UNIVERSIDAD DE SEVILLA - CURSO 2017/2018 - GRADO EN MATEMÁTICAS

TRABAJO FIN DE GRADO

Técnicas no paramétricas y modelos de regresión para datos de tiempo de vida

Departamento :
Estadística e Investigación Operativa

Realizado por :
Lorena Barrenechea López

Tutora :
Inmaculada Barranco Chamorro

Resumen

La característica principal en un estudio de supervivencia es que los sujetos bajo estudio se observan durante un tiempo estipulado, denominado tiempo de seguimiento, y el objetivo es conocer el tiempo en el que tiene lugar el evento o suceso de interés (tiempo de ocurrencia).

Dado que este análisis puede aplicarse a distintos campos de investigación, el evento de estudio será el propio de cada disciplina: en medicina puede ser la muerte de un paciente que sufre una patología concreta o bien un episodio relacionado con la enfermedad; en ingeniería puede referirse al fallo de una máquina o de una de sus piezas y en economía podría ser el inicio de un nuevo empleo después de un periodo de desempleo.

A lo largo de este trabajo, explicaremos algunas de las técnicas utilizadas para el estudio de supervivencia.

En el Capítulo 1, estableceremos algunas definiciones para poder entender las funciones usadas en cada método, y que desarrollaremos en los siguientes capítulos. Hablaremos sobre la censura y el truncamiento, que es una manera de poder continuar con el estudio, cuando un individuo (o una pieza) muere (o falla). Y por último, las técnicas paramétricas que se utilizan.

En el Capítulo 2, presentaremos las técnicas no paramétricas que se utilizan cuando los datos no se ajustan a ninguna distribución conocida. Estas técnicas son Kaplan-Meier, Nelson-Aalen y el test Log-Rank.

En el Capítulo 3, nos centramos en un modelo en el que además de relacionar la tasa de supervivencia con el tiempo, se añaden diferentes covariables explicativas. Es el denominado modelo de regresión de Cox. Vemos los residuos que se utilizan para comprobar que ese modelo es válido. Por último, introducimos lo que se llama "modelo de Cox estratificado" que surge cuando la hipótesis de riesgos proporcionales no se cumple.

Finalmente, en el Capítulo 4, utilizamos el software R para aplicar lo expuesto en este trabajo a un conjunto de datos reales.

Abstract

The main feature in a survival study is that, the subjects under study are observed during a fixed period of time, called follow-up time, and the aim is to determine the time in which the event of interest takes place (time of occurrence).

Since this analysis can be applied to different fields of research, we will adapt the study to each discipline: in the case of medicine, it can be the death of a patient suffering from a specific pathology, or an episode related to the disease; in engineering it can refer to the failure of a machine or of one of its parts, and in economy it could be the start of a new job after a period of unemployment.

Throughout this essay, we will explain some of the techniques used for the survival study.

In Chapter 1, we will establish some definitions to understand the functions used in each method, and we will develop them in the next chapters. We will talk about censoring and truncation, which are two methods that allow us to continue with the study. Both happen when an individual (or an item) dies or has a failure. And finally, we will focus on the parametric techniques that are used.

In Chapter 2, we will show the non-parametric techniques required when the data doesn't fit to any of the known distributions.These techniques are Kaplan-Meier, Nelson-Aalen and the Log-Rank test.

In Chapter 3, we will focus on a model which, apart from associating the survival rate and time, adds different explanatory covariates. This is called Cox regression model. This chapter also analyses the residues that are used to validate the model. Later, we will introduce what is called "stratified Cox model", which arises when the proportional risks hypothesis does not meet the conditions required.

Finally, in Chapter 4, we will use R software to apply what has been previously mentioned in this essay to a set of real data.

Índice general

Índice de figuras

Capítulo 1

Generalidades y Modelos

En este capítulo se presentan los conceptos y resultados básicos a la hora de introducir y comprender el estudio de métodos para hacer inferencia con datos de tiempos de vida censurados, así como diferentes modelos continuos de distribuciones tales como la exponencial, Weibull y Log-Normal, entre otros.

1.1. Funciones básicas

Definición 1.1.1 La ***fiabilidad*** de un dispositivo (componente o sistema), sometido a unas condiciones de trabajo concretas, es la probabilidad de que este funcione correctamente ("sobreviva" sin fallar) durante un determinado periodo de tiempo.

Por lo tanto, la fiabilidad constituye un aspecto fundamental de la calidad de todo dispositivo. Por tal motivo, resulta especialmente interesante la cuantificación de dicha fiabilidad, de forma que sea posible hacer estimaciones sobre la vida útil de un producto.

Ejemplo 1.1.1 En el caso de una avioneta monomotor, puede ser útil conocer la probabilidad de que este falle en diferentes etapas de su vida (tras 500 horas de funcionamiento, 800 horas de funcionamiento, etc.). El obtener una buena estimación de la fiabilidad del motor, posibilitará la toma de decisiones racionales acerca de cuándo conviene revisarlo o cambiarlo por otro nuevo.

Para tener las ideas más claras dividiremos los productos a analizar en dos grupos:

- **Productos no reparables:** en los que puede ocurrir solo un fallo. Ejemplos: bombillas de luz, transistores, motores a propulsión, microprocesadores, etc.

- **Productos reparables:** en los que puede ocurrir más de un fallo. En este caso es importante considerar la disponibilidad del producto reparado (que dependerá de la ocurrencia de fallos y del tiempo de mantenimiento). Ejemplos: automóviles, electrodomésticos, etc.

A lo largo del trabajo trataremos con tiempos de vida de productos no reparables, supondremos además que los datos son independientes e idénticamente distribuidos (i.i.d.).

A partir de ahora iremos definiendo nuevas funciones que nos serán útiles para trabajar con datos censurados.

Notación 1.1.1

- Sea T una variable aleatoria (v.a.) continua y no negativa, que describe los tiempos de fallo de un determinado dispositivo.
- Sea $f(t)$ la función de densidad de T y $F(t)$ su función de distribución. Como ya sabemos, estas funciones están definidas por

$$F(t) = P(T \leq t) = \int_0^t f(x)\,dx$$

$$f(t) = F'(t) \tag{1.1}$$

para $t > 0$.

Propiedad 1.1.1

1. Como $f(t)$ es la función de densidad de la variable aleatoria T definida en el intervalo $(0,+\infty)$, entonces su integral es igual a 1.

$$\int_0^{+\infty} f(x)\,dx = 1 \tag{1.2}$$

2. Por ser T una variable aleatoria continua, su función de distribución $F(t)$ es una función continua.

3. $F(t)$ es monótona creciente, para todo t en el soporte de la v.a. T.

Definición 1.1.2 La ***función de fiabilidad*** $R(t)$ o ***función de supervivencia*** $S(t)$, es la complementaria de la función de distribución $F(t)$ de T, es decir, nos determina la probabilidad de que el dispositivo "sobreviva" al instante t. Esta función determina la proporción de dispositivos iniciales que seguirán funcionando correctamente en el instante t. Se define como:

$$R(t) \equiv S(t) = 1 - F(t) = 1 - P(T \leq t) = P(T > t) \tag{1.3}$$

Proposición 1.1.1 Dada T una variable aleatoria continua no negativa y $S(t)$ su función de supervivencia, se tiene que:

1. $S(0) = 1$.
2. $S(+\infty) = 0$.
3. $S(t)$ es una función continua y monótona decreciente en el soporte de T.

Demostración 1.1.1

1. $S(0) = 1 - F(0) = 1 - \int_0^0 f(x)\, dx = 1 - 0 = 1$.
2. $S(+\infty) = \lim_{t \to +\infty} S(t) = \lim_{x \to +\infty} (1 - \int_0^t f(x)\, dx) = 1 - \int_0^{+\infty} f(x)\, dx \overset{(1.2)}{=}$

   $= 1 - 1 = 0$.
3. Por la expresión dada en (1.3) se tiene $S(t) = 1 - F(t)$ y sabiendo por 2 de la Propiedad 1.1.1 que $F(t)$ es continua entonces, $S(t)$ es una función continua.
   Además, por 3 de la Propiedad 1.1.1 se tiene que $F(t)$ es monótona creciente entonces $S(t)$ es monótona decreciente para todo t en el soporte de T.

□

Definición 1.1.3 Se define ***función de riesgo*** (o hazard rate), $h(t)$, como aquella que nos da la probabilidad instantánea de muerte o fallo en el instante t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1.4}$$

De (1.4) obtenemos trivialmente que podemos aproximar

$$h(t)\Delta t \simeq P(t \leq T < t + \Delta t | T \geq t). \tag{1.5}$$

Este producto nos va a indicar la probabilidad aproximada de muerte en el intervalo $[t, t + \Delta t)$, dado que el individuo sobrevive al instante t.

Proposición 1.1.2 Se tiene que

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} \quad \text{con} \quad F(t) < 1. \tag{1.6}$$

Demostración 1.1.2
Teniendo en cuenta que

$$\lim_{\Delta t \to\ 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \lim_{\Delta t \to\ 0^+} \frac{F(T \leq t + \Delta t) - F(T \leq t)}{\Delta t} = F'(t) = f(t)$$

Llegamos al resultado:

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

$$= \frac{1}{P(T \geq t)} \lim_{\Delta t \to\ 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \frac{f(t)}{S(t)}$$

□

A partir de la representación de la función de riesgo dada en (1.5) se puede aprender acerca de las causas del fallo en un ítem, y acerca de la fiabilidad de un producto.

Su comportamiento se nos presentará en tres formas básicas: creciente, decreciente o constante.

Ejemplo 1.1.1

- Función de riesgo creciente: Indica que es más probable que los elementos fallen con el tiempo.

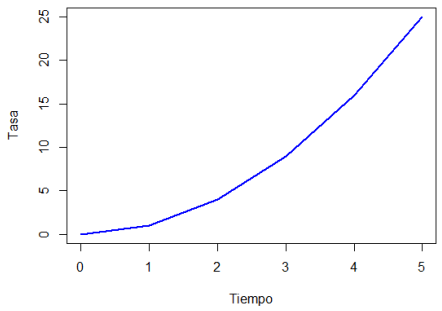

Figura 1.1: Función de riesgo creciente.

- Función de riesgo decreciente: Indica que los fallos son más probables que se produzcan temprano en la vida útil de un producto.

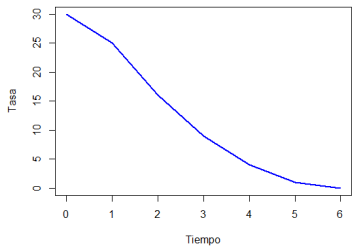

Figura 1.2: Función de riesgo decreciente.

Un ejemplo que represente esta gráfica es la probabilidad de supervivencia cuando se sale de una operación de riesgo. Tras acabar la operación de riesgo hay mucha probabilidad de fallo, pero al ir pasando el tiempo, el individuo va estabilizándose.

- Función de riesgo constante: Indica que los fallos son igual de probables que se produzcan en cualquier momento durante la vida útil del producto.

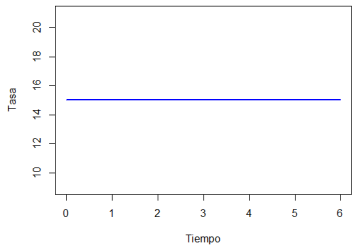

Figura 1.3: Función de riesgo constante.

- Las tres formas de fallo básicas se combinan para generar la denominada ***curva de bañera*** (bathtub curve), curva típica en fiabilidad.
  La tasa de riesgo suele ser alta al principio, baja en el medio y nuevamente alta al final de la vida útil.

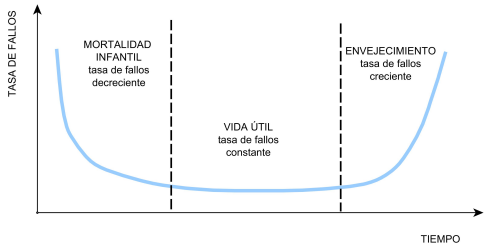

Figura 1.4: Función de riesgo con forma de bañera.

Definición 1.1.4 Se define la ***función de riesgo acumulada*** (cumulative hazard rate), como

$$H(t) = \int_0^t h(x)\,dx, \qquad t > 0$$

Nota 1.1.1
Al representar $H(t)$, nos aparecerá una línea recta si $h(t)$ es constante, una función que crece más rápido que una recta si $h(t)$ es creciente, y una que crece más despacio si $h(t)$ es decreciente.

Proposición 1.1.3 Se tienen las siguientes relaciones con respecto a la función de riesgo y la de supervivencia:

1. $S'(t) = -f(t)$
2. $h(t) = -\dfrac{d}{dt} logS(t)$
3. $S(t) = e^{-H(t)}$

Demostración 1.1.3

1. $S(t) = 1 - F(t) \Longrightarrow S'(t) = -F'(t) \overset{(1.1)}{=} -f(t).$
2. $\dfrac{d}{dt}\{logS(t)\} = \dfrac{S'(t)}{S(t)} = -\dfrac{f(t)}{S(t)} \overset{(1.6)}{=} -h(t) \Longrightarrow h(t) = -\dfrac{d}{dt} logS(t).$
3. Del anterior punto, si integramos de 0 a x respecto de t , tendremos

$$\int_0^x \frac{d}{dt} logS(t)dt = logS(t)\Big]_{t=0}^{t=x} 0 = logS(x) - logS(0) = logS(x) - log1 =$$

$$= logS(x) = -\int_0^x h(t)dt \Longrightarrow logS(x) = -\int_0^x h(t)dt = -H(x) \Longrightarrow$$

$$\Longrightarrow S(x) = e^{-H(x)}.$$

□

Definición 1.1.5 Se denomina ***cuantil de orden*** p, t_p, al valor que cumple

$$P(T \leq t_p) = p, \qquad 0 < p < 1.$$

Es decir, $t_p = F^{-1}(p)$.

Nota 1.1.2
Si hablamos en términos de *percentiles*, t_p es el percentil $100p$.

Nota 1.1.3
El percentil 50, o cuantil de orden 0.50, $t_{0.50}$ es la mediana.

1.2. Censura y Truncamiento

1.2.1. Censura

Se dice que una observación en un estudio,(Lawless 2003, [15]), está ***censurada*** cuando el individuo no presenta el evento de interés durante el tiempo de seguimiento (duración del estudio). Además, como es lógico, el estudio debe tener un punto de inicio y final adaptado a los recursos disponibles. Según si el evento ha tenido lugar antes del punto de inicio del estudio o bien, no se ha ocasionado una vez dado por finalizado, el tiempo de censura será la fecha de inicio o final del estudio.

Dependiendo del ámbito de estudio, el diseño experimental y los resultados obtenidos para la variable, las observaciones censuradas se pueden agrupar en 3 grandes grupos: censuradas por la derecha, censuradas por la izquierda y censuradas en un intervalo. En los siguientes subapartados se explica cada uno de estos grupos de forma detallada.

1.Variables censuradas por la derecha

Una observación está *censurada por la derecha* cuando el evento de interés no tiene lugar durante el período de observación, de modo que no es posible determinar el tiempo de ocurrencia. En función de la causa que ha originado la censura por la derecha, podemos encontrar distintos tipos:

- **Censura de tipo I**

  La finalización del período de observación de los individuos tiene lugar a un tiempo predeterminado, el cual se fija durante el diseño del estudio; así el tiempo de censura es conocido y fijo. La censura de tipo I puede ser:

- FIJA: el tiempo de inicio y final del estudio (censura) es el mismo para todos los individuos.
- PROGRESIVA: los individuos se dividen en grupos y cada uno de los grupos tiene un tiempo de censura concreto.
- GENERALIZADA: cada uno de los individuos tiene un tiempo de entrada al estudio y un tiempo de censura específico.

- **Censura de tipo II**

  La finalización del período de observación de los sujetos no ocurre en un tiempo prefijado, sino que éste continúa hasta que ocurre el suceso de estudio en una proporción establecida de individuos respecto al total. Por tanto, el tiempo de censura no se conoce a priori, sino que se trata de una variable aleatoria, dado que la proporción que se estipula durante el diseño del estudio es la proporción de fallos.

- **Censura de tipo III**

  Este grupo está formado por la *censura aleatoria*, también llamada no informativa. Se denomina así, porque el tiempo de censura lo determina un fenómeno aleatorio no esperado, que tiene lugar durante la consecución del estudio, e impide seguir con la observación del individuo hasta el tiempo final.
  Estos sucesos no esperados pueden ser: la pérdida del sujeto sin más información, el abandono voluntario del estudio o la experimentación de un evento de competencia con el evento de interés, que obliga a eliminar al individuo del estudio.
  Por tanto, en todos estos casos, el tiempo de censura es aleatorio y tiene lugar antes del tiempo de finalización del estudio que se ha estipulado.

2.Variables censuradas por la izquierda

Las observaciones *censuradas por la izquierda* son aquellas en las que el evento de interés ha tenido lugar antes del punto de inicio del estudio. Así, el tiempo de censura en este caso, será el tiempo de inicio del período de seguimiento ya que se conoce que el suceso ha ocurrido previamente, pero no puede saberse con exactitud cuándo (no es cuantificable).
Tal y como se ha indicado en la censura por la derecha, la censura por la izquierda también la encontramos en mediciones analíticas de la variable. En este caso, el valor obtenido en la medición es inferior a un umbral determinado y, por este motivo no es cuantificable.
Un ámbito de estudio donde este tipo de censura es recurrente, es el orientado a la investigación medioambiental, dado que los instrumentos de medida tienen un límite de detección específico y a menudo se detectan observaciones que no lo alcanzan.

3.Variables censuradas en un intervalo

Las observaciones que presentan *censura en un intervalo*, suelen ocasionarse en aquellos estudios donde las mediciones de la variable de interés se realizan de forma periódica, de modo que es posible que el suceso de estudio haya tenido lugar en un tiempo entre dos de las mediciones. En este caso, se sabe que el tiempo de ocurrencia se sitúa entre dos tiempos de censura (el máximo y el mínimo valor que conforman el intervalo), pero se desconoce el tiempo exacto y esto no permite su cuantificación.
Está documentado, que uno de los tipos de estudios donde este fenómeno es más frecuente, son los estudios de vida útil de alimentos, ya que el producto puede deteriorarse entre dos tiempos consecutivos de evaluación. También en problemas dentales, ocurren en un instante indeterminado, entre dos revisiones consecutivas.
Aunque es cierto, que en ellos también podríamos encontrar observaciones censuradas por la derecha (por ejemplo, si un individuo acepta el producto aún superando el tiempo máximo de almacenaje) o por la izquierda (por ejemplo, si se testan productos de la competencia y se desconoce la fecha de producción, pero el individuo no lo acepta desde el primer momento).

1.2.2. Truncamiento

El ***truncamiento*** tiene lugar cuando sólo aquellos sujetos que manifiestan el evento dentro de una ventana observacional se observan, del resto no se realiza ningún seguimiento y, por tanto, no se obtiene información sobre ellos.
El ejemplo más claro de truncamiento lo encontramos en el campo de la astronomía: en una parte del espacio, sólo los elementos suficientemente brillantes pueden observarse desde la Tierra; aquellos cuya intensidad lumínica es inferior a un cierto nivel, no es posible saber de su existencia.

1.3. Técnicas paramétricas

Existen numerosos modelos paramétricos,(Lawless 2003, [15]), que se usan en el análisis de tiempos de vida, y en problemas relacionados con la modelización del envejecimiento y el proceso de fallo. Entre los modelos univariantes, son unas pocas distribuciones las que toman un papel fundamental, dada su demostrada utilidad en casos prácticos. Así se tiene, entre otras: la exponencial, la Weibull y la log-normal.
El método consiste en estimar, por métodos numéricos (máxima verosimilitud o mínimos cuadrados), los parámetros característicos de la distribución, y usar su normalidad asintótica para realizar la estimación por intervalos y resolver contrastes de hipótesis.

1.3.1. Distribución Exponencial

La distribución exponencial tiene un papel fundamental en el Análisis de Fiabilidad, ya que se trata de la distribución más básica en el análisis de datos de tiempo de fallo. Se utiliza para modelar el tiempo transcurrido entre dos sucesos aleatorios, no muy frecuentes, cuando la tasa de ocurrencia, λ, se supone constante.

En fiabilidad, se usa para describir los tiempos de fallo de un dispositivo durante su etapa de vida útil, en la cuál la tasa de fallo es (aproximadamente) constante, es decir, $h(t) = \lambda$.
Una tasa de fallo constante, significa que, para un dispositivo que no haya fallado con anterioridad, la probabilidad de fallar en el siguiente intervalo infinitesimal es independiente de la edad del dispositivo.

La expresión de la función de densidad que sigue una distribución exponencial es:

$$f(t) = \lambda\, exp\{-\lambda t\}, \qquad 0 < t < \infty,\ \lambda > 0.$$

La función de distribución es:

$$F(t) = \int_0^t f(u)\, du = 1 - exp\{-\lambda t\}, \qquad 0 < t < \infty,\ \lambda > 0.$$

La función de supervivencia queda como:

$$S(t) = 1 - F(t) = exp\{-\lambda t\}, \qquad 0 < t < \infty,\ \lambda > 0.$$

La función de riesgo es:

$$h(t) = \frac{f(t)}{S(t)} = \lambda, \qquad 0 < t < \infty,\ \lambda > 0.$$

A continuación, en la Figura 1.5, hemos representado la función de densidad, función de distribución, supervivencia y tasa de riesgo para una distribución exponencial para varios valores de λ.

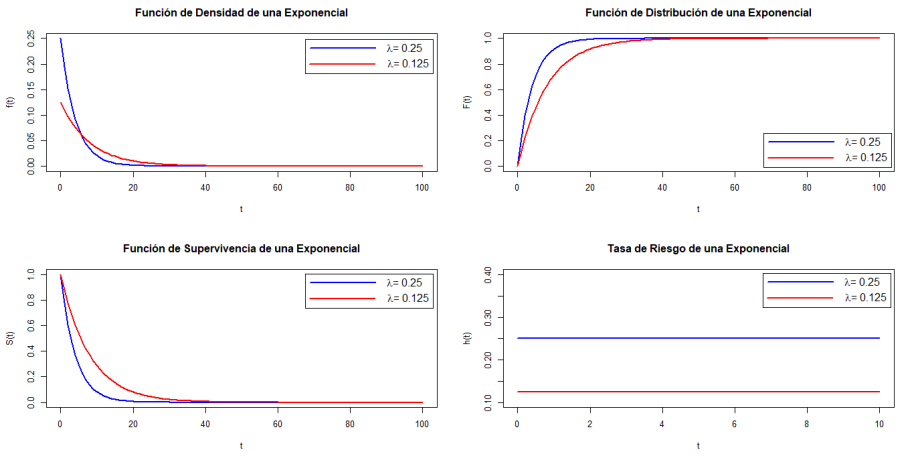

Figura 1.5: Representación de una Exponencial.

1.3.2. Distribución Weibull

Un inconveniente de la distribución exponencial es, que no sirve como modelo para tiempos de vida en los que la razón de fallo no es una función constante, sino que la probabilidad condicional de fallo instantáneo varía con el tiempo. Es decir, mientras que la distribución exponencial supone una razón de fallo constante, la familia de distribuciones Weibull incluyen razones de fallo crecientes y decrecientes. Como muchos fallos que encontramos en la práctica presentan una tendencia creciente, debido al envejecimiento o desgaste, esta distribución es útil para describir los patrones de este tipo de fallo.
La expresión de la función de densidad que sigue una distribución Weibull es:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, \qquad 0 < t < \infty,\ \alpha > 0,\ \beta > 0.$$

La función de distribución es:

$$F(t) = \int_0^t f(u)\,du = 1 - exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, \qquad 0 < t < \infty,\ \alpha > 0,\ \beta > 0.$$

La función de supervivencia queda como:

$$S(t) = 1 - F(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, \qquad 0 < t < \infty,\ \alpha > 0,\ \beta > 0.$$

La función de riesgo es:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}, \qquad 0 < t < \infty,\ \alpha > 0,\ \beta > 0.$$

Esta distribución viene caracterizada por dos parámetros: α (***escala***) y β (***forma***).

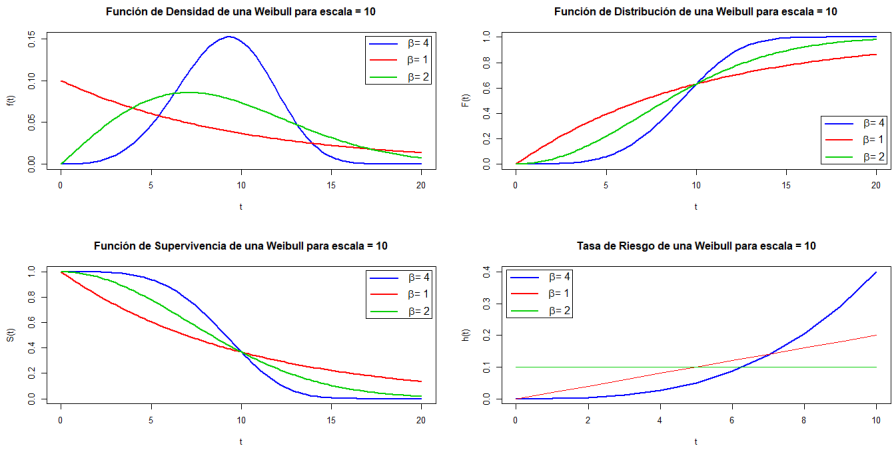

Figura 1.6: Representación de una Weibull.

En la Figura 1.6 se han representado $f(t), F(t), S(t)$ y $h(t)$ para una escala fija ($\alpha = 10$), y distintos valores del parámetro de forma.

1.3.3. Distribución Log-normal

La distribución normal, es sin duda la más importante de las distribuciones estadísticas, sin embargo no resulta de mucho interés a la hora de modelar tiempos de fallo.

Esto es debido al hecho de que la distribución normal admite valores negativos, lo cual contrasta con el hecho de que los tiempos transcurridos hasta el fallo sean siempre valores positivos. Una forma de solventar esta dificultad, es recurrir a la distribución log-normal, relacionada con la normal, y sólo considera valores positivos.

Se dice que una variable aleatoria $T > 0$ tiene un comportamiento log-normal, de parámetros μ y σ, si su logaritmo es una variable aleatoria con distribución normal, es decir, si $ln(T) \sim \mathcal{N}(\mu, \sigma)$.

Si T sigue una distribución log-normal se representa por $T \sim \mathcal{LN}(\mu, \sigma)$, donde μ y σ son los parámetros de ***localización*** y ***dispersión*** respectivamente, de la distribución de $ln(T)$.

La expresión de la función de densidad que sigue una distribución log-normal es:

$$f(t) = \frac{1}{\sigma\ t\sqrt{2\pi}}\ exp\left\{\frac{-1}{2\sigma^2}(ln(t)-\mu)^2\right\}, \qquad t>0.$$

La función de distribución es:

$$F(t) = \int_0^t f(u)\ du = \Phi\left(\frac{lnt-\mu}{\sigma}\right), \qquad t>0.$$

donde $\Phi(z)$ representa la función de distribución de una normal estándar, cuyo cálculo se obtiene con la integral

$$\Phi(z) = \int_{-\infty}^{z} \phi(u)\ du$$

donde $\phi(u)$ es la función de densidad de una Normal(0,1).

La función de supervivencia depende también de $\Phi(z)$. Se tiene que

$$S(t) = 1 - F(t) = P(T>t) = P\left(\frac{lnT-\mu}{\sigma} > \frac{lnt-\mu}{\sigma}\right) = 1-\Phi\left(\frac{lnt-\mu}{\sigma}\right),$$

para $t>0$.

La función de riesgo $h(t) = \dfrac{f(t)}{S(t)}$ tiene valor cero en $t=0$, es creciente hasta un máximo y después decrece muy lentamente sin llegar a 0.

A continuación, en la Figura 1.7 se han representado $f(t), F(t), S(t)$ y $h(t)$ para una distribución log-normal con parámetro de localización cero, $\mu=0$, y distintos valores del parámetro σ. Estos gráficos se han realizado con el paquete 'eha' de R [9].

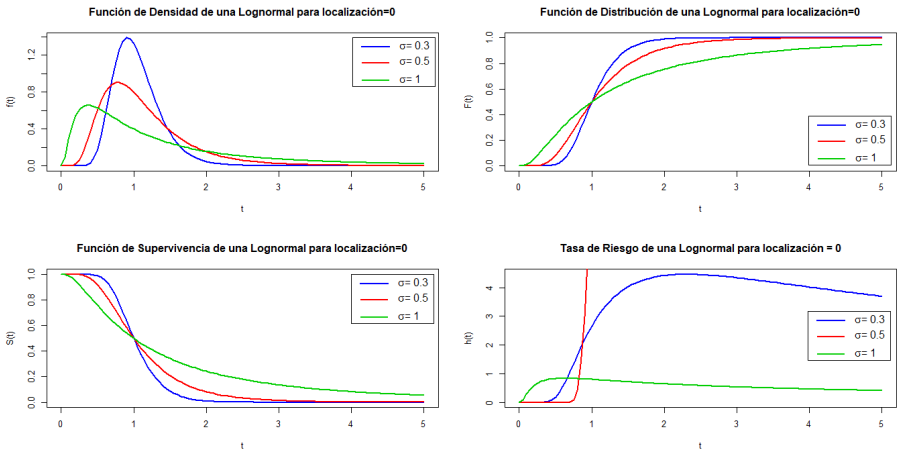

Figura 1.7: Representación de una Log-normal.

Capítulo 2

Técnicas no paramétricas

En este capítulo se introducen las técnicas no paramétricas que se utilizan para solventar el problema de que nuestros datos no se ajusten a ninguna distribución conocida (Exponencial, Weibull, Log-normal, entre otras.), y por tanto, no dependen de ningún parámetro.

Si recordamos lo visto anteriormente en asignaturas como Inferencia Estadística, se suele usar la función de densidad $f(t)$ y la función de distribución $F(t)$ para analizar y resolver un problema no paramétrico. Sin embargo, en el estudio del análisis de tiempos de vida, se usarán las siguientes funciones: $S(t)$ que es la función de supervivencia para el método Kaplan-Meier; $h(t)$ que es la función de hazard (o de riesgo) . Esta es muy irregular y difícil de entender, y por último, $H(t)$ que es la función cumulative hazard (o riesgo acumulado) para el método Nelson-Aalen.

2.1. Introducción

En general, las técnicas no paramétricas de datos de tiempo de vida son útiles para hacer un análisis preliminar, que nos puede ayudar a elegir un modelo paramétrico, haciendo uso de las características de dicho modelo.De esta manera, es mucho más fácil estudiarlo, porque conocemos a qué distribución se ajustan los datos, o bien, en el caso de no poder elegir un modelo, las técnicas no paramétricas nos proporcionan herramientas para estudiar datos de supervivencia que no se ajusten a ningún modelo.

Un ejemplo de ello, es lo que ocurre con datos relacionados con el comportamiento humano. Estos no suelen seguir las mismas pautas, y será más difícil ajustar un modelo, que en el caso de estudios relacionados con características de máquinas.

Comenzaremos estudiando la estimación no paramétrica de la función de supervivencia, $S(t)$, usando datos no censurados, para realizar inferencia en modelos de tiempo de vida.
En primer lugar, hemos de introducir alguna notación para llevarlo a cabo:

- Sean $T_1, T_2, \ldots, T_n$, las variables aleatorias que indican los tiempos de vida de n individuos.

  Se suponen independientes e idénticamente distribuidas (iid).

- Sean $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$, los tiempos (valores) observados que están ordenados de la variable T.

Consideramos la ***función de supervivencia,*** $S(t)$, que se definió como

$$S(t) = 1 - F(t) = P[T > t]$$

y como ***estimador*** de la función de supervivencia $\hat{S}(t)$

$$\hat{S}(t) = 1 - F_n^*(t)$$

donde $F_n^*(t)$ es lo que conocemos como ***función de distribución empírica***.

Definición 2.1.1 La función de distribución empírica $F_n^*(t)$ es un estimador simple de la función de distribución teórica $F(t)$ y se define como la distribución discreta que asigna la probabilidad $1/n$ a cada valor $t_{(i)}$ tal que $i = 1, 2, \ldots, n$. Su fórmula viene dada por:

$$\begin{array}{rccl} F_n^*: & \mathbb{R} & \longrightarrow & [0,1] \\ & t & \longmapsto & F_n^*(t) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \xi_i(t) \quad \text{con } \xi_i(t) = \begin{cases} 1 & \text{si} \quad T_i \leq t \\ 0 & \text{en} \quad \text{cc} \end{cases} \end{array}$$

Proposición 2.1.1 Para un t fijo se da:

1. $\xi_i(t) \sim Be(F(t))$
2. $E[F_n^*(t)] = F(t) \qquad \text{y} \qquad V[F_n^*(t)] = \dfrac{F(t)[1-F(t)]}{n}$
3. $F_n^*(t) \xrightarrow{cs} F(t) \qquad \text{y} \qquad F_n^*(t) \xrightarrow{p} F(t), \quad \text{cuando } n \to \infty.$
4. $\sqrt{n}\dfrac{F_n^*(t) - F(t)}{\sqrt{F(t)[1-F(t)]}} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1), \quad \text{cuando } n \to \infty.$

Veamos qué forma tiene esta función con una gráfica, obtenida a partir de algunos datos.

Ejemplo 2.1.1
Consideramos una muestra de 10 valores ($n = 10$) ordenados de forma creciente.

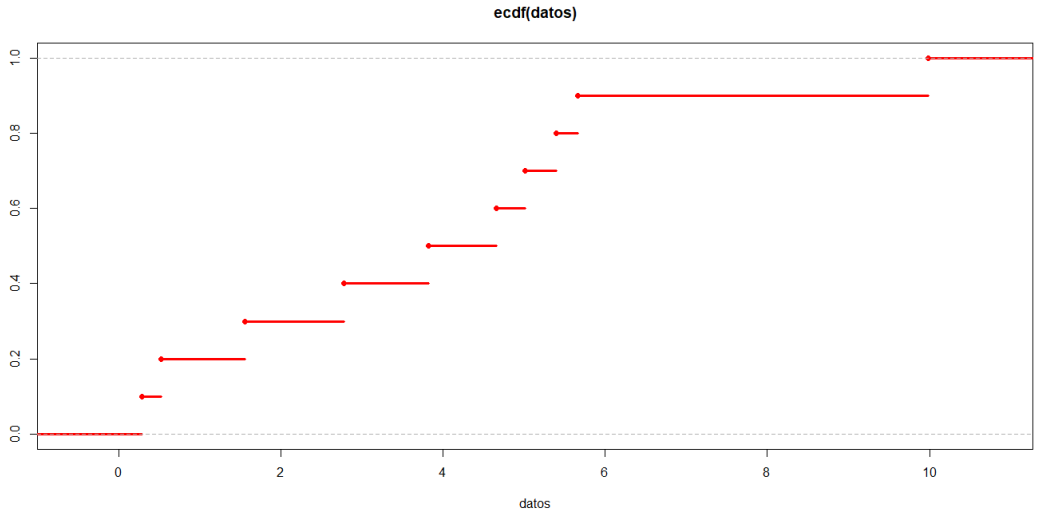

Figura 2.1: Representación de la función empírica de una muestra de valores.

Como se observa, esta es una gráfica en forma de escalera, ascendente de izquierda a derecha, con unos escalones (saltos), todos de altura $1/n$, sobre cada uno de los valores de la muestra, que sube desde 0 hasta 1 a medida que recorremos los valores de la misma. Estas son características generales de cualquier función de distribución empírica.

Volviendo a la búsqueda del estimador de la función de supervivencia, consideramos $\hat{S}(t)$ como:

$$\hat{S}(t_{(i)}) = 1 - p_i \tag{2.1}$$

con p_i= proporción de datos $\leq t_{(i)}$.

Por analogía con la función de distribución empírica se puede tomar la aproximación $p_i = i/n$.
Sin embargo, en la literatura especializada (Lawless [15], Caroni [5], entre otros.), se recomienda tomar una versión corregida de p_i:

$$p_i = \frac{i - a}{n - 2a + 1}, \quad i = 1, \ldots, n$$

con a tal que $0 < a < 1$.

Aunque la elección de a es libre, algunas de las elecciones más comunes son:

- Tomar $a = 0.5$, entonces

$$p_i = \frac{i - 0.5}{n}, \quad i = 1, \ldots, n$$

  Propuesta por Hazen.

- Tomar $a = 3/8$, entonces

$$p_i = \frac{i - 3/8}{n + 1/4}, \quad i = 1, \ldots, n$$

Propiedad 2.1.1
El estimador $\hat{S}(t)$ es una función escalonada en la que se producen las discontinuidades de salto en los tiempos de fallo observados $t_{(i)}$.

A continuación, dibujaremos la gráfica de $\hat{S}(t)$ a partir de algunos datos.
Hay que tener en cuenta que como hemos mencionado antes, la función de distribución empírica $F_n^*(t)$ es creciente de 0 a 1 y al cumplirse $\hat{S}(t) = 1 - F_n^*(t)$ se sabe de antemano que esta gráfica será decreciente de 1 a 0.

Ejemplo 2.1.2
Consideramos un conjunto de datos (obtenidos de Caroni [5]) con 238 pacientes (n=238) en el que aparece *tiempo* (que son los tiempos de fallo) y *censura* (aquellos valores que están censurados y aquellos que no). La finalidad es estudiar el tiempo que siguen un grupo de individuos adictos a la heroína con un tratamiento de metadona.

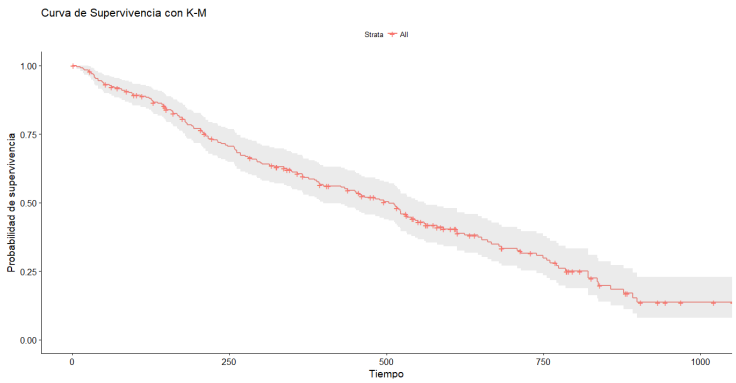

Figura 2.2: Estimación utilizando el método Kaplan-Meier con datos de metadona.

2.2. Método Kaplan-Meier

Las técnicas de estimación no paramétricas con datos censurados, se inicia con los aportes de Kaplan y Meier en el año 1958, quienes publicaron los resultados obtenidos para observaciones ***censuradas por la derecha*** y desarrollaron un estudio de las propiedades básicas de un nuevo estimador, que luego se conoció con el nombre de sus creadores. También conocido como el ***estimador límite producto***.
El estimador de Kaplan-Meier es uno de los más utilizados en los paquetes estadísticos, ya que es un estimador no paramétrico que no supone ninguna estructura para la función de distribución del tiempo de vida de las unidades bajo estudio.

La característica principal del análisis con este método, que hace que sea distintiva, es que la proporción acumulada que sobrevive se calcula para el tiempo de supervivencia individual de cada paciente, y no se agrupan los tiempos en intervalos. Por esta razón, es especialmente útil para estudios que utilizan un número pequeño de pacientes.

Por lo tanto, el método de Kaplan-Meier es útil para buscar un estimador de $S(t)$ con datos censurados por la derecha. Además, incorpora la idea del tiempo en el que ocurren los eventos.

Para llevar a cabo la búsqueda de este estimador, vamos a introducir algunas notaciones:

- Sea $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$ los tiempos (o valores) observados de manera ordenada.
- Sea n los números de objetos (o individuos) bajo estudio.
- Sea k los fallos distintos realmente observados ordenados de manera creciente $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ con $k \leq n$.
- Sea d_i el número de fallos que realmente se observa en el instante $t_{(i)}$.
- Sea n_i el número de unidades en riesgo "justo antes" de $t_{(i)}$, es decir, aquellas que aún están bajo observación en ese momento, todavía no han fallado, y tampoco han sido censuradas.

Ejemplo 2.2.1
Veamos con un gráfico cómo se usan estas notaciones, cuando estudiamos 6 unidades (n=6) y hay 4 fallos observados.
Como se puede observar, tenemos 3 fallos distintos (k=3), ya que las unidades 5 y 6 fallan al mismo tiempo, luego se cuentan como un mismo fallo, los cuales se ordenan de forma $t_{(1)} < t_{(2)} < t_{(3)}$.

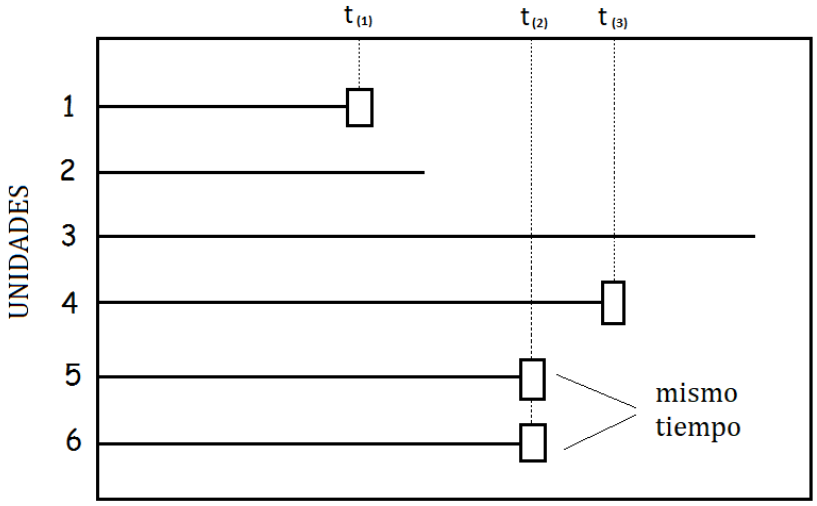

Figura 2.3: Estudio de 6 unidades.

- En $t_{(1)}$: tenemos justo antes 6 unidades que todavía no han fallado ($n_1 = 6$), y en ese instante ha fallado una unidad ($d_1 = 1$).
- Entre $t_{(1)}$ y $t_{(2)}$, la segunda unidad está censurada por la derecha, luego no se considera, y no está en el análisis.
- En $t_{(2)}$: tenemos justo antes 4 unidades ($n_2 = 4$), y en ese instante han fallado dos unidades ($d_2 = 2$).
- En $t_{(3)}$: tenemos justo antes 2 unidades ($n_3 = 2$), y en ese instante ha fallado una unidad ($d_3 = 1$).
- La tercera unidad continúa, pero no sabemos el tiempo en el que falla, luego no hay más cálculos.

Lema 2.2.1 Sea T la variable aleatoria que representa el tiempo de vida de unos ítems y $t_{(i)}$ los tiempos observados. Se tiene:

$$S(t_{(i)}) = P(T > t_{(1)})P(T > t_{(2)}/T > t_{(1)}) \dots P(T > t_{(i)}/T > t_{(i-1)}) \qquad (2.2)$$

Demostración 2.2.1
Para probar la relación dada, hacemos uso de la regla de multiplicación, la cual nos dice que dados $A_1, \dots, A_i$ sucesos se da:

$$P(A_1 \cap \dots \cap A_i) = P(A_1)P(A_2/A_1) \dots P(A_{i-1}/A_1 \cap \dots \cap A_{i-2})P(A_i/A_1 \cap \dots \cap A_{i-1})$$

En nuestro caso, el suceso A_i se define como $\{T > t_{(i)}\}$ y los conjuntos están contenidos uno dentro de otro, $\{T > t_{(i)}\} \subset \{T > t_{(i-1)}\} \subset \ \ldots \ \subset \{T > t_{(1)}\}$ se obtiene:

$$S(t_{(i)}) = P(T > t_{(i)}) \quad \text{por definición.}$$

Podemos expresar

$$P(T > t_{(i)}) = P(T > t_{(1)} \cap \ \ldots \ \cap T > t_{(i-1)} \cap T > t_{(i)})$$

Aplicando la regla de la multiplicación:

$$P(T > t_{(1)} \cap \ \ldots \ \cap T > t_{(i-1)} \cap T > t_{(i)}) =$$

$$= P(T > t_{(1)})P(T > t_{(2)}/T > t_{(1)}) \ldots P(T > t_{(i)}/T > t_{(1)} \cap \cdots \cap T > t_{(i-1)}) =$$

$$= P(T > t_{(1)})P(T > t_{(2)}/T > t_{(1)}) \ldots P(T > t_{(i)}/T > t_{(i-1)})$$

La última igualdad se tiene porque como $\{T > t_{(i-1)}\} \subset \ \ldots \ \subset \{T > t_{(1)}\}$ entonces $\{T > t_{(1)} \cap \ \ldots \ \cap T > t_{(i-1)}\} = \{T > t_{(i-1)}\}$.

Por lo tanto:

$$S(t_{(i)}) = P(T > t_{(1)})P(T > t_{(2)}/T > t_{(1)}) \ldots P(T > t_{(i)}/T > t_{(i-1)}) \qquad (2.3)$$

□

A continuación, vamos a estimar las probabilidades que intervienen en la expresión (2.2).

Por definición sabemos que $P(T > t_{(i)}) = S(t_{(i)})$, y por (2.1) que $\hat{S}(t_{(i)}) = 1 - p_i$, luego:

$$\hat{P}(T > t_{(i)}) = \hat{S}(t_{(i)}) = 1 - p_i$$

Veámos que ocurre en cada caso:

- Para $i = 1$:

  $$\hat{P}(T > t_{(1)}) = \hat{S}(t_{(1)}) = 1 - p_1$$

  donde $p_1 = \dfrac{d_1}{n_1}$ es la estimación de la proporción de fallos en $t_{(1)}$, sabiendo que n_1 unidades están funcionando y fallan d_1.

  Por tanto, $\hat{P}(T > t_{(1)}) = 1 - \dfrac{d_1}{n_1} = \dfrac{n_1 - d_1}{n_1}$.

- Para $i = 2$:

  $$\hat{P}(T > t_{(2)}/T > t_{(1)}) = \hat{S}(t_{(2)}) = 1 - p_2$$

  donde $p_2 = \dfrac{d_2}{n_2}$ es la estimación de la proporción de fallos en $t_{(2)}$, sabiendo que n_2 unidades están funcionando y fallan d_2.

  Por tanto, $\hat{P}(T > t_{(2)}/T > t_{(1)}) = 1 - \dfrac{d_2}{n_2} = \dfrac{n_2 - d_2}{n_2}$.

- En general, para i:

$$\hat{P}(T > t_{(i)}/T > t_{(i-1)}) = \hat{S}(t_{(i)}) = 1 - p_i$$

donde $p_i = \dfrac{d_i}{n_i}$ es la estimación de la proporción de fallos en $t_{(i)}$, sabiendo que n_i unidades están funcionando y fallan d_i.

Por tanto, $\hat{P}(T > t_{(i)}/T > t_{(i-1)}) = 1 - \dfrac{d_i}{n_i} = \dfrac{n_i - d_i}{n_i}$.

La relación anterior motiva la propuesta del método Kaplan-Meier para estimar $S(t)$ de manera:

- Para $t < t_{(1)} \Longrightarrow \hat{S}(t) = 1$ ya que antes de $t_{(1)}$ no ha fallado ninguno, siempre suponiendo que ningún dato antes de $t_{(1)}$ haya sido censurado.
- Para $t \geq t_{(1)} \Longrightarrow$ tomamos i tal que $t_{(i)} \leq t \leq t_{(i+1)}$ y analizamos cada caso:
  - Si $i = 1 \Longrightarrow t_{(1)} \leq t < t_{(2)}$

$$\hat{S}(t_{(1)}) = \hat{P}(T > t_{(1)}) = 1 - p_1 = \frac{n_1 - d_1}{n_1}$$

  - Si $i = 2 \Longrightarrow t_{(2)} \leq t < t_{(3)}$

$$\hat{S}(t_{(2)}) = \hat{P}(T > t_{(1)})\hat{P}(T > t_{(2)}/T > t_{(1)}) = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2}$$

  - En general, sea $i \Longrightarrow t_{(i)} \leq t$

$$\hat{S}(t_{(i)}) = P(T > t_{(1)})P(T > t_{(2)}/T > t_{(1)}) \ldots P(T > t_{(i)}/T > t_{(i-1)}) =$$

$$= \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2} \cdot \ldots \cdot \frac{n_i - d_i}{n_i}$$

En resumen:

$$\hat{S}(t) = \begin{cases} \displaystyle\prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j}, & \text{para} \quad t \geq t_{(1)}. \\ 1, & \text{para} \quad t < t_{(1)}. \end{cases}$$

Calculamos el estimador de Kaplan-Meier de la función de supervivencia, $\hat{S}(t)$, en el siguiente ejemplo.

Ejemplo 2.2.2
Consideramos los datos dados en el Ejemplo 2.2.1 donde aparecen los tiempos y fallos observados junto con las unidades que fallan.

$t_{(j)}$	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}$
$t_{(1)}$	6	1	$\frac{n_1 - d_1}{n_1} = 5/6$	5/6
(c_1)				
$t_{(2)}$	4	2	$\frac{n_2 - d_2}{n_2} = 1/2$	5/6*1/2=5/12
$t_{(3)}$	2	1	$\frac{n_3 - d_3}{n_3} = 1/2$	5/6*1/2*1/2=5/24
(c_2)				

Los datos que han sido censurados se representan por c_i. En este caso, hay un dato censurado c_1 entre $t_{(1)}$ y $t_{(2)}$ y otro dato c_2 después de $t_{(3)}$.

Veamos gráficamente el estimador de Kaplan-Meier que se ajusta a estos datos:

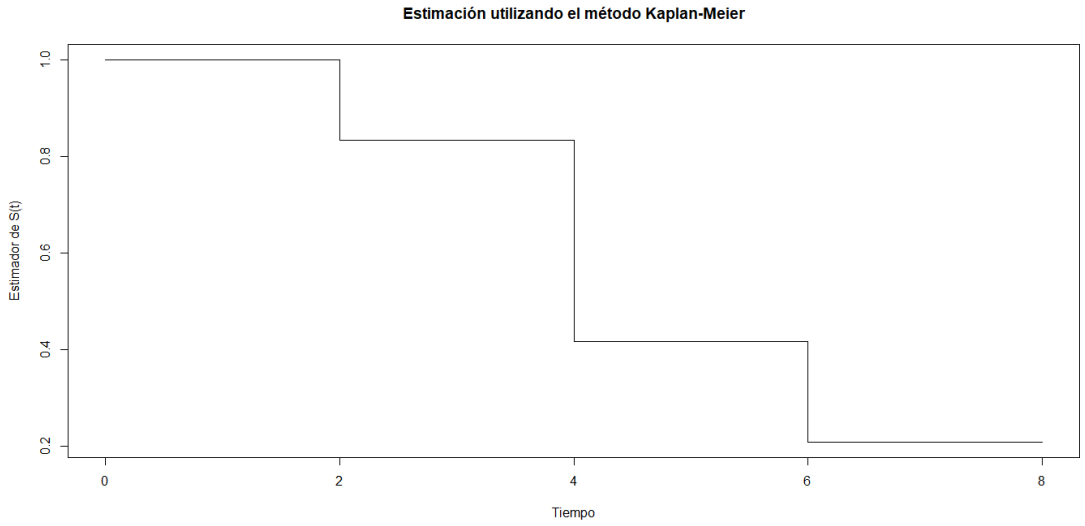

Figura 2.4: Método de Kaplan-Meier estimando S(t).

Para dibujarlo hemos supuesto que el valor de los tiempos de fallo observados son: $t_{(1)} = 2, t_{(2)} = 4$ y $t_{(3)} = 6$.

Propiedad 2.2.1
El estimador Kaplan – Meier es el estimador no paramétrico, máximo verosimil de la función de supervivencia, lo cuál hace que presente algunas de las propiedades de un buen estimador, como ser insesgado, consistente, eficiente y suficiente. Este estimador cumple con varias de estas propiedades ya que es consistente (Efron, 1967) y eficiente (Wellner, 1982). Además, es un estimador de máxima verosimilitud para datos censurados (Peterson, 1997) y es asintóticamente normal (Breslow & Crowley, 1974).
Estas propiedades facilitan el cálculo de este estimador y la utilización en problemas con datos censurados por la derecha. Cuando las estimaciones de $S(t)$ se realizan con datos no censurados, coincide con el estimador no paramétrico de la función de supervivencia. Lo anterior hace que el estimador Kaplan–Meier sea un buen estimador para muestras grandes. Para muestras muy pequeñas, hemos de tener en cuenta, que las propiedades asintóticas ya no se cumplen.

2.2.1. Varianza del estimador de Kaplan-Meier $\hat{S}(t)$

El estimador de Kaplan-Meier da una estimación puntual, o un único valor de la función de supervivencia en cualquier instante t. Por lo tanto, si se desea tener una precisión de este estimador, en diferentes instantes de tiempo, o sobre diferentes muestras, es necesario contar con un buen estimador de la varianza.

Para ello usaremos la ***fórmula de Greenwood*** que nos da una expresión aproximada para la varianza:

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \left\{ \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)} \right\} \tag{2.4}$$

Además, la desviación estándar o error estándar, se calcula como la raíz cuadrada de la varianza dada en (2.4).

$$s.e.(\hat{S}(t)) = \hat{S}(t) \left\{ \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)} \right\}^{1/2}$$

Si ahora suponemos que $\hat{S}(t)$, en un momento determinado t, sigue una distribución normal, con media μ igual al valor verdadero de $S(t)$, y varianza σ^2, entonces un intervalo de confianza aproximado, que es asintótico en el sentido de que k sea grande, sería

$$I.C\ (S(t); 1-\alpha) = \left(\hat{S}(t) - Z_{1-\alpha/2}\ s.e.(\hat{S}(t))\ ;\ \hat{S}(t) + Z_{1-\alpha/2}\ s.e.(\hat{S}(t)) \right)$$

donde $Z_{1-\alpha/2}$ representa el cuantil al nivel $1-\alpha/2$ de $\mathcal{N}(0,1)$.

Si lo queremos al 95 % :

$$1-\alpha = 0.95 \rightarrow \alpha = 0.05 \quad // \quad Z_{1-\alpha/2} = Z_{0.975} = 1.96$$

$$I.C\ (S(t); 0.95) = \left(\hat{S}(t) - 1.96\ s.e.(\hat{S}(t))\ ;\ \hat{S}(t) + 1.96\ s.e.(\hat{S}(t))\right)$$

En el siguiente ejemplo realizamos una aplicación donde se calcula $\hat{S}(t)$,$s.e.(\hat{S}(t))$ e $I.C$ al 95 %, entre otras.

Ejemplo 2.2.3
Consideramos como tiempos de fallo, (obtenidos de Caroni [6]), T= "número de millones de revoluciones". Se estudian 25 rodamientos de bolas de cerámica antes de que se produzca un fallo (n=25). En este caso, tenemos 6 observaciones censuradas por la derecha (representadas con un *):

17.88	28.92	33.00	41.52	42.12	45.60
48.48	51.84	51.96	54.12	55.56	67.80
67.80*	67.80*	68.64	68.64*	68.88*	84.12
93.12	98.64	105.12	105.84*	127.92	128.04
173.40*					

Usando el método de Kaplan-Meier, sabemos que $\hat{S}(t)$ es una función escalonada, con saltos sólo en los tiempos de fallo observados $t_{(i)}$. En este caso, hay 19 saltos (ya que 6 de las observaciones han sido censuradas, por lo tanto, no entran en el análisis), y $\hat{S}$ no cambia en los 6 tiempos en los que las observaciones han sido censuradas.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
17.9	25	1	0.9600	0.0392	0.8862	1.000
28.9	24	1	0.9200	0.0543	0.8196	1.000
33.0	23	1	0.8800	0.0650	0.7614	1.000
41.5	22	1	0.8400	0.0733	0.7079	0.997
42.1	21	1	0.8000	0.0800	0.6576	0.973
45.6	20	1	0.7600	0.0854	0.6097	0.947
48.5	19	1	0.7200	0.0898	0.5639	0.919
51.8	18	1	0.6800	0.0933	0.5197	0.890
52.0	17	1	0.6400	0.0960	0.4770	0.859
54.1	16	1	0.6000	0.0980	0.4357	0.826
55.6	15	1	0.5600	0.0993	0.3956	0.793
67.8	14	1	0.5200	0.0999	0.3568	0.758
68.6	11	1	0.4727	0.1014	0.3105	0.720
84.1	8	1	0.4136	0.1045	0.2521	0.679
93.1	7	1	0.3545	0.1050	0.1984	0.633
98.6	6	1	0.2955	0.1028	0.1494	0.584
105.1	5	1	0.2364	0.0977	0.1051	0.532
127.9	3	1	0.1576	0.0916	0.0504	0.492
128.0	2	1	0.0788	0.0721	0.0131	0.474

Figura 2.5: Información obtenida con el programa R, con los datos de número de millones de revoluciones de bolas de cerámica.

Veamos gráficamente el estimador de Kaplan-Meier de la curva de supervivencia que se ajusta a estos datos:

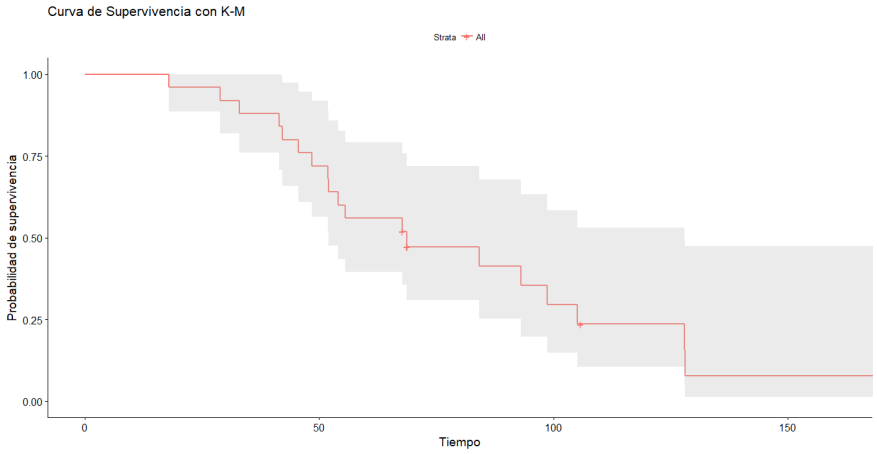

Figura 2.6: Método de Kaplan-Meier estimando S(t).

Para poder dibujarlo, además de la variable *tiempo* (dado en la tabla), hemos introducido *censura* (aquellos valores que están censurados y aquellos que no).

2.3. Método Nelson-Aalen

Este estimador fue propuesto, por primera vez por Nelson W. Aalen (1969), y luego por Altschuler (1970), quién lo descubrió utilizando técnicas de conteo con animales.

Recordemos que, por definición, la ***función de riesgo acumulada*** (o cumulative hazard rate) es de la forma:

$$H(t) = \int_0^t h(u)du \tag{2.5}$$

con $h(t)$ la ***función de riesgo*** (o fallo) en el instante t.

La expresión (2.5) también se considera como la suma de las probabilidades de fallo en el intervalo $(0, t]$ por lo que se sugiere el siguiente estimador:

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

donde $t_{(j)}$ representa los tiempos observados, d_j el número de fallos ocurridos en el instante $t_{(j)}$, y n_j el número de individuos en riesgo antes de $t_{(j)}$.

Propiedad 2.3.1
El estimador de Nelson-Aalen es otra función de escalonada, luego $\hat{H}$ puede ser útil para ayudar a entender lo que está sucediendo, y así poder elegir un modelo.

Ejemplo 2.3.1
Si $\hat{H}$ es aproximadamente lineal entonces h es aproximadamente constante, como correspondería al modelo exponencial.

2.3.1. Varianza del estimador de Nelson-Aalen H

Se propone:

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

A continuación vamos a construir el estimador de Nelson-Aalen, a partir de $\hat{S}(t)$, ya que existe un relación entre ellos.

2.3.2. Estimador de Nelson-Aalen, $\hat{H}(t)$, a partir del estimador de Kaplan-Meier, $\hat{S}(t)$

Usando la Propiedad 3 de la Proposición 1.1.3, vista en el Capítulo 1, se tiene:

$$S(t) = e^{-H(t)} \Longrightarrow H(t) = -ln\ S(t) \tag{2.6}$$

Para el caso de una variable continua, este estimador ha tenido mucha discusión por parte de Nelson (1972), Breslow y Crowley (1974), Efron (1977) y Altschuler (1979). Los cuáles llegaron a la conclusión que en este caso $\hat{H}(t)$ y $\hat{S}(t)$ son asintóticamente equivalentes, con la excepción de valores altos de t, en los que las estimaciones son menos estables.

La diferencia entre ellos será, por lo general pequeña, luego no existe ninguna razón suficiente para escoger alguna de estas.

Una de las utilidades de las estimaciones $\hat{H}(t)$ y $\hat{S}(t)$ es, en la construcción de gráficas para evaluar la selección de una determinada familia paramétrica de distribuciones.

Comenzamos con la construcción del estimador de Nelson-Aalen, paso a paso.

Por (2.6) sabemos que $H(t) = -ln\ S(t)$ luego si:

- $t < t_{(1)} \Longrightarrow \hat{H}(t) = -ln\hat{S}(t) = -ln\ 1 = 0$

- $t \geq t_{(1)} \Longrightarrow \hat{H}(t) = -ln\hat{S}(t) = -ln \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} = -\sum_{j:t_j \leq t} ln\left(\frac{n_j - d_j}{n_j}\right) =$

$$= -\sum_{j:t_j \leq t} ln\left(1 - \frac{d_j}{n_j}\right)$$

Lema 2.3.1 (Aproximación por ***Taylor***)
Dado $u_j < 1$, se tiene que:

$$-ln(1 - u_j) = \sum_{l=1}^{\infty} \frac{u_j^l}{l}.$$

Por tanto, se puede aproximar $-ln(1-u) \approx u$ para $u < 1$, y en nuestro caso se aplicará $u_j = \frac{d_j}{n_j} < 1$, porque $d_j < n_j$.
Finalmente podemos aproximar $-ln\left(1 - \frac{d_j}{n_j}\right) \approx \frac{d_j}{n_j}$.

En resumen, definimos el estimador de Nelson-Aalen como:

$$\hat{H}(t) = \begin{cases} \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j} & , \quad \text{para} \quad t \geq t_{(1)}. \\ 0, & \text{para} \quad t < t_{(1)}. \end{cases} \tag{2.7}$$

luego este estimador, $\hat{H}$, nos ofrece una alternativa al estimador $\hat{S}$.

Por otra parte, sabiendo por (2.7) que, $\hat{H}(t) = \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j}$ y por (2.6) que $S(t) = e^{-H(t)}$, se tiene:

$$\hat{S}(t) = exp\left\{-\hat{H}(t)\right\} = exp\left\{-\sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j}\right\} = \prod_{j:t_{(j)} \leq t} exp\left\{-\frac{d_j}{n_j}\right\}$$

que es el llamado ***estimador de Altshuler*** de S.

Propiedad 2.3.2
Una de las propiedades más importantes de este estimador es:

$$\text{Altshuler} \geq \text{Kaplan-Meier}, \qquad \text{para cada t.}$$

La diferencia entre los dos es muy pequeña, y es equivalente, para t pequeños.

En el siguiente ejemplo demostraremos de qué manera se utiliza el método de Nelson-Aalen, aplicado a unos datos.

Ejemplo 2.3.2
Consideramos los datos utilizados en el Ejemplo 2.2.3 teniendo en cuenta todas las variables que intervienen.

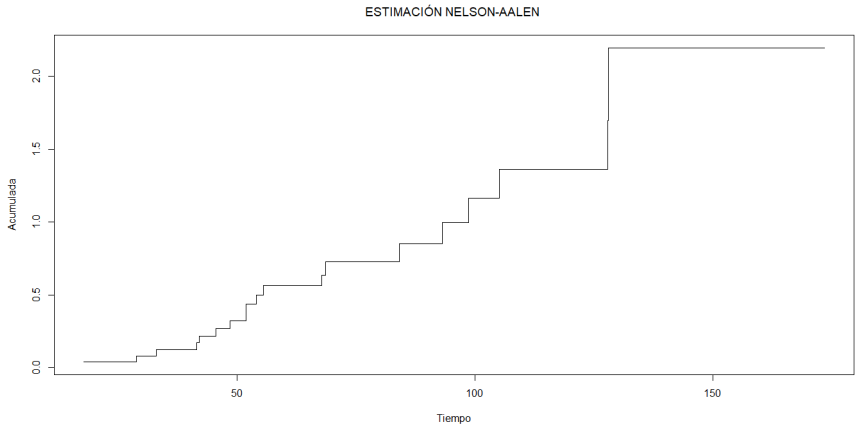

Figura 2.7: Estimación utilizando el método Nelson-Aalen.

2.4. Test Log-Rank

Para comparar la supervivencia de dos o más grupos de observaciones, necesitamos un test estadístico apropiado. El fin es determinar si todos los grupos presentan la misma supervivencia, o qué grupos son distintos.
Las representaciones gráficas de las curvas de supervivencia dan una alerta sobre posibles diferencias entre estas, pero con las pruebas estadísticas, se obtiene si existen diferencias más significativas entre las curvas, indicándonos que el factor considerado influye de forma importante en el riesgo de que falle.

Ejemplo 2.4.1
Un ejemplo de esta situación sería analizar, si los pacientes sobreviven más tiempo con el nuevo tratamiento que con el establecido o si las unidades fabricadas en A duran más que las fabricadas en B.

Para comparar la igualdad de dos o más funciones de supervivencia (fiabilidad) con datos censurados, se presentan los siguientes contrastes no paramétricos:

- Test de Log-Rank (riesgos proporcionales):
  es muy potente para calcular diferencias cuando los logaritmos de las curvas de supervivencias son proporcionales, pero tiene muchos problemas para detectar las diferencias cuando las curvas de supervivencias se cruzan.

- Test de Breslow (test de Wilcoxon generalizado):
  detecta las diferencias cuando las curvas de supervivencia se cruzan o se cortan, pero solamente al principio, por lo cual no es recomendable para un estudio a largo plazo.

- Test de Tarone-Ware :
  es un test intermedio a los otros dos.

En este caso, nos centraremos en el ***Test de Log-Rank***.
Es el test más potente cuando el cociente de las funciones de riesgo es aproximadamente constante.

Supongamos que se va a comparar la supervivencia de dos grupos A_1 y A_2 donde la función de supervivencia es, respectivamente, S_1 y S_2. Además, se tiene una muestra de cada población, con su respectivo tamaño n_1 y n_2, donde $n = n_1 + n_2$ es el número total de datos en la muestra combinada. Los tiempos de fallo se definen como $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$.

La comparación de las dos curvas de supervivencias se efectúa a través de contrastes basados en tablas de contingencia, como la siguiente:

	GRUPO		
EVENTO	A_1	A_2	TOTAL
Muerte	d_{1j}	d_{2j}	d_j
No muerte	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
EN RIESGO	n_{1j}	n_{2j}	n_j

Definimos el contraste de hipótesis de la siguiente manera:

$$\begin{cases} H_0 : S_1(t) = S_2(t) \\ H_1 : S_1(t) \neq S_2(t) \end{cases}$$

donde t es el tiempo total de observación de la muestra.

En este test de Log-Rank se compara el número de fallos observados dentro de cada grupo A_1 y A_2, además del número de fallos esperados bajo la hipótesis nula.

Cuando la hipótesis nula es cierta, es decir la función de supervivencia es igual en ambas poblaciones, la probabilidad condicional de fallo en $t_{(j)}$ es igual para los dos grupos λ_i, por lo tanto, la distribución de probabilidad de (d_{1j}, d_{2j}) está dada de la siguiente manera:

$$\prod_{i=1}^{2}\left[\left(\begin{array}{c} n_{ij} \\ d_{ij} \end{array}\right)\lambda_j^{d_{ij}}(1-\lambda_i)^{n_j-d_{ij}}\right] = \prod_{i=1}^{2}\left[\left(\begin{array}{c} n_{ij} \\ d_{ij} \end{array}\right)\right]\lambda_j^{d_{ij}}(1-\lambda_i)^{n_j-d_{ij}}$$

donde:

d_j = número total de fallos ocurridos en el tiempo $t_{(j)}$.
n_j = número total de unidades en riesgos antes de $t_{(j)}$.
d_{ij} = el número de fallos ocurridos en el tiempo $t_{(j)}$ entre los individuos del grupo $i(=1,2)$.
n_{ij} = el número de unidades en riesgo al principio de $t_{(j)}$ entre los individuos del grupo $i(=1,2)$.

Como las funciones de supervivencia coinciden (debido a la hipótesis nula), la función de riesgo es la misma en ambas poblaciones, por lo que el fallo es independiente del grupo, lo que implica que los fallos esperados en el grupo 1 viene dado por:

$$e_{i1} = \frac{n_{1j}d_j}{n_j}$$

Por tanto, se define el estadístico de Log-Rank:

$$u_i = \sum_{j=1}^{k}(d_{ij} - e_{ij})$$

es decir, los fallos observados menos los esperados.

En el caso de que el valor de k sea lo suficientemente grande se tiene por el ***teorema central del límite*** lo siguiente:

$$\frac{u}{\sqrt{v}} = \frac{\sum_{j=1}^{k}(d_{1j} - e_{1j})}{\sqrt{\sum_{j=1}^{k} v_j}} \sim \mathcal{N}(0,1)$$

donde v_j es la varianza de d_{1j} usando la distribución hipergeométrica de la forma:

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

Por lo tanto, el test estadístico Log-Rank se define:

$$\frac{u^2}{\sqrt{v}} \sim \mathcal{X}_1^2$$

En este test las diferencias observadas en todos los tiempos de fallo tienen igual importancia sin tener en cuenta el número de unidades en riesgo en cada caso. En cambio, es más útil considerar las diferencias observadas en el estadístico al principio de la observación, que las observadas al final, debido que al inicio se observan más casos.

De lo escrito anteriormente, se define una familia de test para los cuales se tiene el siguiente estadístico:

$$u = \frac{\sum_{j=1}^{k} w_j(d_{1j} - e_{1j})}{\sqrt{\sum_{j=1}^{k} w_j^2(d_{1j})}}$$

donde $w = (w_1, w_2, \ldots, w_k)$ es un vector de pesos que pondera las diferencias entre fallos observados y fallos esperados a lo largo del tiempo de observación. Bajo hipótesis nula el estadístico tiene distribución $\mathcal{N}(0, 1)$.
Si se consideran varios vectores de pesos se obtienen diferentes test, de la siguiente manera:

Si $w = 1$, se obtiene el test de Log-Rank.

Si $w_j = n_j$, se obtiene el test de Breslow o wilcoxon generalizado.

Si $w_j = \sqrt{n_j}$, se obtiene el test de Tarone-Ware.

Si $w_j = \prod_{i=1}^{j} \frac{n_i - d_i + 1}{n_i + 1}$, se obtiene el test de Prentice.

Los tres últimos test mencionados anteriormente, presentan grandes problemas porque únicamente detectan diferencias cuando se presenta alguna de las siguientes situaciones $S_1(t) < S_2(t)$ ó $S_1(t) > S_2(t)$ para todo t, de lo contrario el valor del estadístico u sería una suma de valores positivos y negativos lo cuál implica que el resultado se aproxima a cero y no es estadísticamente significativo.
Por lo tanto, estos test son de utilidad, cuando la hipótesis alternativa se corresponde con la hipótesis de riesgo acumulado.

Veámos un ejemplo del test de Log-Rank con los datos de pacientes con linfoma cuyas variables son *grupo* (al que pertenece cada paciente), *dias* (en el que se estudia a cada paciente) y *censura* (aquellos valores que están censurados y aquellos que no).

Ejemplo 2.4.2
Supongamos que queremos comparar las funciones de supervivencia de los pacientes con linfoma de Hodgkin(HL) y aquellos con linfoma no-Hodgkin(NHL). Para ello construimos un gráfico donde se observe las estimaciones de Kaplan y Meier para los pacientes con HL y con NHL.

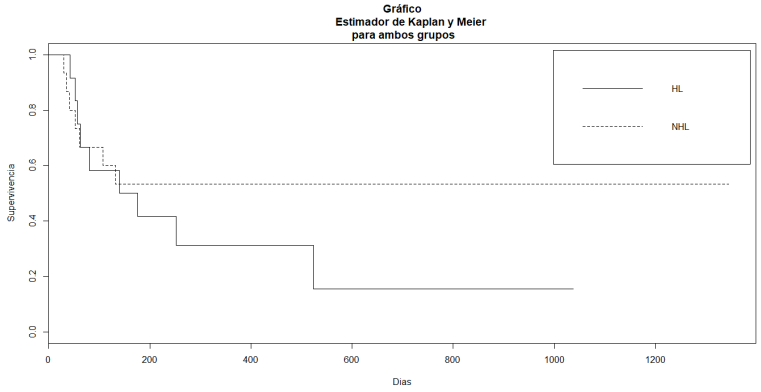

Figura 2.8: Estimación utilizando el método Kaplan-Meier para dos grupos.

De manera gráfica, aparentemente se observan diferencias entre ambas funciones de supervivencia, pero no sabemos si estas son significativas.

De manera analítica, para comparar ambas funciones aplicamos el test de Log-Rank y se obtiene que:

p-valor=0.369 $>$ 0.05 luego se acepta la hipótesis nula de igualdad de funciones de supervivencia (para un nivel de significación del 5 %).
Como información extra, $\mathcal{X}_1^2 = 0.807$.

Capítulo 3

Modelo de Regresión de Cox

Hay veces que es interesante modelizar, no solo la relación entre la tasa de supervivencia y el tiempo(como vimos en el Capítulo 2), sino también la posible relación con diferentes variables recogidas para cada sujeto. Se trata por tanto de expresar la tasa de mortalidad como una función del tiempo y de un conjunto de variables explicativas o predictoras.

Por ejemplo, si deseamos analizar el tiempo de supervivencia de unos individuos, este puede verse afectado por variables como el sexo, la edad, el tratamiento... o en otro caso, si lo que queremos es analizar el tiempo en el que un cliente permanece comprando en una empresa, es decir, se sigue manteniendo fiel a esa empresa, puede estar relacionado con variables como el tipo de negocio del cliente, el tamaño de la empresa (empleados, volumen de negocios),etc.

El modelo de regresión de Cox, propuesto por Cox en 1972, es sin duda, uno de los modelos estadísticos más usuales en el análisis de datos de supervivencia. Este modelo permite modelizar el estudio de los tiempos de vida hasta la ocurrencia de un evento de interés considerando variables de interés $\underline{X}$'s denominadas covariables. De la misma forma, la función de supervivencia en cualquier instante t queda modelizada en presencia de covariables.

A continuación, vamos a introducir alguna notación necesaria para entender el modelo.

Notación 3.0.1

- Sea $\underline{X} = (X_1, \ldots, X_p)'$ un vector aleatorio p-dimensional de covariables o de variables explicativas y $\underline{X}_i = (X_{i1}, \ldots, X_{ip})$ el vector fila que nos da el valor de las p-variables predictoras para el individuo i $(i = 1, \ldots, n)$.

La información muestral para n individuos y p variables se recogen en la siguiente matriz de tamaño (nxp):

$$\begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \\ . \\ . \\ . \\ \underline{X}_n \end{pmatrix} = \begin{pmatrix} X_{11} & \ldots\ldots & X_{1p} \\ X_{21} & \ldots\ldots & X_{2p} \\ . & \ldots\ldots & . \\ . & \ldots\ldots & . \\ . & \ldots\ldots & . \\ X_{n1} & \ldots\ldots & X_{np} \end{pmatrix}$$

- Sea $\underline{\beta} = (\beta_1, \ldots, \beta_p)'$ los coeficientes o parámetros correspondientes a cada covariable que intervienen en el modelo.
- Sea $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ los coeficientes estimados del modelo.

3.1. Formulación del modelo

Existen varias formas de introducir covariables en nuestros modelos,(Boj del Val [1],Caroni [5]). El modelo estadístico más conocido que incorpora las covariables es el ***modelo de regresión lineal estándar***, en el que el valor de una variable dependiente continua 'y', está relacionada con un conjunto de covariables X_j ($j = 1, \ldots, p$) por

$$y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

donde la constante β_0 ha sido absorbida en el vector de regresión tomando $X_1 = 1$. Los errores aleatorios correspondientes a las diferentes observaciones se suponen independientes e idénticamente distribuidos, $\mathcal{N}(0, \sigma^2)$. Suponiendo que las covariables X_j no son conocidas, tenemos que con $\underline{X} = (X_1, \ldots, X_p)'$, $y \sim \mathcal{N}(\mu, \sigma^2)$ donde

$$\mu = \mu(\underline{X}) = \underline{\beta}'\underline{X}$$

Así, las covariables entran en el modelo a través de su efecto sobre el parámetro de la distribución de 'y'.

El modelo de Cox consiste, en expresar la función de hazard (o de riesgo) $h(t)$ en función del tiempo t y de un conjunto de covariables o predictores $\underline{X}$, las cuales definen a un individuo en estudio del siguiente modo

$$h(t; \underline{X}) = h_0(t) exp(\underline{\beta}'\underline{X}) \tag{3.1}$$

equivalente a

$$h(t; X_1, \ldots, X_p) = h_0(t) exp\left(\sum_{j=1}^{p} \beta_j X_j\right)$$

donde $h_0(t)$ se denomina función de riesgo basal o "baseline hazard". Más adelante lo interpretaremos.

Propiedad 3.1.1 (Interpretación de (3.1))

- Si $exp(\underline{\beta}'\underline{X}) > 1$ entonces $h(t;\underline{X}) > h_0(t)$ es decir, la función hazard aumenta con respecto a la función de hazard baseline, lo que implica que la probabilidad de supervivencia disminuye.
- Si $exp(\underline{\beta}'\underline{X}) < 1$ entonces $h(t;\underline{X}) < h_0(t)$ es decir, la función hazard disminuye con respecto a la función de hazard baseline lo que implica que la probabilidad de supervivencia aumente.

Como hipótesis de partida podemos suponer que los tiempos de supervivencia tienen distribuciones continuas.
Para cada individuo i, con $i = 1, \ldots, n$, conocemos el tiempo que ha estado bajo estudio (muerte/fallo) t_i, su estado de fallo o censura c_i (variable codificada con 1 si el dato no está censurado y sigue dentro del análisis y con 0 si el dato ha sido censurado y ya no está en el análisis) y las covariables $\underline{X}_i's$ están fijas. Si incluimos el subíndice i para denotar a un individuo determinado, el modelo (3.1) se podría re-escribir como

$$h(t_i;\underline{X}_i) = h_0(t_i)exp(\underline{\beta}'\underline{X}_i) \tag{3.2}$$

equivalente a

$$h(t_i;(X_{i1},\ldots,X_{ip})) = h_0(t_i)exp\left(\sum_{j=1}^{p}\beta_j X_{ij}\right)$$

La función $h_0(t)$, ***función de baseline hazard***, corresponde al riesgo de un individuo que tiene como valor 0 en todos los predictores y sería el denominado ***individuo de referencia*** de cara a la interpretación posterior de los resultados. Veamoslo sustituyéndolo en el modelo (3.1)

$$h(t;X_1 = 0,\ldots,X_p = 0) = h_0(t)exp\left(\sum_{j=1}^{p}\beta_j 0\right) = h_0(t)e^0 = h_0(t).$$

También se interpreta que la función $h_0(t)$ es aquella función del modelo en el caso de no incorporar predictores.

La función de riesgo basal, $h_0(t)$, es la única parte de la expresión del modelo de Cox que depende del tiempo t. La otra parte $exp(\underline{\beta}'\underline{X})$ sólo depende del vector de covariables $\underline{X}$ de los individuos que suponemos "independientes del tiempo", es decir, sus valores no varían a lo largo del tiempo.

A continuación vamos a dar un ejemplo de lo que sería una variable independiente del tiempo, aunque a priori parezca que sí depende.

Ejemplo 3.1.1
El sexo, la raza o el grupo de tratamiento son variables fijas, sólo toman un valor, el inicial. También podríamos considerar variables como el hecho de ser o no fumador (estado de fumador) como variable independiente del tiempo, ya que aunque el estado de fumador puede variar en el tiempo, para el estudio no varía, porque se parte de un estado inicial, y se supone que no cambia hasta el final, y por lo tanto sólo toma un valor por individuo.

El modelo de Cox, definido en (3.1), se considera un modelo ***semiparamétrico***, debido a que incluye una parte paramétrica y otra parte no paramétrica:

1. La parte ***paramétrica*** se corresponde con $exp(\underline{\beta}'\underline{X})$, es decir, con la exponencial del predictor lineal $\eta = \underline{\beta}'\underline{X}$. En esta parte del modelo se estiman los parámetros o coeficientes $\underline{\beta}$ de la regresión mediante la maximización de la denominada ***función de verosimilitud parcial*** que estudiaremos con detalle más adelante.

2. La parte ***no paramétrica*** es la función de riesgo basal $h_0(t)$. Esta es una función arbitraria condicionada a la estimación de los parámetros $\underline{\beta}$ de la regresión.
   Es por esta componente no paramétrica de la fórmula que el modelo de Cox se considera ***semiparamétrico***.

Una vez estimada la parte paramétrica, $exp(\hat{\beta}'\underline{X})$, y posteriormente la no paramétrica $\hat{h}_0(t)$, tendremos la estimación del modelo semiparamétrico completo:

$$\hat{h}(t;\underline{X}) = \hat{h}_0(t)exp(\underline{\hat{\beta}}'\underline{X})$$

3.2. Hipótesis de riesgos proporcionales, PH

En el modelo de Cox se busca como primer paso, la relación entre los riesgos de muerte de dos individuos expuestos a factores de riesgo diferentes. Para ello, el modelo parte de una hipótesis fundamental, la de que los riesgos son proporcionales.

Definición 3.2.1 Se define la ***razón de riesgo***(o hazard ratio) entre dos sujetos con diferente vector de covariables $\underline{X} = (X_1, \ldots, X_p)'$ y $\underline{X}^* = (X_1^*, \ldots, X_p^*)$ como

$$\frac{h(t;\underline{X}^*)}{h(t;\underline{X})}. \tag{3.3}$$

Se evalúa en el numerador el grupo de mayor riesgo, definido por $\underline{X}^*$, y en el denominador el grupo de menor riesgo, definido por $\underline{X}$. Se espera que la razón de riesgo sea mayor que 1, ya que $h(t;\underline{X}^*) > h(t;\underline{X})$ cuantifica cuántas veces es mayor el riesgo de morir con perfil $\underline{X}^*$ que con $\underline{X}$, luego el tiempo de supervivencia disminuye.

Si sustituimos en la expresión (3.3), el modelo dado en (3.1) obtenemos

$$\frac{h(t;\underline{X}^*)}{h(t;\underline{X})} = \frac{h_0(t)exp(\underline{\beta}'\underline{X}^*)}{h_0(t)exp(\underline{\beta}'\underline{X})} = \frac{exp\left(\sum_{j=1}^{p}\beta_j X_j^*\right)}{exp\left(\sum_{j=1}^{p}\beta_j X_j\right)} =$$

$$= exp\left(\sum_{j=1}^{p}\beta_j(X_j^* - X_j)\right) = exp((\underline{X}^* - \underline{X})'\underline{\beta})$$

Por tanto

$$exp((\underline{X}^* - \underline{X})'\underline{\beta}) \quad \text{es igual al cociente (3.3)} \quad \frac{h(t,\underline{X}^*)}{h(t,\underline{X})}. \tag{3.4}$$

Observamos que el resultado de la razón de riesgo dada en (3.4) no depende de la función de riesgo basal $h_0(t)$, tan solo del valor de los predictores y de las betas estimadas, es decir, no depende del tiempo.

En resumen, en el modelo de Cox se supone la hipótesis de que los riesgos son proporcionales, ya que las covariables se suponen no dependientes del tiempo.

Veamos un ejemplo en el que no se cumple la hipótesis de riesgos proporcionales.

Ejemplo 3.2.1
Consideremos un estudio en el que los pacientes con cáncer se reparten aleatoriamente a radioterapia con o sin cirugía.
Definimos una variable que toma los valores 0 y 1 que indica si ha habido cirugía o no, respectivamente, y supongamos que ésta es la única variable predictora de interés en el modelo de Cox. La razón de riesgo se calcularía como

$$\frac{\hat{h}(t, X = 1)}{\hat{h}(t, X = 0)}$$

Cuando un paciente recibe cirugía para eliminar un tumor canceroso hay, por lo general, un alto riesgo de complicaciones por la propia cirugía, o quizás de muerte temprana, en el proceso de recuperación.

Una vez que el paciente pasa este periodo crítico, es cuando se pueden observar las ventajas de la cirugía.

Supongamos que se dibujan las curvas de riesgo para los dos grupos y que se observa que se cruzan aproximadamente en el día 3, que antes del día 3 el riesgo para el grupo de cirugía es más alto que el riesgo para el grupo sin cirugía, mientras que después de 3 días, el riesgo para el grupo de cirugía es inferior que el riesgo para el grupo sin cirugía.

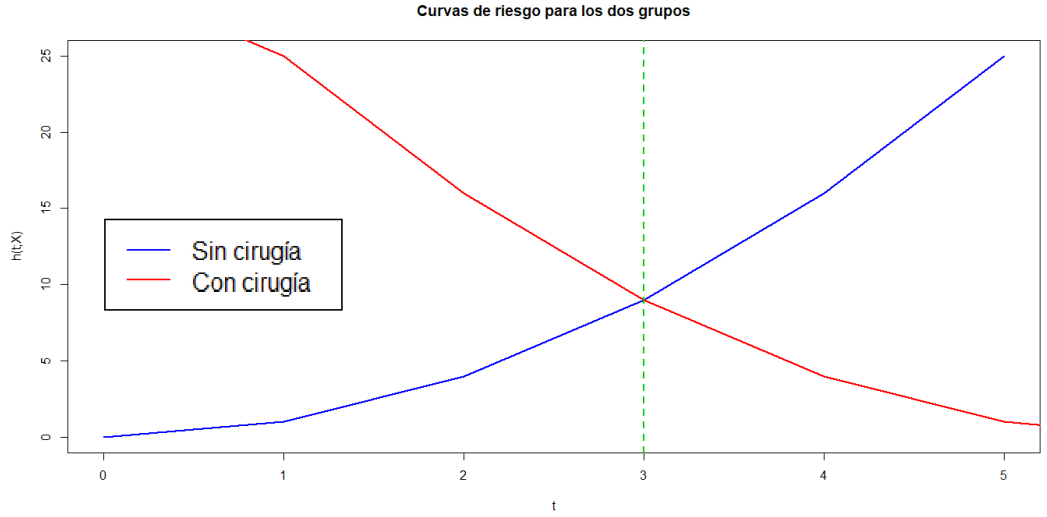

Figura 3.1: Gráfico en el que se representan las curvas de riesgo para el grupo sin cirugía y el grupo con cirugía.

Escrito en términos de la razón de riesgo se tendría lo siguiente:

$$t = 2: \quad \frac{\hat{h}(2; X = 1)}{\hat{h}(2, X = 0)} < 1$$

$$t = 3: \quad \frac{\hat{h}(3; X = 1)}{\hat{h}(3, X = 0)} = 1$$

$$t = 5: \quad \frac{\hat{h}(5; X = 1)}{\hat{h}(5, X = 0)} > 1$$

Se observa que la razón de riesgo que obtenemos no es constante a lo largo del tiempo t.

Se encontrará alguna solución a este problema cuando se estudie el modelo de Cox estratificado, el cual permite la utilización de predictores dependientes del tiempo, y que veremos más adelante.

En resumen, si los riesgos del numerador y denominador de la razón de riesgo se cruzan, la hipótesis de riesgos proporcionales no se cumple, y por lo tanto el modelo (3.1) de riesgos proporcionales no es adecuado.

3.2.1. Diagrama de diagnóstico para PH

En esta sección introducimos una forma de decidir si el modelo es adecuado o no. Para ello utilizaremos el diagrama de diagnóstico para riesgos proporcionales (PH), el cual explicamos a continuación.

Lema 3.2.1 Bajo la hipótesis (PH), la función de supervivencia para cualquier t es:

$$S(t;\underline{X}) = S_0(t)\hat{}\{exp(\underline{\beta}'\underline{X})\} \tag{3.5}$$

Demostración 3.2.1
Por la Propiedad 3 de la Proposición 1.1.3 del Capítulo 1 se tiene

$$S(t;\underline{X}) = exp\{-H(t;\underline{X})\}$$

Sustituyendo el modelo dado en (3.1) en la expresión anterior

$$S(t;\underline{X}) = exp\{-H_0(t)exp(\underline{\beta}'\underline{X})\} = \{exp(-H_0(t))\}\hat{}\{exp(\underline{\beta}'\underline{X})\}$$

Si llamamos

$$S_0(t) = exp(-H_0(t)) \tag{3.6}$$

nos queda lo siguiente

$$S(t;\underline{X}) = S_0(t)\hat{}\{exp(\underline{\beta}'\underline{X})\}$$

□

Una vez demostrada la expresión (3.5), continuamos aplicando logaritmo neperiano a ambos lados de la igualdad:

$$lnS(t,\underline{X}) = ln[S_0(t)\hat{}\{exp(\underline{\beta}'\underline{X})\}] = \{exp(\underline{\beta}'\underline{X})\}lnS_0(t)$$

Como $S(t,\underline{X}) < 1$ entonces $lnS(t,\underline{X}) < 0$. Para solucionar este problema, multiplicamos todo por (-1), para convertirlo en positivo, $(-lnS(t,\underline{X}) < 0)$, y obtenemos así:

$$-lnS(t,\underline{X}) = exp(\underline{\beta}'\underline{X})\{(-lnS_0(t)\}$$

Volvemos a aplicar logaritmo neperiano:

$$ln\{-lnS(t,\underline{X})\} = ln[exp(\underline{\beta}'\underline{X})\{(-lnS_0(t))\}] =$$
$$= ln\{exp(\underline{\beta}'\underline{X})\} + ln\{-lnS_0(t)\} = \underline{\beta}'\underline{X} + ln\{-lnS_0(t)\}$$

Por la definición dada en (2.6) sabemos que $H_0(t) = -lnS_0(t)$ luego podemos escribir:

$$ln\{-lnS(t, \underline{X})\} = \underline{\beta}'\underline{X} + ln\{H_0(t)\}$$

lo que significa que la curva $ln\{-lnS(t, \underline{X})\}$ para cada $\underline{X}$ debe ser paralela a $ln\{H_0(t)\}$ en el tiempo, luego si se tiene la hipótesis (PH) se mantiene las curvas de supervivencia para diferentes $\underline{X}'s$ deben ser paralelas.

Veamos un ejemplo en el que se cumple la hipótesis de riesgos proporcionales.

Ejemplo 3.2.2
Consideramos los tiempos de fallo (en millones de revoluciones) de rodamientos de bolas de cerámica dividiéndolas, según hayan salido de la fábrica 1 o de la fábrica 2 (datos extraídos del Ejemplo 2.2.3):

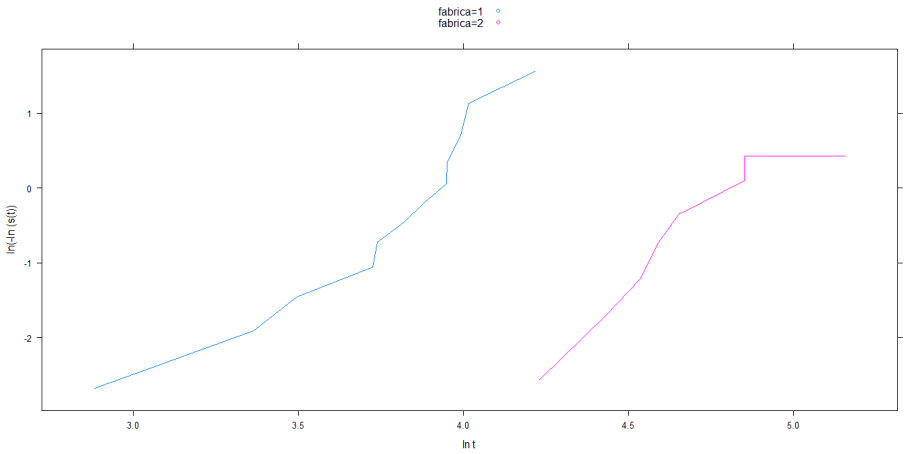

Figura 3.2: Riesgos proporcionales del rodamiento de bolas cerámicas.

En este caso, el gráfico sugiere, que la hipótesis PH es satisfactoria, ya que ambas curvas son paralelas.

3.3. Estimación de los coeficientes

En el modelo de regresión de Cox los parámetros $\underline{\beta} = (\beta_1, \ldots, \beta_p)'$ se estiman maximizando el logaritmo de la denominada ***función de verosimilitud parcial***. La maximización de dicha función se realiza mediante métodos numéricos, obteniendo de esta forma la estimación $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$.
Con la estimación de estos parámetros ya tendremos la componente paramétrica totalmente especificada en el modelo

$$h(t, \underline{X}) = h_0(t)exp(\underline{\hat{\beta}}'\underline{X})$$

y consecuentemente, podremos hacer inferencia sobre dicho vector de parámetros, para calcular la razón de riesgos, de interés en el estudio.

La función de verosimilitud parcial, que a continuación vamos a definir, se denomina parcial, debido a que, se tiene en cuenta únicamente, en la función de verosimilitud, las probabilidades de los tiempos de muerte/fallo, y no incluye las probabilidades de los tiempos de datos censurados.
Sin embargo, en el cálculo de las probabilidades de los tiempos, se tiene en cuenta a todos los individuos (censurados o no a posteriori) en riesgo, al inicio de los diferentes tiempos de muerte/fallo.

Denominamos $\mathcal{L} \equiv \mathcal{L}(\beta_1, \ldots, \beta_p)$ a la *función de verosimilitud parcial*. Supongamos que tenemos k tiempos de muerte, y que no hay empates. Así, tendremos $n - k$ tiempos censurados.

Notación 3.3.1

- Sea $t_{(1)}, \ldots, t_{(k)}$, los tiempos de muerte ordenados.
- Sea R_i para $i = 1, \ldots, k$, el conjunto de los individuos que están en riesgo antes del tiempo $t_{(i)}$.

Definimos por $\mathcal{L}_i \equiv \mathcal{L}_{t_{(i)}}(\beta_1, \ldots, \beta_p)$ para $i = 1, \ldots, k$ a cada porción de la verosimilitud parcial anterior perteneciente a los diferentes tiempos de muerte $t_{(1)}, \ldots, t_{(k)}$.

Construiremos la función de verosimilitud parcial como el producto de cada una de las aportaciones de los k tiempos de muerte:

$$\left.\begin{array}{rcl} i = 1 \longrightarrow & \mathcal{L}_1 & \equiv \mathcal{L}_{t_{(1)}}(\beta_1, \ldots, \beta_p) \\ i = 2 \longrightarrow & \mathcal{L}_2 & \equiv \mathcal{L}_{t_{(2)}}(\beta_1, \ldots, \beta_p) \\ & \vdots & \\ i = k \longrightarrow & \mathcal{L}_k & \equiv \mathcal{L}_{t_{(k)}}(\beta_1, \ldots, \beta_p) \end{array}\right\} \Longrightarrow \mathcal{L} = \prod_{i=1}^{k} \mathcal{L}_i$$

Veamos cuanto vale exactamente cada una de las $\mathcal{L}_i \equiv \mathcal{L}_{t_{(i)}}(\beta_1, \ldots, \beta_p)$ para $i = 1, \ldots, k$.

Para cada unidad $l \in R_i$

$$h(t_{(i)}; X_l)\delta t = P(\text{fallo en } \left[t_{(i)}, t_{(i)} + \delta t\right))$$

dado que la función de riesgo en $t_{(i)}$ da la probabilidad instantánea de fallo condicionada a haber alcanzado $t_{(i)}$ cuyas unidades en R_i lo han hecho

$$P(\text{de que un individuo } i \text{ falle/a que haya un fallo en } t_{(i)}) =$$

$$= \frac{h(t_{(i)}, \underline{X}_i)}{\sum_{l \in R_i} h(t_{(i)}, \underline{X}_l)} = \frac{h_0(t_{(i)}) exp(\underline{\beta}' \underline{X}_i)}{\sum_{l \in R_i} h_0(t_{(i)}) exp(\underline{\beta}' \underline{X}_l)} = \frac{h_0(t_{(i)}) exp(\underline{\beta}' \underline{X}_i)}{h_0(t_{(i)}) \sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l)}$$

Esta última igualdad se tiene porque $h_0(t_{(i)})$ no depende de l, lo que implica que se anule tanto en el numerador, como en el denominador.
$\underline{X}_i$ es el vector de covariables para el individuo con tiempo de muerte $t_{(i)}$ y $\underline{X}_l$, para $l \in R_i$, el vector de covariables de cada uno de los individuos de R_i.

Por tanto nos queda la expresión:

$$\mathcal{L}_{t_{(i)}}(\beta_1, \ldots, \beta_p) = \frac{exp(\underline{\beta}' \underline{X}_i)}{\sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l)}$$

Se observa que la función de verosimilitud parcial, así calculada, no depende de las cuantías de los tiempos, sino tan solo de su ordenación, y de si el dato estaba o no censurado. Como consecuencia se podría obtener las mismas estimaciones de $\underline{\beta}$ para distintos datos, siempre que estos tengan el mismo patrón de orden y censura en los tiempos de supervivencia.

En resumen, la función de verosimilitud parcial queda de la siguiente manera:

$$\mathcal{L}(\underline{\beta}) = \prod_{i=1}^{k} \left\{ \frac{exp(\underline{\beta}' \underline{X}_i)}{\sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l)} \right\}$$

Una vez que tenemos la función de verosimilitud parcial construida, procedemos a calcular $\hat{\underline{\beta}}$, las estimaciones de los coeficientes del modelo $\underline{\beta}$:

$$\text{Sea } \hat{\underline{\beta}} = \text{máx } \ ln\ \mathcal{L}(\underline{\beta}), \text{ calculamos:} \tag{3.7}$$

$$ln\ \mathcal{L}(\underline{\beta}) = \sum_{i=1}^{k} ln \left\{ \frac{exp(\underline{\beta}' \underline{X}_i)}{\sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l)} \right\} = \sum_{i=1}^{k} \left\{ ln\ exp(\underline{\beta}' \underline{X}_i) - ln \sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l) \right\} =$$

$$= \sum_{i=1}^{k} \underline{\beta}' \underline{X}_i - \sum_{i=1}^{k} ln \left\{ \sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l) \right\}$$

A continuación derivamos:

$$\frac{\partial ln\ \mathcal{L}}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{k} \underline{\beta}' \underline{X}_i - \sum_{i=1}^{k} ln\left\{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)\right\}\right) =$$

$$= \underbrace{\frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{k} \underline{\beta}' \underline{X}_i\right)}_{(1)} - \underbrace{\frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{k} ln\left\{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)\right\}\right)}_{(2)}$$

Calculamos (1):

$$\frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{k} \underline{\beta}' \underline{X}_i\right) = \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{k} \beta_j X_{ij}\right) = \sum_{i=1}^{k} \frac{\partial}{\partial \beta_j}\left(\beta_j X_{ij}\right) = \sum_{i=1}^{k} X_{ij}$$

Calculamos (2):

$$\frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{k} ln\left\{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)\right\}\right) = \sum_{i=1}^{k} \frac{\partial}{\partial \beta_j}\left(ln\left\{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)\right\}\right) =$$

$$= \sum_{i=1}^{k}\left(\frac{\sum_{l\in R_i} \frac{\partial}{\partial \beta_j}\left\{exp\left(\beta_j X_{jl}\right)\right\}}{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)}\right) = \sum_{i=1}^{k}\left(\frac{\sum_{l\in R_i} X_{jl}\left\{exp(\underline{\beta}' \underline{X}_l)\right\}}{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)}\right)$$

Por tanto:

$$\frac{\partial ln\ \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{k} X_{ij} - \sum_{i=1}^{k}\left(\frac{\sum_{l\in R_i} X_{jl}\left\{exp(\underline{\beta}' \underline{X}_l)\right\}}{\sum_{l\in R_i} exp(\underline{\beta}' \underline{X}_l)}\right) \tag{3.8}$$

Derivando por segunda vez

$$\frac{\partial^2 ln\ \mathcal{L}}{\partial \beta_i \partial \beta_j}, \tag{3.9}$$

Si igualamos (3.8) a 0, para $j = 1, \ldots, p$, obtendremos las ecuaciones que nos permitirán obtener las estimaciones de $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ mediante la utilización de algún método numérico.

De (3.9), comprobamos que realmente es un máximo, y podemos obtener, como ocurre cuando se trabaja en general con una función de verosimilitud, la ***matriz de información (observada)***, $\mathbf{I}(\underline{\beta})$, donde cada elemento se iguala a:

$$I_{ij}(\underline{\beta}) = -\frac{\partial^2 ln\ \mathcal{L}}{\partial\beta_i\partial\beta_j}.$$

Así, la matriz de varianzas y covarianzas estimada (pxp) es $\hat{\Sigma} = I^{-1}(\underline{\hat{\beta}})$.
Cabe notar que este estimador, obtenido a partir de la maximización de la función de verosimilitud parcial, es asintóticamente no sesgado, eficiente y normal. Además, aunque el estimador $\underline{\hat{\beta}}$ estime consistentemente el vector de parámetros $\underline{\beta}$ no es completamente eficiente. Finalmente, la distribución de $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ es aproximadamente normal, de media $(\beta_1, \ldots, \beta_p)$ y matriz de varianzas y covarianzas Σ.

3.3.1. Contrastes de hipótesis

Tras el ajuste del modelo de Cox, se ha de comprobar si las variables del modelo son significativas. Para ello, existen pruebas que se encargan de validar las correspondientes hipótesis. En ellas, se considera el vector de parámetros estimados $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$, y la matriz de varianzas y covarianzas estimada, $\hat{\Sigma}$, lo que nos permite utilizar tests análogos a los utilizados en un modelo lineal, o lineal generalizado.

Para contrastar la hipótesis $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$, podemos utilizar el ***estadístico de Wald*** dado por:

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim N(0,1) \quad \text{asintóticamente.}$$

La fórmula para calcular un intervalo de confianza aproximado al nivel $(1-\alpha)$ para el coeficiente β_j es la siguiente:

$$\hat{\beta}_j \pm Z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\beta}_j)} = \hat{\beta}_j \pm Z_{1-\alpha/2}\ s.e.(\hat{\beta}_j).$$

En cambio, si lo que queremos hacer es el test $H_0 : \underline{\beta} = \underline{\beta_0}$ vs. $H_1 : \underline{\beta} \neq \underline{\beta_0}$ se suelen utilizar tres contrastes:

1. **El contraste de Wald.** Este contraste se basa en que los coeficientes $\underline{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ siguen una distribución aproximadamente normal de media $(\beta_1, \ldots, \beta_p)$ y matriz de varianzas y covarianzas $\hat{\Sigma} = I^{-1}(\underline{\hat{\beta}})$. El estadístico se define como:

   $$\mathcal{X}^2_W = (\underline{\hat{\beta}} - \underline{\beta}_0)' I(\underline{\hat{\beta}})(\underline{\hat{\beta}} - \underline{\beta}_0),$$

   que bajo la hipótesis nula sigue una distribución $\mathcal{X}^2$ con p grados de libertad.

2. **El contraste de la razón de verosimilitud.** En este contraste se utiliza el valor de la función de verosimilitud parcial evaluada en $\underline{\hat{\beta}}$ ($\mathcal{L}(\underline{\hat{\beta}})$) y evaluada en $\underline{\beta}_0$ ($\mathcal{L}(\underline{\hat{\beta}}_0)$):

$$\mathcal{X}^2_{LR} = -2\{ln\mathcal{L}(\underline{\beta}_0) - ln\mathcal{L}(\underline{\hat{\beta}})\},$$

que bajo la hipótesis nula sigue una distribución $\mathcal{X}^2$.

3. **El contraste del "score" (Log Rank).** En este contraste se utiliza el gradiente (derivadas) del logaritmo de la función de verosimilitud parcial evaluada en la hipótesis nula y supone que bajo la hipótesis nula el vector scores:

$$\mathcal{X}^2_{SC} = \left(\frac{\partial\mathcal{L}(\underline{\beta}_0)}{\partial\underline{\beta}}\right)' \left(-\frac{\partial\mathcal{L}^2(\underline{\beta}_0)}{\partial\underline{\beta}\partial\underline{\beta}'}\right)^{-1} \frac{\partial\mathcal{L}(\underline{\beta}_0)}{\partial\underline{\beta}},$$

sigue una distribución aproximada $\mathcal{X}^2$.

Nota 3.3.1
El contraste de Wald tiene una interpretación más directa, que el contraste de verosimilitud y el del score, sin embargo no es invariante ante diferentes parametrizaciones y los otros dos, sí. Con el contraste del score sólo hace falta maximizar bajo la hipótesis nula, con lo que si hay que realizar el test para varios parámetros, es más rápido computacionalmente. Sin embargo, el test de la máxima verosimilitud converge más rápido hacia la distribución normal.
Ante la duda de cuál utilizar, es recomendable decantarse por el test de la razón de verosimilitudes.

3.4. Residuos en el análisis de supervivencia

Una de las ventajas que han surgido del enfoque del análisis de supervivencia es la posibilidad de efectuar análisis de residuos.
Los residuos se pueden utilizar para:

1. Descubrir la forma funcional correcta de un predictor continuo.
2. Identificar los sujetos que están pobremente pronosticados por el modelo.
3. Identificar los puntos o individuos de influencia.
4. Verificar el supuesto de riesgo proporcional.

Existen cinco tipos de residuos de interés en el modelo de Cox: los de Cox-Snell, los de martingala, los de desvíos(*Deviances*) y los de Schoenfeld. De ellos pueden derivarse otros dos: los residuos escalados de Schoenfeld y los *dfbetas*.

Para poder explicar cada residuo, se consideramos un conjunto de n sujetos independientes de tal manera que el proceso de conteo $N_i \equiv \{N_i(t), t \geq 0\}$ para el i-ésimo sujeto es el número de eventos observados hasta el tiempo t, y suponemos que la función de intensidad para $N_i(t)$ viene dada por la expresión:

$$\alpha_i(t) = D_i(t)dH(t, \underline{X}_i(t)) = D_i(t) \cdot exp(\underline{\beta}^t \underline{X}_i(t))dH_0(t)$$

donde:

- $D_i(t)$ es un proceso 0-1 que indica si el i-ésimo sujeto está en riesgo en el tiempo t.
- $\underline{\beta}$ es un vector de coeficientes de regresión.
- $\underline{X}(t)$ es un vector p-dimensional de procesos de las covariables.

Pero antes de pasar al análisis de los cinco tipos de residuos, veamos previamente otra forma de estimar la función de supervivencia.
Como vimos en (3.5), bajo la hipótesis de riesgos proporcionales, la función de supervivencia S de los individuos (o unidades) con sus respectivas covariables $\underline{X}$ se define como:

$$S(t, \underline{X}) = S_0(t)\hat{}\{exp(\underline{\beta}'\underline{X})\} \tag{3.10}$$

Por (3.7) tenemos el estimador de $\underline{\beta}$ luego necesitamos estimar la función $S_0(t)$ ó análogamente la función $H_0(t)$ porque por (3.6):

$$S_0(t) = exp\{-H_0(t)\} \Longrightarrow H_0(t) = -ln\ S_0(t) \tag{3.11}$$

Por tanto, vamos a estimar $H_0(t)$ usando el estimador Breslow que explicaremos a continuación.

Estimador Breslow:

$$\hat{H}_0(t) = \sum_{t_{(i)} \leq t} \frac{d_i}{\sum_{l \in R_i} exp(\underline{\hat{\beta}}'\underline{X}_l)}$$

que se reduce al estimador de Nelson-Aalen, cuando no existe ninguna covariable, es decir, cuando $\underline{\beta} = \underline{0}$:

$$\hat{H}_0(t) = \sum_{t_{()} \leq t} \frac{d_i}{\sum_{l \in R_i} exp(0)} = \sum_{t_{(i)} \leq t} \frac{d_i}{\sum_{l \in R_i} 1} = \sum_{t_{(i)} \leq t} \frac{d_i}{n_i}$$

porque R_i son todas las unidades que están en riesgo justo antes de $t_{(i)}$, es decir, n_i.

3.4.1. Residuos de Cox-Snell

Si un analista está interesado en evaluar el ajuste global del modelo planteado, los residuos más comunes utilizados por este tipo de análisis son los de Cox-Snell. Si el modelo de CPH dado por (3.1) se mantiene, entonces las estimaciones del tiempo de supervivencia del modelo planteado vienen dadas por un estimador de la función de supervivencia $\hat{S}(t_{(i)})$ que debe ser muy similar al verdadero valor de $S(t_{(i)})$.
Para evaluar esto, se calcula los residuos de Cox-Snell definidos de la forma:

$$r_{cs_i} = \hat{H}(t_{(i)}, \underline{X}_i) \overset{(3.11)}{=} -ln\hat{S}(t_{(i)}, \underline{X}_i) \overset{(3.10)}{=} -ln[\hat{S}_0(t_{(i)})\hat{}\{exp(\underline{\hat{\beta}}'\underline{X}_i)\}] =$$

$$= \{exp(\underline{\hat{\beta}}'\underline{X}_i)\}[-ln\hat{S}_0(t_{(i)})] \overset{(3.11)}{=} \hat{H}_0(t_{(i)})\{exp(\underline{\hat{\beta}}'\underline{X}_i)\}$$

Un resultado importante es que si el modelo apropiado se ajusta bien a los datos, entonces los r_{cs_i} tendrán para cada i un valor exp(1), es decir, distribución exponencial con razón o tasa de riesgo igual a 1.
Para probar si los residuos de Cox-Snell están o no aproximadamente distribuidos de forma exponencial, tenemos que construir su gráfico de residuos. La lógica de este método es sencilla. Si los residuos de Cox-Snell están, de hecho, distribuidos de forma exponencial, entonces si dibujamos el estimador de Nelson-Aalen de la tasa de riesgo acumulado de los residuos de Cox-Snell contra r_{cs_i}, debería aproximarse a una línea recta que pasa por el origen con pendiente igual a 1 para que el modelo planteado se ajuste bien a los datos.

3.4.2. Residuos de martingala

Los residuos de martingala se definen como:

$$M_i(t) = N_i(t) - E_i(t) = N_i(t) - \int_0^t D_i(s) \cdot exp(\underline{\beta}'\underline{X}_i(s))dH_0(s), \quad i = 1, \ldots, n.$$

Sea $\underline{\beta}$ estimada por la función de máxima verosimilitud parcial y el riesgo acumulado H_0 por el estimador del riesgo base de Breslow entonces el residuo de martingala se estima de la forma:

$$\hat{M}_i(t) = N_i(t) - \hat{E}_i(t) = N_i(t) - \int_0^t D_i(s) \cdot exp(\underline{\hat{\beta}}'\underline{X}_i(s))d\hat{H}_0(s), \quad i = 1, \ldots, n.$$

3.4.3. Residuos basados en el estadístico Deviance

Los residuos de desvíos se obtienen mediante una transformación de normalización de los de martingala, y son similares en forma a los residuos de desvíos (*deviances*) en la regresión de Poisson. Si todas las covariables son fijas en el tiempo, los residuos de desvíos toman la forma:

$$de_i = \text{signo}(\hat{M}_i) \cdot \sqrt{-\hat{M}_i - N_i log((N_i - \hat{M}_i/N_i)}$$

Los residuos de desvíos se utilizan para la detección de valores atípicos (*outliers*).

3.4.4. Residuos de Schoenfeld

Para desarrollar este residuo, hay que recordar:

$$P(\text{de que un individuo } k \text{ falle/a que haya un fallo en } t_{(i)}) =$$

$$= \frac{h(t_{(i)}, \underline{X}_k)}{\sum_{l \in R_i} h(t_{(i)}, \underline{X}_l)} = p_k$$

Luego

$$E[\underline{X}/R_i] = \sum_{k \in R_i} \underline{X}_k \, p_k = \frac{\sum_{k \in R_i} \underline{X}_k \, h(t_{(i)}, \underline{X}_k)}{\sum_{l \in R_i} h(t_{(i)}, \underline{X}_l)} \overset{(3.2)}{=} \frac{\sum_{k \in R_i} \underline{X}_k \, exp(\underline{\beta}' \underline{X}_k)}{\sum_{l \in R_i} exp(\underline{\beta}' \underline{X}_l)}$$

Por tanto, definimos el residuo de Schoenfeld (parcial) como

$$\underline{r}_i = \underline{X}_i - \hat{E}[\underline{X}/R_i]$$

donde $\hat{E}$ indica que $\underline{\beta}$ ha sido reemplazado por su estimador $\underline{\hat{\beta}}$.
Esto es lo que se conoce como "observado menos esperado" de una forma residual, pero aplicada a las variables predictoras $\underline{X}$'s y no a la variable dependiente t.

Nota 3.4.1

1. Se trata de un vector con una componente para cada elementos de $\underline{X}$.

2. Los residuos de Schoenfeld solo están definidos para aquellas unidades con tiempos de fallo observados.

3.4.5. Residuos escalados de de Schoenfeld

Sea $\hat{V}(\underline{r}_i)$ la matriz de covarianzas estimada de $\underline{r}_i$. Entonces $\{\hat{V}(\underline{r}_i)\}^{-1}$ es una versión estandarizada del residuo de Schoenfeld. Se considera mejor que el residuo no estandarizado para detectar problemas en el modelo de Cox.

Es más simple usar la aproximación:

$$\{\hat{V}(\underline{r}_i)\}^{-1} \sim r\hat{V}(\hat{\underline{\beta}})$$

donde r es el número total de observaciones sin censura y $\hat{V}(\hat{\underline{\beta}})$ es la matriz de covarianza de las estimaciones de $\underline{\beta}$.

Esto da los residuos escalados de de Schoenfeld:

$$\underline{r}_i^* = r\hat{V}(\hat{\underline{\beta}})\ \underline{r}_i$$

3.4.6. Residuos *dfbeta*

Los residuos *dfbeta* sirven para determinar la influencia de cada observación en la estimación de los coeficientes de regresión.
Este residuo calcula el cambio aproximado en el j-ésimo coeficiente (es decir, la j-ésima covariable) si la observación i-ésima se elimina del conjunto de datos y se vuelve a estimar el modelo sin esta observación. Así para el paciente i el valor *dfbeta* correspondiente a la variable j es el siguiente:

$$dfbeta_i(\beta_j) = \hat{\beta}_j - \hat{\beta}_j(i)$$

Los valores *dfbeta* pueden estandarizarse dividiendo por el error estándar del coeficiente correspondiente.
En su representación gráfica se suelen mostrar los valores de los residuos *dfbeta* estandarizados para cada covariable del modelo frente a los índices de paciente (número de orden). Si la supresión de una observación hace que el coeficiente incremente, el residuo *dfbeta* es negativo y viceversa.

3.5. Modelo estratificado

Como vimos en el Ejemplo 3.2.1, hay casos en que puede violarse la presunción de riesgos proporcionales para alguna covariable. En tal caso, puede ser posible estratificar esa covariable y utilizar el modelo de riesgos proporcionales dentro de cada estrato y considerando las otras covariables.

En este caso, los sujetos en el estrato m-ésimo tienen una función baseline hazard arbitraria, $h_{0m}(t)$, y el efecto de otras covariables explicativas sobre la

función de riesgo puede representarse por un modelo de riesgos proporcionales en ese estrato de la forma

$$h_m(t;\underline{X}) = h_{0m}(t)exp(\underline{\beta}'\underline{X}), \qquad m = 1,\ldots,s.$$

En este modelo, los coeficientes de regresión se supone que son los mismos en todos los estratos, aunque las funciones de baseline hazard pueden ser diferentes, y no estar relacionadas.

De hecho, el logaritmo de la función de verosimilitud parcial del modelo no estratificado (3.1), es

$$ln\ \mathcal{L}(\underline{\beta}) = \sum_{i=1}^{k} \underline{\beta}'\underline{X}_i - \sum_{i=1}^{k}\ ln\left\{\sum_{l\in R_i} exp(\underline{\beta}'\underline{X}_l)\right\}$$

se aplica a cada estrato por separado. Así, en general, para los s-estratos obtenemos:

$$ln\ \mathcal{L}(\underline{\beta}) = \sum_{m=1}^{s}\sum_{i=1}^{k_m} \underline{\beta}'\underline{X}_{im} - \sum_{m=1}^{s}\sum_{i=1}^{k_m}\ ln\left\{\sum_{l\in R_{im}} exp(\underline{\beta}'\underline{X}_{lm})\right\}$$

donde $\underline{X}_{im}$ es la covariable de la unidad que está fallando en el i-ésimo tiempo dentro del estrato $m = 1,\ldots,s$ cuando el conjunto de riesgo en ese estrato es R_{im}.

Por tanto, todo lo que cambia en la estimación $\underline{\beta}$ (común a todos los estratos) es que se necesita el doble de sumas.

Capítulo 4

Aplicación práctica con el software R

Introducción

En este capítulo se describen brevemente las herramientas y técnicas descritas a lo largo de todo el trabajo para la realización del análisis estadístico que consiste en estudiar el tiempo que siguen un grupo de individuos adictos a la heroína con un tratamiento de metadona (datos contenidos en Caplehorn y Bell [3]). Tenemos una muestra de 238 pacientes (clientes) en relación con la siguientes variables explicativas:

- **cliente**: Indica el número de observación. Es una etiqueta que identifica al individuo.
- **clinica**: Indica el tipo de clínica en el que ha estado el paciente. Toma el valor 0 si ha estado en la clínica A y 1 si ha estado en la clínica B.
- **censura**: Es una variable que está codificada como 1 si el cliente continua en el tratamiento, o en caso contrario, 0 si el cliente ha abandonado el tratamiento.
- **tiempo**: Muestra los días en los que el paciente ha sido tratado con metadona.
- **prision**: Indica si el cliente ha estado o no en prisión. Toma el valor 1 si ha estado y, 0 sino.
- **dosis**: Indica la cantidad de metadona que se le suministra a cada cliente.

Preparación de los datos

En primer lugar, para realizar el estudio sobre dicho conjunto de datos, los cargamos con la función *read.table()*:

```
cdat<-read.table("heroina.txt", header=TRUE,row.names = 1)
head(cdat)

##   cliente clinica censura tiempo prision dosis
## 1     244       0       0      2       1    60
## 2     202       0       1      7       1    40
## 3     190       0       1     17       1    40
## 4     247       0       1     19       1    40
## 5     220       0       0     28       0    50
## 6     230       0       0     28       0    50
```

Veamos, con la función *str()*, el aspecto de estos datos:

```
attach(cdat)
str(cdat)

## 'data.frame': 238 obs. of  6 variables:
##  $ cliente: num  244 202 190 247 220 230 203 212 261 248 ...
##  $ clinica: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ censura: num  0 1 1 1 0 0 1 1 1 1 ...
##  $ tiempo : num  2 7 17 19 28 28 29 30 33 35 ...
##  $ prision: num  1 1 1 1 0 0 1 0 1 0 ...
##  $ dosis  : num  60 40 40 40 50 50 60 60 60 60 ...
```

La información que nos devuelve es que es una muestra de 238 datos con 6 variables.

El objeto Surv

Hemos de preparar los datos para realizar un estudio de supervivencia con el paquete estadístico *survival* mediante la función *survfit()*

```
install.packages("survival")
library(survival)
```

Un objeto Surv no es más que la combinación de información entre los tiempos y su censura.

Es necesario trabajar con los datos en este formato para más tarde aplicar algunas técnicas.

```
sup<-Surv(tiempo,censura)
```

Vemos como obtener si la observación del evento no está censurada

```
Surv(5,1)

## [1] 5
```

Si la observación del evento está censurada

```
Surv(5,0)

## [1] 5+
```

Si queremos los primeros 6 términos

```
head(sup)

## [1]  2+  7  17  19  28+ 28+
```

como vemos, se representa con un "+" a la derecha del dato, aquel que está censurado.

Estimación no paramétrica de la función de supervivencia

Estimador de Kaplan-Meier

Recordamos por teoría, visto en el Capítulo 2, la finalidad de este método es que la proporción acumulada que se mantiene en el tratamiento se calcula para el tiempo de supervivencia individual de cada paciente y no se agrupan los tiempos en intervalos. Además, matemáticamente su fórmula viene dada por:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j}$$

El estimador de Kaplan-Meier para la función de supervivencia se obtiene a través del paquete estadístico *survival* (cargado anteriormente) mediante la función *survfit()*. Para ello, consideramos los estratos (clinica A y B) para poder compararlos.

```
outp<-survfit(Surv(tiempo,censura)~strata(clinica),
              type="kaplan-meier", data=cdat)
attach(outp)
```

Veamos un resumen estadístico, mostrando los 10 primeros términos de cada estrato

```
summary(outp)

## Call: survfit(formula = Surv(tiempo, censura) ~ strata(clinica),
##               data = cdat, type = "kaplan-meier")
##

##
##                 strata(clinica)=clinica=0
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    7    162       1   0.9938 0.00615      0.98184        1.000
##   17    161       1   0.9877 0.00868      0.97080        1.000
##   19    160       1   0.9815 0.01059      0.96094        1.000
##   29    157       1   0.9752 0.01223      0.95155        0.999
##   30    156       1   0.9690 0.01366      0.94258        0.996
##   33    155       1   0.9627 0.01493      0.93390        0.992
##   35    154       1   0.9565 0.01609      0.92545        0.989
##   37    153       1   0.9502 0.01716      0.91719        0.984
##   41    152       1   0.9440 0.01815      0.90907        0.980
##   47    151       1   0.9377 0.01907      0.90107        0.976
##
##                 strata(clinica)=clinica=1
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   13     74       1    0.986  0.0134        0.961        1.000
##   26     73       1    0.973  0.0189        0.937        1.000
##   35     72       1    0.959  0.0229        0.916        1.000
##   41     71       1    0.946  0.0263        0.896        0.999
##   79     68       1    0.932  0.0294        0.876        0.991
##  109     66       1    0.918  0.0321        0.857        0.983
##  122     65       1    0.904  0.0346        0.838        0.974
##  143     64       1    0.890  0.0368        0.820        0.965
##  149     62       1    0.875  0.0389        0.802        0.955
##  161     61       1    0.861  0.0408        0.785        0.945
```

Esta salida devuelve los siguientes valores:

- **time**: tiempo de supervivencia de cada cliente dado en días.

- **n.risk**: número de elementos en riesgo en ese instante.
- **n.event**: número de elementos que fallan en ese momento.
- **survival**: es la estimación de Kaplan-Meier de la función de supervivencia en el instante correspondiente, $\hat{S}(t)$.
- **std.err**: es el error estándar asociado a cada $\hat{S}(t)$.
- **lower 95 % CI**: es el extremo inferior del intervalo de confianza para $S(t)$ al nivel 95 %.
- **upper 95 % CI**: es el extremo superior del intervalo de confianza para $S(t)$ al nivel 95 %.

Representamos la estimación calculada de $S(t)$. Para ello, usaremos la función *ggsurvplot()* contenida en el paquete *survminer*, (Kasambara y Kosinski [14]).

```
install.packages("survminer")
library(survminer)
```

```
ggsurvplot(fit = outp, data = cdat, conf.int = T, title = "Curva
           de Supervivencia", xlab = "Tiempo", ylab = "Probabilidad
           de supervivencia")
```

En esta figura presentamos las gráficas de las funciones de supervivencia estimadas, con las bandas de confianza asociadas por el método de Kaplan-Meier, para los dos tipos de clínicas. Notamos que las curvas estimadas son de tipo escalonada, ambas con datos censurados (prueba de ello es la aparición de símbolos "+" en cada una de ellas).

Observamos diferencias en las curvas de supervivencia de las dos clínicas. En general, $\hat{S}(t)$ es mayor para la clínica B (codificada con el valor 1).

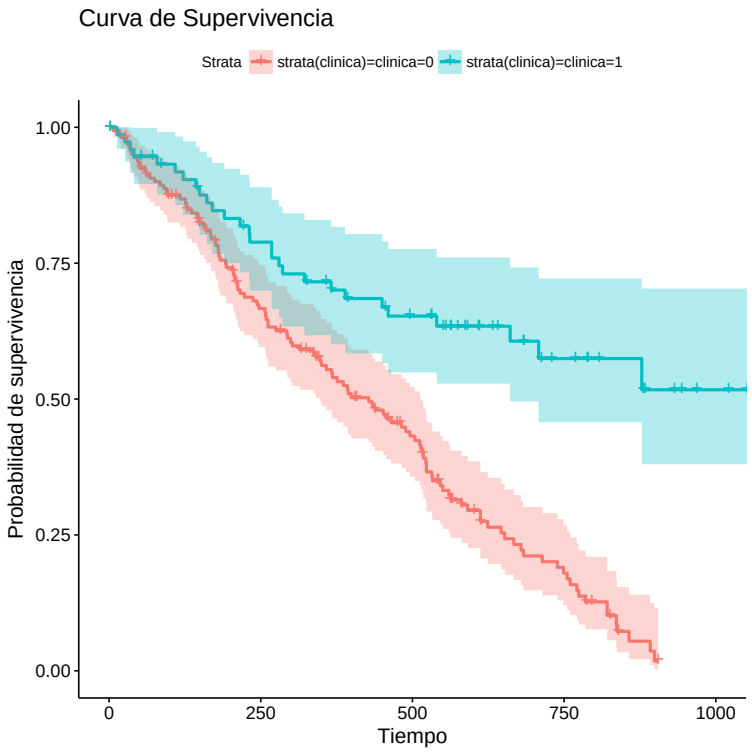

Figura 4.1: Estimador de Kaplan-Meier.

Estimador de Nelson-Aalen

Recordar que, $H(t)$ es la función de hazard acumulada (o cumulative hazard) definida como:

$$H(t) = \int_0^t h(u)\, du$$

que representa la suma de las probabilidades de fallo (en nuestro caso, de abandonar el tratamiento) en el intervalo $(0, t]$.
Matemáticamente, su fórmula viene dada por:

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

Instalamos los siguientes paquetes que nos harán falta para calcular este estimador.

```
install.packages("ggfortify")
library(ggfortify)
```

```
install.packages("dplyr")
library(dplyr)
```

En este caso, la información se extrae calculando la suma acumulada del número de eventos entre las personas en riesgo como se vió en el Capítulo 2.

```
mod<- survfit(Surv(tiempo,censura)~clinica, cdat)
R <- mod %>% fortify %>% group_by(strata) %>% mutate(CumHaz =
      cumsum(n.event/n.risk))
R
```

```
## # A tibble: 215 x 10
## # Groups:   strata [2]
##  time n.risk n.event n.censor  surv std.err upper lower strata
## <dbl>  <dbl>   <dbl>    <dbl> <dbl>   <dbl> <dbl> <dbl> <fct>
## 1     2    163       0        1 1     0       1     1     1
## 2     7    162       1        0 0.994 0.00619 1     0.982 1
## 3    17    161       1        0 0.988 0.00878 1     0.971 1
## 4    19    160       1        0 0.981 0.0108  1     0.961 1
## 5    28    159       0        2 0.981 0.0108  1     0.961 1
## 6    29    157       1        0 0.975 0.0125  0.999 0.952 1
## 7    30    156       1        0 0.969 0.0141  0.996 0.943 1
## 8    33    155       1        0 0.963 0.0155  0.992 0.934 1
## 9    35    154       1        0 0.956 0.0168  0.989 0.925 1
## 10    37    153       1        0 0.950 0.0181  0.984 0.917 1
## # ... with 205 more rows
##
##
## # A tibble: 215 x 10
## # Groups:   strata [2]
##  CumHaz
##   <dbl>
##   0
##   0.00617
##   0.0124
##   0.0186
##   0.0186
##   0.0250
##   0.0314
##   0.0379
##   0.0444
##   0.0509
## # ... with 205 more rows
```

Cuya gráfica asociada es la siguiente:

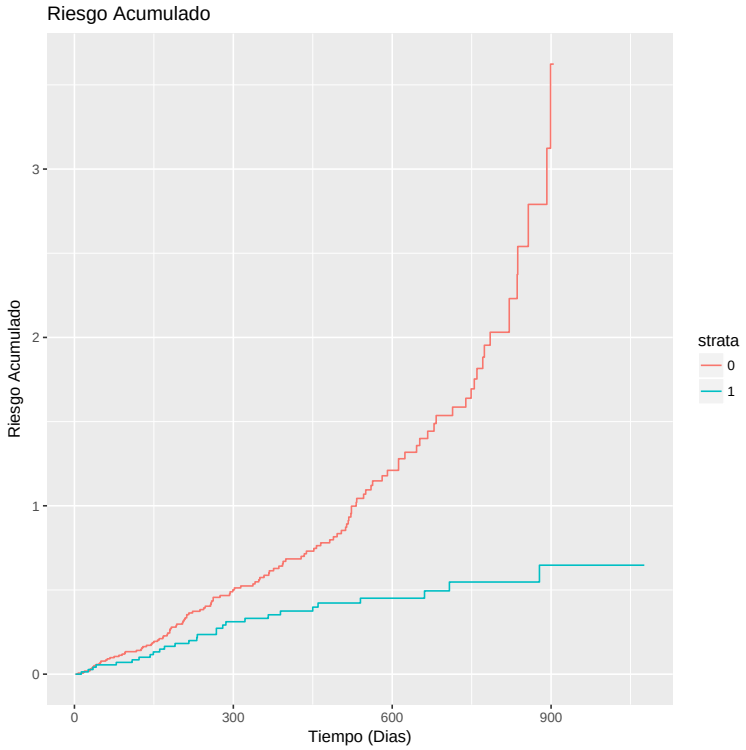

Figura 4.2: Estimador de Nelson-Aalen.

De nuevo observamos bastantes diferencias entre las clínicas. Se aprecia que hay un menor $\hat{H}(t)$ para la clínica B (codificada con el valor 1) lo que es coherente con los resultados obtenidos al aplicar Kaplan-Meier puesto que $\hat{H}(t) = -ln\hat{S}(t)$.

Para formalizar estas apreciaciones realizamos, a continuación, un test de hipótesis.

Test de Log-Rank

Para aplicar este test, definimos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 & : & S_1(t) & = & S_2(t) \\ H_1 & : & S_1(t) & \neq & S_2(t) \end{cases}$$

Para comparar ambas funciones de supervivencia, aplicamos el test de Log-Rank:

```
out1<-survdiff(Surv(tiempo, censura) ~ clinica,data=cdat)
out1

## Call:
## survdiff(formula = Surv(tiempo, censura) ~ clinica, data = cdat)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## clinica=0 163      122     90.9      10.6      27.9
## clinica=1  75       28     59.1      16.4      27.9
##
##  Chisq= 27.9  on 1 degrees of freedom, p= 1.28e-07
```

En base a los valores obtenidos en la tabla anterior, el p-valor $p = 1.28 \cdot 10^{-07} < 0.05$ luego se rechaza la hipótesis nula de igualdad de funciones de supervivencia (para un nivel de significación del 5 %).

En consecuencia, podemos concluir que existe una clara evidencia de desigualdad entre las curvas de supervivencia. Además, $\mathcal{X}_1^2 = 27.9$.

Como son distintas, los datos generados permiten a su vez realizar una estimación del riesgo $hr = \dfrac{O0/E0}{O1/E1}$
donde:

- $O0$ representa el valor "Oberved" en la clínica A (valor 0)
- $E0$ representa el valor "Expected" en la clínica A (valor 0)
- $O1$ representa el valor "Oberved" en la clínica B (valor 1)
- $E1$ representa el valor "Oberved" en la clínica B (valor 1)

luego

```
hr<-(122/90.9)/(28/59.1)
hr

## [1] 2.832862
```

Por tanto, los clientes que están en la clínica A se mantienen en el tratamiento 2.832862 veces más que los de la clínica B.

Ajuste del modelo Cox

Consideramos un modelo de Cox con las covariables *prision* y *dosis*, usando la covariable *clinica* como estrato para diferenciar el comportamiento entre clínicas. El modelo es de la forma:

$$h_m(t;\underline{X}) = h_{0m}(t)exp(\underline{\beta}'\underline{X}), \qquad m = 1, 2.$$

donde $h_1(t;\underline{X})$ referencia a la clínica A y $h_2(t;\underline{X})$ referencia a la clínica B. Hay que tener en cuenta que observando dicha expresión, los coeficientes correspondientes a cada covariable $\underline{\beta} = (\beta_1, \beta_2)$ son iguales en ambos modelos. Lo que varía respecto a las clínicas, es la función baseline hazard $h_{0m}(t)$.

Creamos el modelo de Cox con *prision* y *dosis* como covariables y *clinica* como estrato, dándole el nombre *mod()*.

```
mod<-coxph(Surv(tiempo,censura)~prision+dosis+strata(clinica),
           data=cdat, method="breslow")
```

Para ver si el modelo es correcto, veamos si se cumple la hipótesis de riesgos proporcionales. Al hacerlo, podemos obtener uno de estos dos casos:

1. Si las curvas se cruzan $\Longrightarrow$ se rechaza la hipótesis de riesgos proporcionales y el modelo resultante es con la covariable *clínica* como estrato.

2. Si las curvas son paralelas $\Longrightarrow$ se acepta la hipótesis de riesgos proporcionales y el modelo resultante es introduciendo *clínica* como covariable.

Para ello, calculamos las funciones baseline hazard para cada tiempo, mostrando los 10 primeros términos para cada estrato:

```
bh <- basehaz(mod, centered=TRUE)
bh

##          hazard time    strata
## 1   0.000000000    2 clinica=0
## 2   0.005302894    7 clinica=0
## 3   0.010677622   17 clinica=0
## 4   0.016126156   19 clinica=0
## 5   0.016126156   28 clinica=0
## 6   0.021724919   29 clinica=0
## 7   0.027363076   30 clinica=0
## 8   0.033028253   33 clinica=0
## 9   0.038733766   35 clinica=0
```

```
## 10  0.044466952   37 clinica=0
##
##
## 148 0.000000000    2 clinica=1
## 149 0.012484863   13 clinica=1
## 150 0.025167324   26 clinica=1
## 151 0.038130669   35 clinica=1
## 152 0.051531857   41 clinica=1
## 153 0.051531857   53 clinica=1
## 154 0.051531857   72 clinica=1
## 155 0.065994531   79 clinica=1
## 156 0.065994531   86 clinica=1
## 157 0.081399362  109 clinica=1
```

Calculamos los logaritmos:

```
lnhazard<-log(bh[,1])
lntime<-log(bh[,2])
```

Veamos si son paralelos o no, para ello instalamos el paquete, que se necesita para dibujarlo:

```
install.packages("lattice")
library(lattice)
```

```
xyplot(lnhazard~lntime, group=strata,auto.key = TRUE,data=bh,
       xlab = "ln(t)",ylab = "ln(-ln(S(t))", main = "Clinica
       como estrato",type="l")
```

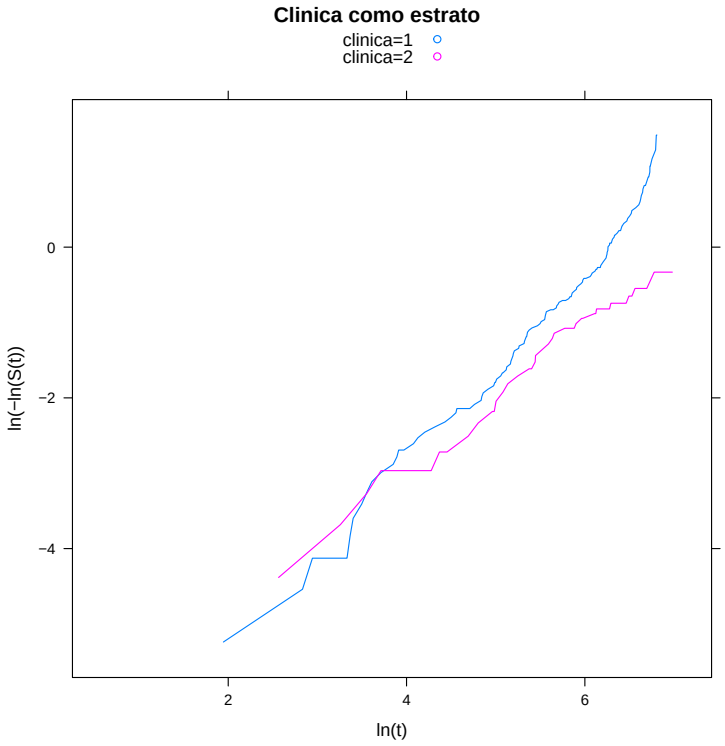

Figura 4.3: Modelo con clinica como estrato.

Como no son paralelos, podemos concluir que los errores en la clínica A y en la B no son proporcionales. Por tanto, es erróneo poner *clinica* como variable en el modelo, es mejor usarlo como estratificación.
Por tanto, nos quedamos con el siguiente modelo, en el que la *clinica* se considera un estrato.
Analizamos el modelo:

```
summary(mod2)

## Call:
## coxph(formula = Surv(tiempo, censura) ~ prision + dosis +
##       strata(clinica), data = cdat, method = "breslow")
##
##   n= 238, number of events= 150
##
##              coef exp(coef)  se(coef)      z Pr(>|z|)
## prision  0.388788  1.475192  0.168915  2.302   0.0214 *
## dosis   -0.035145  0.965465  0.006465 -5.436 5.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##         exp(coef) exp(-coef) lower .95 upper .95
## prision    1.4752     0.6779    1.0594    2.0541
## dosis      0.9655     1.0358    0.9533    0.9778
##
## Concordance= 0.651  (se = 0.034 )
## Rsquare= 0.133   (max possible= 0.994 )
## Likelihood ratio test= 33.94  on 2 df,   p=4.268e-08
## Wald test            = 32.69  on 2 df,   p=7.968e-08
## Score (logrank) test = 33.36  on 2 df,   p=5.693e-08
```

Este resumen presenta información acerca de las pruebas individuales para verificar que cada coeficiente, es significativamente distinto de cero. Las columnas proporcionan información para cada covariable como sigue:

- **coef**: Valor del coeficiente de regresión estimado.
- **exp(coef)**: Función exponencial evaluada en el coeficiente de regresión estimado. Indica el cambio en la función de riesgo por cada unidad que se incremente la covariable asociada.
- **se(coef)**: Error estándar del coeficiente de regresión estimado.
- **z**: Corresponde al valor del estadístico, obtenido dividiendo el valor del coeficiente de regresión estimado entre el error estándar estimado.
- **p**: p-valor proveniente de una distribución normal con media cero y varianza uno.

Además, se obtiene información correspondiente al siguiente contraste de hipótesis:

$$\begin{cases} H_0 & : & \underline{\beta} & = & \underline{\beta}_0 \\ H_1 & : & \underline{\beta} & \neq & \underline{\beta}_0 \end{cases}$$

- Test de razón de verosimilitud

  El estadístico del contraste, denotado por $\mathcal{X}^2_{LR}$, es 33.94 y con una distribución $\mathcal{X}^2$ con 2 grados de libertad, tiene un p-valor $p = 4.268 \cdot 10^{-08}$, lo que significa que se rechaza la hipótesis nula por ser menor que 0.05 (nivel de significación).

- Test de Wald

  El test de Wald, denotado anteriormente por $\mathcal{X}^2_W$, es 32.69 y con una distribución $\mathcal{X}^2$ con 2 grados de libertad, tiene un p-valor $p = 7.968 \cdot 10^{-08}$. De igual modo, se rechaza la hipótesis nula por la misma razón.

- Test de score

  La prueba de score, denotada anteriormente por $\mathcal{X}^2_{SC}$, es 33.36 y con una distribución $\mathcal{X}^2$ con 2 grados de libertad, se obtiene un p-valor $p = 5.693 \cdot 10^{-08}$. Se tienen las mismas conclusiones que en los casos anteriores.

Para las tres pruebas anteriores se aprecia un p-valor significativamente pequeño (inferior a 0.05), lo cual es evidencia de que, bajo las pruebas realizadas, los coeficientes del modelo son significativamente distintos de cero, y por tanto, se podría considerar que el modelo tiene sentido con las variables explicativas consideradas.
Con las salidas obtenidas, se puede verificar si son significativos, o no, cada uno de los coeficientes estimados para las covariables. Específicamente, para cada covariable, realizamos el contraste

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

con el siguiente estadístico, cuya distribución asintótica es $\mathcal{N}(0,1)$

$$z_j = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \ .$$

La región crítica de este contraste, a nivel α, es

$$R.C. : \ |z_j| \geq z_{1-\alpha/2}$$

donde $z_{1-\alpha/2}$ es el cuantil de orden $1-\alpha/2$ de la distribución $\mathcal{N}(0,1)$. Tomaremos como nivel de significación $\alpha = 0.05$, con lo cual rechazaremos H_0, y consideraremos que el correspondiente coeficiente es significativo, si $|z_j| \geq z_{0.975} = 1.96$.

- En nuestro caso tenemos que para la covariable *prision* se tiene:

  $$prision \Rightarrow |z| = |2.302| > 1.96 \rightarrow p - valor = 0.0214 < 0.05$$

  entonces la variable *prision* es significativa.
  A la otra variable *dosis* le ocurre lo mismo:

  $$dosis \Rightarrow |z| = |-5.436| > 1.96 \rightarrow p - valor = 5.44e-08 < 0.05$$

  entonces la variable *dosis* es significativa.

- Mirando la columna de $exp(coef)$ y sabiendo por teoría que:

  $$exp((\underline{X}^* - \underline{X})'\underline{\beta}) = \frac{h(t, \underline{X}^*)}{h(t, \underline{X})}$$

Si mantenemos constante la covariable *dosis*:

$$exp(\beta_{prision}) = 1.4752 = \frac{h(t, \underline{X}^*)}{h(t, \underline{X})}$$

lo que implica que cuando en $\underline{X}^*$, la *prision* aumenta una unidad, entonces el riesgo relativo en $\underline{X}$ es 1.4752. Es decir, el riesgo aumenta.

Si mantenemos constante la covariable *prision*:

$$exp(\beta_{dosis}) = 0.9655 = \frac{h(t, \underline{X}^*)}{h(t, \underline{X})}$$

lo que implica que al aumentar la dosis de metadona en una unidad, el riesgo disminuye, específicamente esto ocurre de la forma:
$h(t; x + 1) = 0.9655\ h(t; x)$.

- El coeficiente de determinación es $R^2 = 0.13$ nos indica que, aproximadamente el 13 % de la variación de los tiempos que están en tratamiento t de individuos adictos a la heroína, quedan estadísticamente explicados por las dos covariables mencionadas.

Finalmente, el modelo estimado queda de la siguiente manera:

$$\hat{h}(\underline{X}, t) = \hat{h}_0(t) \cdot exp\left\{0.388788(prision) - 0.035145(dosis)\right\} \Longrightarrow$$
$$\Longrightarrow \frac{\hat{h}(\underline{X}, t)}{\hat{h}_0(t)} = exp\left\{0.388788(prision) - 0.035145(dosis)\right\}$$

Verificación del modelo a través de los residuos

Las pruebas y los diagnósticos gráficos para el modelo de riesgos proporcionales pueden basarse en los residuos vistos en el Capítulo 3.

Con mayor comodidad, la función *cox.zph* calcula la prueba de riesgos proporcionales para cada covariable.

Vamos a probar el supuesto de riesgos proporcionales de nuestro ajuste del modelo de regresión de CPH con el siguiente contraste de hipótesis:

$$\begin{cases} H_0 & : \quad \text{proporcionalidad del modelo de Cox} \\ H_1 & : \text{no proporcionalidad del modelo de Cox} \end{cases}$$

```
cox.zph(mod2)
```

```
##             rho  chisq     p
## prision -0.0209 0.0652 0.798
## dosis    0.0849 0.9692 0.325
## GLOBAL       NA 0.9944 0.608
```

Por lo que, no existen evidencias significativas al 5 % de que se viole el supuesto de riesgos proporcionales para ninguna de las dos covariables ni globalmente.

Residuos de Cox-Snell

Después de ajustar el modelo, es útil estudiar los residuos de Cox-Snell con el fin de evaluar el ajuste del modelo de riesgos proporcionales.
Si el modelo es correcto y la estimación de los β's son cercanas a los valores reales, entonces estos residuos deberían comportarse como una muestra censurada de observaciones de una distribución exponencial.

Hemos calculado el estimador de Nelson-Aalen de la tasa de riesgo acumulado de los residuos de Cox-Snell. Si una distribución exponencial se ajusta a los datos, entonces, este estimador debería describir aproximadamente una línea de pendiente igual a 1.

```
estado<-cdat$censura
mresi<-residuals(mod2,type="martingale")
csresi<-censura-mresi
hazard.csresi<-survfit(Surv(csresi,censura)~strata(clinica),
                       type="fleming-harrington")
plot(hazard.csresi$time,-log(hazard.csresi$surv),
     xlab='residuos de Cox-Snell', ylab='riesgo acumulado',
     lty = 1:4,main="Representación de los residuos de Cox-Snell")
lines(c(0,5),c(0,5))
```

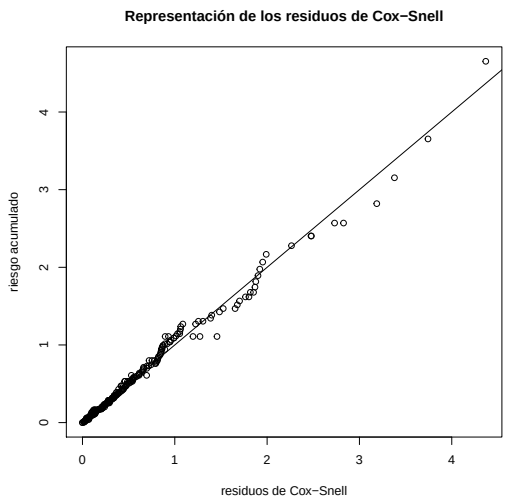

Figura 4.4: Salida de los residuos de Cox-Snell.

La gráfica anterior nos sugiere que este modelo se ajusta bien a los datos.

Residuos de martingala

Los residuos tipo martingala para la covariable prision pueden generarse a través de la sentencia:

```
mres1<-residuals(mod2, type=c("martingale"))
plot(cdat[,1], mres1, xlab=c("prision")[1],
     ylab="Residuos martingale", main="Residuos de Martingala")
abline(h=0, lty=2)
lines(lowess(cdat[,1], mres1, iter=0))
```

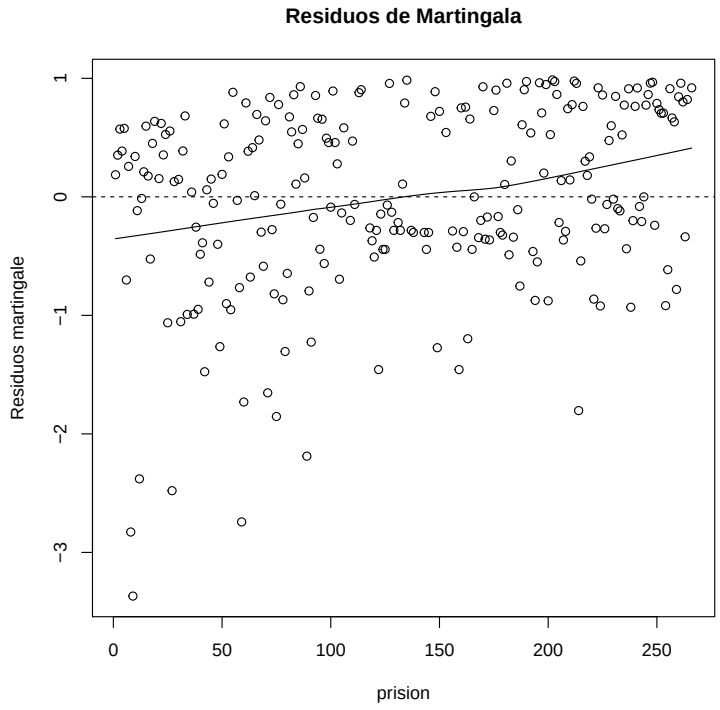

Figura 4.5: Salida de los residuos de martingala para prision.

Podemos ver claramente, una tendencia curva creciente. Estos residuos presentan una forma funcional definida.

Residuos basados en el estadístico Deviance

Los residuos tipo deviance sirven para comprobar la existencia de outliers. Veamoslo

```
devresi <- resid(mod2, type="deviance")
plot(mod2$linear.predictor, devresi, ylab="Residuos de Deviance",
     main='Residuos de deviance')
abline(h=0,lty=2, col='black')
```

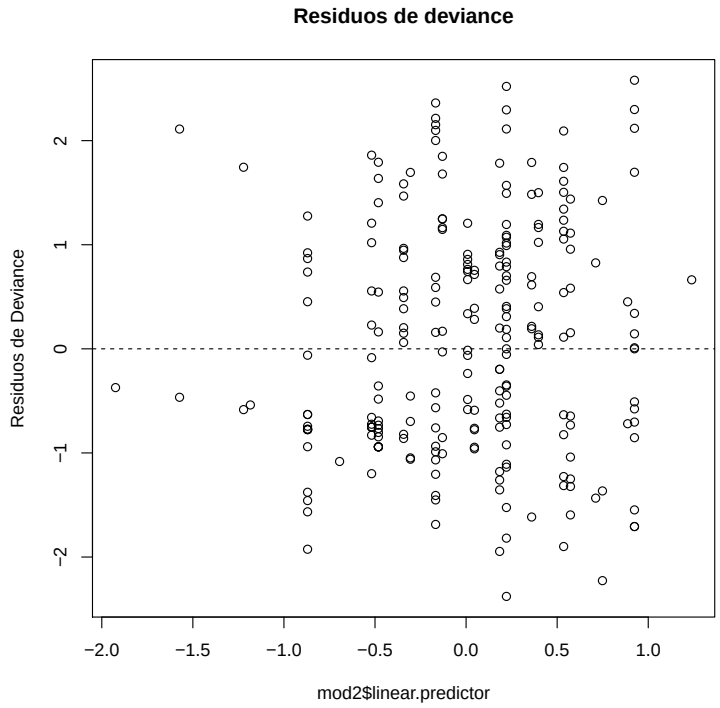

Figura 4.6: Salida de los residuos del estadístico Deviance.

Como se comprueba, no apreciamos patrones definidos pero si valores alejados del origen.

Residuos escalados de Schoenfeld

Ahora estamos interesados en evaluar la hipótesis de riesgos proporcionales del modelo de CPH, examinando si el impacto de una o más covariables sobre el riesgo de los adictos a la heroína puede variar con el tiempo.
Calculamos los residuos escalados de Schoenfeld para nuestro caso de la forma:

```
par(mfrow=c(1,2))
plot(cox.zph(mod2))
```

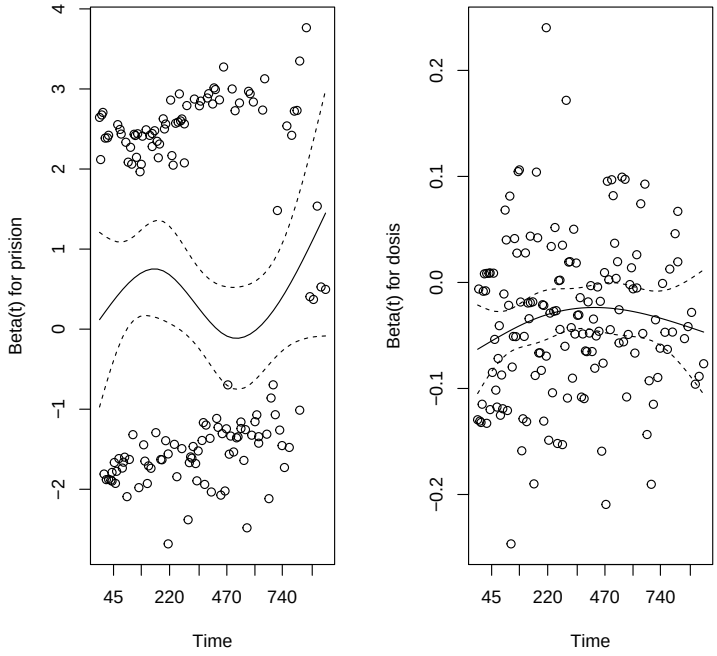

Figura 4.7: Salida de los residuos escalados de Schoenfeld.

Las tendencias en los diagramas de dispersión de los residuos escalados de Schoenfeld a menudo son difíciles de determinar, especialmente con las covariables binarias (como *prision* en nuestro caso) donde sólo hay dos bandas horizontales de residuos presentes.

Residuos dfbeta

```
dfbeta <- residuals(mod2, type="dfbetas")
par(mfrow=c(1,2))
for (j in 1:2){
   plot(dfbeta[,j],
   ylab=names(coef(mod2))[j])
   abline(h=0, lty=2, col='black')
   lines(c(0,0),c(0,0))
}
```

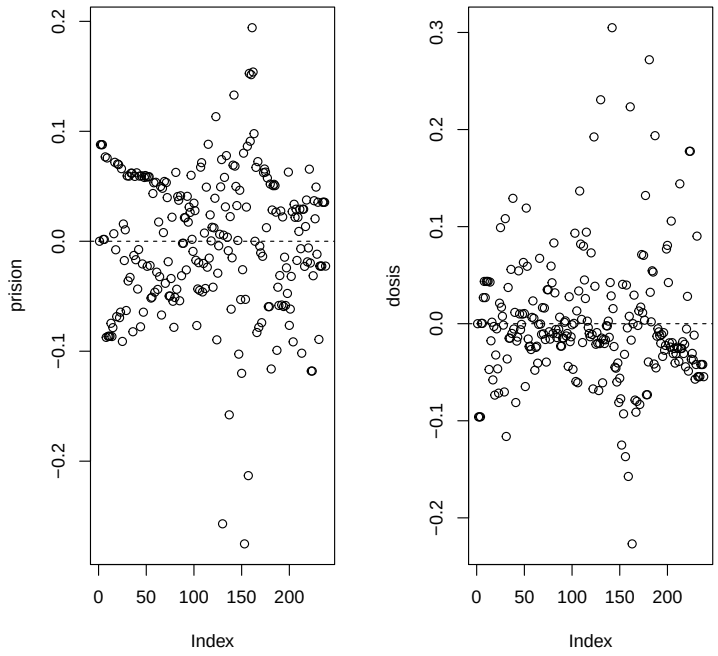

Figura 4.8: Salida de los residuos dfbeta.

En estas figuras se nos muestran los residuos *dfbeta* del modelo. Como vemos estos residuos están centrados con respecto al origen, y no presentan patrones definidos. Se nos presentan algunos datos demasiados alejados del origen en ambas figuras.

4.1. Conclusiones al estudio de los datos de *metadona*

Los datos analizados, corresponden a programas de rehabilitación realizados en distintas clínicas australianas durante los años 1970. Individuos adictos a la heroína fueron sometidos a tratamientos con metadona. Trabajamos con los datos del estudio realizado por Caplehorn y Bell (1991), cuyo objetivo era verificar si existían diferencias significativas entre el tipo de tratamiento aplicado a los pacientes en las etiquetadas como clínica A y B (la política de estas clínicas fue diferente) . Se consideran como covariables de interés en el estudio: dosis (dosis diaria de metadona administrada durante el tratamiento), y prisión (si el individuo ha estado antes en prisión o no).

Aplicando la metodología propuesta en el trabajo hemos obtenido los siguientes resultados:

1. Se han estimado las funciones de supervivencia, $S(t)$, y de hazard acumulada, $H(t)$, en cada clínica.

2. Hemos enconttrado que existen diferencias significativas entre las funciones de supervivencia en las dos clínicas, aplicando el test de log-rank.

3. Se ha discutido por qué es adecuado utilizar un modelo de Cox estratificado en este caso.

4. El modelo estratificado de Cox se ha aplicado considerando como estratos las clínicas A y B. Esto nos ha permitio estimar las funciones basales para cada clínica, y estimar el efecto de las covariables (dosis y prisión) en el tiempo de permanencia en tratamiento. Se incluye el análisis de los residuos, para ilustrar cómo se calculan estos en el modelo de Cox.

Bibliografía

[1] Boj del Val,Eva. "El modelo de regresión de Cox". Departamento de Matemática Económica, Financiera y Actuarial de Mayo de 2017 (Universidad de Barcelona).

[2] Borges P.,Rafael Eduardo. "Análisis de supervivencia de pacientes con diálisis peritoneal". Revista Colombiana de Estadística de Diciembre de 2005 (Volumen 28 No 2. pp. 243 a 259).

[3] Caplehorn, J.R.M. and Bell. "Methadone dosage and retention of clients in maintenance treatment". Med.J.Aust. 154, pp. 195-199. (1991).

[4] Cardona Hurtado,Diego Alejandro and Trujillo Bonilla,Jenny Carolina. "Aspectos básicos de estimación no paramétrica en análisis de sobrevivencia. Una aplicación a un estudio de deserción estudiantil". Trabajo para optar al título de Profesional en Matemáticas con Énfasis en Estadística de 2013 (Universidad Del Tolima) Ibagué, Colombia.

[5] Caroni,Chrys. "Lifetime data analysis. Reliability and survival analysis". (2002).

[6] Caroni,Chrys. "The Correct "Ball Bearings" Data". Lifetime Data Analysis, 8, 395-399. Kluwer Academic Publishers. (2002).

[7] García-Hinojosa,Cristina Pruenza. "Estudio de análisis de supervivencia". Trabajo Fin de Grado de Mayo 2014 (Universidad Autónoma de Madrid).

[8] Grambsch, P. and Therneau, T.M.. "Proportional hazards tests and diagnostics based on weighted residuals". Biometrika. 81,515-26. (1994).

[9] Broström, Göran. "eha: Event History Analysis". R package version 2.5.1.(2017).

[10] Hernández Domínguez,Ana María. "Análisis estadístico de datos de tiempos de fallo en R". Máster Universitario en Estadística Aplicada de 2010 (Universidad de Granada).

[11] Hess, K.R.. "Graphical Methods for assessing violations of the proportional hazards assumption in Cox regression". Statistics in Medicine, 14, 1707-1723. (1995).

[12] Jiménez,Pablo Moreno. "Inferencia estadística para datos censurados. Métodos y aplicaciones". Trabajo Fin de Grado de Junio de 2014 (Universidad de Sevilla).

[13] Kalbfleisch, John D. and Prentice, Ross L. "The Statistical Analysis of Failure Time Data". Second Edition. Wiley & Sons. (2002).

[14] Kassambara, Alboukadel and Kosinski, Marcin. "survminer: Drawing Survival Curves using 'ggplot2' ". (2018).

[15] Lawless, J.F. "Statistical Models and Methods for Lifetime Data". Second Edition. Wiley & Sons. (2003).

[16] López Montoya,Antonio Jesús. "Comparación de dos modelos de Regresión en fiabilidad". Máster Universitario en Estadística Aplicada de 2011 (Universidad de Granada).

[17] Martinez,Javier. "Análisis de Supervivencia en R". 22 de mayo de 2017.

[18] Quintanilla Casas,Beatriz. "Estadística en variables con censura: Aplicación a datos medioambientales". Máster en Bioinformática y Bioestadística de Junio de 2017 (Universitat Oberta de Catalunya).

[19] R Core Team, "R: A Language and Environment for Statistical Computing". R Foundation for Statistical Computing. Vienna, Austria. (2018).

[20] Roque Roque,Daniel Octavio, "Forma funcional de covariables en el modelo de Cox". Tesis de 2009 (Universidad Nacional Mayor de San Marcos) Lima-Perú.

[21] Sarkar, Deepayan. "lattice: Multivariate Data Visualization with R". Springer. New York. (2008).

[22] Tang, Yuan; Horikoshi, Masaaki and Li, Wenxuan "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages". (2016).

[23] Therneau, Terry M. and Grambsch, Patricia M.. "survival: Modeling Survival Data: Extending the Cox Model". Springer, New York.(2000).

[24] Wickham, Hadley ; François, Romain ; Henry, Lionel and Müller, Kirill. "dplyr: A Grammar of Data Manipulation". (2018).