UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Introducción al Análisis de Supervivencia con R

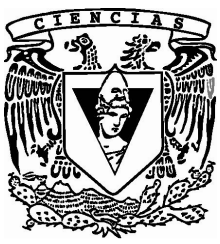
T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

ANGEL MANUEL GODOY AGUILAR



DIRECTOR DE TESIS:
MAT. MARGARITA ELVIRA CHÁVEZ CANO

2009

Tabla de Contenido

TABLA DE CONTENIDO

Introducción

En muchos campos de estudio es importante conocer el momento de ocurrencia de algún evento que sea de interés, frecuentemente lo que se pretende inferir es el tiempo que tarda en ocurrir un evento a partir de un momento en particular, a esto se le denomina *tiempo de falla*, pero hay que tener presente que a pesar de que la unidad de medida más común o intuitiva sea el tiempo, esta puede estar dada en longitud, dinero, volumen, etc. variando de acuerdo al contexto del evento. El análisis de supervivencia es una forma de modelar el tiempo de falla utilizando información de eventos que han ocurrido con anterioridad en circunstancias similares.

La utilización de *software* para realizar el análisis de supervivencia se ha vuelto imprescindible en la práctica, pues cuando se tiene gran cantidad de información a manejar, realizar el análisis sin una herramienta computacional sería una tarea poco envidiable. Pero no sólo esto, en algunos casos, la escasez de datos es la que ha motivado el desarrollo de los modelos de supervivencia utilizando información parcial de los tiempos de falla, de manera que el problema es el contrario al anteriormente mencionado, en este caso, la utilización de software permite obtener información precisa y confiable del modelo de supervivencia. Estas razones han sido la motivación para incluir conjuntamente con la teoría de supervivencia, la manera de realizar el análisis utilizando un paquete estadístico.

A lo largo de este trabajo, conforme se definen y desarrollan los elementos del modelo de supervivencia, se explica la manera de obtenerlos con el paquete estadístico **R**. La decisión de usar este paquete está sustentada por la coherencia de los métodos utilizados por el paquete con la teoría desarrollada en este trabajo, así como las opciones que ofrece para el manejo de datos, obtener y graficar la información de forma precisa, contribuyendo notablemente al estudio del modelo de supervivencia al facilitar su análisis.

En el primer capítulo de este trabajo se explican los elementos necesarios para ajustar un modelo de supervivencia y se formaliza la definición de

tiempo de falla. Un modelo de supervivencia se elabora con datos tales que la información que proporcionan del tiempo de falla del sujeto en estudio es total o parcialmente conocida. En este segundo caso, se plantean varios escenarios de la forma en que los datos aportan esta información parcial llamados *categorías de censura*, dado que cada escenario implica hipótesis distintas en los modelos de supervivencia, desarrollados con la finalidad de optimizar la información con que se cuenta. Otra característica llamada *truncamiento*, es la forma en que los datos son considerados o no en el estudio. Se debe tomar en cuenta que de identificar las características de la información que se tiene, se podrá utilizar el modelo adecuado para estudiar el tiempo de falla de interés.

En el segundo capítulo se explican las características de las funciones que aportan más información al modelar el tiempo de falla, así como las relaciones que hay entre ellas. Por la información que cada una de estas funciones aporta al análisis de supervivencia, la interpretación y manejo adecuado de éstas llevan al investigador a obtener informacón precisa del modelo. Además, se presentan un acercamiento a los modelos de supervivencia paramétricos más comunes centrando el interés en interpretar la forma de su funcón de riesgo por la ayuda que ésta aporta a identificar un modelo adecuado.

En el capítulo tercero se estima la llamada *función de supervivencia*, la cual, por la información que ésta aporta, y la manera en que se relaciona con las demás funciones de interés, es la función principal en el análisis de supervivencia. Esta función es el modelo más significativo del tiempo de falla, razón por la cual se incluye en este trabajo el desarrollo de bandas de confianza que permiten apreciar la relación entre el modelo estimado y el tiempo de falla real.

En el último capítulo se desarrolla un modelo que permite comparar el tiempo de falla entre individuos que pertenecen a distintos grupos. Esta es una herramienta útil en la práctica, dada la frecuencia con la que los investigadores recurren al análisis de supervivencia para identificar si hay diferencia en el tiempo de falla entre dos o más grupos. Su interés frecuentemente radica en identificar las características por las cuales difiere el tiempo de falla entre grupos y tomar decisiones de acuerdo a sus intereses.

Este trabajo tiene la finalidad de dar un acercamiento a la teoría general de supervivencia mediante el úso de un paquete estadístico como herramienta actuarial, con la finalidad de promover los estudios posteriores para la utilización de modelos de supervivencia, profundizando en las particularidades que los modelos requieran en las diversas áreas de aplicación actuarial.

Capítulo 1

Características de los datos de supervivencia

1.1. Introducción

En el análisis de supervivencia, el interés se centra en un grupo o varios grupos de individuos para cada uno de los cuales (o del cual) hay un evento puntual definido, llamado falla, que ocurre después de un tiempo llamado tiempo de falla. La falla puede ocurrir a lo más una vez en cualquier individuo. El término supervivencia se debe a que en las primeras aplicaciones de este método de análisis, se utilizaba como evento, la muerte de un paciente.

El tiempo de supervivencia se define como el tiempo transcurrido desde la entrada al estudio o estado inicial hasta el estado final o el tiempo que transcurre hasta la ocurrencia del evento de interés.

Ejemplos del tiempo de falla incluyen el tiempo de vida de componentes de máquinas en confiabilidad industrial, la duración de huelgas o periodos de desempleo en economía, los tiempos que toman los sujetos para completar tareas específicas en experimentación psicológica y comúnmente a los tiempos de supervivencia de pacientes en un ensayo clínico. Es importante tomar siempre en cuenta que el análisis de supervivencia tiene un amplio campo de aplicación en cualquier disciplina si es adecuadamente utilizado.

Para determinar el tiempo de falla de forma precisa, son necesarios tres requerimientos: un tiempo origen que debe ser definido sin ambigüedad, una escala para medir el paso del tiempo que debe ser acorde a las necesidades del estudio y finalmente, el significado de falla debe ser totalmente claro.

Algunas veces, es de interés solamente la distribución de los tiempos de falla de un solo grupo. Mas frecuentemente, el interés está en comparar los tiempos de falla en dos o más grupos para ver, por ejemplo, si los tiempos de falla de los individuos son más grandes sistemáticamente en el segundo grupo que en el primero. Alternativamente, los valores de las variables explicativas deben estar disponibles para cada individuo, estas variables están pensadas para que estén relacionadas con la supervivencia. Ilustrando esto, el tiempo de vida de una máquina puede estar influenciado por el esfuerzo ejercido sobre ésta, el material del cual está hecho, las sustancias con que tenga contacto o la temperatura del área de trabajo en la cual funciona, por tanto, estas condiciones mencionadas, pueden tomar el papel de variables explicativas en la supervivencia de la máquina que será el sujeto de estudio. En la práctica clínica, es muy común que de forma rutinaria se colecte una gran cantidad de información (capturada en variables) para cada paciente, dándose el investigador a la tarea poco envidiable de resumir el efecto conjunto de estas variables explicativas, sobre la supervivencia del paciente.

1.2. Tiempo de falla

El tiempo de origen debe ser definido de manera precisa para cada individuo. Es también deseable que, sujeto a cualesquier diferencias sobre las variables explicativas, todos los sujetos de estudio, sean tan comparables como sea posible en sus tiempos de origen. El tiempo de origen no necesita ser y usualmente no está en el mismo tiempo calendario para cada individuo. En la mayoría de los estudios se presentan entradas escalonadas, de tal forma que los sujetos entran a lo largo de un periodo, posiblemente largo, al estudio, por tanto, el tiempo de falla para cada sujeto es usualmente medido desde su propia fecha de entrada. La figura 1.1 ilustra esta situación.

La evaluación de programas de examen para la detección de cáncer de seno proporciona un ejemplo instructivo de las dificultades en la elección de un tiempo origen. El propósito del examen es detectar la enfermedad en una etapa temprana de su desarrollo, que de otra forma sería imposible. Aún en la ausencia de tratamiento efectivo, los pacientes con enfermedad detectada en el examen, se esperaría que sobrevivan más tiempo que los pacientes cuya enfermedad es detectada sin la ayuda del examen. Este sesgo complica seriamente cualquier comparación de los tiempos de falla de los dos. Quizá la única forma satisfactoria para evaluar el efecto del examen en la reducción de la mortalidad, es comparar la tasa de mortalidad en la población en la que se realiza el examen con otra en la que no se tiene acceso

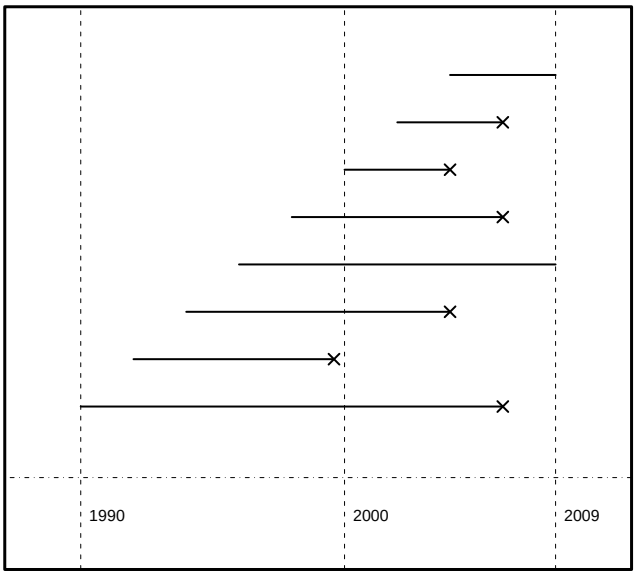


Figura 1.1: Tiempos de falla

a dicho examen.

La escala para medir el tiempo de falla es frecuentemente el tiempo reloj (tiempo real), aunque hay otras posibilidades, tales como el kilometraje en un auto, o defectos en hilos textiles, donde el tiempo de falla será la longitud medida hasta el primer defecto.

El significado de evento puntual de falla debe ser definido de forma precisa. En algún tratamiento médico, falla podría significar muerte, muerte por una causa específica como el cáncer de pulmón, la primera recurrencia de una enfermedad después del tratamiento, o la incidencia de una enfermedad nueva. En algunas aplicaciones hay poca o ninguna arbitrariedad en la definición de falla. En otras, por ejemplo, en algunos contextos industriales, la falla se define como el primer momento en el cual el desempeño, medido de alguna forma cuantitativa, cae por debajo de un nivel aceptable previamente establecido.

1.3. Censura y Truncamiento

Los datos de supervivencia se pueden presentar en diferentes formas que crean problemas con su análisis. Características particulares que usualmente se presentan son la Censura y el Truncamiento. En este trabajo se utilizará un tipo particular de datos de supervivencia, datos que pueden presentan censura por la derecha. Dada la importancia que tiene identificar las carácterísticas de los datos con que se trabaja, se explica a continuación los distintos tipos de censura y la diferencia con el truncamiento.

1.4. Censura

De forma general, la censura ocurre cuando se sabe que algunos tiempos de falla han ocurrido solamente dentro de ciertos intervalos y el resto de los tiempos de vida son conocidos exactamente. Hay varias categorías de censura, principalmente *Censura por la derecha*, *Censura por la izquierda* y *Censura por intervalo*. Para identificar adecuadamente el tipo de censura que presentan los datos, se tiene que considerar la forma en que han sido obtenidos los datos de supervivencia. Cada tipo de censura puede corresponder a diferente función de verosimilitud, la cual puede ser la base para la inferencia en la modelación.

1.4.1. Censura por la derecha

Primero se tiene que considerar la *Censura Tipo I* donde el evento es observado solamente si éste ocurre antes de un tiempo predeterminado, independientemente del tamaño de muestra. Un ejemplo de este tipo de censura se puede exhibir en un estudio de animales que comienza con un número fijo de éstos, a los cuales se les aplica uno o varios tratamientos, siendo la muerte de los animales el evento de interés. Debido al tiempo o por las consideraciones de costos, el investigador tiene que terminar el estudio antes de que se presente el evento de interés en todos los animales, sacrificando a los que no han fallecido. Los tiempos de supervivencia registrados para los animales que murieron durante el periodo de estudio son los tiempos desde el inicio del estudio hasta su muerte. Estos son llamados observaciones *exactas* o *no censuradas*. Los tiempos de supervivencia de los animales sacrificados no son conocidos exactamente, pero son registrados como al menos la longitud del estudio. Estas son llamadas observaciones censuradas. Algunos animales podrían perderse o morir accidentalmente y sus tiempos de supervivencia hasta

el momento de perderse o morir, son también observaciones censuradas, pero no corresponderán a la *Censura Tipo I*.

En censura por la derecha es conveniente usar la siguiente notación. Para un individuo específico bajo estudio, se supone que éste tiene un tiempo de vida X y un tiempo fijo de censura C_r (C_r por el nombre en inglés "*right censoring*"), donde las X´s para cada individuo se suponen como variables aleatorias independientes e idénticamente distribuídas con función de densidad $f(x)$. De este modo, el tiempo de vida exacto de un individuo puede ser conocido si y sólo si $X \leq C_r$. Si $X > C_r$, el individuo es un sobreviviente y su tiempo de vida es censurado en C_r.

Los datos del estudio pueden estar convenientemente representados por las parejas de variables (T, δ) donde δ es una variable indicadora definida como sigue:

$$\delta = \begin{cases} 1 & \text{si el tiempo de vida } X \text{ corresponde a un evento} \\ 0 & \text{si el tiempo de vida } X \text{ es censurado} \end{cases}$$

y T es igual a X si el tiempo de vida es observado o igual a C_r si es censurado, i.e., $T = min(X, C_r)$. Por construcción cada T para cada individuo es una variable aleatoria.

Cuando los sujetos de estudio tienen diferentes tiempos de censura, fijados previamente, esta forma de censura es llamada: *Censura Tipo I progresiva*. Este tipo de censura se puede representar mediante el siguiente ejemplo que presenta dos diferentes tiempos de censura.

Suponga que se tienen 20 ratones en un experimento donde el evento de interés es la muerte. Suponga que se han marcado 10 ratones de color rojo y los restantes 10 de color azul, de manera que se ha determinado a cada grupo de ratones, tiempos de censura de 42 y 104 semanas respectivamente. De modo que los ratones con marca roja que sobrevivan 42 semanas serán sacrificados, así como los ratones marcados de color azul que lleguen vivos a las 104 semanas.

Una forma de ampliar la perspectiva de la *Censura Tipo I* es cuando los individuos entran al estudio a diferentes tiempos, y el punto terminal de estudio predeterminado por el investigador es el mismo para todos. En este caso, el tiempo de censura para cada sujeto es conocido en el momento en que entra al estudio, de manera que cada individuo tiene fijo y especificado su

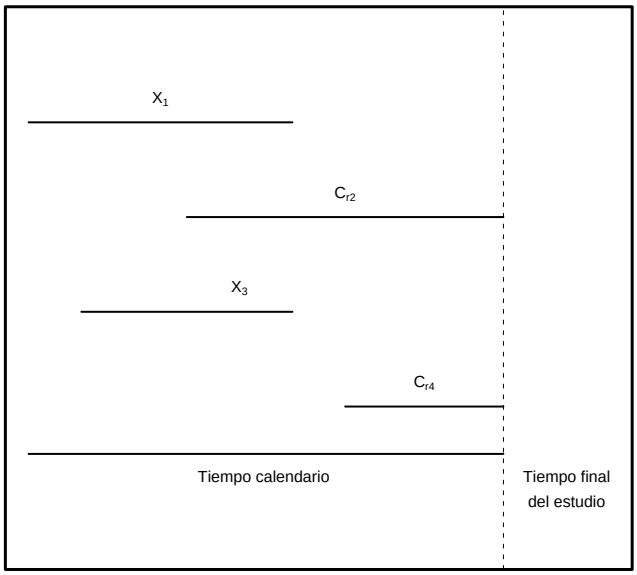


Figura 1.2: Censura tipo 1 generalizada para 4 individuos. $T_1 = X_1$.- Tiempo de falla para el primer individuo ($\delta_1 = 1$). $T_2 = C_{r2}$.- Tiempo censurado por la derecha para el segundo individuo ($\delta_2 = 0$). $T_3 = X_3$.- Tiempo de falla para el tercer individuo ($\delta_3 = 1$). $T_4 = C_{r4}$.- Tiempo censurado por la derecha para el cuarto individuo ($\delta_4 = 0$).

propio tiempo de censura. Este tipo de censura ha sido denominado *Censura de Tipo I generalizada*. Este tipo de censura es ilustrado en la figura 1.2.

Una representación conveniente de la *Censura de Tipo I generalizada* se da al reescalar la entrada al estudio de cada individuo al tiempo cero como se muestra en la fugura 1.3.

Otro metodo de representación es mediante el diagrama de Lexis. Donde el tiempo calendario se encuentra en el eje horizontal, y la longitud del tiempo de vida es representada por una linea de 45°. El tiempo que un individuo pasa en el estudio es representado por la altura del rayo en el eje vertical. Esto es ilustrado en la figura 1.4.

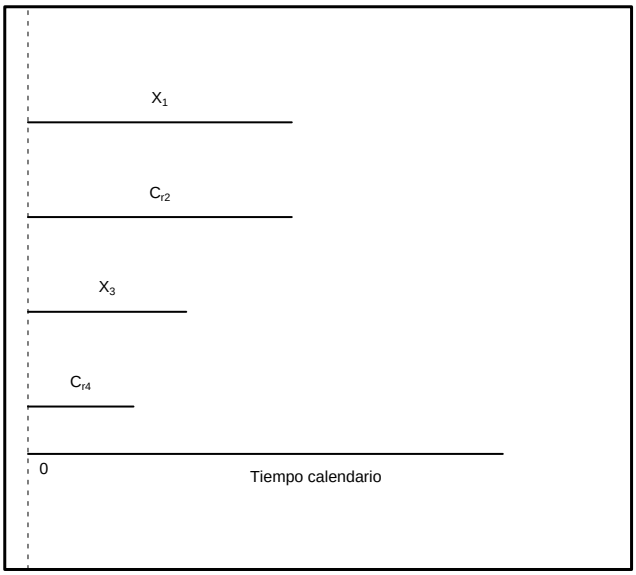


Figura 1.3: Censura tipo 1 generalizada para 4 individuos reescalada al tiempo cero

Un segundo tipo de censura por la derecha es la *Censura tipo II*, en la cual hay dependencia del tamaño de muestra (denotado por n) y las fallas que se observen. Aquí, todos los individuos son puestos en estudio al mismo tiempo y se da el término de éste cuando r de los n individuos han presentado el evento de interés. Donde r es un número entero positivo determinado previamente por el investigador, tal que $r < n$. La notación conveniente para este tipo de censura se presenta como sigue. Sean $T_1, T_2, \ldots, T_n$ los tiempos de falla de los n individuos y sean $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ sus respectivas estadísticas de orden. Entonces el final del estudio queda dado de forma aleatoria por $T_{(r)}$, la $r-$ésima estadística de orden. Por tanto, $(n-r)$ observaciones serán censuradas y fijadas al tiempo $T_{(r)}$.En este caso, el timepo de censura es aleatorio, pues $(n-r)$ observaciones serán censuradas al tiempo dado por la r-ésima falla, la cual no se sabe cuando ocurrira. De modo que esto marca una diferencia importante entre la *Censura de Tipo I* y la *Censura tipo II*.

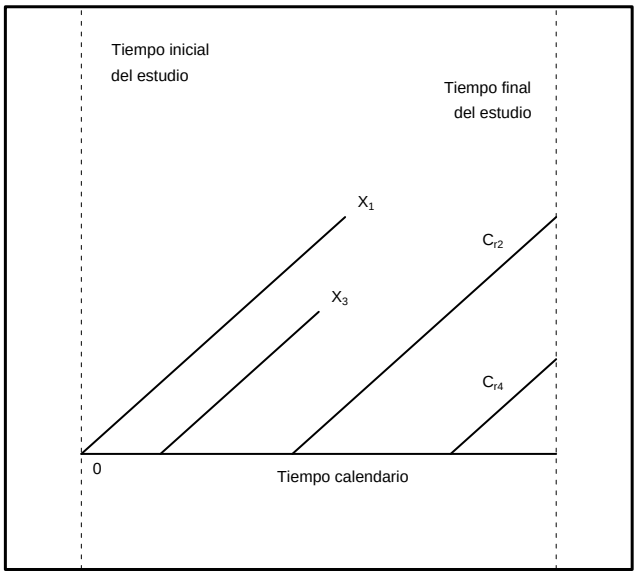


Figura 1.4: Diagrama de Lexis para la censura de tipo 1 generalizada

$$Censura = \begin{cases} Tipo\ I & C \text{ determinístico} \\ Tipo\ II & C \text{ aleatorio} \end{cases}$$

Una generalización de la *Censura tipo II* es similar a la generalización en la censura *Censura tipo I*, con diferentes tiempos de censura. Esta es llamada *Censura tipo II progresiva*. Aquí, el investigador debe fijar los siguientes elementos antes de comenzar el estudio. K será el número de diferentes tiempos de censura que se realizarán a lo largo del estudio en una muestra de tamaño n. $r_1, r_2, \ldots, r_k$ (k números enteros positivos) serán el número de sujetos que deberán presentar el evento de interés para determinar el respectivo tiempo de censura y $n_1, n_2, \ldots, n_k$ (k números enteros positivos tales que $n_1 + n_2 + ... + n_k = n$) serán el número de individuos que deben estar fuera del estudio a cada tiempo de censura. Con estos elementos, el estudio será realizado de la siguiente forma:

Al presentarse los primeros r_1 eventos de interés, también denotados por

fallas, $n_1 - r_1$ individuos serán retirados de los $n - r_1$ individuos sobrevivientes, quedando $n - n_1$ individuos en el estudio. Cuando se presenten las siguientes r_2 fallas, $n_2 - r_2$ individuos serán retirados de los $(n - n_1) - r_2$ individuos sobrevivientes, quedando $n - (n_1 + n_2)$ individuos en el estudio. Y asi sucesivamente hasta que al tener r_k fallas de los $n - (n_1 + n_2 + ... + n_{(k-1)}) = n_k$ individuos sobrevivientes en el estudio, los $(n - n_1 - n_2 ... - n_{(k-1)}) - r_k = n_k - r_k$ individuos restantes sean eliminados, dando por terminado el experimento. De este modo, si T_i denota el tiempo del $i-$ésimo sujeto en presentar el evento de interés (lo cual excluye a los sujetos removidos intencionalmente), los K tiempos de censura serán las v.a. T_{r_1} , $T_{n_1+r_2}$, $T_{n_1+n_2+r_3}$,... ,$T_{n_1+n_2+...+n_{k-1}+r_k}$.

La *Censura tipo II progresiva* puede ser representada mediante el siguiente ejemplo. Suponga que se tienen 100 ratones en un experimento donde el evento de interés es la muerte. Se definen 3 tiempos de falla distintos. El primer tiempo de falla se dará cuando mueran 15 ratones, en ese momento, se retirarán del estudio 15 ratones de los 85 vivos, continuando en el estudio 70 ratones. El segundo tiempo de falla se dará cuando mueran 20 ratones de los 70 en estudio, en ese momento, se retirarán 10 ratones de los 50 vivos, quedando 40 ratones en estudio. El tercer tiempo de falla quedara determinado cuando mueran 30 ratones de los 40 en estudio y se sacrificarán en ese momento los 10 ratones supervivientes. De este modo, en el primer tiempo de falla se obtendrán 15 eventos y 15 censuras, en el segundo tiempo de falla se obtendrán 20 eventos y 10 censuras, y en el tercer tiempo de falla se obtendrán 30 eventos y 10 censuras.

Otro tipo de censura es la *Censura tipo III* o también llamada *Censura aleatoria*, la cual surge cuando los sujetos salen del estudio sin presentar la falla por razones no controladas por el investigador. Por ejemplo, en un estudio donde el evento de interés es la muerte por alguna razón específica, un sujeto puede presentar *Censura aleatoria* si fallece por alguna razón ajena a la de interés, o si el investigador pierde acceso al sujeto y éste sale del estudio.

1.4.2. Censura por la izquierda

Un tiempo de vida X asociado con un individuo específico en el estudio, es considerado *censurado por la izquierda*, si éste es menor que un tiempo de censura C_l (C_l por el nombre en inglés "*left censoring*"). Esto es, que el evento de interés le ha ocurrido al sujeto en estudio, antes de que el sujeto haya sido observado por el investigador al tiempo C_l. Para estos individuos, se sabe que han presentado el evento algún tiempo antes de C_l. El dato proveniente de

una muestra censurada por la izquierda puede ser representado por la pareja de variables aleatorias (T, ϵ) donde $T = X$ si el tiempo de vida es observado o $T = C_l$ si es censurado y ϵ indica cuando el tiempo de vida exacto es observado ($\epsilon = 1$) o no ($\epsilon = 0$).

Algunas veces, si la censura por la izquierda ocurre en el estudio, la censura por la derecha puede ocurrir también y los tiempos de vida son considerados *doblemente censurados*. De nuevo, los datos pueden ser representados por una pareja de variables (T, δ) donde $T = max\,[min(X, C_r), C_l]$ es el tiempo de estudio y δ es una variable indicadora definida como sigue:

$$\delta = \begin{cases} 1 & \text{si } T \text{ es el tiempo de ocurrencia del evento} \\ 0 & \text{si } T \text{ es el tiempo censurado por la derecha} \\ -1 & \text{si } T \text{ es el tiempo censurado por la izquierda} \end{cases}$$

1.4.3. Censura por intervalo

Este es un tipo de censura más general que ocurre cuando el tiempo de vida se sabe que ocurre solamente dentro de un intervalo. Este tipo de censura se presenta cuando se tiene un estudio longitudinal donde el seguimiento del estado de los sujetos se realiza periódicamente y por tanto, la falla sólo puede conocerse entre dos periodos de revisión, generando un intervalo de la forma (L_i, R_i) para cada sujeto en el estudio.

1.5. Truncamiento

Una segunda característica que puede presentarse en algunos estudios de supervivencia, son los datos truncados.

El truncamiento es definido como una condición que presentan ciertos sujetos en el estudio y el investigador no puede considerar su existencia.

Cuando los datos presentan truncamiento, solamente individuos a los que les ocurre algún evento particular, antes del evento de interés o la censura, son considerados en el análisis por el investigador.

1.5.1. Truncamiento por la izquierda

Este ocurre cuando los sujetos entran al estudio a una edad particular (no necesariamente el origen del evento de interés), y son observados desde este "tiempo retrasado de entrada", hasta que el evento ocurra o hasta que el evento es censurado.

Si Y es el momento de ocurrencia del evento que trunca a los sujetos en estudio, entonces para muestras truncadas por la izquierda, solo los individuos tales que $X \geq Y$ serán considerados.

El tipo mas común de truncamiento por la izquierda ocurre cuando los sujetos entran al estudio a una edad aleatoria y son observados por este "tiempo retrasado de entrada", hasta que el evento ocurre o hasta que el sujeto es censurado por la derecha. En este caso, todos los sujetos que presenten el evento de interés antes del "tiempo retrasado de entrada", no serán considerados para el experimento. Note que esto es opuesto a la censura por la izquierda, donde se tiene información parcial de individuos que presentan el evento de interés antes de su edad de entrada al estudio, para truncamiento por la izquierda, estos individuos no serán considerados para ser incluidos en el estudio.

1.5.2. Truncamiento por la derecha

Este ocurre cuando sólo individuos que han presentado el evento son incluidos en la muestra y ningún sujeto que haya presentado aún el evento será considerado. Un ejemplo de muestras que presentan truncamiento por la derecha, son los estudios de mortalidad basados en registros de muerte.

1.6. Objeto de supervivencia en R

Para utilizar las funciones de supervivencia de **R** en este trabajo, es necesario cargar la librería *Survival* mediante la siguiente instrucción

```
> library("survival")
```

Las rutinas en la librería *Survival* trabajan con objetos de la forma *Surv*, los cuales son una estructura de datos que combinan información de tiempo y censura. Cada objeto es construido usando la función *Surv*, esta función

crea un objeto de supervivencia en el cual se captura el tiempo observado de la variable y su estatus (censurado o no), donde la estructura para datos que presentan censura por la derecha es como sigue

```
Surv(time,event)
```

Esta función presenta los argumentos *time* y *event*. El argumento *time* corresponde al tiempo desde que el sujeto entra al estudio, hasta que éste presenta la falla o censura. El argumento *event* es una variable binaria uqe indica el estatus del tiempo registrado, donde, de forma predeterminada corresponde a

$$\begin{aligned} 0 \leftarrow &\quad \text{si el dato es censurado} \\ 1 \leftarrow &\quad \text{si la falla es observada} \end{aligned}$$

Para ilustrar esto, suponga que se tiene un individuo con tiempo de falla en 5. El objeto de la forma Surv para este individuo corresponde a

```
> Surv(5,1)
[1] 5
```

En caso de que el tiempo de este individuo fuera censurado, quedaría como sigue

```
> Surv(5,0)
[1] 5+
```

Como se puede apreciar, el tiempo censurado queda denotado por un signo +, indicando, por la naturaleza de la censura, que la falla se pudo presentar en un momento futuro.

Si se tiene una base de datos donde se indique el tiempo, y el estatus no esté determinado por 1 y 0. El argumento *event* puede ser modificado, de modo que el valor que tome la variable al ocurrir el evento puede ser especificado explícitamente por un número o un caractér, y la ausencia de éste será entendido como censura.

Para ejemplificar esto, sean 3 individuos que presentan fallas en los tiempos $1, 2, 3$ y sean 3 individuos que presentan censura en los tiempos $4, 5, 6$, donde la base de datos tiene la estructura siguiente

1.6. OBJETO DE SUPERVIVENCIA EN R

```
> tiempo <- c(1,2,3,4,5,6)

> estatus <- c("falla","falla","falla","censura",
               "censura","censura")
```

El objeto *Surv* para este conjunto de datos puede ser creado como sigue

```
> Surv(tiempo,estatus == "falla")
[1] 1  2  3  4+ 5+ 6+
```

Si de forma alternativa, la estructura de los datos fuera distinta, donde la falla está indicada por el número 7 y las censuras por distintos caractéres o números, como sigue

```
> tiempo <- c(1,2,3,4,5,6)
> estatus <- c(7,7,7,"Censura","NA",0)
```

El objeto *Surv* para este nuevo conjunto de datos puede ser creado como sigue

```
> Surv(tiempo,estatus == 7)
[1] 1  2  3  4+ 5+ 6+
```

Con los objetos creados de esta forma se puede trabajar con las funciónes en **R**, pues cuentan con la información del tiempo de ocurrencia de la falla o la cencura de cada dato.

Capítulo 2

Modelo de supervivencia

Un modelo de supervivencia es caracterizado por variables aleatorias no negativas, de modo que la variable aleatoria T será tomada para denotar el tiempo de falla o también llamado tiempo de supervivencia. Por la naturaleza de T, se tiene que $T \geq 0$. La distribución de ésta variable puede ser caracterizada por almenos 4 funciones básicas:

1 Función de supervivencia
2 Función de riesgo
3 Función de densidad de probabilidad
4 Vida media residual

Éstas funciones son matemáticamente equivalentes, en el sentido de que si una de ellas está dada, pueden derivarse las otras tres. Otra función relacionada con las anteriores es la *función de riesgo acumulado* que puede resultar útil en el análisis de supervivencia.

En la práctica, las cuatro funciones principales en supervivencia pueden ser utilizadas para ilustrar diferentes aspectos de los datos. Un aspecto básico en el análisis de supervivencia, es la estimación de estas funciones a partir de los datos muestrales y extraer inferencias acerca del patrón de supervivencia en la población.

2.1. Función de supervivencia

La función básica empleada para describir los fenómenos de tiempo-evento es la función de supervivencia denotada por $S(t)$, también llamada *tasa de supervivencia acumulativa*. Esta función es la probabilidad de que un sujeto

en estudio no experimente el evento de interés (sobreviva) antes de un momento dado, por tanto, sea T una variable aleatoria no negativa (de tiempo de falla) con función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(\text{un individuo sobreviva mas allá de } t). \end{aligned}$$

O visto de otra forma:

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - P(T \leq t) \\ &= 1 - P(\text{un individuo falle antes del tiempo } t). \end{aligned}$$

Por tales características, $S(t)$ es una función no creciente tal que:

$$S(0) = 1 \quad \text{y} \quad S(t) = 0 \text{ cuando } t \to \infty.$$

Esto es, la probabilidad de sobrevivir al menos al tiempo cero es uno y la de sobrevivir un tiempo infinito es cero.

Cuando T es una variable aleatoria continua, la función de supervivencia es la integral de la función de densidad de probabilidad, esto es,

$$S(t) = P(T > t) = \int_t^\infty f(t)dt.$$

Para describir el recorrido de la supervivencia, se hace la representación gráfica de $S(t)$. Esta gráfica es llamada curva de supervivencia. Muchos tipos de curvas de supervivencia pueden presentarse y analizarse de manera particular, pero es importante notar que todas tienen las mismas propiedades básicas, son monótonas no crecientes, igual a uno en cero y a cero cuando el tiempo tiende a infinito. La tasa de decrecimiento, varía de acuerdo al riesgo que experimente el evento al tiempo t pero es difícil determinar en esencia la modelación de la falla solamente observando la curva de supervivencia. No obstante, el uso de esta curva representa un análisis importante en la práctica, y es usual comparar dos o más curvas de supervivencia para comprender el comportamiento que tienen entre ellas a lo largo del tiempo.

En la representación gráfica, una curva de supervivencia empinada, como la que se muestra en la figura 2.1-(a) representa baja tasa de supervivencia o corto tiempo de supervivencia. Una curva de supervivencia plana o gradual como la que se muestra en la figura 2.1-(b) representa alta tasa de

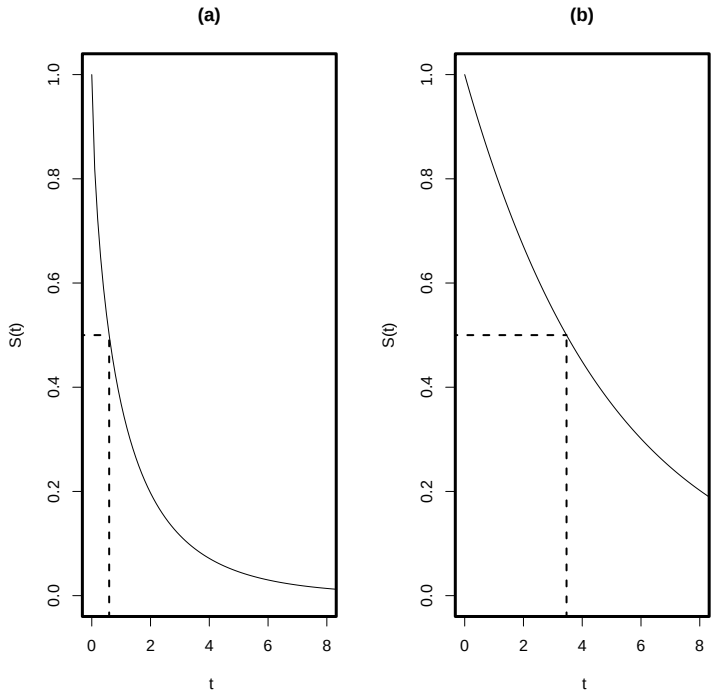


Figura 2.1: Curvas de supervivencia

supervivencia o mayor supervivencia. La curva de supervivencia puede ser utilizada para encontrar el percentil 50 (la mediana) y otros percentiles o cuantiles (por ejemplo el 25^{avo} y el 75^{avo}) del tiempo de supervivencia. La mediana de los tiempos de supervivencia en las figuras 2.1-(a) y 2.1-(b) son aproximadamente .6 y 3.5 unidades de tiempo, respectivamente. La media es utilizada para describir la tendencia central de una distribución, pero en las distribuciones de supervivencia la mediana es frecuentemente mejor, debido a que un pequeño número de sujetos con tiempo de vida excepcionalmente largos o cortos va a causar que la media del tiempo de supervivencia sea desproporcionadamente grande o pequeña.

Cuando T es una variable aleatoria discreta, diferentes técnicas son requeridas. En análisis de supervivencia, estas variables surgen cuando los tiempos de falla están agrupados en intervalos o cuando los tiempos de vida hacen referencia a unidades en números enteros positivos. Sea T una v.a. discreta que toma valores t_j con $j = 1, 2, ...$, con función de masa de probabilidad $f(t_j) = P(T = t_j)$ donde $t_1 < t_2 < \ldots$.

La función de supervivencia para una variable aleatoria discreta T está da-

da por

$$S(t) \;=\; P(T > t) \;= \sum_{t_j > t} f(t_j).$$

2.2. Función de densidad

Como cualquier variable aleatoria, el tiempo de supervivencia T tiene una función de densidad de probabilidad. En el caso continuo

$$S(t) \;=\; P(T > t) \;=\; \int_t^{\infty} f(u)du$$

entonces,

$$f(t) \;=\; -\frac{dS(t)}{dt}. \tag{2.1}$$

De la relación 2.1, $f(t)dt$ puede ser, aunque de manera aproximada, la probabilidad de que un evento pueda ocurrir al tiempo t y $f(t)$ es una función no negativa con área bajo $f(t)$ igual a uno.

De la función de densidad puede ser encontrada la proporción de individuos que cae en cualquier intervalo de tiempo y el pico de frecuencia más alto de fallas. La curva de densidad en la figura 2.2-(a) da un patrón de alta tasa de fallas al principio del estudio y una tasa decreciente de fallas cuando se incrementa el tiempo. En la figura 2.2-(b), el pico de frecuencia alta de fallas ocurre a aproximadamente 1.7 unidades de tiempo. La proporción de individuos que cae entre 1 y 2 unidades de tiempo es igual al área sombreada que aparece en las figuras.

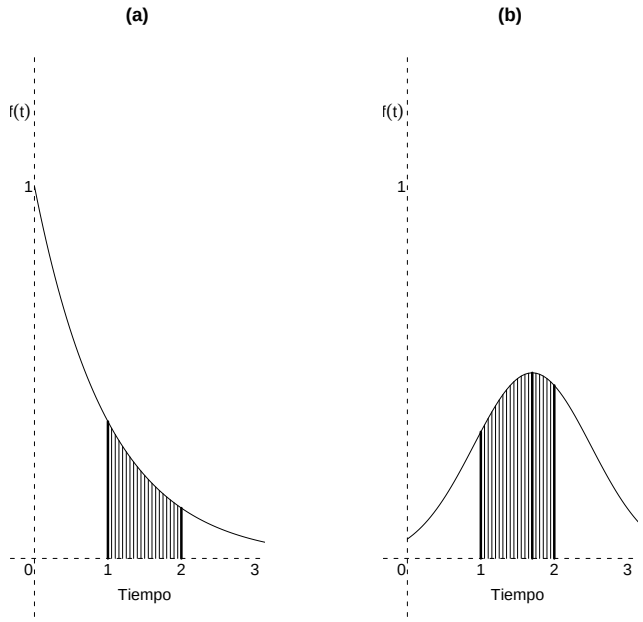


Figura 2.2: Curvas de densidad

En el caso discreto, como es mencionado anteriormente

$$f(t_j) = P(T = t_j).$$

2.3. Función de riesgo

La función de riesgo del tiempo de supervivencia T da la tasa de falla condicional. Esta se define como la probabilidad de falla durante un intervalo de tiempo muy pequeño, suponiendo que el sujeto de estudio ha sobrevivido hasta el inicio del intervalo, o como el límite de la probabilidad de que un sujeto falle en un intervalo muy corto, t a $t + \Delta t$ dado que el individuo ha sobrevivido hasta el tiempo t. La función de riesgo queda definida como

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

$$\lim_{\Delta t \to 0^+} \frac{P(\text{ un sujeto de edad } t \text{ falle en el intervalo de tiempo } (t, t+\Delta t))}{\Delta t P(\text{ un individuo sobreviva mas allá de } t)}.$$

Si T es una variable aleatoria continua, entonces

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log S(t).$$

Una función relacionada se define como la *función de riesgo acumulado* dada por

$$H(t) = \int_0^t h(u)du = -\ln(S(t)). \tag{2.2}$$

Por tanto, usando la relación 2.2 relaciones se tiene que

$$S(t) = exp\left[-H(t)\right] = exp\left[-\int_0^t h(u)du\right].$$

La función de riesgo es también conocida como *tasa instantánea de mortalidad, fuerza de mortalidad, tasa de mortalidad condicional, tasa de falla de edad específica* y demás nombres relacionados con el tema que se esté tratando y la interpretación que se tenga dentro de éste. Es una medida de propención a falla como una función de la edad del individuo en el sentido de que la cantidad $\Delta t h(t)$ es la proporción esperada de individuos de edad t que fallarán en el intervalo t a $t + \Delta t$, o una aproximación a la probabilidad de que un individuo de edad t experimente un evento en el siguiente instante.

La función de riesgo describe la forma en que cambia la tasa instantánea de la ocurrencia de un evento de interés al paso del tiempo y la única restricción para esta función es que tiene que ser no negativa, es decir $h(t) \geq 0$. La función de riesgo puede crecer, decrecer, permanecer constante o tener un proceso más complicado. En la figura 2.3 están graficados varios tipos de función de riesgo.

Para ilustrar las funciones de riesgo se presentan algunos escenarios en la figura 2.3, por ejemplo, pacientes con leucemia que no responden al tratamiento tienen una tasa de riesgo creciente $h_1(t)$. $h_2(t)$ es una función de riesgo decreciente que puede indicar el riesgo de soldados heridos por bala que fueron sometidos a cirugía. El peligro principal es la operación misma y este peligro decrece si la cirugía es exitosa. Una función de riesgo constante como en $h_3(t)$ es el riesgo de individuos saludables entre 18 y 40 años de edad

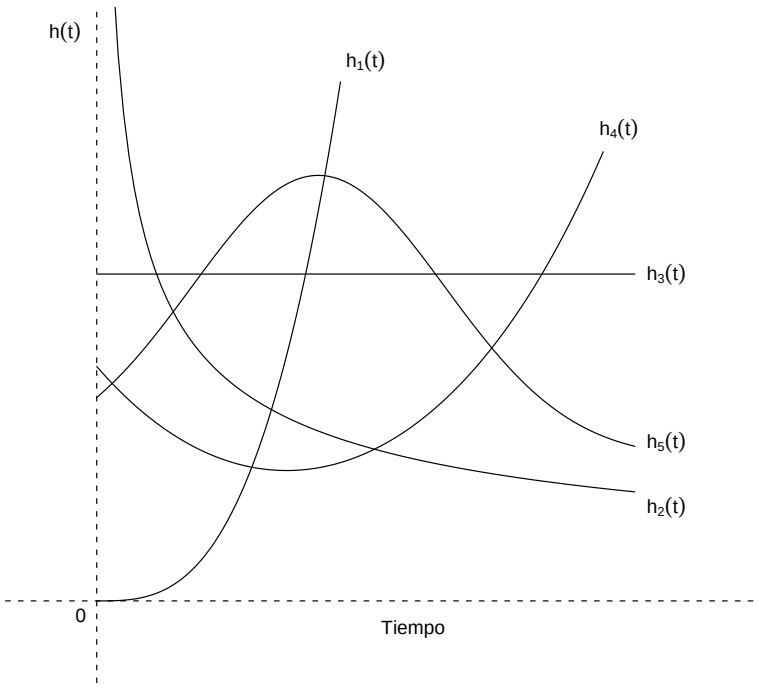


Figura 2.3: Funciones de riesgo

cuyos riesgos principales de muerte son los accidentes. La llamada curva de tubo de baño $h_4(4)$ describe el proceso de vida humana, durante el periodo inicial el riesgo es alto (alta mortalidad infantil), subsecuentemente el riesgo permanece aproximadamente constante hasta un cierto tiempo, después del cual crece debido a fallas por deterioro. Finalmente, pacientes con tuberculosis tienen riesgos que se incrementan inicialmente, luego decrecen después de tratamiento. Este incremento y luego decremento se muestra en la función de riesgo $h_5(t)$.

En el caso discreto. Sea T una v.a. discreta que toma valores t_j con $j = 1, 2, ...$ La función de riesgo se define para los valores t_j y proporciona la probabilidad condicional de falla al tiempo $t = t_j$, dado que el individuo estaba vivo antes de t_j, por lo tanto se tiene que

$$\begin{aligned} h(t_j) &= P(T = t_j | T \geq t_j) \\ &= \frac{P(T = t_j)}{P(T \geq t_j)} \\ &= \frac{f(t_j)}{S(t_j-)}. \end{aligned}$$

Donde t_j- corresponde a un instante antes de t_j y por tanto

$P(T \geq t_j) = 1 - P(T < t_j) = S(t_j-) \neq S(t_j)$ en el caso discreto.

Notemos que:

$$f(t_j) = S(t_{j-1}) - S(t_j).$$

Por tanto:

$$h(t_j) = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}. \tag{2.3}$$

Despejando a $S(t_j)$ en la ecuación 2.3 tenemos:

$$S(t_j) = [1 - h(t_j)]\, S(t_{j-1}). \tag{2.4}$$

Donde podemos ver que:

$$S(t_1) = [1 - h_1]S(0) = [1 - h_1]$$

$$S(t_2) = [1 - h(t_2)]\, S(t_1) = [1 - h(t_2)]\, [1 - h(t_1)]$$

$$S(t_3) = [1 - h(t_3)]\, S(t_2) = [1 - h(t_3)]\, [1 - h(t_2)]\, [1 - h(t_1)]$$

...

Por tanto se tiene que:

$$S(t) = \prod_{t_j \leq t} (1 - h(t_j)). \tag{2.5}$$

Y consecuentemente de la ecuación 2.4 se tiene que:

$$S(t) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})}.$$

2.3.1. Función de riesgo acumulado

La función de riesgo acumulado es denotada por $H(t)$. En el caso continuo corresponde a

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

y en el caso discreto,

$$H(t) = \sum_{t_j \leq t} h(t_j).$$

Pero esta definición tiene un inconveniente con la relación

$$S(t) = exp\left[-H(t)\right]$$

pues esta definición en el caso discreto no es cierta, aunque es utilizada como una aproximación, sucede que

$$S(t) = exp\left\{-H(t)\right\} = e^{h(t_1)}e^{h(t_2)}...e^{h(t_j)}$$

con $t_j \leq t$.

Lo cual no corresponde con la relación entre $S(t)$ y $h(t)$ de la ecuación 2.5 en el caso discreto. Por este motivo se prefiere definir a la función de riesgo acumulado en el caso discreto como

$$H(t) = -\sum_{t_j \leq t} ln\left[1 - h(t_j)\right].$$

Expresión que está bien definida dado que $0 < h(t_j) < 1$, pues

$$h(t_j) = 1 - \frac{S(t_j)}{S(t_{j-1})}$$

y para los valores t_j donde $S(t_j)$ tiene sentido en el caso discreto, sucede que

$$S(t_j) > S(t_{j+1}).$$

De tal modo que

$$S(t) = exp\left\{-H(t)\right\}$$

$$= exp\left\{+\sum_{t_j \leq t} ln\left[1 - h(t_j)\right]\right\} = \prod_{t_j \leq t}(1 - h(t_j))$$

Lo cual concuerda con la relación entre $S(t)$ y $h(t)$ de la ecuación 2.5 en el caso discreto.

En ambos casos, tanto el discreto como el continuo, esta función como su nombre lo indica, acumula el riesgo al paso del tiempo. De tal manera que corresponde a una función no decreciente y de acuerdo a su forma de incrementarse, se podrá tener información del comportamiento del riesgo a lo largo del tiempo, lo cual es una ventaja en el análisis de supervivencia.

2.4. Vida media residual

La cuarta función básica en el análisis de supervivencia es la función de *Vida media residual* al tiempo t denotada como $mrl(t)$ (por el nombre en inglés *mean residual life*). Para los sujetos de edad t, esta función mide la esperanza de tiempo de vida restante, o el tiempo esperado antes de la ocurrencia del evento de interés. Por tanto, queda definida por

$$mrl(t) = E(T - t|T > t)$$

Para el caso continuo, por definición de esperanza condicional se tiene que

$$E(T - t|T > t) = \int_0^\infty (u - t) f(u|u > t) du$$

$$= \int_0^\infty (u - t)\frac{f(u)}{S(t)} I_{(t,\infty)}(u) du = \int_t^\infty \frac{(u - t)f(u)}{S(t)} du.$$

Por lo cual la función de *Vida media residual* al tiempo t queda definida por

$$mrl(t) = E(T - t|T > t) = \frac{\int_t^\infty (u - t) f(u) du}{S(t)}.$$

Por lo que se puede apreciar que la vida media residual es el área bajo la curva de supervivencia a la derecha de t dividida entre $S(t)$. De tal modo

que la vida media $\mu = E(T) = E(T - 0|T > 0) = mrl(0)$, es el área total de la curva de supervivencia, es decir,

$$\mu = E(T) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt.$$

De esta forma, para el caso de variables aleatorias continuas se tiene la relación

$$mrl(t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}.$$

Además, la varianza de T está relacionada con la función de supervivencia por

$$Var(T) = E\left[T^2\right] - E\left[T\right]^2$$

$$= 2\int_0^\infty uS(u)du - \left[\int_0^\infty S(u)du\right]^2.$$

Dado que, integrando por partes se tiene

$$E(T^2) = \int_0^\infty t^2 f(t)dt$$

$$= -t^2 S(t)\,|_0^\infty - \int_0^\infty (-)2tS(t)dt = 2\int_0^\infty tS(t)dt.$$

El p-ésimo cuantil, también llamado $100p$ percentil de la distribución de T es el valor x_p tal que

$$S(x_p) = 1 - p.$$

La mediana del tiempo de vida es el $50p$ percentil $x_{0.5}$ de la distribución de T. De esto se sigue que la mediana del tiempo de vida es $x_{0.5}$ tal que

$$S(x_{0.5}) = 0.5.$$

En el caso discreto la función de *Vida media residual* está dada por

$$E(T - t|T > t) = \sum_{r=0}^{\infty}(t_r - t)P[T = t_r|T > t] = \sum_{r=0}^{\infty}\frac{(t_r - t)P[T = t_r, T > t]}{S(t)}$$

sea $t_i \leq t < t_{i+1}$ para alguna i con $i = 0,1,2,...$

$$= \frac{(t_{i+1}-t)P\left[T=t_{i+1}\right]+(t_{i+2}-t)P\left[T=t_{i+2}\right]+(t_{i+3}-t)P\left[T=t_{i+3}\right]+...}{S(t)}$$

$$= \frac{(t_{i+1}-t)\left[S(t_i)-S(t_{i+1})\right]+(t_{i+2}-t)\left[S(t_{i+1})-S(t_{i+2})\right]+(t_{i+3}-t)\left[S(t_{i+2})-S(t_{i+3})\right]+...}{S(t)}$$

$$= \frac{(t_{i+1}-t)S(t_i)-(t_{i+1}-t)S(t_{i+1})+(t_{i+2}-t)S(t_{i+1})-(t_{i+2}-t)S(t_{i+2})+...}{S(t)}$$

$$= \frac{(t_{i+1}-t)S(t_i)+(t_{i+2}-t_{i+1})S(t_{i+1})+(t_{i+3}-t_{i+2})S(t_{i+2})+...}{S(t)}$$

Por tanto el caso discreto la función de *Vida media residual* queda expresada como

$$mrl(t) = \frac{(t_{i+1}-t)S(t_i)+\sum_{j\geq i+1}(t_{j+1}-t_j)S(t_j)}{S(t)} \text{ para } t_i \leq t < t_{i+1}.$$

Donde el p-ésimo cuantil de la distribución de T es el valor x_p tal que

$$F(x_p) \geq p \text{ y } S(x_p) \geq 1-p.$$

La relación entre la función de *Vida media residual* y la función de riesgo está dada por

$$h(x) = \left[\frac{\frac{dmrl(x)}{dx}+1}{mrl(x)}\right]. \tag{2.6}$$

Esto se exhibe desarrollando la parte derecha de la ecuación como sigue:

Utilizando la definición de $mrl(x)$

$$\left[\frac{\frac{dmrl(x)}{dx}+1}{mrl(x)}\right] = \left[\frac{\frac{d\frac{1}{S(x)}\int_x^\infty S(u)du}{dx}+1}{mrl(x)}\right]. \tag{2.7}$$

Utilizando que

$$\frac{d\int_x^\infty S(u)du}{dx} = \lim_{r\to\infty} S(r)\frac{dx}{dx} - S(x)\frac{dx}{dx} = -S(x)$$

y el producto de las derivadas para desarrollar el diferencial de la ecuación 2.7 se tiene que

$$= \frac{\frac{-S'(x)}{S(x)^2}\int_x^\infty S(u)du + \frac{-S(x)}{S(x)} + 1}{mrl(x)} = \frac{\frac{-S'(x)}{S(x)^2}\int_x^\infty S(u)du}{\int_x^\infty \frac{S(u)}{S(x)}du}$$

utilizando que $-S'(x) = f(x)$

$$= \frac{f(x)\int_x^\infty S(u)du S(x)}{S(x)^2\int_x^\infty S(u)du} = \frac{f(x)}{S(x)} = h(x).$$

Con lo cual se concluye la igualdad de la relación 2.6.

La relación entre la función de *Vida media residual* y la función de supervivencia está dada por

$$S(x) = \frac{mrl(0)}{mrl(x)} exp\left[-\int_0^x \frac{1}{mrl(u)}du\right]. \tag{2.8}$$

Esto se exhibe desarrollando la parte izquierda de la ecuación como sigue:

$$\begin{aligned}
S(t) &= exp\left[-\int_0^x h(u)du\right] \\
&= exp\left[-\int_0^x \frac{mrl'(u)+1}{mrl(u)}du\right] \\
&= exp\left[-\int_0^x \frac{mrl'(u)}{mrl(u)}du - \int_0^x \frac{1}{mrl(u)}du\right] \\
&= exp\left[-\int_0^x \frac{1}{mrl(u)}du\right] exp\left[-\int_0^x \frac{mrl'(u)}{mrl(u)}du\right] \\
&= exp\left[-\int_0^x \frac{1}{mrl(u)}du\right] exp\left[-\int_0^x \frac{dln(mrl(u))}{du}du\right] \\
&= exp\left[-\int_0^x \frac{1}{mrl(u)}du\right] exp\left[ln(mrl(0)) - ln(mrl(x))\right] \\
&= exp\left[-\int_0^x \frac{1}{mrl(u)}du\right] \frac{mrl(0)}{mrl(x)}
\end{aligned}$$

Con lo cual se concluye la igualdad en la relación 2.8

La relación entre la función de *Vida media residual* y la función de densidad está dada por

$$f(t) = \left(\frac{d}{dt}mrl(t) + 1\right)\left(\frac{mrl(0)}{mrl(t)^2}\right) exp\left[-\int_0^t \frac{du}{mrl(u)}\right] \tag{2.9}$$

Usando que $f(t) = h(t)S(t)$ y las ecuaciones 2.6, 2.8 se puede obtener la igualdad en 2.9.

2.5. Relaciones entre las funciones de supervivencia

Es importante tomar en cuenta que a partir de alguna de las cuatro funciones básicas en supervivencia se pueden obtener las demás. Esto es útil en el momento en que se desea interpretar el comportamiento de la variable aleatoria T, que indica el tiempo de supervivencia, con cada una de estas funciones en un contexto apropiado al análisis que se realice.

2.5.1. Caso continuo

Las relaciones entre las cuatro funciónes básicas en supervivencia en el caso en que T es una variable aleatoria continua puede ser resumida como sigue[1]:

[1]Klein [6]. pag. 35.

$$
\begin{aligned}
S(t) &= \int_t^{\infty} f(t)dt \\
&= exp\left\{-\int_0^t h(u)du\right\} \\
&= exp\left\{-H(t)\right\} \\
&= \frac{mrl(0)}{mrl(t)} exp\left[-\int_0^t \frac{du}{mrl(u)}\right] \\
f(t) &= -\frac{d}{dt}S(t) \\
&= h(t)S(t) \\
&= \left(\frac{d}{dt}mrl(t)+1\right)\left(\frac{mrl(0)}{mrl(t)^2}\right) exp\left[-\int_0^t \frac{du}{mrl(u)}\right] \\
h(t) &= -\frac{d}{dt} ln\left[S(t)\right] \\
&= \frac{f(t)}{S(t)} \\
&= \left[\frac{\frac{d}{dt}mrl(t)+1}{mrl(t)}\right] \\
mrl(t) &= \frac{\int_t^{\infty} S(u)du}{S(t)} \\
&= \frac{\int_t^{\infty}(u-t)f(u)du}{S(t)}
\end{aligned}
$$

2.5.2. Caso discreto

Las relaciones entre las cuatro funciónes básicas de supervivencia en el caso en que T una v.a. discreta que toma valores t_j con $j = 1, 2, ...$, puede ser resumida como sigue[2]:

$$\begin{aligned} S(t) &= \sum_{t_j > t} f(t_j) \\ &= \prod_{t > t_j} (1 - h(t_j)) \\ f(t_j) &= S(t_j) - S(t_{j+1}) \\ &= h(t_j) S(t_{j-1}) \\ h(t_j) &= \frac{f(t_j)}{S(t_{j-1})} \\ mrl(t) &= \frac{(t_{i+1} - t) S(t_i) + \sum_{j \geq i+1} (t_{j+1} - t_j) S(t_j)}{S(t)}. \end{aligned}$$

para $t_i \leq t < t_{i+1}$

Las relaciónes dadas para el caso continuo y discreto, y la interpretación que cada una tiene por separado forman una herramienta útil en el análisis de datos de supervivencia, por la información que de éstas se puede obtener para interpretar el tiempo de falla que presentan los sujetos de estudio.

2.6. Modelos paramétricos comunes

Algunos tiempos de falla pueden ser caracterizados por familias de distribuciones específicas que solo dependen de uno o varios parámetros desconocidos, los cuales proporcionan las características especificas del modelo en estudio. La selección de un modelo paramétrico es usualmente mediante la función de riesgo, pues de acuerdo a la información que el investigador tenga del fenómeno que causa la falla, puede determinar las características que el modelo debe seguir en la forma de la tasa de riesgo conforme avanza el tiempo. Por ejemplo, puede ser que el riesgo de muerte de un paciente después de someterse a alguna cirugía sea creciente las primeras horas y después (si sobrevive), su salud se estabilice hasta lograr su recuperación. En este caso, una función de riesgo creciente en valores pequeños del tiempo, que alcance

[2]Klein [6]. pag. 36.

un máximo y luego sea decreciente puede ser conveniente para modelar este fenómeno.

Utilizar un modelo paramétrico es restrictivo en el sentido de que se pueden exigir formas específicas del riesgo en el tiempo. Por ejemplo, el modelo exponencial que presenta riesgo constante, resultaría inadecuado para modelar el tiempo que tarda un individuo en morir cuando se le ha detectado una enfermedad terminal, pues en este caso, el riesgo debe ser claramente creciente. No obstante, puede haber situaciones donde se tenga evidencia para suponer que el riesgo puede ser constante en el tiempo, si fuera de interés modelar el tiempo que tarda en romperse la cuerda del violín de un concertista, puede ser que éste dependa de la dificultad de las piezas que el concertista tenga que tocar y el tiempo que invierta en practicar para perfeccionar el sonido, de modo que podría pensarse que la falla de la cuerda puede suceder en cualquier momento, independiente del tiempo que lleve colocada en el instrumento.

Debido a los criterios mencionados para seleccionar los modelos paramétricos adecuados, se presentan a continuación las distribuciones más comunes en modelos de supervivencia y una explicación detallada de la forma de su función de riesgo por la utilidad que ésta tiene en la selección del modelo.

2.6.1. Modelo Exponencial

Sa función de supervivencia está dada por:

$$S(t) = e^{-\lambda t}.$$

donde $\lambda > 0$.

Su función de densidad es:

$$f(t) = \lambda e^{-\lambda t}.$$

y es caracterizada por su función de riesgo constante,

$$h(t) = \lambda.$$

La distribución exponencial tiene la propiedad de pérdida de memoria, expresada como:

$$P(T \geq t + z | T \geq t) = P(T \geq z).$$

de la cual se sigue que la función de vida media residual es constante dada por:

$$E(T - t|T > t) = E(T) = \frac{1}{\lambda}.$$

De modo que el tiempo de ocurrencia de un evento no depende de lo que haya sucedido en el pasado, esta propiedad también es conocida como: propiedad de "*no-aging*" o como "*old as good as new*". La propiedad de pérdida de memoria también es reflejada en la interpretación de la función de riesgo constante, donde la probabilidad de falla a un tiempo t, dado que el evento no ha ocurrido antes, no tiene dependencia sobre t. Además, la distribución exponencial ha sido históricamente popular, pues la tasa de riesgo constante aparece de forma restrictiva en aplicaciones industriales y de salud.

Dado que la distribución exponencial es un caso particular de las distribuciones $Weibull$ y $Gamma$ consideradas más adelante, hereda propiedades de estas dos distribuciones.

2.6.2. Modelo Weibull

La función de supervivencia está dada por:

$$S(t) = e^{-\lambda t^{\alpha}}.$$

En esta distribución, $\lambda > 0$ es un parámetro de escala y $\alpha > 0$ es un parámetro de forma. La distribución exponencial es un caso particular cuando $\alpha = 1$.

Su función de densidad es:

$$f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}}.$$

y su función de riesgo esta dada por:

$$h(t) = \alpha\lambda t^{\alpha-1}.$$

Como se puede apreciar en la figura 2.4, esta función es convenientemente flexible siendo creciente (cuando $\alpha > 1$), decreciente (cuando $\alpha < 1$) y constante (cuando $\alpha = 1$), lo cual favorece a modelar el tiempo de falla para distintas formas del riesgo a través del tiempo. Es evidente que la forma de la distribución $Weibull$ depende del parámetro α, y ésta es la razón por la cual se le denomina parámetro de forma.

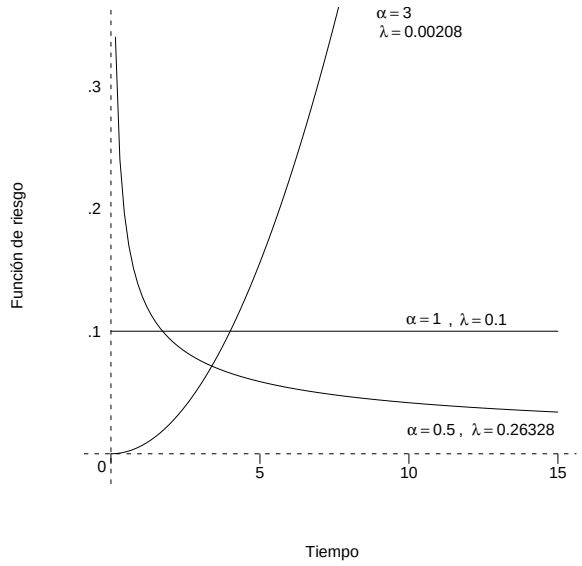


Figura 2.4: Función de riesgo de la distribución Weibull.

2.6.3. Modelo Log-normal

Se dice que la distribución de una variable aleatoria T es log-normal, cuando la distribución de su logaritmo $Y = \ln(T)$ tiene una distribución normal.

Su función de densidad queda completamente especificada por los parámetros μ y σ, los cuales corresponden a la media y varianza de Y, y está dada por:

$$\begin{aligned} f(t) &= \frac{\exp\left[-\frac{1}{2}\left(\frac{\ln t-\mu}{\sigma}\right)^2\right]}{t\,(2\pi)^{\frac{1}{2}}\,\sigma} \\ &= \frac{\phi\left(\frac{\ln t-\mu}{\sigma}\right)}{t}. \end{aligned}$$

donde $-\infty < \mu < \infty$ y $0 < \sigma < \infty$.

La función de supervivencia está dada por:

$$S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right).$$

donde Φ es la función de distribución acumulativa de una variable normal estándar.

La función de riesgo de la distribución log normal tiene una forma de "joroba", dado que toma el valor cero al tiempo cero, después crece a un máximo y decrece a cero cuando t tiende a infinito, esto se puede apreciar en la figura 2.5. Esta distribución ha sido criticada para modelar tiempos de falla dado que la función de riesgo es decreciente para valores grandes de t, lo cual es inaceptable en muchas situaciones. El modelo puede ser factible cuando valores grandes del tiempo no son de interés.

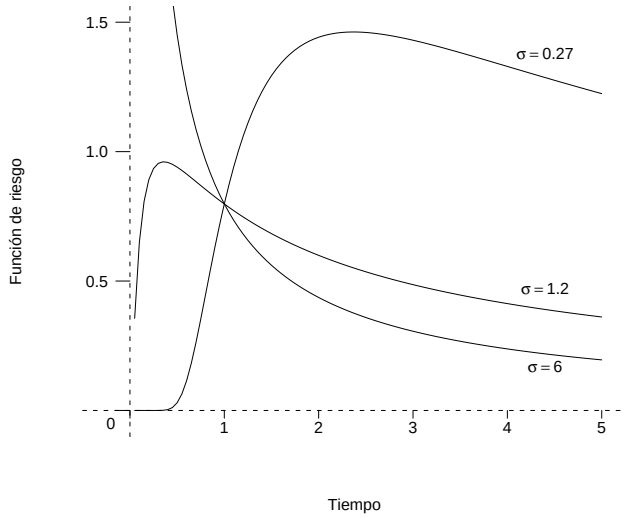


Figura 2.5: Función de riesgo de la distribución Log-normal con el parámetro $\mu = 0$.

2.6.4. Modelo Log-logístico

Cuando una variable aleatoria Y tiene distribución logística, su función de densidad está dada por:

$$f(y) = \frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^2}.$$

con $-\infty < y < \infty$ y donde μ y σ son, respectivamente, la media y el parámetro de escala de Y.

Una variable aleatoria T se dice que tiene una distribución log-logistic cuando su logaritmo, $Y = \ln T$ tiene una distribución logistic. Su función de supervivencia está dada por:

$$S(t) = \frac{1}{1+\lambda t^{\alpha}}.$$

Su función de densidad está dada por:

$$f(t) = \frac{\alpha\lambda t^{\alpha-1}}{\left[1+\lambda t^{\alpha}\right]^2}.$$

Y su función de riesgo está dada por:

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}.$$

donde $\alpha = \frac{1}{\sigma} > 0$ y $\lambda = \exp(-\frac{\mu}{\sigma})$.

El numerador de la función de riesgo es el mismo que el de la función de riesgo de la distribución Weibull, pero el denominador provoca que la función de riesgo posea las siguientes características: monótona decreciente cuando $\alpha \leq 1$, para $\alpha > 1$ es creciente inicialmente hasta alcanzar un máximo al tiempo $\left[(\alpha-1)/\lambda\right]^{\frac{1}{\alpha}}$ y entonces decrece hasta cero cuando el tiempo tiende a infinito como se puede ver en la figura 2.6.

2.6.5. Modelo Gamma

La distribución gamma tiene propiedades muy parecidas a las de la distribución *Weibull*, pero ésta no es matemáticamente fácil de tratar. Su función de densidad está dada por:

$$f(t) = \frac{\lambda^{\beta} t^{\beta-1}\exp(-\lambda t)}{\Gamma(\beta)}.$$

donde $\lambda > 0$, $\beta > 0$ y $\Gamma(\beta)$ es la función gamma dada por:

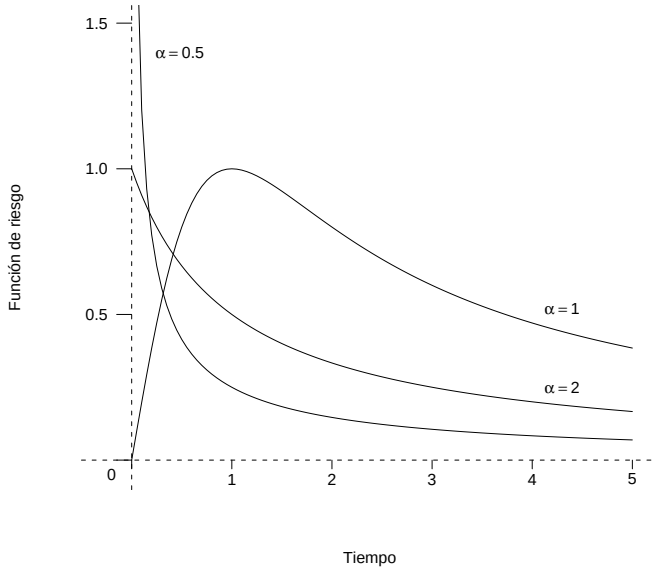


Figura 2.6: Función de riesgo de la distribución Log-logística con el parámetro $\lambda = 1$.

$$\Gamma(u) = \int_0^\infty u^{\alpha-1} \exp(-u) du.$$

Por razones similares a las de la distribución $Weibull$, λ es un parámetro de escala y β es llamado parámetro de forma. Esta distribución, al igual que la $Weibull$ incluye el caso exponencial cuando $\beta = 1$, se aproxima a una distribución normal cuando $\beta \to \infty$, y es una función $ji - cuadrada$ con v grados de libertad cuando $v = 2\beta$ (con β entero) y $\lambda = \frac{1}{2}$.

La función de supervivencia de una distribución gamma es expresada como:

$$\begin{aligned} S(t) &= \frac{\int_t^\infty \lambda(\lambda x)^{\beta-1} \exp(-\lambda x) dx}{\Gamma(\beta)} \\ &= 1 - \frac{\int_0^{\lambda t} (u)^{\beta-1} \exp(-u) du}{\Gamma(\beta)} \\ &= 1 - I(\lambda t, \beta). \end{aligned}$$

donde I es la función gamma incompleta dada por:

$$I(t,\beta) = \frac{\int_0^t (u)^{\beta-1}\exp(-u)du}{\Gamma(\beta)}.$$

La función de riesgo de la distribución gamma es monótona creciente para $\beta > 1$, con $h(0) = 0$ y con $h(t) \to \lambda$ cuando $t \to \infty$, y es monótona decreciente cuando $\beta < 1$, con $h(0) = \infty$ y $h(t) \to \lambda$ cuando $t \to \infty$, como se puede apreciar en la figura 2.7. Donde la expresión de ésta función de riesgo está dada por:

$$h(t) = \frac{\lambda^\beta t^{\beta-1}\exp(-\lambda t)}{\Gamma(\beta)\left[1 - I(\lambda t,\beta)\right]}.$$

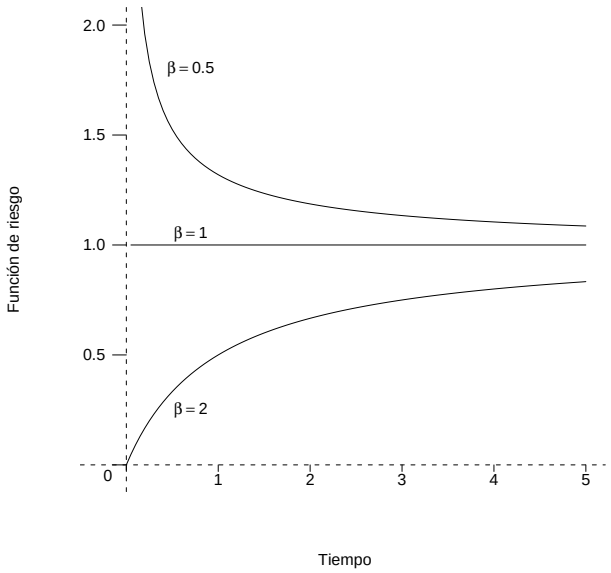


Figura 2.7: Función de riesgo de la distribución gamma con el parámetro $\lambda = 1$.

2.6.6. Modelo Erlang

La distribución Erlang se puede ver como un caso particular de la distribución gamma cuando $\beta = n$ con n entero. Por lo que este modelo hereda

las propiedades ya mencionadas en el modelo *gamma* bajo las respectivas restricciones. Su función de supervivencia está dada por:

$$S(t) = \exp(-\lambda t) \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!}.$$

Su función de densidad está dada por:

$$f(t) = \lambda \exp(-\lambda t) \frac{(\lambda t)^{n-1}}{(n-1)!}.$$

Donde la función de riesgo está dada como sigue:

$$h(t) = \lambda(\lambda t)^{n-1} \left[(n-1)! \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \right]^{-1}.$$

Cuadro 2.1: Función de riesgo, función de supervivencia y función de densidad de algunos modelos paramétricos de supervivencia comunes.

Distribución	Función de riesgo $h(t)$	Función de supervivencia $S(t)$	Función de densidad $f(t)$
Exponencial $\lambda > 0$ $t \geq 0$	λ	$\exp[-\lambda t]$	$\lambda \exp[-\lambda t]$
Weibull α , $\lambda > 0$ $t \geq 0$	$\alpha\lambda t^{\alpha-1}$	$\exp^{-\lambda t^{\alpha}}$	$\alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}$
Gamma β , $\lambda > 0$ $t \geq 0$	$\frac{f(t)}{S(t)}$	$1 - I(\lambda t, \beta)$	$\frac{\lambda^{\beta} t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}$
Log-normal $\mu \in R$,$\sigma > 0$ $t \geq 0$	$\frac{f(t)}{S(t)}$	$1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$	$\frac{\phi\left(\frac{\ln t - \mu}{\sigma}\right)}{t}$
Log-logistic α,$\lambda > 0$ $t \geq 0$	$\frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}$	$\frac{1}{1+\lambda t^{\alpha}}$	$\frac{\alpha\lambda t^{\alpha-1}}{[1+\lambda t^{\alpha}]^2}$
Gompertz θ, $\alpha > 0$ $t \geq 0$	$\theta e^{\alpha t}$	$\exp\left[\frac{\theta}{\alpha}\left(1 - e^{\alpha t}\right)\right]$	$\theta e^{\alpha t} \exp\left[\frac{\theta}{\alpha}\left(1 - e^{\alpha t}\right)\right]$
Pareto θ,$\lambda > 0$ $t \geq \lambda$	$\frac{\theta}{t}$	$\frac{\lambda^{\theta}}{t^{\theta}}$	$\frac{\theta\lambda^{\theta}}{t^{\theta+1}}$
Gamma generalizada β, $\lambda > 0$ $\alpha > 0$,$t \geq 0$	$\frac{f(t)}{S(t)}$	$1 - I(\lambda t^{\alpha}, \beta)$	$\frac{\alpha\lambda^{\beta} t^{\alpha\beta-1} \exp(-\lambda t^{\alpha})}{\Gamma(\beta)}$

Capítulo 3

Estimación de la función de supervivencia

En el análisis de datos, se presentan estadísticas que resumen la información tales como la media, el error estándar para la media, etc. En el análisis de datos de supervivencia sin embargo, debido a posibles censuras, las estadísticas que resumen la información pueden no tener las propiedades estadísticas deseadas, tales como insesgamiento. Por ejemplo, la media muestral ya no es un estimador insesgado de la media poblacional (del tiempo de supervivencia). De esta forma necesitamos usar otros métodos para presentar los datos. Una forma de estimar la verdadera distribución subyacente. Cuando esta distribución es estimada, podemos entonces estimar otras cantidades de interés tales como la media, la mediana, etc.

3.1. Caso sin censura

Suponga que se tiene una muestra de tiempos de supervivencia donde ninguna de las observaciones está censurada. La función de supervivencia $S(t)$ es la probabilidad de que un individuo sobreviva un tiempo mayor o igual a t. Esta función puede ser estimada por la función de supervivencia empírica dada por,

$$\tilde{S}(t) = \frac{\text{Número de individuos que sobreviven más alla de } t}{\text{Número total de individuos en el conjunto de datos}}$$

Equivalentemente $\tilde{S}(t) = 1 - \tilde{F}(t)$ donde $\tilde{F}(t)$ es la función de distribución empírica, la cual es el número de individuos que han fallado al tiempo t entre el número total de individuos. De esta definición empírica para la función de

supervivencia se tiene que $S(t) = 1$ para valores de t antes de la primera falla y que $S(t) = 0$ después del tiempo de la última muerte.

La función de supervivencia estimada $\tilde{S}(t)$ considerada constante entre dos tiempos de falla adyacentes, por lo que la gráfica de $\tilde{S}(t)$ es una función escalonada que decrece inmediatamente después de cada tiempo de supervivencia observado.

3.2. Estimador de Kaplan-Meier

Para determinar el estimador de Kaplan-Meier[1] de una función de supervivencia de una muestra que contiene datos censurados por la derecha, se forma una serie de intervalos de tiempo, donde cada intervalo contiene un tiempo de falla y se considera que esta falla ocurre al inicio del intervalo.

Por ejemplo, sean $t_{(1)}$, $t_{(2)}$ y $t_{(3)}$ tres tiempos de supervivencia observados, ordenados de tal modo que $t_{(1)} < t_{(2)} < t_{(3)}$ y sea c un tiempo de supervivencia censurado que ocurre entre $t_{(2)}$ y $t_{(3)}$. Los intervalos construidos comienzan en $t_{(1)}$, $t_{(2)}$ y $t_{(3)}$, y cada intervalo incluye un tiempo de falla, además estos pueden contener a más de un individuo que falle en algún tiempo en particular. Note que ningún intervalo comienza en el tiempo de censura c. Esta situación es ilustrada en la figura 3.1, en la cual D representa una falla y C un tiempo censurado, en ésta figura se tiene que dos individuos fallan en $t_{(1)}$, uno falla en $t_{(2)}$ y tres fallan en $t_{(3)}$.

El tiempo de origen es denotado por t_0, en el cual comienza el periodo inicial y termina justo antes de $t_{(1)}$, el tiempo de la primera muerte. Esto significa que el intervalo desde t_0 hasta $t_{(1)}$ no incluirá ninguna muerte. El primer intervalo construido comienza en $t_{(1)}$ y termina justo antes de $t_{(2)}$, este intervalo contiene por tanto al tiempo de muerte $t_{(1)}$. El segundo intervalo comienza en $t_{(2)}$ y termina justo antes de $t_{(3)}$ e incluye al tiempo de muerte $t_{(2)}$ y el tiempo de censura c. El tercer intervalo comienza en $t_{(3)}$ y contiene al mas largo tiempo de supervivencia, $t_{(3)}$.

En general, sean n individuos de los cuales son observados los tiempos de supervivencia t_1 , t_2,...t_n. Algunas de estas observaciones pueden ser censuradas por la derecha y también puede suceder que más de un individuo presente el mismo tiempo de supervivencia. Por tanto, sea r el número de tiempos de falla entre los individuos, donde $r \leq n$. Los r tiempos de falla

[1]Collet [2]. pag. 19.

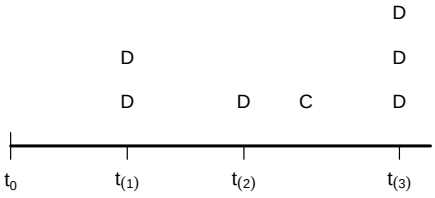


Figura 3.1: Tiempos de supervivencia para 7 individuos en tres tiempos de falla distintos

ordenados de forma ascendente serán $t_{(1)}$, $t_{(2)}$,...$t_{(r)}$ y el j-ésimo tiempo de falla será denotado por $t_{(j)}$ con $j = 1, 2, ..., r$. El número de individuos quienes están vivos justo antes del tiempo $t_{(j)}$, incluyendo aquéllos que están próximos a fallar a este tiempo, estarán denotados por n_j para $j = 1, 2, ..., r$, y d_j denotará el número de fallas a este tiempo. El tiempo del intervalo desde $t_{(j)} - \delta$ hasta $t_{(j)}$ donde δ es un intervalo de tiempo infinitesimal, tiene incluido un tiempo de muerte. Como hay n_j individuos que están vivos justo antes de $t_{(j)}$ y $d_{(j)}$ muertes al tiempo $t_{(j)}$, la probabilidad de que un individuo falle durante el intervalo $t_{(j)} - \delta$ a $t_{(j)}$, es estimado por $d_{(j)}/n_{(j)}$. El correspondiente estimador de la probabilidad de sobrevivir a través de este intervalo es entonces $\left(n_{(j)} - d_{(j)}\right)/n_{(j)}$. Algunas veces hay observaciones censuradas al mismo tiempo en que ocurre algún tiempo de falla, sucediendo que el tiempo de falla y el tiempo de supervivencia censurado ocurren simultáneamente. En este caso, el tiempo de supervivencia censurado es tomado como si ocurriera inmediatamente después del tiempo de muerte cuando se calcula el valor de $n_{(j)}$.

De la manera en la que los intervalos de tiempo han sido construidos, el intervalo de $t_{(j)}$ a $t_{(j+1)} - \delta$, el tiempo inmediatamente antes del siguiente tiempo de falla, no contiene fallas. Por tanto la probabilidad de sobrevivir de $t_{(j)}$ a $t_{(j+1)} - \delta$ es uno, y la probabilidad conjunta de sobrevivir de $t_{(j)} - \delta$ a $t_{(j)}$ y de $t_{(j)}$ a $t_{(j+1)} - \delta$ puede ser estimada por $\left(n_{(j)} - d_{(j)}\right)/n_{(j)}$. En el

limite cuando δ tiende a cero, $\left(n_{(j)} - d_{(j)}\right)/n_{(j)}$ llega a ser un estimador de la probabilidad de sobrevivir de $t_{(j)}$ a $t_{(j+1)}$.

Haciendo la suposición de que las fallas de los individuos en la muestra ocurren independientemente una de otra, el estimador para la función de supervivencia para cualquier tiempo en los k intervalos de tiempo construidos desde $t_{(k)}$ hasta $t_{(k+1)}$, k = 1,2,...,r, donde $t_{(r+1)}$ es definido como ∞, puede ser estimada la probabilidad de sobrevivir mas allá de $t_{(k)}$. Esta es la probabilidad de sobrevivir a través del intervalo desde $t_{(k)}$ a $t_{(k+1)}$ y todos los intervalos anteriores. Este es el estimador de Kaplan-Meier de la función de supervivencia , el cual está dado por

$$\tilde{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right)$$

Para $t_{(k)} \leq t \leq t_{(k+1)}$, k = 1,2,...,r, con $\tilde{S}(t) = 1$ para $t < t_{(1)}$ y donde $t_{(r+1)}$ es tomado como ∞. En sentido estricto, si la observación más grande es un tiempo de supervivencia censurado, $t*$, se dice que $\tilde{S}(t)$ está indefinida para $t > t*$. Por otro lado, si la observación más grande es un tiempo de supervivencia observado, $t_{(r)}$, es una observación no censurada, $n_r = d_r$, entonces $\tilde{S}(t)$ es cero para $t > t_{(r)}$. La gráfica del estimador Kaplan-Meier para la función de supervivencia es una función escalonada en la cual, la probabilidad de supervivencia estimada es constante entre tiempos adyacentes de falla y decreciente en cada tiempo de falla.

El estimador Kaplan-Meier también es conocido como el *estimador producto limite*. Es conveniente notar que si no hay tiempos censurados en el conjunto de datos, $n_j - d_j = n_{j+1}$, $j = 1, 2, ..., k$ y por tanto se tiene que

$$\tilde{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j=1}^{k} \frac{n_{j+1}}{n_j} = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times ... \times \frac{n_{k+1}}{n_k} = \frac{n_{k+1}}{n_1}$$

para k = 1,2,...,$r-1$. Donde $\tilde{S}(t) = 1$ para $t < t_{(1)}$ y $\tilde{S}(t) = 0$ para $t > t_{(r)}$. Ahora n_1 es el número de individuos en riesgo justo antes del primer tiempo de falla, el cual es el número de individuos en la muestra y n_{k+1} es el número de individuos tales que su tiempo de supervivencia es mayor o igual a $t_{(k+1)}$. Consecuentemente, se tiene que en ausencia de censura, $\tilde{S}(t)$ es simplemente la función de supervivencia empírica definida anteriormente.

3.3. Estimación de la función de supervivencia con R

El estimador de Kaplan-Maier para la función de supervivencia es obtenido en el paquete estadístico **R** mediante la función *survfit*. Esta función en su forma más sencilla, solo requiere un objeto de supervivencia creado por la función *Surv* , denotado por *formula*, y retorna un objeto *survfit*.

```
survfit(formula)
```

Suponga que se tiene la estructura de datos siguiente

```
> Datos
 [1]  1   2   3   4   5   6+  7+  8+  9+ 10+
```

El objeto *survfit* con la información del estimador de Kaplan-Meier para la función de supervivencia es obtenida como sigue

```
> survfit (Datos)
Call: survfit(formula = Datos)

      n  events  median 0.95LCL 0.95UCL
   10.0     5.0     7.5     3.0     Inf
```

La función por si sola, retorna relativamente poca información, n corresponde al número de individuos en estudio, *events* corresponde al numero de fallas presentadas en los n individuos (por tanto, el número de censuras está dado por $n - events$), *median* es el tiempo mediano antes de que se presente la falla con respecto a la curva de la función de supervivencia estimada (el tiempo t tal que $S(t) = .5$), $0.95LCL$ es limite inferior de una banda estimada con un 95 % de confianza para el valor de *median* y $0.95UCL$ es limite superior de una banda estimada con un 95 % de confianza para el valor de *median* (esta banda es estimada por el método ordinario que se verá más adelante).

En el caso del ejemplo, se tienen 10 sujetos de estudio, de los cuales 5 presentan falla y 5 presentan censura, el tiempo mediano antes de presentar falla es de 7.5 y un intervalo con un 95 % de confianza está dado en su limite inferior por 3 y en su limite superior por 1. (en los intervalos de confianza obtenidos por la función *survfit* , Inf toma el valor de 1 y $-inf$ toma el valor de 0).

Para obtener más información que se encuentra en el objeto creado por la función *survfit*, se puede usar la función *summary* como sigue (por comodidad, el objeto creado por la función *survfit* será guardado en la variable K_M).

```
> K_M <- survfit (Datos)

> summary(K_M)
Call: survfit(formula = Datos)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     10       1      0.9  0.0949        0.732        1.000
    2      9       1      0.8  0.1265        0.587        1.000
    3      8       1      0.7  0.1449        0.467        1.000
    4      7       1      0.6  0.1549        0.362        0.995
    5      6       1      0.5  0.1581        0.269        0.929
```

La información que se despliega al utilizar la función *summary* resulta más completa, *time* representa el tiempo para el que se presenta la información (anteriormente, la información se presentaba para el tiempo dado por *median*), *n.risk* indica la cardinalidad del conjunto en riesgo o el número de individuos que continua en estudio al tiempo correspondiente, *n.event* corresponde al número de fallas que se presentan entre cada tiempo, *survival* indica el valor que toma la función de supervivencia estimada por el método Kaplan-Meier en el tiempo correspondiente, *std.err* corresponde al error estándar estimado para la función de supervivencia en el tiempo respectivo (el error estandar se menciona más adelante) y finalmente $lower95\,\%CI$ y $upper95\,\%CI$ denotan el intervalo de confianza para la función de supervivencia en cada tiempo como se mencionó anteriormente.

Para obtener la información dada por *summary* en tiempos específicos, se guardan dichos tiempos en un vector, y se usa como segundo argumento al usar la función *summary*. Suponiendo que es de interés la información en los tiempos $0, .5, 1, 2, 3.7, 4, 5, 6, 10, 11, 20$ esta se obtiene de la siguiente manera

```
> time <- c(0,.5,1,2,3.7,4,5,6,10,11,20)

> summary(K_M,time)
```

```
Call: survfit(formula = Datos)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  0.0     10       0      1.0  0.0000        1.000        1.000
  0.5     10       0      1.0  0.0000        1.000        1.000
  1.0     10       1      0.9  0.0949        0.732        1.000
  2.0      9       1      0.8  0.1265        0.587        1.000
  3.7      7       1      0.7  0.1449        0.467        1.000
  4.0      7       1      0.6  0.1549        0.362        0.995
  5.0      6       1      0.5  0.1581        0.269        0.929
  6.0      5       0      0.5  0.1581        0.269        0.929
 10.0      1       0      0.5  0.1581        0.269        0.929
```

Es importante notar que se deja de presentar la información para los tiempo mayores a 10 puesto que este es el tiempo mayor en la base de datos, además es conveniente apreciar que la información se puede obtener para cualquier número mayor o igual a cero.

Una forma de obtener la información contenida en el objeto *survfit*, que facilite su almacenamiento en variables para posibles necesidades del usuario se presenta como sigue.

Utilizando la función *names* se pueden ver los atributos del objeto *survfit* (en este caso, el objeto *survfit* tiene el nombre de K_M).

```
> names(K_M)
 [1] "n"       "time"   "n.risk"  "n.event"   "surv"     "type"
 [7] "std.err" "upper"  "lower"   "conf.type" "conf.int" "call"
```

De modo que para obtener los tiempos de falla se puede usar la siguiente combinación:

```
> summary(K_M)$time
[1] 1 2 3 4 5
```

O para obtener los datos de supervivencia de los tiempos de falla se hace como sigue:

```
> summary(K_M)$surv
[1] 0.9 0.8 0.7 0.6 0.5
```

La ventaja de usar esta función, es que los datos obtenidos quedan guardados en la variable de nombre $summary(K\_M)\$surv$, en este caso, los valores de la función de supervivencia para los tiempos de falla.

En **R** se puede obtener la gráfica de la función de supervivencia aplicando la función *plot* al objeto *survfit* como se ilustra a continuación

```
> plot(K_M, conf.int=F )
```

obteniendo la gráfica de la figura 3.2, que corresponde a la forma más sencilla de presentar la función de supervivencia estimada. Se usa como segundo argumento $conf.int = F$ debido a que no se desea que muestre el intervalo de confianza estimado para la función de supervivencia.

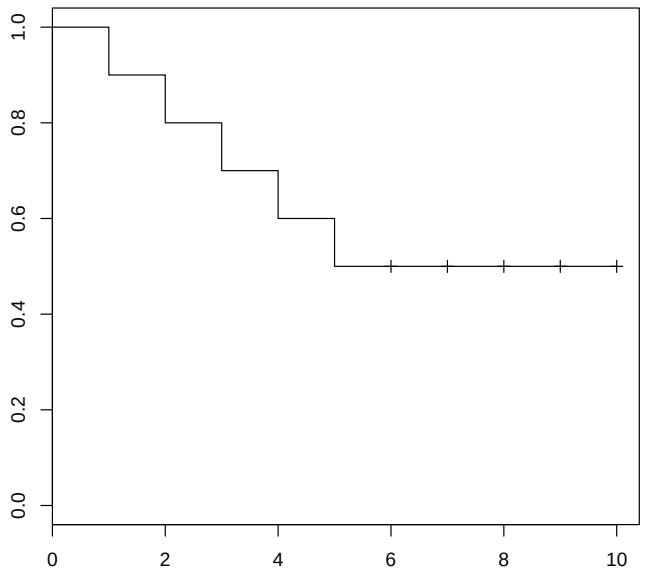


Figura 3.2: Función de supervivencia estimada.

Para mejorar la presentación de la gráfica de la función de supervivencia, se pueden utilizar todos los recursos que tiene la función *plot* para obtener un

formato deseado. Esto se ilustra en la figura 3.3 donde se utilizó el siguiente código:

```
plot(K_M,conf.int=F, main = "Función de supervivencia",
xlab= "Tiempo ",ylab=" Probabilidad de supervivencia",lwd=2,
col ="blue")
box(lwd=3, col = 'black')
axis(1, seq(0,10,1))
axis(2, seq(0,1,.1))
abline(h = seq(0,1,.1), v =seq(0,10,.5),lty=3,
col ="gray")
legend("bottomleft",c("Curva de supervivencia"),lty=1,
lty=1,col="blue")
```

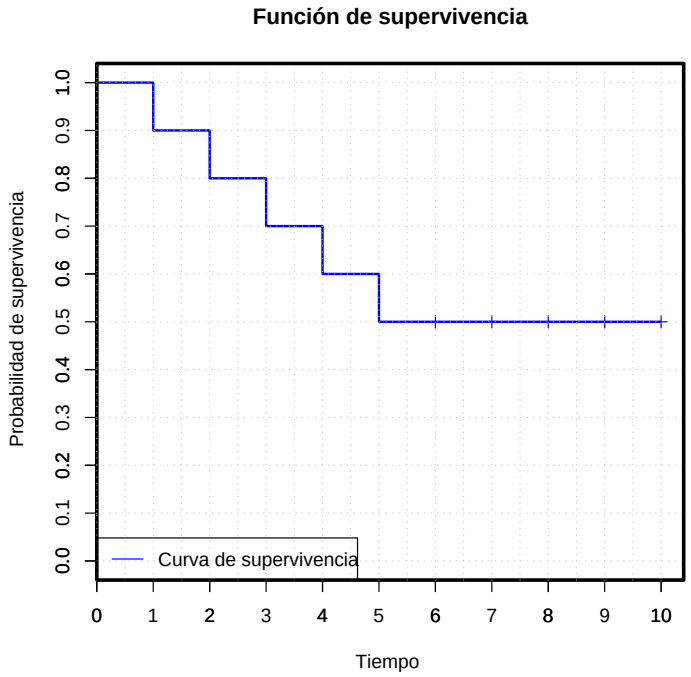


Figura 3.3: Función de supervivencia estimada con formato de presentación.

Para obtener la gráfica de la función acumulativa de riesgo, se utiliza como argumento de la función *plot*, $fun =$ "*cumhaz*" como se ilustra en el siguiente código, obteniendo la figura 3.4.

```
plot(K_M,conf.int=F,fun="cumhaz")
```

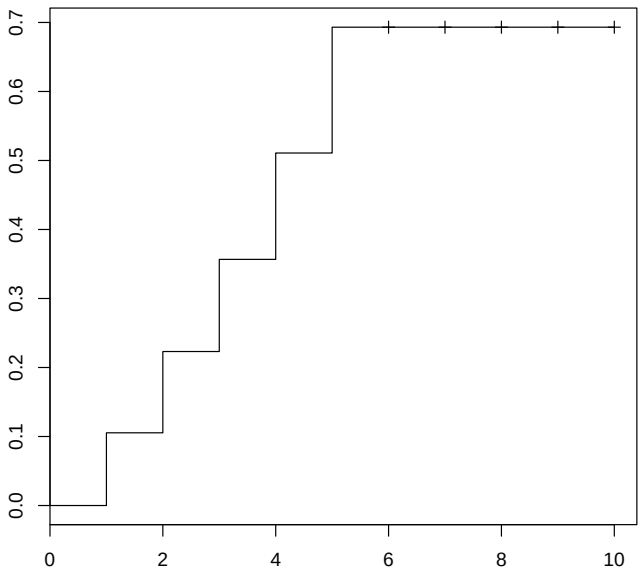


Figura 3.4: Función de riesgo acumulado para el modelo de supervivencia estimado.

Donde la figura 3.4 puede ser mejorada en cuanto a presentación de forma similar a la figura 3.3.

3.3. ESTIMACIÓN DE LA FUNCIÓN DE SUPERVIVENCIA CON R

Puede ser de interés observar las curvas de supervivencia de distintos grupos en una misma gráfica y esto se puede hacer de la siguiente manera:

Suponga que se tiene una base de datos de la siguiente forma, donde se tiene la información de tiempo evento para tres distintos grupos.

Tiempo	status	grupo	Tiempo	status	grupo	Tiempo	status	grupo
3	0	A	6	1	B	12	1	C
6	1	A	4	0	B	6	0	C
5	1	A	3	1	B	18	1	C
8	1	A	8	0	B	14	1	C
9	0	A	6	1	B	16	1	C
3	0	A	12	0	B	19	0	C
4	1	A	11	1	B	14	1	C
8	1	A	15	0	B	17	1	C
5	0	A	7	1	B	13	0	C

Los datos son puestos en un objeto que **R** identifique como tabla de la siguiente forma:

```
> tabla
   Tiempo status grupo
1       3      0     A
2       6      1     A
3       5      1     A
4       8      1     A
5       9      0     A
6       3      0     A
7       4      1     A
8       8      1     A
9       5      0     A
10      6      1     B
11      4      0     B
12      3      1     B
13      8      0     B
14      6      1     B
```

```
15      12      0      B
16      11      1      B
17      15      0      B
18       7      1      B
19      12      1      C
20       6      0      C
21      18      1      C
22      14      1      C
23      16      1      C
24      19      0      C
25      14      1      C
26      17      1      C
27      13      0      C
```

La manera de obtener la curva de supervivencia para estos tres grupos vistos como un solo grupo de estudio es como sigue:

```
plot(survfit(Surv(Tiempo,status)),conf.int=F,
     main="Funcion de supervivencia",
     xlab= "Tiempo de falla",
     ylab="Probabilidad de supervivencia")
```

Obteniendo la gráfica de la figura 3.5.

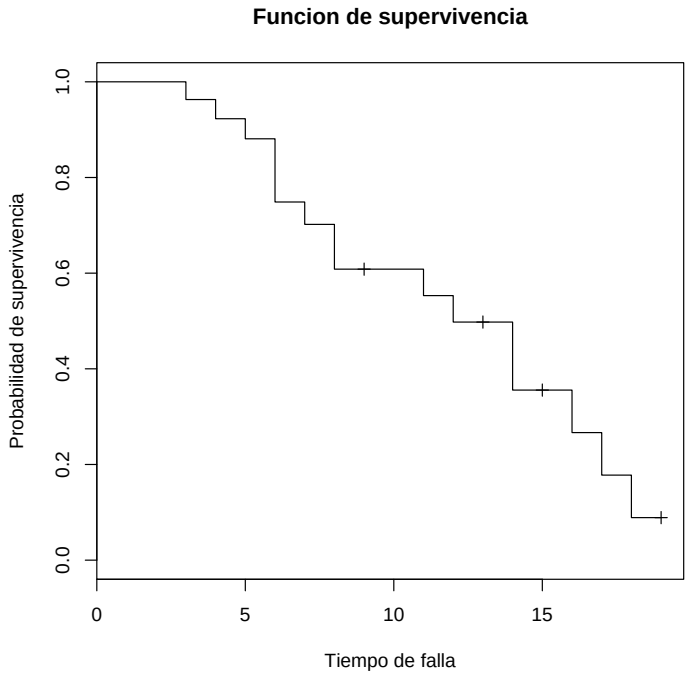


Figura 3.5: Función de supervivencia estimada considerando los datos como un solo grupo.

La manera de obtener las curvas de supervivencia para estos tres grupos por separado es como sigue:

```
plot(survfit(Surv(Tiempo,status)~grupo), col=2:4,
    main="Funcion de supervivencia por grupo",
    xlab= "Tiempo de falla" ,
    ylab="Probabilidad de supervivencia")
    legend("bottomleft",c("Grupo A","Grupo B","Grupo C"),
    lty=c(1,1,1),col=2:4)
```

Obteniendo la gráfica de la figura 3.6.

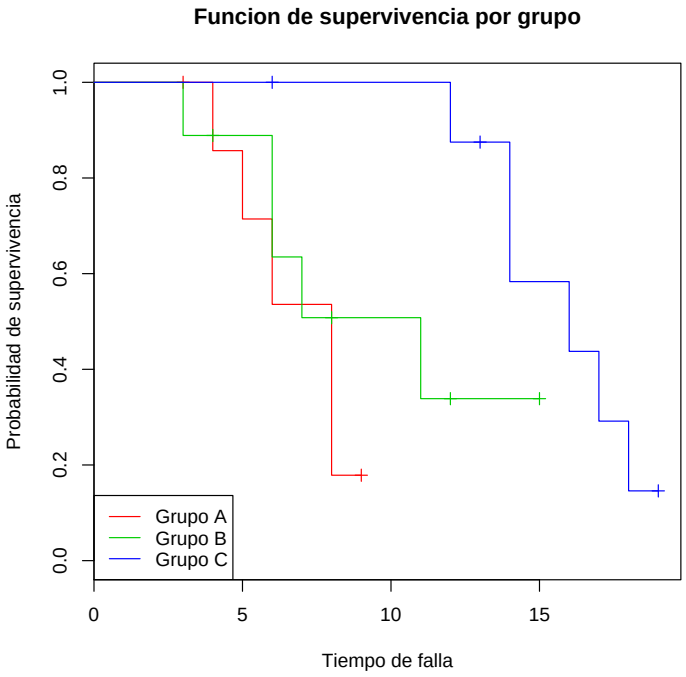


Figura 3.6: Función de supervivencia estimada para los tres grupos por separado.

3.4. Bandas de confianza para la función de supervivencia

Se pueden obtener intervalos de confianza puntuales para la función de supervivencia. Estos intervalos son válidos para un sólo valor del tiempo para el cual se va a hacer la inferencia.

En algunas aplicaciones resulta de interés obtener bandas de confianza que garanticen, con un nivel dado de confiabilidad, que la función de supervivencia caiga dentro de la banda para toda t en un intervalo.

3.4.1. Error estándar del estimador Kaplan-Meier

El estimador Kaplan-Meier de la función de supervivencia para cualquier valor de t en el intervalo que comienza en t_k y termina en t_{k+1} puede ser escrito como

$$\tilde{S}(t) = \prod_{j=1}^{k} \tilde{p_j}$$

para k = 1,2,...,r donde $\tilde{p_j} = \frac{n_j - d_j}{n_j}$ es la probabilidad estimada de que un individuo sobreviva a través del intervalo de tiempo que comienza en $t_{(j)}$ para j = 1,2,...,r.

Tomando logaritmo

$$\log \tilde{S}(t) = \sum_{j=1}^{k} \log \tilde{p_j}.$$

Y entonces, dado que la probabilidad de falla es independiente entre cada intervalo de tiempo que comienza en $t_{(j)}$ para j = 1,2,...,k cuando t está en el intervalo que comienza en t_k y termina en t_{k+1}, la varianza de $\log \tilde{S}(t)$ está dada por

$$Var\left\{\log \tilde{S}(t)\right\} = \sum_{j=1}^{k} Var\left\{\log \tilde{p_j}\right\}.$$

El número de individuos que sobreviven a través del intervalo que comienza en $t_{(j)}$ puede suponerse que sigue una distribución *Binomial* con parámetros n_j y p_j, donde p_j es la probabilidad de sobrevivir a través del intervalo. El número observado de sobrevivientes está dado por $n_j - d_j$ y utilizando el resultado de que la varianza de una variable aleatoria *Binomial* con parámetros n y p es $np(1-p)$, entonces, la varianza de $n_j - d_j$ toma la siguiente expresión

$$Var\left(n_j - d_j\right) = n_j p_j (1 - p_j).$$

Dado que $\tilde{p_j} = \frac{n_j - d_j}{n_j}$, se tiene que

$$Var\left(\tilde{p_j}\right) = Var\left(\frac{n_j - d_j}{n_j}\right) = \frac{Var\left(n_j - d_j\right)}{n_j^2} = \frac{p_j(1-p_j)}{n_j}.$$

Por tanto, la varianza de $\tilde{p_j}$ puede ser estimada por

$$\frac{\tilde{p_j}(1-\tilde{p_j})}{n_j}.$$

Para obtener la varianza de $\log \tilde{p_j}$, se hará uso de un resultado general para la aproximación de la varianza de una función de una variable aleatoria. De acuerdo con este resultado, la varianza de una función $g(X)$, donde X es una variable aleatoria, está dada por

$$Var\left\{g(X)\right\} \approx \left\{\frac{dg(X)}{dX}\right\}^2 Var(X).$$

Este resultado es conocido como la *aproximación de series de Taylor*[2] para la varianza de una función de una variable aleatoria.

Usando este resultado se tiene que

$$Var\left\{\log \tilde{p_j}\right\} \approx \left\{\frac{d\log \tilde{p_j}}{d\tilde{p_j}}\right\}^2 Var(\tilde{p_j}) = \left\{\frac{1}{\tilde{p_j}}\right\}^2 Var(\tilde{p_j}).$$

De modo que al utilizar la estimación de la varianza de $\tilde{p_j}$, la aproximación de la varianza estimada de $\log \tilde{p_j}$ queda expresada como

$$\left\{\frac{1}{\tilde{p_j}}\right\}^2 \frac{\tilde{p_j}\left(1-\tilde{p_j}\right)}{n_j},$$

de modo que al sustituir $\tilde{p_j}$ por $\frac{n_j-d_j}{n_j}$ se tiene que la varianza estimada de $log\tilde{p_j}$ queda aproximada por

$$\frac{d_j}{n_j\left(n_j-d_j\right)},$$

de modo que

$$Var\left\{\log \tilde{S}(t)\right\} = \sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j-d_j\right)}.$$

Aplicando nuevamente la *aproximaciòn de series de Taylor* se tiene

$$Var\left\{\log \tilde{S}(t)\right\} \approx \frac{1}{\left[\tilde{S}(t)\right]^2} Var\left\{\tilde{S}(t)\right\},$$

de modo que

[2]Collet [2].pag. 23.

$$Var\left\{\tilde{S}(t)\right\} \approx \left[\tilde{S}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)}.$$

Finalmente, el error estándar del estimador Kaplan-Meier para la función de supervivencia queda definido como

$$s.e.\left\{\tilde{S}(t)\right\} \approx \left[\tilde{S}(t)\right] \left\{\sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)}\right\}^{\frac{1}{2}}$$

para los valores de t tal que $t_{(k)} \leq t \leq t_{(k+1)}$. Este resultado es conocido como la *fórmula de Greenwood*[3].

La aproximación de la varianza de $log(-log(\tilde{S}(t)))$ es necesaria para obtener la banda de confianza "log log", y se obtiene utilizando nuevamente la *aproximación de series de Taylor* para la varianza de una variable aleatoria dada por

$$Var\left\{g(X)\right\} \approx \left\{\frac{dg(X)}{dX}\right\}^2 Var(X)$$

donde $X = \log \tilde{S}(t)$ y $g(X) = \log(-X)$, de modo que

$$Var\left[\log(-\log(\tilde{S}(t)))\right] \approx \left[\frac{1}{\left(\log \tilde{S}(t)\right)\tilde{S}(t)}\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)}$$

Por tanto, el $s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}$ está dado por la raiz cuadrada de la $Var\left[\log(-\log(\tilde{S}(t)))\right]$ de modo que

$$s.e.\left\{\log(-\log(\tilde{S}(t)))\right\} \approx \frac{1}{\left(\log \tilde{S}(t)\right)\tilde{S}(t)} \left[\sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)}\right]^{\frac{1}{2}}.$$

3.4.2. Intervalo de confianza para la función de supervivencia

Este intervalo de confianza para el verdadero valor de la función de supervivencia al tiempo t, es obtenido con el supuesto de que el valor estimado de

[3]Collet [2]. pag. 23.

la función de supervivencia al tiempo t, $\tilde{S}(t)$, se distribuye asintoticamente, normal con media $S(t)$ y varianza estimada dada por $Var\left\{\tilde{S}(t)\right\}$. Por tanto, un intervalo de un $100(1-\alpha)\,\%$ de confianza para $S(t)$, para un valor específico de t está dado por

$$\left(\tilde{S}(t) - z_{1-\alpha/2} s.e.\left\{\tilde{S}(t)\right\}, \tilde{S}(t) + z_{1-\alpha/2} s.e.\left\{\tilde{S}(t)\right\}\right)$$

donde el $s.e.\left\{\tilde{S}(t)\right\}$ está dado por la *fórmula de Greenwood* y $z_{1-\alpha/2}$ es el cuantil que acumula una probabilidad de $1-\alpha/2$ en una distribución normal estándar.

3.4.3. Intervalo de confianza para la función de supervivencia obtenido con R

La manera de especificar el tipo de intervalo de confianza que se desea calcular, es por medio del argumento *conf.type* dentro de la función *survfit*.

La estructura para obtener la función de supervivencia estimada por medio del estimador de Kaplan-Meier y su la banda de confianza , es como sigue

```
survfit(formula, conf.type = "plain")
```

De modo que si se tiene la siguiente estructura de datos de supervivencia:

```
> Datos
 [1]  1   2   3   4   5   6+  7+  8+  9+ 10+
```

La función estimada y su intervalo de confianza se obtiene como sigue:

```
K_M.ordinario <- survfit (Datos,conf.type = "plain" )
```

Y la información se puede obtener como sigue:

```
> summary(K_M.ordinario)
Call: survfit(formula = Datos, conf.type = "plain")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	10	1	0.9	0.0949	0.714	1.000
2	9	1	0.8	0.1265	0.552	1.000
3	8	1	0.7	0.1449	0.416	0.984
4	7	1	0.6	0.1549	0.296	0.904
5	6	1	0.5	0.1581	0.190	0.810

Estos resultados se pueden obtener a pesar de omitir el argumento $conf.type$ = "$plain$" en la función $survfit$, debido a que éste intervalo es usado por el programa de forma predeterminada. No obstante, es conveniente indicarlo con la finalidad de tener claro el tipo de intervalo que se está estimando y contemplando el hecho de que será modificado posteriormente.

Esta banda de confianza se puede observar gráficamente, conjuntamente con la función de supervivencia estimada mediante la siguiente instrucción:

```
plot(K_M.ordinario,xlab= "Tiempo de falla",
     ylab="Probabilidad de supervivencia")
```

Obteniendo la gráfica de la figura 3.7

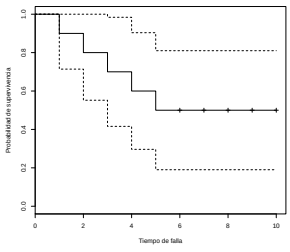


Figura 3.7: Función de supervivencia estimada con su banda de confianza.

Las bandas de confianza son calculadas de forma predeterminada con un 95 % de confianza, pero esto puede ser modificado mediante el argumento $conf.int$ en la función $survfit$, obteniendo la estructura siguiente:

```
survfit(formula, conf.type = "plain",conf.int = .95 )
```

Donde el .95 puede ser cambiado por cualquier número entre cero y uno que corresponda a la confianza deseada por el investigador.

3.4.4. Intervalo de confianza usando la transformación log

Una forma alternativa de estimar un intervalo de confianza para la función de supervivencia es transformando el valor de $\tilde{S}(t)$ en el rango $(-\infty, \infty)$ con la transformación $\log(S(t))$ y obtener un intervalo de confianza para este valor transformado. El resultado induce un intervalo de confianza para $S(t)$.

Este intervalo de confianza para el verdadero valor del logaritmo de la función de supervivencia al tiempo t, es obtenido con el supuesto de que sigue una distribución asintótica normal con media $\log(S(t))$ y varianza estimada por $\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}$ es una aproximación de la distribución de $\log(\tilde{S}(t))$. Por tanto, un intervalo de un $100(1-\alpha)\,\%$ de confianza para $\log(S(t))$, para un valor específico de t está dado por:

$$\left(\log(\tilde{S}(t)) - z_{1-\alpha/2} s.e.\left\{\log(\tilde{S}(t))\right\}, \log(\tilde{S}(t)) + z_{1-\alpha/2} s.e.\left\{\log(\tilde{S}(t))\right\}\right)$$

donde

$$s.e.\left\{\log(\tilde{S}(t))\right\} = \left[\sum_{j=1}^{k} \frac{d_j}{n_j\,(n_j - d_j)}\right]^{\frac{1}{2}}$$

y $z_{1-\alpha/2}$ es el cuantil que acumula $1-\alpha/2$ de probabilidad en una distribución normal estándar.

Este intervalo induce una banda de confianza para $S(t)$ dado que

$$\log(\tilde{S}(t)) - z_{1-\alpha/2} s.e.\left\{\log(\tilde{S}(t))\right\}$$

$$\leq \log(S(t))$$

$$\leq \log(\tilde{S}(t)) + z_{1-\alpha/2} s.e.\left\{\log(\tilde{S}(t))\right\}.$$

Desigualdad que al aplicarle la función exponencial se tiene que

$$exp\left[\log(\tilde{S}(t)) - z_{1-\alpha/2}s.e.\left\{\log(\tilde{S}(t))\right\}\right]$$

$$\leq S(t)$$

$$\leq exp\left[\log(\tilde{S}(t)) + z_{1-\alpha/2}s.e.\left\{\log(\tilde{S}(t))\right\}\right],$$

y por tanto

$$\tilde{S}(t)exp\left[-z_{1-\alpha/2}s.e.\left\{\log(\tilde{S}(t))\right\}\right]$$

$$\leq S(t)$$

$$\leq \tilde{S}(t)exp\left[+z_{1-\alpha/2}s.e.\left\{\log(\tilde{S}(t))\right\}\right].$$

De modo que, en éste caso, una banda de confianza para el verdadero valor de $S(t)$ de un $100(1-\alpha)\,\%$ de confianza para un valor específico de t está dado por

$$\left(\tilde{S}(t)exp\left[-z_{1-\alpha/2}s.e.\left\{\log(\tilde{S}(t))\right\}\right], \tilde{S}(t)exp\left[+z_{1-\alpha/2}s.e.\left\{\log(\tilde{S}(t))\right\}\right]\right)$$

Este intervalo es una forma de crear una banda de confianza para la función acumulativa de riesgo $H(t)$ puesto que $H(t) = -\log(S(t))$.

3.4.5. Intervalo de confianza usando la transformación log obtenido con R

La estructura para obtener la función de supervivencia estimada por medio del estimador Kaplan-Meier y la banda de confianza usando la transformación log, es como sigue:

```
survfit(formula, conf.type = "log")
```

De modo que si se tiene la siguiente estructura de datos de supervivencia

```
> Datos
 [1]  1   2   3   4   5   6+  7+  8+  9+ 10+
```

La función estimada y su intervalo de confianza se obtiene como sigue

```
K_M.log <- survfit (Datos,conf.type = "log" )
```

Y la información se puede obtener como sigue

```
> summary(K_M.log)
Call: survfit(formula = Datos, conf.type = "log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     10       1      0.9  0.0949        0.732        1.000
    2      9       1      0.8  0.1265        0.587        1.000
    3      8       1      0.7  0.1449        0.467        1.000
    4      7       1      0.6  0.1549        0.362        0.995
    5      6       1      0.5  0.1581        0.269        0.929
```

Esta banda de confianza se puede observar gráficamente, conjuntamente con la función de supervivencia estimada mediante la siguiente instrucción

```
plot(K_M.log,xlab= "Tiempo de falla",
      ylab="Probabilidad de supervivencia")
```

Obteniendo la gráfica de la figura 3.8.

Las bandas de confianza son calculadas de forma predeterminada con un 95 % de confianza, pero esto puede ser modificado mediante el argumento $conf.int$ en la función $survfit$, obteniendo la estructura siguiente

```
survfit(formula, conf.type = "plain",conf.int = .95 )
```

Donde el .95 puede ser cambiado por cualquier número entre cero y uno que corresponda a la confianza deseada por el investigador.

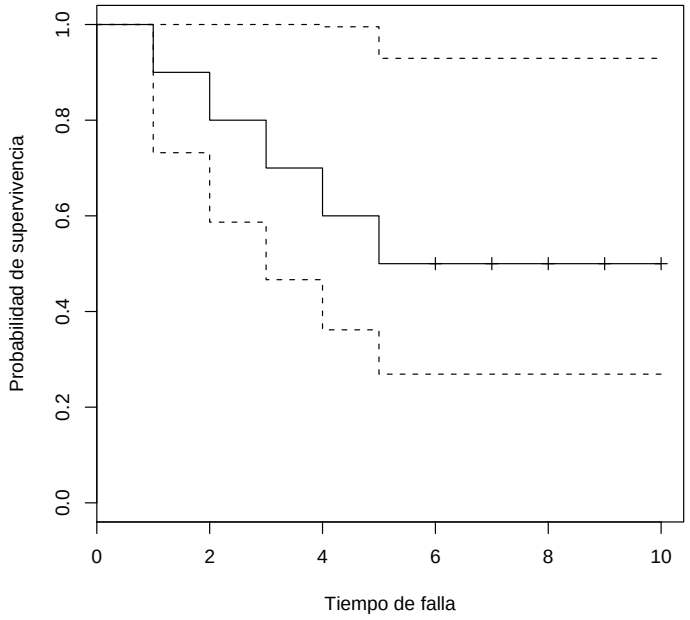


Figura 3.8: Función de supervivencia estimada con su banda de confianza, usando la transformación log.

3.4.6. Intervalo de confianza usando la transformación log-log

Otra forma alternativa de estimar un intervalo de confianza para la función de supervivencia es transformando el valor de $\tilde{S}(t)$ en el rango $(-\infty, \infty)$ con la transformación *log-log* dada por: $\log(-\log(S(t)))$, y obtener un intervalo de confianza para este valor transformado. El resultado induce un intervalo de confianza para $S(t)$.

Este intervalo de confianza para el verdadero valor del $\log(-\log(S(t)))$ al tiempo t, es obtenido con el supuesto de que una distribución asintótica normal con media $log(-log(S(t)))$ y desviación estandar estimada por la expreción de $s.e\log(-\log(\tilde{S}(t)))$ es una aproximación de la distribución de $\log(-\log(\tilde{S}(t)))$. Por tanto, un intervalo de un $100(1-\alpha)\,\%$ de confianza para $\log(-\log(S(t)))$, para un valor específico de t está dado por

$$\Big(\log(-\log(\tilde{S}(t))) - z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}$$

$$, \log(-\log(\tilde{S}(t))) + z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\Big)$$

donde $z_{1-\alpha/2}$ es el cuantil que acumula $1-\alpha/2$ de probabilidad en una distribución normal estándar y

$$s.e.\left\{\log(-\log(\tilde{S}(t)))\right\} \approx \frac{1}{\log \tilde{S}(t)} \left[\sum_{j=1}^{k} \frac{d_j}{n_j\,(n_j - d_j)}\right]^{\frac{1}{2}}.$$

Este intervalo induce una banda de confianza para $S(t)$ dado que:

$$\log(-\log(\tilde{S}(t))) - z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}$$

$$\leq \log(-\log(S(t)))$$

$$\leq \log(-\log(\tilde{S}(t))) + z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}$$

Desigualdad que al aplicarle la función exponencial se tiene que:

$$-\log(\tilde{S}(t)) exp\left[-z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]$$

$$\leq -\log(S(t))$$

$$\leq -\log(\tilde{S}(t)) exp\left[+z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]$$

multiplicando por menos y usando que $a * \log(b) = \log(b^a)$ cuando a y b son constantes, se tiene que

$$\log\left(\tilde{S}(t)^{exp\left[+z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]}\right)$$

$$\leq \log(S(t))$$

$$\leq \log\left(\tilde{S}(t)^{exp\left[-z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]}\right)$$

y por tanto, aplicando nuevamente la función exponencial se tieme que

$$\tilde{S}(t)^{exp\left[+z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]} \leq S(t) \leq \tilde{S}(t)^{exp\left[-z_{1-\alpha/2} s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]}$$

De modo que, en este caso, una banda de confianza para el verdadero valor de $S(t)$ de un $100(1-\alpha)\,\%$ de confianza para un valor especifico de t esta dado por

$$\left(\tilde{S}(t)^{exp\left[+z_{1-\alpha/2}s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]}, \tilde{S}(t)^{exp\left[-z_{1-\alpha/2}s.e.\left\{\log(-\log(\tilde{S}(t)))\right\}\right]}\right).$$

3.4.7. Intervalo de confianza usando la transformación log-log obtenido con R

La estructura para obtener la función de supervivencia estimada por medio del estimador Kaplan-Meier y la banda de confianza usando la transformación log-log $\log\log$, es como sigue:

```
survfit(formula, conf.type = "log-log")
```

De modo que si se tiene la siguiente estructura de datos de supervivencia

```
> Datos
 [1]  1   2   3   4   5   6+  7+  8+  9+ 10+
```

La función estimada y su intervalo de confianza se obtiene como sigue:

```
K_M.loglog <- survfit (Datos,conf.type = "log-log" )
```

Y la información se puede obtener como sigue:

```
> summary(K_M.loglog)
Call: survfit(formula = Datos, conf.type = "log-log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     10       1      0.9  0.0949        0.473        0.985
    2      9       1      0.8  0.1265        0.409        0.946
    3      8       1      0.7  0.1449        0.329        0.892
    4      7       1      0.6  0.1549        0.253        0.827
    5      6       1      0.5  0.1581        0.184        0.753
```

Esta banda de confianza se puede observar gráficamente, conjuntamente con la función de supervivencia estimada mediante la siguiente instrucción

```
plot(K_M.loglog,xlab= "Tiempo de falla",
     ylab="Probabilidad de supervivencia")
```

Obteniendo la gráfica de la figura 3.9.

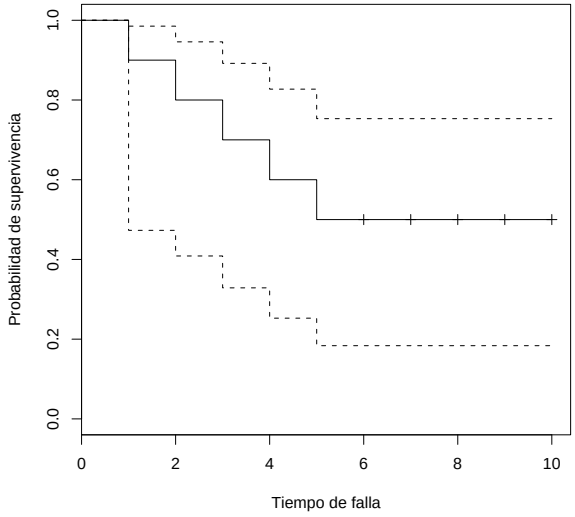


Figura 3.9: Función de supervivencia estimada con su banda de confianza, usando la transformación log-log.

Las bandas de confianza son calculadas de forma predeterminada con un 95 % de confianza, pero esto puede ser modificado mediante el argumento $conf.int$ en la función $survfit$, obteniendo la estructura siguiente:

```
survfit(formula, conf.type = "plain",conf.int = .95 )
```

Donde el .95 puede ser cambiado por cualquier número entre cero y uno que corresponda a la confianza deseada por el investigador.

Capítulo 4

Modelo de Riesgos proporcionales

Existen diversos modelos de supervivencia que involucran covariables al incorporar la manera en que éstas afectan el tiempo de falla del sujeto en estudio y uno de estos es el modelo de riesgos proporcionales. Si bien es cierto que el modelo de riesgos proporcionales es el más usado en bioestadística y en muchas disciplinas más, por ser el más entendido e implementado y los resultados que éste proporciona al ser utilizado adecuadamente, en varios casos no es un modelo apropiado por los supuestos que se tienen que cumplir (Como el supuesto de que la proporción de las funciones de riesgo es invariante al tiempo y esto no siempre sucede) y entonces es necesario estudiar modelos alternativos [1].

En la comparación del tiempo de falla de dos grupos, es de interés el caso en el que el riesgo de falla, en cualquier tiempo dado para un individuo en un grupo, es proporcional al riesgo en ese tiempo para un individuo en el otro grupo. Esta es la hipótesis de riesgos proporcionales, la cual es el fundamento de un gran número de métodos de análisis en el área de supervivencia.

Sea $h_1(t)$ el riesgo de falla al tiempo t para un individuo en el grupo I y $h_2(t)$ el riesgo en el mismo tiempo en el grupo II. Si estos dos riesgos son proporcionales, entonces se puede escribir:

$$h_1(t) = \psi h_2(t)$$

donde ψ es una constante que no depende del tiempo t. Integrando ambos lados de la igualdad, multiplicando por -1 y exponenciando se tiene

[1] Cox and Oakes [4].

$$exp\left\{-\int_0^t h_1(u)du\right\} = exp\left\{-\int_0^t \psi h_2(u)du\right\}.$$

Dado que la relación entre $S(t)$ y $h(t)$ está dada por

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

y si $S_1(t)$ y $S_2(t)$ son las funciones de supervivencia para los grupos I y II respectivamente, entonces, al suponer la hipótesis de riesgos proporcionales se tiene que

$$S_1(t) = [S_2(t)]^{\psi}.$$

Dado que la función de supervivencia toma valores entre cero y uno, este resultado permite ver que $S_2(t)$ es mayor o igual que $S_1(t)$ si ψ es menor o igual a uno, en el tiempo t. Esto significa que si dos funciones de riesgo son proporcionales, sus respectivas funciones de supervivencia no se cruzan. Esta es una condición necesaria pero no suficiente en la hipótesis de riesgos proporcionales.

Una verificación informal de la probable validez de la hipótesis de riesgos proporcionales puede realizarse al dibujar en una misma gráfica las dos funciones de supervivencia estimadas para los dos grupos de datos de supervivencia. De tal manera que si las funciones de supervivencia estimadas no se cruzan, la hipótesis de riesgos proporcionales puede estar justificada. Por supuesto que los estimadores de las funciones de supervivencia basados en la muestra pueden cruzarse a pesar de que las correspondientes funciones de riesgo reales sean proporcionales, pues habría que considerar las bandas de confianza de las funciones en la gráfica y no olvidar que estas corresponden a una estimación de las funciones reales, por lo cual se debe tener cuidado con la interpretación de tales gráficas.

Usualmente se considera que el grupo I en estudio está constituido por individuos en una situación estándar de la cual se tiene ya alguna información y al grupo II como una variación de éste, y se pretende inferir acerca de este nuevo grupo en comparación con el estándar. En medicina se puede considerar un conjunto de individuos que presentan alguna enfermedad en particular que es dividido en dos grupos, formando parte del grupo I aquellos que reciban un tratamiento usual o estándar y los restantes formando el grupo II, recibiendo un nuevo tratamiento.

El valor de ψ es el cociente de los riesgos de falla en cualquier tiempo t para un individuo en el grupo II relativo a un individuo en el grupo I, por esto ψ es conocida como el *riesgo relativo o razón de riesgo*. Si $\psi < 1$, el riesgo de falla en t es menor para un individuo en el nuevo grupo, relativo a un individuo en el grupo estándar. Por otro lado si $\psi > 1$ el riesgo de falla en t es mayor para un individuo en el nuevo grupo, relativo a un individuo en el grupo estándar.

Una forma alternativa de expresar el modelo $h_1(t) = \psi h_2(t)$ lleva a un modelo que puede ser más fácilmente generalizado. Con el supuesto de que se tienen disponibles los datos de supervivencia de n individuos y denotamos a la *i-ésima* función de riesgo por $h_i(t)$ con $i = 1,2,...,n$ donde $h_0(t)$ corresponderá a la función de riesgo de un individuo en el caso estándar, de modo que la función de riesgo para un individuo con riesgo proporcional al estándar está dada por $\psi h_0(t)$.

El riesgo relativo ψ no puede ser negativo, dado que las funciones de riesgo son mayores o iguales a cero, entonces resulta conveniente expresarlo como sigue

$$\psi = exp(\beta)$$

y por tanto, se tiene que

$$\beta = ln\psi$$

donde cualquier valor de β en el rango $(-\infty, \infty)$ llevará a un valor positivo de ψ. Es conveniente notar que se obtienen valores positivos de β cuando la razón de riesgos, ψ, es mayor que uno, esto es, cuando el riesgo de falla en un nuevo grupo es inferior al riesgo en el grupo estándar.

Sea X una variable indicadora que toma valores de cero si un individuo está en el grupo estándar y uno si el individuo está en el grupo alternativo. Si x_i es el valor de X para el *i-ésimo* individuo en el estudio con $i = 1,2,...,n$, la función de riesgos para este individuo puede ser escrita como

$$h_i(t) = e^{\beta x_i} h_0(t)$$

donde $x_i = 1$ si el *i-ésimo* individuo está en el grupo alternativo y $x_i = 0$ de cualquier otra forma. Este es el modelo de riesgos proporcionales.

4.1. Modelo general de riesgos proporcionales

El modelo $h_i(t) = e^{\beta x_i} h_0(t)$ establecido anteriormente se generaliza a la situación donde el riesgo de falla en un tiempo particular depende de los valores $x_1, x_2, ..., x_p$ de p variables explicativas $X_1, X_2, ..., X_p$. Con el supuesto de que los valores de estas variables han sido registradas en el tiempo de origen del estudio.

El conjunto de variables explicativas en el modelo de riesgos proporcionales será representado por el vector $\mathbf{x}$, de esta forma, $\mathbf{x} = (x_1, x_2, ..., x_p)'$. Sea $h_0(t)$ la función de riesgo para un individuo para el cual los valores de todas las variables explicativas que forman el vector x, son cero. La función $h_0(t)$ es llamada la *función de riesgo inicial*. La función de riesgo para el *i-ésimo* individuo puede ser escrita como

$$h_i(t) = \psi(\mathbf{x}_i) h_0(t)$$

donde $\psi(\mathbf{x}_i)$ es una función de los valores del vector de variables explicativas para el *i-ésimo* individuo. La función $\psi(\cdot)$ puede ser interpretada como el riesgo en el tiempo t para un individuo cuyo vector de variables explicativas es $\mathbf{x}_i$, relativo al riesgo para un individuo para quien $\mathbf{x} = \mathbf{0}$.

Debido a que el riesgo relativo $\psi(\mathbf{x}_i)$ no puede ser negativo, es conveniente escribirlo como $exp(\eta_i)$, donde η_i es una combinación lineal de las p variables explicativas x_i, de modo que

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots, + \beta_p x_{pi} = \sum_{j=1}^{p} \beta_j x_{ji}.$$

En notación matricial, $\eta_i = \boldsymbol{\beta}' \mathbf{x}_i$, donde $\boldsymbol{\beta}$ es el vector de los coeficientes de las variables explicativas $X_1, X_2, ..., X_p$ en el modelo. La cantidad η_i es llamada el *componente lineal* del modelo, o también es conocido como *puntaje de riesgo* para el *i-ésimo* individuo. Por tanto, el modelo general de riesgos proporcionales se puede escribir como:

$$h_i(t) = exp\left(\beta_1 x_{1i} + \beta_2 x_{1i} + \ldots, + \beta_p x_{1i}\right) h_0(t)$$

y puede ser re-expresado en la forma:

$$ln\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{1i} + \ldots + \beta_p x_{1i}$$

Donde se puede apreciar que el modelo de riesgos proporcionales puede también ser considerado como un modelo lineal para el logaritmo natural

de la razón de riesgos. Es importante notar que en el modelo de riesgos proporcionales,η_i contiene terminos lineales en cada β_j, pero cada x_{ji} puede provenir de transformaciones de las covariables originales como puede ser: $x_{ki} \times x_{ji}$, x_{ji}^2,...,etc.. Existen otras formas posibles para $\psi(\mathbf{x}_i)$ como $1 + \boldsymbol{\beta}'\mathbf{x}_i$ o el llamado logistic $log(1 + e^{\boldsymbol{\beta}'\mathbf{x}_i})$, pero la expresión $\psi(\mathbf{x}_i) = exp(\boldsymbol{\beta}'\mathbf{x}_i)$ lleva al modelo más comúnmente utilizado para los datos de supervivencia. Es importante notar que no hay término constante en la componente lineal del modelo de riesgos proporcionales. Si fuera incluído un término constante β_0, la función de riesgo inicial podría simplemente ser cambiada de escala dividiendo $h_0(t)$ entre $exp(\beta_0)$, y el término constante sería cancelado.

4.1.1. Inclusión de variables y factores al modelo

Hay dos tipos de variables de las que una función de riesgo puede depender, *variables* y *factores*. Una variable es tal que toma valores numéricos que frecuentemente están en una escala de medida continua, tales como la edad, temperatura o estatura. Un factor es una variable que toma un conjunto limitado de valores, que son conocidos como los *niveles* de los factores. Por ejemplo, el sexo es un factor con dos niveles.

Consideremos ahora cómo variables y factores pueden ser incorporados en la componente lineal de un modelo de riesgos proporcionales.

Inclusión de una variable

Las variables solas son fácilmente incorporadas en un modelo de riesgos proporcionales. Cada variable aparece en el modelo con un coeficiente β correspondiente. Por ejemplo, en una situación en la cual la función de riesgo depende de dos variables X_1 y X_2. Los valores de estas variables para el $i-$ésimo de n individuos es :

$$h_i(t) = exp(\beta_1 x_{1i} + \beta_2 x_{2i}) h_0(t)$$

En modelos como este, la función de riesgo inicial, $h_0(t)$, es la función de riesgo para un individuo para el cual todas las variables incluidas en el modelo toman el valor cero.

Inclusión de un factor

Sea el caso en el que se modela la dependencia de la función de riesgo de un solo factor A, donde A tiene a niveles. El modelo para un individuo para quien el nivel A es j, necesitará incorporar el término α_j que representa

el efecto debido al $j-$ésimo nivel del factor. Los términos $\alpha_1, \alpha_2, ..., \alpha_a$ son conocidos como los *efectos principales* del factor A. De acuerdo al modelo de riesgos proporcionales, la función de riesgo para un individuo con el factor A en el nivel j es $exp(\alpha_j)h_0(t)$. Ahora, la función de riesgo inicial $h_0(t)$ ha sido definida como el riesgo para un individuo con valores de todas las variables explicativas iguales a cero. Para ser consistentes con esta definición, una de las α_j debe ser tomada como cero. Una posibilidad es adoptar la restricción $\alpha_1 = 0$, que corresponde a tomar la función de riesgo inicial como el riesgo para un individuo para quien A está en el primer nivel. Esta es la restricción más usada en la práctica.

Los modelos que contienen términos que corresponden a factores, pueden ser expresados como combinaciones lineales de variables explicativas definiendo *variables indicadoras o mudas* para cada factor.

Si se adopta la restricción $\alpha_1 = 0$, el término α_j puede ser incluido definiendo $a-1$ variables indicadoras $X_2, X_3, ..., X_a$ que toman los valores mostrados en la tabla siguiente.

Nivel de A	X_2	X_3	$\ldots$	X_a
1	0	0	$\ldots$	0
2	1	0	$\ldots$	0
3	0	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	0	0	$\ldots$	1

El término α_j puede entonces ser incorporado en la parte lineal del modelo de riesgos proporcionales incluyendo en el término lineal, las $a-1$ variables explicativas $X_2, X_3, ..., X_a$ con coeficientes $\alpha_2 x_2 + \alpha_3 x_3 + ... + \alpha_a x_a$, donde x_j es el valor de X_j para un individuo para el cual A está en el nivel j, $j = 2, 3, ..., a$. Contando entonces con $a-1$ parámetros asociados con el efecto principal del factor A.

4.2. Estimación del modelo de riesgos proporcionales

La estimación o ajuste del modelo de riesgos proporcionales dada la expresión

$$h_i(t) = exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}\right) h_0(t)$$

a un conjunto observado de datos de supervivencia, involucra la estimación de coeficientes desconocidos de las variables explicativas $X_1, X_2, ..., X_p$ en el componente lineal del modelo, $\beta_1, \beta_2, ..., \beta_p$. Puede necesitarse también la estimación de la función de riesgo inicial, $h_0(t)$. Pero estas dos componentes del modelo pueden ser estimadas por separado. Se estiman primero las β´s y estos estimadores son utilizados para construir un estimador de la función de riesgo inicial. Este es un hecho importante, ya que significa que para hacer inferencias acerca de los efectos de p variables explicativas $X_1, X_2, ..., X_p$ sobre el riesgo relativo, $h_i(t)/h_0(t)$, no necesitamos un estimador de $h_0(t)$.

4.2.1. Estimación del modelo por máxima verosimilitud

Los coeficientes β´s en el modelo de riesgos proporcionales, que son parámetros desconocidos en el modelo, pueden ser estimados usando el método de maxima verosimilitud como sigue.

4.2.2. Función de verosimilitud (sin censura)

Suponga que se tienen disponibles los datos de n individuos donde sus tiempos de falla están denotados por $t_1, t_2, ..., t_n$ respectivamente, donde t_i es el *i-ésimo* tiempo de falla sin ordenar. Sean $\tau_1 < \tau_2 < ... < \tau_n$ los tiempos de falla ordenados de los n individuos, así que τ_j es el *j-ésimo* tiempo de falla ordenado y J_j denota al sujeto que falla en τ_j. Asi $J_j = i$ si y sólo si $t_i = \tau_j$ con i = 1,2,3,...,n.

Sea $R(\tau_j) = \{i : t_i \geq \tau_j\}$ el conjunto de individuos que están en riesgo al tiempo τ_j, así que $R(\tau_j)$ es el conjunto de individuos que no han presentado la falla al tiempo exactamente anterior a τ_j. Este conjunto es llamado *conjunto de riesgo* y su tamaño es denotado por r_j.

El principio básico de la derivación de la verosimilitud es como sigue:

Sean $\{\tau_j\}$ y $\{J_j\}$ conjuntamente equivalentes a los datos originales, es decir, los tiempos t_i de falla sin ordenar.

Con $h_0(t)$ desconocida, las τ_j sólo pueden dar poca o ninguna información acerca de $\boldsymbol{\beta}$, pues su distribución depende fuertemente de $h_o(t)$. Como ejemplo de esto, $h_0(t)$ puede ser idénticamente cero excepto en vecindades pequeñas de las τ_j. Esto es porque la función de riesgo tiene una forma arbitraria, y entonces es posible que $h_0(t)$ sea cero en aquellos intervalos de

tiempo en los cuales no hay fallas. Esto significa que estos intervalos no dan información acerca de los valores de los parámetros $\boldsymbol{\beta}$. Por tanto, la atención debe estar enfocada sobre los J_j.

En el presente caso, la distribución conjunta de $P(i_1, i_2, ..., i_n)$ sobre el conjunto de todas las posibles permutaciones de $(1, 2, ..., n)$ puede ser derivado explícitamente. Donde $P(i_1, i_2, ..., i_n)$ es la verosimilitud del orden de falla de los individuos, de modo que

i_1 denota al individuo que falla primero
i_2 denota al individuo que falla en segundo lugar
$\vdots$
i_n denota al individuo que falla en n-ésimo lugar

donde el orden de muerte de los n individuos se puede dar en todas las permutaciones de $(1, 2, ...n)$.

La derivación de $P(i_1, i_2, ..., i_n)$ se da como sigue:

Sea $H_j = \{\tau_1, \tau_2, ..., \tau_j, i_1, i_2, ..., i_{j-1}\}$ la variable que denota la historia de los tiempos de falla, tal que la información que contiene es:

τ_1 es el tiempo de falla del primer individuo (denotado por i_1)
τ_2 es el tiempo de falla del segundo individuo (denotado por i_2)
$\vdots$
τ_{j-1} es el tiempo de falla del individuo anterior al *j-ésimo* (denotado por i_{j-1})

y τ_j es el tiempo de falla del siguiente individuo en fallar que aún es desconocido en la historia H_j.

La probabilidad condicional de que $J_j = i$ dada la historia entera $H_j = \{\tau_1, \tau_2, ..., \tau_j, i_1, i_2, ..., i_{j-1}\}$ hasta el *j-ésimo* tiempo de falla τ_j , es la probabilidad condicional de que el individuo i falle en τ_j dado que un individuo del conjunto de riesgo $R(\tau_j)$ falle en τ_j. Esto se puede escribir, por definición de probabilidad condicional, como

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

Donde

A: El individuo i falle en τ_j.

B: Un individuo del conjunto de riesgo $R(\tau_j)$ falle en τ_j.

para denotar que

$$P\left[i \text{ falle en } \tau_j | \text{ Un individuo de } R(\tau_j) \text{ falle en } \tau_j\right]$$

$$= \frac{P\left[i \text{ falle en } \tau_j \text{ cuando } i \in R(\tau_j)\right]}{P\left[\text{ Un individuo de } R(\tau_j) \text{ falle en } \tau_j\right]}.$$

Donde se desarrolla el lado derecho de la ecuación como sigue.

$$\frac{P\left[i \text{ falle en } \tau_j \text{ cuando } i \in R(\tau_j)\right]}{P\left[\text{ Un individuo de } R(\tau_j) \text{ falle en } \tau_j\right]} \tag{4.1}$$

$$= \frac{\lim_{\Delta\tau_j \to 0} P\left[i \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\lim_{\Delta\tau_j \to 0} P\left[k \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j) \forall k \in R(\tau_j)\right]}. \tag{4.2}$$

Pues en el numerador de la ecuación 4.1, dado que τ_j es un tiempo de falla e i es un individuo en el conjunto de riesgo de τ_j, se tiene que el límite del numerador de la ecuación 4.2 existe y es distinto de cero. Dado que $(A \cap B) \subset B$, se tiene el resultado en probabilidad de que $P\left[A \cap B\right] \leq P\left[B\right]$ y por tanto el límite en el denominador de la ecuación 4.2 es distinto de cero, y se exhibe su existencia al generalizar a k individuos el caso del numerador:

$$= \frac{\lim_{\Delta\tau_j \to 0} P\left[i \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\lim_{\Delta\tau_j \to 0} \sum_{k \in R(\tau_j)} P\left[k \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}. \tag{4.3}$$

Puesto que el conjunto B, como quedó definido anteriormente, se puede particionar en $B_1, B_2, ..., B_{r_j}$ eventos ajenos e independientes, donde B_k es que el individuo $k \in R(\tau_j)$ falle en $(\tau_j, \tau_j + \Delta\tau_j)$, y dado que r_j es la cardinalidad de $R(\tau_j)$, se tiene el resultado de probabilidad de que $P\left[B\right] = \sum_{k \in R(\tau_j)} P\left[B_k\right]$ teniendo así que el denominador de la ecuación 4.3 se deduce del denominador de la ecuación 4.2:

$$= \lim_{\Delta\tau_j \to 0} \frac{\frac{P\left[i \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}{\frac{\sum_{k \in R(\tau_j)} P\left[k \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}} \tag{4.4}$$

Dado que los límites en 4.3 existen y el límite del denominador es distinto de cero, utilizando el resultado de cálculo, donde con estas hipótesis se tiene que el cociente de los límites es el límite de los cocientes, se tiene la ecuación 4.4, además, ésta es multiplicada por un 1 conveniente de la forma $\left(\frac{\Delta\tau_j}{\Delta\tau_j}\right)$.

$$= \lim_{\Delta\tau_j \to 0} \frac{\frac{P\left[i \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}{\sum_{k \in R(\tau_j)} \frac{P\left[k \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}. \tag{4.5}$$

Dado que $\Delta\tau_j$ es un incremento constante para la suma $\sum_{k \in R(\tau_j)}$, puede cambiarse la suma de los términos, entre el incremento, por la suma de cada término entre el incremento, y así obtenerse el cociente en 4.5 a partir de 4.4.

$$= \frac{\lim_{\Delta\tau_j \to 0} \frac{P\left[i \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}{\lim_{\Delta\tau_j \to 0} \sum_{k \in R(\tau_j)} \frac{P\left[k \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}. \tag{4.6}$$

Por el mismo argumento que se obtiene 4.4 de 4.3, se tiene 4.6 de 4.5, pero se exhibe la existencia de los límites y el hecho de que el denominador es distinto de cero en el desarrollo de los límites siguientes.

$$= \frac{lim_{\Delta\tau_j \to 0} \frac{P\left[i \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}{\sum_{k \in R(\tau_j)} lim_{\Delta\tau_j \to 0} \frac{P\left[k \in R(\tau_j) \text{ falle en } (\tau_j, \tau_j + \Delta\tau_j)\right]}{\Delta\tau_j}}. \tag{4.7}$$

Se obtiene el denominador en 4.7 del denominador en 4.6. Dado que la suma $\sum_{k \in R(\tau_j)}$ es finita, de hecho, de cardinalidad r_j como se menciona anteriormente, se puede intercambiar el límite por la suma en 4.6, obteniendo así la ecuación 4.7.

$$= \frac{h_i(\tau_j)}{\sum_{k \in R(\tau_j)} h_k(\tau_j)}. \tag{4.8}$$

La ecuación 4.8 se obtiene de 4.7 utilizando la definición de la función de riesgo. Así, aquí se puede observar que los límites de la ecuación 4.6 existen y que su denominador es distinto de cero, hecho que se utilizó anteriormente.

Usando que $h_i(t) = \psi(\mathbf{x}_i) h_0(t)$ se tiene que

$$\frac{h_i(\tau_j)}{\sum_{k \in R(\tau_j)} h_k(\tau_j)} = \frac{\psi(\mathbf{x}_i) h_0(\tau_j)}{\sum_{k \in R(\tau_j)} \psi(\mathbf{x}_k) h_0(\tau_j)}$$

y dado que $h_0(\tau_j)$ es constante para la suma sobre $k \in R(\tau_j)$ se obtiene

$$\frac{\psi(\mathbf{x}_i) h_0(\tau_j)}{h_0(\tau_j) \sum_{k \in R(\tau_j)} \psi(\mathbf{x}_k)},$$

donde al cancelar las funciones de riesgo inicial se obtiene

$$\frac{\psi(\mathbf{x}_i)}{\sum_{k\in R(\tau_j)}\psi(\mathbf{x}_k)}$$

Esta última expresión fue derivada de la probabilidad condicional de que $J_j = i$ dada la historia H_j, de hecho, es funcionalmente independiente de $\tau_1, \tau_2, ..., \tau_j$. Por lo tanto es igual a $P_j(i|i_1, i_2, ..., i_{j-1})$, la distribución condicional de J_j dado solamente $J_1 = i_1$, $J_2 = i_2$,...,$J_{j-1} = i_{j-1}$.

La distribución conjunta $P(i_1, i_2, ..., i_n)$ puede por lo tanto ser obtenida por la regla de la cadena usual para probabilidad condicional como:

$$P(i_1, i_2, ..., i_n) = \prod_{j=1}^{n} P_j(i_j|i_1, i_2, ..., i_{j-1}) = \prod_{j=1}^{n} \frac{\psi(\mathbf{x}_{(j)})}{\sum_{k\in R(\tau_j)}\psi(\mathbf{x}_k)}$$

que por definición corresponde a

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{exp(\boldsymbol{\beta}'\mathbf{x}_{(j)})}{\sum_{k\in R(\tau_j)} exp(\boldsymbol{\beta}'\mathbf{x}_k)}.$$

Para aclarar la estructura de la verosimilitud, considere la figura 4.1,

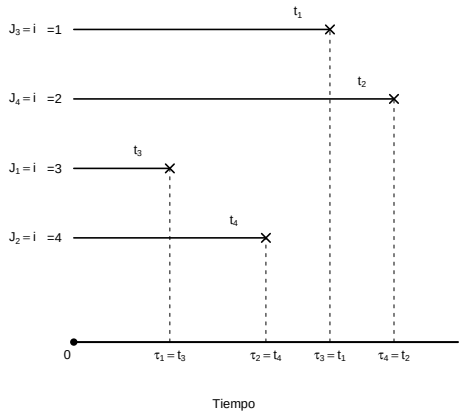


Figura 4.1: Falla de cuatro individuos sin censura. Las fallas ocurren en los instantes τ_1, τ_2, τ_3, τ_4, que corresponden con los tiempos de vida t_1, t_2, t_3, t_4. Los conjuntos de riesgo son $R(\tau_1) = \{1,2,3,4\}$, $R(\tau_2) = \{1,2,4\}$, $R(\tau_3) = \{1,2\}$ y $R(\tau_4) = \{2\}$. Se puede apreciar que $J_j = i$ *sii* $\tau_j = t_i$ para los 4 casos.

donde el escenario planteado en la figura 4.1, tiene la siguiente función de verosimilitud

$$P(3,4,1,2) = \frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)} \times \frac{\psi(4)}{\psi(1)+\psi(2)+\psi(4)} \times \frac{\psi(1)}{\psi(1)+\psi(2)} \times \frac{\psi(2)}{\psi(2)}.$$

4.2.3. Función de verosimilitud (con censura)

Suponga una muestra de tamaño n donde se observan d fallas, por lo que se tienen $n-d$ datos censurados. Sean $\tau_1 < \tau_2 < ... < \tau_d$ los tiempos ordenados de falla observados. Como antes, sea $J_j = i$ si el sujeto i falla al instante τ_j, y sea $R(\tau_j) = \{i : t_i \geq \tau_j\}$ el correspondiente conjunto de riesgo de tamaño r_j.

Con esto, se obtiene como antes que la probabilidad condicional de que $J_j = i$ dada la historia H_j es

$$\frac{\psi(\mathbf{x}_i)}{\sum_{k\in R(\tau_j)} \psi(\mathbf{x}_k)}.$$

Donde H_j incluye tanto las censuras como las fallas en el intervalo $(0, \tau_j)$ y de hecho, estas no censuras pueden ocurrir en (τ_{j-1}, τ_j) haciendo que el conjunto de riesgo, y por tanto la ecuación anterior no dependa de τ_j.

Combinaciones de estas probabilidades condicionales dan la llamada *verosimilitud parcial* (ver Apéndice A)

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{exp(\boldsymbol{\beta}'\mathbf{x}_{(j)})}{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}.$$

Donde la correspondiente función de *log verosimilitud parcial* es

$$\ln\left(L(\boldsymbol{\beta})\right) = \sum_{j=1}^{d} \left[\boldsymbol{\beta}'\mathbf{x}_{(j)} - \ln\left(\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)\right)\right].$$

De modo que los estimadores máximo verosímiles de los parámetros $\boldsymbol{\beta}$ en el modelo de riesgos proporcionales, se encuentran maximizando la función de *verosimilitud parcial* o de forma equivalente, maximizando la función de *log verosimilitud parcial* para $\beta_1, \beta_2, ..., \beta_p$.

Los métodos más eficientes de maximización se realizan tomando las derivadas parciales de primero y segundo orden que son obtenidas a continuación.

Sea $U_\xi(\boldsymbol{\beta})$ con $\xi = 1, 2, ..., p$ la derivada parcial de la función de *log verosimilitud parcial* con respecto a β_ξ, se tiene que

$$U_\xi(\boldsymbol{\beta}) = \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_\xi} = \sum_{j=1}^{d} \left[\frac{\partial \boldsymbol{\beta}'\mathbf{x}_{(j)}}{\partial \beta_\xi} - \frac{\partial \ln\left(\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)\right)}{\partial \beta_\xi}\right]$$

$$= \sum_{j=1}^{d} \left[x_{(\xi j)} - \frac{\sum_{k\in R(\tau_j)} x_{(\xi j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}\right]$$

Sea $J_{\xi\eta}(\boldsymbol{\beta})$ menos la derivada de $U_{\xi}(\boldsymbol{\beta})$ con respecto de β_{η} con $\eta = 1, 2, ..., p$, con la finalidad de obtener las derivadas parciales de segundo orden de la función de *log verosimilitud parcial*, se tiene que

$$J_{\eta\xi}(\boldsymbol{\beta}) = -\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\beta_{\xi}\partial\beta_{\eta}} = -\sum_{j=1}^{d}\left[\frac{\partial x_{(\xi j)}}{\partial\beta_{\eta}} - \frac{\partial\frac{\sum_{k\in R(\tau_j)} x_{(\xi j)}exp(\boldsymbol{\beta}'\mathbf{x}_k)}{\sum_{k\in R(\tau_j)} exp(\boldsymbol{\beta}'\mathbf{x}_k)}}{\partial\beta_{\eta}}\right]$$

$$= \sum_{j=1}^{d}\left[\frac{\sum_{k\in R(\tau_j)} x_{(\xi k)}x_{(\eta k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right) \times \sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\left(\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)\right)^2}\right.$$

$$\left. - \frac{\sum_{k\in R(\tau_j)} x_{(\xi k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right) \times \sum_{k\in R(\tau_j)} x_{(\eta k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\left(\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)\right)^2}\right]$$

$$= \sum_{j=1}^{d}\left[\frac{\sum_{k\in R(\tau_j)} x_{(\xi k)}x_{(\eta k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}\right.$$

$$\left. - \frac{\sum_{k\in R(\tau_j)} x_{(\xi k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right) \times \sum_{k\in R(\tau_j)} x_{(\eta k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\left(\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)\right)^2}\right]$$

Newton-Raphson es el método de maximización más usual, utilizando de forma iterativa las ecuaciones

$$U_{\xi}(\boldsymbol{\beta}) = \sum_{j=1}^{d}\left[x_{(\xi j)} - \frac{\sum_{k\in R(\tau_j)} x_{(\xi j)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}\right]$$

y

$$J_{\xi\eta}(\boldsymbol{\beta}) = \sum_{j=1}^{d}\left[\frac{\sum_{k\in R(\tau_j)} x_{(\xi k)}x_{(\eta k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}\right.$$

$$\left. - \frac{\sum_{k\in R(\tau_j)} x_{(\xi k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right) \times \sum_{k\in R(\tau_j)} x_{(\eta k)}exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)}{\left(\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}'\mathbf{x}_k\right)\right)^2}\right].$$

La matriz de información está dada por, menos la matriz de segundas derivadas parciales de la función de *log verosimilitud parcial* y es denotada por $I(\boldsymbol{\beta}) = \left[J_{\xi\eta}(\boldsymbol{\beta})\right]_{p\times p}$ donde el la entrada (ξ, η) está dada por $J_{\xi\eta}(\boldsymbol{\beta})$.

4.2.4. Estimación de la función de riesgo inicial

Datos provenientes de experimentos que presentan censura por la derecha, pueden ser convenientemente expresados por parejas de variables aleatorias (T, δ) donde δ indica cuando el tiempo de vida X es observado ($\delta = 1$) o no ($\delta = 0$), y T es igual a X si el tiempo de vida es observado y C_r si es censurado por la derecha, i.e. $T = min\{X, C_r\}$.

Detalles de la construcción de la función de verosimilitud para *Censura Tipo I*, son como sigue.

Para $\delta = 0$ puede verse como

$$P[T, \delta = 0] = P[T = C_r|\delta = 0]P[\delta = 0] = P[\delta = 0] = P[X > C_r] = S(C_r).$$

Y para $\delta = 1$ como

$$P[T, \delta = 1] = P[T = X|\delta = 1]P[\delta = 1] = P[X = T|X \leq C_r]P[X \leq C_r]$$

$$= \left[\frac{f(t)}{1 - S(C_r)}\right][1 - S(C_r)] = f(t).$$

Estas expresiones pueden ser combinadas como sigue

$$P[t, \delta] = [f(t)]^{\delta}[S(t)]^{1-\delta}$$

y al tener una muestra aleatoria de parejas de la forma (T_i, δ_i), $i = 1, 2, ..., n$, la función de verosimilitud está dada por

$$L = \prod_{i=1}^{n} P[t_i, \delta_i] = \prod_{i=1}^{n} [f(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i}.$$

Debido a que se tiene la relación $f(t_i) = h(t_i)S(t_i)$, se tiene que

$$\begin{aligned} L &= \prod_{i=1}^{n} [h(t_i)S(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i)]^{\delta_i}[S(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i). \end{aligned}$$

Y dado que $S(t_i) = exp\{-H(t_i)\}$

$$L = \prod_{i=1}^{n} [h(t_i)]^{\delta_i} \, exp\{-H(t_i)\}.$$

En el modelo de riesgos proporcionales, esta expresión se puede ver como

$$L[\boldsymbol{\beta}, h_0(t)] = \prod_{j=1}^{n} h(T_j|\mathbf{x}_j)^{\delta_j} S(T_j|\mathbf{x}_j)$$

$$= \prod_{j=1}^{n} h_0(T_j)^{\delta_j} \, [exp(\boldsymbol{\beta}'\mathbf{x}_j)]^{\delta_j} \, exp[-H_0(T_j)exp(\boldsymbol{\beta}'\mathbf{x}_j)].$$

Ahora, los β se fijan y se considera la maximización de esta verosimilitud como una función de $h_0(t)$ solamente. La función a maximizar es

$$L_{\boldsymbol{\beta}}(h_0(t)) = \left[\prod_{i=1}^{d} h_0(t_i)exp(\boldsymbol{\beta}'\mathbf{x}_{(i)})\right] exp\left[-\sum_{j=1}^{n} H_0(T_j)exp(\boldsymbol{\beta}'\mathbf{x}_j)\right].$$

Esta función alcanza su máximo cuando $h_0(t) = 0$ excepto para tiempos en los cuales ocurren eventos. Sea $h_{0i}(t) = h_0(t_i)$, $i = 1, 2, ..., d$ y entonces $H_0(T_j) = \sum_{t_i \leq T_j} h_{0i}$. Entonces

$$L_{\boldsymbol{\beta}}(h_{01}, ..., h_{0d}) = \prod_{i=1}^{d} h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_{(i)}) \cdot exp\left[-\sum_{j=1}^{n}\sum_{t_i \leq T_j} h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_j)\right]$$

$$= \prod_{i=1}^{d} h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_{(i)}) \cdot \prod_{i=1}^{d} exp\left[-\sum_{t_i \leq T_j} h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_j)\right]$$

$$= \prod_{i=1}^{d} h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_{(i)})exp\left[-\sum_{t_i \leq T_j} h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_j)\right].$$

Dado que $exp(\boldsymbol{\beta}'\mathbf{x}_{(i)})$ es constante y el conjunto dado por $\{t_i \leq T_j\}$ es equivalente a $\{j \in R(\tau(i))\}$ sobre $h_{0i}exp(\boldsymbol{\beta}'\mathbf{x}_j)$, se tiene la proporción dada por

$$L_{\boldsymbol{\beta}}(h_{01}, ..., h_{0d}) \propto \prod_{i=1}^{d} h_{0i}exp\left[-h_{0i}\sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}'\mathbf{x}_j)\right].$$

Tomando logaritmo

$$Ln(L_{\boldsymbol{\beta}}(h_{01},...,h_{0d})) \propto \sum_{i=1}^{d} Ln\left\{ h_{0i} exp\left[-h_{0i} \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j) \right] \right\}.$$

Derivando con respecto a h_{0i} se tiene que

$$\frac{\partial Ln(L_{\boldsymbol{\beta}}(h_{01},...,h_{0d}))}{\partial h_{0i}} \propto$$

$$\frac{exp\left[-h_{0i} \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j)\right] - h_{0i} exp\left[-h_{0i} \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j)\right] \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j)}{h_{0i} exp\left[-h_{0i} \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j)\right]}$$

$$= \frac{1 - h_{0i} \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j)}{h_{0i}} = \frac{1}{h_{0i}} - \sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j).$$

De modo que igualando a cero la derivada del logaritmo de la función de verosimilitud, se tiene que el estimador de h_{0i} está dado por

$$\hat{h}_{0i} = \frac{1}{\sum_{j \in R(\tau(i))} exp(\boldsymbol{\beta}' \mathbf{x}_j)}.$$

4.3. Modelo de riesgos proporcionales con R

Los estimadores de los coeficientes de regresión del modelo de riesgos proporcionales de Cox, son obtenidos mediante la función *coxph*. Esta función en su forma más sencilla, requiere un objeto de supervivencia creado por la función *Surv* (la cual contiene la información de tiempo de falla y censura de los individuos en estudio) y la información de las covariables de cada individuo. Esta información es ordenada de forma específica y es denotada por *fórmula* en el argumento de la función:

```
coxph(formula)
```

Suponga que se tiene la información de los tiempos de supervivencia de un grupo de individuos en las variables: tiempo, estatus, cov1, cov2,..., covn. Donde cov1,cov2,..., covn corresponden a n variables explicativas en el modelo. La estructura de la formula queda como sigue:

```
coxph(Surv(tiempo,estatus) ~ cov1 + cov2 + ... + covn )
```

Suponga que se tiene la estructura de datos siguiente:

```
> tiempo
 [1]  1  2  3  4  5  6  7  8  9 10
> estatus
 [1] 1 1 1 1 1 0 0 0 0 0
> cov1
 [1] 0 1 1 1 0 0 1 0 1 0
> cov2
 [1] 23 54 76 57 97 34 65 23 45 76
```

Los estimadores de los 2 coeficientes de regresión del modelo Cox para cov1 y cov2 se obtienen como sigue:

```
> coxph(Surv(tiempo,estatus) ~ cov1 + cov2)
Call:
coxph(formula = Surv(tiempo, estatus) ~ cov1 + cov2)


        coef exp(coef) se(coef)     z    p
cov1 0.58261      1.79   0.9339 0.624 0.53
cov2 0.00905      1.01   0.0216 0.418 0.68

Likelihood ratio test=0.56  on 2 df, p=0.755  n= 10
```

Los coeficientes están dados en la columna *coef*, en el renglón correspondiente a cada covariable, en este caso, el coeficiente para las covariables cov1 y cov2 son 0.58261 y 0.00905 respectivamente. En la columna denotada por *exp(coef)* se muestra el valor de cada coeficiente bajo la función exponencial, dado que este valor, es el que mide el impacto de cada variable explicativa en la curva de supervivencia en la interpretación del modelo.

La información adicional que se despliega por default en la función *coxph* tiene que ver con la significancia de las covariables en el modelo.

4.4. Contraste de hipótesis del modelo de riesgos proporcionales

Una vez que se ha ajustado un modelo de Cox, se verifica que sean significativas las variables del modelo.

4.4.1. Significancia de las variables del modelo

Sea $\mathbf{b} = (b_1, b_2, ..., b_p)^t$ el vector que tiene por componentes los estimadores máximo verosímiles de $\boldsymbol{\beta}$ y sea $\mathbf{I}(\boldsymbol{\beta})$ la matriz de información evaluada en $\boldsymbol{\beta}$. Existen tres pruebas principales para probar la hipótesis de que son significativos los parámetros $\boldsymbol{\beta}$ para el modelo de riesgos proporcionales[2]:

1. Prueba de Wald

[2]klein [6]. pag. 253

2. Prueba de razón de verosimilitudes

3. Prueba de puntajes

Estas tres pruebas son asintóticamente equivalentes.

4.4.2. Prueba de Wald

Esta prueba está basada en la distribución asintóticamente normal de los estimadores de máxima verosimilitud. Esto es que para muestras grandes, $\mathbf{b}$ es un vector aleatorio de distribución normal con media $\boldsymbol{\beta}$ y matriz de varianzas y covarianzas estimado por $\mathbf{I}^{-1}(\mathbf{b})$.

Esta prueba tiene la hipótesis nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ para la cual la estadística de prueba es:

$$X_W^2 = (\mathbf{b} - \boldsymbol{\beta}_0)^t \, \mathbf{I}(\mathbf{b}) \, (\mathbf{b} - \boldsymbol{\beta}_0) \, . \tag{4.9}$$

La cual tiene distribución Ji-cuadrada con p grados de libertad si H_0 es cierta para muestras grandes.

4.4.3. Prueba de razón de verosimilitudes

Esta prueba tiene la hipótesis nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ para la cual la estadística de prueba es:

$$X_{LR}^2 = 2 \left[\ln \left(L(\mathbf{b}) \right) - \ln \left(L(\boldsymbol{\beta}_0) \right) \right] . \tag{4.10}$$

La cual tiene distribución Ji-cuadrada con p grados de libertad si H_0 es cierta para muestras grandes.

4.4.4. Prueba de puntajes

Esta prueba está basada en $\mathbf{U}(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), U_2(\boldsymbol{\beta}), ..., U_p(\boldsymbol{\beta}))^t$ donde $U_\xi(\boldsymbol{\beta})$ es la derivada parcial de la función de *log verosimilitud parcial* con respecto a β_ξ con $\xi = 1, 2, ..., p$. Para muestras grandes, $\mathbf{U}(\boldsymbol{\beta})$ tiene distribución asintótica normal con el vector cero por media y matriz de varianzas y covarianzas dada por $\mathbf{I}(\boldsymbol{\beta})$ cuando H_0 es cierta. Donde la hipótesis nula es $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ y la estadística de prueba está dada por:

$$X_{SC}^2 = U(\boldsymbol{\beta}_0)^t \mathbf{I}^{-1}(\boldsymbol{\beta}_0) U(\boldsymbol{\beta}_0). \tag{4.11}$$

La cual, para muestras grandes, bajo H_0 tiene distribución Ji-cuadrada con p grados de libertad.

4.4.5. Pruebas locales

Usualmente es de interés hacer pruebas sobre subconjuntos de β´s. La hipótesis es entonces $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1_0}$ donde $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)^t$. Aquí, $\boldsymbol{\beta}_1$ es un vector de $q \times 1$ de las β´s de interés y $\boldsymbol{\beta}_2$ es el vector de las restantes $p - q$ β´s.

La prueba de Wald, con $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1_0}$ está basada en los estimadores de máxima verosimilitud parcial de $\boldsymbol{\beta}$. Suponga que la partición de la matriz de información $\mathbf{I}$ está dada por:

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix}.$$

Donde $\mathbf{I}_{11}$ y $\mathbf{I}_{22}$ son las $q \times q$ y $(p-q) \times (p-q)$ sub-matrices de segundas derivadas parciales de menos la *log verosimilitud parcial* con respecto a $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ respectivamente, y $\mathbf{I}_{12}$ como $\mathbf{I}_{21}$ las sub-matrices de las segundas derivadas parciales mezcladas. Donde, la estadística de la prueba de Wald es:

$$X_W^2 = \left(\mathbf{b}_1 - \boldsymbol{\beta}_{1_0}\right)^t \mathbf{I}^{11}(\mathbf{b}) \left(\mathbf{b}_1 - \boldsymbol{\beta}_{1_0}\right) \tag{4.12}$$

donde $\boldsymbol{\beta}^{11}(b)$ es la sub-matriz superior de $q \times q$, de $\mathbf{I}^{-1}(\mathbf{b})$ (ver Apéndice B). Para muestras grandes, esta estadística tiene distribución Ji-cuadrada con q grados de libertad bajo H_0.

Sea $\mathbf{b}_2(\boldsymbol{\beta}_{1_0})$ el estimador de máxima verosimilitud parcial de $\boldsymbol{\beta}_2$ basado en la log verosimilitud con los primeros q β´s fijos en un valor de $\boldsymbol{\beta}_{1_0}$. La prueba de razón de verosimilitud es de $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1_0}$ tiene por estadística de prueba a:

$$X_{LR}^2 = 2\left[\ln\left(Lik(\mathbf{b})\right) - \ln\left(Lik\left[\boldsymbol{\beta}_{1_0}, \mathbf{b}_2(\boldsymbol{\beta}_{1_0})\right]\right)\right]. \tag{4.13}$$

La cual, para muestras grandes, bajo H_0 tiene distribución Ji-cuadrada con q grados de libertad.

La prueba $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1_0}$ usando la estadística de puntajes se tiene como sigue, sea $\mathbf{U}_1\left[\boldsymbol{\beta}_{1_0}, \mathbf{b}_2(\boldsymbol{\beta}_{1_0})\right]$ el vector $q \times 1$ de puntajes de $\boldsymbol{\beta}_1$, evaluado en el valor, bajo H_0, de $\boldsymbol{\beta}_{10}$ y en el estimador de máxima verosimilitud parcial restringido para $\boldsymbol{\beta}_2$. Entonces

$$X_{SC}^2 = \mathbf{U}_1\left[\boldsymbol{\beta}_{1_0}, \mathbf{b}_2(\boldsymbol{\beta}_{1_0})\right]^t \mathbf{I}^{11}(\boldsymbol{\beta}_{1_0}, \mathbf{b}_2(\boldsymbol{\beta}_{1_0}))\mathbf{U}_1\left[\boldsymbol{\beta}_{1_0}, \mathbf{b}_2(\boldsymbol{\beta}_{1_0})\right]. \tag{4.14}$$

La cual, para muestras grandes, bajo H_0 tiene distribución Ji-cuadrada con q grados de libertad.

4.5. Verificación de la significancia de las variables del modelo con R

Las pruebas con las cuales se verifica la significancia de las variables en el modelo son hechas por la función *coxph*, y para ser vista esta información, es necesario utilizar la función *summary* como se presenta enseguida

Suponga que se tiene la estructura de datos siguiente:

```
> tiempo
 [1]  1  2  3  4  5  6  7  8  9 10
> estatus
 [1] 1 1 1 1 1 0 0 0 0 0
> cov1
 [1] 0 1 1 1 0 0 1 0 1 0
> cov2
 [1] 23 54 76 57 97 34 65 23 45 76
```

El modelo de regresión estimado y las pruebas para exhibir su significancia se muestran con la siguente instrucción

```
  > summary(coxph(Surv(tiempo,estatus) ~ cov1 + cov2))
Call:
coxph(formula = Surv(tiempo, estatus) ~ cov1 + cov2)

  n= 10
       coef exp(coef) se(coef)     z    p
cov1 0.58261      1.79   0.9339 0.624 0.53
cov2 0.00905      1.01   0.0216 0.418 0.68

     exp(coef) exp(-coef) lower .95 upper .95
cov1      1.79      0.558     0.287     11.17
cov2      1.01      0.991     0.967      1.05

Rsquare= 0.055   (max possible= 0.873 )
Likelihood ratio test= 0.56  on 2 df,   p=0.755
Wald test            = 0.51  on 2 df,   p=0.777
Score (logrank) test = 0.52  on 2 df,   p=0.77
```

4.5. VERIFICACIÓN DE LA SIGNIFICANCIA DE LAS VARIABLES DEL MODELO CON R

La primera tabla presenta información acerca de las pruebas locales para verificar que cada coeficiente es significativamente distinto de cero. Las columnas proporcionan información para cada covariable como sigue:

coef: El valor del coeficiente de regresión estimado.

exp(coef): La función exponencial evaluada en el coeficiente de regresión estimado.

se(coef): el error estándar del coeficiente de regresión estimado.

z: Corresponde a la estadística de prueba, obtenida dividiendo el valor del coeficiente de regresión estimado entre el error estándar estimado.

p: El $p-value$ que corresponde a dos veces el área acumulada a la derecha del cuantil z en una distribución normal con media cero y varianza uno.

La segunda tabla presenta información acerca de los coeficientes de regresión estimados.

exp(coef): La función exponencial evaluada en el coeficiente de regresión estimado.

exp(-coef): La función exponencial evaluada en, menos el coeficiente de regresión estimado.

lower .95: El límite inferior de un intervalo del 95 % de confianza para el coeficiente de regresión estimado.

upper .95: El límite superior de un intervalo del 95 % de confianza para el coeficiente de regresión estimado.

En los últimos tres renglones se puede ver la información correspondiente a probar la hipótesis nula de que el vector de variables del modelo son cero, i.e. $H_0 : \boldsymbol{\beta} = \bar{\mathbf{0}}$.

Prueba de razón de verosimilitudes

La estadística de prueba, denotada anteriormente por X^2_{LR} es 0.56 y con una distribución χ^2 con 2 grados de libertad, tiene un $p-value$ de $p = 0.755$.

Prueba de Wald

La estadística de prueba, denotada anteriormente por X^2_W es 0.51 y con una distribución χ^2 con 2 grados de libertad, tiene un $p-value$ de $p = 0.777$.

Prueba de puntajes

La estadística de prueba, denotada anteriormente por X^2_{SC} es 0.52 y con una distribución χ^2 con 2 grados de libertad, tiene un $p-value$ de $p = 0.77$.

De modo que al observar valores "pequeños"en el p-value, se tendrá evidencia de que, bajo las pruebas realizadas, los coeficientes del modelo son significativamente distintos de cero, y por tanto, se podrá considerar que el modelo tiene sentido para las variables explicativas consideradas.

4.6. Función de supervivencia para el modelo de riesgos proporcionales

Una vez que ya se tienen los estimadores de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$ se obtiene el estimador de h_j, la función de riesgo inicial entre t_{j-1} y t_j ($t_0 := 0$) con $j = 1, 2, ..., d$ como

$$\hat{h}_j = \frac{1}{t_j - t_{j-1}} \frac{1}{\sum_{k \in R(\tau_j)} exp\left(\boldsymbol{\beta}' \mathbf{x}_k\right)}.$$

De tal modo que la función de riesgo acumulado, para la función de riesgo inicial, puede ser estimada de la siguiente manera

$$\int_0^t \hat{h}_0(u)du = \sum_{j=1}^{l} \left(t_j - t_{j-1}\right) \hat{h}_j + \left(t - t_l\right) \hat{h}_{l+1}.$$

Dado que la función de riesgo $\hat{h}_0$ se considera constante con valor $\hat{h}_j$ en el intervalo que comienza en t_{j-1} y termina en t_j. Por tanto, la integral se transforma en la suma de las áreas de rectángulos con altura $\hat{h}_j$ y base $(t_j - t_{j-1})$ y el rectángulo de altura t_{l+1} con base $(t - t_l)$, al considerar que el límite superior de la integral dado por t cae en el intervalo (t_l, t_{l+1}). Es importante notar que la integral de la función de riesgo puede ser estimada solamente para valores de t tales que $t \leq t_d$ pues la función de riesgo queda indefinida para $t > t_d$.

Utilizando la relación entre la función de supervivencia y la función de riesgo acumulado en el modelo de riesgos proporcionales se tiene que

$$\hat{S}(t, \mathbf{x}) = exp\left\{-exp\left\{\hat{\boldsymbol{\beta}}' \mathbf{x}\right\} \int_0^t \hat{h}_0(u)du\right\}.$$

Por tanto, $S(t, \mathbf{x})$ puede ser estimada por:

$$\hat{S}(t, \mathbf{x}) = exp\left[-exp(\hat{\boldsymbol{\beta}}' \mathbf{x}) \left\{\sum_{j=1}^{l} \left(t_j - t_{j-1}\right) \hat{h}_j + \left(t - t_l\right) \hat{h}_{l+1}\right\}\right]$$

$$= exp\left[-exp(\hat{\boldsymbol{\beta}}' \mathbf{x}) \left\{\sum_{j=1}^{l} \frac{1}{\sum_{k \in R(\tau_j)} exp\left(\boldsymbol{\beta}' \mathbf{x}_k\right)} + \frac{t - t_l}{t_{l+1} - t_l} \frac{1}{\sum_{k \in R(\tau_{l+1})} exp\left(\boldsymbol{\beta}' \mathbf{x}_k\right)}\right\}\right].$$

4.7. Función de supervivencia para el modelo de riesgos proporcionales obtenida con R

La función de supervivencia para los tiempos de falla de un grupo de individuos bajo el modelo de riesgos proporcionales, se obtiene aplicando la función *survfit* a un objeto creado por la función *coxph* como se muestra a continuación:

```
survfit(coxph(... )
```

Suponga que se tiene la estructura de datos siguiente:

```
> tiempo
 [1]  1  2  3  4  5  6  7  8  9 10
> estatus
 [1] 1 1 1 1 1 0 0 0 0 0
> cov1
 [1] 0 1 1 1 0 0 1 0 1 0
> cov2
 [1] 23 54 76 57 97 34 65 23 45 76
```

La función de supervivencia estimada para el modelo de riesgos proporcionales se obtiene con la siguiente instrucción:

```
> survfit(coxph(Surv(tiempo,estatus) ~ cov1 + cov2))
```

Donde por default se despliega la siguiente información:

```
Call:
survfit.coxph(object=coxph(Surv(tiempo,estatus)~cov1 + cov2))

      n  events  median 0.95LCL 0.95UCL
     10       5     Inf       4     Inf
```

Para obtener la información detallada de la función de supervivencia bajo este modelo, así como su gráfica, se puede proceder de la misma forma que para un objeto de la forma *Surv*, visto en el capítulo 4.

4.8. Bandas de confianza para la función de supervivencia en el modelo de riesgos proporcionales

Para obtener las bandas de confianza de la función de supervivencia para el modelo de riesgos proporcionales, se presenta la derivación de la varianza asintótica de $\hat{S}(t, \mathbf{x})$ como sigue:

$\hat{S}(t, \mathbf{x})$ es afectada por dos fuentes de variación, primero, la variación de la estimación de la integral de la función de riesgo, y segundo, la variación de la estimación de β por $\hat{\beta}$. Para considerar esto, $\hat{S}(t, \mathbf{x})$ se aproxima por la expansión en serie de Taylor de primer orden

$$\hat{S}(t, \mathbf{x}) \approx \hat{S}(t, \mathbf{x}) \mid_{\boldsymbol{\beta}} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \frac{\partial \hat{S}(t, \mathbf{x})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

Donde $\hat{S}(t, \mathbf{x}) \mid_{\boldsymbol{\beta}}$ es $\hat{S}(t, \mathbf{x})$ evaluado en $\boldsymbol{\beta}$ y $\frac{\partial \hat{S}(t, \mathbf{x})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ es el vector de primeras derivadas parciales de $\hat{S}(t, \mathbf{x})$ con respecto de $\boldsymbol{\beta}$. Debido a que $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$ es asintóticamente independiente de $\hat{S}(t, \mathbf{x}) \mid_{\boldsymbol{\beta}}$ y $\frac{\partial \hat{S}(t, \mathbf{x})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ es asintóticamente constante, la varianza de $\hat{S}(t, \mathbf{x})$ se aproxima por

$$Var\left\{\hat{S}(t, \mathbf{x})\right\} \approx Var\left\{\hat{S}(t, \mathbf{x}) \mid_{\boldsymbol{\beta}}\right\} + \frac{\partial \hat{S}(t, \mathbf{x})}{\partial \boldsymbol{\beta}'} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} Var\left(\boldsymbol{\beta}\right) \frac{\partial \hat{S}(t, \mathbf{x})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

El primer término, presenta la variación de la estimación de la integral de la función de riesgo dada anteriormente y el segundo término presenta la variación de la estimación de $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$.

La varianza asintótica de $\boldsymbol{\beta}$ está dada por

$$Var(\hat{\beta}) = -\left\{\frac{\partial^2 Lik(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_p}\right\}^{-1}.$$

De modo que resta calcular la primera derivada parcial de $\hat{S}(t, \mathbf{x})$ con respecto a $\boldsymbol{\beta}$, lo cual se presenta como sigue

$$\frac{\partial \hat{S}(t, \mathbf{x})}{\partial \hat{\beta}_m} =$$

$$exp\left[-exp(\hat{\boldsymbol{\beta}}'\mathbf{x})\left\{\sum_{j=1}^{l} \frac{1}{\sum_{k \in R(\tau_j)} exp\left(\boldsymbol{\beta}' x_k\right)} + \frac{t - t_l}{t_{l+1} - t_l} \frac{1}{\sum_{k \in R(\tau_{l+1})} exp\left(\boldsymbol{\beta}' x_k\right)}\right\}\right]$$

$$\left(exp(\hat{\boldsymbol{\beta}}'\mathbf{x}) \left[\sum_{j=1}^{l} \frac{\sum_{k\in R(\tau_j)} x_{km} exp\left(\boldsymbol{\beta}' x_k\right)}{\left\{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}' x_k\right)\right\}^2} + \frac{t-t_l}{t_{l+1}-t_l} \frac{\sum_{k\in R(\tau_{l+1})} x_{km} exp\left(\boldsymbol{\beta}' x_k\right)}{\left\{\sum_{k\in R(\tau_{l+1})} exp\left(\boldsymbol{\beta}' x_k\right)\right\}^2} \right] \right.$$

$$\left. - x_m exp(\hat{\boldsymbol{\beta}}'\mathbf{x}) \left\{ \sum_{j=1}^{l} \frac{1}{\sum_{k\in R(\tau_j)} exp\left(\boldsymbol{\beta}' x_k\right)} + \frac{t-t_l}{t_{l+1}-t_l} \frac{1}{\sum_{k\in R(\tau_{l+1})} exp\left(\boldsymbol{\beta}' x_k\right)} \right\} \right)$$

$$= \hat{S}(t,\mathbf{x}) \hat{D}_m(t,\mathbf{x})$$

con $m = 1, 2, ..., p$

Donde $\hat{\mathbf{D}}'(t,\mathbf{x}) = (\hat{D}_1(t,\mathbf{x}), \hat{D}_2(t,\mathbf{x}), ..., \hat{D}_p(t,\mathbf{x}))$ es el vector de primeras derivadas parciales de la integral de la función de riesgo con respecto de $\hat{\beta}$.

Por lo tanto,

$$\frac{\partial \hat{S}(t,\mathbf{x})}{\partial \boldsymbol{\beta}'} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} Var\left(\boldsymbol{\beta}\right) \frac{\partial \hat{S}(t,\mathbf{x})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

$$= \hat{S}^2(t,\mathbf{x}) \hat{\mathbf{D}}'(t,\mathbf{x}) \hat{Var}\left(\boldsymbol{\beta}\right) \hat{\mathbf{D}}(t,\mathbf{x})$$

La varianza de $\hat{S}(t,\mathbf{x})$ dado β cuando $\beta = 0$ es aproximada mediante la Formula de Greenwood

$$\hat{Var}\left(\prod \hat{p}_i\right) \approx \hat{P}^2(t) \left\{ \prod \left(1 + \frac{\hat{q}_i}{N_i \hat{p}_i}\right) - 1 \right\}$$

donde las $\hat{p}_i$ son probabilidades independientes binomiales de una muestra de tamaño N_i, $\hat{q}_i = 1 - \hat{p}_i$ y $\hat{P}(t) = \prod \hat{p}_i$. La forma usual de la fórmula de Greenwood es $\hat{Var}(\prod \hat{p}_i) \approx \hat{P}^2(t) \sum \hat{q}_i / N_i \hat{p}_i$ obtenida anteriormente usando la *aproximaciòn de series de Taylor para la varianza de una variable aleatoria*, pero esta expresión es justamente el término lineal de la expresión anterior, dado que $\hat{q}_i / N_i \hat{p}_i$ es el término lineal del desarrollo de $(1 + \hat{q}_i / N_i \hat{p}_i) - 1$.

Dado que al utilizar la fórmula de Greenwood, N_i es el número de individuos en el conjunto de riesgo, $\hat{P}(t) = \hat{S}(t,\mathbf{x})$ y dado que $\hat{P}(t) = \prod \hat{p}_i$, en la ecuación

$$\hat{S}(t,\mathbf{x}) =$$

$$exp\left[-exp(\hat{\boldsymbol{\beta}}'\mathbf{x})\left\{\sum_{j=1}^{l}\frac{1}{\sum_{k\in R(\tau_j)}exp\,(\boldsymbol{\beta}'x_k)}+\frac{t-t_l}{t_{l+1}-t_l}\frac{1}{\sum_{k\in R(\tau_{l+1})}exp\,(\boldsymbol{\beta}'x_k)}\right\}\right].$$

Se puede ver que

$$\hat{p_i}=exp\left\{\frac{-exp(\hat{\boldsymbol{\beta}}'\mathbf{x})}{\sum_{k\in R(\tau_j)}exp\,(\boldsymbol{\beta}'x_k)}\right\}\quad i=1,\ldots,l.$$

$$\hat{p_{l+1}}=exp\left\{\frac{-exp(\hat{\boldsymbol{\beta}}'\mathbf{x})\left(\frac{t-t_l}{t_{l+1}-t_l}\right)}{\sum_{k\in R(\tau_j)}exp\,(\boldsymbol{\beta}'x_k)}\right\}$$

De modo que $\hat{S}(t,\mathbf{x})=\prod_j^{l+1}\hat{p_i}$, donde $\hat{p_i}$ puede ser tratado como condicionalmente independiente, dados $\boldsymbol{\beta}$ fijos.

Utilizando lo anterior, se tiene que la estimaciòn de la varianza asintótica de $\hat{S}(t,\mathbf{x})$ está dada por[3]:

$$Var\left\{\hat{S}(t,\mathbf{x})\right\}\approx\hat{S}^2(t,\mathbf{x})\left\{\prod_{i=1}^{l+1}\left(1+\frac{\hat{q_i}}{N_i\hat{p_i}}\right)-1+\hat{\mathbf{D}}'(t,\mathbf{x})\hat{Var}\,(\boldsymbol{\beta})\hat{\mathbf{D}}(t,\mathbf{x})\right\}.$$

Definiendo el error estándar estimado de $\hat{S}(t,\mathbf{x})$ como

$$s.e.\left\{\hat{S}(t,\mathbf{x})\right\}\approx\hat{S}(t,\mathbf{x})\left\{\prod_{i=1}^{l+1}\left(1+\frac{\hat{q_i}}{N_i\hat{p_i}}\right)-1+\hat{\mathbf{D}}'(t,\mathbf{x})\hat{Var}\,(\boldsymbol{\beta})\hat{\mathbf{D}}(t,\mathbf{x})\right\}^{\frac{1}{2}}$$

se tiene que un intervalo de un $100(1-\alpha)\,\%$ de confianza para $S(t,\mathbf{x})$ está dado por:

$$\left(\hat{S}(t,\mathbf{x})-z_{1-\alpha/2}s.e.\left\{\hat{S}(t,\mathbf{x})\right\},\hat{S}(t,\mathbf{x})+z_{1-\alpha/2}s.e.\left\{\hat{S}(t,\mathbf{x})\right\}\right)$$

donde $z_{1-\alpha/2}$ es el cuantil que acumula $1-\alpha/2$ de probabilidad en una distribución normal con media cero y varianza uno.

[3]Link [8]. pag. 604

4.9. Bandas de confianza para la función de supervivencia en el modelo de riesgos proporcionales con R

El intervalo de confianza para el modelo de riesgos proporcionales se obtendrá de la misma manera que para un objeto de la forma *Surv*.

Una vez que se ha obtenido la función de supervivencia para los tiempos de falla de un grupo de individuos, bajo el modelo de riesgos proporcionales, aplicando la función *survfit* a un objeto creado por la función *coxph*, los intervalos de confianza para esta función de supervivenca son generados automáticamente.

Suponga que se tiene la estructura de datos siguiente:

```
> tiempo
 [1]  1  2  3  4  5  6  7  8  9 10
> estatus
 [1] 1 1 1 1 1 0 0 0 0 0
> cov1
 [1] 0 1 1 1 0 0 1 0 1 0
> cov2
 [1] 23 54 76 57 97 34 65 23 45 76
```

La función de supervivencia estimada para el modelo de riesgos proporcionales se obtiene con la siguiente instrucción:

```
 > survfit(coxph(Surv(tiempo,estatus) ~ cov1 + cov2))
```

Donde por default se despliega la siguiente información:

```
Call:
survfit.coxph(object=coxph(Surv(tiempo,estatus)~cov1 + cov2))

      n  events  median 0.95LCL 0.95UCL
     10       5     Inf       4     Inf
```

4.9. BANDAS DE CONFIANZA PARA LA FUNCIÓN DE SUPERVIVENCIA EN EL MODELO DE RIESGOS PROPORCIONALES CON R

La información que aparece es la siguiente: n corresponde al número de individuos en estudio, *events* corresponde al número de fallas presentadas en los n individuos (por tanto, el número de censuras está dado por $n - events$), *median* es el tiempo mediano antes de que se presente la falla con respecto a la curva de la función de supervivencia estimada (el tiempo t tal que $S(t) = .5$), $0.95LCL$ es límite inferior de una banda estimada con un 95 % de confianza para el valor de *median* y $0.95UCL$ es límite superior de una banda estimada con un 95 % de confianza para el valor de *median*.

Para obtener más información de la banda de confianza generada, se pueden seguir las instrucciónes vistas en el capítulo 4, con la finalidad de obtener los valores del intervalo de confianza en uno o varios tiempos específicos, cambiar el nivel de confianza (que por default es del 95 %), así como graficar la función de supervivencia estimada con sus bandas de confianza.

4.10. Interpretación del modelo

Cuando el modelo de riesgos proporcionales es utilizado, los coeficientes de las variables explicativas en el modelo, pueden ser interpretados como el logaritmo del cociente de la función de riesgo y la función de riesgo inicial. La interpretación de los parámetros correspondientes a los diferentes tipos de variables explicativas del modelo de riesgos proporcionales depende si éstas son variables o factores.

Si el coeficiente de regresión β corresponde a una variable, la cantidad $exp\beta$ es el cambio en la función de riesgo por cada unidad que se incremente la variable, dado que las otras covariables están fijas, esto se puede escribir como sigue:

$$\begin{aligned}
&\frac{h(t, x_1, \ldots, x_i + 1, \ldots, x_p, \beta_1, \ldots, \beta_p)}{h(t, x_1, \ldots, x_i, \ldots, x_p, \beta_1, \ldots, \beta_p)} \\
&= \frac{h_0(t) exp\left\{\beta_1 x_1 + \cdots + \beta_i (x_i + 1) + \cdots + \beta_p x_p\right\}}{h_0(t) exp\left\{\beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_p x_p\right\}} \\
&= exp\left\{\beta_i\right\}.
\end{aligned}$$

Si el coeficiente de regresión β corresponde a un factor, la interpretación es como sigue. Suponga que se tiene un factor A con a niveles, y sean $\beta_2, \ldots, \beta_a$ los coeficientes de regresión de los correspondientes niveles, considerando que el primer nivel del factor A es cero. La cantidad $exp\left\{\beta_i - \beta_j\right\}$ es el cociente de la función de riesgo para el sujeto con el nivel i y el nivel j de las variables explicativas $(i, j = 2, \ldots, a)$, dado que las otras variables explicativas toman valores iguales. La cantidad $exp\left\{\beta_i\right\}$ corresponde al cambio relativo en la función de riesgo para sujetos con variables explicativas con nivel i $(i = 2, \ldots, a)$ y nivel 1.

Ejemplo

La base de datos *"cucaracha"* contiene la información del tiempo de falla de 150 cucarachas (donde 20 son censuradas), que fueron divididas en 3 grupos (A,B,C) y sometidas a un tipo de insecticida diferente. El experimento, donde el tiempo de falla es la muerte de la cucaracha por el insecticida al que fue expuesto, consistió en aplicarle a 50 cucarachas el insecticida A, a 50 el insecticida B y a 50 el insecticida C y esperar a la muerte de la cucaracha. El tiempo de falla está medida en días y el experimento duró 50 días, por lo que las cucarachas que sobrevivieron al día 50 fueron eliminadas y y sus tiempos de falla censurados en ese momento, por lo cual los datos presentan censura por la derecha. La base de datos contiene para cada cucaracha la siguente información:

- Tiempo de falla.
- Estatus de censura (0 si es censurado y 1 si la falla es observada).
- Peso de la cucaracha en gramos.
- Tipo de insecticida que se le aplicó a la cucaracha (A, B, ó C).

Es de interés el hecho de que se tenga el peso de la cucaracha y no otro dato. Es posible que el especialista considere que el peso de la cucaracha es una variable que da información de su posible edad, longitud o resistencia y sea una variable que influye en su tiempo de falla.

Información de la población de estudio

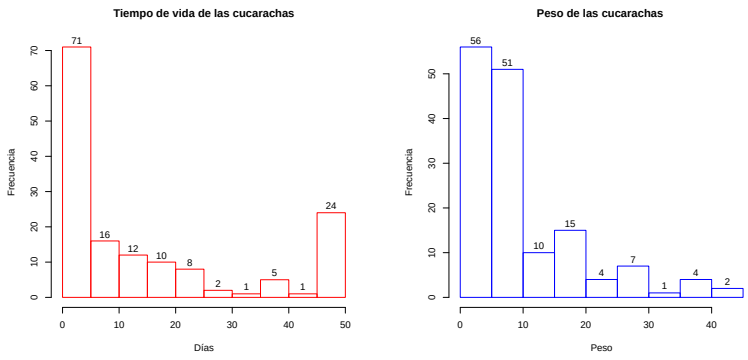
Para poder realizar un análisis de supervivencia adecuado, es importante conocer la estructura de la población con que se está trabajando. En este caso es de interés saber si los insecticidas fueron aplicados a muestras de cucarachas con las mismas características de peso para evitar sesgamiento en la efectividad del insecticida aplicado.

Con la finalidad de conocer la estructura de la población de cucarachas en estudio se presentan algunas estadísticas básicas.

Se obtiene información de la base de datos con el paquete estadístico **R**, y se obtiene siguiente información:

```
     death            status           weight       group
 Min.   : 1.00   Min.   :0.0000   Min.   : 0.055   A:50
 1st Qu.: 1.00   1st Qu.:1.0000   1st Qu.: 2.459   B:50
 Median : 7.00   Median :1.0000   Median : 6.316   C:50
 Mean   :15.17   Mean   :0.8667   Mean   : 9.390
 3rd Qu.:21.00   3rd Qu.:1.0000   3rd Qu.:11.955
 Max.   :50.00   Max.   :1.0000   Max.   :42.090
```

De la primera columna se puede observar que el tiempo máximo de vida de una cucaracha fue de 50 días y el tiempo mínimo fue de un día. De la segunda columna se puede saber que aproximadamente el 86.67 % de la población no presenta censura. De la tercera columna se observa que el peso mínimo fue de 0.055 gramos, el máximo fue de 42.090 gramos y que en promedio, la poblacion pesa 9.390 gramos. De la última columna se tiene que hay tres grupos denotados por A, B y C, los cuales constan de 50 cucarachas cada uno.

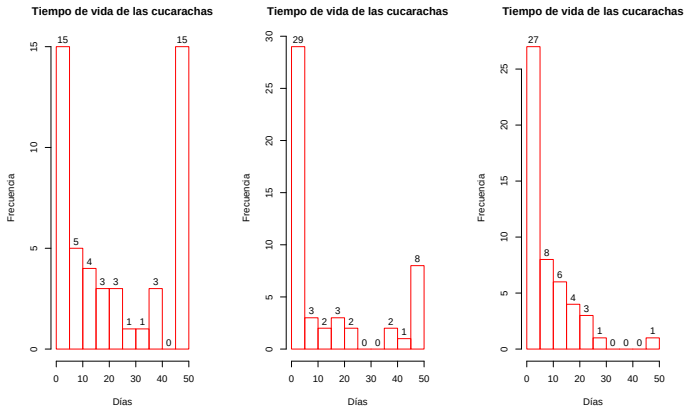


(a) (b)

Figura 4.2: Histograma del tiempo de vida y el peso de las cucarachas.

Se obtiene información más clara de las variables de interés mediante los histogramas de la figura 4.2, donde se presenta la frecuencia de los tiempos de

supervivencia y la estructura del peso de la población. Se puede observar que la mayoría de las cucarachas tienen un peso menor a 10 gramos y en general se presenta la falla los primeros 5 días aunque un número considerable sobrevive entre 45 y 50 días.

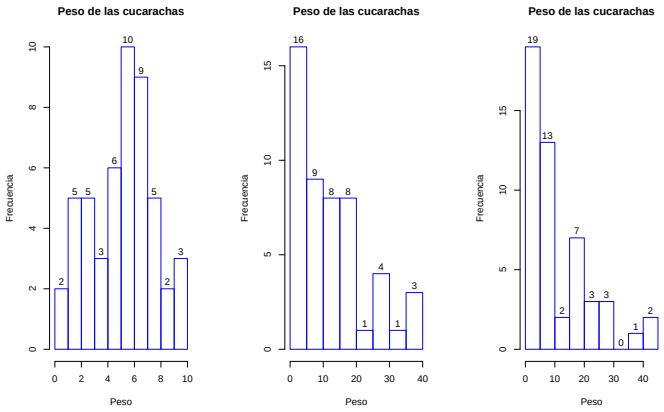


(a) Insecticida A (b) Insecticida B (c) Insecticida C

Figura 4.3: Histogramas del tiempo de vida de las cucarachas.

La comparación por grupo del tiempo de falla es ilustrada en la figura 4.3, donde se muestra el histograma de los tiempos de falla de las cucarachas separadas por grupo. Esto es un primer acercamiento a la relación de los tiempos de vida y el tipo de insecticida utilizado.

En la figura 4.4 se muestra la estructura en peso que tiene cada grupo al que se le aplicó el insecticida, el cual no parece uniforme aunque esta sería una característica deseable para evitar sesgamiento de los resultados del modelo de supervivencia.



(a) Insecticida A (b) Insecticida B (c) Insecticida C

Figura 4.4: Histogramas del peso de las cucarachas.

En el análisis de la información de la población en estudio, se puede profundizar tanto como se desee y solo se ha dado un acercamiento a la forma en que éste se puede hacer, aunque existe una gran cantidad de herramientas estadísticas y alternativas que no se mencionan por no ser el interés principal del trabajo.

Modelo de supervivencia

En la figura 4.5 se muestra el estimador de Kaplan-Meier de la función de supervivencia con la banda de confianza utilizando la transformación $log-log$ para las 50 cucarachas sin distinción de grupo, pues puede ser de interés, analizar el tiempo de falla de una cucaracha si no se tiene información de cual fue el insecticida que se le aplicó.

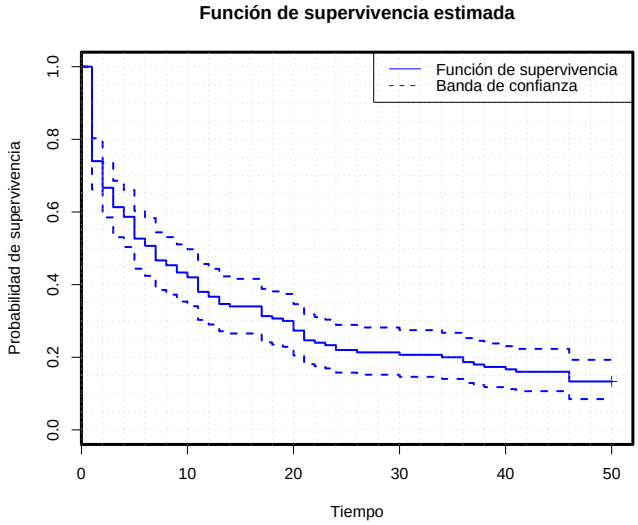


Figura 4.5: Función de supervivencia estimada sin distinción de insecticida aplicado.

A continuación se presenta la información del modelo de supervivencia solicitada de manera específica cada 5 días, con la cual se puede analizar la evolución de la probabilidad de falla en el modelo con su respectivo intervalo de confianza y considerar la proporción de individuos que presentan la falla de los que están en riesgo entre cada periodo de 5 días.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	150	0	1.000	0.0000	1.0000	1.000
5	88	71	0.527	0.0408	0.4438	0.603
10	65	16	0.420	0.0403	0.3404	0.497
15	51	12	0.340	0.0387	0.2654	0.416
20	45	10	0.273	0.0364	0.2047	0.346
25	33	8	0.220	0.0338	0.1576	0.289
30	32	2	0.207	0.0331	0.1461	0.275
35	30	1	0.200	0.0327	0.1404	0.267
40	26	5	0.167	0.0304	0.1122	0.231
45	24	1	0.160	0.0299	0.1066	0.223
50	20	4	0.133	0.0278	0.0848	0.193

En la figura 4.6 se presentan los estimadores de Kaplan-Meier de las funciones de supervivencia para los tres grupos por separado con la finalidad de hacer un análisis comparativo visual de la probabilidad de supervivencia con distintos insecticidas. De esta gráfica se puede concluir que es significativamente más efectivo el insecticida C para eliminar las cucarachas.

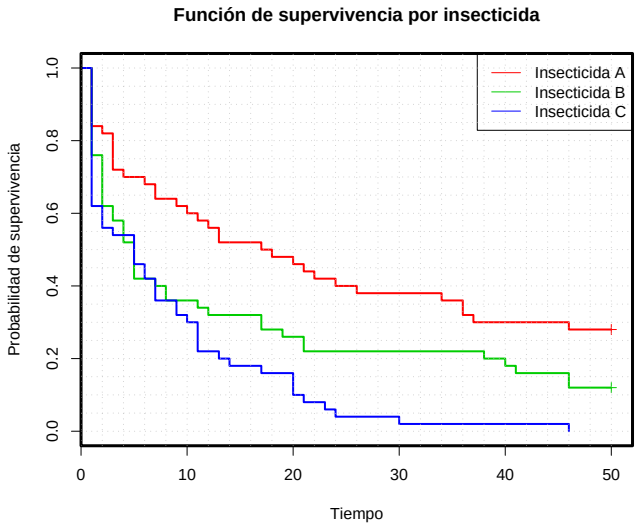


Figura 4.6: Funciones de supervivencia estimadas distinguiendo el tipo de insecticida aplicado.

En la figura 4.7 se presenta la función de riesgo acumulado estimada para los tres grupos por separado donde se puede apreciar que el riesgo es mayor para las cucarachas tratadas con el insecticida C mientras transcurre el tiempo. Cada salto de estas funciones escalonadas corresponde al riesgo de falla diario de cada cucaracha tratada con el correspondiente insecticida, de manera que se puede interpretar con detalle la evolución del riesgo para cada insecticida por separado y en comparación con los demás mediante la longitud de los saltos de la función de riesgo acumulado.

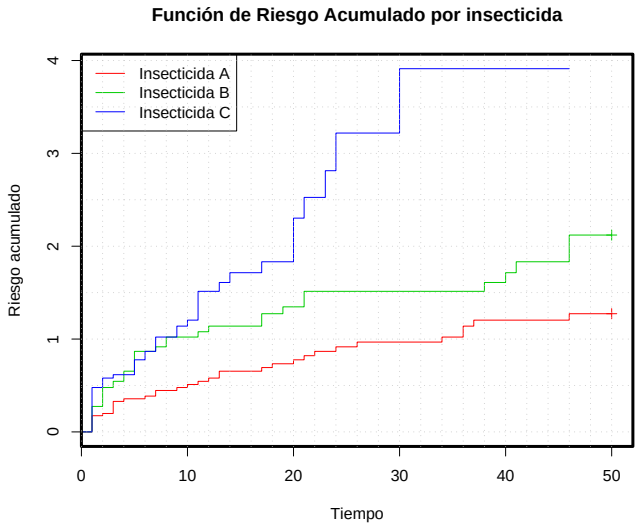


Figura 4.7: Funciones de riesgo acumulado distinguiendo el tipo de insecticida aplicado.

Modelo de riesgos proporcionales

Como se ha dejado entrever a lo largo del ejemplo, es de interés saber si está relacionado el peso de una cucaracha con su tiempo de supervivencia. Lo que se hace a continuación es implementar el modelo de riesgos proporcionales utilizando como variable regresora el peso de las cucarachas, y saber si esta variable influye en el tiempo de supervivencia de las cucarachas. El modelo se hace sin distinción del insecticida aplicado, con la idea de que el investigador puede estar interesado en este momento en la resistencia de las cucarachas de acuerdo al peso y suponiendo que ha satisfecho la inquietud de comparar los tipos de insecticida con el análisis anterior.

Enseguida se muestra el modelo de riesgos proporcionales estimado, con lo que se puede apreciar que el coeficiente del peso de las cucarachas es de 0.0188, y bajo la función exponencial, este coeficiente tiene el valor de 1.02, el cual es el que mide el impacto de la variable regresora en el tiempo de falla.

```
         coef exp(coef) se(coef)    z     p
weight 0.0188      1.02  0.00944 2.00 0.046

Likelihood ratio test=3.7  on 1 df, p=0.0545  n= 150
```

A continuación se muestran las tres pruebas de hipótesis presentadas en este trabajo, las cuales verifican que la variable explicativa sea significativa en el modelo.

```
  n= 150
         coef exp(coef) se(coef)    z     p
weight 0.0188      1.02  0.00944 2.00 0.046

       exp(coef) exp(-coef) lower .95 upper .95
weight      1.02      0.981         1      1.04

Rsquare= 0.024   (max possible= 0.999 )
Likelihood ratio test= 3.7  on 1 df,   p=0.0545
Wald test            = 3.98  on 1 df,   p=0.046
Score (logrank) test = 4.01  on 1 df,   p=0.0451
```

Para las tres pruebas se aprecia un p-value significativamente pequeño, lo cual es evidencia de que el peso de las cucarachas es una variable explicativa con un coeficiente significativamente distinto de cero en el modelo de riesgos proporcionales, no olvidando que el nivel de confianza está sujeto a la percepción del investigador.

En la figura 4.8 se muestra la gráfica de supervivencia para el modelo de riesgos proporcionales, la cual no presenta variaciones a la función de supervivencia estimada de Kaplan-Meier de la figura 4.5. Esto se debe a que el peso de las cucarachas influye de manera muy particular en su tiempo de falla. Como se puede ver en la figura 4.9, las cucarachas con un peso mayor a 10 gramos, en su mayoría tienen un tiempo de falla menor a 15 días salvo 4 sujetos de estudio. Las cucarachas cuyo peso es menor a 10 gramos, se encuentran en una de las tres situaciones siguientes:

1. Presentan la falla el primer día.

2. Se registra el tiempo de falla en el día 50, el cual es la fecha del término del experimento y el dato es censurado.

3. Presentan la falla a lo largo del estudio.

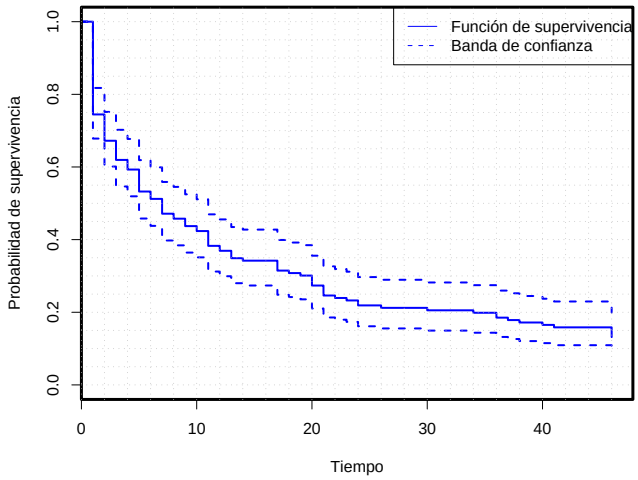
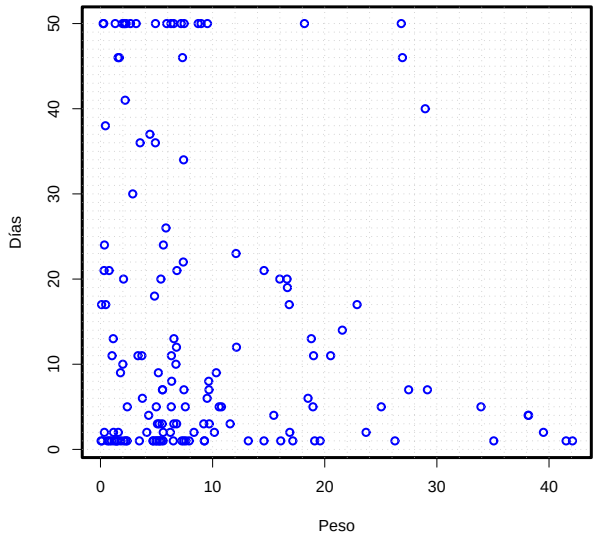


Figura 4.8: Funciones de supervivencia del modelo de riesgos proporcionales considerando el peso de las cucarachas como variable explicativa.

La razón de clasificar el tiempo de falla de esta manera es que las situaciónes 1 y 2 representan un porcentaje considerablemente alto. Esto quiere decir que las cucarachas con peso menor a 10 gramos que no presentan la falla el primer día, puede ser que no presenten la falla como consecuencia de tener contacto con un insecticida, pues en el estudio, solo se tiene registrado un tiempo de supervivencia censurado. Lo cual puede resultar de interés para el especialista pues esto puede ser consecuencia de alguna explicación biológica de las cucarachas. No obstante, el impacto del peso de las cucarachas fue exhibido en las pruebas realizadas para el modelo de riesgos proporcionales y se ha identificado que el peso es influyente en el tiempo de falla de las cucarachas.

Figura 4.9: Tiempo de falla para cada cucaracha en el estudio.



Con este ejemplo, se da un acercamiento a la forma de abordar el análisis de supervivencia. Se dice acercamiento pues de acuerdo a los intereses del investigador, se puede profundizar el análisis con las herramientas que ofrece el paquete estadístico **R** presentadas en este trabajo y analizar con cuidado las funcíones de supervivencia que mejor expliquen las inquietudes que se pretendan resolver con un modelo de supervivencia.

Comentarios finales

El presente trabajo es una introducción al análisis de supervivencia. Se ha dado la motivación para recurrir a esta herramienta al explicar los conceptos principales del análisis de supervivencia y la manera de estimar e interpretar un modelo de supervivencia utilizando el paquete estadístico R, el cual tiene la ventaja de ser *software* libre, lo cual ha propiciado la aportación de muchos investigadores a nivel mundial al paquete. En particular, el análisi de supervivencia se ve favorecido por esta situación pues la librería "*survival*" utilizada en este trabajo es continuamente actualizada con rutinas que utilizan métodos fundamentados en modelos de supervivencia mejorados por investigadores de gran parte del mundo.

A lo largo del trabajo se presenta la forma de realizar el análisis de supervivencia utilizando el paquete estadístico R, siendo una herramienta que permite estimar el modelo de supervivencia y obtener la información que éste contiene. No sólo se explica la manera de utilizar el paquete con las funciones básicas, se ha tenido cuidado de explicar como obtener la mayor información posible de éstas, así como la manera de mejorar la presentación de la información gráfica para su mejor interpretación y entendimiento al considerar que la aportación visual facilita el análisis del modelo.

En cada capítulo se ha buscado una interpretación y justificación de la existencia de los elementos utilizados con la finalidad de proporcionar las herramientas necesarias para la utilización adecuada de este trabajo, esperando que sea motivación para profundizar el estudio en el área de supervivencia, pues existen más modelos con covariables, el desarrollo del modelo de supervivencia mediante procesos de conteo, por mencionar algunos tópicos que no se tocan en este trabajo por pertenecer a un trabajo más extenso en el estudio de supervivencia.

Un comentario final está dirigido a la percepción usual que se tiene de la aplicación del análisis de supervivencia, pues justificado por su origen y

utilidad, las disciplinas de medicina y biología han sido las más favorecidas por la aplicación de los modelos de supervivencia, de hecho esta es la razón por la cual se llama indistintamente tiempo de vida al tiempo de falla. Los modelos actuariales de supervivencia tienen usos operativos en seguros, donde se pretende modelar la probabilidad de ocurrencia de algún evento de interés con la finalidad de calcular la prima y suma asegurada de algún seguro particular, ya sea de vida o daños. En sistemas de pensiones el interés se centra en la permanencia en las actividades laborales o en seguros de vida, en la permanencia de una persona en un grupo asegurado. El modelo de supervivencia es utilizado tambien como una herramienta actuarial para la creación de tablas de mortalidad, las cuales tienen un papel fundamental en seguros y pensiones. La intención de este comentario es exhortar a la aplicación, en mayor medida, de los modelos de supervivencia en más campos de estudio, pues como se ha visto, la adecuación de un modelo de supervivencia a una situación de interés proporcionará una herramienta útil al análisis de una situación de interés.

Apéndice A

Verosimilitud parcial

El concepto de verosimilitud parcial está dado por la generalización de las ideas de verosimilitud parcial y marginal.

Sea y un vector de observaciones representadas por una variable aleatoria Y con función de densidad $f_Y(y;\theta)$, donde Y es transformada en dos nuevas variables aleatorias (V, W) donde la transformación no depende de parámetros desconocidos. Se llama a $f_V(v;\theta)$ la verosimilitud marginal basada en V y a $f_{W|V}(w|v;\theta)$ la verosimilitud condicional basada en W dado $V = v$, ambas funciones de *theta*.

Generalizando esta idea, sea Y una variable aleatoria transformada en la sucesión

$$(X_1, S_1, X_2, S_2, ..., X_m, S_m).....(1)$$

donde los componentes pueden ser ellos mismos vectores.

La función de verosimilitud completa de la sucesión (1) es

$$f_{X_1}(x_1;\theta) f_{S_1|X_1}(s_1|x_1;\theta)$$

$$f_{X_2|(X_1,S_1)}(x_2|x_1, s_1;\theta)$$

$$\vdots$$

$$f_{S_m|(X_1,S_1,X_2,...X_m,S_{m-1})}(S_m|x_1, s_1, x_2 \ldots x_m, s_{m-1};\theta)...(2)$$

Que puede ser escrita como

$$\prod_{j=1}^{m} f_{X_j|X^{j-1},S^{j-1}}(x_j|x^{j-1}, s^{j-1}; \theta) \prod_{j=1}^{m} f_{S_j|X^j,S^{j-1}}(s_j|x^j, s^{j-1}; \theta)...(3)$$

dode $X^{(j)} = (X_1, X_2, ..., X_j)$, $x^{(j)} = (x_1, x_2, ..., x_j)$, $S^{(j)} = (S_1, S_2, ..., S_j)$ y $s^{(j)} = (s_1, s_2, ..., s_j)$.

Donde el segundo producto de la ecuación (3) es la verosimilitud parcial basada en S en la sucesión $\{X_j, S_j\}$.

Apéndice B

Pruebas basadas en la teoría de verosimilitud para muestras grandes

Muchos de los desarrollos usados en las pruebas, en análisis de supervivencia, están basadas en propiedades asintóticas de la verosimilitud o de la verosimilitud parcial. Estos desarrollos están basados en la verosimilitud maximizada misma (prueba de razón de verosimilitudes) o en los estimadores estandarizados mediante el uso de la matriz de información (prueba de Wald) o en las primeras derivadas de la log verosimilitud (prueba de puntajes).

Sea $\mathbf{Y}$ el vector de datos y $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ el vector de parámetros. Sea $L(\boldsymbol{\theta} : \mathbf{Y})$ la expresión que denota a la función de verosimilitud o la función de verosimilitud parcial. El estimador máximo verosímil de $\boldsymbol{\theta}$ es la función de los datos que maximiza la verosimilitud, esto es $\hat{\boldsymbol{\theta}}(\mathbf{Y}) = \hat{\boldsymbol{\theta}}$ es el valor de $\boldsymbol{\theta}$ que maximiza a $L(\boldsymbol{\theta} : \mathbf{Y})$ o equivalentemente maximiza a $logL(\boldsymbol{\theta} : \mathbf{Y})$.

Asociado con la función de verosimilitud, se tiene el vector de puntajes $\mathbf{U}(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), U_2(\boldsymbol{\theta}), ..., U_p(\boldsymbol{\theta}))$ definido por:

$$U_j(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} lnL(\boldsymbol{\theta} : \mathbf{Y}).$$

En casos regulares, el estimador máximo verosímil es la solución de la ecuación $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$. El vector de puntajes tiene la propiedad de que su esperanza es cero cuando la esperanza es tomada con respecto al verdadero valor de $\boldsymbol{\theta}$.

La segunda cantidad clave en la teoría de verosimilitud para muestras grandes es la matriz de información de Fisher definida por:

APÉNDICE B. PRUEBAS BASADAS EN LA TEORÍA DE VEROSIMILITUD PARA MUESTRAS GRANDES

$$\mathbf{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\mathbf{U}(\boldsymbol{\theta})^t\mathbf{U}(\boldsymbol{\theta})\right] = E_{\boldsymbol{\theta}}\left[-\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta})\right]$$

$$= \left\{-E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} lnL(\boldsymbol{\theta} : \mathbf{Y})\right]\right\}; j = 1, .., p, k = 1, .., p.$$

El cálculo de la esperanza en $\mathbf{I}(\boldsymbol{\theta})$ es muy difícil en la mayoría de las aplicaciones en la teoría de verosimilitud, entonces, es usado un estimador consistente de $\mathbf{I}$. Este estimador es la información observada, $\mathbf{I}(\boldsymbol{\theta})$, donde el (j, k)-ésimo elemento está dado por

$$I_{j,k}(\boldsymbol{\theta}) = \frac{\partial^2 lnL(\boldsymbol{\theta} : \mathbf{Y})}{\partial\theta_j\partial\theta_k}, j, k = 1, .., p.$$

El primer conjunto de pruebas basadas en la verosimilitud, están dadas por la hipótesis nula simple, $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. La primera prueba es la prueba de razón de verosimilitudes basada en la estadística

$$X^2_{LR} = 2\left[lnL(\boldsymbol{\theta}_0 : \mathbf{Y}) - lnL(\hat{\boldsymbol{\theta}} : \mathbf{Y})\right].$$

Esta estadística tiene una distribución asintótica Ji-cuadrada con p grados de libertad bajo la hipótesis nula.

Una segunda prueba, llamada prueba de Wald, está basada en la distribución para muestras grandes de los estimadores de máxima verosimilitud. Para muestras grandes, $\hat{\boldsymbol{\theta}}$ tiene una distribución normal multivariada con media $\boldsymbol{\theta}$ y matriz de covarianzas dada por $\mathbf{I}^{-1}(\boldsymbol{\theta})$ de modo que en forma cuadrática, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^t\mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ tiene una distribución $Ji - cuadrada$ con p grados de libertad para muestras grandes. Usando la información observada como un estimador para la matriz de información de Fisher, la estadística de Wald queda expresada como

$$X^2_W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^t\mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

La cual, cuando H_0 es cierta, tiene una distribución $Ji - cuadrada$ con p grados de libertad para muestras grandes.

La tercera prueba llamada prueba de puntajes, está basada en las estadísticas de puntajes. Cuando $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, el vector de puntajes $\mathbf{U}\ (\boldsymbol{\theta}_0)$ tiene, para muestras grandes, una distribución normal multivariada con media cero y matriz de covarianzas $\mathbf{I}(\boldsymbol{\theta})$. A esto sigue que la estadística de prueba está dada por

$$X_S^2 = \mathbf{U}^t(\boldsymbol{\theta}_0)\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{U}(\boldsymbol{\theta}_0).$$

Como en la prueba de Wald, la información de Fisher es reemplazada en muchas aplicaciones por la información observada. Entonces la estadística de prueba queda dada por

$$X_S^2 = \mathbf{U}^t(\boldsymbol{\theta}_0)\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{U}(\boldsymbol{\theta}_0).$$

De nuevo, esta estadística tiene una distribución $Ji - cuadrada$ con p grados de libertad para muestras grandes cuando H_0 es cierta.

Estas tres estadísticas pueden ser usadas para probar otras hipótesis. Se particiona el vector de parámetros $\boldsymbol{\theta}$ en dos vectores Ψ y Φ de longitudes p_1 y p_2 respectivamente. Se desea probar la hipótesis $H_0 : \Psi = \Psi_0$. Sea $\hat{\Phi}(\Psi_0)$ el estimador máximo verosímil de Φ obtenido al maximizar la verosimilitud con respecto de Φ, con Ψ fijo en Ψ_0. Esto es, $\hat{\Phi}(\Psi_0)$ maximiza $lnL\left[(\Psi_0, \Phi) : \mathbf{Y}\right]$ con respecto de Φ . La partición de la matriz de información $\mathbf{I}$ queda dada por

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{\Psi\Psi} & \mathbf{I}_{\Psi\Phi} \\ \mathbf{I}_{\Phi\Psi} & \mathbf{I}_{\Phi\Phi} \end{pmatrix}.$$

Donde $\mathbf{I}_{\Psi\Psi}$ es de dimensión $p_1 \times p_1$, $\mathbf{I}_{\Phi\Phi}$ es de dimensión $p_2 \times p_2$, $\mathbf{I}_{\Psi\Phi}$ es de dimensión $p_1 \times p_2$ y $\mathbf{I}_{\Phi\Psi}^t = \mathbf{I}_{\Psi\Phi}$.

Note que la matriz de información particionada, tiene una inversa, la cual es también una matriz particionada dada por

$$\mathbf{I}^{-1} = \begin{pmatrix} \mathbf{I}^{\Psi\Psi} & \mathbf{I}^{\Psi\Phi} \\ \mathbf{I}^{\Phi\Psi} & \mathbf{I}^{\Phi\Phi} \end{pmatrix}.$$

Con esta modificación, las tres estadísticas para probar $H_0 : \Psi = \Psi_0$ están dadas por:

Prueba de razón de verosimilitudes:

$$X_{LR}^2 = -2\left\{lnLik(\Psi_0, \hat{\Phi}(\Psi_0) : \ \mathbf{Y}) - lnLik(\hat{\Phi} : \mathbf{Y})\right\}.$$

Prueba de Wald:

$$X_W^2 = \left(\hat{\Psi} - \Psi_0\right)^t \left[I^{\Psi\Psi}(\hat{\Psi}, \hat{\Phi})\right]^{-1} \left(\hat{\Psi} - \Psi_0\right).$$

Prueba de puntajes:

APÉNDICE B. PRUEBAS BASADAS EN LA TEORÍA DE VEROSIMILITUD PARA MUESTRAS GRANDES

$$X_S^2 = U_\Psi^t \left[\Psi_0, \hat{\Phi}(\Psi_0)\right] \left[I^{\Psi\Psi}(\Psi_0, \hat{\Phi}(\Psi_0))\right] U_\Psi \left[\Psi_0, \hat{\Phi}(\Psi_0)\right].$$

Estas tres estadísticas tienen una distribución $Ji-cuadrada$ con p_1 grados de libertad cuando la hipótesis nula es cierta.

Apéndice C

Datos utilizados

- Base de datos *"cucaracha"*. Tiempo de falla de 150 cucarachas expuestas a los insecticidas A, B y C en un experimento con tiempo final de censura de 50 días. Fuente: Curso de estadística III de la profesora Margarita E. Chávez Cano. Facultad de Ciencias. UNAM.

	death	status	weight	group
1	20	1	5.385	A
2	34	1	7.413	A
3	1	1	9.266	A
4	2	1	6.228	A
5	3	1	5.229	A
6	3	1	9.699	A
7	50	0	1.973	A
8	26	1	5.838	A
9	1	1	2.088	A
10	50	0	0.237	A
11	21	1	6.814	A
12	3	1	5.502	A
13	13	1	1.137	A
14	11	1	6.323	A
15	22	1	7.384	A
16	50	0	8.713	A
17	50	0	7.458	A
18	1	1	1.424	A
19	50	0	1.312	A
20	9	1	5.162	A
21	50	0	7.187	A
22	1	1	4.677	A

23	13	1	6.548	A
24	50	0	5.903	A
25	50	0	2.113	A
26	1	1	7.617	A
27	6	1	3.737	A
28	50	0	8.972	A
29	50	0	6.523	A
30	50	0	2.165	A
31	36	1	4.895	A
32	3	1	6.538	A
33	46	1	1.674	A
34	10	1	6.726	A
35	50	0	2.671	A
36	1	1	4.949	A
37	18	1	4.819	A
38	3	1	5.080	A
39	36	1	3.532	A
40	37	1	4.406	A
41	50	0	6.286	A
42	7	1	5.529	A
43	1	1	2.270	A
44	1	1	5.245	A
45	7	1	9.675	A
46	24	1	5.610	A
47	4	1	4.297	A
48	50	0	3.179	A
49	12	1	6.776	A
50	17	1	0.466	A
51	1	1	0.626	B
52	1	1	1.221	B
53	1	1	0.124	B
54	21	1	0.320	B
55	50	0	2.282	B
56	50	0	0.287	B
57	1	1	3.468	B
58	46	1	7.314	B
59	50	0	4.901	B
60	1	1	5.418	B
61	8	1	6.344	B
62	2	1	1.163	B
63	12	1	12.126	B

64	3	1	11.561	B
65	2	1	8.333	B
66	1	1	0.055	B
67	5	1	10.583	B
68	50	0	9.534	B
69	1	1	13.182	B
70	2	1	10.156	B
71	2	1	16.881	B
72	4	1	15.452	B
73	17	1	16.831	B
74	5	1	18.947	B
75	1	1	19.099	B
76	11	1	19.000	B
77	8	1	9.652	B
78	1	1	1.544	B
79	5	1	10.786	B
80	2	1	4.130	B
81	41	1	2.200	B
82	5	1	7.567	B
83	21	1	14.581	B
84	1	1	26.259	B
85	38	1	0.440	B
86	50	0	18.188	B
87	3	1	6.789	B
88	19	1	16.669	B
89	4	1	38.177	B
90	7	1	29.154	B
91	1	1	14.578	B
92	46	1	1.569	B
93	2	1	0.345	B
94	5	1	33.929	B
95	40	1	28.958	B
96	4	1	38.139	B
97	50	0	26.822	B
98	2	1	39.501	B
99	1	1	9.264	B
100	17	1	22.880	B
101	7	1	27.480	C
102	1	1	35.069	C
103	5	1	4.974	C
104	1	1	41.521	C

105	1	1	42.090	C
106	5	1	25.037	C
107	6	1	9.509	C
108	2	1	23.682	C
109	24	1	0.352	C
110	1	1	19.589	C
111	1	1	7.426	C
112	1	1	7.913	C
113	1	1	2.370	C
114	7	1	5.533	C
115	13	1	18.800	C
116	6	1	18.508	C
117	11	1	3.343	C
118	46	1	26.926	C
119	5	1	2.388	C
120	14	1	21.567	C
121	2	1	5.594	C
122	1	1	17.150	C
123	20	1	15.986	C
124	2	1	1.588	C
125	20	1	2.055	C
126	1	1	16.074	C
127	23	1	12.086	C
128	11	1	20.524	C
129	1	1	6.493	C
130	1	1	7.258	C
131	20	1	16.635	C
132	9	1	10.324	C
133	1	1	5.228	C
134	1	1	0.784	C
135	1	1	5.587	C
136	1	1	5.011	C
137	7	1	7.441	C
138	11	1	3.690	C
139	1	1	4.708	C
140	3	1	9.207	C
141	1	1	1.400	C
142	5	1	6.309	C
143	9	1	1.784	C
144	21	1	0.767	C
145	10	1	1.993	C

146	11	1	1.030	C
147	30	1	2.875	C
148	1	1	1.820	C
149	1	1	0.974	C
150	17	1	0.100	C

Bibliografía

[1] Collet, D. (1994). Modelling Survival Data in Medical Research. Chapman & Hall.

[2] Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B 34, 187-220.

[3] Cox, D. R. (1975). Partial Likelihood. Biometrika 62, 264-276.

[4] Cox, D. R. and Oakes, D. (1984). Analysis of Survival Data. Chapman & Hall.

[5] Dalgaard, Peter. (2002). Introductory Statistics whith R. Springer.

[6] Klein, John P. and Moeschberger, Melvin L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer.

[7] Korosteleva, O. (2009). Clinical Statistics. Introducing Clinical Trials, Survival Analysis and Longitudinal Data Analysis. Jones ana Bartlett Publishers.

[8] Link, Carol L. (1984). Confidence Intervals for the Survival Function using Cox´s Proportional-Hazard Model with Covariates. Biometrics 40, 601-610.