosimilitud general, ya que las estimaciones obtenidas son muy parecidas.

Para la estimación de $S_0(t)$, recurrimos a la relación que tiene con $H_0(t)$.

Proposición 2.4. *Un estimador para* $H_0(t)$ *es* $\hat{H}_0(t) = \sum_{j \leq t} \left[\frac{d(j)}{\sum_{l=1}^{n} Y_l(j) e^{\hat{\beta}' x_l}} \right]$.

Dem. Notar que un individuo l, con vector de covariables $\mathbf{x}_l$, que esté vivo y no censurado antes del tiempo j, tiene una probabilidad $h(j|\mathbf{x}_l)$ de morir en j. Así podemos escribir el número esperado de muertes en j como:

$$\sum_{l=1}^{n} Y_l(j) h(j|\mathbf{x}_l) = \sum_{l=1}^{n} Y_l(j) h_0(j) e^{\beta' \mathbf{x}_l}$$

Igualando el número esperado de muertes en j al observado, d(j), y despejando obtenemos un estimador para $h_0(j)$:

$$\hat{h}_0(j) = \frac{d(j)}{\sum_{l=1}^{n} Y_l(j) e^{\hat{\beta}' \mathbf{x}_l}}$$

Como $H_0(t) = \sum_{j \leq t} h_0(j)$, se sigue el resultado. □

Notar que para $\hat{\beta} = 0$, este estimador coincide con el estimador de NA. Obtenemos en consecuencia, un estimador para la función de supervivencia basal dado por $\hat{S}_0(t) = exp(-\hat{H}_0(t-))$.

Una vez que hemos obtenido una estimación para β y para $S_0(t)$, ya tenemos la estimación para $S(t|\mathbf{x})$, que queda $\hat{S}(t|\mathbf{x}) = \hat{S}_0(t)^{exp(\hat{\beta}'\mathbf{x})}$.

2.3.2. Razón de riesgos

El modelo de regresión de Cox también recibe el nombre de modelo de riesgos proporcionales. Este nombre viene de que si hacemos el cociente entre dos funciones de riesgo para dos individuos distintos obtenemos una cantidad que no depende de t:

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} = \frac{h_0(t) exp(\beta' \mathbf{x}_i)}{h_0(t) exp(\beta' \mathbf{x}_j)} = \frac{exp(\beta' \mathbf{x}_i)}{exp(\beta' \mathbf{x}_j)} = exp(\beta'(\mathbf{x}_i - \mathbf{x}_j))$$

Denotamos el cociente anterior como **razón de riesgos, HR**, entre dos individuos. Nos proporciona cuánto incrementa el riesgo por unidad que se incrementa la covariable.
Notar que si tomamos logaritmos; $\ln HR = \beta'(\mathbf{x}_i - \mathbf{x}_j)$. Si obtenemos HR=1, $\ln HR = 0$, lo que significa que el suceso de interés tiene la misma probabilidad de darse para ambos individuos. Un HR>1 proporciona un mayor riesgo a experimentar el suceso para el individuo i que para el j, y un HR<1, lo contrario. Por ejemplo, si obtenemos un HR=3 y los individuos todavía no han experimentado el suceso de interés, esto quiere decir que el individuo i tiene el triple de probabilidad de experimentarlo que el j en el siguiente periodo de tiempo.
Supongamos que tenemos dos individuos, i y j, con los mismas valores para las covariables, salvo uno

que difiere en una unidad. Es decir, $(x_{i1},..,x_{ih},..,x_{ik})$ y $(x_{j1},..,x_{jh},..,x_{jk})$ con $x_{ih} = x_{jh} + 1$, y el resto iguales. En tal caso, $HR = e^{\beta_h}$ indica cuánto varía la razón de riesgos cuando la variable aumenta en una unidad. Si la incrementáramos en c unidades obtendríamos un $HR = e^{c\beta_h}$.

Además, cabe destacar que esta razón de riesgos no depende de la función de riesgo basal, únicamente de las covariables y los β's. Así mismo, no depende del tiempo, luego es constante. La interpretación de la razón de riesgos es similar a la de los 'odd ratio' en regresión logística [8].

2.3.3. Regresión de Cox para comparar la supervivencia en grupos.

Los modelos de regresión de Cox son de gran utilidad a la hora de comparar dos o más distribuciones de tiempos de vida. Por ejemplo, la de los hombres con la de las mujeres.
Supongamos que tenemos muestras aleatorias independientes censuradas de tiempos de vida y queremos comparar m distribuciones: $S_1(t),\ldots,S_m(t)$. Supongamos que es un modelo proporcional de riesgos, de manera que $S_0(t) = S_m(t)$ y $S_r(t) = S_0(t)^{exp(\beta_r)}$ para r=1...m-1, y por tanto, $S_r(t) = S_m(t)^{exp(\beta_r)}$ para r=1...m-1.
Queremos contrastar la igualdad de las funciones de supervivencia. En tal caso, nuestra hipótesis es $H_0 : S_1(t) = \ldots = S_m(t)$, que resulta ser lo mismo que contrastar $H_0 : \beta = 0$, con $\beta = (\beta_1 \ldots \beta_{m-1})$ vector de parámetros.
Para ello usaremos un test de la función Score o vector de primeras derivadas, $U(\beta)$. Tomaremos la función de 'verosimilitud parcial', de manera que si la interpretamos como una función de verosimilitud ordinaria obtenemos; $U(\beta) = (U_1(\beta),...,U_k(\beta))$ donde $U_i(\beta) = \frac{\partial(\log L(\beta))}{\partial \beta_i}$. $U(\beta)$ tiene media 0 y matriz de covarianzas $I(\beta)$, que es la matriz de información de Fisher, y cuyos elementos vienen dados por $I_{ij}(\beta) = E[-\frac{\partial^2(\log L(\beta))}{\partial \beta_i \partial \beta_j}]$, para i,j=1,... Bajo la hipótesis nula $U(0)'I^{-1}(0)U(0)$ tiene distribución asintótica χ^2_{m-1}. [14]
Un test para $H : \beta = 0$ es también el que se basa en el estadístico Wald. Como β tiene distribución asintóticamente normal con vector de media 0 y matriz de covarianzas $\widehat{Var}(\hat{\beta}) = I^{-1}(\hat{\beta})$, el estadístico de Wald es $W = \hat{\beta}' I(\hat{\beta})\hat{\beta}$, que tiene distribución asintótica χ^2_{m-1} bajo la hipótesis nula. Valores grandes de W permiten rechazar la hipótesis de igualdad en las distribuciones.

2.3.4. Comprobación de la hipótesis de riesgos proporcionales.

La principal hipótesis del modelo de Cox es la de riesgos proporcionales. Existen varias formas para comprobarla. El método gráfico más común consiste en realizar la curva de supervivencia 'log(-log)'. Representamos gráficamente la función $-\ln(-\ln(\hat{S}(t|\mathbf{x})))$.
Notar que si realizamos el gráfico 'log(-log)' y obtenemos dos curvas paralelas, se cumplen las hipótesis del modelo de Cox o modelo de riesgos proporcionales. La diferencia entre las dos curvas sería una constante $\hat{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)$, que no depende de t. Notar que es un método subjetivo y no rechazamos la hipótesis de proporcionalidad a no ser que se vea muy claro que las curvas no son paralelas.
Existen otros métodos menos comunes, como el de comparar las curvas de supervivencia 'observadas' y 'esperadas' para las distintas covariables por separado. Obtenemos las funciones de supervivencia 'observadas' de las funciones de supervivencia de Kaplan-Meier para cada variable y las 'esperadas' del modelo de Cox. Si las curvas obtenidas son similares aceptamos la hipótesis de riesgos proporcionales.

Otro método más objetivo sería realizar un estudio de la bondad de ajuste para cada variable del modelo de riesgos proporcionales. La hipótesis nula es la de proporcionalidad, entonces un valor grande del p-valor nos lleva a aceptar la hipótesis de riesgos proporcionales, mientras que si este es pequeño, la rechazamos. Algunos de los test para comprobar la hipótesis de proporcionalidad se pueden consultar en [2].

Capítulo 3

Modelos de riesgos competitivos.

En este capítulo vamos a estudiar los modelos en presencia de riesgos competitivos, los cuales resultan de gran interés ya que tratan el hecho de que los individuos pueden fallar por diferentes causas. Hasta ahora hemos hablado del estudio del tiempo hasta que un determinado suceso tiene lugar, o en caso contrario, es censurado. Sin embargo, el evento de interés puede tener lugar por diferentes motivos. Por ejemplo, en un estudio sobre el fallo de un vehículo, este puede deberse a un accidente, al fallo del motor...

En un entorno de supervivencia clásico, la función de supervivencia se estima mediante el método de Kaplan-Meier (2.2), el cual estima la probabilidad de experimentar el suceso de interés antes de un determinado tiempo. Este procedimiento es correcto en ausencia de riesgos competitivos. Sin embargo, si un individuo ha podido fallar por otra causa, el estimador de Kaplan-Meier no se puede interpretar de la misma forma. Veámoslo con un ejemplo que posteriormente estudiaremos en el capítulo 4. Realizamos un estudio sobre el tiempo de vida en UCI de pacientes ingresados por COVID. Los pacientes pueden salir de la UCI, o bien porque reciben el alta, o bien porque mueren. Las observaciones de los pacientes serán censuradas cuando tras la finalización del estudio el paciente continúe en la UCI, y por tanto, no llegamos a observar su resultado. Sin embargo, notar que puede darse otro caso. Los pacientes pueden recibir el alta y por tanto, ya no están en el grupo de riesgo para la muerte, y viceversa. Decimos entonces que el alta es un riesgo competitivo para la muerte.

En el análisis de riesgos competitivos, las personas que experimentan un riesgo competitivo tienen cero posibilidades de experimentar el suceso de interés. En el estimador de Kaplan-Meier se supone que todos los individuos tienen la misma probabilidad de experimentar el suceso de interés. Esto conduciría a una sobreestimación de la incidencia acumulada, proporción de individuos sanos que desarrollan la enfermedad a lo largo de un periodo determinado. En ausencia de riesgos competitivos, el riesgo y la incidencia acumulada están directamente relacionadas de manera que, un aumento del riesgo conlleva un tiempo de supervivencia más corto. En el ejemplo anterior, si consideramos el alta como censura, el estimador de Kaplan-Meier nos dará una estimación del tiempo hasta la muerte mayor que el real.

3.1. Características básicas y especificación del modelo.

Supongamos un conjunto de n individuos. Consideramos m diferentes causas de fallo. Suponemos que el individuo falla por una sola causa. Cada individuo está caracterizado por el par (T,V) donde T es el tiempo de fallo, T>0, y V el modo de fallo, $V \in \{1,...,m\}$. La distribución conjunta de (T,V) puede especificarse a través del riesgo específico para la k-ésima causa, $h_k(t)$, o también a través de la función de incidencia acumulada, $F_k(t)$, con k=1,...,m.
Definimos la **función de riesgo específico de la k-ésima causa** como:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t, V = k | T \geq t)}{\Delta t} \quad k = 1,..,m \tag{3.1}$$

donde la probabilidad del numerador representa la probabilidad de morir por la causa k en $(t, t+\Delta t)$ con todos las causas de fallo presentes y dado que ha sobrevivido hasta el instante t. Es decir, representa la tasa instantánea de ocurrencia del suceso de interés en un entorno en el que los individuos también pueden experimentar el resto de sucesos.
Así mismo definimos la **función de riesgo marginal** como:

$$h(t) = \sum_{k=1}^{m} h_k(t) = \lim_{\Delta t \to 0} \frac{P(T < t+\Delta t | T \geq t)}{\Delta t}$$

También podemos definir el **riesgo acumulado para la k-ésima causa** como: $H_k(t) = \int_0^t h_k(u)\,du$, k=1,..,m. Entonces el **riesgo acumulado marginal** es:

$$H(t) = \int_0^t h(u)\,du = \sum_{k=1}^{m} H_k(t) \tag{3.2}$$

Por otro lado, tenemos la **función de subsupervivencia para la k-ésima causa**: $S_k(t) = P(T \geq t, V = k)$, k=1,..,m, que representa la probabilidad de experimentar el fallo k en un tiempo mayor que t. De manera que, la correspondiente **función de supervivencia marginal** queda:

$$S(t) = P(T \geq t) = e^{-H(t)} = e^{-\sum_{k=1}^{m} H_k(t)} = \prod_{k=1}^{m} e^{-H_k(t)} = \prod_{k=1}^{m} S_k(t)$$

Notar que S(t) puede estimarse con el estimador de Kaplan-Meier. Hay que destacar que, las $S_k(t)$ no son una función de supervivencia ya que $S_k(0) = P(V = k)$, valor que es estrictamente inferior a 1 si hay más de una causa de fallo.
Así mismo, definimos la **función de incidencia acumulada, CIF,** como:

$$F_k(t) = P(T \leq t, V = k) = \int_0^t h_k(u) S(u)\,du \quad k = 1,..,m \tag{3.3}$$

Notar que, $S_k(t) \neq 1 - F_k(t)$. La CIF denota la probabilidad de experimentar el suceso antes de un determinado tiempo y antes de que tenga lugar cualquier otro riesgo competitivo. En consecuencia, los individuos que han experimentado un riesgo competitivo, distinto al suceso de interés, ya no pueden experimentar este. Notar que en tal caso, el número de individuos en riesgo, disminuye más rápido con el tiempo.
Notar que tenemos también la **subfunción de densidad para cada modo de fallo k**: $f_k(t) = F_k'(t) = h_k(t)S(t)$. Ambas, $F_k(t)$ y $f_k(t)$, especifican también la distribución del par (T,V).
Obtenemos en consecuencia, $F(t) = P(T \leq t) = \sum_{k=1}^{m} F_k(t)$ y $\pi_k = P(V = k) = \lim_{t \to +\infty} F_k(t)$, k=1,..,m, que representa la distribución de probabilidad de los diferentes modos de fallo.

3.2. Función de verosimilitud.

Tomamos una muestra aleatoria de n individuos. Los datos observados para el individuo i son de la forma (T_i, V_i), tiempo y modo de fallo, respectivamente. Notar que si el individuo i es censurado en Z_i no conocemos su modo de fallo. Así, estamos considerando dos tipos de observaciones: $(T_i = Z_i, V_i)$ y $T_i > Z_i$.

Proposición 3.1. *Sea δ_i un indicador de censura, que toma valor 1 si Z_i es un tiempo de fallo y 0 si es un tiempo de censura. La función de verosimilitud queda:*

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

donde k representa el modo de fallo del individuo i. Además, la verosimilitud se factoriza como producto de las verosimilitudes para las distintas causas de fallo, L_k con k=1,...,m.

Dem. Demostraremos este resultado en el caso discreto.
Sea una muestra de n individuos con $T_1,..,T_n$ tiempos de vida discretos e independientes y $V_1,..,V_n$ modos de fallo para los correspondientes individuos. Recordar que $Z_i = min(T_i, C_i)$, con $C_i > 0$ constante. Por otro lado, $\delta_i = I(T_i \leq C_i)$. Definimos ahora $W_i = \delta_i V_i$ que proporciona el modo de fallo para aquellos individuos que no son censurados y 0 para los censurados. Luego, los datos observados para los n individuos son de la forma (Z_i, δ_i, W_i).
Notar que un individuo puede contribuir a la verosimilitud de dos formas. Por un lado, puede ocurrir que el individuo i falle en Z_i por la causa $V_i = k$, de manera que observamos $(T_i, 1, k)$. Su contribución a la verosimilitud será $f_k(Z_i)$. En caso contrario, si el individuo i es censurado en Z_i observamos $(C_i, 0, 0)$, de forma que su contribución a la verosimilitud es $P(T_i > C_i)$. Así la distribución conjunta queda:

$$f_k(Z_i)^{\delta_i} P(T_i > C_i)^{1-\delta_i}, \tag{3.4}$$

y la función de verosimilitud se sigue de la independencia de los T_i.
Veamos ahora que la verosimilitud se factoriza como producto de funciones de las distintas causas de fallo, L_k con k=1,...,m.
Por un lado, tenemos que $S(t) = \prod_{k=1}^{m} S_k(t)$ funciones que no tienen las propiedades de las funciones de supervivencia, ya que no cumplen $S_k(0) = 1$. Sea, por otro lado, $\delta_{ik} = I(V_i = k)$, de tal manera que $\delta_i = \sum_{k=1}^{m} \delta_{ik}$. Y, por último, definimos $g_k(t)$ como $g_k(t) = -\frac{dS_k(t)}{dt} = h_k(t)S_k(t)$ dado que,

$$\frac{dS_k(t)}{dt} = \frac{d(exp(-H_k(t)))}{dt} = -\frac{dH_k(t)}{dt} exp(-H_k(t)) = -h_k(t)S_k(t)$$

En tal caso, a partir de (3.4) tenemos;

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\prod_{k=1}^{m} f_k(Z_i)^{\delta_{ik}} \right) \left(\prod_{k=1}^{m} S_k(Z_i)^{1-\delta_i} \right) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{m} \left(h_k(Z_i)^{\delta_{ik}} \left[\prod_{l=1}^{m} S_l(Z_i)^{\delta_{ik}} \right] S_k(Z_i)^{1-\delta_i} \right) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{m} h_k(Z_i)^{\delta_{ik}} S_k(Z_i)^{\delta_{ik}} S_k(Z_i)^{\sum_{l \neq k} \delta_{il} + 1 - \delta_i} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{m} g_k(Z_i)^{\delta_{ik}} S_k(Z_i)^{1-\delta_{ik}} \\ &= \prod_{k=1}^{m} L_k \end{aligned} \tag{3.5}$$

Obtenemos la verosimilitud como producto de m funciones de verosimilitud, una para cada modo de fallo. Cada L_k tiene como función de masa de probabilidad $g_k(t)$ y como función de supervivencia $S_k(t)$, que como ya hemos dicho antes no corresponden a la de ninguna variable aleatoria observable. □

La forma de la función de verosimilitud nos muestra que tanto $h_k(t)$ como $H_k(t)$ se pueden estimar de los datos (T,V).

3.3. Métodos no paramétricos.

Sabemos de (3.1) que $h_k(t)$ representa la probabilidad de fallar en el tiempo t por la causa k habiendo llegado hasta el tiempo t. Luego, podemos estimar esta función por el cociente entre el número de individuos que presenta el fallo k en el tiempo j entre los individuos que están en riesgo en j:

$$\hat{h}_k(j) = \frac{d_k(j)}{n(j)}$$

Conociendo la relación (3.2) podemos obtener un estimador para la función de riesgo acumulada para la k-ésima causa:

$$\hat{H}_k(j) = \sum_{i \leq t} \frac{d_k(j)}{n(j)} \quad k = 1,..,m$$

Una expresión para la estimación de su varianza sería, $\widehat{Var}[\hat{H}_k(j)] = \sum_{j \leq t} \frac{d_k(j)}{n(j)^2}$, k=1,..,m.

Además, sabemos de (3.3) que $F_k(t)$ representa la probabilidad de fallar por la causa k antes del tiempo t. Podemos entonces estimarla por:

$$\hat{F}_k(j) = \sum_{j \leq t} \hat{h}_k(j)\hat{S}(j) = \sum_{j \leq t} \hat{S}(j)\frac{d_k(j)}{n(j)}, \quad k = 1,..,m \tag{3.6}$$

donde $\hat{S}(t)$, la función de supervivencia marginal, se puede estimar por Kaplan-Meier, juntando todos los modos de fallo.
Existen tests para comparar la igualdad de CIF's entre varios grupos, que se pueden consultar en [4].

3.4. Métodos semiparamétricos.

En presencia de riesgos competitivos, también puede ocurrir que el tiempo de supervivencia de un individuo dependa de un conjunto de variables explicativas o covariables. Vamos a extender el modelo de riesgos proporcionales de Cox, para representar el efecto de estas variables.
Sea T la variable de tiempo de vida continua y sea $\mathbf{x}$ un vector de dimensión px1 de covariables fijas. Tomando cada modo de fallo separadamente, la función de riesgo para la k-ésima causa dado $\mathbf{x}$ viene dado de la forma:

$$h_k(t|\mathbf{x}) = h_{0k}(t)exp(\beta_k'\mathbf{x}), \quad k = 1,..,m$$

y donde β_k es un vector de dimensión px1 formado por los coeficientes de regresión para el modo de fallo k. Llamamos a $h_{0k}(t)$ función de riesgo basal asociada a la k-ésima causa, que se corresponde con el riesgo de un individuo que falla por la causa k y dado que todas las covariables toman valor 0. Al igual que cuando no considerábamos riesgos competitivos, el modelo es semiparamétrico. Además, la hipótesis de riesgos proporcionales debe comprobarse para cada causa.

En primer lugar, para la estimación de los parámetros β_k para cada modo de fallo, vamos a basarnos en el método de la verosmilitud parcial de Cox (2.5). Sea una muestra aleatoria censurada de n individuos, donde las observaciones vienen dadas de la forma (Z_i, δ_{ik}) y donde hay a tiempos de fallo por la causa k ($t_1 < .. < t_a$) y n-a tiempos censurados o individuos que no fallan por la causa k. Sea $R_{ik} = R_k(t_i)$ el conjunto de individuos vivos y no censurados justo antes de t_i que fallan por la causa k. La verosimilitud parcial para la estimación de β_k viene dada por (3.1):

$$L(\beta) = \prod_{k=1}^{m} L_k(\beta_k) = \prod_{i=1}^{a}\prod_{k=1}^{m}\left(\frac{e^{\beta_k'\mathbf{x}_i}}{\sum_{l \in R_{ik}} e^{\beta_k'\mathbf{x}_l}}\right)^{\delta_{ik}}$$

Teniendo en cuenta el hecho de que un fallo de causa k tiene lugar en t_i, la probabilidad de que el individuo $l \in R_{ik}$ sea quien falle en ese tiempo por la causa k viene dada por:

$$\frac{h_k(t_i|\mathbf{x}_i)}{\sum_{l \in R_{ik}} h_k(t_i|\mathbf{x}_l)} = \frac{exp(\beta_k'\mathbf{x}_i)}{\sum_{l \in R_{ik}} exp(\beta_k'\mathbf{x}_l)}$$

Y aunque como ocurría para el caso en el que no considerábamos los riesgos competitivos, esta no es una función de verosimilitud, si puede usarse como tal para la estimación de los β_k. Maximizando $L_k(\beta_k)$ mediante métodos numéricos, como el método de Newton-Raphson, primero tomando el logaritmo y

luego derivando respecto al parámetro β_k e igualando a 0 y despejando, obtenemos un estimador, $\hat{\beta}_k$, para k=1,...,m, que es consistente y asintóticamente normal. Notar que el estimador de β_k es el mismo que si consideramos cada causa por separado tomando el resto como censura.

En segundo lugar, procediendo análogamente al caso en el que no considerábamos riesgos competitivos (2.4) obtenemos un estimador para la función de riesgo basal para la k-ésima causa:

$$\hat{H}_{0k}(t) = \sum_{j \leq t} \left[\frac{d_k(j)}{\sum_{l=1}^{n} Y_l(j) e^{\hat{\beta}_j' x_l}} \right]$$

Como conocemos la relación $S(t|\mathbf{x}) = exp(-\sum_{k=1}^{m} H_k(t)) = exp(-\sum_{k=1}^{m} H_{0k}(t) e^{\beta_k' \mathbf{x}})$, y sabemos como obtener un estimador para β_k y otro para $H_{0k}(t)$, la estimación para la función de supervivencia marginal dado un vector de covariables queda:

$$\hat{S}(t|\mathbf{x}) = exp[-\sum_{k=1}^{m} \hat{H}_{0k}(t) e^{\hat{\beta}_k' \mathbf{x}}]$$

Finalmente, sustituyendo estos estimadores en la fórmula que tenemos para la función de incidencia acumulada, $F_k(t|\mathbf{x}) = \int_0^t S(u|\mathbf{x}) h_k(u|\mathbf{x}) du = \int_0^t S(u|\mathbf{x}) dH_k(u|\mathbf{x})$, obtenemos un estimador para esta:

$$\hat{F}_k(t|\mathbf{x}) = \sum_{j \leq t} \hat{S}(j|\mathbf{x}) \frac{d_k(j)}{\sum_{l=1}^{n} Y_l(j) e^{\hat{\beta}_k' x_l}} e^{\hat{\beta}_k' x_l} \quad k = 1, .., m. \tag{3.7}$$

Capítulo 4

Aplicación a un conjunto de datos reales.

Para ilustrar lo que acabamos de abordar acerca del análisis de supervivencia, vamos a aplicarlo a un conjunto de datos reales. En particular, vamos a estudiar el tiempo en UCI de pacientes ingresados por COVID. Disponemos de una muestra de 500 pacientes, desde el inicio de la pandemia hasta el fin de 2020.

Como sabemos, el COVID-19 es una enfermedad infecciosa causada por el SARS-CoV-2. En casos leves produce síntomas similares a los de la gripe, además de perdida de olfato y gusto. Sin embargo, en casos graves puede producir neumonia, dificultad respiratoria y lesión e inflamación de órganos principales. Estos sintómas pueden acarrear el paso por UCI de determinados pacientes.

Nuestro conjunto de datos consta de las variables; edad ($edad$), sexo ($sexo$), días en UCI ($dias\_uci$), hipertensión arterial (hta) y suceso ($resultado$). Distinguimos 2 posibles sucesos además de la censura ($censurados$); pacientes que han fallecido antes de la finalización del estudio ($exitus$) y pacientes que han sido dados de alta antes de la finalización del estudio ($alta$). Notar que los pacientes que siguieron en la UCI cuando el estudio finalizó constituyen el conjunto de datos censurados.

Si nos fijamos en esta última variable, notar que estamos en presencia de riesgos competitivos. El suceso de interés es la salida de la UCI, puede darse por el alta del paciente o por su fallecimiento.

Hemos elegido el sexo, la edad y la hipertensión arterial, entre otras muchas, como variables para el estudio de su efecto en el tiempo de supervivencia, ya que sabemos que hay estudios, tales como [13] donde se ha comprobado su influencia.

Para realizar el análisis estadístico de modelos de riesgos competitivos hacemos uso del programa R, el cual incluye varios paquetes que permiten realizar un estudio suficientemente completo de estos modelos. Algunos resultados obtenidos y que vamos a comentar se encuentran en el Anexo I.

4.1. Análisis descriptivo de las variables.

Previamente a la aplicación del análisis de supervivencia en presencia de riesgos competitivos al conjunto de datos, vamos a hacer un análisis descriptivo de las variables.

- Sexo. Variable de tipo factor. Nuestra muestra consta de 348 hombres y 152 mujeres.
- Edad. Variable numérica. Podemos ver un resúmen en la tabla 6. La moda de esta variable es de 71 años y la mediana de 66 años. Además, veamos su diagrama de cajas en la figura 4.1. Creamos en este caso una nueva variable, gr_edad, que divide a los individuos en dos grupos; uno para los de 65 años o menores, de los cuales tenemos 236, y otro para los mayores de 65, de los cuales tenemos 264.

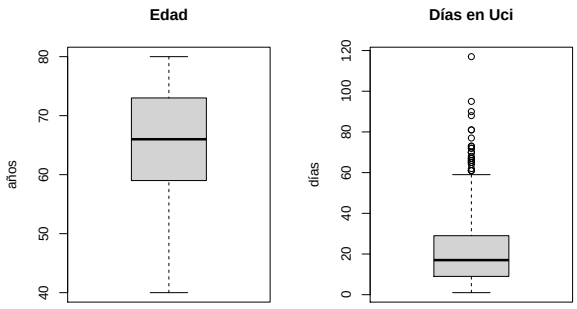

Figura 4.1: Diagrama de cajas para edad y dias en UCI.

- Días en la UCI. Variable numérica de la que podemos ver un resumen en la tabla 7. Nos proporciona el tiempo en UCI de cada paciente y viene dada en días. Tiene una moda de 9 días y una mediana de 17 días, veamos su diagrama de cajas en la figura 4.1.
- Hipertensión arterial. Variable factor que nos proporciona información sobre si el individuo tiene hipertensión arterial o no. Hay 262 pacientes que sí tienen y 238 que no.
- Estado. Variable factor que hemos comentado previamente. En la muestra; 275 pacientes fueron dados de alta, 182 murieron y 43 quedaron censurados.

Podemos observar alguna de las relaciones entre las diferentes variables. Por ejemplo, haciendo un boxplot, Figura 4.2, entre edad-estado y días en UCI-estado. Podemos ver que las edades en las que se

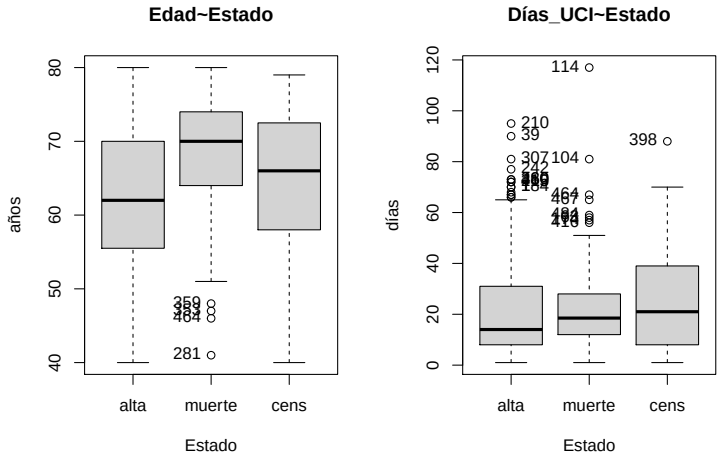

Figura 4.2: Relación entre la edad y los días en UCI con el estado.

concentran las muertes, son más altas que aquellas de los que reciben el alta. Por otro lado, podemos ver que los días en UCI para los que finalmente mueren toman unos valores más concentrados que los del alta.
Realizando otro boxplot, Figura 4.3, entre edad-hta y edad-sexo, observamos que la presencia de hipertensión arterial es más común en pacientes de edades superiores, variables que podrían estar relacionadas. Además, cabe destacar que para la edad entre los hombres y las mujeres que entran en la UCI no se aprecia gran diferencia.

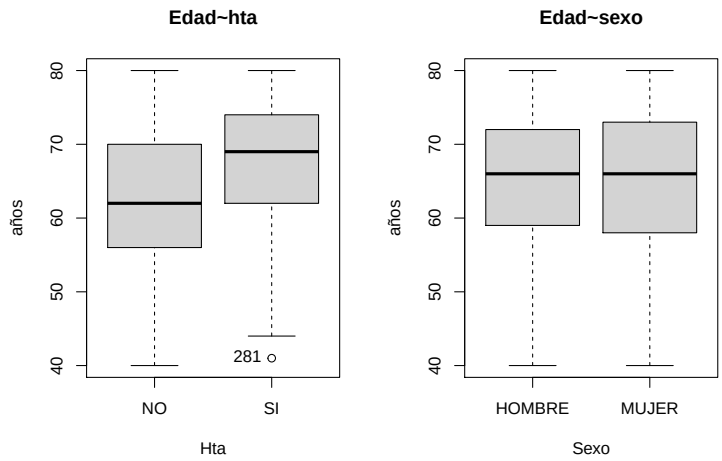

Figura 4.3: Relación entre la edad con el hta y el sexo.

Podemos obtener también la tabla de porcentajes 8, donde ya podemos apreciar que la mayor diferencia es con respecto a los grupos de edad. Se producen más proporción de muertes para pacientes de edades superiores a 65 años y altas para los de edades inferiores. Notar que también una mayor proporción de muertes tiene lugar para los hombres que para las mujeres. Con respecto a la hipertensión arterial, notar que la proporción de pacientes con hta que mueren es mayor que para los que no tienen. Para el alta los porcentajes son semejantes aunque ligeramente superior sobre aquellos que no tienen hta.

4.2. Estimación de la función de incidencia acumulada.

Hablamos de riesgos competitivos cuando observamos más de un resultado posible en el seguimiento del tiempo de supervivencia. En nuestro caso, el suceso de interés es la salida de la UCI. Notar que los individuos pueden salir de la UCI o bien porque se les ha dado el alta o bien porque han fallecido. En presencia de riesgos competitivos, la posible dependencia entre los diferentes sucesos de interés hace de este tipo de análisis diferente del que no los considera.
La función de incidencia acumulada proporciona la probabilidad de fallo por la causa dada antes de un tiempo concreto, y antes de que tenga lugar cualquier otro riesgo competitivo. Sin riesgos competitivos utilizaríamos el estimador de KM (2.2), pero al considerar riesgos competitivos este puede proporcionar resultados sesgados. Utilizamos la función **cuminc()** del paquete cmprsk [3] que estima las CIF para las diferentes causas de fallo (3.6) y permite comparaciones entre grupos.
En primer lugar, aplicamos esta función al conjunto de datos, sin considerar diferencias entre grupos de sexo, edad... Los resultados obtenidos están en la tabla 9.
Nos proporciona los estimadores para los tiempos t=20,40,60,80,100. Sabemos que los días en UCI varían de 1-117. Los estimadores obtenidos nos proporcionan, por ejemplo, la siguiente información:

- La probabilidad de recibir el alta de la UCI antes de 40 días es de 0.475.
- La probabilidad de salir de fallecer antes de 40 días, se entiende que el paciente no ha recibido el alta con anterioridad, es de 0.344.

Utilizando la orden **ggcompetingrisks()** del paquete survminer [5] podemos obtener un gráfico de la CIF y así contrastar visualmente los resultados que acabamos de comentar. Únicamente necesita de un objeto de clase cuminc() como argumento. Realizando el gráfico 4.4 de la CIF para la estimación que hemos obtenido previamente, vemos que existe una mayor probabilidad de obtener el alta que de

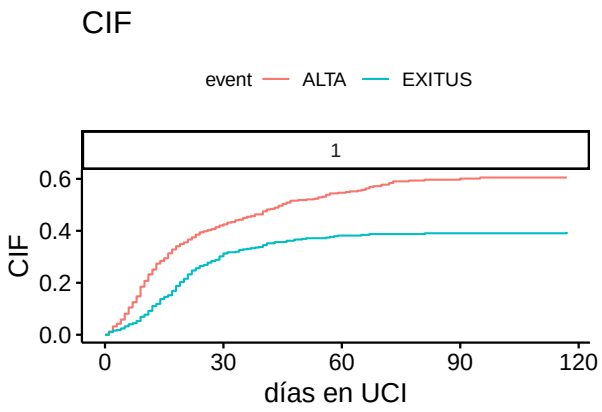

Figura 4.4: Función de incidencia acumulada para cada causa.

fallecer, independientemente del tiempo en UCI.

Además, con el argumento 'group' de la función cuminc() podemos realizar comparaciones entre grupos. En este caso, no solo obtenemos los estimadores de la CIF, sino también un test del estadístico de la χ^2 y sus respectivos p-valores para cada causa, que permiten comparar la CIF para cada causa en función de determinadas covariables (3.7). Vamos a realizar comparaciones con el sexo, con el grupo de edad y con la hipertensión arterial.

- Sexo. Distinguimos entre hombres y mujeres. En la salida Tests 10, la primera columna representa el estadístico χ^2 para el test entre grupos y la segunda su respectivo p-valor. Por un lado, el p-valor para el alta es 0.0015<0.05, luego es significativo. Por otro lado, el p-valor para la muerte es 0.013<0.05, también significativo.
  Notar en la tabla 11 que para los hombres la probabilidad recibir el alta antes de 40 días es de 0.430 y para las mujeres de 0.576. Sin embargo, los hombres tienen una probabilidad de morir antes de 40 días de 0.380 y las mujeres de 0.263. Como podemos observar en el gráfico 4.5

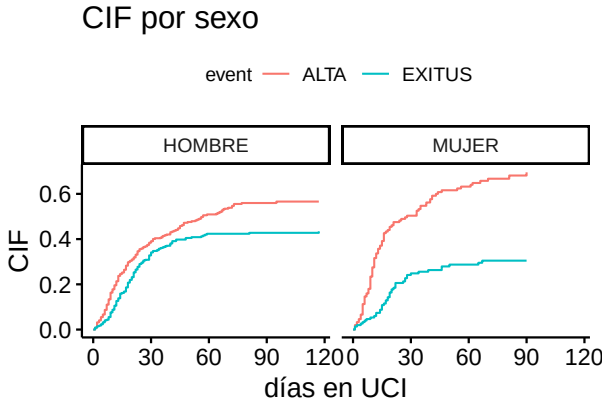

Figura 4.5: CIF por sexo.

  comprobamos que efectivamente, el sexo es significativo. Los hombres tienen probabilidades parecidas, aunque superior para el alta, de experimentar ambos sucesos. Mientras que las mujeres tienen una probabilidad bastante superior de recibir el alta que de morir.

- Hipertensión arterial. Distinguimos entre los que tienen y los que no. Obtenemos en el test 12 un p-valor para el alta de 0.0906>0.05, que no es significativo. Y para la muerte, de 0.0296<0.05, que si es significativo, aunque ambos valores están al límite. Luego, la hipertensión arterial si influye

a la hora de morir, una vez entrado en la UCI. Aquellos pacientes con hta tienen más probabilidad de morir que aquellos que no.
Observamos en 13 que la probabilidad de no tener hta y recibir el alta antes de 40 días es de 0.506 y en caso de tenerla, de 0.447. Por otro lado, la probabilidad de morir antes de 40 días, sin haber recibido el alta, es de 0.297 para los que tienen hta y de 0.386 para los que no. Podemos visualizarlo en el gráfico 4.6.

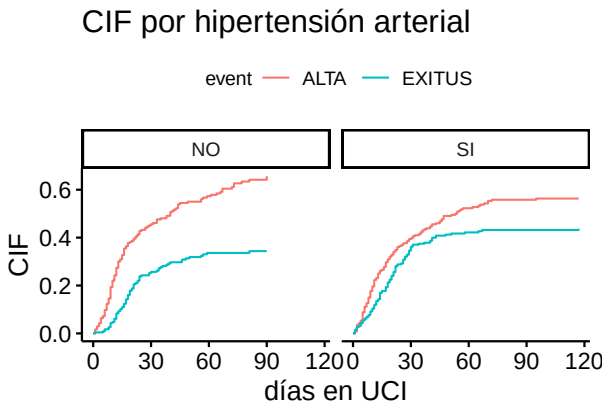

Figura 4.6: CIF por hipertensión arterial.

- Edad. En este caso, usamos la variable *gr_edad*, para así poder hacer la comparación para dos grupos. Con el test 14 obtenemos un p-valor para el alta de 1.565e-10<0.05, luego es significativo. Y para la muerte, 8.609e-10<0.05, que también es significativo.
  Notar que obtenemos de la tabla 15 que la probabilidad de ser menor de 65 y recibir el alta antes de 40 días es de 0.632 y para los mayores de 65 de 0.337. Por otro lado, la probabilidad de fallecer antes de 40 días es 0.207 para los menores de 65 y 0.464 para los mayores. Apreciamos

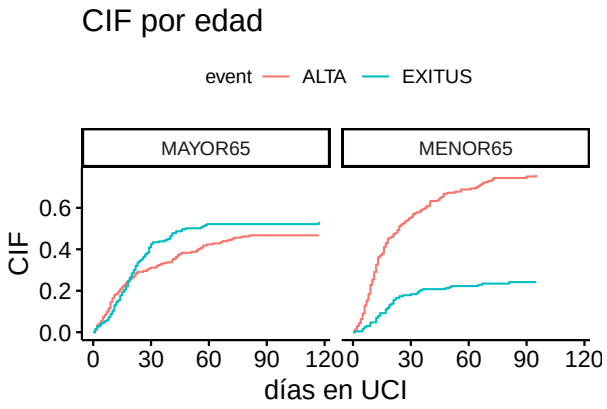

Figura 4.7: CIF por grupo de edad.

en el gráfico 4.7 una gran diferencia. Para los mayores de 65 años las probabilidades de recibir el alta y de morir son semejantes, aunque ligeramente superior la de morir. Sin embargo, para los del grupo contrario es bastante superior la del alta que la de morir.

4.3. Modelo de regresión en presencia de riesgos competitivos.

Aplicamos los resultados de la sección 3.4 para construir un modelo que ajuste las funciones de riesgo dependiendo de las covariables.

La función **CSC()** del paquete riskRegression [1] produce automáticamente modelos de riesgo específicos de la causa para cada causa. El argumento principal es una fórmula; Hist o una lista de estas. Hist permite tratar los datos censurados por la derecha y por intervalos en presencia de riesgos competitivos. Vamos a ajustar el siguiente modelo:

$$CSH <- CSC(Hist(dias\_uci, resultado\_num) \sim sexo + edad + hta, data = datos\_elisa\_s)$$

en el que se incluyen los efectos principales de todas las variables. Notar que hemos incluido la variable edad como continua, y no por grupos. Obtenemos dos resultados, uno para cada causa.

- Alta. Como podemos observar en la tabla 4.1 para el caso del alta son significativas tanto la edad como el sexo.

	coef	exp(coef)	se(coef)	z	p-valor
sexo	0.335	1.398	0.126	2.649	0.00807 **
edad	-0.035	0.965	0.006	-5.463	4.68e-08 ***
hta	0.110	1.116	0.123	0.894	0.371

Tabla 4.1: Salida de la función CSC() para el alta.

- Muerte. En este caso, en la tabla 4.2 para la muerte, la única variable significativa es la edad.

	coef	exp(coef)	se(coef)	z	p-valor
sexo	-0.252	0.777	0.175	-1.439	0.150
edad	0.049	1.051	0.009	4.996	5.86e-07 ***
hta	0.210	1.234	0.152	1.377	0.169

Tabla 4.2: Salida de la función CSC() para la muerte.

Así, podemos ajustar un modelo cuyo argumento principal sea una lista. De manera que, para el alta, las variables que entran en el modelo son la edad y el sexo y, para la muerte, únicamente la edad.

$$CSCajus <- CSC(formula = list(Hist(dias\_uci, resultado\_num) \sim edad + sexo,$$

$$Hist(dias\_uci, resultado\_num) \sim edad), data = datos\_elisa\_s)$$

Este modelo ajustado se puede utilizar para predecir el riesgo de un determinado individuo dadas sus covariables y proporciona las tablas 16 y 17, para el alta y muerte, respectivamente.

Podemos darle una interpretación a los coeficientes obtenidos. Para el alta, el exp(coef) para la edad toma valor 0.966, con lo cual, un individuo tiene 3,4 % menos de probabilidad de recibir el alta en un día determinado que otro que tiene un año menos y el exp(coef) para el sexo toma valor 1.392, es decir, una mujer tiene aproximadamente un 40 % más de probabilidad de recibir el alta un día determinado que un hombre. Por otro lado, para la muerte, el exp(coef) para la edad toma valor 1.051, por lo que un individuo tiene un 5 % más de probabilidad de morir un día determinado que uno que tenga un año menos.

Por otra parte, notemos que aunque el sexo no está incluido en el modelo del riesgo de muerte, esto no quiere decir que no tenga influencia en la probabilidad de muerte de los pacientes, ya que sí está incluido en el riesgo de alta. Como veremos a continuación, el hecho de que las mujeres tengan mayor riesgo de alta, hace que su probabilidad de muerte disminuya respecto a los hombres.

sexo	edad	hta	tiempo	probAlta	probMuerte
mujer	45	si	40	0.849	0.0754
hombre	45	si	40	0.753	0.0994
mujer	75	si	40	0.390	0.4725
hombre	75	si	40	0.303	0.5229

Tabla 4.3: Predicción.

Una vez tenemos el modelo ajustado, podemos utilizarlo para predecir nuevas observaciones para individuos teniendo el valor de sus covariables y para un tiempo dado. La función predict.CauseSpecificCox del paquete riskRegression [1] permite realizar esta predicción. Tomamos, por ejemplo, una mujer y un hombre de 45 años y una mujer y un hombre de 75 años.

Como vemos en la tabla 4.3 para el día 40 tanto el hombre como la mujer de 45 años tienen mucha más probabilidad de recibir el alta que de morir. Sus probabilidades de morir son bastante pequeñas y similares. Mientras que para el alta, la mujer tiene un 85 % de probabilidad de recibirla y el hombre un 75 %. Por otro lado, los de 75 años, tienen ambos mayor probabilidad de morir que de recibir el alta. Además, la mujer tiene una mayor probabilidad de recibir el alta, un 39 %, que el hombre, un 30 %, y una menor probabilidad de morir, un 47 %, que el hombre, un 52 %.
Además, como hemos mencionado antes, los hombres tienen una ligera probabilidad mayor que las mujeres de morir antes de 40 días.

4.4. Comprobación de la hipótesis de riesgos proporcionales.

La hipótesis de riesgos proporcionales es una hipótesis que ha de comprobarse para cada causa. Podemos recurrir entonces a la función **cox.zph()** del paquete survival [12] para ver si esta se cumple para cada causa por separado. Esta función nos aporta dos resultados. Por un lado, un test de hipótesis para ver si el efecto de cada covariable varía con el tiempo y, por otro, una prueba global de todas las covariables.

- Para el alta. La hipótesis nula es la de proporcionalidad; en la tabla 4.4 obtenemos un p.valor de 0.057, por lo que trabajando al nivel de significación de 0,05, no rechazamos la hipótesis.

	Chisq	Df	p-valor
edad	3.78	1	0.052
sexo	2.50	1	0.114
GLOBAL	5.73	2	0.057

Tabla 4.4: Test de proporcionalidad para el alta.

- Para la muerte. Para el caso de la muerte, vemos en la tabla 4.5 que el p-valor que obtenemos es de 0.92>0.05, con lo cual tampoco rechazamos la hipótesis nula.

	Chisq	Df	p-valor
edad	0.0108	1	0.92
GLOBAL	0.0108	1	0.92

Tabla 4.5: Test de proporcionalidad para la muerte.

En ninguno de los casos rechazamos la hipótesis nula. Luego, podemos suponer que se cumple el supuesto de riesgos proporcionales.

Bibliografía

[1] GERDS T.A., BLANCHE P., MORTENSEN R., WRIGHT M., TOLLENAAR N., MUSCHELLI J., MOGENSEN U.B., OZENNE B., *Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*, 2009. https://cran.r-project.org/web//packages/riskRegression/riskRegression.pdf

[2] GRAMBSCH P.M., THERNEAU, T. M. , *Proportional hazards tests and diagnostics based on weighted residuals*, Biometrika, Vol. 81, No. 3, 1994, 515-526. https://doi.org/10.1093/biomet/81.3.515

[3] GRAY B., *Subdistribution Analysis of Competing Risks* , 2009. https://cran.r-project.org/web//packages/cmprsk/cmprsk.pdf

[4] GRAY R.J., *A class of K-sample tests for comparing the cumulative incidence of a competing risk.*, The Annals of statistics, Vol.16, No. 3, 1988, 1141-1154. https://www.jstor.org/stable/2241622?seq=1#metadata_info_tab_contents

[5] KASSAMBARA A, KOSINSKI M, BIECEK P, FABIAN S, *Drawing Survival Curves using 'ggplot2'*, 2009. https://cran.r-project.org/web/packages/survminer/survminer.pdf

[6] KLEINBAUM D.G., KLEIN M., *Survival Analysis: A Self-Learning Text*, Springer-Verlag, New York, 2012.

[7] LAWLESS J.F., *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York, 1982.

[8] MARTÍNEX-GONZÁLEZ M.A., ALONSO A., LÓPEZ-FIDALGO J., *¿Qué es una hazard ratio? Nociones de análisis de supervivencia*, Medicina Clínica, Volume 131, No. 2, June 2008, 65-72. https://www.researchgate.net/profile/Jesus-Lopez-Fidalgo/publication/246617804_Que_es_una_hazard_ratio_Nociones_de_analisis_de_supervivencia/links/5c121de44585157ac1be754f/Que-es-una-hazard-ratio-Nociones-de-analisis-de-supervivencia.pdf

[9] RAMÍREZ J, REGINO E., GUERRERO S.Y., *Comparación de métodos de estimación de regresión de Cox*, Comunicaciones en Estadística, Vol. 10, No. 1, 2017, 101-112. https://dialnet.unirioja.es/servlet/articulo?codigo=6765746

[10] SCHUSTER N.A., HOOGENDIJK E.O., KOK A.A.L., TWISK J.W.R., HEYMANS M.W., *Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risks analysis*, Journal of Clinical Epidemiology, Volume 122, 2020, 42-48. https://www.jclinepi.com/article/S0895-4356(19)31061-3/fulltext

[11] THERNEAU T.M., GRAMBSCH P.M., *Modelling survival data: Extending the Cox model*, Springer-Verlag, 2000.

[12] THERNEAU T.M., LUMLEY T., ATKINSON E., CROWSON C., *Survival Analysis* , 2021. https://cran.r-project.org/web/packages/survival/survival.pdf

[13] VICENZI M., DI COSOLA R., RUSCICA M., RATTI A., ROTA I., ROTA F., BOLLATI V., ALIBERTI S., BLASI F., *The liaison between respiratory failure and high blood pressure: evidence from COVID-19 patients*, European Respiratory Journal, 2020, 56:2001157. https://erj.ersjournals.com/content/56/1/2001157.full

[14] YI Y., WANG X., *Comparison of Wald, Score, and likelihood ratio tests for response adaptive designs*, Journal of Statistical Theory and Applications, Vol. 10, No.4, 2011, 553-569. http://www.mscs.mu.edu/~jsta/issues/10(4)/JSTA10(4)p2.pdf

[15] ZHANG Z, *Survival analysis in the presence of competing risks*, Annals of Translational Medicine, Volume 5, No. 3, 2017, 47. https://atm.amegroups.com/article/view/11637/13922

Anexo I

Este anexo se presenta como complemento al Capítulo 4, incluyendo alguna de las tablas comentadas en esa sección.

	Min-	1st Qu.	Median	Mean	3rd Qu.	Max.
edad	40.00	59.00	66.00	64.82	73.00	80.00

Tabla 6: Resumen de la variable edad.

	Min-	1st Qu.	Median	Mean	3rd Qu.	Max.
días	1.0	9.0	17.0	22.1	29.0	117.0

Tabla 7: Resumen de la variable días en UCI.

		Alta	Censura	Muerte
Edad	≤ 65	69.0	8.8	22.0
	>65	42.4	8.3	49.2
Sexo	Hombre	51.1	8.8	39.9
	Mujer	63.8	7.8	28.2
Hta	Si	53.0	5.3	41.6
	No	57.1	12.1	30.6

Tabla 8: Porcentaje de pacientes que experimentan un suceso en función de las variables.

		20	40	60	80	100
Est	Alta	0.355	0.475	0.546	0.593	0.605
	Muerte	0.215	0.344	0.381	0.387	0.390

Tabla 9: Estimación de la CIF.

	χ^2	p-valor	df
Alta	10.0111	0.0015	1
Muerte	6.1223	0.0133	1

Tabla 10: Test de la CIF por sexo.

		20	40	60	80	100
Est	Hom. Alta	0.309	0.430	0.509	0.559	0.565
	Muj. Alta	0.461	0.576	0.632	0.667	NA
	Hom. Muerte	0.231	0.380	0.423	0.423	0.427
	Muj. Muerte	0.177	0.263	0.287	0.304	NA

Tabla 11: Estimación de la CIF por sexo.

	χ^2	p-valor	df
Alta	2.8632	0.0906	1
Muerte	4.7311	0.0296	1

Tabla 12: Test de la CIF por hipertensión.

		20	40	60	80	100
Est	No Alta	0.386	0.506	0.572	0.634	NA
	Si Alta	0.328	0.447	0.522	0.558	0.563
	No Muerte	0.187	0.297	0.336	0.336	NA
	Si Muerte	0.240	0.386	0.421	0.431	0.431

Tabla 13: Estimación de la CIF por hta.

	χ^2	p-valor	df
Alta	40.9457	1.565e-10	1
Muerte	37.6169	8.609e-10	1

Tabla 14: Test de la CIF por grupo de edad.

		20	40	60	80	100
Est	$\leq$ 65 Alta	0.460	0.632	0.688	0.743	NA
	>65 Alta	0.263	0.337	0.423	0.461	0.467
	$\leq$ 65 Muerte	0.137	0.207	0.222	0.234	NA
	>65 Muerte	0.284	0.464	0.521	0.521	0.521

Tabla 15: Estimación de la CIF por grupo de edad.

	coef	exp(coef)	se(coef)	z	p-valor
edad	-0.034	0.966	0.006	-5.390	7.04e-08 ***
sexo	0.330	1.392	0.126	2.615	0.00892 **

Tabla 16: Salida de la función CSCajus() para el alta.

	coef	exp(coef)	se(coef)	z	p-valor
edad	0.050	1.051	0.009	5.122	3.03e-07 ***

Tabla 17: Salida de la función CSCajus() para el muerte.

Anexo II

En este anexo incluimos el código implementado en R para obtener los resultados del Capítulo 4.

```r
library(riskRegression)
library(car)
library(survival)
library(KMsurv)
library(survMisc)
library(survminer)
library(ggfortify)
#install.packages("flexsurv")
library(flexsurv)
library(ggplot2)
library(dplyr)
library(htmlwidgets)

#ANÁLISIS DESCRIPTIVO
load("C:/Users/cpa5e/OneDrive/Escritorio/4/TFG/R/datos_elisa_def.RData")
str(datos_elisa_s)

summary(datos_elisa_s$edad)
range(datos_elisa_s$edad)
table(datos_elisa_s$gr_edad)

summary(datos_elisa_s$dias_uci)
range(datos_elisa_s$dias_uci)

summary(datos_elisa_s$resultado)
table(as.factor(datos_elisa_s$resultado))/length(datos_elisa_s$resultado)*100

summary(datos_elisa_s$hta)
table(as.factor(datos_elisa_s$hta))/length(datos_elisa_s$hta)*100

summary(as.factor(datos_elisa_s$sexo))
table(as.factor(datos_elisa_s$sexo))/length(datos_elisa_s$sexo)*100

library(modeest)
mlv(datos_elisa_s$edad, method="mfv")
mlv(datos_elisa_s$dias_uci, method="mfv")

par(mfrow = c(1, 2))
boxplot(datos_elisa_s$edad, main="Edad", ylab="años")
boxplot(datos_elisa_s$dias_uci, main="Días en Uci", ylab="días")

par(mfrow = c(1, 2))
Boxplot(datos_elisa_s$edad~datos_elisa_s$resultado, main="Edad~Estado", ylab=
  "años", xlab="Estado", names=c("alta","muerte","cens"))
Boxplot(datos_elisa_s$dias_uci~datos_elisa_s$resultado, main="Días_UCI~Estado
  ", ylab="días", xlab="Estado", names=c("alta","muerte","cens"))

par(mfrow = c(1, 2))
Boxplot(datos_elisa_s$edad~datos_elisa_s$hta, main="Edad~hta", ylab="años",
  xlab="Hta")
```

```
48 Boxplot(datos_elisa_s$edad~datos_elisa_s$sexo, main="Edad~sexo", ylab="años",
     xlab="Sexo")
49
50 datos_elisa_s$resultado_num<-as.numeric(factor(datos_elisa_s$resultado,
    levels=c("CENSURA","ALTA","EXITUS")))-1
51
52 datos_elisa_s$sexo_num<-as.numeric(factor(datos_elisa_s$sexo, levels=c("
    HOMBRE","MUJER")))-1
53
54 datos_elisa_s$hta_num<-as.numeric(factor(datos_elisa_s$hta, levels=c("SI","NO
    ")))-1
55
56 #ESTIMACIÓN DE LA FUNCIÓN DE INCIDENCIA ACUMULADA
57 library(cmprsk)
58 cif<-cuminc(datos_elisa_s$dias_uci, datos_elisa_s$resultado, cencode='CENSURA
    ')
59 cif
60 library(ggfortify)
61 ggcompetingrisks(cif)
62
63 #Realicemos ahora comparaciones por grupos.
64 #SEXO.
65 cif_sexo<-cuminc(datos_elisa_s$dias_uci, datos_elisa_s$resultado,datos_elisa_
    s$sexo, cencode='CENSURA')
66 cif_sexo
67 ggcompetingrisks(cif_sexo, main = "CIF por sexo",xlab='días en UCI', ylab='
    CIF')
68 #HIPERTENSIÓN ARTERIAL.
69 cif_hta<-cuminc(datos_elisa_s$dias_uci, datos_elisa_s$resultado,datos_elisa_s
    $hta, cencode='CENSURA')
70 cif_hta
71 ggcompetingrisks(cif_hta)
72 #EDAD
73 datos_elisa_s$gr_edad<-cut(datos_elisa_s$edad, breaks=c(0,65,100), labels = c
    ("MENOR65","MAYOR65"))
74 cif_edad<-cuminc(datos_elisa_s$dias_uci, datos_elisa_s$resultado,datos_elisa_
    s$gr_edad, cencode='CENSURA')
75 cif_edad
76 ggcompetingrisks(cif_edad)
77
78 #MODELO DE REGRESIÓN EN PRESENCIA DE RIESGOS COMPETITIVOS
79 CSH <-CSC(Hist(dias_uci,resultado_num)~sexo+edad+hta,data=datos_elisa_s)
80 CSH
81
82 #Modelo ajustado
83 CSCajus <- CSC(formula = list(Hist(dias_uci,resultado_num)~edad+sexo,
84 Hist(dias_uci,resultado_num)~edad), data = datos_elisa_s)
85 CSCajus
86
87 #PREDICCIÓN DEL MODELO
88 nuevosdatos<-data.frame(sexo=c("MUJER","HOMBRE","MUJER","HOMBRE"),hta=factor(
    c("SI","SI","SI","SI"),levels=levels(datos_elisa_s$hta)),edad=c(45,45,75,75)
    )
89 nuevosdatos
90 nuevosdatosmat<-model.matrix(~sexo+hta+edad, data=nuevosdatos)[,-1]
91 nuevosdatosmat
92
93 pred1<-predict(CSCajus, nuevosdatos, 40, cause=1)
94 pred1
95
96 pred2<-predict(CSCajus, nuevosdatos, 40, cause=2)
97 pred2
98
```

```
99  #COMPROBACIÓN DEL MODELO
100 csh_alta<- coxph(Surv(dias_uci,resultado=='ALTA')~edad+sexo,data=datos_elisa_
      s)
101 summary(csh_alta)
102 cox.zph(csh_alta, transform = "rank")
103
104 csh_muerte<- coxph(Surv(dias_uci,resultado=='EXITUS')~edad,data=datos_elisa_s
      )
105 summary(csh_muerte)
106 cox.zph(csh_muerte, transform = "rank")
```