Universidad de Granada
Departamento de Estadística e I.O.

ANÁLISIS ESTADÍSTICO DE DATOS DE TIEMPOS DE FALLO EN R

Ana María Hernández Domínguez

Tutores: M.Luz Gámiz Pérez y M.Dolores Martínez Miranda

Granada, 2010

ÍNDICE GENERAL

CONTENIDOS DEL TRABAJO

El contenido de este trabajo está dividido claramente en tres partes.

El Capítulo I, *Introducción al Análisis de Fiabilidad*, es sin lugar a dudas el bloque más importante de este trabajo. En él se exponen los modelos teóricos más importantes del Análisis de Fiabilidad como son: la distribución Exponencial o la Weibull para los modelos paramétricos; el Estimador de Kaplan-Meier para los no paramétricos; o el Modelo de Cox dentro de los semiparamétricos.

Como la característica principal en Fiabilidad es que la observación completa de la variables no es posible, veremos las causas principales de este problema que dan lugar a los distintos tipos de datos en este análisis.

El Capítulo II, *Análisis computacional de datos en tiempos de fallo*, es la parte más técnica del trabajo. Tras estudiar en el primer bloque los modelos teóricos más importantes en Fiabilidad, en este bloque vemos como esas funciones podrían ser calculadas de un modo computacional, y para ello, nosotros veremos algunas de las funciones para este tipo de análisis del lenguaje *R*.

El entorno de análisis estadístico *R* tiene varias librerías que permiten el cálculo del análisis de supervivencia, pero nosotros nos referiremos exclusivamente a la librería *Survival*. Así que explicaremos la sintaxis y sus argumentos para las funciones más importantes de esta librería.

El Capítulo III, es una aplicación real donde se puede ver como son empleados algunos de los modelos que hemos visto en el primer bloque y su tratamiento computacional usando algunas de las funciones del bloque II.

Para este estudio, se ha tomado una red de suministro de agua de una ciudad española de la que se pretende obtener información y registro del verdadero estado de deterioro de la red y su fiabilidad. Mediante tramos individuales de tubería se ha estudiado el fallo teniendo en cuenta diversas características físicas, y la fuerte censura y truncamiento que presentan los datos. Por este motivo, veremos algunos de sus factores y su impacto sobre el riesgo de fallo, y su efecto sobre la razón de fallo mediante el Modelo de Riesgos Proporcionales de Cox (RPC).

Capítulo I

Introducción al Análisis de Fiabilidad

El Análisis de Fiabilidad o Análisis de Supervivencia es un conjunto de técnicas que se emplean para analizar los datos en los que la variable de interés es el tiempo que transcurre desde un instante inicial bien definido, hasta la ocurrencia de un determinado suceso o instante final. Es decir, obtener información sobre la variable tiempo de vida, tiempo hasta la muerte o curación, probabilidad de fallo en cada instante, riesgo de fallo, etc.

Cuando hablamos de tiempo de vida nos estamos refiriendo a la longitud de tiempo hasta la ocurrencia de un suceso de interés (que suele ser el fallo de una pieza, o la muerte o recaída de un paciente) desde un punto prefijado. En otras ocasiones, el término tiempo de vida se usa en sentido figurado. Matemáticamente, el tiempo de vida es una variable aleatoria no negativa. Klein y Moeschberger (1997), Andersen, Borgan, Gill y Keiding (1993), Cox y Oakes (1984), Lawless (1982).

En el tratamiento del conjunto de datos es de importancia la utilización de modelos específicos paramétricos, modelos no paramétricos y semi-paramétricos que describen bien el comportamiento del conjunto de datos.

Una de las características más importantes en el análisis de los datos en un estudio de fiabilidad de un dispositivo es la presencia de censura, es decir, cuando el suceso de interés no ha sido observado en ese sujeto. Esta peculiaridad, entre otras, de los datos en

análisis de fiabilidad hace que los métodos clásicos de estimación tengan que ser adaptados como veremos.

A continuación, estudiaremos los modelos usados a menudo en estudios de Fiabilidad y posteriormente los aplicaremos en un caso práctico.

1. Variable aleatoria Tiempo de vida

Llamaremos *T* al tiempo hasta la ocurrencia de un suceso de interés (generalmente el fallo de un mecanismo). Nos referimos a este tiempo como tiempo de fallo, tiempo de vida o tiempo de supervivencia. Consideraremos *T* como una variable continua. Además, es una variable no negativa, $T \geq 0$.

1.1. Función de Fiabilidad o de Supervivencia

Se define como la probabilidad de que el dispositivo sobreviva más allá del instante *t.*

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

donde *F(t)* es su función de distribución, que asumimos es una función absolutamente continua.

Esta función cuantifica la capacidad que tiene un dispositivo para cumplir con éxito su función, en un intervalo de tiempo *(0,t]*.

Sus propiedades son:

- Monótona, decreciente y continua.
- $S(0) = 1$ y $\lim_{t \to \infty} S(t) = 0$
- $S(t) = P[T > t] = \int_t^{\infty} f(x)dx \qquad 0 \leq t \leq \infty$ donde *f* es la función de densidad asociada a *T.*

1.2. Función de Riesgo

La *función de riesgo o tasa de fallo*, denotada por *r(t)*, se define como el cociente entre la función de densidad y la función de supervivencia. Esta función tiene un interés particular en aplicaciones de Supervivencia ya que indica la disposición inmediata al fallo en un intervalo de tiempo pequeño, como una función de la edad. Su expresión es:

$$r(t) = \frac{f(t)}{S(t)}$$

Desde esta expresión podemos dar una interpretación del significado físico de la función de riesgo. Supongamos un individuo cuyo tiempo de vida viene dado por la variable aleatoria T. Sea Δt pequeño, entonces

$$r(t)\cdot\Delta t = \frac{f(t)\Delta t}{S(t)} \approx \frac{P(\text{aparezca un fallo en } (t, t+\Delta t))}{P(T>t)} = P(t < T \le t+\Delta t \mid T > t)$$

Así, $r(t)\cdot\Delta t$ representa la probabilidad de fallo durante $(t,\ t+\Delta t]$, dado que el sistema está funcionando en t.

En algunos casos, se interpreta también como la velocidad de degradación del dispositivo, o la intensidad con que se presentan los fallos en el dispositivo, justo en el instante t. Por eso la estimación de la función de riesgo se basa esencialmente en consideraciones físicas.

Cuando el fenómeno físico que se está estudiando es la evolución de un individuo, se puede estimar esta función considerando una serie de individuos en condiciones similares, de modo que si se quiere modelizar matemáticamente el tiempo de vida, una estimación natural de la función de riesgo es el número de recaídas por unidad de tiempo ocurridos en cada intervalo de observación. Siguiendo esta línea, y con el fin de elegir un modelo adecuado vía la función de riesgo, conviene tener en cuenta la existencia de tres tipos de "fallos" (sucesos en general) que presentan características esencialmente temporales.

El primer tipo, llamado *fallo inicial,* se manifiesta al principio de la vida del individuo y va desapareciendo conforme se desarrolla el periodo inicial. Un ejemplo de este tipo puede observarse en las tablas de mortalidad humana, en las cuales se supone que en la vida de un individuo hay presentes, al principio de la misma, ciertos problemas de carácter hereditario que pueden provocar desenlaces fatales y que van desapareciendo conforme el individuo crece, y aproximadamente están presentes hasta una edad de 10 años.

El segundo tipo de fallo se llama *fallo accidental*, y ocurre durante el periodo en el que el individuo presenta una función de riesgo constante, generalmente menor que la que prevalece durante su periodo inicial. En las tablas de mortalidad humana, se supone que las muertes ocurridas entre los 10 y 30 años son debidas a accidentes.

El tercer tipo, llamado *fallo de desgaste*, se asocia con un deterioro gradual del individuo. En las tablas de mortalidad humana hay, a partir de los 30 años una proporción creciente de muertes atribuidas al envejecimiento del individuo.

Si en la evolución de un individuo únicamente estuviesen presentes estos tres tipos de fallo, el modelo seleccionado para representar matemáticamente su tiempo de funcionamiento debe tener una función de riesgo que por su forma es conocida con el nombre de Curva de Bañera, y se asemeja a la siguiente figura:

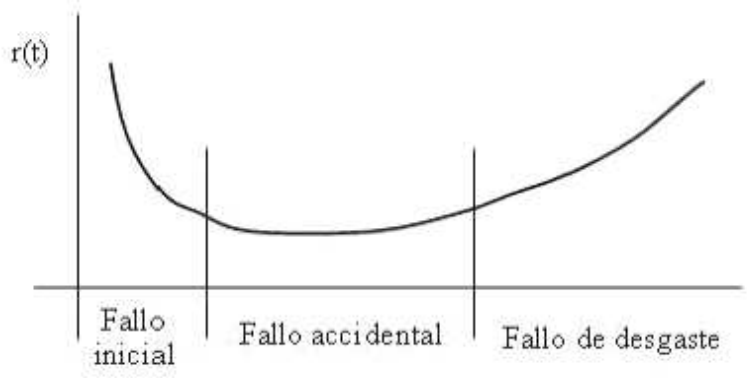

Figura 1.1. Curva de Bañera

1.3. Función de Riesgo Acumulada

La *Función de Riesgo Acumulada, R(t),* se define como:

$$R(t) = \int_0^t r(u)du = -\log S(t)$$

Esta función es importante en la medición de la frecuencia con que ocurren los fallos en el tiempo, en la construcción de papeles probabilísticos y en el análisis de residuos en el ajuste de algunos modelos.

Todas estas funciones *f(t), F(t), S(t), r(t)* y *R(t)* sirven para caracterizar la distribución de probabilidad de *T* y, además, son equivalentes entre sí, con lo que conocida una de ellas, las restantes son fáciles de obtener a través de alguna de las siguientes relaciones:

$$f(t) = -\frac{d}{dt}S(t)$$

$$r(t) = -\frac{d}{dt}\ln S(t)$$

$$S(t) = \exp\left\{\int_0^t r(u)du\right\} = e^{-R(t)}$$

$$R(t) = -\ln S(t)$$

2. Algunos modelo aleatorios de fiabilidad

Existen numerosos modelos paramétricos que son usados en el análisis de tiempo 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 dado su demostrada utilidad en casos prácticos. Así se tiene, entre otras, la exponencial, gamma, la Weibull, la normal y log-normal y la log-logística.

El método consiste en estimar, por métodos robustos (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.

A continuación, presentamos algunas de las distribuciones referidas anteriormente.

2.1. Distribución Exponencial

La distribución exponencial tiene un papel fundamental en el Análisis de Fiabilidad ya que se trata de la distribución más empleada en el análisis de datos de tiempo de fallo.
Se suele utilizar para modelizar el tiempo transcurrido entre dos sucesos aleatorios 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 cual la tasa de fallo es aproximadamente constante, $r(t) = \lambda$. 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:

$$f(t) = \lambda e^{-\lambda t} \qquad t \geq 0$$

donde λ es la tasa de fallo, y es una constante positiva.

La función de distribución es:

$$F(t) = 1 - e^{-\lambda t} \qquad t \geq 0$$

La función de fiabilidad queda como:

$$S(t) = 1 - F(t) = e^{-\lambda t} \qquad t \geq 0$$

La función de riesgo es:

$$r(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \qquad \text{t} \geq 0$$

<u>Propiedad de no memoria</u>

La distribución exponencial se caracteriza por la *propiedad de no memoria*, y se enuncia como sigue:

Sea X una variable aleatoria con distribución exponencial entonces

$$\text{P}\{\text{X>x+t|X>t}\} = P\{X > t\}$$

para todo $x, y \geq 0$.

Es decir, si interpretamos X como el tiempo de vida de un individuo, esta igualdad establece que la probabilidad de que el individuo sobreviva al tiempo $x+t$ dado que ha sobrevivido al tiempo x, es igual a la probabilidad de que sobreviva al tiempo t al principio de su vida (o cuando está nuevo para el caso de un objeto).
La condición anterior puede también escribirse en términos de la función de supervivencia

$$S(x+t) = S(x)S(t)$$

Esta propiedad describe el proceso de vida sin envejecimiento, y caracteriza a la función exponencial.

2.2. Distribución Gamma

Una generalización de la distribución exponencial es la *distribución Gamma*. Una variable aleatoria de tiempo de vida se dice que se distribuye según una Gamma de parámetros α y λ, $G(\alpha,\lambda)$, cuando su función de densidad viene dada por

$$f(t) = \frac{\lambda^{\alpha} t^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda t} \qquad \text{t} \geq 0, \quad \alpha, \lambda > 0$$

donde α es el parámetro de forma, λ es el parámetro de escala y Γ representa la función gamma que se define

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad , \quad \text{para todo} \quad \alpha > 0$$

y verifica que $\Gamma(n)=(n-1)!$, para todo $n \in \mathrm{N}$. Cuando $\alpha=1$, se tiene la distribución exponencial.

La función de riesgo es:

$$r(t) = \frac{t^{\alpha-1} e^{-\lambda t}}{\int_0^{\infty} x^{\alpha-1} e^{-x} dx}$$

es una función decreciente en t si $\alpha<1$, es constante si $\alpha=1$ y es creciente en t si $\alpha>1$.

Además $r(t)$ se aproxima asintóticamente a $1/\lambda$ cuando $t \to \infty$, lo cual sugiere que la distribución Gamma puede ser útil como un modelo de población cuando los individuos que sufren determinada enfermedad son sometidos a un programa de seguimiento regular. La razón de fallo puede crecer o decrecer algo inicialmente, pero después de algún tiempo la enfermedad tiende a estabilizarse y a partir de ahí la recaída es tan probable en un intervalo de tiempo como en otro de la misma amplitud.

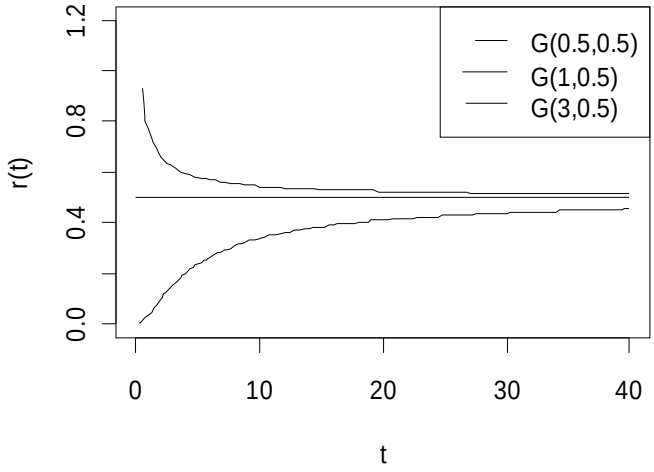

Figura 1.2. Funciones de riesgo de distintas distribuciones Gamma

2.3. 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 de 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.

Definiremos la distribución de Weibull a partir de su función razón de fallo, *r(t)*. Sea T una variable aleatoria tiempo de vida tal que la correspondiente razón de fallo viene dada por:

$$r(t) = \lambda\beta(\lambda t)^{\beta-1}$$

con $\beta > 0$, $\lambda > 0$, $t > 0$; β es el parámetro de forma y λ el de escala. Notamos a la distribución como $W(\lambda, \beta)$. Cuando $\beta = 2$, es conocida como distribución de Rayleigh.

Gráficamente, representamos las funciones de riesgo de distribuciones para W(1, 1.5), W(1,1) y W(1,0.5).

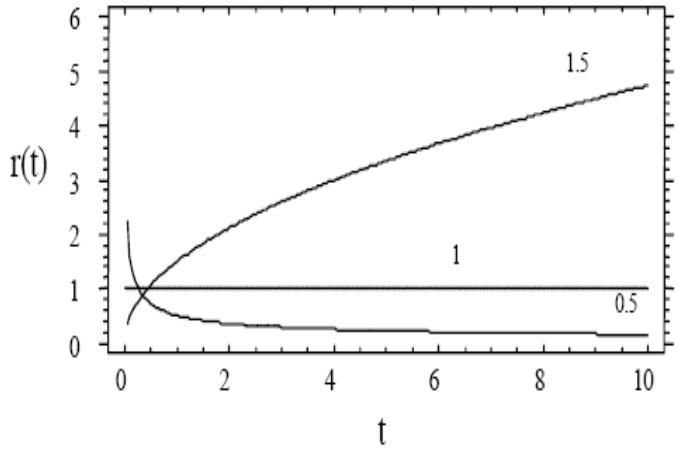

Figura 1.3. Funciones de riesgo de distribuciones W(1, 1.5), W(1,1) y W(1, 0.5)

Propiedades

Dada la importancia de la distribución Weibull ya que es una familia más amplia que contiene como caso particular a la Exponencial y está formada por aquellas distribuciones con función de riesgo de tipo potencial, vamos a ver algunas de sus propiedades más importantes:

Propiedad 1: Si $T_1, T_2, ..., T_n$ son n variables aleatorias independientes e idénticamente distribuidas con distribución común Weibull, entonces $T_{(1)} = \min\{T_1, T_2, ..., T_n\}$ es también Weibull. La función de supervivencia es

$$S_{T_{(1)}}(t) = \left(S(t)\right)^n = \exp\left\{-n\left(\lambda t\right)^\beta\right\}$$

y los parámetros pueden identificarse a partir de aquí fácilmente.

Propiedad 2: Si X tiene distribución $exp(\lambda)$, entonces $T=X^p$ tiene distribución Weibull, $W(\lambda^p, 1/p)$.

De donde se deduce que la distribución exponencial es un caso particular de la distribución Weibull para $\beta=1$.

Propiedad 3: La función de riesgo es creciente para $\beta>1$ y $r(t)\to\infty$ cuando $t\to\infty$. Es decreciente para $\beta<1$ y tiende asintóticamente a 0 cuando $t\to\infty$.

Propiedad 4: Si $(\lambda t)\beta << 1$, una buena aproximación de F(t) es $(\lambda t)\beta$.

Propiedad 5: Si $\beta \to \infty$, se tiene un tiempo de vida constante pues $S(t) \to 0$.

Propiedad 6: El i-ésimo momento de la distribución es igual a

$$E\left[T^i\right] = \Gamma\left(1 + i/\beta\right)\lambda^{-i}$$

siendo $\Gamma(\cdot)$ la función Gamma.

2.4. Distribución Normal

Una variable aleatoria T se dice que tiene distribución *Normal*, N(μ, σ), si la función de densidad es

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/2\sigma^2} \qquad t \in R,\ \ \mu \in R,\ \ \sigma > 0$$

Si llamamos $\Phi(t)$ a la función de distribución de la distribución N(0,1), entonces

$$F(t) = P\{T \le t\} = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

se puede comprobar que la función de distribución de la N(μ, σ) es
y, por tanto, la función de riesgo viene dada por

$$r(t) = \frac{1}{\sigma} \cdot \frac{\phi\left((t-\mu)/\sigma\right)}{1 - \Phi\left((t-\mu)/\sigma\right)}$$

Si $r_\phi(t)$ representa la función de riesgo de la normal estándar, se tiene la siguiente igualdad

$$r(t) = \frac{1}{\sigma} r_\phi\left(\frac{t-\mu}{\sigma}\right)$$

El siguiente gráfico muestra la función de riesgo de la N(0,1). Es una función creciente para todo t y se aproxima asintóticamente a $r(t) = t$.

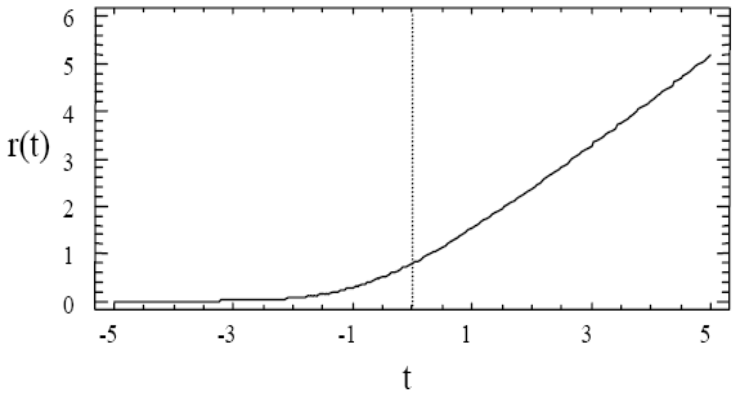

Figura 1.4. Razón de fallo de la distribución N(0,1)

Cuando una variable aleatoria tiene distribución normal pero con una cota superior y/o inferior para los valores de la variable, la distribución resultante se llama *distribución normal truncada*. Cuando sólo hay una cota inferior (superior), la distribución se dice truncada a la izquierda (derecha), si existen las dos cotas se dice doblemente truncada.

La distribución normal truncada en 0 es a menudo usada como distribución de tiempo de vida.

2.5. Distribución Lognormal

La distribución normal, sin duda la más importante de las distribuciones estadísticas, 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. Aunque como hemos visto anteriormente, este problema se podría solucionar truncado la distribución Normal, otra forma de solventar esta dificultad, es recurrir a la distribución log-normal, derivada de la normal, que sólo considera valores positivos.

Se dice que una variable aleatoria T tiene un comportamiento lognormal, de parámetros μ y σ si su logaritmo es una variable aleatoria con distribución Normal, es decir, si

$$\ln(T) \rightarrow N(\mu, \sigma)$$

Si T sigue una distribución lognormal se representa por $T \sim LN(\mu, \sigma)$, donde μ y σ son los parámetros de localización y dispersión de la distribución de $ln(T)$.

Sus funciones de densidad y de distribución vienen dadas respectivamente por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-(\ln t - \mu)^2 / 2\sigma^2} \qquad t \geq 0$$

$$F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right) \qquad t \geq 0$$

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 fiabilidad $S(t)$ y la función de riesgo $r(t)$ dependen también de $\Phi(z)$, como se puede ver en la siguiente expresión

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

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

El modelo lognormal es muy fácil de utilizar si no hay censura (este concepto se define en secciones posteriores), pero con censura los cálculos y operaciones se complican.

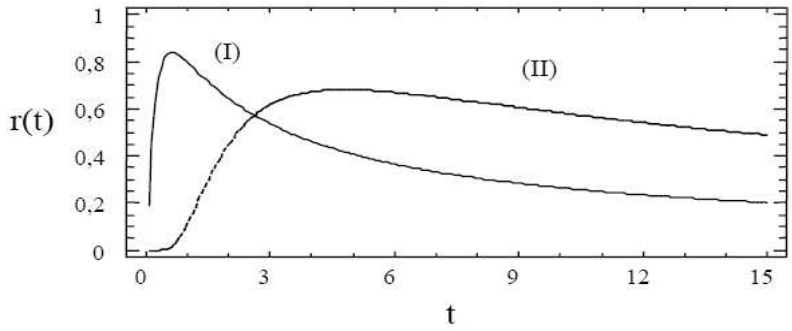

Figura 1.5. Razones de fallo de distribuciones lognormales
μ=0 y σ=1; y (II) μ=1 y σ=0.5

La distribución lognormal se ajusta a ciertos tipos de fallos (fatiga de componentes metálicos, vida de los aislamientos eléctricos), procesos continuos (procesos técnicos) y puede ser una buena representación de la distribución de los

tiempos de reparación.

La distribución lognormal es importante en la representación de fenómenos de efectos proporcionales, tales como aquellos en los que un cambio en la variable en cualquier punto de un proceso es una proporción aleatoria del valor previo de la variable.

2.6. Distribución Log-Logística

La importancia de la distribución log-logística, al igual que la lognormal, está en que proporcionan modelos para tiempos de vida que no tienen función de riesgo monótona, sino que pueden (según cómo se elijan los parámetros) presentar diferentes formas. Las funciones de riesgo de estas dos distribuciones con muy parecidas pero la de la log-logística es mucho más manejable.

La función de supervivencia $S(t)$ es

$$S(t) = \frac{1}{1+e^{\lambda}t^{k}}$$

La función de riesgo es

$$r(t) = \frac{e^{\lambda}kt^{k-1}}{1+e^{\lambda}t^{k}}$$

para $t \geq 0$. El siguiente gráfico representa una función de riesgo para un modelo log-logístico, que crece inicialmente y a partir de un momento determinado cambia de sentido y decrece.

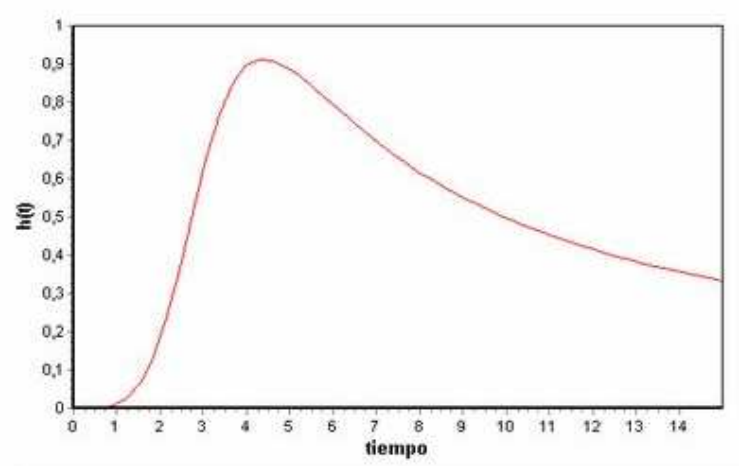

Figura 1.6. Función de riesgo de una log-logística

Algunas de las ventajas de esta función con respecto a la lognormal, es que

cuando $t \to \infty$, la función de riesgo se aproxima a cero. Además, la forma de su función de riesgo es muy parecida a la lognormal pero algebraicamente es mucho más manejable.

3. Tipos de datos en Fiabilidad

Un ensayo clásico de Fiabilidad consiste en observar a lo largo del tiempo una muestra de dispositivos o piezas en funcionamiento. La variable de interés suele ser el tiempo de fallo del dispositivo, pero existen ocasiones en las que no es posible la observación completa de esta variable porque no sea factible continuar el experimento hasta que fallen todos los dispositivos. Las dos causas principales son la censura y el truncamiento.

En la censura destacamos dos tipos: *Censura Tipo I,* en la cual los dispositivos son observados hasta un tiempo determinado; y la *Censura Tipo II*, en la que los individuos son observados hasta que ocurran un número determinado de fallos. La determinación del tiempo para el Tipo I y el número de fallos para el Tipo II deben establecerse antes de iniciar el experimento, y no durante el transcurso del mismo.

Atendiendo a las causas que dan lugar a la censura o al truncamiento, se distinguen los siguientes tipos.

Censura por la derecha

Este tipo de censura se considera dentro de la Censura Tipo I. Se presenta cuando finalizado el periodo de observación de un determinado dispositivo, aún no ha ocurrido el evento que se desea observar (generalmente el fallo del dispositivo). Es decir, puesto un dispositivo en funcionamiento y fijado un cierto valor Y (duración del seguimiento o periodo de observación), se dice que una observación T (tiempo de vida del dispositivo) es censurada a la derecha si se desconoce el valor de T, y sólo se sabe que es mayor que Y.

Censura por la izquierda

Se presenta cuando el evento que se desea observar ha ocurrido antes de que el dispositivo o individuo se incluya en el estudio. Por tanto, se desconoce el valor exacto de la observación T_i.

Este tipo de censura suele confundirse con el truncamiento por la izquierda o la entrada tardía.

Censura doble

Se presenta cuando los datos están censurados tanto por la izquierda como por la derecha.

Censura aleatoria

Se produce cuando en el transcurso de un estudio, algunas unidades experimentan otros sucesos independientes del de interés que provocan la salida del estudio.

Esto puede deberse fundamentalmente a varias razones: a que hasta el momento de la finalización del estudio no haya ocurrido el evento (si es que el periodo de seguimiento es finito), a que el individuo abandone el estudio, o en el caso de que ocurra en el individuo o dispositivo otro evento que imposibilite la ocurrencia del evento que se desea observar.

El modelo aleatorio de censura a la derecha se representa con el par *(X; δ)*, donde:

$$X = \min\{T, Y\} \qquad \text{y} \qquad \delta = 1_{\{T \le Y\}} \begin{cases} 1 & \text{si} \quad \text{T} \le \text{Y} \\ 0 & \text{si} \quad \text{T} > \text{Y} \end{cases}$$

donde δ es la función aleatoria indicadora de censura. Cada valor de la muestra será un par (x_i, δ_i), con δ_i=1 cuando el valor observado de X corresponde con el tiempo de fallo del dispositivo y δ_i=0 cuando el valor registrado de X es un tiempo de censura, es decir, no se ha observado el tiempo de fallo del dispositivo porque un suceso aleatorio e independiente del fallo del mecanismo ha ocurrido en el tiempo x_i interrumpiendo la observación de la vida del mecanismo hasta su fallo. En este caso la información muestral que se registra acerca de la variable de interés es que ésta tomaría un valor superior al tiempo registrado.

Truncamiento por la izquierda

Sucede cuando los individuos entran en el estudio a edades aleatorias, por lo que el origen del tiempo de vida precede al origen de estudio.

Para aquellos sujetos en los que el fallo tiene lugar antes del inicio del estudio serán ignorados y no entrarán a formar parte del estudio. La información que se registra se refiere por tanto no a la variable de interés tiempo de vida tal cual, sino a esta variable condicionada a que el individuo sobrevivió para entrar en el estudio.

Truncamiento por la derecha

Ocurre cuando sólo los individuos que presentan el evento o fallo son incluidos en el estudio. En este caso la información que se registra también corresponde a una

variable condicionada a que el tiempo de fallo fue anterior a la finalización del estudio. Individuos que no cumplen esta condición son ignorados por el experimento.

Modelo de truncamiento por la izquierda y censura por la derecha

Este modelo teórico se recoge en problemas que presentan datos que son truncados y censurados al mismo tiempo.

Sea (X,T,C) un vector aleatorio, donde X es el tiempo de truncamiento, T es el tiempo de fallo, y C es el tiempo de censura. Se supone que bajo un modelo de truncamiento por la izquierda y censura a la derecha, las observaciones muestrales serán observaciones del vector (X,Y,δ), donde $Y = min\{T,C\}$ y δ es el indicador de censura, es decir, $\delta = 1_{\{T \leq C\}}$ si T es un tiempo de fallo ó 0 si T es un tiempo de censura. Si $Y < X$, no hay observaciones. Es decir, que la muestra está constituida por n observaciones de tipo (x_i, y_i, δ_i) en los que $x_i \leq y_i$ para todo $i=1, 2, \ldots, n$.

Las condiciones habituales en este modelo son:

- T es independiente de (X,C)
- X,T y C son mutuamente independientes.

3.2. Ejemplos de tipos de datos

Para poder entender mejor los diferentes tipos de experimentos muestrales que hemos descrito anteriormente, vamos a ver un ejemplo, que aunque no cubre todos los casos, refleja los más importantes.

EJEMPLO 1

Imaginemos un estudio en el que se quiere analizar la supervivencia en pacientes con trasplante de corazón. El periodo de estudio son 52 semanas, donde el evento de interés es la muerte del paciente, aunque no todos los pacientes tienen por qué morir en ese periodo de tiempo. Así que, el estudio lleva consigo el registro de información muestral incompleta en uno u otro sentido.

En el siguiente gráfico se ilustra mediante líneas de distinto tipo las observaciones en 6 pacientes. La línea gruesa indica el periodo de observación del individuo; el círculo grueso representa la observación del evento; el recuadro indica el tiempo de censura; y el círculo vacío representa la ocurrencia del evento pero que no queda registrada en el estudio.

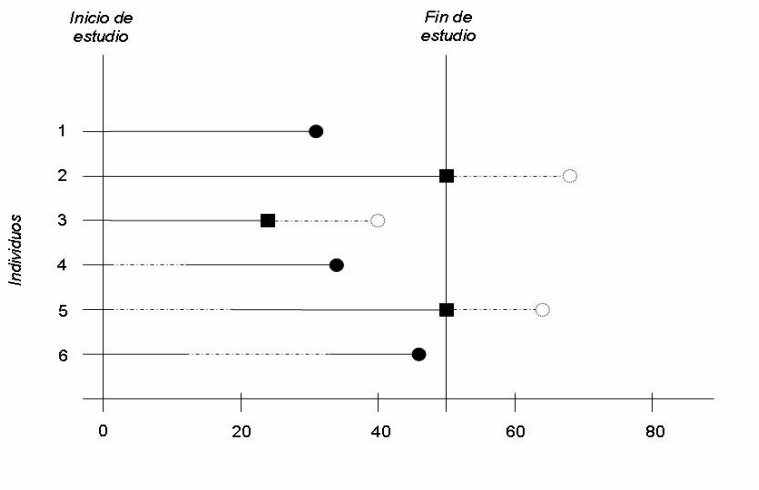

Figura 1.7. Ejemplo de censura y truncamiento en 6 pacientes

a) El sujeto 1 entra en el estudio en el momento del trasplante y muere después de 40 semanas. Por tanto, este dato no tiene censura, ya que el dato es observado desde el comienzo hasta su muerte.

b) El sujeto 2 entra en el estudio en el momento del trasplante y sobrevive más de 52 semanas, por lo que deja de ser observado. A las 90 semanas fallece, sin embargo, esta observación no es tenida en cuenta en el estudio ya que finalizó a las 52 semanas.

   Éste es un ejemplo de censura a la derecha. La censura por la derecha suele ocurrir cuando el estudio tiene una fecha determinada para la finalización.

c) El sujeto 3 entra en el estudio en el momento del trasplante, pero abandona el estudio en la semana 30. Por lo tanto, este es un ejemplo de censura aleatoria a la derecha.

   Aunque el individuo fallece en la semana 50, el dato no es observado.

d) El sujeto 4 entra en el estudio en la semana 15 después del trasplante y muere a las 35 semanas. Estamos ante el caso de una entrada tardía al estudio. Se dice que este dato presenta truncamiento a la izquierda.

e) El sujeto 5 entra en el estudio en la semana 30 después de ser trasplantado, y a la finalización del estudio en la semana 52 aun no ha experimentado el evento. Por lo que es un dato que presenta truncamiento a la izquierda y censura a la

derecha.

f) El sujeto 6 entra en el estudio en el momento del trasplante pero abandona el estudio en la semana 10. La información se pierde hasta la semana 35 en la que vuelve a entrar en el estudio, para finalmente fallecer en la semana 45. Por tanto, este dato es un ejemplo de múltiples intervalos de observación.

EJEMPLO 2

Un ejemplo de datos truncados por la derecha lo encontramos en los tiempos de incubación en enfermos de SIDA. Dado que sólo se registran individuos que han desarrollado SIDA, el tiempo de incubación (T), que es el tiempo que transcurre desde el diagnóstico del VIH y SIDA, sólo se tendrá en cuenta individuos cuyo tiempo de incubación sea menor o igual que la duración total del estudio (τ). Es decir, sólo entran en estudio aquellos individuos que cumplan $T \leq \tau$, por tanto las observaciones son truncadas a la derecha. Un individuo que no haya desarrollado SIDA cuando acabe el periodo de observación no entra en la muestra, no se puede registrar su tiempo de incubación y por lo tanto es ignorada.

EJEMPLO 3

Para el caso de la censura a la izquierda se puede utilizar como ejemplo un estudio en adolescentes que empiezan a fumar mariguana (Klein & Moeschberger, 1997). Se selecciona una muestra de adolescentes con 16 años que ya fuman mariguana y se les pregunta a qué edad comenzaron. La variable de interés es el tiempo hasta que fuman por primera vez (T). Algunos lo recuerdan con exactitud (a los 14 años, a los 15 años, …) por lo que proporcionan información completa sobre la variable en estudio. Sin embargo, hay otros jóvenes que no recuerdan en qué momento fumaron por primera vez, pero dado que ahora fuman y tienen 16 años, la información que aportan es un dato censurado a la izquierda, ya que se sabe que $T \leq 16$.

4. Estimación no paramétrica de las características de fiabilidad

Los modelos no paramétricos son métodos analíticos y gráficos que permiten interpretar los datos obtenidos sin asumir ningún tipo concreto de modelo probabilístico para los tiempos de fallo y las funciones básicas (riesgo, fiabilidad), por lo que se estiman directamente de los datos. En ocasiones, esto puede resultar ventajoso pues no se

requieren de grandes supuestos previos sobre el modelo.

A continuación, describiremos algunos de los métodos no paramétricos empleados en la estimación de las características en fiabilidad:

- Función de Fiabilidad Empírica
- Estimador de Kaplan-Meier o Estimador Producto Límite
- Estimador de la función de riesgo acumulada: Estimador de Nelson-Aalen

4.1. Función de Fiabilidad Empírica

Supongamos n ítems que son observados hasta que ocurre el fallo, y t_j el tiempo de ocurrencia del fallo, por lo que tendremos $t_1 < t_2 < \ldots < t_k$ con $k \leq n$; y sea d_j el número de fallos ocurridos en el tiempo t_j, con $j = 1, 2, \ldots, k$.
Se define la **función de fiabilidad empírica** como

$$S_n(t) = \frac{\text{nº de observaciones después de t}}{n}$$

es decir,

$$S_n(t) = \frac{n - \sum_{i=1}^{j} d_j}{n}, \qquad t_j \leq t \leq t_{j+1}$$

para $j = 0, 1, \ldots, k$, siendo $t_0 = 0$ y $t_{k+1} = \infty$.

Propiedades

- Es no creciente
- Toma valor 0 en todo t menor que el primer tiempo de fallo observado, t_1.
- Toma valor 1 en todo t mayor que el último tiempo de fallo observado, t_k.
- Es continua a la derecha. Permanece constante entre dos observaciones consecutivas y presenta un salto de magnitud d_j/n en la observación j-ésima.
- Si F es la función de distribución que describe la variable aleatoria T, tiempo de vida del sistema en estudio, F = 1 - S, se tiene el conocido Teorema de Glivenko-Cantelli, según el cual, si

$$D_n = \sup_{-\infty<t<\infty} \left| F(t) - F_n(t) \right|$$

entonces, $P(D_n \to 0) = 1$. O dicho de otra forma, $F_n(t)$ converge uniformemente hacia $F(t)$, casi seguramente. En este enunciado entendemos por $F_n = 1 - S_n$.

Sin embargo, la función de fiabilidad estimada tiende a subestimar la función de fiabilidad, por lo que no es un buen estimador cuando en la muestra aparecen observaciones censuradas. Esto es así, ya que estamos entendiendo que los ítems en cuestión fallan en el tiempo de censura, con lo cual estamos falseando la información porque lo único que sabemos es que hasta ese instante de censura el ítem no había fallado, pero tras ese tiempo no se tiene información.

Si la muestra procede de un plan de ensayo en el que la duración del test está prefijada, es decir, es Tipo I, transcurrido un tiempo T_0 no hay observaciones, por lo que el estimador S_n está definido únicamente en el intervalo $[0,T_0]$. Fuera de ese intervalo no puede estimarse la función de fiabilidad ya que no hay información muestral.

Si en cambio, se lleva a cabo un plan de ensayo Tipo II, es decir, se observa la muestra hasta la ocurrencia del r-ésimo fallo, construimos un estimador S_n hasta que alcanza el valor $(n-r)/n$. De esta forma, estará definida en el intervalo $[0,t_r]$.

Cuando las muestras son multicensuradas, entonces se emplearán otros métodos más sofisticados como el estimador de Kaplan-Meier que veremos a continuación.

4.2. Estimador de Kaplan- Meier o Estimador Producto Límite

Las técnicas de estimación no paramétrica con datos censurados se inicia con los aportes de Kaplan y Meier (1958), quienes publicaron algunos resultados obtenidos para observaciones censuradas a la derecha y añaden un estudio de las propiedades básicas de un nuevo estimador que se conocerá con el nombre de sus autores. A partir de entonces, el **Estimador de Kaplan-Meier** o **Estimador de Producto Límite** se convierte en uno de lo métodos no paramétricos más utilizados para estimar la función de fiabilidad con datos no agrupados en presencia de censura.

Definición

El Estimador de Kaplan-Meier se define de la siguiente forma. Supongamos que se observa una muestra de n piezas en funcionamiento, de forma que de cada pieza i se registra o bien su tiempo de fallo o bien su tiempo de censura. El objetivo es estimar la función de fiabilidad de la variable aleatoria T, tiempo de fallo. Gráficamente, la información de la muestra vendría dada en la forma:

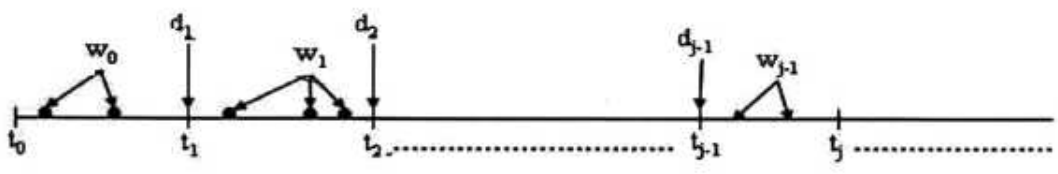

Figura 1.8. Observación de n piezas

Se supone que se registran un total de k $(k \leq n)$ distintos tiempos $0 = t_0 < t_1 < t_2 < \ldots < t_k$ en los cuales ocurren los fallos. Por tanto, decimos que se registran k fallos, con lo cual $w = n - k$ es el número de datos censurados. El método de Kaplan-Meier estima la probabilidad de supervivencia en el tiempo t_j mediante el producto de la probabilidad de supervivencia en el tiempo t_{j-1} por la probabilidad condicionada de sobrevivir al tiempo t_j si se ha sobrevivido hasta el tiempo t_{j-1}. A continuación, definimos las siguientes cantidades:

- d_j : es el número de piezas que han fallado en el instante t_j, si no hay empates, $d_j = 1$;
- w_j : es el número de piezas censuradas en el intervalo (t_j, t_{j+1});
- n_j : es el número de piezas en riesgo inmediatamente antes del instante t_j, y se calcula de la forma:

$$n_0 = n;$$
$$n_1 = n - w_0;$$
$$n_2 = n_1 - w_1 - d_1;$$
$$\ldots\ldots\ldots\ldots$$
$$n_j = n_{j-1} - w_{j-1} - d_{j-1};$$
$$\ldots\ldots\ldots\ldots$$

Dado que conocemos exactamente dónde ocurren los fallos, la probabilidad de supervivencia permanecerá constante entre dos fallos consecutivos. Las censuras deben ser tenidas en cuenta a la hora de calcular el número de unidades en riesgo en cada punto de estimación.

Así, el **Estimador de Kaplan-Meier** se define de la siguiente forma

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

Si el último dato registrado es un tiempo de censura t^*, quiere decir que no hemos observado todos los fallos de las piezas en estudio, por tanto, la estimación de la función de supervivencia no vale cero en ningún momento, de modo que no podemos estimar esta curva hasta $+\infty$, ya que no sería una curva de supervivencia propiamente dicha. En este caso, construimos el estimador sólo hasta este tiempo de censura, t^*. El último intervalo sería por tanto $[t_k, t^*)$, y en este caso, $\hat{S}_k > 0$.

Propiedades

El estimador de Kaplan-Meier es **el estimador no paramétrico máximo verosímil de la función de fiabilidad** por lo que presenta varias de las propiedades deseables en un buen estimador.

Estas propiedades proporcionan facilidad de cálculo y la ventaja de poder utilizarlo en problemas con datos censurados a la derecha. Además, cuando las estimaciones de $S(t)$ se hacen con datos en ausencia de censura, su expresión coincide con la del estimador no paramétrico de la función de fiabilidad, es decir la función de fiabilidad empírica definida anteriormente. Por lo tanto, este estimador tiene muchas de las propiedades deseables en un estimador para muestras grandes. Sin embargo, para muestras pequeñas ya no son tan robustas. En particular, es sesgado para muestras finitas y la magnitud del sesgo es inversamente proporcional al tamaño de la muestra. Por otro lado, la eficiencia asintótica del estimador de Kaplan-Meier es inferior a la del estimador paramétrico de máxima verosimilitud si el nivel de censura es alto o la supervivencia está próxima a cero.

Varianza del Estimador

El estimador de Kaplan-Meier de $S(t)$ da una estimación puntual o un único valor para esta función en cualquier instante t. Por lo tanto, si se desea tener una medida de la precisión de este estimador en diferentes instantes de tiempo o sobre diferentes muestras, es necesario contar con un buen estimador de su varianza, el cual viene dado por la fórmula de Greenwood (1926).

Para calcular la varianza usaremos la aproximación por el método delta (ver Meeker y Escobar, 1998). El desarrollo en serie de Taylor hasta el primer sumando para $\hat{S}_i$, considerada como función de $\hat{q}_i$, nos daría:

$$\hat{S}(t_i) = S(t_i) + \sum_{j=1}^{i} \left.\frac{\partial \hat{S}}{\partial \hat{p}_j}\right|_{p_j} \left(\hat{p}_j - p_j\right)$$

y, a partir de aquí, la varianza vendría dada, usando el método delta, por

$$Var\left[\hat{S}_i\right] = E\left[\left(\hat{S}_i - S_i\right)^2\right] = \sum_{j=1}^{i}\sum_{k=1}^{i} \frac{\partial \hat{S}_i}{\partial \hat{p}_j}\frac{\partial \hat{S}_i}{\partial \hat{p}_k} Cov\left(\hat{p}_j, \hat{p}_k\right)$$

Además, cuando n es grande las $\hat{q}_i$ son incorreladas, esto es para $i \neq j$,

$$Cov\left(\hat{p}_j, \hat{p}_k\right) = 0$$

con lo cual

$$Var\left[\hat{S}_i\right] = \sum_{j=1}^{i}\left(\frac{\partial \hat{S}_i}{\partial \hat{p}_j}\right)^2 Var\left(\hat{p}_i\right)$$

y, sustituyendo, se obtiene la fórmula de Greenwood para la estimación de la varianza

$$\hat{Var}\left[\hat{S}(t_i)\right] \approx \left[\hat{S}(t_i)\right]^2 \sum_{j=1}^{i}\frac{\hat{q}_j}{\hat{p}_j n_j} = \left[\hat{S}(t_i)\right]^2 \sum_{j=1}^{i}\frac{d_j}{n_j(n_j - d_j)}$$

Intervalos de confianza

A partir del estimador del error estándar que hemos obtenido anteriormente podemos construir intervalos de confianza basándonos en que la distribución de

$$Z_{\hat{S}} = \frac{\hat{S}(t_i) - S(t_i)}{\hat{E}E_{\hat{S}(t_i)}}$$

es asintóticamente normal estándar, con lo que

$$\left[S_1(t_i); S_2(t_i)\right] = \hat{S}(t_i) \pm \hat{E}E_{\hat{S}(t_i)} \cdot Z_{\alpha/2}$$

donde $z_{\alpha/2}$ es el cuantil de orden $1-\alpha/2$ de la distribución N(0,1).

Cuando el tamaño muestral no es grande, la ley normal puede no ser una aproximación adecuada para la distribución de $Z_{\hat{S}}$, particularmente en las colas de la distribución, por ejemplo, es posible que $S_1(t) < 0$ ó $S_2(t) > 1$, resultados que están fuera del rango de posibles valores de una función de fiabilidad. Podría obtenerse una mejor aproximación usando la transformación *logit*, esto es:

$$\log it(p) = \log\left[\frac{p}{1-p}\right]$$

y basándose en la distribución de

$$Z_{\log it(\hat{S})} = \frac{\log it\left[\hat{S}(t_i)\right] - \log it\left[S(t_i)\right]}{\hat{E}E_{\log it\left[\hat{S}(t_i)\right]}}$$

Esta transformación de $\hat{S}(t_i)$, verifica

$$-\infty < \log it\left[\hat{S}(t_i)\right] < \infty$$

con lo cual $Z_{\log it(\hat{S})}$ está más próxima a una N(0,1). Con esto construimos el intervalo de confianza

$$\left[S_1(t_i); S_2(t_i)\right] = \left[\frac{\hat{S}(t_i)}{\hat{S}(t_i) + \left(1 - \hat{S}(t_i)\right) \times w}, \frac{\hat{S}(t_i)}{\hat{S}(t_i) + \left(1 - \hat{S}(t_i)\right) / w}\right]$$

con

$$w = \exp\left\{z_{\alpha/2} \frac{\hat{E}E_{\hat{S}(t_i)}}{\hat{S}(t_i) + \left(1 - \hat{S}(t_i)\right)}\right\}.$$

4.3. Estimador de la Función de Riesgo Acumulada: Estimador de Nelson-Aalen

Dado que la función de supervivencia se puede expresar en términos de la función de riesgo, $R(t) = \square\ ln\ S(t)$, si se desea estimar la función de riesgo acumulada, entonces se tiene que, $\hat{R}(t) = -\ln \hat{S}(t)$ donde $\hat{S}(t)$ es el estimador de Kaplan-Meier de la función de supervivencia.

Otro posible estimador de $R(t)$ puede obtenerse mediante las sumas acumuladas de la estimación empírica de la función de riesgo

$$\tilde{R}(t) = \tilde{\Delta}(t) = \sum_{j, t_j \le t} \frac{d_j}{n_j}$$

donde d_j representa el número de fallos ocurridos en el instante t_j; y n_j es el número de individuos en riesgo en t_j (estas cantidades se definen del mismo modo que en el

estimador de Kaplan-Meier, ver Sección 1.4.2.).

A este estimador se le llama **Estimador de Nelson -Aalen** y fue introducido por primera vez por Nelson (1969, 1972). De un modo independiente, Altshules (1970) redescubrió el mismo estimador para procesos de conteo con animales. Posteriormente, adoptando la formulación del proceso de conteo, Aalen (1978) extendió sus usos más allá de la superviviencia para estudiar sus propiedades usando martingalas.

El cociente d_j/n_j, proporciona una estimación de la probabilidad condicionada de que una unidad que sobrevive hasta el instante inmediatamente anterior a t_j, falle en el instante t_j.

A partir de la relación logarítmica entre R(t) y S(t), se obtiene un estimador alternativo $\tilde{S}(t)$, de la función de supervivencia, conocido como estimador de Fleming-Harrington, de la función de supervivencia, cuya relación es:

$$\tilde{S}(t) = e^{-\tilde{R}(t)} = e^{-\sum_{j,j\le t}\frac{d_j}{n_j}}$$

Este estimador es de gran utilidad en la construcción de gráficas, para evaluar la selección de una determinada familia paramétrica de distribuciones, cuando se trata de modelizar la distribución del tiempo de vida de una unidad y realizar unas primeras estimaciones de los parámetros del modelo seleccionado.

Cuando T es una variable continua, $\tilde{S}(t)$ y $\hat{S}(t)$, son estimadores asintóticamente equivalentes. En realidad, $\tilde{S}(t)$ es la aproximación lineal de primer orden de la función $\hat{S}(t)$, ya que

$$R(t) = -\sum_{j,t_j\le t}\ln(1-\hat{q}_j) \approx \sum_{j,t_j\le t}\hat{q}_j = \sum_{j,t_j\le t}\frac{d_j}{n_j} = \tilde{R}(t)$$

5. Comparación no paramétrica de curvas

Un problema típico en fiabilidad consiste en comparar las funciones de supervivencia de k (≥ 2) poblaciones que se diferencian en algún factor. Es decir, se trata de detectar posibles diferencias, en cuanto a las funciones de supervivencia, que dicho factor pudiese haber inducido. Aunque la representación gráfica de las curvas de supervivencia alerte sobre posibles diferencias, sería interesante disponer de una prueba estadística que permita establecer diferencias significativas objetivas entre las curvas, lo que indicaría que el factor considerado es determinante en el riesgo de fallo.

Aunque existen numerosos tests no paramétricos, para comparar 2 o más curvas

de supervivencia, nosotros estudiaremos el **Test de Savage o Test de log-rank.**

Vamos a aplicar el procedimiento a la comparación de la supervivencia de dos grupos de ítems, G_1 y G_2, cuya función de supervivencia es respectivamente S_1 y S_2. Este test está indicado, y en este caso es el más potente, cuando el cociente de las funciones de riesgo es aproximadamente constante, esto es, la hipótesis alternativa es la de riesgos proporcionales.

Supongamos que disponemos de una muestra de cada población, de tamaños respectivos n_1 y n_2. De modo que $n = n_1 + n_2$ es el número total de datos en la muestra combinada, para la que denotamos $t_1 < t_2 < ... < t_k$, los tiempos de fallo distintos observados y ordenados.

Formulamos la hipótesis nula de la siguiente forma:

$$H_o : S_1(t) = S_2(t) \qquad \nabla t \leq \tau$$

siendo τ el tiempo total de observación de la muestra, es decir, sería el máximo de los t_j.

El test consiste en comparar el número de fallos observados dentro de cada grupo (G_1 y G_2) y el número de fallos esperado bajo la hipótesis nula.

Definimos las siguientes cantidades:

d_j: el número total de fallos ocurridos en t_j;

n_j: el número total de ítems en riesgo justo antes de t_j;

d_{ij}, i =1, 2: el número de fallos ocurridos en el tiempo t_j entre los individuos del grupo i;

n_{ij}: i =1, 2: el número de ítems en riesgo al principio de t_j entre los individuos del grupo i.

Si la hipótesis nula es cierta, la probabilidad condicional de fallo en t_j es igual para los dos grupos, λ_j, por lo tanto, la distribución de probabilidad (d_{1j}, d_{2j}) sería

$$\prod_{i=1}^{2}\left[\binom{n_{ij}}{d_{ij}}\lambda_j^{d_{ij}}\left(1-\lambda_j\right)^{n_j-d_{ij}}\right]=\left[\prod_{i=1}^{2}\binom{n_{ij}}{d_{ij}}\right]\lambda_j^{d_{ij}}\left(1-\lambda_j\right)^{n_j-d_{ij}}$$

A partir de aquí definimos, para cada i, el estadístico log-rank:

$$U_i = \sum_{j=1}^{k}\left(d_{ij} - e_{ij}\right),$$

se tiene, por el teorema central del límite, cuando *k* es suficientemente grande, que

$$\frac{U_i}{\sqrt{Var(U_1)}} = \frac{\sum_{j=1}^{k}\left(d_{ij} - n_{ij}\right)}{\sqrt{\sum_{j=1}^{k} Var(d_{ij})}} \to N(0,1)$$

y, por lo tanto

$$\frac{U_i}{\sqrt{Var(U_1)}} \to \chi^2(1) .$$

Basándonos en estas dos distribuciones podemos construir tests equivalentes para evaluar las diferencias entre los grupos.

Este test da la misma importancia a las diferencias observadas en todos los tiempos de fallo, independientemente del número de ítems en riesgo en cada caso. No obstante, resulta más apropiado considerar que las diferencias registradas al principio del intervalo de observación deben pesar más en el estadístico que las diferencias que se registran al final, ya que las primeras cuentan con más casos. A partir de este razonamiento se define una familia de tests cuyo estadístico tiene la siguiente forma:

$$U = \frac{\sum_{j=1}^{k} w_j \left(d_{ij} - e_{ij}\right)}{\sqrt{\sum_{j=1}^{k} w_j^2 \left(d_{ij}\right)}}$$

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 de los tiempos observados. Este estadístico tiene, bajo la hipótesis nula, distribución N(0,1). Considerando distintos vectores peso, se obtienen diferentes tests, para **w = 1** obtenemos el test de **log-rank**.

Test de Gehan, Breslow o de Wilcoxon generalizado: $w_j = n_j$.

Test de Tarone, Tarone-Ware: $w_j = \sqrt{n_j}$

Test de Prentice: $w_j = \prod_{i=1}^{j} \frac{n_i - d_i + 1}{n_i + 1}$

El problema de este grupo de tests es que únicamente detectan diferencias en las curvas de supervivencia cuando la situación es una de las siguientes: $S_1(t) < S_2(t)$ ó $S_1(t) > S_2(t)$ para todo *t,* en otro caso, el numerador de *U* sería una suma de términos

positivos y negativos que podría resultar en un valor próximo a 0 y por tanto no sería estadísticamente significativo. Es decir, este grupo de tests resulta especialmente eficaz cuando la hipótesis alternativa se corresponde con la hipótesis de riesgos proporcionales.

6. Estimación con datos truncados a la izquierda y censurados a la derecha

En el caso de truncamiento a la izquierda y censura a la derecha, para el caso de la función de supervivencia, se produce un importante sesgo debido a la subestimación de la supervivencia. Para poder entender mejor como sería un caso de este tipo, pongamos un ejemplo.

Channing House es un centro de retiro de la localidad Palo Alto, California. Se registraron datos correspondientes (Klein & Moeschberger, 1997) a las edades en el momento de fallecimiento de 462 individuos que vivían en la residencia durante el periodo de enero de 1964 a julio de 1975. Dado que un individuo debe sobrevivir un tiempo suficiente (65 años) para entrar en un centro de estas características, todos los individuos que fallecieron a edades tempranas no entrarán en el centro y, por lo tanto, quedan fuera del alcance de las investigaciones, es decir, tales individuos son excluidos del estudio dado que no viven el tiempo suficiente para entrar en el centro. De modo que si la información muestral se restringe a los habitantes de la residencia, no estamos registrando información sobre la variable T = tiempo de vida de un individuo de Palo Alto, sino de la variable T condicionada al suceso $T \geq 65$.

Los datos con trocamiento a la izquierda y censura a la derecha se presentan de la siguiente forma.

Sea (X,T,C) un vector aleatorio, donde X es el tiempo de truncamiento, T es el tiempo de fallo, y C es el tiempo de censura. Se supone que bajo un modelo de truncamiento por la izquierda y censura a la derecha, las observaciones muestrales serán observaciones del vector (X,Y,δ), donde $Y = min\{T,C\}$ y δ es el indicador de censura, es decir, $\delta = 1_{\{T \leq C\}}$ si T es un tiempo de fallo ó 0 si T es un tiempo de censura. Si $Y < X$, no hay observaciones. Es decir, que la muestra está constituida por n observaciones de tipo (x_i, y_i, δ_i) en los que $x_i \leq y_i$ para todo $i=1, 2, ..., n$.

Ahora, nuestro objetivo es obtener el estimador de la función de supervivencia $S(t) = P\{T > t\}$. Este estimador fue propuesto por Turnbull (1976) y Tsai (1987) que definen el estimador de la función de supervivencia para un modelo bajo truncamiento a la izquierda y censura a la derecha (TICD) de la forma

$$\hat{S}(t)=\prod_{i=1}^{n}\left(1-\frac{1_{\{t_i\le t,\delta_i=1\}}}{\sum_{j=1}^{n}1_{\{x_j\le t_i\le t_j\}}}\right)=\prod_{t_i\le t}\left(1-\frac{d_i}{n_i}\right)$$

donde 1_A es el indicador de la función definida por la condición A; d_i denota el número de fallos en y_i; y r_i el número de individuos en riesgo en y_i.

Esta última expresión da el estimador no paramétrico de máxima verosimilitud de la función de supervivencia bajo este supuesto. Se reduce al estimador de Kaplan-Meier cuando no hay datos truncados y a la función de supervivencia (o fiabilidad) empírica cuando no hay ni truncamiento ni censura. Además, bajo ciertas condiciones, verifica propiedades deseables tales como la consistencia y la convergencia a una ley normal.

Sin embargo, este estimador puede presentar serios problemas en caso de pequeñas muestras o muestras grandes con algunos datos truncados en los periodos iniciales. En esta situación, el estimador dado en la última ecuación puede subestimar severamente las probabilidades de supervivencia cerca del origen. Sea $y_1 < y_2 < \ldots < y_k$, podría ocurrir que $d_i = r_i$ para algún y_i, con $i < k$. Si esta situación ocurre, entonces $\hat{S}(t)=0$ para algún $t \ge y_i$, incluso si hay supervivencias y fallos observados más allá de y_i. Esta situación es en parte resuelta por Pan y Chappell (1998) que introducen el estimador de Nelson extendido para la función de riesgo acumulada con datos truncados a la izquierda y censurados a la derecha, y constituyen, basándose en éste, un estimador no paramétrico para la función de supervivencia.

Bajo censura a la derecha, Nelson propuso el siguiente estimador de la función de riesgo acumulada

$$\tilde{H}(t)=\sum_{i=1}^{n}\frac{1_{\{y_i\le t,\delta_i=1\}}}{\sum_{j=1}^{n}1_{\{y_i\le y_j\}}}=\sum_{i=1}^{n}\frac{d_i}{\tilde{r}_i}$$

Basándose en esta expresión, es posible estimar la función de supervivencia por medio de $\tilde{S}(t)=\exp\left(-\tilde{H}(t)\right)$.

Pan y Chappell (1998) extienden este estimador para datos truncados por la izquierda. Para ello, redefinen algunas cantidades involucradas en la última. Específicamente, $\tilde{r}_i$, esto es, el número de individuos en riesgo en el instante y_i. Esta cantidad es obtenida mediante $r_i=\sum_{j=1}^{n}1_{\{x_j\le y_i\le y_j\}}$. En consecuencia, el estimador de

Nelson-Aalen extendido, se expresa

$$\tilde{H}_e(t) = \sum_{i=1}^{n} \frac{1_{\{y_i \le t, \delta_i = 1\}}}{\sum_{j=1}^{n} 1_{\{x_j \le y_i \le y_j\}}} = \sum_{y_i \le t} \frac{d_i}{r_i}$$

y, por tanto, el estimador de la función de supervivencia puede obtenerse $\tilde{S}_e(t) = \exp\left(-\tilde{H}_e(t)\right)$.

7. Modelos Semiparámetricos

Hasta ahora, con los modelos descritos anteriormente, lo que hemos estado estudiando ha sido la relación entre la tasa de fallo (o la función de supervivencia) y el tiempo. Sin embargo, los modelos semiparamétricos permiten realizar ese mismo estudio evaluando el efecto de covariables sobre la función de riesgo. Uno de estos modelos es el Modelo de Riesgos Proporcionales de Cox.

7.1. Modelo de Riesgos Proporcionales de Cox

El modelo de riesgos proporcionales introducido por Cox (1972) es el modelo de regresión más utilizado en análisis de fiabilidad, pero no es a partir del desarrollo del enfoque basado en los procesos de recuento, que este modelo logra su completa madurez, este enfoque ha permitido la verificación de los supuestos de riesgos proporcionales y el estudio de los residuos.

El modelo de riesgos proporcionales permite ver la posible relación con diferentes variables registradas para cada sujeto y no sólo la relación entre la tasa de fallo y el tiempo. Por tanto, se trata de calcular la tasa de mortalidad o de fallo, como una función del tiempo y de un determinado conjunto de variables explicativas o covariables.

7.1.1. El modelo de Cox

El modelo de riesgos proporcionales de Cox, que es un tipo de modelo semiparamétrico, en el que se asume la forma paramétrica únicamente para el efecto de las variables de predicción o de pronóstico e incluye una función de riesgo arbitraria de referencia $r_0(t)$ con forma sin especificar.

De esta forma, el modelo de riesgos proporcionales de Cox se presenta en su forma más habitual como:

$$r(t;Z) = r_0(t)e^{b^T Z} = r_0(t)e^{b_1 z_1 + b_2 z_2 + \ldots + b_p z_p}$$

donde t es la edad alcanzada por la unidad; y $Z^T = (z_1, z_2, \ldots, z_p)$ es la traspuesta de un vector de covariables y b es el vector de parámetros de regresión.

Asumiendo que $r_0(t)$ es la función de riesgo de una unidad con vector de covariables Z=0 (nivel base), en el modelo de riesgos proporcionales de Cox, la función de fiabilidad condicional para T, dado un vector de covariables Z, será:

$$S(t,Z) = e^{-\int_0^t r(u;z)du} = e^{-\int_0^t r_0(u)e^{b^T Z}du} = \left[e^{-\int_0^t r_0(u)du} \right]^{e^{b^T Z}} = \left[S_0(t) \right]^{e^{b^T Z}}$$

donde S_0 es la función de fiabilidad base.

Como conclusión, cuando el objetivo es evaluar la influencia de covariables sobre la función de riesgo, una buena alternativa es el modelo de Cox. Pero cuando lo que se pretende es predecir futuros fallos dentro de un cierto horizonte de tiempo a largo plazo, es conveniente hacer una hipótesis paramétrica sobre la forma de $h_0(t)$.

7.1.2. Modelo de Riesgos Proporcionales Estratificado

Cuando se configuran grupos o estratos en función a algún criterio, en un conjunto de datos, la formulación más general del modelo de Cox estratificado para la función de riesgo de un sistema en el j-ésimo estrato (j = 1, …, p), suele expresarse por

$$h_j(t;Z) = h_{0j}(t)e^{b^T Z}$$

Donde:

- las funciones de riesgo básicas $h_{0j}(t)$ en cada uno de los j estratos son arbitrarias y distintas.
- el vector de coeficientes b^T es el mismo en todos los estratos.

Si la hipótesis de proporcionalidad no se verifica en los p estratos de un factor, se suele considerar este modelo. En este modelo se asume que las funciones de riesgo son proporcionales dentro del mismo estrato, pero no necesariamente a través de los p estratos.

7.1.3. Residuos de Cox-Snell

En este apartado y el siguiente, presentamos diversas técnicas que permiten valorar la bondad del ajuste del modelo de riesgos proporcionales a un conjunto de datos.
La primera técnica que observamos está enfocada a valorar la bondad del ajuste del modelo de Cox de manera global.

Los residuos generalizados de Cox-Snell se definen, para el caso de datos completos, es decir, sin censura, por

$$e_i = -\ln R\left(t_i; Z, \hat{b}\right)$$

donde e_i es el residuo i-ésimo para la unidad i de la muestra; y $R(t_i; Z, \hat{b})$ es la fiabilidad estimada evaluada en t_i con vector de covariables Z.

Si el modelo de Cox es correcto y los valores estimados de los parámetros de regresión están próximos a los reales, los residuos obtenidos deben ajustarse a una distribución exponencial de parámetro 1.
Estos residuos son utilizados si los parámetros de un modelo seleccionado han sido estimados por el método de la máxima verosimilitud con datos censurados, por lo que la bondad del ajuste del modelo de Cox a los datos observados puede basarse en criterios gráficos utilizando residuos.

En observaciones con censura, Cox y Snell propusieron una corrección de primer orden de estos residuos. Estos son los residuos ajustados de Cox -Snell, que son de la forma:

$$e_i = -\ln R(t_i; Z, \hat{b}) + 1$$

7.1.4. Diagnósticos de Regresión

A continuación presentamos diversas técnicas que permiten valorar la bondad del ajuste del modelo de riesgos proporcionales a un conjunto de datos. Para esto vamos a basarnos en análisis de diferentes tipos de residuos generados en el ajuste.

Los residuos se pueden utilizar para:

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

Existen cuatro tipos de residuos de interés en el modelo de Cox: los residuos de martingala, *deviance*, los de puntuaje (score) y los de Schoenfeld. De estos cuatro residuos se pueden derivar otros dos: los residuos dfbetas y los residuos escalados de

Schoenfeld.
A continuación explicamos brevemente cada uno de estos residuos.

Residuos de martingala

Los residuos de martingala se usan para estudiar la forma funcional con que una covariable debería ser introducida en el modelo, y se definen como

$$\hat{M}_i(t) = N_i(t) - \hat{E}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} d\hat{R}_0(\beta, s)$$

donde $\hat{R}_0(\beta, s)$ es el estimador del riesgo base de Breslow (o de Nelson y Aalen) definido como

$$\hat{R}_0(\beta, s) = \int_0^t \frac{\sum_{i=1}^{n} dN_i(s)}{\sum_{i=1}^{n} Y_i(s) e^{\beta' Z_i(s)}}$$

y están basados en la martingala de un proceso de conteo para el i-ésimo individuo, $M_i(t) = N_i(t) - E_i(t)$, definida mediante:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} r_0(s) ds$$

Los residuos de martingala son muy asimétricos y con una cola muy larga hacia la derecha.

Residuos *deviance*

Los residuos de desvíos se utilizan para la detección de valores atípicos (outliers). Estos residuos se obtienen mediante una transformación de normalización de los desvíos de martingala.
Los residuos de desvíos se definen de la siguiente manera: si todas las covariables son fijas en el tiempo, los residuos toman la forma:

$$d_i = signo\left(\hat{M}_i\right) * \sqrt{-\hat{M}_i - N_i \log\left(\left(N_i - \hat{M}_i\right)/N_i\right)}$$

Residuos de puntajes (scores)

Los residuos de puntajes se utilizan para verificar la influencia individual y para la estimación robusta de la varianza. Se definen como:

$$U_{ij} = U_{ij}\left(\hat{\beta}, \infty\right)$$

donde $U_{ij}\ (\beta,t),\ j = 1,...,p$ son las componentes del vector fila de longitud p obtenido a través del proceso de puntaje para el i-ésimo individuo:

$$U_i(\beta) = \int_0^t \left[Z_i(t) - \bar{Z}(\beta, t) \right] dN_i(t)$$

Residuos de Schoenfeld

Los residuos de Schoenfeld son útiles para la verificación del supuesto de riesgo proporcional en el modelo de Cox, y se definen de la forma:

$$s_{ij}(\beta) = Z_{ij}(t_i) - \bar{Z}_j(\beta, t_i)$$

con una fila por individuo y una columna por covariable, donde i y t_i son los individuos y el tiempo de ocurrencia del evento respectivamente.

$$e_i = -\ln S\left(t_i; Z, \hat{b}\right) + 1$$

Capítulo II

Análisis Computacional de Datos de Tiempos de Fallo

Una vez que hemos visto los modelos teóricos más importantes en Análisis de Fiabilidad, el siguiente paso es poder calcular esas funciones mediante métodos informáticos que faciliten su cálculo. Para ello, expondremos en este capítulo las principales herramientas para llevar a cabo un análisis de tiempos de vida mediante el uso del lenguaje de programación R.

El Análisis de Supervivencia en el entorno de programación estadística R puede hacerse a través de diferentes librerías especializadas aunque en nuestro caso nos centraremos en el estudio de la librería *survival,* que es tal vez la principal para realizar Análisis de Supervivencia. Esta librería permite realizar un análisis de datos que presentan características habituales en contextos de Fiabilidad y Supervivencia como son la censura y el truncamiento.

A continuación, se presentan algunas de las funciones contenidas en la librería *survival*, de las que describiremos los argumentos de su sintaxis, junto a algunos ejemplos con sus salidas en R, que nos ayudarán a comprender mejor su empleo.

1. La función *Surv*

- **Definición**

  La función Surv permite crear objetos de tipo survival, por lo general usando como una variable respuesta en la fórmula del modelo.

- **Sintaxis**

```
Surv(time, time2, event,
        type=c('right', 'left', 'interval', 'counting', 'interval2'),
                origin=0)
is.Surv(x)
```

- **Descripción de los argumentos**
  - *time*. Representa el tiempo de inicio de la observación. Para datos de intervalo, el primer argumento es el extremo inicial del intervalo.

  - *event.* Representa el indicador de estado, normalmente 0 = vivo (censurado), 1 = muerto (no censurado). Otras opciones son VERDADERO/FALSO (VERDADERO = la muerte) o 1/2 (2 = muerto). Para datos con censura de intervalo, el indicador de estado es 0 = censura a la derecha, 1 = suceso ocurrido en *time*, 2 = censura a la izquierda, 3 = censura de intervalo. Este indicador puede ser omitido en el caso de que se asuma que todos los sujetos tienen el mismo estado.

  - *time2*. Representa el tiempo de finalización de la observación para un intervalo censurado o proceso de conteo. Se asume que los intervalos están abiertos a la izquierda y cerrados a la derecha, (el principio, el final]. Para un proceso de recuento de datos *event* indica si un acontecimiento ocurrió al final del intervalo.

  - *type*. Es una cadena de caracteres que especifica el tipo de censura. Valores posibles son "derecha", "izquierda", "conteo", "intervalo", o "intervalo2". Por defecto, suele ser censura por la derecha o conteo, dependiendo de si el argumento *time2* está ausente o presente, respectivamente.

  - *origin*. Es una utilidad que permite trabajar bajo el enfoque de los

procesos de recuento. Esta opción es usada en un modelo que contiene estratos dependientes del tiempo, para enumerar los sujetos correctamente que cambian de un estrato a otro. En raras ocasiones se suele emplear.

- *x*. Cualquier objeto de R.

➢ **Ejemplo (1)**

```
with(lung, Surv(time, status))
                    # cuando se observan longitudes de tiempo

Surv(heart$start, heart$stop, heart$event)
                    # cuando se registran tiempos de entrada y salida
```

El primer conjunto de datos usado en este ejemplo "*lung*", hace referencia a la superviencia de pacientes de cáncer de pulmón procedentes del North Central Cancer Treatment Group. El conjunto de datos contiene 228 registros, cada uno de los cuales está compuesto por 10 variables, desde el código del paciente en la institución, el tiempo de observación de cada paciente, el estado al final del estudio (vivo o fallecido), el sexo, la edad, el peso perdido en los últimos 6 meses, así como cierta información indicando el grado de satisfacción con que el individuo realiza tareas cotidianas.

```
> lung
     inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1       3  306      2  74   1       1       90       100     1175      NA
2       3  455      2  68   1       0       90        90     1225      15
3       3 1010      1  56   1       0       90        90       NA      15
4       5  210      2  57   1       1       90        60     1150      11
5       1  883      2  60   1       0      100        90       NA       0
6      12 1022      1  74   1       1       50        80      513       0
7       7  310      2  68   2       2       70        60      384      10
8      11  361      2  71   2       2       60        80      538       1
9       1  218      2  53   1       1       70        80      825      16
```

Figura 2.1. Datos de pacientes de cáncer de pulmón (Loprinzi et al., 1994).

Tras ejecutar el código anterior se crea el objeto de R representado en la Figura 2.2, donde se indican los tiempos de participación de cada individuo resaltando con el símbolo + los datos censurados.

```
> with(lung, Surv(time, status))
  [1]  306   455  1010+  210   883  1022+  310   361   218   166   170   654   728    71   567   144   613   707    61    88   301    81   624   371
 [25]  394   520   574   118   390    12   473    26   533   107    53   122   814   965+   93   731   460   153   433   145   583    95   303   519
 [49]  643   765   735   189    53   246   689    65     5   132   687   345   444   223   175    60   163    65   208   821+  428   230   840+  305
 [73]   11   132   226   426   705   363    11   176   791    95   196+  167   806+  284   641   147   740+  163   655   239    88   245   588+   30
 [97]  179   310   477   166   559+  450   364   107   177   156   529+   11   429   351    15   181   283   201   524    13   212   524   288   363
[121]  442   199   550    54   558   207    92    60   551+  543+  293   202   353   511+  267   511+  371   387   457   337   201   404+  222    62
[145]  458+  356+  353   163    31   340   229   444+  315+  182   156   329   364+  291   179   376+  384+  268   292+  142   413+  266+  194   320
[169]  181   285   301+  348   197   382+  303+  296+  180   186   145   269+  300+  284+  350   272+  292+  332+  285   259+  110   286   270    81
[193]  131   225+  269   225+  243+  279+  276+  135    79    59   240+  202+  235+  105   224+  239   237+  173+  252+  221+  185+   92+   13   222+
[217]  192+  183   211+  175+  197+  203+  116   188+  191+  105+  174+  177+
```

Figura 2.2. La función *Surv* para datos censurados a la derecha.

El segundo conjunto de datos usado como ejemplo en este caso es el conjunto "*heart"*, contiene información sobre supervivencia de 172 pacientes en la lista de espera del programa de transplantes de corazón del hospital de Standford. El registro se muestra como sigue en la siguiente figura.

```
> heart
    start   stop event          age        year surgery transplant  id
1     0.0   50.0     1 -17.15537303  0.12320329       0           0   1
2     0.0    6.0     1   3.83572895  0.25462012       0           0   2
3     0.0    1.0     0   6.29705681  0.26557153       0           0   3
4     1.0   16.0     1   6.29705681  0.26557153       0           1   3
5     0.0   36.0     0  -7.73716632  0.49007529       0           0   4
6    36.0   39.0     1  -7.73716632  0.49007529       0           1   4
7     0.0   18.0     1 -27.21423682  0.60780287       0           0   5
8     0.0    3.0     1   6.59548255  0.70088980       0           0   6
9     0.0   51.0     0   2.86926762  0.78028747       0           0   7
10   51.0  675.0     1   2.86926762  0.78028747       0           1   7
```

Figura 2.3. Datos de pacientes de corazón (Crowley y Hu, 1977).

En este caso cada individuo se incorpora al estudio en el instante que indica la variable *start*, y abandona el estudio en el instante indicado por *stop*. Es decir los pacientes entran en el estudio de manera escalonada, esta es la forma habitual en que los individuos son incorporados en estudios de Supervivencia, a diferencia de en estudios de Fiabilidad, donde todos los sujetos entran en estudio a la vez. No sólo estamos interesados en la longitud del intervalo que determina el tiempo de vida si también de dónde está localizado. La variable *event* nos dice de cada individuo si está vivo o falleció al final del estudio. Además se incluye otro tipo de información como la edad del individuo (*age*), si se le ha realizado un transplante o no (*transplant*) y si ha recibido otro tipo de cirugía antes (*surgery*). Ahora el resultado de la función *Surv* es el objeto que se muestra, en parte, a continuación.

```
> Surv(heart$start, heart$stop, heart$event)
  [1] (  0.0,  50.0 ] (  0.0,   6.0 ] (  0.0,   1.0+] (  1.0,  16.0 ] (  0.0,  36.0+]
 [10] ( 51.0, 675.0 ] (  0.0,  40.0 ] (  0.0,  85.0 ] (  0.0,  12.0+] ( 12.0,  58.0 ]
 [19] ( 17.0,  81.0 ] (  0.0,  37.0+] ( 37.0,1387.0 ] (  0.0,   1.0 ] (  0.0,  28.0+]
 [28] (  0.0,  37.0 ] (  0.0,  18.0+] ( 18.0,  28.0 ] (  0.0,   8.0+] (  8.0,1032.0 ]
 [37] (  0.0,  83.0+] ( 83.0, 219.0 ] (  0.0,  25.0+] ( 25.0,1800.0+] (  0.0,1401.0+]
 [46] (  0.0,  16.0+] ( 16.0, 852.0 ] (  0.0,  16.0 ] (  0.0,  17.0+] ( 17.0,  77.0 ]
 [55] (  0.0,  12.0 ] (  0.0,  46.0+] ( 46.0, 100.0 ] (  0.0,  19.0+] ( 19.0,  66.0 ]
 [64] (  0.0,  41.0+] ( 41.0,1408.0+] (  0.0,  58.0+] ( 58.0,1322.0+] (  0.0,   3.0 ]
 [73] (  0.0,   2.0+] (  2.0, 996.0 ] (  0.0,  21.0+] ( 21.0,  72.0 ] (  0.0,   9.0 ]
 [82] (  0.0,  32.0+] ( 32.0, 285.0 ] (  0.0, 102.0 ] (  0.0,  41.0+] ( 41.0, 188.0 ]
 [91] ( 67.0, 942.0+] (  0.0, 149.0 ] (  0.0,  21.0+] ( 21.0, 343.0 ] (  0.0,  78.0+]
```

Figura 2.4. La función Surv con datos con entrada escalonada.

2. La función *survfit*

- **Definición**

  Esta función permite crear curvas de supervivencia utilizando el método de Kaplan- Meier (opción por defecto) o de Fleming y Harrington. También permite predecir la función de supervivencia para modelos de Cox, o un modelo de tiempo de vida acelerada (ver Klein y Moeschberger (1997), por ejemplo).

- **Sintaxis**

  survfit(formula, ...)

- **Descripción de los argumentos**
  - *formula*. Objeto para la fórmula.
  - *data.* Los datos se enmarcan para interpretar las variables llamadas en los argumentos *formula*, *subset* y *weights*.
  - *weights*. Pesos del caso.
  - *subset.* Expresión que indica un subconjunto de las filas de *data* para ser usado en la estimación. Puede ser un vector lógico (de longitud igual al número de observaciones), un vector numérico indicando el número de observaciones que deben ser incluidas (o excluidas si es negativo), o un vector de caracteres para incluir el nombre de las filas. Todas las observaciones son incluidas por defecto.
  - *na.action*. Función para filtrar datos faltantes. Esto es aplicado al marco modelo después de que haya sido aplicado *subset.* Por defecto es *options* () *$na.action*. Un posible valor para *na.action* es *na.omit*, que suprime las observaciones que contienen uno o varios valores perdidos.
  - *times.* Vector de tiempos en el que la curva de supervivencia es evaluada. Por defecto, el resultado será evaluado en cada valor diferente del vector de tiempos suministrados en *formula*.
  - *type*. Una cadena de caracteres que especifica el tipo de curva de

supervivencia. Los posibles valores son: *"kaplan-meier", "fleming-harrington" o "fh2"* si se da una formula.

- *error*. Una cadena de caracteres especificando el error del estimador. Como valores posibles se tiene *"greenwood"* para la formula de Greenwood o *"tsiatis"* para la fórmula de Tsiatis, (es suficiente con el primer carácter).

- *conf.type*. Puede ser *"none", "plain", "log"* (por defecto), o *"log-log"*. La primera opción no calcula intervalos de confianza. La segunda calcula los intervalos estándar $curve \pm k * se(curve)$, donde k se determina *conf.int*. La opción *log* option calcula intervalos basados en la función de riesgo acumulado o *log*(*survival*).

- *start.time.* Valor numérico que especifica un instante de tiempo donde empezar a calcular la información sobre la supervivencia. La curva resultante es la curva de supervivencia condicional a sobrevivir por encima de *start.time.*

- *conf.int*. El nivel para intervalos de confianza bilaterales. Por defecto es 0.95.

- *se.fit.* Un valor lógico indicando si los errores estándar deben ser calculados. Por defecto es *TRUE.*

- **Resultados**

  El resultado es un objeto de clase *survfit* conteniendo una o varias curvas de supervivencia. Además proporciona la información siguiente.

n número total de sujetos en cada curva.
time Los instantes temporales en que la curva salta.
n.risk Número de sujetos en riesgo en cada tiempo t.
n.event Número de sucesos ocurridos en cada tiempo t.
n.enter Para datos de procesos de recuento solo, el número de sujetos que entran en el tiempo t.
n.censor Para procesos de recuento solo, el número de sujetos que salen del conjunto de riesgo, sin sufrir el suceso, en el tiempo t.
surv La supervivencia estimada en el tiempo t+0.

std.err El error estándar de la función de riesgo acumulado.
upper Límite superior de confianza para la curva de supervivencia.
lower Límite inferior de confianza para la curva de supervivencia.
strata Si hay múltiples curvas, esta componente da el número de elementos de time, etc. correspondientes a la primera curva, las segunda curva etc. Los nombres de los elementos son etiquetas para las curvas.

- **Ejemplo (2)**

```
leukemia.surv<-survfit(Surv(time,status)~x,data=aml)
plot(leukemia.surv,lty=2:3)
legend(100,.9,c("Maintenance","No Maintenance"),lty=2:3)
title("Kaplan-Meier Curves\nfor AML Maintenance Study")

lsurv2<-survfit(Surv(time,status)~x,aml, type='fleming')
plot(lsurv2,lty=2:3,fun="cumhaz",xlab="Months",ylab="Cumulative
Hazard")
```

El conjunto de datos empleados en este ejemplo, llamado *'aml'*, hace referencia a la supervivencia en pacientes con enfermedad aguda de leucemia donde "time" es el tiempo de supervivencia o censura; "status", el estado de censura; y "x" indica el mantenimiento o no de la quimioterapia.
Para estos datos, cabría preguntarse si el tratamiento de quimioterapia debería de ampliarse o no para ciclos adicionales.

time	*status*		*x*
1	*9*	*1*	*Maintained*
2	*13*	*1*	*Maintained*
3	*13*	*0*	*Maintained*
....			
20	*30*	*1*	*Nonmaintained*
21	*33*	*1*	*Nonmaintained*
22	*43*	*1*	*Nonmaintained*
23	*45*	*1*	*Nonmaintained*

Resultado

Observando la curva de supervivencia que hemos obtenido (Figura 2.5), vemos

como en los 50 primeros meses de tratamiento de la enfermedad con quimioterapia, la supervivencia de estos pacientes aumentaba progresivamente, hasta mantenerse casi estable a partir de ese tiempo.

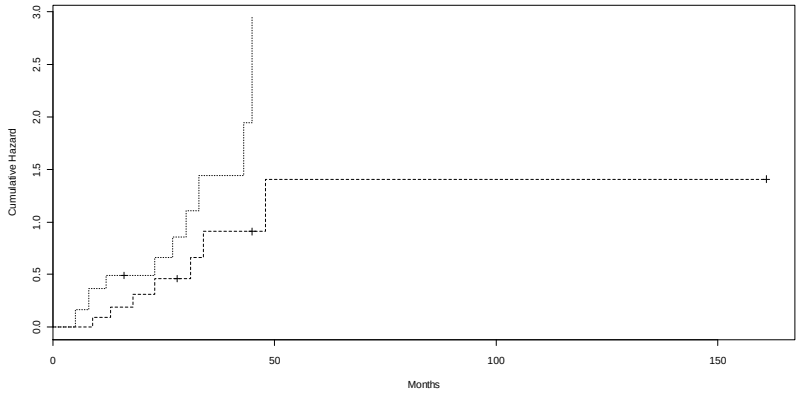

Figura 2.5. Curva de supervivencia del Ejemplo (2)

3. La función *survexp*

- **Definición**

  Devuelve la supervivencia esperada de una cohorte de sujetos, o la supervivencia esperada para cada sujeto de forma individual.

- **Sintaxis**

  survexp(formula, data, weights, subset, na.action, times, cohort=TRUE, conditional=FALSE, ratetable=survexp.us, scale=1, npoints, se.fit, model=FALSE, x=FALSE, y=FALSE)

- **Descripción de los argumentos**

  - *Formula*. Fórmula o modelo antes descrito.
  - .... Otros argumentos al método específico.

- **Ejemplo (3.1)**

```
# Estimador de la supervivencia condicional
```

```
survexp(futime ~ ratetable(sex="male", year=accept.dt,
                age=(accept.dt-birth.dt)), conditional=TRUE, data=jasa)
```

El conjunto de datos empleados en este ejemplo, llamado *'jasa'*, hace también referencia a la supervivencia en pacientes que están en lista de espera en el programa de transplante de corazón en Stanford. Las variables principales incluidas en este conjunto de datos son:

birth.dt: fecha de nacimiento
accept.dt: aceptación en programa
tx.date: fecha de trasplante
fu.date: final de seguimiento del paciente
fustat: estado al finalizar el seguimiento (muerto o vivo)
surgery: cirugía previa
age: edad (en días)
futime: tiempo de seguimiento
wait.time: tiempo de espera hasta el trasplante
trasplant: indicador de trasplante

Resultado

```
Call:
survexp(formula = futime ~ ratetable(sex = "male", year = accept.dt,
 age = (accept.dt - birth.dt)), data = jasa, conditional = TRUE)
 age ranges from 8.8 to 64.4 years
 male: 103  female: 0
 date of entry from 1967-09-13 to 1974-03-22

Time n.risk survival
0    102    1.000
1    102    1.000
2     99    1.000
4     96    1.000
5     94    1.000
...
1321      7    0.974
```

1386 6 0.972
1400 5 0.972
1407 4 0.972
1571 3 0.969
1586 2 0.969
1799 1 0.967

- **Ejemplo (3.2)**

  # Estimador de la supervivencia condicional estratificada a priori
  survexp(futime ~ surgery + ratetable(sex="male", year=accept.dt, age=(accept.dt-birth.dt)), conditional=TRUE, data=jasa)

Resultado

Call:
survexp(formula = futime ~ surgery + ratetable(sex = "male", year = accept.dt, age = (accept.dt - birth.dt)), data = jasa, conditional = TRUE)

age ranges from 8.8 to 64.4 years
male: 103 female: 0
date of entry from 1967-09-13 to 1974-03-22

surgery=0
Time n.risk survival
0 87 1.000
1 87 1.000
2 85 1.000
4 82 1.000
5 80 1.000
...
1400 4 0.972
1407 3 0.972
1571 3 0.969
1586 2 0.969
1799 1 0.967

surgery=1

Time	*n.risk*	*survival*
0	*15*	*1.000*
1	*15*	*1.000*
2	*14*	*1.000*
4	*14*	*1.000*
5	*14*	*1.000*
……		
1400	*1*	*0.971*
1407	*1*	*0.971*
1571	*0*	*0.971*
1586	*0*	*0.971*
1799	*0*	*0.971*

- **Ejemplo (3.3)**

```
#Comparación de las curvas de supervivencia para el modelo estimado
de #Cox
pfit <-coxph(Surv(time,status>0) ~ trt + log(bili) + log(protime) + age +
                                  platelet, data=pbc)
plot(survfit(Surv(time, status>0) ~ trt, data=pbc))
lines(survexp( ~ trt, ratetable=pfit, data=pbc), col='purple')
```

Resultado

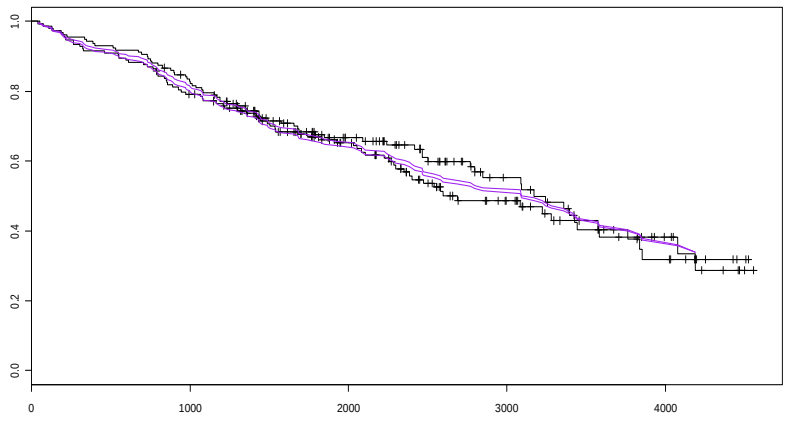

Figura 2.6. Curva de Supervivencia para el Ejemplo (3)

El gráfico muestra como a medida que pasa el tiempo, la supervivencia de los pacientes que están esperando un transplante disminuye.

4. La función *survdiff*

- **Definición**

  Esta función permite realizar contrastes de hipótesis para verificar si hay diferencia entre dos o más curvas de supervivencia basadas en las familias de pruebas G de Harrington y Fleming (1982), o para una sola curva contra una alternativa conocida.

- **Sintaxis**

  survdiff(formula, data, subset, na.action, rho=0)

- **Descripción de los argumentos**
  - *formula*. Expresión de la fórmula como para otros modelos de supervivencia, de la forma *Surv (time, status) ~ predictors*. Para el test de una muestra, los predictores deben consistir en un solo término *offset (sp)*, donde *sp* es un vector que da la probabilidad de supervivencia de cada sujeto. Para el test de *k*-muestas, una combinación única de predictores definidas para un subgrupo. Un término *strata* debe ser usado para producir una prueba estratificada. En el caso de valores perdidos en las estimaciones deben ser tratados como un grupo separado, más bien que omitidos, usa la función *strata* con su argumento *na.group=T*.

  - *subset.* Esta expresión indica el subconjunto de filas de datos que deben ser usadas en la estimación.

  - *rho*. Parámetro escalar que controla el tipo del test.

- **Resultados**
  - *n*. Número de sujetos en cada grupo.

  - *obs*. Número observado de acontecimientos en cada grupo. Si hay estratos, esto será una matriz con una columna por estrato.

- *exp*. Número esperado de acontecimientos en cada grupo. Si hay estratos, esto será una matriz con una columna por estrato.

- *chisq*. Estadístico chis.cuadrado para una prueba de igualdad.

- *var*. Matriz de varianzas de la prueba.

- *strata*. Número de sujetos contenido en cada estrato.

- **Ejemplo (4.1)**

```
# Test para dos muestras
survdiff(Surv(futime, fustat) ~ rx,data=ovarian)
```

El conjunto de datos empleados en este ejemplo, llamado *'ovarian'*, compara la supervivencia en 2 grupos de pacientes con cáncer de ovarios que siguen tratamientos diferentes. Las variables de este conjunto de datos son

	futime	*fustat*	*age*	*resid.ds*	*rx*	*ecog.ps*
1	59	1	72.3315	2	1	1
2	115	1	74.4932	2	1	1
3	156	1	66.4658	2	1	2
4	421	0	53.3644	2	2	1
5	431	1	50.3397	2	1	1
……						
22	268	1	74.5041	2	1	2
23	329	1	43.1370	2	1	1
24	353	1	63.2192	1	2	2
25	365	1	64.4247	2	2	1
26	377	0	58.3096	1	2	1

futime: tiempo de supervivencia o censura.
fustat: estado (censura o muerte)
age: años
rx: grupo de tratamiento

Resultado

Call:
survdiff(formula = Surv(futime, fustat) ~ rx, data = ovarian)

	N	*Observed*	*Expected*	*(O-E)^2/E*	*(O-E)^2/V*
rx=1	13	7	5.23	0.596	1.06
rx=2	13	5	6.77	0.461	1.06

Chisq= 1.1 *on* 1 *degrees of freedom, p*= 0.303

- **Ejemplo (4.2)**

Test para 7 muestras estratificadas
survdiff(Surv(time, status) ~ pat.karno + strata(inst), data=lung)

Resultado

Call:
survdiff(formula = Surv(time, status) ~ pat.karno + strata(inst),
data = lung)

n=224, 4 *observations deleted due to missingness.*

	N	*Observed*	*Expected*	*(O-E)^2/E*	*(O-E)^2/V*
pat.karno=30	2	1	0.692	0.1372	0.15752
pat.karno=40	2	1	1.099	0.0089	0.00973
pat.karno=50	4	4	1.166	6.8831	7.45359
pat.karno=60	30	27	16.298	7.0279	9.57333
pat.karno=70	41	31	26.358	0.8174	1.14774
pat.karno=80	50	38	41.938	0.3698	0.60032
pat.karno=90	60	38	47.242	1.8080	3.23078
pat.karno=100	35	21	26.207	1.0345	1.44067

Chisq= 21.4 *on* 7 *degrees of freedom, p*= 0.00326

- **Ejemplo (4.3)**

```
# Expectativas de supervivencia en pacientes de transplante de corazón
expect <- survexp(futime ~ ratetable(age=(accept.dt - birth.dt),
        sex=1,year=accept.dt,race="white"), jasa, cohort=FALSE,
        ratetable=survexp.usr)
survdiff(Surv(jasa$futime, jasa$fustat) ~ offset(expect))
```

Resultado

```
Call:
survdiff(formula = Surv(jasa$futime, jasa$fustat) ~ offset(expect))

Observed Expected       Z       p
 75.000    0.587 -97.119    0.000
```

5. La función *survreg*

- **Definición**

  Permite ajustar modelos de regresión paramétricos en análisis de supervivencia. Éstos son modelos localización y escala para transformaciones de la variable tiempo. Las distribuciones que se pueden modelar directamente a través de la función survreg son la Weibull, la exponencial, la Normal, la lognormal, la logística y la log-logística.

- **Sintaxis**

```
survreg(formula, data, weights, subset,
        na.action, dist="weibull", init=NULL, scale=0,
        control,parms=NULL,model=FALSE, x=FALSE,
        y=TRUE, robust=FALSE, score=FALSE, ...)
```

- **Descripción de los argumentos**
  - *formula.* Expresión de la fórmula como para otros modelos de regresión.

- *na.action*. Esta función, usada después de *subset,* filtra los valores perdidos. Por defecto es *opcions()$na.action*.

- *dist*. Distribución asumida para la variable y. Si el argumento es una cadena de caracteres, entonces se asume que llama a un elemento de *survreg.distributions*. Estos incluyen "weibull", "exponencial", "normal", "logístico", "lognormal" y "loglogistica".

- *parms*. Contiene una lista de parámetros fijos. Por ejemplo, para la distribución de *t de Student*, indica los grados de libertad. La mayoría de las distribuciones no tienen ningunos parámetros.

- *INIT*. Vector opcional para valores iniciales de los parámetros.

- *scale*. Valor fijo opcional para la escala. Si es <=0 entonces la escala es estimada.

- *control*. Contiene una lista de valores de control, en el formato producido por *survreg.control*.

- *score*. Devuelve el vector resultante.

- *robust*. Usa errores estándar robustos, basados en la independencia de individuos si no hay ningún término *cluster ()* en la fórmula, basada en la independencia de grupos si los hubiere.

- **Resultados**

  Devolverá un objeto de la clase *survreg*.

- **Ejemplo (5.1)**

  survreg(Surv(futime, fustat) ~ ecog.ps + rx, ovarian, dist='weibull', scale=1)

  Resultado

Call:
survreg(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian,
dist = "weibull", scale = 1)

Coefficients:
(Intercept) ecog.ps rx
6.9618376 -0.4331347 0.5815027

Scale fixed at 1

Loglik(model)= -97.2 *Loglik(intercept only)*= -98
Chisq= 1.67 *on* 2 *degrees of freedom, p*= 0.43
n= 26

- **Ejemplo (5.2)**

survreg(Surv(futime, fustat) ~ ecog.ps + rx, ovarian,
dist="exponential")

Resultado

Call:
survreg(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian,
dist = "exponential")

Coefficients:
(Intercept) ecog.ps rx
6.9618376 -0.4331347 0.5815027
Scale fixed at 1

Loglik(model)= -97.2 *Loglik(intercept only)*= -98
Chisq= 1.67 *on* 2 *degrees of freedom, p*= 0.43
n= 26

- **Ejemplo (5.3)**

survreg(Surv(time, status) ~ ph.ecog + age + strata(sex), lung)

Resultado

Call:
survreg(formula = Surv(time, status) ~ ph.ecog + age + strata(sex),
data = lung)

Coefficients:
(Intercept) ph.ecog age
6.73234505 -0.32443043 -0.00580889

Scale:
sex=1 *sex*=2
0.7834211 0.6547830

Loglik(model)= -1137.3 *Loglik(intercept only)*= -1146.2
Chisq= 17.8 *on* 2 *degrees of freedom, p*= 0.00014
n=227 *(*1 *observation deleted due to missingness)*

- **Ejemplo (5.4)**

```
# survreg's scale = 1/(rweibull shape)
# survreg's intercept = log(rweibull scale)
y <- rweibull(1000, shape=2, scale=5)
survreg(Surv(y)~1, dist="weibull")
```

Resultado

Call:
survreg(formula = Surv(y) ~ 1, dist = "weibull")

Coefficients:
Intercept=1.628072
Scale= 0.4883554

Loglik(model)= -2206.4 *Loglik(intercept only)*= -2206.4
n= 1000

- **Ejemplo (5.5)**

```
tobinfit <- survreg(Surv(durable, durable>0, type='left') ~ age + quant,
data=tobin, dist='gaussian')
survreg.control
```

Resultado

```
function (maxiter = 30, rel.tolerance = 1e-09, toler.chol = 1e-10,
    iter.max, debug = 0, outer.max = 10)
{
 if (missing(iter.max)) {
      iter.max <- maxiter
   }
else maxiter <- iter.max
 list(iter.max = iter.max, rel.tolerance = rel.tolerance,
 toler.chol = toler.chol, debug = debug, maxiter = maxiter,
outer.max = outer.max)
}
<environment: namespace:survival>
```

Para esta función existen otras que permiten obtener valores predichos, como es la función *'Predict.survreg'* que proporciona un vector o una matriz con valores pronosticados. Y la función *'Resifuals.survreg'* devuelve un vector o matriz de residuos para objetos que provienen de *survreg*.

6. La función *coxph*

- **Definición**

  Permite ajustar modelos de regresión de Cox. Permite también ajustar modelos con variables dependientes del tiempo, modelos estraficados, modelos de múltiples eventos por individuo y otras extensiones derivadas del enfoque basado en los procesos de conteo de Andersen y Gill.

- **Sintaxis**

coxph(formula, data=, weights, subset,
na.action, init, control,
method=c("efron","breslow","exact"),
singular.ok=TRUE, robust=FALSE,
model=FALSE, x=FALSE, y=TRUE, ...)

- **Descripción de los argumentos**
  - *control.* El objeto coxph.control se emplea específicamente para la iteración de los límites y otras opciones de control.

  - *method.* Es una cadena de caracteres que especifica el método para tratar los empates. Si no hay tiempos de muerte iguales, todos los métodos son equivalentes. Casi todos los programas de regresión de Cox usan por defecto el método Breslow, pero éste no. El método exacto calcula la probabilidad exacta parcial, que es equivalente a un modelo logístico condicional.

  - *singular.ok*. Valor lógico que indica como manejar la matriz del modelo. Si es *TRUE*, el programa automáticamente saltará sobre las columnas de la matriz X que son las combinaciones lineales de las primeras columnas. En este caso, los coeficientes para tales columnas serán NA (valores perdidos) y la matriz para la varianza contendrá ceros.

  - *robust.* Si es TRUE, devolverá una estimación de la varianza robusta.

- **Ejemplo (6.1)**

coxph(Surv(futime,fustat)~age+resid.ds+rx+ecog.ps,data=ovarian)

Resultado

Call:
coxph(formula = Surv(futime, fustat) ~ age + resid.ds + rx +
ecog.ps, data = ovarian)

	coef	*exp(coef)*	*se(coef)*	*z*	*p*
age	0.125	1.133	0.0469	2.662	0.0078
resid.ds	0.826	2.285	0.7896	1.046	0.3000
rx	-0.914	0.401	0.6533	-1.400	0.1600
ecog.ps	0.336	1.400	0.6439	0.522	0.6000

Likelihood ratio test=17.0 *on* 4 *df, p*=0.00190 *n*= 26

Para el conjunto de datos *ovarian* la variable más influyente es la edad, como refleja su estadístico.

Una función particular asociada a '*coxph*' es la función *'predict.coxph'* que calcula valores estimados y términos de regresión para un modelo estimado de *coxph*. Un ejemplo para esta función es el siguiente:

- **Ejemplo (6.2)**

```
fit <- coxph(Surv(time, status) ~ age + ph.ecog + strata(inst), lung)
mresid <- lung$status - predict(fit, type='expected') #Martingale resid
predict(fit,type="lp")
predict(fit,type="risk")
predict(fit,type="expected")
predict(fit,type="terms")
predict(fit,type="lp",se.fit=TRUE)
predict(fit,type="risk",se.fit=TRUE)
predict(fit,type="expected",se.fit=TRUE)
predict(fit,type="terms",se.fit=TRUE)
```

7. La función *survfit.coxph*

- **Definición**

  Calcula la función de supervivencia predicha para un modelo de riesgos proporcionales de Cox.

- **Sintaxis**

survfit(formula, newdata,
se.fit=TRUE, conf.int=.95,
individual=FALSE,
type,vartype,
conf.type=c("log","log-log","plain","none"),...)

- **Descripción de los argumentos**
  - *formula*. Es un objeto *coxph*.

  - *newdata*. Los datos se enmarcan con los mismos nombres de las variables que aparecen en la fórmula *coxph*. La curva producida será representativa de una cohorte cuyas covariables corresponden a los valores en *newdata*. Por defecto es la media de las covariables usadas en el estimador *coxph*.

  - *individual*. Valor lógico que indica si cada fila de *newdata* representa a un individuo distinto (por defecto, FALSE), o si cada fila del marco de datos representa diferentes tiempos para sólo un individuo (TRUE).

  - *conf.int*. Nivel de confianza para un intervalo de confianza sobre la curva de supervivencia. Por defecto es 0.95.

  - *type, vartype.* Cadena de caracteres que especifica el tipo de curva de supervivencia. Los valores posibles son *"aalen"* o *"kaplan-meier"* (sólo los dos primeros carácteres son necesarios). Por defecto es *"aalen"*. Para versiones más recientes de *survfit*, para conseguir el estimador *"aalen"* se usa *type="tsiatis"*.

  - *conf.type*. Una de las opciones *"none"*, *"plain"*, *"log"* (por defecto), o *"log-log"*. La primera opción no genera intervalos de confianza. El segundo crea los intervalos estándar *curve* ± *k *se (curve)*, donde *k* es determinada por *conf.int*. La opción logarítmica calcula intervalos basados en el riesgo acumulativo o log(survival). La última opción está basada en intervalos para el logaritmo del riesgo o log (- log (survival)).

- **Resultados**

  El resultado es un objeto de clase *survfit*.

- **Ejemplo (7.1)**

```
#estima una función Kaplan-Meier y su gráfica
fit <- survfit(Surv(time, status) ~ x, data = aml)
plot(fit, lty = 2:3)
legend(100, .8, c("Maintained", "Nonmaintained"), lty = 2:3)
```

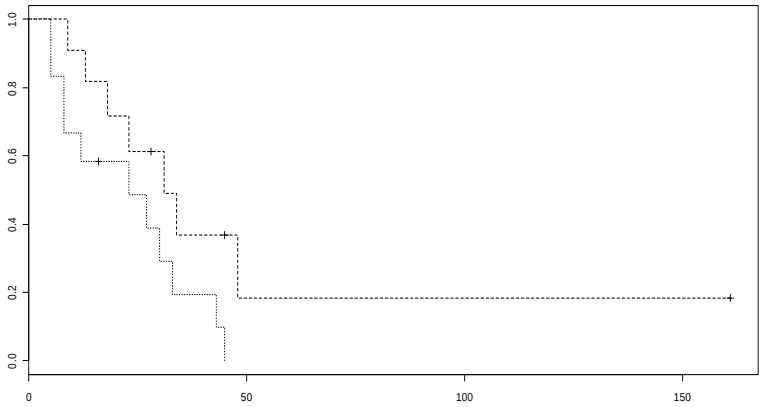

Figura 2.7. Gráfica de Kaplan-Meier del Ejemplo (7.1)

- **Ejemplo (7.2)**

```
#estima un modelo de riesgos proporcionales de Cox y su gráfica
#predice la supervivencia para una edad de 60 años
fit <- coxph(Surv(futime, fustat) ~ age, data = ovarian)
plot(survfit(fit, newdata=data.frame(age=60)),
     xscale=365.25, xlab = "Years", ylab="Survival")
```

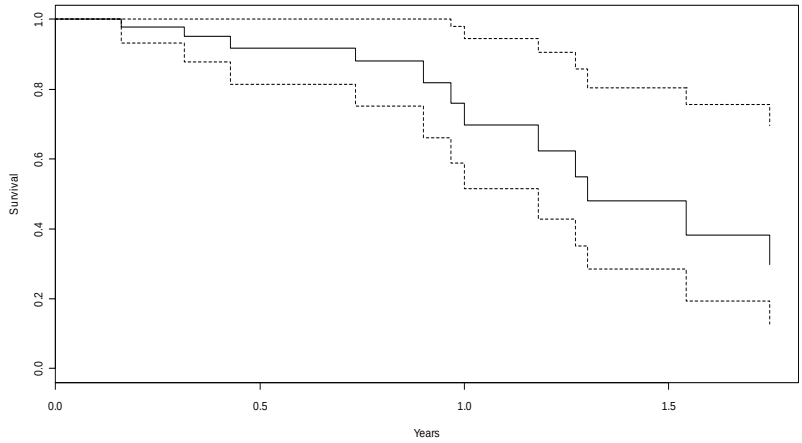

Figura 2.8. Curva de riesgos proporcionales de Cox del Ejemplo (7.2)

Este último gráfico, muestra como la supervivencia en individuos con cáncer de mama es mayor con uno de los tratamientos, mientras que para el otro tratamiento, la supervivencia disminuye a partir de los 60 años de un modo progresivo.

8. La función *basehaz*

- **Definición**

  Calcula la curva de supervivencia para un modelo de Cox.

- **Sintaxis**

  basehaz(fit, centered = TRUE)

- **Descripción de los argumentos**
  - *centered.* Si toma el valor TRUE, la curva que se obtiene es para un hipotético sujeto cuyos valores de la covariable son los correspondientes a la media de los datos originales. En otro caso vector de medias será cero.

- **Resultados**
  - *time.* El vector es clasificado para tiempos exactos (aquellos en los cuales ocurrió un acontecimiento).

  - *hazard*. Curva de la función de riesgo.

- *strata*. Si *fit* fue un modelo de Cox estratificado, se corresponderá con el estrato. Habrá una curva de supervivencia por estratos.

9. La función *residuals.coxph*

- **Definición**

  Calcula los diferentes residuos generados al ajustar un modelo de riesgos proporcionales de Cox

- **Sintaxis**

```
residuals(object,
        type=c("martingale", "deviance", "score", "schoenfeld",
        "dfbeta", "dfbetas", "scaledsch","partial"),
        collapse=FALSE, weighted=FALSE, ...)
```

- **Descripción de los argumentos**
  - *type*. Cadena de caracteres que indica el tipo de residuos deseados. Los valores posibles son *"martingale"*, *"deviance"*, *"score"*, *"schoenfeld"*, *"dfbeta"*, *" dfbetas"* y *"scaledsch"*. Sólo requiere una cadena determinada al ajuste requerido.

  - *collapse*. Vector que indica las filas. En modelos tiempo-dependientes más de una fila puede pertenecer a un solo individuo. Si hubiera 4 individuos representados por 3, 1, 2 y 4 filas de datos respectivamente, entonces *collapse=c (1,1,1, 2, 3,3, 4,4,4,4)* podría ser usado para obtener residuos por sujeto más bien que por observación.

- **Resultados**

  Para los residuos martingala y deviance, el objeto devuelto es un vector con un elemento para cada sujeto (sin *collapse*). Para los residuos de puntajes es una matriz con una fila por sujeto y una columna por variable. El orden de filas coincide con el orden en que los datos fueron introducidos en el ajuste original . Para los residuos de Schoenfeld, el objeto devuelto es una matriz con una fila para cada evento y una columna por variable.

Las filas son ordenadas por tiempo dentro de estratos, y el atributo *strata* contiene el número de observaciones en cada estratos.

- **Ejemplo (9)**

```
fit <- coxph(Surv(start, stop, event) ~ (age + surgery)* transplant,
             data=heart)
mresid <- resid(fit, collapse=heart$id)
```

Nota. el símbolo * significa que se está considerando covariables dependientes. Pero este tema queda fuera del estudio que hacemos aquí.

El conjunto de datos empleados en este ejemplo, llamado *'heart'*, hace referencia a la supervivencia en pacientes que pertenecen a una lista de espera para un programa de transplante de corazón de Stanford.

10. La función *cox.zph*

- **Definición**

  Test para contrastar la hipótesis de riesgos proporcionales después de ajustar un modelo de Cox (*coxph*).

- **Sintaxis**

```
cox.zph(fit, transform="km", global=TRUE)
```

- **Descripción de los argumentos**
  - *fit*. Representa el resultado para ajustar un modelo de regresión de Cox, usando la función de coxph.
  - *transform*. Representa una cadena de caracteres que especifica cómo los tiempos de supervivencia deberían ser transformadas antes de realizar el test. Los posibles valores que toma son: *"km", "rank", "identity"* o una función de un argumento.
  - *global. A*demás de los test por variable, proporciona el test chi-cuadrado

de forma global.

- **Resultados**
  - *table*. Una matriz con una fila para cada variable y opcionalmente, una última fila para la prueba global. Las columnas de la matriz contienen el coeficiente de correlación entre el tiempo de supervivencia transformado y los residuos de Schoenfeld, un coeficiente chi-cuadrado y el p-valor.

  - *x*. La transformación para el eje del tiempo.

  - *y*. Matriz de residuos de Schoenfeld. Habrá una columna por variable y una fila por suceso ocurrido. Las etiquetas de las filas contienen el número de acontecimientos originales (estos valores serán los mismos que en x).

  - *call*. Es la secuencia que se llama para la rutina. Los cálculos requieren la matriz original x del modelo de Cox ajustado. Ésto hace ahorrar tiempo, ya que si la opción x=TRUE es usada en *coxph*. Esta función por lo general sería seguida tanto por la gráfica como por el resultado obtenido. El gráfico da una estimación para el coeficiente tiempo-dependiente *beta(t)*. Si se asumen riesgos proporcionales, entonces *beta(t)* será una línea horizontal. La impresión del test viene dado por *slope=0.*

Los cálculos requieren la matriz original x del modelo de Cox ajustado, de modo que se ahorra tiempo si la opción x=TRUE es usada en *coxph*. A esta función debe seguir por lo general una representación gráfica y presentación de los resultados numéricos obtenidos. El gráfico da una estimación para el coeficiente tiempo-dependiente *beta(t)*. Si se asumen riesgos proporcionales, entonces *beta(t)* será una línea horizontal (pendiente 0). Los resultados numéricos se refieren al test para contrastar la hipótesis *slope=0.*

- **Ejemplo (10.1)**

```
fit <- coxph(Surv(futime, fustat) ~ age + ecog.ps,
             data=ovarian)
temp <- cox.zph(fit)
```

print(temp)

Resultado

	rho	*chisq*	*p*
age	-0.243	0.856	0.355
ecog.ps	0.520	2.545	0.111
GLOBAL	*NA*	3.195	0.202

plot(Temp.)

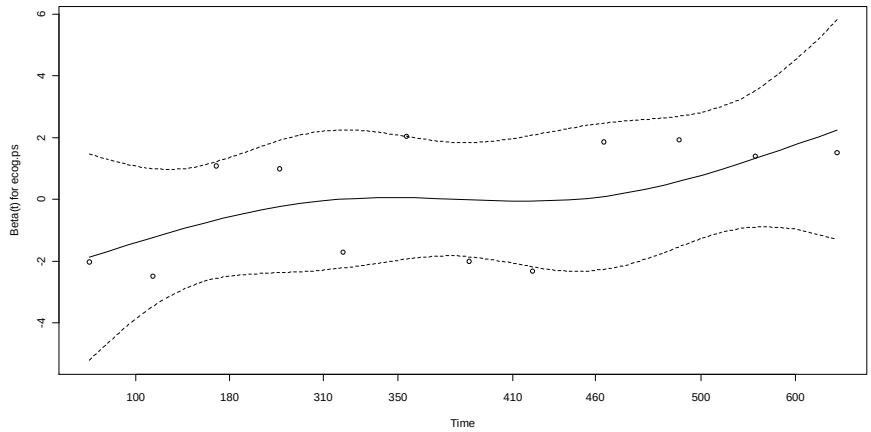

Figura 2.9. Función de riesgo de un modelo de regresión de Cox del Ejemplo (10.1)

Una función complementaria de '*cox.zph*' es la función *'Plot.cox.zph'* que muestra un gráfico de residuos de Schoenfeld escalado, además de una estimación lineal suave. Vemos a continuación un ejemplo para esta función, en el que se emplea el conjunto de datos *"veteran"* que registra las observaciones en dos tipos de tratamiento para el cáncer pulmonar.

- **Ejemplo (10.2)**

```
vfit <- coxph(Surv(time,status) ~ trt + factor(celltype) +
              karno + age, data=veteran, x=TRUE)
temp <- cox.zph(vfit)
plot(temp, var=5)
plot(temp[5])
abline(0, 0, lty=3)
```

```
abline(lm(temp$y[,5] ~ temp$x)$coefficients, lty=4, col=3)
title(main="VA Lung Study")
```

Resultado

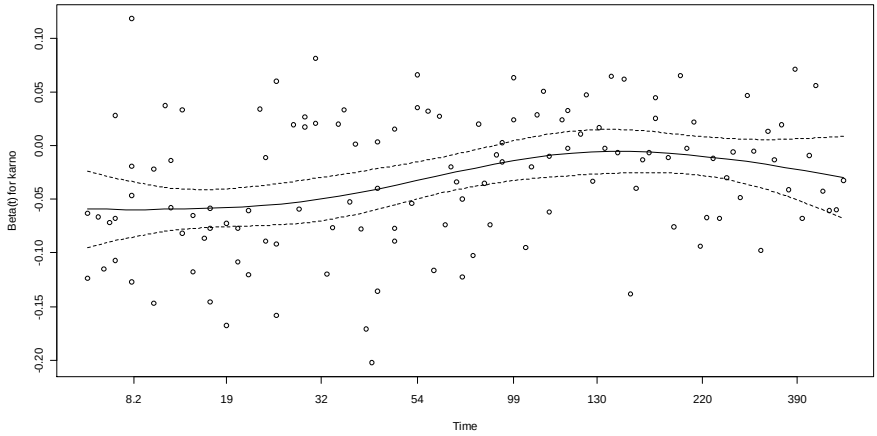

Figura 2.10. Gráfico de residuos de Schoenfeld

11. La función *strata*

- **Definición**

  Esta es una función especial usada en el contexto del modelo de supervivencia de Cox. Esta función identifica las variables de estratificación cuando ellas aparecen a la derecha de una fórmula.

- **Sintaxis**

  ```
  strata(..., na.group=FALSE, shortlabel=FALSE, sep=', ')
  ```

- **Descripción de los argumentos**
  - .... Cualquier número de variables. Todas deben ser de la misma longitud.
  - *na.group*. Variable lógica, que si es TRUE, entonces los valores perdidos son tratadas con un nivel distinto cada variable.
  - *shortlabel*. Si es TRUE, omite nombres de variables para pasar a etiquetas del factor.

- *sep.* Sirve para separar grupos al crear etiquetas.

- **Resultados**
  El resultado es idéntico al de la función *interaction*, pero para el etiquetado de los factores.

- **Ejemplo (11)**

```
a<-factor(rep(1:3,4))
b<-factor(rep(1:4,3))
levels(strata(a))
levels(strata(a,b,shortlabel=TRUE))
coxph(Surv(futime, fustat) ~ age + strata(rx), data=ovarian)
```

Resultado

```
Call:
coxph(formula = Surv(futime, fustat) ~ age + strata(rx), data = ovarian)

       coef  exp(coef)   se(coef)  z       p
age   0.137     1.15       0.0474  2.9     0.0038

Likelihood ratio test=12.7  on 1 df, p=0.000368  n= 26
```

Capítulo III

Aplicación real

1. Introducción

Hasta ahora, hemos estudiamos dentro del Capítulo I algunos de los modelos teóricos más importantes que se aplican en el Análisis de Fiabilidad o Análisis de Supervivencia para datos que presentan algún tipo de censura o truncamiento. En el segundo capítulo, se vieron las principales funciones de la librería *survival* del programa R para el análisis computacional de esos modelos teóricos.

En este tercer y último bloque, lo que pretendemos es ilustrar la teoría que hemos visto hasta ahora. Para ello, analizaremos los datos relativos a fallos registrados en una red de suministro de agua para evaluar la probabilidad de fallo de los tubos que componen la red y el impacto que algunos factores tienen sobre el riesgo de fallo de la infraestructura.

El caso que nosotros estudiaremos es el problema que provoca el paso del tiempo en infraestructuras de Sistemas de Suministro de Agua en una ciudad de tamaño medio española. La idea es obtener información y un registro del verdadero estado del deterioro de la red, y su fiabilidad mediante la modelización estadística. Con los

resultados obtenidos, se permitirá realizar un diseño, construcción y mantenimiento más eficiente de las redes de suministro de agua.

Para esto, se registraron los fallos en tramos individuales de dichas tuberías, y así, se pretende evaluar la probabilidad de fallo de los tubos. Además, se consideraron sólo los fallos en condiciones normales, excluyendo otras causas por acontecimientos anormales. Las observaciones se llevaron a cabo entre los años 2000 y 2005, aunque la vida de algunas de las tuberías comenzó tiempo atrás. El hecho de que los datos que se tienen sean para un periodo concreto y tan reducido, introduce la idea del truncamiento por la izquierda, ya que no se considera la información para fallos anteriores al año 2000; y censura a la derecha, para tubos que fallaron después de 2005.

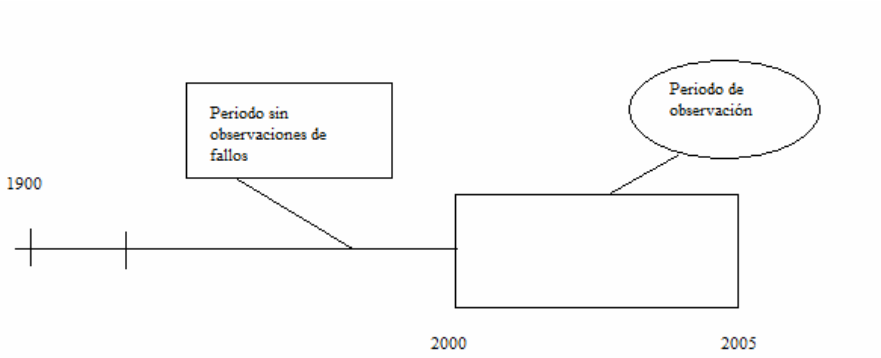

Figura 3.1. Duración del estudio

Como ya hemos explicado con anterioridad, la forma más común de truncamiento a la izquierda ocurre cuando los individuos o dispositivos entran en el estudio a edades aleatorias, esto es, el origen de los tiempos de vida precede al origen del estudio y después de esta entrada en estudio los dispositivos son seguidos hasta que el fallo ocurra o hasta que sean censurados a la derecha. De este modo, los dispositivos que fallan antes del tiempo de entrada al estudio no serán tenidos en cuenta por el investigador.

En el caso que a nosotros nos ocupa, consideraremos una observación como muestra la Figura 3.1, en la que se ve que la información disponible no es completa, y sólo son registrados fallos ocurridos en el periodo 2000-2005. Vemos el truncamiento por la izquierda para las tuberías que fallaron antes del 2000, y que por tanto, no son consideradas dentro de la muestra. Y por otro lado, las tuberías que fallaron después del 2005 consideran censura a la derecha. En esta situación, hemos observado un alto porcentaje de sujetos (por encima del 98%), lo que supone una fuerte censura que

limitará el procedimiento de estimación en este caso.

Definimos las siguientes cantidades:

- Tiempo de truncamiento a la izquierda: $X = max\ \{0,\ 2000 - fecha\ de\ instalación\}$. Si X = 0, indica que no hay truncamiento para ese sujeto, el tramo correspondiente se instaló después del año 2000 y por tanto se ha observado durante toda su vida hasta el fallo o hasta el momento de censura en el año 2005.

- Tiempo de fallo: $T = fecha\ del\ fallo\ registrado - fecha\ de\ instalación.$ Con función de distribución F.

- Tiempo de censura: $C = 2006 - fecha\ de\ instalación.$ Con función de distribución H.

X, T, C se suponen mutuamente independientes y no negativas. En el modelo con truncamiento a la izquierda y censura a la derecha, las observaciones consisten en el vector (X, Y, δ), donde $Y \geq X$, $Y = min\ \{T,C\}$ y $\delta = I\ (T \leq C)$ es un indicador de censura. En el caso $Y \leq X$, no hay observaciones.

El conjunto muestral será representado por $\{(x_i, y_i, \delta_i), i=1,2,\ldots,n\}$.

Los sujetos en estudio en este trabajo corresponden a tramos de tubería de distintas características. Los datos registrados, entre otros, son los correspondientes al año de instalación, longitud de sección, diámetro de sección, material del tubo, condiciones de tráfico y datos de fallos. Algunas características de la muestra registrada se presentan en la Tabla 3.1.

Dato geográfico	Ciudad de la Costa Mediterránea española de 330000 habitantes
Longitud total	1028.8 km
Unidad experimental	Sección de tubo
Número de registros (secciones de tubo)	32387
Periodo de instalación de tubos	1890-2005
Registro de fallos	2000-2005
Número de fallos registrados	1487

Tabla 3.1. Descripción de la muestra

Como hemos dicho antes, el interés de este estudio se centra en analizar las roturas registradas en una red de suministro de agua para evaluar el estado actual de la red y pronosticar su futura degradación. Por lo que se usarán herramientas cuantitativas en el mantenimiento del Sistema de Suministro de Agua para evaluar el estado actual y pronosticar la futura degradación de las infraestructuras. Por este motivo, emplearemos aproximaciones no paramétricas y semiparamétricas, que permiten identificar la influencia de factores en los fallos de los tubos.

2. Estimación de la Función de Fiabilidad

Vamos a considerar en primer lugar el problema de obtener un estimador de la función de fiabilidad o supervivencia $S(t) = \mathrm{P}\{T > t\}$ siguiendo las pautas indicadas en el apartado 6 del Capítulo I de este trabajo.

La función *survfit* aplicada a nuestros datos teniendo en cuenta el truncamiento por la izquierda, e ignorando los distintos grupos de riesgo que vendrían determinados por las distintas covariables que se especifican en los siguientes apartados, produce los resultados presentados en la siguiente figura.

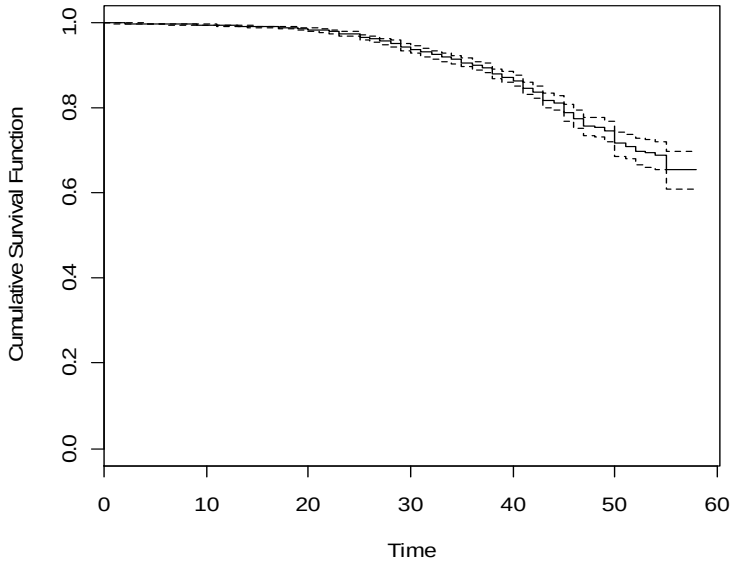

Figura 3.2. Estimador de la función de supervivencia.

En la Figura 3.3 se incluyen estimaciones de la función de supervivencia, en primer lugar, ignorando el truncamiento presente en la muestra (curva superior) y a continuación considerando el método de Nelson-Aalen extendido para estimar la función de riesgo acumulada y obtener el correspondiente estimador de la función de supervivencia (el mismo de la Figura 3.2).

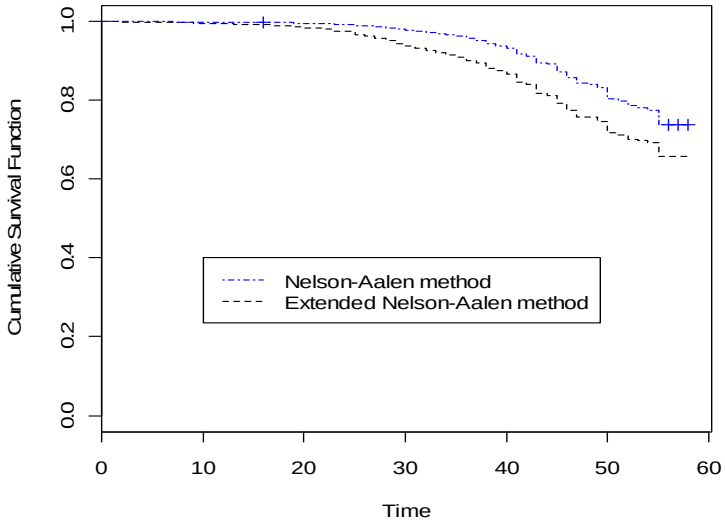

Figura 3.3. Estimación de la supervivencia

La presencia de fuerte censura en los datos es la razón por la que la función de supervivencia no puede ser estimada a partir del 70%. En la Figura 3.2 comparamos las probabilidades de supervivencia estimadas si no se tiene en cuenta el truncamiento y vemos que existe un problema de sobreestimación si sólo consideramos la censura a la derecha en los datos.

El error de sobreestimación cometido con el estimador de Nelson-Aalen (el estimador de Kaplan-Meier reproduce este error en la misma magnitud) puede deberse al hecho de no considerar el truncamiento a la izquierda de la muestra, es decir, no tiene en cuenta que muchos tubos podrían haber fallado antes del año 2000. En este estudio, un fallo producido antes del año 2000 no es observado, pero esto no significa que no haya ocurrido. El estimador de Kaplan-Meier asume que no ocurre ningún fallo durante el periodo de truncamiento, por lo que se asume que el tiempo de supervivencia es mayor que el de truncamiento. Y por tanto, los fallos que se producen en el truncamiento no están registrados. Así que el estimador de Nelson-Aalen extendido refleja mejor la situación real de los datos.

En la Figura 3.3 se muestran los resultados para la función de riesgo y la función de riesgo acumulada basada en la vida de la cañería. En ambos casos se tiene en cuenta el fenómeno del truncamiento y se han usado funciones de R que han sido modificadas para atender a este hecho.

En el gráfico de la izquierda vemos como es insignificante el riesgo de fallo al principio de la vida de las cañerías (los 20 primeros años) y ese riesgo aumenta con el paso de los años. Con el gráfico de la derecha se puede confirmar que el número de

fallos se acumula con el paso de los años.

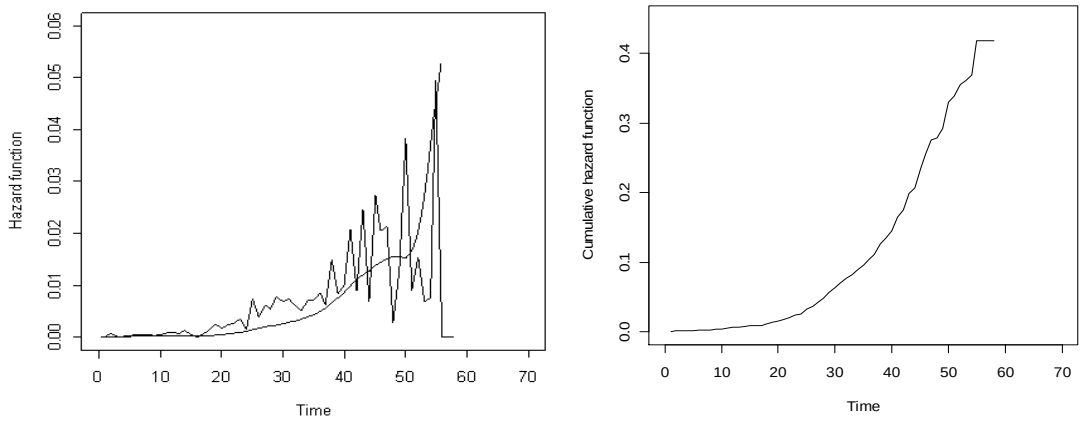

Figura 3.4. Estimación función de riesgo y función de riesgo acumulada

3. Estudio de las Covariables. Modelos Estratificados

Se consideran diversos factores que afectan potencialmente a las transformaciones de una Red de Suministro de Agua tal y como se describe a continuación. Específicamente, se establecen grupos de riesgo de acuerdo con las características físicas del agua, como son el *material*, la *longitud* (en metros) y el *diámetro* (en milímetros) de la sección de tubo, y las *condiciones del tráfico rodado* que afectan a la sección de tubo. Los resultados se presentan en el siguiente apartado.

3.1. Materiales

De acuerdo con la base de datos, se han usado 4 tipos diferentes de materiales. Aunque hay que tener en cuenta que hay una secuencia en tiempo con el uso de los materiales. Por ejemplo, el cemento *Asbetos* está actualmente fuera de uso, después de que fuese el material más usado en las décadas de los años 50 y 60.

Respecto al tipo de material del que están hechos los tubos, se consideran 4 categorías: Ductile Cast Iron (DI), Gray Cast Iron (CI), Polyethylene (PE) y Asbestos Cement (AC). De acuerdo a la Figura 3.4, el material que presenta menor tendencia de fallo es DI. Los tubos fabricados con DI tienen aproximadamente un 80% de supervivencia a los 55 años.

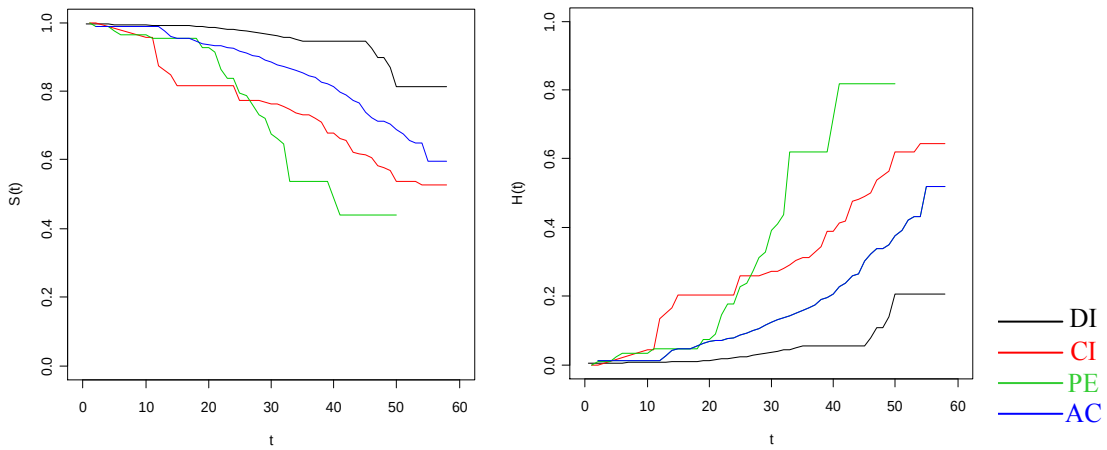

Figura 3.5. Función de supervivencia para grupos de riesgo por material (izquierda). Función de riesgo acumulada para grupos de riesgo por material (derecha).

Por otro lado, el tipo de material más vulnerable a fallos tempranos (antes de los 23 años) es CI, mientras que después de 30 años el polietileno es más sensible a roturas.

3.2. Longitud (en metros, m)

Esta variable se caracteriza de acuerdo a la siguiente clasificación: tubos de menos de 2 m de longitud; tubos de longitud entre [2, 10); tubos con valores de longitud comprendidos entre [10, 50); y tubos de más de 50 m de longitud.

Como se puede apreciar en la Figura 3.5, las secciones de tubo más largas tienen mayor tendencia a sufrir roturas tempranas. Por otra parte, las secciones de tubo más cortas proporcionan mayor resistencia a fallos, con lo cual, se podría decir que la longitud causa un efecto negativo respecto al riesgo de fallo.

Las diferencias entre las longitudes comienzan a ser relevantes a partir de los 20 años. Los tubos con menos de 2 m de longitud tienen una probabilidad igual a 1 de supervivencia a los 55 años, mientras que esta probabilidad decrece hasta el 40% para la clase de tubos con longitud mayor que 50 m.

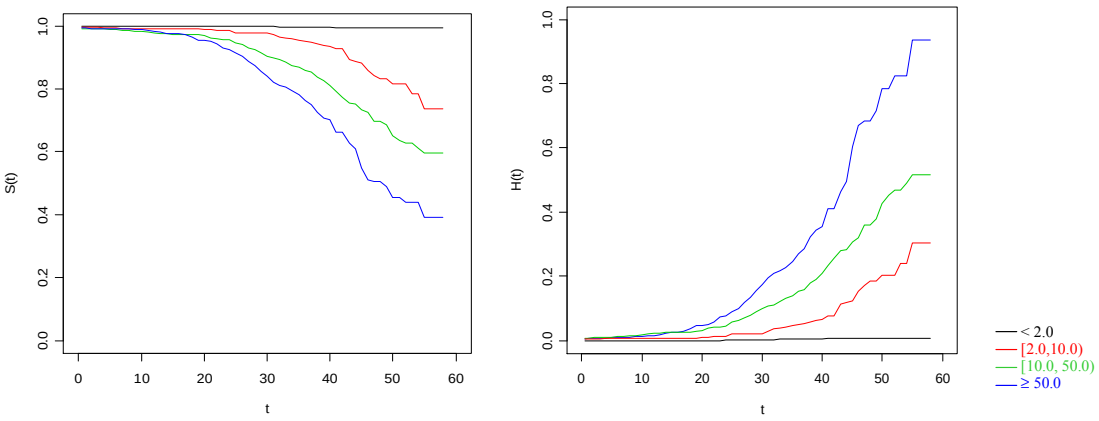

Figura 3.6. Función de supervivencia para grupos de riesgo por longitud (izquierda). Función de riesgo acumulada para grupos de riesgo por longitud (derecha).

3.3. Diámetro (en milímetros, mm)

El análisis de la situación para la variable *diámetro* es similar al caso de la variable *longitud*. Los tubos se agrupan de acuerdo a diferentes magnitudes de diámetro. Entonces, los 4 grupos de riesgo se establecen del siguiente modo: los tubos con menos de 90 mm de diámetro; los que tienen un diámetro comprendido entre (90, 175]; tubos con diámetro dentro del intervalo (175, 300]; y tubos de más de 300 mm de diámetro.

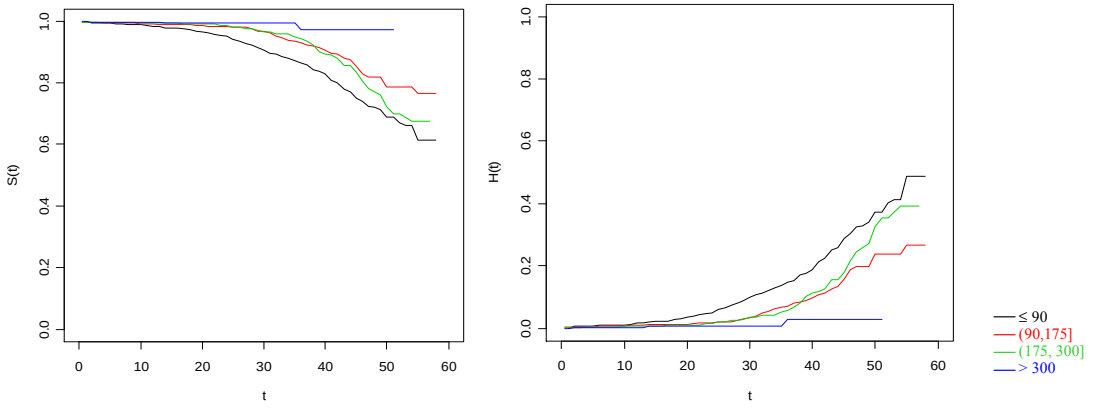

Figura 3.7. Función de supervivencia para grupos de riesgo por diámetro (izquierda). Función de riesgo acumulada para grupos de riesgo por diámetro (derecha).

Observando la Figura 3.7, vemos que la variable *diámetro* causa un efecto pasitivo con respecto al riesgo de rotura en el sentido de que un tubo más ancho tiene una probabilidad más alta de supervivencia. Como ejemplo, se puede ver en los gráficos

anteriores, que los tubos más anchos de 300 mm tienen una probabilidad de supervivencia de más de 50 años de un 95%, mientras que para los tubos con menos de 90 mm de diámetro, está probabilidad decrece por debajo del 60%.

4. Modelo de Cox

Una vez vistos los anteriores gráficos, parece razonable, para estimar el efecto que las covariables tienen sobre la razón de fallo de un tubo, suponer para nuestros datos un modelo de riesgos proporcionales de Cox (RPC). La hipótesis clave en el modelo de riesgos proporcionales de Cox es la proporcionalidad de la razón de riesgo, es decir, el cociente de razones de riesgo de dos sujetos descritos por dos conjuntos de covariables diferentes es constante en el tiempo. Esto podría significar, entre otras cosas, que las curvas de supervivencia correspondientes a diferentes grupos de riesgo no se cortan. Si miramos los gráficos anteriores, esta afirmación puede mantenerse, excepto para la covariable que indica los distintos tipos de material.

A continuación, estimamos el modelo para nuestros datos. Esto es, consideramos variables artificiales definiendo los niveles de las covariables: *materiales* y *tráfico rodado*. De forma que, estimaremos un modelo con el siguiente conjunto de variables explicativas:

Z_1 Longitud (m)
Z_2 Diámetro (mm)
Z_3 Tipo de material (1-hierro dúctil; 0-otros)
Z_4 Tipo de material (1-hierro colado; 0-otros)
Z_5 Tipo de material (1-polietileno; 0-otros)
Z_6 Nivel de presión de tráfico rodado (1-acera; 0-otros)
Z_7 Nivel de presión de tráfico rodado (1-normal; 0-otros)

El modelo semiparamétrico para el caso de datos truncados a la izquierda, debe ser formulado del siguiente modo

$$r(t \mid Z, T > X) = r_0(t \mid T > X) \exp\left(\sum_{k=1}^{7} \beta_k Z_k \right)$$

donde la parte izquierda de la función denota la velocidad de degradación en el tiempo *t* para una sección de tuberías con un vector de covariables $Z=(Z_1, Z_2,\ldots, Z_7)'$, r_0 es una

curva de riesgo base arbitraria y $\beta=(\beta_1, \beta_2,\ldots, \beta_7)'$ es un vector de parámetros tal que β_k mide el efecto de la covariable k sobre la probabilidad instantánea de fallo. Cuando este valor estimado es 0, se deduce que la correspondiente covariable no tiene efecto, y en consecuencia debe ser eliminada del modelo.

Para estimar el modelo usamos el método de la probabilidad parcial adaptada a datos truncados a la izquierda del mismo modo que en se explicó en la sección correspondiente del Bloque I para datos en los que no hay el condicionamiento introducido por el truncamiento. De este modo, los coeficientes de regresión son estimados modificando las probabilidades parciales que explican el retraso en la entrada del grupo de riesgo. Esto es, construimos la función de verosimilitud parcial

$$L(\beta)=\prod_{i=1}^{D}\frac{\exp\left[\sum_{k=1}^{m}\beta_k Z_{i,k}\right]}{\sum_{j\in r_i}\exp\left[\sum_{k=1}^{m}\beta_k Z_{i,k}\right]}$$

donde $r_i=\sum_{j=1}^{m}1_{\{x_j\le y_i\le y_j\}}$ es el tamaño del conjunto de sujetos en riesgo en el instante y_i, y D es el número de tiempos de fallo distintos. Para estimar el modelo tenemos que adaptar algunas funciones que el programa R incorpora dentro de la librería *survival*.

4.1. Construcción del Modelo

La primera aproximación que desarrollaremos considera todo el grupo de covariables que fueron dadas anteriormente. Hemos realizado una serie de análisis que no se muestran aquí sobre los tests de significación de los coeficientes involucrados y la valoración de diferentes plots de residuos que nos han llevado a considerar la versión final del modelo cuyos resultados mostramos en la tabla siguiente.

	coef	exp(coef)	se(coef)	z	p
$Z'_1=\log(Z_1)$	0.532	1.702	0.0361	14.71	0.0e+00
$Z'_2=(Z_2/100)\text{^}2$	-0.065	0.937	0.0184	-3.52	4.2e-04
$Z'_3=Z_4$	0.473	1.605	0.1720	2.75	6.0e-03
$Z'_4=Z_5$	0.759	2.136	0.2109	3.60	3.2e-04
$Z'_5=Z_4\times Z_8$	1.453	4.276	0.5577	2.61	9.2e-03
$Z'_6=Z_3\times Z_8$	-1.563	0.210	0.2763	-5.66	1.5e-08

Tabla 3.2. Estimaciones del modelo de riesgos proporcionales de Cox

Vamos a explicar el resultado de la Tabla 3.3 examinando cada variable una a una.

4.1.1. Tráfico rodado (Z_6, Z_7)

En primer lugar, tenemos que considerar el factor de riesgo especificado por los niveles de estrés del tráfico rodado como un factor de estratificación. Hemos comprobado que el modelo estratificado es más adecuado que un Modelo de Cox donde el tráfico rodado se incorpora por medio de variables *dummy* (variables 0-1 que se introducen para representar los diferentes niveles de tráfico en la zona), donde no todos los factores se ajustan bien en el modelo.

4.1.2. Longitud (Z_1)

La longitud es la variable más importante como queda reflejado con su estadístico (ver Tabla 3.2), que ha sido introducido usando una transformación logarítmica debido a su asimetría (rango 0’02559-1869’67544, mediana = 7’50058 y media = 28033404). Desde este punto de vista, el riesgo relativo para una sección de tubería de longitud l_1 teniendo un fallo comparada con una sección de tubo de longitud l_2 viene dada por: $\left(\frac{l_1}{l_2}\right)^{0'532}$.

Por ejemplo, el primer y el tercer cuartil de la variable longitud son 0.5 y 33.98273 respectivamente, lo que se traduce en un riesgo relativo de (33.98273/0.5)^0.532 = 9.43577 para una sección de tubería promedio (en el resto de covariables) en la cola superior frente a una sección de tubería promedio en la cola inferior de la distribución de longitud.

También hemos estudiado los residuos martingala para comprobar que una transformación logarítmica de esta variable parece conveniente. Los resultados se pueden ver en la Figura 3.8. Se incluye un ajuste suavizado de los residuos, que da una idea de la transformación de la covariable en caso de que no se sostiene linealmente.

En nuestro caso, la linealidad del gráfico indica que la transformación logarítmica de la covariable *Longitud* es apropiada.

4.1.3. Diámetro (Z_2)

Los valores de la muestra para esta covariable han sido reescalados dividiendo entre 100. A continuación, examinamos los residuos martingala de este factor para determinar la forma funcional con la cual este factor debería ser presentado en el modelo. Después de algunos exámenes, concluimos que la mejor forma funcional para este factor es una transformación cuadrática (Figura 3.9), así que consideramos la covariable definida

como $Z_2^{'} = \left(\frac{Z_2}{100}\right)^2$. Esta transformación está relacionada con las dimensiones de una sección de tubo particular. Específicamente, prácticamente refleja el área de un corte transversal de tubo.

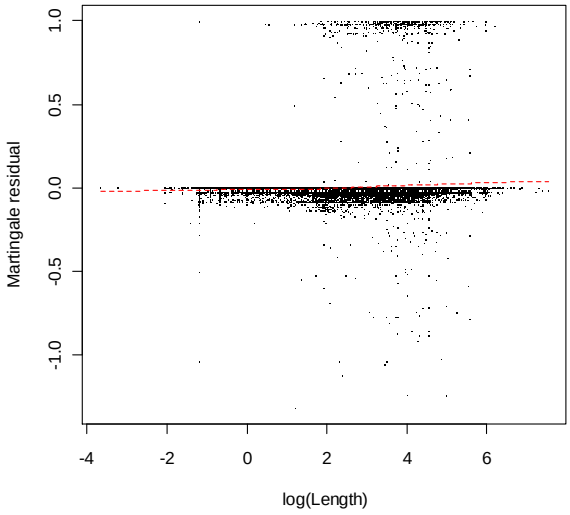

Figura 3.8. Residuos martingala para la covariable *log(longitud)*

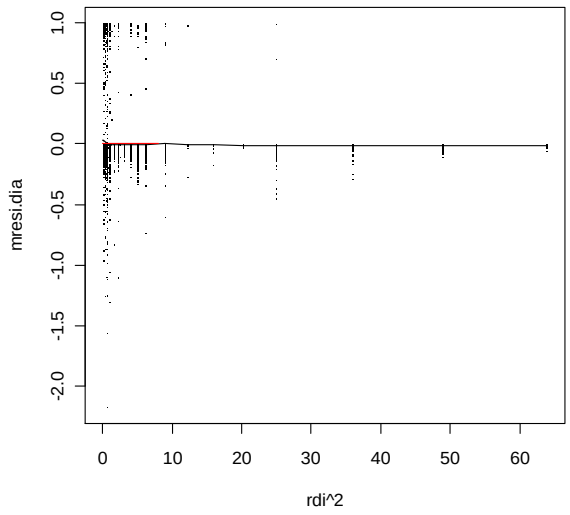

Figura 3. 9 Residuos martingala para la covariable Diámetro

4.1.4. Material (Z_3, Z_4, Z_5)

La Figura 3.10 sugiere que hay una componente de tiempo en la covariable material, de forma que los tubos hechos de *Polyethylene* (PE) o *Ductile Iron* (DI) que han sido instalados recientemente (la mayoría de ellos han sido instalados en los últimos 25 años) mientras que los tubos hechos de *Cast Iron* (CI) o *Asbestos Cement* (AC) fueron fabricados en años anteriores, de modo que la sección de tubo más vieja de la muestra sea hecha de este tipo de material. Así hemos definido una variable dependiente del tiempo de la siguiente forma:

$$Z_8 = \begin{cases} 1, & \text{si la instalación es posterior a 1980 (la edad de entrada en el estudio está en la mayoría en 20 años)} \\ 0, & \text{en otro caso} \end{cases}$$

Entonces, para mostrar la dependencia del tiempo del factor material construimos un modelo en el cual las interacciones entre las variables Z_3, Z_4, Z_5 y Z_8 han sido tenidas en cuenta, revelando que sólo los términos incluidos en la Tabla 3.2 eran significativos, con lo que es $Z'_3 = Z_4$, $Z'_5 = Z_4 * Z_8$ y $Z'_6 = Z_3 * Z_8$.

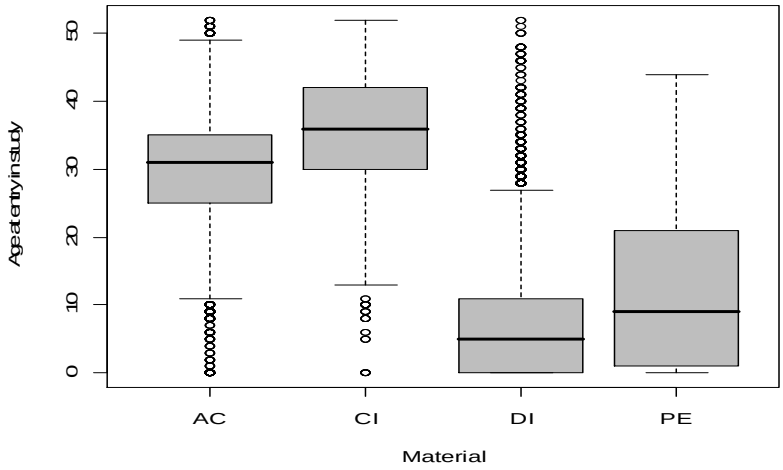

Figura 3.10. Gráfico de cajas para los grupos de materiales

4.2. Estudio de la bondad de ajuste: Residuos de Cox-Snell

Después del ajuste del modelo, hemos calculado los residuos de Cox-Snell para evaluar de una manera global el ajuste del modelo de riesgos proporcionales. Si el modelo es correcto y las estimaciones de β se aproximan a los verdaderos valores, entonces estos residuos deberían parecerse a una muestra censurada para una distribución exponencial de parámetro la unidad.

Hemos calculado el estimador de Nelson-Aalen para la función de riesgo acumulado de los residuos de Cox-Snell. Si la distribución exponencial se ajusta a los datos, entonces, este estimador debería describir aproximadamente una línea recta con pendiente igual a 1. La gráfica representada en la Figura 3.11 sugiere que este modelo es aceptable excepto en la cola de la derecha donde las estimaciones son muy inestables entre otras cosas por la fuerte censura que presentan los datos (sobre el 98%).

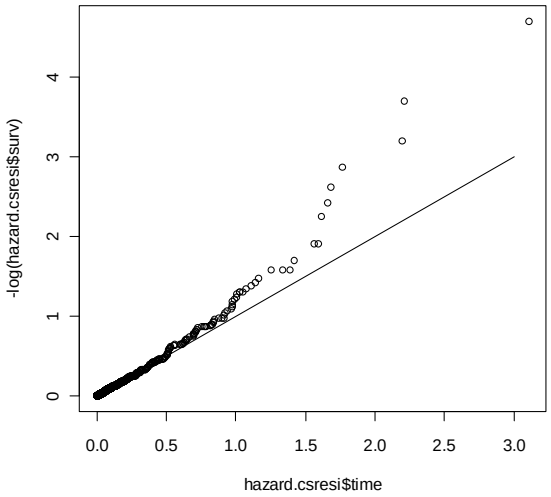

Figura 3.21. Residuos de Cox-Snell

4.3. Comprobación de la hipótesis de Riesgos Proporcionales en las covariables: Residuos de Shoenfeld Escalados

Ahora nos interesaremos en evaluar el supuesto de riesgos proporcionales de Cox examinando si el impacto de una o varias covariables sobre el riesgo de fallo puede variar con el tiempo. Por ejemplo, una tubería de agua hecha con un tipo particular de material puede degradarse con el tiempo en el sentido que el correspondiente coeficiente β puede no ser constante, esto es $\beta(t)$. Si, por el contrario, si la hipótesis de riesgos proporcionales es admisible, una representación gráfica de $\beta(t)$ frente al tiempo describirá una línea horizontal. La Figura 3.12 presenta los gráficos para cada covariable donde se puede apreciar que ningún coeficiente es dependiente del tiempo en nuestro caso.

El estudio también puede ser abordado numéricamente. Como alternativa a la hipótesis de riesgos proporcionales, Therneau y Gramsch (2000) consideran que los coeficientes de regresión están dados por funciones de tiempo de la forma $\beta(t) = \beta + \theta\, g(t)$, siendo $g(t)$ una función suave. La Tabla 3.3 muestra los resultados de los tests que

contrastan la hipótesis nula H_0: θ=0, es decir, el modelo de riesgos proporcionales es correcto. Los *p*-valores en la tercera columna indican que la hipótesis nula puede ser aceptada con bastante contundencia en todos los casos.

	θ	Chisq	p
$Z'_1 = \log(Z_1)$	-0.02468	0.20095	0.654
$Z'_2 = (Z_2/100)^2$	0.01347	0.08234	0.774
$Z'_3 = Z_4$	-0.03242	0.55276	0.457
$Z'_4 = Z_5$	-0.00377	0.00835	0.927
$Z'_5(t) = Z_4 \times Z_8(t)$	-0.01924	0.14671	0.702
$Z'_6(t) = Z_3 \times Z_8(t)$	-0.01297	0.09284	0.761
GLOBAL	NA	1.11896	0.981

Tabla 3.3. Residuos de Schoenfeld escalados

4.2. Comprobación de la existencia de datos outlers: Residuos dfbeta y de desvíos

Ahora exploramos los residuos para determinar la influencia de cada observación sobre el modelo estimado. Calculamos, mediante los residuos *dfbeta*, implementados en el programa R, el cambio aproximado en el coeficiente *k*-ésimo (es decir, la *k*-ésima covariable) si la observación *i*-ésima hubiera sido eliminada del conjunto de datos y el modelo es estimado de nuevo sin aquella observación. Para cada covariable, hemos representado las observaciones (en el orden de los tiempos de fallo registrados) frente al cambio escalado (dividiendo por el error estándar de cada coeficiente) en el coeficiente después de eliminar dicha observación del modelo. Si la eliminación de una observación implica un incremento en el coeficiente, el residuo *dfbeta* es negativo y viceversa.

La Figura 3.13 muestra los residuos *dfbeta* escalados para el modelo. Si consideramos la escala para el eje Y en los diferentes gráficos, se puede apreciar que, excepto para las covariables Z'_2 y Z'_5, ninguna de las observaciones ejerce un cambio mayor a aproximadamente un 20% del error estándar. Si se comparan con el valor del coeficiente de regresión estimado y su error estándar podemos concluir que ninguna de las observaciones es influyente individualmente.

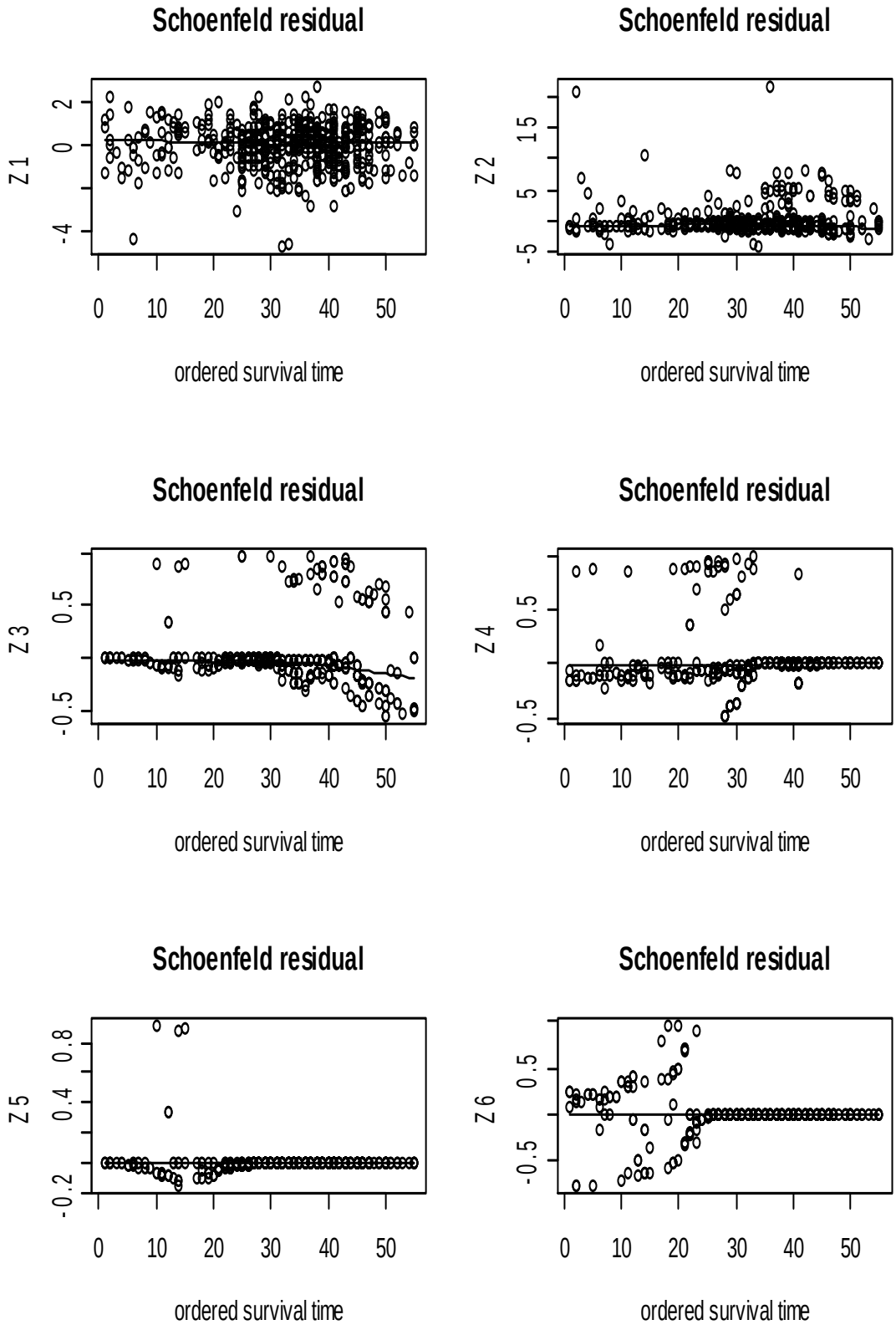

Figura 3.12. Residuos de Schoenfeld

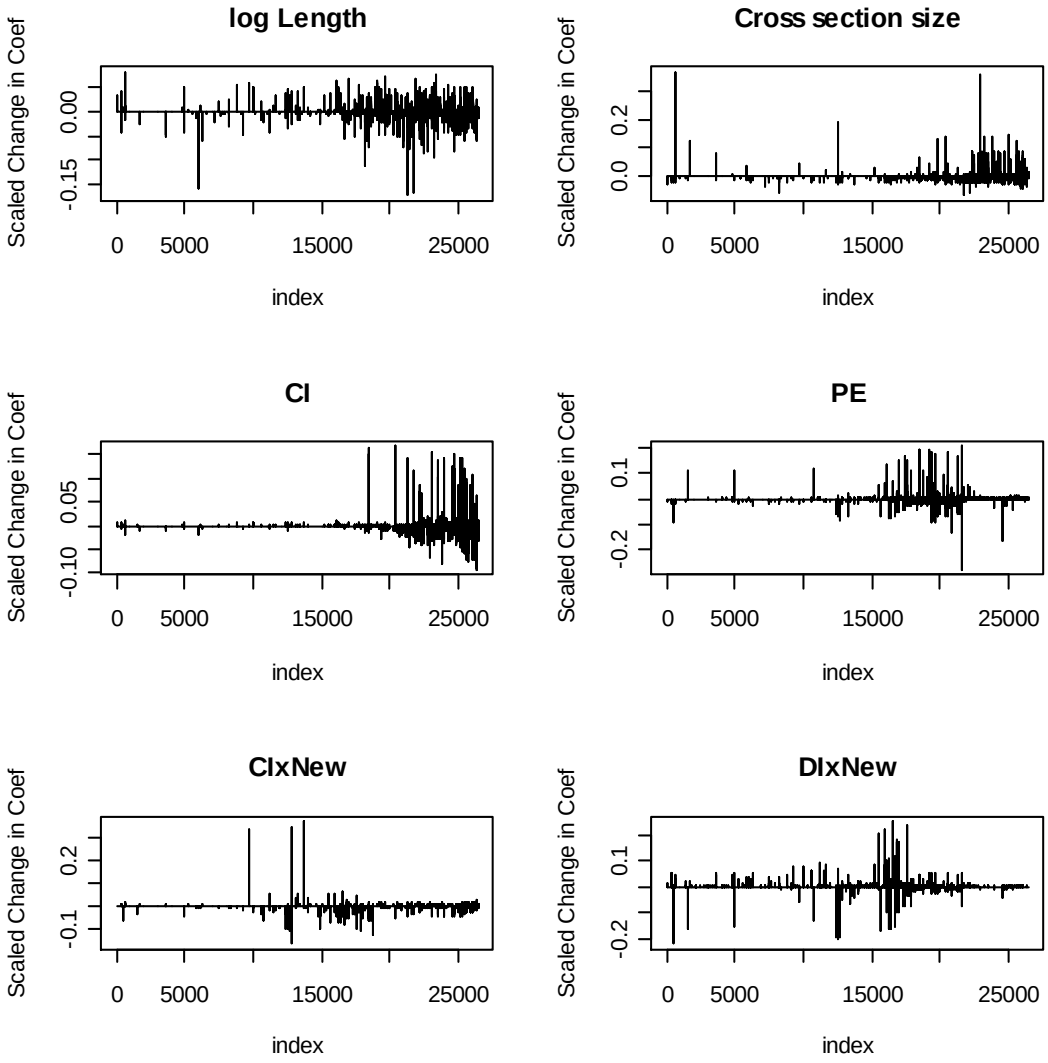

Figura 3.13. Residuos *dfbeta*

Con respecto a la covariable Z'_2, hay dos observaciones con *dfbeta* mayor que el 30%, en concreto el registro correspondiente al sujeto 641 da un residuo *dfbeta* igual a 0.3685178 y el registro número 22950 presenta un residuo *dfbeta* igual a 0.3580343. La Tabla 3.4 muestra la información registrada para estos dos casos.

Indice	Edad a la entrada	Edad en el fallo	δ	Long. (*m*)	Diam. (*mm*)	Material	Trafico
641	0	2	1	161.4118	500	DI	Pesado
22950	32	36	1	395.8230	500	AC	Pesado

Tabla 3.4. Observación con un alto residuo *dfbeta*

Hemos investigado también los residuos de desvíos para establecer si estos dos registros presentan un valor extremo también en este sentido. Los valores obtenidos para los residuos de desvíos son 2.987455 y 1.7449 respectivamente. El residuo correspondiente al caso 641 es bastante alto como para deducir que podría ser considerado un *outlier*, esto es, un caso mal predicho, lo que indica una sección de tubo que falló demasiado pronto en comparación con la predicción dada por el modelo estimado. Según los valores de sus covariables, el valor correspondiente de diámetro grande pone esta observación en un riesgo bajo, aunque falló muy pronto.

Para el caso número 22950, la magnitud grande de longitud (que afecta negativamente al riesgo de fallo) compensa el valor del diámetro en cuanto al valor predicho, esto podría explicar el menor valor de residuo de desvío para esta observación.

Finalmente, la covariable Z'_5 está asociada a una estimación del coeficiente de regresión de 4.276 (0.5577). Este valor es bastante alto comparado con los residuos *dfbeta*. Así que podemos incluir que ninguna de las observaciones es influyente en gran medida.

Para terminar, en la Figura 3.13 observamos un gráfico donde se representa los residuos de desvíos, para detectar valores *outliers*. Si se observa un diagrama de puntos alrededor del 0 sin patrón se puede concluir un buen ajuste del modelo. Se podría considerar que los residuos que se exceden la magnitud en 3 aproximadamente pertenecen a individuos mal predichos. Observemos 33 puntos en el gráfico que están colocados encima de la línea horizontal $y = 3$. Ninguno de estos puntos está asociado a observaciones influyentes. A pesar de esto, después de ajustar una línea a los residuos, se puede detectar una leve tendencia, indicando que algunos casos pueden alcanzar tiempos de fallo demasiado pronto comparado con lo estimado por el modelo.

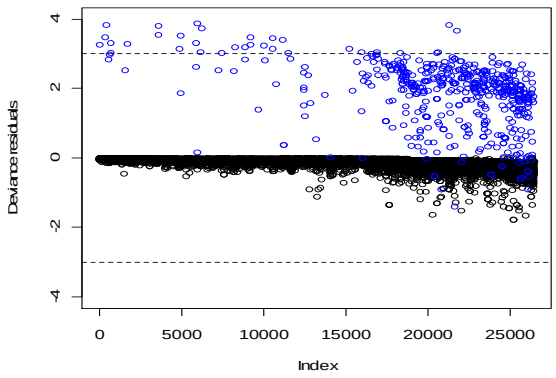

Figura 3.3. Residuos deviance. Fallos observados (en azul)

5. Conclusiones

Hemos explorado las propiedades de fiabilidad de una red de suministro de agua instalada en una ciudad de tamaño medio de la costa mediterránea española. El estudio es válido para cualquier otro sistema de suministro de agua de características similares, y el objetivo principal es el de usar instrumentos cuantitativos en el mantenimiento de estos sistemas para evaluar su estado actual, así como pronosticar el futuro de las infraestructuras.

Hemos usado métodos no paramétricos y semiparamétricos que han sido adaptados a las particulares características de la muestra que estamos tratando. En particular, la muestra que usamos en nuestro estudio está caracterizada por la presencia de truncamiento a la izquierda y censura a la derecha.

Por un lado, hemos obtenido una estimación no paramétrica de las propiedades de fiabilidad del sistema de cañerías y, por otro lado, hemos estimado el impacto de algunos factores sobre el riesgo de fallo de un tubo particular. Nuestro análisis mostró que los factores que afectan la supervivencia en tubos son algunas características del tubo (longitud, diámetro, material) y tráfico rodado. Los tubos que eran más propensos al fallo tenían las características siguientes: una longitud larga, un diámetro pequeño y un material de hierro (teniendo en cuenta que este material fue instalado después de 1980), y son colocados bajo acera.

Se han realizado algunos análisis de residuos para validar el modelo. Primero, una evaluación global de la hipótesis de riesgos proporcionales usando los residuos de Cox-Snell, lo que mostraba un buen ajuste, excepto para el caso de datos en la cola de la derecha donde las estimaciones son bastantes inestables entre otras cosas debido a la presencia de fuerte censura en el conjunto de datos.

A continuación tratamos la evaluación de la hipótesis de riesgos proporcionales de Cox por covariables, examinando si el impacto de cada covariable sobre el riesgo puede variar con el tiempo, usando los residuos de Schoenfeld. Ninguna covariable dependiente del tiempo ha sido descubierta en nuestro caso. Finalmente, se ha comprobado la presencia de valores influyentes, y sólo para la covariable Z'_2 se han encontrado dos valores influyentes.

Bibliografía

[1] Aalen, O. (1978). Nonparametric estimation of parcial transition probabilities in multiple decrement models.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. y Keiding, N. (1993) *Statistical models based on counting processes.* Springer-Verlag, New York.

[3] Borgan, Ø. (1997). *Three contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aalen-Johansen*. Universidad de Oslo.

[4] Borges Peña, R.E. (2005). *Análisis de Sobrevivencia utilizando el lenguaje R.* Universidad de Los Andes.

[5] Breslow, N.(1972), "Discussion on Professor Cox's paper", *J. Roy. Statist. Soc. A*, vol. 34, pg. 2160–2170

[6] Carrión, A., Debon, A., Cabrera, E., Gámiz, M.L. y Solano, H. (2008). "Failure risk análisis in water supply Networks". ESREL 2008.

[7] Carrión, A., Solano, H., Gámiz, M.L. y Debon, A., (2010). "Evaluation of the reliability of a water supply network from right-censored and left-truncated

break data", *Water Resources Management*, 24 (12), 2917-2935.

[8] Cox, D.R. y Oakes, D. (1984). *Analysis of survival data*. Chapman&Hall.

[9] Crawley, M.J. (2007). *The R book*. John Wiley & Sons Inc.

[10] Crowley, J. y Hu, M. (1977), Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72**, 27–36.

[11] Fleming, T.R. y Harrington D.P. (1984), "Nonparametric estimation of the survival distribution in censored data", *Comm. Statist.–Theor. Meth.*, vol. 20, 2469–2486.

[12] Fox, J. (2008), "Cox proportional-hazards regression for survival data". Appendix to *An R and S-PLUS Companion to Applied Regression.* Sage Publications, 2008.

[13] Gámiz Pérez, M.L. (2000). *Introducción a los procesos estocásticos. Cadenas de Markov y Procesos de Renovación.* Universidad de Granada.

[14] Iglesias Pérez, M.C. (2003). *Estimación de la función de distribución condicional en presencia de censura y truncamiento: una aplicación al estudio de la mortalidad en pacientes diabéticos.* Tesis doctoral, Universidad de Vigo.

[15] Klein, J., y Moeschberger, M. (1997). *Survival analysis: techniques for censored and truncated data*. Springer, New York.

[16] Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*, 2nd Edition. Wiley.

[17] Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. "Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group". *Journal of Clinical Oncology*. 12(3):601-7, 1994.

[18] Lumley, T. (2007). The *Survival* Package.

[19] Meeker, W. Q. y Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. John Wiley & Sons.

[20] Pan, W. y Chappell, R. (1998), "A nonparametric estimator of survival functions for arbitrarily truncated and censored data", *Lifetime Data Analysis*, pg. 187–202.

[21] Pérez Ocón, R., Gámiz Pérez, M.L. y Ruíz Castro, J.E. (1998). *Métodos estocásticos en Teoría de la Fiabilidad.* Proyecto Sur de Ediciones, S.L.

[22] Solano Hurtado, H.(2008). *Análisis de supervivencia en Fiabilidad. Predicción en condiciones de alta censura y truncamiento: el caso de las redes de suministro de agua potable.* Tesis doctoral, Universidad de Valencia.

[23] Therneau, T.M. y Gramsch, P.M. (2000), *Modeling survival data. Extending the Cox model.* Springer.