Métodos estocásticos en medicina: ANÁLISIS DE SUPERVIVENCIA

Rocío Aznar Gimeno

Trabajo de fin del grado de Matemáticas

Universidad de Zaragoza

Summary

The survival analysis is a set of techniques whose aim is the study of the variable "time to the ocurrence of a given event". This event can be, for example, the development of a disease, response to a treatment or the death. This random continuos variable is called *Survival Time* and is denoted by T.
Survival analysis is an essential tool to the study the evolution of diseases but also has great influence in other areas of knowledge, for example, engineering.

Our work focuses on survival analysis applied to medicine. Firstly, we introduce the basic concepts of survival analysis (chapter 1); secondly, we explain the usual nonparametric methods for estimating and comparating and semiparametric models of the survival analysis (chapter 2 and chapter 3) for, finally, apply them to a particular group of individuals with prostate cancer (chapter 4).

Introduction of survival analysis

One of the most important feature of survival data is the lack information of the data. This occurs when some subjects in the study have not experienced the event of interest at the end of the study. For example, some patients may still be alive or disease-free at the end of the study period. This is because, in the practice, it is almost impossible to have lengthy studies that allow collect full information on all patients. So, the exact survival times of these subjects are unknown and these are called *censored observations or censored times.*
It is said that an individual *survives* if has not suffered yet to the event. In this case, it is said that the individual *is at risk.* Furthermore, it is said that an individual *has failed* if he has suffered the event ; consequently it is called failure time to time when the event occurred.

The distribution of survival times is usually described or characterized by three functions: the survivorship function, the density function and the hazard function. These three functions are mathematically equivalent.
The survivorship function, denoted by S(t), is defined as the probability that an individual survives longer than t, that is,

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

The hazard function, denoted by h(t), is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval, per unit time, that is,

$$h(t) = \lim_{\triangle t \to 0^+} \frac{P\left(t < T \leq t + \triangle t \mid T > t\right)}{\triangle t}.$$

Furthermore, the cumulative hazard function is defined as:

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

The function of density of T is defined as usual.

In addition to the censoring problem, we do not usually know specific hypotheses about the distribution of the random variable T, so we do not dispose of a parametric model for the survival functions . For that reason, we need to incorporate nonparametric methods (not assumed any particular form in the variable T) of estimating survival functions.

Nonparametric estimators

The Kaplan-Meier estimator is a nonparametric estimator of the survivorship function $S(t)$, introduced by Edward L.Kaplan and Paul Meier, and which includes information on all data, censored and uncensored. In addition, Kaplan and Meier assume independence between the time of entry into the study and the probability of failure.
Suppose a study of n individuals where their failure times or censorship (uncensored and censored times) are known. We denoted $t_{(1)}, \ldots, t_{(s)}$ with $s \leq n$ to the different ordered times of failure. The Kaplan-Meier estimator $S(t)$ is defined as

$$\hat{S}(t) = \prod_{j, t_{(j)} \leq t} \frac{n_j - d_j}{n_j}, \quad j = 1, \ldots, s \qquad y \qquad \hat{S}(t) = 1 \quad si \quad t < t_{(1)}.$$

where n_j is the number of individuals that are at risk at the time $t_{(j)}$ and d_j the number of individuals that fail at the time $t_{(j)}$.
The Kaplan-Meier estimator can be derivated as a maximum likelihood estimator [8]. In this way, confidence intervals for large samples for the function $S(t)$ can be obtained. The variance of the Kaplan-Meier estimator can be obtained by applying the Delta method [7].

Nelson y Aalen propose a nonparametric estimator of the cumulative hazard function given by

$$\tilde{H}(t) = \sum_{j; t_{(j)} \leq t} \frac{d_j}{n_j},$$

where $t_{(j)}, d_j, n_j$ are defined as before.
As for the Kaplan-Meier estimator, the estimator can be derived also as a maximum likelihood estimator [1].

Nonparametric test. Test de Logrank

The problem of comparing survival distributions arises often in medical research. Suppose, for example, that we want to compare the tumor-free times for two groups of people, G_1, G_2, that received treatments 1 and 2, respectively. To compare these groups, we explain and use the Logrank test, a nonparametric test that can be used for data with and without censored observations. The null hypothesis is $H_0 : S_1(t) = S_2(t)$, that is, treatments 1 and 2 are equally effective.
In this situation, we know ([11]) that the statistical

$$Z = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed number of patients who relapse and E_i are the expected number of patients who relapse in the group G_i $i = 1, 2$, follows, under the null hypothesis of equality of the survival functions, a distribution χ^2_1 (asymptotically).

Until now, we have only considered the estimate of the survival function as a function of time. Now we propose a model that analyze the influence of other variables.

Cox model

Cox was the first to propose a semiparametric regression model, called Cox model or proportional hazards model, that is represented as follows:

$$h(t, X, \beta) = h_0(t) \exp(X\beta),$$

where X is the covariate, β the unknown coefficient asociated to the covariate and $h_0(t)$ a function called baseline hazard function. Note that the model is valid for several covariables. Cox regression allows us to evaluate, from a set of independent variables, which of them have a significant influence on the hazard function. To evaluate the influence of each variable, we use the Wald test, whose null hypothesis is $H_0 : \beta = 0$.

The main element in the Cox regression is the hazard ratio that allows us to compare the hazard function between two groups of patients. Unlike Logrank test, the hazard ratio provides us information about the magnitude of the differences between groups.
In the Cox model, the hazard ratio is expressed as:

$$HR(x_i, x_j) = \frac{h(t, x_i, \beta)}{h(t, x_j, \beta)} = \exp(\beta(x_i - x_j)),$$

where x_i and x_j indicate the value of the covariable of the two groups.
The above expression shows that the hazard ratio does not depend of the time; this is the proportional hazards assumption .

The coefficients β are unknown coefficients that we must estimate. To estimate the coefficients, Cox [4] proposed a partial likelihood function. Later, it was shown that the partial likelihood function had the same properties as the maximum likelihood estimators [2]. This fact allow to get confidence intervals for large samples for the coefficientes β.

Application

In the last chapter, we apply all the previous concepts to a real problem.
We will do a survival analysis to a group of 359 individuals with prostate cancer. The data belong to the hospital "Miguel Servet". The individuals are incorporated to the study after surgery (radical prostatectomy). The event of interest occurs when the value of PSA $> 0,4 ng/mL$.

The aim of our analysis is to study the influence of some preoperative variables and the age of the patient in the operation, on the survival time, this is, the free time of cancer. The study will carry out, mainly, with the SSPS statistical program and the R program.
The results of this analysis are shown in the work (chapter 4).

Índice general

Capítulo 1

Introducción al análisis de supervivencia

El análisis de supervivencia es un conjunto de técnicas que tienen como objeto estudiar la variable 'tiempo hasta la ocurrencia de un suceso de interés' y su dependencia con otras posibles variables. A esta variable aleatoria se le conoce como **tiempo de supervivencia** y se denota por T.

1.1. Introducción histórica

El análisis de supervivencia tiene su origen en la construcción de tablas de mortalidad, de donde proviene el término de 'supervivencia' pues el suceso de interés allí era la muerte. Las primeras tablas de mortalidad fueron construidas por el astrónomo inglés Edmond Halley (1656-1742) a partir de los registros de nacimientos y funerales de la ciudad de Breslau (1693).

Posteriormente, el análisis de supervivencia fue extendido al campo de la ingeniería para analizar la fiabilidad de los diferentes elementos que forman una máquina. La Segunda Guerra Mundial aceleró el desarrollo de estas técnicas para aplicarlas a la industria militar.

El auge del análisis de supervivencia en medicina empezó en la década de los 70 jugando un papel muy importante en el estudio de la evolución de enfermedades y más tarde también gracias al avance tecnológico.

En las últimas décadas, los modelos estadísticos para el análisis de supervivencia han continuado progresando extendiéndose su aplicación a otras áreas como son la economía, criminología o las ciencias sociales y del comportamiento. De esta manera, hoy en día, el suceso de interés para el estudio no es necesariamente la muerte como lo era en su origen si no que puede ser la recaída o desarrollo de una cierta enfermedad, fallo en una pieza de un máquina, una determinada subida en la prima de riesgo o la salida de la universidad, por ejemplo.

Así, el análisis de supervivencia se extiende y se aplica a otras áreas de conocimiento además del área de la medicina que es en la que nos centramos en nuestro trabajo.

1.2. Conceptos básicos del análisis de supervivencia

Con el fin de facilitar la visualización de los conceptos básicos que definiremos a continuación vamos a considerar la siguiente situación:

Supongamos que se realiza un estudio durante un año observando cada mes a un grupo de 6 personas que padecieron un tipo de enfermedad y que, tras someterlas a una intervención quirúrgica y quedar, aparentemente, libres de dicha enfermedad, se quiere estudiar la distribución de la variable aleatoria tiempo de supervivencia, siendo el evento la recaída en dicha enfermedad, esto es, el tiempo libre de enfermedad.

En la práctica, suele ser imposible hacer un estudio que permita conocer los tiempos de supervivencia exactos de los pacientes por lo que se consideran los tiempos de observación en el estudio como tiempos de supervivencia. De esta manera, en ocasiones, debido a la forma en la que observamos a los pacientes, los tiempos de supervivencia se miden de forma discreta como ocurre, por ejemplo, en la situación anterior donde los tiempos de supervivencia podrán tomar el valor $t \in \mathbb{N}$ con $t = 1, \ldots, 12$. Notemos que es importante la unidad de tiempo que se considera pues, dependiendo de ésta, la información puede ser más o menos refinada.

1.2.1. Terminología básica

Hoy en día, se dice que un individuo *sobrevive* si no ha sufrido todavía el suceso. En este caso, se dirá que dicho individuo *está en riesgo*. Por otro lado, se dice que un individuo *ha fallado* si ha sufrido el suceso de interés; consecuentemente se denomina *tiempo de fallo* al tiempo en el que ha ocurrido el suceso.

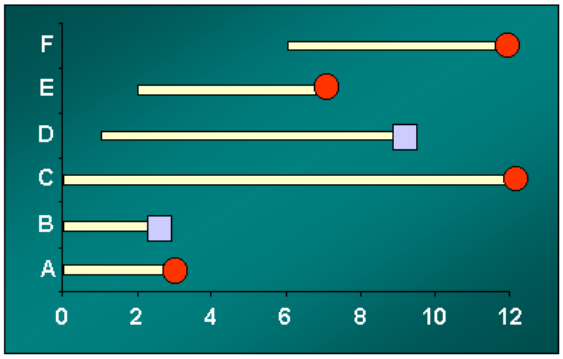
Datos censurados

En un estudio, como el de la situación anterior, puede que todos los pacientes hayan fallado y se conozcan sus tiempos de ocurrencia, es decir, contamos con la información completa de todos sus tiempos de supervivencia. Los datos de supervivencia conocidos con exactitud reciben el nombre de **exactos o no censurados**.
Sin embargo, la obtención de muestras con información completa de todos los pacientes suele requerir estudios demasiado largos que no se dan en la práctica. De esta forma, habitualmente no conocemos los tiempos de supervivencia exactos de todos los pacientes. En este caso, algunos tiempos de supervivencia son conocidos parcialmente (no conocemos su tiempo de fallo) y reciben el nombre de **datos u observaciones censuradas**. Por ejemplo, en el caso de la situación anterior, puede ocurrir que algunos pacientes puedan estar libres de la enfermedad al final del periodo de estudio o incluso puede ocurrir que se pierda su seguimiento (por abandono o muerte por otras causas, por ejemplo). En estos casos, se conoce solamente una cota de sus tiempos de supervivencia. En este sentido, existen tres tipos de censura en los datos que pueden darse en el análisis de supervivencia:

- **Censura por la derecha**: Es la censura que tiene lugar cuando se desconoce el tiempo exacto en el que se produce el suceso pero se sabe que ocurre después de un cierto tiempo t. Es la censura más habitual y es con la que trabajaremos.

- **Censura por la izquierda**: Es la censura que tiene lugar cuando se desconoce el tiempo exacto en el que se produce el suceso pero se sabe que ocurre antes de un cierto tiempo t.
- **Censura por intervalo**: Es la censura que tiene lugar cuando se sabe que el ha ocurrido entre los tiempos a y b.

Volviendo a la situación descrita al inicio, supongamos que la información sobre los seis individuos en los doce meses de estudio se recoge en el esquema siguiente:



Supongamos que el círculo rojo representa un dato censurado y el cuadrado gris un dato no censurado, esto es, un tiempo de fallo. Bajo dicha suposición, la interpretación del esquema sería la siguiente:

Los individuos A,B y C entran en el estudio (se les interviene quirúrgicamente) al inicio de éste mientras que los individuos D, E y F entran en el estudio después del primer, segundo y sexto mes de estudio, respectivamente. Los individuos A, C, E y F proporcionan datos censurados con censura a derecha. Sus tiempos de supervivencia son 3+, 12+, 7+ y 12+ respectivamente, donde $t+$ denota un tiempo de censura. Los individuos A y E son individuos a los que se les ha perdido su seguimiento en el estudio y los individuos C y F siguen libres de enfermedad al final del estudio. Por otro lado, los individuos B y D presentan un tiempo de supervivencia exacto de 3 y 9 meses respectivamente; esto es, permanecen, respectivamente, 3 y 9 meses libres de enfermedad.

1.2.2. Funciones de tiempos de supervivencia

El tiempo de supervivencia T es una variable aleatoria continua. La distribución de supervivencia se describe generalmente por tres funciones que serán equivalentes matemáticamente; si conocemos una de ellas, las otras dos se derivan de ésta. Así, para conocer la distribución de la variable aleatoria T, que es en lo que estamos interesados, bastará con conocer o estimar una de las funciones que se describen a continuación.

La función de supervivencia

Definición 1.1. *La función de supervivencia, denotada por $S(t)$, se define como la probabilidad de que un individuo sobreviva más de un cierto tiempo t, es decir:*

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

La gráfica de S(t) se denomina **curva de supervivencia**.

Observación 1.2. *Obsérvese, en primer lugar, que la función $S(t)$ es complementaria a la función de distribución de T pues, por definición, $F(t) = P(T \leq t) = 1 - S(t)$. Por otro lado, notemos que $S(t)$ es una función no creciente y verifica que $S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$.*

En la práctica, si no hay observaciones censuradas, la función de supervivencia se estima como la proporción de pacientes que sobreviven más de un cierto tiempo t. Esto es,

$$\widehat{S}(t) = \frac{\text{número de individuos que sobreviven más de un cierto tiempo t}}{\text{número total de individuos}}. \tag{1.1}$$

Kaplan y Meier ([8]) nos proporcionan un estimador para la función de supervivencia cuando existen datos censurados que describiremos en el capítulo siguiente.

La función de densidad

Definición 1.3. *La función de densidad, denotada por $f(t)$, se define de manera usual como el limite de la probabilidad de que un individuo falle en un corto intervalo de tiempo, de t a $t + \triangle t$, por unidad de tiempo; es decir:*

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

En la práctica, si no hay observaciones censuradas, la función de densidad $f(t)$ se estima como la proporción de pacientes en riesgo en un intervalo de tiempo por unidad de tiempo. En el caso de observaciones censuradas, a partir del estimador de Kaplan-Meier para la función de supervivencia se puede obtener una estimación para $f(t)$.

Función de riesgo

Definición 1.4. *La función de riesgo, denotada por $h(t)$, se define como la probabilidad de fallo durante un intervalo pequeño de tiempo, de t a $t+\triangle t$, asumiendo que el individuo está en riesgo en el tiempo t, por unidad de tiempo; es decir:*

$$h(t) = \lim_{\triangle t \to 0^+} \frac{P\left(t < T \leq t + \triangle t \mid T > t\right)}{\triangle t}.$$

Por otro lado, se define la ***función de riesgo acumulada*** *como:*

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

En este caso, la función *h(t)* puede decrecer, crecer ó ser constante. Ejemplos de los diversos tipos de comportamiento y su interpretación puede verse en [10].

De manera similar a la estimación de $f(t)$, la función $h(t)$ se estima, cuando no hay datos censurados, como la proporción de pacientes que fallan en un intervalo de tiempo, dado que habían sobrevivido al comienzo del intervalo, por unidad de tiempo.
Nelson y Aalen ([1] y [3]) nos proporcionan un estimador de la función de riesgo acumulada cuando hay datos censurados que describiremos en el capítulo siguiente.

Relaciones entre las funciones de supervivencia

Proposición 1.5. *La función de riesgo, de densidad y de supervivencia definidas anteriormente son matemáticamente equivalentes.*

Demostración. En efecto, en primer lugar, es obvio que

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}. \tag{1.2}$$

Por otro lado,

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}[1 - S(t)],$$

es decir,

$$f(t) = -S'(t). \tag{1.3}$$

Sustituyendo ahora (1.3) en (1.2) e integrando de 0 a t obtenemos

$$H(t) = -\log(S(t)), \tag{1.4}$$

puesto que S(0)=1, y tomando exponenciales a ambos lados de la igualdad resulta

$$S(t) = \exp\left(-H(t)\right) = \exp\left(-\int_0^t h(t)dt\right). \tag{1.5}$$

Por último, sustituyendo (1.5) en (1.2) obtenemos

$$f(t) = h(t)\exp\left(-H(t)\right) = h(t)\exp\left(-\int_0^t h(t)dt\right).$$

□

Capítulo 2

Métodos no paramétricos

Generalmente, en la práctica, no se conocen hipótesis concretas acerca de la distribución de la variable T. Por tanto, no se dispone de un modelo paramétrico para las funciones de supervivencia. En estos casos, empleamos métodos no paramétricos para su estimación.

En nuestro trabajo nos ceñimos al estudio de métodos no paramétricos de estimación y comparación de funciones de supervivencia (Capítulo 2) y de modelos semiparamétricos, donde se asume una forma paramétrica en parte del modelo (Capítulo 3). En [9] y [7] se pueden consultar algunos modelos paramétricos y su estimación.

2.1. Estimador para la función de supervivencia

En el capítulo anterior se daba una estimación de la función de supervivencia $S(t)$ cuando no existen datos censurados (1.1). Sin embargo, cuando existen observaciones censuradas el estimador (1.1) no es el adecuado. En efecto, si consideramos los tiempos de censura como tiempos de fallo en ese instante tendemos a subestimar la función de supervivencia y si ignoramos los datos censurados y trabajamos con los demás, estamos prescindiendo de parte de la información que nos proporcionan los datos. De esta manera, existe una necesidad de introducir un nuevo estimador de la función de supervivencia que incluya la información de los datos censurados y evite ese sesgo.

Existen dos procedimientos no paramétricos para estimar $S(t)$: la estimación actuarial mediante tablas de vida, en la que los datos vienen agrupados en intervalos, y el estimador de Kaplan-Meier o estimador producto-límite basado en observaciones individuales. Nosotros trabajaremos con datos obtenidos de forma individual por lo que describiremos a continuación el estimador de Kaplan-Meier, que a su vez puede verse como un caso particular del anterior. Si se desea más detalle sobre el análisis de tablas de vida puede consultarse [10].

2.1.1. Estimador de Kaplan-Meier

El estimador de Kaplan-Meier es el estimador no paramétrico de la función de supervivencia $S(t)$ cuando disponemos de tiempos de supervivencia individuales, y fue introducido por

Edward L.Kaplan y Paul Meier. Incluye la información de todos los datos, censurados y no censurados y además tiene buenas propiedades como se verá en la siguiente sección (Teorema 2.2). Por otro lado, Kaplan y Meier asumen independencia entre el tiempo de entrada en el estudio y la probabilidad de fallo.

Definición

Consideremos un estudio en el que se observa a n individuos y de los que se conoce sus tiempos de fallo o el instante de censura (tiempos no censurados y censurados). Sean $t_{(1)}, \ldots, t_{(s)}$ con $s \leq n$ los tiempos de fallo distintos ordenados. Notar que es posible que en la muestra se produzcan empates en los tiempos de fallo debido a la forma en la que se observan los datos. Sean:

- n_j: número de individuos en riesgo en el instante $t_{(j)}$.
- d_j: número de individuos que fallan en el tiempo $t_{(j)}$.

El estimador de Kaplan-Meier de $S(t)$ se define como

$$\hat{S}(t) = \prod_{j, t_{(j)} \leq t} \frac{n_j - d_j}{n_j}, \quad j = 1, \ldots, s \qquad y \qquad \hat{S}(t) = 1 \;\; si \;\; t < t_{(1)}. \tag{2.1}$$

Notar que el estimador está bien definido pues $n_j \geq 1$ y $n_j \geq d_j$.

Observaciones

- La idea de empate está incluida en la definición de d_j.
- Kaplan-Meier incluye la información de los datos censurados a través de la definición de n_j. En efecto, por definición, n_j es el número de individuos en riesgo en el tiempo $t_{(j)}$, esto es, número de individuos con tiempo de supervivencia de al menos $t_{(j)}$, donde se incluyen también los individuos con instante de censura $t_{(j)}$.
- La función $\hat{S}(t)$ permanece constante entre los tiempos entre sucesos. Así, la función $\hat{S}(t)$ será una función escalonada.
- Cuando el último tiempo observado de la muestra ordenada $t_{(s)}$ es un tiempo de fallo, el estimador toma el valor cero a partir de ese instante de tiempo. Sin embargo, si el último corresponde a un dato censurado, $\hat{S}(t)$ no toma valor cero a partir de ese instante. En esta situación, es habitual considerar que $\hat{S}(t)$ no está definido para $t > t_{(s)}$.
- Denotemos por $\hat{p}_j = \dfrac{n_j - d_j}{n_j}$ a la probabilidad estimada de sobrevivir en un tiempo $t_{(j)}$. De esta forma, el estimador de Kaplan-Meier puede escribirse como $\hat{S}(t_{(j)}) = \hat{p}_1 \times \cdots \times \hat{p}_j$, y de forma recursiva $\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)})\hat{p}_j$.

Nota 2.1. *Desde un punto de vista teórico, si asumimos que el tiempo de supervivencia es continuo, no es posible que, con probabilidad positiva, se produzcan empates en tiempos de fallo. Sin embargo, en la práctica este hecho se puede dar debido a la forma en la que*

tomamos las observaciones. Esto es, tomar los tiempos de supervivencia como tiempos de observación. Una forma de romper esos empates podría consistir en considerar que realmente no han ocurrido en un tiempo t sino que han ocurrido de manera secuencial en instantes de tiempo muy próximos (infinitesimalmente) al tiempo t. Este esquema evita los empates pero es inmediato comprobar que el factor que contribuye a la estimación de la función de supervivencia en el tiempo t es el mismo que el dado por Kaplan-Meier. En consecuencia es innecesario hacer ajustes del tipo romper los empates considerándolos consecutivos en tiempos muy próximos en el estimador de Kaplan-Meier.

El estimador de Kaplan-Meier admite una expresión alternativa, menos intuitiva pero más sencilla de calcular, como es:

$$\hat{S}(t) = \prod_{r, t_{(r)} \leq t} \frac{n-r}{n-r+1},$$

donde $r \in \mathbb{N}$ es el lugar que ocupa el tiempo de fallo observado $t_{(r)}$ con $t_{(1)}, \ldots, t_{(n)}$ los n tiempos de supervivencia ordenados (censurados y no censurados). Una aplicación práctica con esta expresión puede verse en [10].

Veamos ahora un ejemplo que ilustra la idea del estimador de Kaplan Meier.

Ejemplo

Supongamos que 10 pacientes se unen a un estudio clínico al principio del año 2000. Durante el año, 6 pacientes mueren y 4 sobreviven. Al final de ese año, 20 pacientes más se unen al estudio. En el 2001, 3 pacientes de los que entraron al principio del 2000 y 15 de los que entraron al final del año mueren quedando 1 y 5 supervivientes respectivamente.
Supongamos que el estudio termina al final del 2001 y queremos estimar $S(2)$, esto es, la proporción de pacientes que sobreviven a la muerte 2 o más años.

De los 10 pacientes que comienzan en el estudio a principios del año 2000, 6 tienen tiempo de supervivencia 1 (mueren a final de año) y 4 de al menos 1 (tiempo de supervivencia censurado que se denota por 1+). De los 20 individuos que se unen al final de este año, 15 tienen tiempo de supervivencia 1 y 5 tienen tiempo de supervivencia 1+. De los 4 individuos que sobreviven el primer año, 3 mueren en el segundo año.
Así, $n_1 = 10 + 20 = 30$, $d_1 = 6 + 15 = 21$, $n_2 = 4$ y $d_2 = 3$.
Los pacientes que sobreviven dos años pueden ser considerados como los que sobreviven el primer año y de éstos los que sobreviven un año más. Esto es,

$$\begin{aligned}\widehat{S}(2) &= P(\text{sobreviven el primer año y entonces sobreviven un año más}) \\ &= P(\text{sobreviven 2 años} \mid \text{sobreviven el primer año}) \times P(\text{sobreviven el primer año}).\end{aligned}$$

Luego,

$$\begin{aligned}\widehat{S}(2) &= (\text{proporción de pacientes sobrevivendo dos años dado que sobreviven el primer año}) \times \\ &\times (\text{proporción de pacientes que sobreviven un año}) = \frac{1}{4} \times \frac{4+5}{10+20} = \frac{n_2 - d_2}{n_2} \times \frac{n_1 - d_1}{n_1},\end{aligned}$$

que es el estimador de Kaplan-Meier (2.1) para $t = 2$.

Intervalos de confianza para $S(t)$

Una vez construido el estimador de Kaplan-Meier de $S(t)$ es necesario tener una medida de su precisión. Para ello, se requiere una estimación de su varianza que nos permitirá obtener intervalos de confianza para $S(t)$.

Teorema 2.2. *Un intervalo de confianza de* $S(t)$ *para un tiempo fijo t, para muestras grandes, a un nivel del* $100(1-\alpha)\,\%$ *viene dado por*

$$\left(\hat{S}(t) - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{S}(t))}}, \hat{S}(t) + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\hat{S}(t))}}\right) \tag{2.2}$$

donde $z_{1-\frac{\alpha}{2}}$ *es el cuantil correspondiente a la distribución normal estándar y*

$$\widehat{Var(\widehat{S}(t))} \approx [\widehat{S}(t)]^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{2.3}$$

conocida como la ***fórmula de Greenwood*** *y que resulta de aplicar el Método Delta.*

Demostración. El estimador no paramétrico de Kaplan-Meier de la función de supervivencia se puede deducir también como un estimador máximo verosímil [8]. De esta manera, el estimador de Kaplan-Meier posee buenas propiedades respecto a los conceptos de consistencia, sesgo, eficiencia y suficiencia, entre otros. Además, por otro lado, las propiedades asintóticas de los estimadores de máxima verosimilitud garantizan la normalidad asintótica del estimador de Kaplan-Meier. Luego, podemos construir el intervalo de confianza para $S(t)$ como en (2.2).

Demostremos ahora la fórmula de Greenwood (2.3). El *Método Delta* [7] está basado en la aproximación de primer orden por el desarrollo en serie de Taylor. Consideremos la función de la variable aleatoria X denotada por $f(X)$. Usando los dos primeros términos de la serie de Taylor, en torno a la media de la variable, para aproximar el valor de la función tenemos

$$f(X) \approx f(\mu) + (X - \mu) \times \frac{\partial f(X)}{\partial X} \mid_{X=\mu} .$$

Así, se sigue que la varianza de la función es aproximadamente

$$Var(f(X)) \approx Var(X - \mu) \times [f'(\mu)]^2 \approx \sigma^2 \times [f'(\mu)]^2.$$

El estimador del Método Delta se obtiene cuando usamos las estimaciones de σ^2 y μ en la ecuación anterior. Esto es,

$$\widehat{Var(f(X))} \approx \widehat{\sigma^2} \times [f'(\widehat{\mu})]^2.$$

Como ya se ha visto, $\widehat{S}(t)$ puede verse como producto de proporciones, así por comodidad estimaremos primero la varianza del logaritmo del estimador Kaplan-Meier.
Considerando $f(X) = \ln(X)$ tenemos que

$$\widehat{Var(\ln(X))} \approx \widehat{\sigma^2} \times \frac{1}{\widehat{\mu}^2}.$$

Suponiendo primero que las observaciones de supervivencia entre los n_i sujetos en riesgo son independientes Bernoulli con probabilidad constante $\hat{p}_i$ queda:

$$\widehat{Var(\ln(\hat{p}_i))} \approx \frac{\hat{p}_i(1-\hat{p}_i)}{n_i} \times \frac{1}{\hat{p}_i^2} = \frac{d_i}{n_i(n_i - d_i)}.$$

La segunda suposición es que las observaciones en diferentes conjuntos de riesgo son independientes, esto es, que un individuo sobreviva 1 año es independiente a que otro individuo distinto sobreviva 2 años. Luego tomando $X = \ln(\widehat{S}(t))$,

$$\widehat{Var(\ln(\widehat{S}(t)))} = \sum_{t_i \leq t} \widehat{Var(\ln(\hat{p}_i))} \approx \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

Por último, aplicando otra vez lo mismo para $f(X) = \exp(X)$ siendo $X = \ln(\widehat{S}(t))$ obtenemos la fórmula de Greenwood (2.3). □

Nota 2.3. *Notemos que el intervalo de confianza (2.2) para valores extremos de t puede incluir valores fuera del rango (0,1). Además, debido a la hipótesis asintótica de normalidad, el intervalo de confianza no será demasiado satisfactorio para muestras pequeñas. Una manera de corregir estos problemas consiste en considerar una transformación biyectiva, g, que evite las restricciones en el rango y que mejore la aproximación normal en muestras pequeñas. Se calcularía el intervalo de confianza para $g(S(t))$ (calculando la estimación de la varianza de $g(S(t))$ por el Método Delta) y, tras aplicar la transformación inversa, g^{-1}, a los extremos del intervalo de confianza obtenido, obtendríamos el intervalo de confianza para $S(t)$. Un ejemplo de estas tranformaciones es $g(S(t)) = \log(-\log(S(t)))$ (página 43 de [7]).*

2.2. Estimador para la función de riesgo acumulada

A partir de la estimación de Kaplan-Meier pueden derivarse estimaciones para otras funciones de interés, por ejemplo, para la función de riesgo acumulado. Recuérdese que (1.4) relaciona la función de riesgo acumulado $H(t)$ y la función de supervivencia $S(t)$; en consecuencia, un estimador de la función de riesgo acumulado es

$$\hat{H}(t) = -\log \hat{S}(t), \tag{2.4}$$

siendo $\hat{S}(t)$ el estimador de Kaplan-Meier dado en (2.1).

2.2.1. Estimador de Nelson Aalen

Nelson y Aalen ([1] y [3]) propusieron otro estimador cuya expresión es

$$\tilde{H}(t) = \sum_{j;t_{(j)} \leq t} \frac{d_j}{n_j}, \tag{2.5}$$

con $t_{(j)}, d_j, n_j$ definidos como en (2.1). Es un estimador no paramétrico de $H(t)$ y se puede presentar una derivación formal de este estimador en términos de la teoría de procesos de conteo y Martigalas [1], pero queda fuera del alcance de este trabajo.

Desde el punto de vista teórico, no hay argumentos para preferir un estimador al otro si bien el estimador de Nelson-Aalen tiene la ventaja de la sencillez de cálculo.

A partir del estimador de Kaplan-Meier hemos obtenido un estimador para la función de supervivencia. Sin embargo, se puede invertir el esquema, esto es, derivar un estimador para la función de supervivencia a partir del estimador de Nelson-Aalen. De esta manera, Nelson y

Aalen proponen un estimador para la función de supervivencia usando la relación (1.5), esto es,

$$\tilde{S}(t) = \exp\left(-\tilde{H}(t)\right), \tag{2.6}$$

Teorema 2.4. *Sea T variable continua. Los estimadores (2.4) y (2.5) son asintóticamente equivalentes, siendo el estimador de Nelson-Aalen, $\tilde{H}(t)$, la aproximación lineal de primer orden de la función $\hat{H}(t)$. Por otro lado, los estimadores (2.1) y (2.6) son también asintóticamente equivalentes, siendo el estimador de Kaplan-Meier, $\hat{S}(t)$, la aproximación lineal de primer orden de la función $\tilde{S}(t)$.*

Demostración. Usando Taylor y por la definición del estimador de Kaplan-Meier (2.1),

$$\tilde{H}(t) = -\log \hat{S}(t) = -\sum_{j;t_{(j)}\leq t} \log\left(1 - \frac{d_j}{n_j}\right) \approx \sum_{j;t_{(j)}\leq t} \left(\frac{d_j}{n_j}\right), \quad con \quad d_j << n_j.$$

Por un procedimiento análogo se tiene la equivalencia para los estimadores de la función de supervivencia.
Salvo para valores altos de t, la diferencia entre ambos estimadores es pequeña por lo general. □

Intervalos de confianza para $H(t)$

El estimador de Nelson-Aalen puede deducirse como un estimador de máxima verosimilitud ([1]). De esta manera, podemos construir un intervalo de confianza de para la función de riesgo acumulado, en un tiempo fijo t para muestras grandes, a un nivel del $100(1-\alpha)\,\%$. Dicho intervalo viene dado por:

$$\left(\tilde{H}(t) - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\tilde{H}(t))}}, \tilde{H}(t) + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\tilde{H}(t))}}\right),$$

donde una estimación de la varianza viene dada por

$$\widehat{Var\tilde{H}}(t) = \sum_{j;t_{(j)}\leq t} \left(\frac{d_j(n_j - d_j)}{n_j^3}\right),$$

expresión que se obtiene de forma análoga al Teorema 2.2 (Método Delta).
Otra estimación de la varianza puede verse en [1]:

$$\widehat{Var\tilde{H}}(t) = \sum_{j;t_{(j)}\leq t} \frac{d_j}{n_j^2}.$$

2.3. Comparación de funciones de supervivencia

A menudo estamos interesados en comparar distribuciones de supervivencia de dos o más grupos de pacientes. Entre los distintos tests no paramétricos para comparar distribuciones de supervivencia (ver en [10]), describiremos el test de Logrank (introducido originalmente por Mantel en 1966) pues es apropiado cuando disponemos de datos censurados con censura a derecha.

2.3.1. Test de Logrank

El test de Log-rank es un método no paramétrico que compara las funciones de supervivencia de dos grupos de individuos y el marco de trabajo es el mismo que cuando calculamos el estimador Kaplan Meier. No ofrece ninguna información sobre la magnitud de las diferencias entre los grupos o un intervalo de confianza. Para conocer este tipo de información se utiliza el cociente de riesgos que se explica en el siguiente capítulo (Capítulo 3).

La idea en la que se basa este test es la misma que cuando en Estadística se intenta comparar dos distribuciones a través del test χ^2. Ahora sólo se considera dos clases: individuos que han sufrido el suceso y los que no. Así, si representamos en una tabla de contingencia esta situación tendríamos:

Población	SI	NO	Total
P_1	O_1	$n_1 - O_1$	n_1
P_2	O_2	$n_2 - O_2$	n_2

donde n_1 y n_2 representan el tamaño de cada una de las poblaciones, O_1 y O_2 el número de sucesos observados en las poblaciones P_1 y P_2 respectivamente.
En esta situación, se sabe ([11]), que el estadístico

$$Z = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i},$$

donde E_i es el número esperado de sucesos en la población P_i, sigue, bajo la hipótesis nula de igualdad de las funciones de supervivencia de cada población, esto es, $H_0 : S_1(t) = S_2(t)$, una distribución χ^2_1, asintóticamente. Así, cuando calculemos el valor del estadístico Z anterior, si el p-valor correspondiente es suficientemente pequeño asumiremos que las funciones de supervivencia son distintas.

Capítulo 3

Modelos semiparamétricos

Hasta ahora solo se ha considerado la estimación de la supervivencia (funciones de supervivencia) en función del tiempo. Ahora nos planteamos cómo pueden estar influyendo otras variables. En otras palabras, ¿cómo incluir covariables en la estimación de las funciones de supervivencia?.
Por otra parte, el proceso de envejecimiento que está presente cuando los individuos se siguen en el tiempo es lo que distingue el tiempo de supervivencia de otras variables aleatorias. De las funciones que manejamos que describen la distribución del tiempo de supervivencia, la función de riesgo es la que mejor captura la esencia de este proceso de envejecimiento.

Modelos semiparamétricos

Consideramos un modelo de regresión en el que la función de riesgo depende del tiempo y de otras covariables que describen los sujetos. Para facilitar la notación, consideraremos, en primer lugar, que solo hay una covariable que denotaremos por X y siendo β el coeficiente de regresión desconocido asociado a la covariable.
El modelo de regresión se representa como sigue:

$$h(t, X, \beta) = h_0(t) r(X, \beta), \tag{3.1}$$

La función de riesgo, tal como se expresa en la fórmula anterior, es el producto de dos funciones elegidas de forma que $h(t, X, \beta) > 0$.
La función $h_0(t)$ no toma ninguna forma paramétrica en particular, es la única parte del modelo que depende del tiempo y caracteriza pues el cambio en la función de riesgo en función del tiempo. La función $h_0(t)$ sólo depende del tiempo, es decir, toma el mismo valor en un tiempo t para todos los pacientes. Notemos que la función $h_0(t)$ es la función de riesgo cuando $r(X, \beta) = 1$. Cuando la función $r(X, \beta)$ es tal que $r(X = 0, \beta) = 1$, a $h_0(t)$ se le denomina *función de riesgo base*.
Por otra parte, la otra función, $r(X, \beta)$, caracteriza el cambio en la función de riesgo en función de las covariables. En este modelo no se hace ningún supuesto sobre la forma específica de la función $h_0(t)$, por lo que este es un modelo semiparamétrico en el sentido de que sólo se asume una forma paramétrica para el efecto de las covariables.

A menudo estamos interesados en comparar la función de riesgo entre dos grupos de

pacientes. En este contexto juega un papel importante el *cociente de riesgos* definido como:

$$HR(t, x_i, x_j) = \frac{h(t, x_i, \beta)}{h(t, x_j, \beta)},$$

para dos grupos de individuos con valores de la covariable denotados por x_i y x_j, respectivamente, en un tiempo t. Además, por (3.1),

$$HR(x_i, x_j) = \frac{r(x_i, \beta)}{r(x_j, \beta)}.$$

Así, la razón de riesgo (HR) no depende del tiempo; depende solo de la función $r(X, \beta)$.

En el caso de que se dispone de más covariables, el modelo (3.1) se representa de la misma manera sólo que ahora X representa un vector de covariables y β el correspondiente vector de parámetros asociados a ellos.

3.1. Modelo de Cox

Cox fue el primero en proponer el modelo en (3.1) tomando $r(X, \beta) = \exp(X\beta)$, modelo que se denomina *modelo de Cox* o *modelo de riesgos proporcionales* y que, si tenemos p covariables, se representa por:

$$h(t, X, \beta) = h_0(t) \exp(X\beta) = h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p), \tag{3.2}$$

Nótese que, en este modelo, $h_0(t)$ es lo que hemos denominado como función de riesgo base.

Por otro lado, podemos también dar una expresión para la función de supervivencia a través del modelo de Cox. Por (1.5) tenemos que

$$S(t, X, \beta) = \exp(-H(t, X, \beta))$$

y, por (3.2),

$$H(t, X, \beta) = \int_0^t h(u, X, \beta) du = \exp(X\beta) H_0(t),$$

donde $H_0(t) = \int_0^t h_0(u)$ se define como la *función de riesgo base acumulada*.
De esta forma, la función de supervivencia para el modelo de Cox queda representada como sigue:

$$S(t, X, \beta) = [S_0(t)]^{\exp(X\beta)},$$

donde $S_0(t)$ se define como la *función base de supervivencia*.

Cociente de riesgos

El elemento principal en la regresión de Cox es el cociente de riesgos pues relaciona las dos funciones de riesgo en función de los cambios en la variable independiente, esto es, permite comparar la función de riesgo entre dos grupos de pacientes. En el modelo de Cox el cociente de riesgos es:

$$HR(X_i, X_j) = \exp(\beta(x_i - x_j)).$$

La expresión anterior muestra que el cociente de riesgos no depende del tiempo. Esto equivale a la denominada hipótesis de riesgos proporcionales.

Asumiendo que la función de riesgo viene dada por (3.2), el primer tema que debemos abordar es la estimación de los parámetros β. Posteriormente debemos analizar si las covariables son realmente significativas para nuestro modelo y de cuáles podemos prescindir.

3.1.1. Estimación de los parámetros

Como acabamos de decir, una vez propuesto el modelo que vamos a considerar debemos estimar, a partir de los datos, los parámetros de dicho modelo, esto es, los coeficientes $\beta_1, \ldots, \beta_p$ de las covariables $X_1, \ldots, X_p$.

Para la estimación de los coeficientes $\beta_1, \ldots, \beta_p$, Cox [4] propuso una función de verosimilitud parcial que depende solo del parámetro vectorial de interés y de la que se obtienen los coeficientes estimados. Cox especuló que los estimadores de los parámetros obtenidos de la función de verosimilitud parcial tendrían las mismas propiedades que los estimadores de máxima verosimilitud. Más tarde se demostró esta conjetura [2]. Cox asumía que no había empates en los tiempos de supervivencia; sin embargo, sabemos que en la práctica los empates en los tiempos de supervivencia son comunes y la función de verosimilitud parcial de Cox fue modificada para poder manejarlos [5]. Con el fin de facilitad la exposición, en lo que sigue presentaremos la función de verosimilitud parcial para el caso de que no haya empates.

Supongamos que tenemos n individuos con $t_1, \ldots, t_n$ sus tiempos de supervivencia. Sean $t_{(1)} < \ldots < t_{(r)}$ los tiempos de fallo ($r \leq n$) y sea $R(t_{(j)})$ el conjunto de riesgo en el instante $t_{(j)}$, esto es, el conjunto de personas cuyos tiempos de supervivencia son de al menos $t_{(j)}$. La función de verosimilitud parcial del modelo de Cox viene dada por:

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' z_i(t_{(j)}))}{\sum_{l \in R(t_{(j)})} \exp(\beta' z_l)}, \tag{3.3}$$

donde $z_i(t_{(j)})$ es el vector de valores de las covariables para el individuo i que muere en el instante $t_{(j)}$, z_l es el vector de valores de las covariables para el individuo l del conjunto $R(t_{(j)})$ y $\beta' = (\beta_1, \ldots, \beta_p)$.
La expresión (3.3) está expresada sólo en función de los tiempos de fallo. La función de verosimilitud parcial expresada en función de todos los tiempos es

$$L(\beta) = \prod_{j=1}^{n} \left[\frac{\exp(\beta' z_i(t_{(j)}))}{\sum_{l \in R(t_{(j)})} \exp(\beta' z_l)} \right]^{\delta_j},$$

donde δ_j es el indicador del evento, tomando el valor 0 si el dato j-ésimo es censurado.
La estimación $\hat{\beta}$ del vector de coeficientes β cumplirá que $\hat{\beta} = \underset{\forall \beta}{\text{máx}}(l(\beta))$, siendo $l(\beta) = logL(\beta)$, el logaritmo de la función de verosimilitud. Por tanto, $\hat{\beta}$ es la solución de las siguientes ecuaciones simultáneas:

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0 \quad j = 1, 2, \ldots, p.$$

De esta forma, la solución $\hat{\beta}$ se obtiene mediante métodos numéricos como, por ejemplo, el de Newton-Raphson multivariable [6].

Intervalo de confianza para β_i

Por otra parte, es posible construir intervalos de confianza para β_i a partir de la estimación de la correspondiente matriz de covarianzas [7]. En concreto, el intervalo de confianza, para muestras grandes, de nivel $100(1-\alpha)\,\%$ para un β_i viene dado por

$$(\hat{\beta}_i - z_{1-\frac{\alpha}{2}}\sqrt{v_{ii}}, \hat{\beta}_i + z_{1-\frac{\alpha}{2}}\sqrt{v_{ii}}),$$

donde v_{ii} es el elemento (i,i) de dicha matriz (inversa de la matriz de Fisher).

Relevancia de las covariables

Una vez estimado el modelo debemos analizar si todas las variables son relevantes para el modelo, es decir, si podemos prescindir de alguna de ellas. En concreto, nos planteamos un test de la forma:

$$\begin{cases} H_0 : \beta_j = 0 \\ \\ H_1 : \beta_j \neq 0 \end{cases}$$

En caso de aceptar H_0, el test nos dice que podemos prescindir de la covariable x_j. Para ello usaremos el test de Wald [10].
Debe notarse que si prescindimos de alguna variable debemos reevaluar los coeficientes del modelo.

Interpretación del coeficiente β

En general, las variables con las que trabajamos son continuas o categóricas. Por simplificidad nos centraremos en el caso de variables dicotómicas (dos valores) y variables continuas.

- Caso de una variable dicotómica:

  Supongamos que estamos interesados en comparar la función de riesgo de dos grupos de pacientes con cáncer de próstata según la edad. Consideremos la covariable dicotómica, X, que toma el valor 0 en los pacientes con edad menor a una cierta edad dada y el valor 1 en los pacientes con una edad mayor o igual a la dada. Así, la función de riesgo para el i-ésimo paciente se representa como:

  $$h_i(t, X, \beta) = h_0(t)\exp(X_i\beta) = \begin{cases} h_0(t)\exp(\beta) & si \quad X_i = 1 \\ \\ h_0(t) & si \quad X_i = 0 \end{cases}$$

  .
  Si queremos comparar la función de riesgo entre los individuos con edad menor a una dada y los individuos con edad mayor o igual que la dada, el cociente de riesgos queda:

  $$HR = \frac{h_0(t)\exp(\beta)}{h_0(t)} = \exp(\beta).$$

Si β tomara el valor ln(2) la interpretación sería que el riesgo de fallo en los pacientes con mayor edad que una cierta dada es el doble que en los pacientes con menor edad que la dada.

De esta forma, el coeficiente β representa el aumento en el logaritmo de la función de riesgo cuando pasamos de una categoría a otra.

- Caso de una variable continua:

  Supongamos que los dos grupos a comparar toman los valores $x + a$ y x en la variable X, siendo a un valor real cualquiera. El cociente de riesgos quedaría

  $$HR(x + a, x) = \exp(a\beta).$$

  El coeficiente β representaría el incremento en el logaritmo de la función de riesgo por cada incremento a en la variable.

Cuando se trabaja con variables categóricas con más de dos clases, por ejemplo n, lo que se hace es generar $n - 1$ variables denominadas de diseño que representan las n categorías posibles. Por ejemplo, en el caso de tener tres categorías A, B y C se definen dos variables de diseño D_1 y D_2 de manera que, para los individuos de la categoría A ambas variables toman el valor 0, para los de la B la variable D_1 toma el valor 1 y la D_2 el valor 0 y para los de la categoría C, la variable D_1 toma el valor 0 y la D_2 el valor 1. La interpretación de los coeficientes es análoga a lo comentado anteriormente.

Capítulo 4

Análisis de supervivencia aplicado al cáncer de próstata

4.1. Cáncer de próstata

La próstata es la glándula sexual masculina encargada de producir el semen. Es del tamaño de una nuez y se encuentra debajo de la vejiga urinaria, rodeando a la uretra.
El cáncer de próstata es el tercer tumor más frecuente en varones españoles y supone la tercera causa de muerte por cáncer en España. La enfermedad se desarrolla más frecuentemente en individuos mayores de 50 años y la incidencia de ésta aumenta con la edad. A diferencia de otro tipo de cáncer, el cáncer de próstata se caracteriza por evolucionar de forma muy lenta y es extremadamente frecuente. De hecho, la mayoría de los hombres con cáncer de próstata mueren muchos años después de su detección por causas naturales sin que el cáncer les afecte en la calidad de vida.

Detección del cáncer

La detección del cáncer se suele llevar a cabo principalmente por la prueba en sangre del antígeno prostático específico (PSA) o por la exploración física de la glándula prostática (tacto rectal). El antígeno prostático específico es una proteína producida por la próstata y su elevación en plasma es proporcional a la masa tumoral presente, de ahí que se utilice como test para detectar el cáncer. Los valores de PSA que consideraremos normales en nuestro estudio (siguiente sección) son los $PSA < 4ng/mL$. Los pacientes con PSA mayor presentan, en principio, mayor riesgo. Sin embargo, los niveles de PSA en sangre pueden elevarse por otras razones como puede ser el agrandamiento de próstata, lo que se denomina hiperplasia prostática benigna (HPB), que es un problema común en casi todos los hombres a medida que envejecen. Por esto, podrían considerarse diferentes niveles de PSA en función de la edad del paciente.

Gradación histológica del cáncer

Si los resultados de poseer el cáncer son sospechosos se procede a la extracción de una muestra tisular de la próstata (biopsia prostática) que es examinada en microscopio. Una vez realizada la biopsia, si se encuentra el cáncer, el patólogo emplea dos sistemas de gradación del cáncer de próstata: la escala de Gleason y el estadio clínico. El procedimiento de la escala de Gleason consiste en seleccionar dos zonas de la muestra y, basándose en la observación al microscopio

de las características que presentan las células, asignar a cada una de ellas un número del 1 al 5 en relación con el grado de diferenciación de las células y, por tanto, con la agresividad del cáncer. Posteriormente se suman los dos valores obtenidos de las dos zonas de la muestra obteniéndose un número comprendido entre el 2 y el 10. Este valor es el valor conocido como la escala de Gleason. Un valor menor que 6 corresponde a un cáncer con escasa agresividad, un valor 7 con un cáncer de agresividad intermedia y un valor mayor que 8 con un cáncer de alta agresividad.
Por otro lado, el estadio clínico basado en el resultado de pruebas como, por ejemplo, imágenes o biopsias describe la extensión local del tumor de la próstata clasificándolo en las siguientes categorías y subcategorías:

- T1: Tumor clínicamente indetectable, no se puede palpar ni observar por imágenes.
  - T1a: El tumor se encuentra en menos del 5 % del tejido extirpado.
  - T1b: El tumor se encuentra en más del 5 % del tejido extirpado.
  - T1c: El tumor se encuentra mediante biopsia por aguja.
- T2: Tumor clínicamente detectable, se puede palpar u observar por imágenes, pero está confinado a la próstata.
  - T2a: El tumor se encuentra en la mitad o menos de un solo lado de la próstata.
  - T2b: El tumor se encuentra en más de la mitad de un solo lado de la próstata.
  - T2c: El tumor se encuentra a ambos lados de la próstata.
- T3: Tumor extendido fuera de la próstata. Pudo haberse propagado a las vesículas seminales.
- T4: Tumor extendido a tejidos adyacentes a la próstata (además de las vesículas seminales), como por ejemplo los esfínteres externos, el recto, la vejiga,los músculos elevadores o la pared pélvica.

Tratamiento y seguimiento
Con el fin de tratar el cáncer de próstata, se le aplica un tratamiento al paciente. Si el cáncer no se ha propagado por fuera de la glándula prostática el tratamiento más común es la prostatectomía radical, que es la cirugía consistente en extirpar toda la glándula prostática y algunos tejidos alrededor de ésta. La cirugía debe eliminar las células cancerosas, sin embargo, el cáncer es posible que pueda reaparecer ya que, en la práctica, suele ser prácticamente imposible extirpar con éxito todas las células cancerosas, por lo que es frecuente hacer chequeos regulares como, por ejemplo, pruebas en sangre del PSA. El nivel de PSA debería bajar a valores muy próximos a 0 ng/mL. El punto de corte de PSA establecido para considerar recidiva bioquímica en nuestro estudio es $> 0,4\ ng/mL$.

4.2. Estudio de un grupo de pacientes con cáncer

Realizaremos el estudio del análisis de supervivencia a un grupo de 359 individuos a que se les ha detectado un cáncer de próstata, en el hospital “Miguel Servet” de Zaragoza.

Los individuos entran en el estudio en el momento que se les interviene quirúrgicamente con una prostatectomía radical hasta que se produce la recidiva bioquímica o no se produce (censura). Los pacientes en estudio son pacientes de "bajo riesgo", en el sentido de que todos pertenecen a las categorías T1 o T2 del estado clínico.

El objetivo de nuestro estudio es estudiar la influencia de las variables preoperatorias (nivel de PSA, escala de Gleason, estadio clínico), junto a la edad del paciente a la intervención quirúrgica, sobre el tiempo de supervivencia, es decir, el tiempo libre de cáncer. Como se ha indicado anteriormente, se considera que se ha producido una recidiva del cáncer cuando $PSA > 0{,}4 ng/mL$ (recidiva bioquímica).

El estudio se llevará a cabo, principalmente con el programa estadístico SSPS y con el programa R.

Las variables con las que trabajaremos son las siguientes:

- Meses_hasta_último_seguimiento: Es la variable continua "tiempo de supervivencia" siendo la unidad de tiempo los meses. El suceso de interés es la recidiva bioquímica, esto es, un valor de PSA$> 0,4\ ng/mL$.
- PSA_04_Dummy: Es una variable dicotómica donde el valor 0 indica que no se ha dado el suceso (recidiva bioquímica) en el individuo, esto es, el dato es censurado y el valor 1 que si ha dado, esto es, dato no censurado.
- Edad_a_la_prostatectomía: Es la variable continua que expresa la edad del paciente a la prostatectomía radical.

  Las variables siguientes son variables preoperatorias.
- PSA: Es la variable continua que expresa el valor de PSA del paciente en las pruebas de sangre para la detección del cáncer realizadas antes de la operación.
- Gl_Bx_Cat: Es la variable categórica que indica el grado en la escala de Gleason.
- Est_Clin_Cat_Rec: Es la variable categórica que indica el estadio clínico del paciente, el valor "1" para la categoría T1. Toma el valor "4" para las categorías T2a y T2b y el valor "5" para la categoría T2c.

4.2.1. Análisis descriptivo de las variables

Empezaremos el estudio con un análisis descriptivo de nuestras variables con la ayuda del programa estadístico SSPS. Las tablas de frecuencias de las variables categóricas o los estadísticos descriptivos de las variables continuas se muestran en el Anexo.

Los tiempos de supervivencia de los pacientes varían entre 1 y 15 años y el tiempo medio es de aproximadamente 7 años (Figura 4.7). Por otro lado, recordando que el valor 'SI' denota que se ha producido el suceso (recidiva bioquímica), se observa en la Figura 4.8 que, en más de la mitad de los casos, 65,7 %, no se ha producido la recidiva durante el estudio.

Como se había indicado en el inicio del capítulo, la enfermedad se desarrolla más frecuentemente a partir de los 50 años; en nuestro caso, la media de los pacientes en la intervención

quirúrgica es de 64 años. El más joven tiene 43 y el más mayor 74 (Figura 4.9). Los pacientes mayores que 74 años tendrán, como se ha dicho antes, un tiempo de supervivencia medio de unos 7 años por lo que se decide no incluirlos en el estudio.
En media, los pacientes presentan valores altos (superiores al valor tipificado como normal) de PSA (Figura 4.10), lo que tiene sentido pues los pacientes del estudio son pacientes que padecen cáncer de próstata. Sin embargo, el 87,7 % de los casos presentan un cáncer de escasa agresividad y solo 5 pacientes presentan un cáncer de agresividad alta (Figura 4.11). Por último, se observa que los pacientes se reparten de manera casi proporcional en las categorías T1 y T2 (Figura 4.12).

4.2.2. Curva de supervivencia

Usando el programa Rcommander hemos representado gráficamente (ver Anexo, Figura 4.13) la función de supervivencia estimada por Kaplan y Meier (2.1) y los intervalos de confianza para cada tiempo t con la fórmula de Greenwood (2.2),(2.3), a un nivel del 95 %, esto es, $\alpha = 0,05$.
Observamos que la gráfica no toma el valor cero en su último tiempo de observación, es decir, el tiempo de supervivencia máximo corresponde a un dato censurado. Por otro lado, observamos que, por ejemplo, la probabilidad estimada de estar libre de enfermedad más de 4 años es casi del 0,8.

En realidad, lo que aparenta gráficamente ser una banda de confianza no lo es. Lo que representa el programa son los intervalos de confianza en cada punto. Así, la "aparente" banda de confianza es la unión de los puntos extremos de estos intervalos de confianza individuales. Obviamente, las bandas de confianza reales serán más amplias que estas bandas "aparentes" "proporcionadas" por los intervalos de confianza individuales pues tienen que asegurar el nivel de confianza en cualquier punto.

4.2.3. Test de Logrank

Con el fin de valorar la influencia de los valores que toman las variables sobre el tiempo de supervivencia del paciente, en esta sección aplicaremos en SSPS y en R el test de Logrank (sección 2.3.1). Para ello, hemos considerado las siguientes variables dicotómicas, calculadas a partir de las variables que se nos proporcionan:

- Edad65: Toma el valor 0 si el paciente tiene una edad menor o igual a 65 y el valor 1 para el resto de pacientes.
- PSA4: Toma el valor 0 en los pacientes con un valor de PSA preoperatorio menor o igual a 4 ng/mL (valor de PSA considerado normal) y el valor 1 en el resto.
- GleasonBiopsia6: Toma el valor 0 en los pacientes con una escala de Gleason menor o igual a 6, esto es, cáncer con escasa agresividad, y el valor 1 para el resto (cáncer más agresivo).
- EstadioclinicoT1T2: Toma el valor 0 en el paciente con un estadio clínico T1 y el valor 1 en el paciente con un estadio clínico T2.

Las tablas de frecuencias de estas variables dicotómicas pueden verse en el Anexo. El valor α que tomamos es $\alpha = 0,05$.

Test de Logrank para la variable 'Edad65'

En este caso, las poblaciones son: "pacientes con una edad menor o igual a 65 años" y "pacientes con edad mayor que 65 años". El propósito del test será conocer si tener una edad mayor o menor que 65 años influye significativamente sobre el tiempo de supervivencia. Los resultados del estadístico en R y SSPS son los siguientes:

```
                N Observed Expected (O-E)^2/E
Edad65=0      199       61     69.9      1.13
Edad65=1      160       62     53.1      1.49

 Chisq= 2.6  on 1 degrees of freedom, p= 0.105
```

Comparaciones globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	2,622	1	,105

Prueba de igualdad de distribuciones de supervivencia para los distintos niveles de Edad65.

Figura 4.1: Test de Log Rank.Edad65

El p-valor es mayor que 0,05 luego no se rechaza la hipótesis nula de igualdad de las funciones de supervivencia. Esto es, tener más o menos de 65 años no influye significativamente en el tiempo de supervivencia.

Test de Logrank para la variable 'PSA4'

Los resultados del estadístico en R y SSPS se muestran en la figura siguiente:

```
                N Observed Expected (O-E)^2/E
PSA4=0         24        3     9.78     4.699
PSA4=1        335      120   113.22     0.406

 Chisq= 5.1  on 1 degrees of freedom, p= 0.0238
```

Comparaciones globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	5,111	1	,024

Prueba de igualdad de distribuciones de supervivencia para los distintos niveles de PSA4.

Figura 4.2: Test de Log Rank.PSA4

El p-valor es menor que 0,05 luego se rechaza la hipótesis nula de igualdad de las funciones de supervivencia. Es una variable significativa, es decir, estar por debajo o por encima del valor de PSA 4 ng/mL influirá en el tiempo de supervivencia. Como es lógico y, como se puede ver en el Anexo en la (Figura 4.18), un paciente con un valor menor que 4 tendrá mayor probabilidad de sobrevivir más de un cierto tiempo t que un paciente un valor mayor a 4.

Test de Logrank para la variable 'GleasonBiopsia6'

Los resultados del estadístico en R y SSPS son los siguientes:

```
                          N Observed Expected (O-E)^2/E
GleasonBiopsia6=0       315       96    112.2      2.34
GleasonBiopsia6=1        44       27     10.8     24.25

 Chisq= 26.7  on 1 degrees of freedom, p= 2.43e-07
```

Comparaciones globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	26,656	1	,000

Prueba de igualdad de distribuciones de supervivencia para los distintos niveles de GleasonBiopsia6.

Figura 4.3: Test de Log Rank.GleasonBiopsia6

El p-valor es menor que 0,05 luego se rechaza la hipótesis nula; es una variable significativa, luego la agresividad del cáncer antes de la intervención quirúrgica (pertenecer a un grupo con un gleason menor que 6 o a un grupo con un gleason mayor que 6) influirá en el tiempo de supervivencia. Como en el caso anterior, esto puede verse reflejado en sus curvas de supervivencia (Figura 4.19), en el Anexo.

Test de Logrank para la variable 'EstadioClínicoT1T2'

El propósito del test consistirá en predecir si existe diferencia significativa en los tiempos de supervivencia dependiendo de la extensión local del tumor que presentaban los pacientes

antes de la cirugía. Los resultados del estadístico en R y SSPS son los siguientes:

```
                            N Observed Expected (O-E)^2/E
EstadioClinicoT1T2=0      168       50     59.2      1.44
EstadioClinicoT1T2=1      191       73     63.8      1.34

 Chisq= 2.8  on 1 degrees of freedom, p= 0.0956
```

Comparaciones globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	2,778	1	,096

Prueba de igualdad de distribuciones de supervivencia para los distintos niveles de EstadioClinicoT1T2.

Figura 4.4: Test de Log Rank.EstadioclinicoT1T2

El p-valor, mayor que 0,05, indica que la variable no es significativa. No se rechaza la hipótesis nula de igualdad de las funciones de supervivencia de las poblaciones "pacientes con estadio clínico T1" y "pacientes con estadio clínico T2".

4.2.4. Modelo de Cox

Comenzaremos planteando un modelo de Cox con todas la covariables que hemos considerado en el estudio, es decir, 'Edad_la_prostatectomía', 'PSA', 'Gl_Bx_Cat' y 'Est_Clin_Cat_Rec' y el test de Wald valorará la influencia de cada una de ellas, rechazándola o no como variable explicativa del modelo. Los resultados son:

```
                              coef exp(coef) se(coef)     z Pr(>|z|)
edad_a_la_prostatectomã.a  0.02966   1.03010  0.01651 1.796  0.07253 .
psa                        0.03109   1.03157  0.01159 2.681  0.00733 **
est_clin_cat_rec[T.4       ] 0.19125   1.21077  0.21355 0.896  0.37046
est_clin_cat_rec[T.5       ] 0.56923   1.76691  0.23322 2.441  0.01466 *
gl_bx_cat[T.7]             0.78313   2.18832  0.24262 3.228  0.00125 **
gl_bx_cat[T.8-10]          2.81155  16.63571  0.49088 5.728 1.02e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                           exp(coef) exp(-coef) lower .95 upper .95
edad_a_la_prostatectomã.a      1.030    0.97078    0.9973     1.064
psa                            1.032    0.96939    1.0084     1.055
est_clin_cat_rec[T.4       ]   1.211    0.82592    0.7967     1.840
est_clin_cat_rec[T.5       ]   1.767    0.56596    1.1187     2.791
gl_bx_cat[T.7]                 2.188    0.45697    1.3602     3.521
gl_bx_cat[T.8-10]             16.636    0.06011    6.3563    43.539
```

Figura 4.5: Modelo de Cox

A la vista de los p-valores, es evidente que las variables 'Est_Clin_Cat' y 'Edad_la_prostatectomía' pueden ser eliminadas del modelo. En consecuencia, ajustaremos un nuevo modelo sin esas variables.
Podríamos haber deducido también esta afirmación fijándonos en los intervalos de confianza. En efecto, en el test de Wald [10] la hipótesis nula es $\beta = 0$ (variable no significativa) luego, para el valor usual $\alpha = 0,05$ podemos fijarnos el intervalo de confianza para el coeficiente β o, equivalentemente, para $\exp\beta$ que es el que aparece directamente en la Figura 4.5. Así, si el valor $\exp(\beta) = 1$ pertenece al intervalo , no se rechaza la hipótesis nula; en caso contrario, se rechaza, es decir, la variable permanece en el modelo. Por ejemplo, para la variable 'PSA',

el intervalo de confianza para $\exp\beta$ incluye el valor 1, lo que significa que la variable entra en el modelo.
En coherencia con los comentarios que acabamos de hacer, hemos considerado dos modelos incluyendo, o bien la variable 'Edad_la_prostatectomía', o la 'Est_Clin_Cat' en el modelo. En ambos casos, hemos obtenido que ninguna de esas dos variables era estadísticamente significativa, por lo que finalmente hemos seleccionado un modelo con sólo las variables 'PSA' y 'Gl_Bx_Cat'.

Los resultados correspondientes a este modelo se muestran en la tabla siguiente, donde se observa que todos los p-valores son significativos:

```
                     coef exp(coef) se(coef)     z Pr(>|z|)
psa                0.03167   1.03218  0.01145 2.766 0.005677 **
gl_bx_cat[T.7]     0.82638   2.28502  0.23937 3.452 0.000556 ***
gl_bx_cat[T.8-10]  2.52069  12.43711  0.47489 5.308 1.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
psa                    1.032     0.9688     1.009     1.056
gl_bx_cat[T.7]         2.285     0.4376     1.429     3.653
gl_bx_cat[T.8-10]     12.437     0.0804     4.903    31.546
```

Figura 4.6: Modelo de Cox. Modelo final

4.2.5. Conclusiones de los resultados

Los pacientes en el estudio son pacientes a los que se les detectó el cáncer pero categorizados en "bajo riesgo". Esto se refleja en el análisis descriptivo de las variables 'Gl_Bx_Cat' y 'Est_Clin_Cat'. Por otro lado, no hay datos censurados por pérdidas en el estudio y los datos censurados (no recidiva) presentan el 65,7 % de los casos. Esto, junto al análisis descriptivo de la variable 'tiempo de supervivencia', nos permite decir que los pacientes, en general, tienen un buen pronóstico del cáncer.
Por otro lado, se ha demostrado que no existe influencia en ser mayor o menor de 65 años ni en pertenecer a la categoría clínica T1 o T2, sobre el tiempo hasta la recidiva bioquímica. Sí influye, en cambio, sobre el tiempo de supervivencia, tener un diagnóstico preoperatorio de escasa o mucha agresividad del cáncer (Gleason) o un PSA mayor o menor que 4 ng/mL antes de la operación.
Por último, las variables que aparecen en el modelo de Cox, esto es, que mejor explican la variable 'tiempo de supervivencia' son las variables preoperatorias 'Gl_Bx_Cat' y 'PSA'.
Bajo estos resultados, diríamos que los exámenes de PSA o Gleason son exámenes fundamentales y posiblemente decisivos para una futura recidiva del cáncer.

Bibliografía

[1] Aalen, O.O; Borgan, O and Gjessing, H.H (2008), *Survival and Event History Analysis*, Springer, New York.

[2] Andersen, P.K. and Borgan, O. and Gill, R.D. and Keiding,N. (1993), *Statistical Model Based on Counting Processes*, Springer Series in Statistics, New York.

[3] Borgan, O; Institute of Mathematics, University of Oslo (1997), *Three contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aalen-Johansen estimators.*

[4] Cox, D.R (1975), *Partial likelihood.* Biometrika,**62**, 269-276.

[5] Efron, B. (1977), *The Efficiency of Cox's Likelihood Function for Censored data*, Journal of the American Statistical Association, **72**, 557-565.

[6] Gupta, S.K. (1995), *Numerical Methods for Engineers,* New Age International(P) Ltd., Publishers, Reprint 2003, New Delhi.

[7] Hosmer, D.W. and Lemeshow,S (1999), *Applied Survival Analysis.Regression modeling of time to event data*, Wiley Series in Probability and Statistics, New York.

[8] Kaplan, E. L.; Meier, P.(1958); *Nonparametric estimation from incomplete observations.* J. Amer. Statist. Assn. 53:457–481.

[9] Le, C.T. (1997), *Applied survival analysis*, Wiley Series in Probability and Statistics, New York.

[10] Lee, E.T. and Wang, J.W. (2003), *Statistical Methods for Survival Data Analysis*, Wiley Series in Probability and Statistical, 3rd edition, New Jersey.

[11] Rohatgi, V.K (1939), *An Introduction to Probability Theory and Mathematical Statistics*, Wiley, New York.