Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas
Postgrado en Ciencias Matemáticas

Inferencia sobre puntos de cambio en el modelo de riesgo proporcional para un problema de trasplante

TESIS

PARA OBTENER EL TÍTULO DE:

Maestra en Ciencias Matemáticas

Presenta:

Lic. Sonia Venancio Guzmán

Director de tesis:

Dr. Bulmaro Juárez Hernández

Puebla, Puebla, ENERO 2020

Inferencia sobre puntos de cambio en el modelo de riesgo proporcional para un problema de trasplante

Venancio Guzmán Sonia

Dedicatoria

A mis padres, Esperanza y Albino,
por todo el apoyo que me han dado.

A mis hermanos,
Flor, Eduardo, Leticia,
Jesús, Julieta, Marisol y
Albino.

DRA. HORTENSIA J. REYES CERVANTES Presidente
DR. VÍCTOR HUGO VÁZQUEZ GUEVARA Secretario
DRA. GLADYS LINARES FLEITES Vocal
DR. FERNANDO VELASCO LUNA Vocal
DR. BULMARO JUÁREZ HERNÁNDEZ Director de Tesis

P R E S E N T E.

Me permito comunicar a usted que la fecha programada para realizar el **EXAMEN** de *MAESTRÍA EN CIENCIAS (MATEMATICAS),* del (a) **C. SONIA VENANCIO GUZMÁN,** con la tesis titulada:

"Inferencia sobre puntos de cambio en el modelo de Riesgo Proporcional para un problema de trasplante"

es el día **14** de **noviembre** del año **2019** a las **14:00 hrs.**, en el(la) **Sala Audiovisual II (FM9/109)** de esta Facultad.

Por la atención que se sirva dar al presente, anticipo a usted las más sinceras gracias.

A T E N T A M E N T E
"PENSAR BIEN, PARA VIVIR MEJOR"
H. Puebla de Z., a 7 de noviembre de 2019

DRA. LIDIA AURORA HERNÁNDEZ REBOLLAR
SECRETARIA DE INVESTIGACIÓN Y ESTUDIOS DE POSGRADO

DRA.LAHR/mtrv

Facultad de Ciencias Físico Matemáticas
Av. San Claudio y 18 sur, edif. FM1
Ciudad Universitaria, Col. San Manuel, Puebla, Pue. C.P. 72570
01 (222) 229 55 00 Ext. 7550 y 7552

Agradecimientos

Agradezco a cada una de las personas que están en mi vida porque cada una de ellas siempre me enseñan algo nuevo y porque siempre están impulsandome a crecer cada día. Especialmente agradezco a mis padres por todo el apoyo y amor que siempre me han brindado sin esperar nada a cambio.

Agradezco también a cada una de las personas e instituciones que me apoyaron en este trabajo de tesis:

A mi asesor el Dr. Bulmaro Juárez Hernández por brindarme sus grandes conocimiento acerca de la especialidad, por su gran paciencia y sobre todo por el tiempo dedicado al presente trabajo.

A la Dra. Hortensia Josefina Reyes Cervantes, al Dr. Víctor Hugo Vázquez Guevara, a la Dra. Gladys Linares Fleites y al Dr. Fernando Velasco Luna, por su apoyo en cada una de las observaciones hechas acerca de la tesis y por el tiempo dedicado, ya que gracias a ello se obtuvo una mejor versión.

A la Facultad de Ciencias Físico Matemáticas de la BUAP por brindarme la oportunidad de realizar la Maestría en Ciencias Matemáticas.

Al Consejo Nacional de Ciencia y Tecnología (Conacyt) por el apoyo económico que me brindo durante la estancia en la maestría.

Introducción

Los antecedentes del Análisis de Supervivencia se pueden situar alrededor del siglo $XVII$, en la construcción de las primeras tablas de mortalidad por el astrónomo Edmon Halley, las cuales publicó a partir del registro de funerales y nacimientos en la ciudad de Breslau. Sin embargo, originalmente se aplicó en la Ingeniería, cuando se trataba de analizar la duración, la resistencia y la fiabilidad de materiales y/o componentes de una máquina. Así pues, el Análisis de Supervivencia se convierte en una herramienta muy importante de estudio, al permitir generalizar no sólo el análisis de respuestas tales como son: "si/no","falleció/sobrevivió", sino también porque permite el análisis del tiempo en que se tarda en ocurrir un determinado suceso [2].

El Análisis de Supervivencia es por tanto, una técnica estadística que tiene como objetivo primordial modelar el tiempo hasta ocurrir un determinado suceso para un grupo de individuos, el cual cabe aclarar, no es necesariamente hasta el tiempo de muerte, ya que muchas veces suele ocurrir que durante el análisis, individuos que se van incorporando en el estudio desaparezcan o bien que al entrar terminen antes de que el suceso que se pretende estudiar se llegue a realizar, esto en realidad es una característica que permite diferenciar al Análisis de Supervivencia de otros análisis estadísticos.

Cuando se tiene información incompleta de la supervivencia de algunos individuos se da lugar a lo que se conoce como individuos o datos censurados. El Análisis de Supervivencia se caracteriza también por incluir información que proporcionan estos individuos, por lo cual son considerados como datos relevantes durante el análisis. Para realizar un estudio de supervivencia, es necesario contar con un inicio y un fin bien definidos, el inicio no necesariamente es el mismo para cada individuo, de hecho, los períodos de seguimiento en este tipo de análisis son casi siempre diferentes, ya que en la práctica es muy frecuente ver que nuevos individuos se van incorporando en el estudio durante todo el período de la observación, por lo que estos últimos en hacerlo, son igualmente analizados, aunque durante un período de tiempo más pequeño que los que empezaron al principio.

Dos conceptos muy importantes para el Análisis de Supervivencia son dos tipos de probabilidades: de supervivencia y de riesgo. La probabilidad de supervivencia o función de supervivencia (comúnmente denotada por $S(t)$) hace referencia a la probabilidad de que un individuo sobreviva a partir de su fecha de entrada, hasta un momento determinado. Es decir, esta función se centra más en la no ocurrencia del evento. Mientras que la función de riesgo (llamada también tasa de fallo condicional), es la probabilidad, por unidad de tiempo,

de que a un individuo que está siendo observado en un tiempo determinado (entre el tiempo t y $t + \Delta t$) le suceda el evento esperado, dado que el sujeto vivía en el tiempo t.

Por otra parte, también es necesario hablar acerca de los métodos utilizados para realizar dicho análisis, estos pueden llevarse a cabo a través de modelos paramétricos o bien, a través de estudios no paramétricos. Las pruebas paramétricas, requieren que la distribución de la población sea caracterizada por ciertos parámetros, sin embargo, las pruebas no paramétricas no parten de este supuesto, en este caso, se parte del supuesto de que el conjunto de datos no proviene de una distribución ya conocida.

En particular, existen dos métodos no paramétricos más utilizados: el de Kaplan-Meier y de Riesgos Proporcionales de Cox, el cual en realidad es un método semiparamétrico. El método de Kaplan-Meier permite comparar supervivencia entre grupos o entre categorías, por ejemplo, considerando la categoría sexo y grupos con la edad, se pueden obtener las curvas de supervivencia de hombres y mujeres, o bien , las curvas considerando intervalos de edad de los sujetos en estudio. Sin embargo, este método de Meier no permite modelar la influencia de predictores, como es el caso del Modelo de riesgos proporcionales. Este modelo de Cox, es muy utilizado por su flexibilidad y puesto que es mucho más simple que otros modelos a la hora de interpretar los coeficientes. Otra característica importante de este modelo, es que es usado para encontrar las relaciones entre el riesgo que se produce en un determinado individuo y algunas variables independientes y las variables explicativas, por lo que este modelo nos permite evaluar dentro de un conjunto de variables cuáles tienen relación o influencia sobre la función de riesgo y en consecuencia también en la función de supervivencia ([2], [10]).

Durante las últimas décadas el Análisis de Supervivencia se ha convertido en una herramienta muy importante, más en el ámbito clínico [8], ya que nos permite modelar el comportamiento de la supervivencia de los pacientes que tienen alguna enfermedad. Sin embargo, a pesar de que es un campo ya bastante desarrollado, sigue siendo objeto de estudio, pues en la realidad se cuenta con una gran diversidad de datos disponibles.

Este trabajo de tesis está enfocado al Análisis de Supervivencia, su principal objetivo es mostrar una aplicación de dicha técnica estadística, con un grupo de pacientes con insuficiencia renal a los cuales se les ha practicado un trasplante de riñón y se les tiene en observación para ver si presentan rechazo al injerto realizado. Así, se cuenta con una base de datos de pacientes del Instituto Mexicano del Seguro Social (IMSS) del estado de Puebla, de dichos pacientes se cuenta con diversa información clínica ya que estas son necesarias a la hora de realizar el análisis, por lo cual dichas variables son evaluadas para ver cuales de ellas son significativas en el estudio.

Por otro lado, cabe resaltar que información de casos clínicos difícilmente está al alcance de todos para poder ser analizados, ya que es información delicada de tratar, y aún cuando se cuenta con dichos datos, es difícil de analizarla

por completo puesto que valores de algunas variables de los pacientes no son registradas o sólo dan una idea muy baga de lo que se pretende registrar, por lo cual, en el presente trabajo, las variables que se consideran para realizar el análisis fueron aquellas que si tenían en mayor proporción información completa, dejando aun lado aquellas que tenía demasiados datos faltantes, aunque en el estudio pudiera ocurrir que algunas de estas variables son también significativas.

Índice general

Capítulo 1

CONCEPTOS BÁSICOS

El Análisis de Supervivencia es un conjunto de técnicas estadísticas, que consisten en la medición del tiempo que transcurre entre un evento inicial (inclusión en el estudio de un individuo) hasta que se llega a producir un evento de interés para cierto grupo de individuos (por lo general, este evento es de falla, muerte, recaída o el desarrollo de determinada enfermedad), es decir, es el conjunto de métodos estadísticos para analizar datos de supervivencia. Al periodo de tiempo transcurrido entre estos dos eventos se le denomina tiempo de falla, tiempo de supervivencia o tiempo de vida.

Para determinar el tiempo de falla, se requieren tres requisitos. Primero, el tiempo de origen debe estar bien definido para cada sujeto en estudio (ya que no necesariamente es el mismo) por lo cual su tiempo de falla se mide desde su propia fecha de entrada. Segundo, la escala con la cual se medirá el tiempo debe estar fijada, ya que este puede variar dependiendo del problema que se esté atendiendo, su único requisito es que el tiempo de falla de los individuos sean no negativos. Por último, el significado de falla debe ser claro ([3], [18]).

1.0.1. Distribución en los datos de supervivencia

Uno de los objetivos principales para realizar el análisis a través del tiempo, es llegar a conocer la distribución del tiempo de supervivencia de un sólo grupo de individuos en observación, sin embargo, en muchos casos es importante llegar a realizar una comparación de los tiempos de supervivencia de dos o más grupos.

En muchos problemas, se parte del supuesto de que la distribución de los datos es normal, sin embargo, existen muchos casos en que esto no siempre ocurre, pues en su gran mayoría los datos no tienen un comportamiento simétrico como es el caso de la distribución normal. De hecho, en la práctica suele ser difícil encontrar un buen modelo que se ajuste adecuadamente a un conjunto de datos, esto motiva a no sólo trabajar con modelos paramétricos, si no también con modelos semiparamétricos, ya que estos no presuponen una distribución conocida para los datos con los que se cuenta.

1.0.2. Censura y Truncamiento

Una característica importante del Análisis de Supervivencia es el hecho de que se puede trabajar con información incompleta sobre los datos con los que se cuenta ([9], [18]).
Suponga que se establece el tiempo de seguimiento de cierto grupo de individuos, esto es, se establece el tiempo de origen t_0 de manera clara y se define t_f como el tiempo en que finaliza el estudio. Durante el periodo de seguimiento puede ocurrir que uno o varios de los individuos abandonen el estudio al tiempo C de modo que la observación sobre estos individuos cesa en este tiempo aunque no se haya experimentado la falla, provocando con ello que sólo se registre información parcial de dichos individuos de su tiempo de supervivencia, este puede ser mayor que C o incluso que t_f, a estos tipos de datos se le conoce como datos censurados y a los individuos que llegan a experimentar dicho evento se les conoce como datos no censurados.

Observación 1.1. Al igual que el tiempo de falla, la censura es un evento puntual, además, en algunos casos el tiempo de censura llega a ser conocida (fija) o puede convertirse en una variable aleatoria.

Definición 1.1.1. (**Censura**) Se dice que el tiempo de supervivencia de un individuo está censurado cuando el evento de interés no ha sido observado durante el tiempo de seguimiento.

Mecanismos de censura

Las observaciones censuradas son indudablemente una de las características más importantes de los datos de supervivencia, la censura puede presentarse de diversas formas. Por lo cual se describen sus distintas clasificaciones.

Censura por la derecha

Esto ocurre cuando el individuo no experimenta el evento de interés durante todo el tiempo de observación, por lo cual el tiempo de falla no es conocido, sin embargo, se conoce una cota inferior para dicho tiempo. Por otra parte, en función de la causa podemos encontrar distintos tipos de censura por la derecha, los cuales son:

i. **Censura tipo I.** Este tipo de censura ocurre cuando el tiempo final t_f del estudio es prefijado y por tanto el tiempo de censura es conocida y fijo.

    1. *Fija.* Aquí el tiempo de origen t_0 y el tiempo final t_f del estudio es el mismo para cada uno de los individuos.
    2. *Progresiva.* Los datos se dividen en grupos y cada uno de los grupos tiene un tiempo de censura asociado.
    3. *Generalizada.* A cada individuo se le asocia un tiempo de origen y un tiempo de censura correspondiente.

ii. **Censura tipo II.** Un individuo experimenta censura tipo II cuando el tiempo de finalización del estudio no ocurre en un tiempo prefijado, sino que se continúa el estudio hasta que ocurre el evento de interés para cierta proporción de la población, de modo que el tiempo de duración del estudio es aleatorio pero el número de observaciones censuradas y no censuradas es fijo. Esto es, suponga que se tiene una muestra aleatoria $T_1, T_2, \ldots, T_n$, para esta muestra se considera que el tiempo en el que se debe de terminar el estudio es cuando para cierto número de sujetos predeterminado, digamos r, se presenta el evento de interés, así, consideramos los primeros r tiempos de falla más pequeños:

$$T_{(1)}, T_{(2)}, \ldots, T_{(r-1)}, T_{(r)}.$$

Con esto, los $n - r$ individuos restantes son censuradas por la derecha por el tiempo de falla de la r-ésima observación de esta muestra de estadísticos de orden.

Notación 1. Sea $T_1, T_2, \ldots, T_n$ una muestra aleatoria y sea C_i el tiempo de censura por la derecha de la i-ésima información, observamos los pares:

$$(T_1^*, \delta_1), \ , (T_2^*, \delta_2), \ldots, (T_n^*, \delta_n),$$

donde

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1, & \text{si } T_i \leq C_i \text{ (no censurado)}, \\ 0, & \text{si } T_i > C_i \text{ (censurado)} \end{cases}$$

y

$$T_i^* = \text{mín}(T_i, C_i) = \begin{cases} T_i, & \text{si } \delta_i = 1, \\ C_i, & \text{si } \delta_i = 0. \end{cases}$$

A la función δ se le conoce como función indicadora de censura, esta toma el valor 0 si el individuo está censurado y 1 si no es censurado.

Censura aleatoria

Este tipo de censura es también conocida como censura no informativa que se encuentra determinado por un fenómeno aleatorio no esperado, lo cual impide el seguimiento de la observación del individuo en estudio, es decir, aquí se asume que cada individuo tiene un tiempo de falla T y un tiempo de censura asociado C, donde T y C son *v.a.'s* continuas e independientes cuyas funciones de supervivencia son también independientes.

Las principales razones por las cuales puede haber dicha censura son:

i. *Que haya pérdida del seguimiento.* Esto ocurre comúnmente cuando el individuo decide dejar el estudio por alguna razón, de modo que la única información que proporciona es la del tiempo en el que fue observado por última vez, sin haber experimentado el evento de interés.

ii. *Es removido del estudio.* En ocasiones algunos individuos llegan a experimentar algún otro evento que puede o no ser ajeno al de interés, en caso de ser completamente ajeno, el individuo es retirado del estudio.

iii. *Terminación del estudio.* Si el sujeto llega al final del tiempo de estudio sin haber experimentado el evento de interés.

Censura por la izquierda

Este tipo de censura ocurre cuando, dado $t > 0$ se sabe que el tiempo de falla ocurre dentro del intervalo $[0, t]$, sin embargo, no se sabe con exactitud en qué momento ocurrió.

Dada una muestra aleatoria $T_1, T_2, \ldots, T_n$, la función indicadora de censura está dada por

$$\delta_i = I(T_i \geq C_i) = \begin{cases} 1, & \text{si } T_i \geq C_i \text{ (no censurado)}, \\ 0, & \text{si } T_i < C_i \text{ (censurado)}. \end{cases}$$

Para ilustrar esquemáticamente la censura, considere la Figura 1.1. Aquí puede ver el tiempo de estudio para 8 individuos en un ensayo clínico, en donde el tiempo de ingreso al estudio está representado por un (•), los individuos que mueren (fallan) son representados por (**M**), los individuos que se pierden durante el seguimiento por (**P**), y los que todavía están vivos al final del período de observación por (**C**) [2].

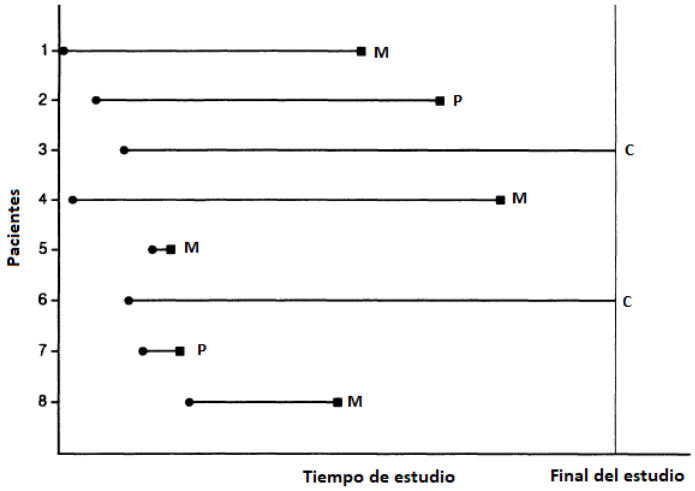


Figura 1.1: Tipos de censura; sujetos que fallan (M), perdidos (P) o censurados (c). Fuente: [2].

Censura por intervalo

Considere una sucesión ordenada de tiempos ascendentes $t_0 < t_1 < \cdots < t_{n-1} < t_n$, se dice que un individuo presenta censura por intervalo si se sabe que su tiempo de falla ocurre en algún intervalo de la sucesión, digamos $(t_{j-1}, t_j]$, $j \in \{1, 2, \ldots, n-1\}$, pero al igual que en la censura por la izquierda se desconoce el tiempo exacto.

Truncamiento

El truncamiento de los datos de supervivencia se presenta sólo cuando se analizan aquellos individuos en estudio que llegan a experimentar el evento de interés y que se encuentran dentro de un intervalo tiempo $(U, V]$, así, si T es una variable aleatoria continua y positiva tal que en el tiempo t_f ocurre el evento de interés, entonces sólo nos interesa si $t_f \in (U, V]$, con lo cual se puede ver que la inferencia de los datos truncados se limitan a la estimación condicional. Los individuos cuyo tiempo de falla no se encuentran dentro del intervalo no son considerados, por lo cual no aportan ninguna información al análisis. Esta es una característica importante que lo diferencia de la censura, puesto que datos censurados proporcionan información parcial al investigador.

Al igual que la censura, existen tipos de truncamiento. Uno de ellos es el truncamiento a la izquierda, el cual ocurre cuando el límite superior del intervalo está dado por $V = \infty$, es decir, cuando el tiempo de falla $t_f > U$. Otro, es el truncamiento a la derecha que se da cuando $U = 0$, por lo cual decimos que $0 < t_f \leq V$ [10].

1.1.1. Distribución continua y función de supervivencia

Sean (Ω, S, P) un espacio de probabilidad y $T : \Omega \to \mathbb{R}^+$ una variable aleatoria $(v.a.)$ definida sobre este espacio. Consideré la función de distribución $(f.d.d.)$ de T,

$$F(t) = P(T \leq t) = P(\{\omega \in \Omega : T(\omega) \leq t\}),$$

donde, $\{\omega \in \Omega : T(\omega) \leq t\} \in S$ para cada $t \in \mathbb{R}^+$. Para el caso continuo se tiene que:

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(x)dx,$$

con $f(\cdot)$ la función de densidad $(f.d.p.)$ de T. Definimos ahora, la función de supervivencia, la cual describe a la distribución del tiempo de supervivencia T, esto es, esta función cuantifica la capacidad que tiene un individuo de sobrevivir más allá del tiempo t [3].

Definición 1.1.2. La función de supervivencia *(f.s.)* denotada por $S(\cdot)$ se define como:

$$S(t) = P(T > t) = 1 - F(t) = \int_{t}^{\infty} f(x)dx.$$

Con la igualdad anterior y haciendo uso del teorema fundamental del cálculo, se obtiene que:

$$f(t) = \frac{dF(t)}{dt} = -\frac{d}{dt}(1 - F(t)) = -\frac{d}{dt}S(t). \tag{1.1}$$

Por otro lado, la función de supervivencia cumple con las siguientes propiedades:

1. $0 \leq S(t) \leq 1, \forall\, t \in \mathbb{R}^+$,

2. es monótona decreciente,

3. $S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$,

4. $S(t)$ es continua por la derecha.

1.1.2. Función de riesgo

La función de riesgo también conocida como tasa de falla y denotada por $h(\cdot)$, es una función muy importante en el estudio de Análisis de Supervivencia, ya que nos indica la tasa instantánea de falla en el tiempo t, condicionado a que no ha ocurrido el evento de interés en el estudio, es decir, nos va dando una idea de cómo va cambiando la probabilidad del evento de interés en un instante de tiempo, dado que aún no se ha presentado el evento al inicio de este. Formalmente se da su definición a continuación.

Definición 1.1.3. La función de riesgo (*f.r.*), denotada por $h(\cdot)$ se define como

$$h(t) = \lim_{\Delta\to 0^+} \frac{P(t \leq T < t + \Delta | T > t)}{\Delta}. \tag{1.2}$$

Para ver la relación de la función de riesgo con la función de supervivencia, hacemos el siguiente análisis. Consideremos la probabilidad condicional de que el objeto en estudio falle en el intervalo de tiempo $[t, t + \Delta)$ para Δ suficientemente pequeño, dado que se ha sobrevivido hasta el tiempo t, esto es:

$$\begin{aligned} \frac{P(t \leq T < t + \Delta | T > t)}{\Delta} &= \frac{P(\{t \leq T < t + \Delta\} \cap \{T > t\})}{\Delta P(T > t)} \\ &= \frac{P(t \leq T < t + \Delta)}{\Delta P(T > t)}, \end{aligned}$$

de modo que

$$\begin{aligned} h(t) &= \frac{1}{P(T > t)} \lim_{\Delta\to 0^+} \frac{P(t \leq T < t + \Delta)}{\Delta} = \frac{1}{S(t)} \lim_{\Delta\to 0^+} \frac{F(t + \Delta) - F(t)}{\Delta} \\ &= \frac{1}{S(t)} F'(t) = \frac{f(t)}{S(t)}. \end{aligned}$$

Esto motiva a definir también a la función de riesgo como:

$$h(t) = \frac{f(t)}{S(t)}, \ \ t > 0. \tag{1.3}$$

Esto nos lleva a relacionar la *f.s.* con la *f.r.* de la siguiente manera.

Teorema 1.1.4. *Sea T una v.a. continua y sea $S(\cdot)$ su función de supervivencia. La f.s. cumple que:*

$$S(t) = \exp\left(- \int_0^t h(x)dx \right), \ \ \forall\, t \geq 0. \tag{1.4}$$

Demostración:

Se tiene de (1.1) y (1.3) que

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

de modo que por el teorema fundamental del cálculo se obtiene:

$$log(S(t)) - log(s(0)) = -\int_0^t h(x)dx,$$

por tanto, de la propiedad 3 de la *f.s.* se obtiene que:

$$log(S(t)) = -\int_0^t h(x)dx,$$

en consecuencia,

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

■

Como consecuencia de este resultado y de (1.3) se obtiene que:

$$f(t) = h(t)\exp\left(-\int_0^t h(x)dx\right) = h(t)\exp(-H(t)),$$

donde $H(t) = \int_0^t h(x)dx$ es llamada función de riesgo acumulado (*f.r.a.*). También podemos ver algunas de las propiedades que satisface dicha función, las cuales se enuncian a continuación.

Corolario 1.1.5. *Sea T una v.a. continua con función de densidad $f(\cdot)$ y sea $h(\cdot)$ la función de riesgo correspondiente. Entonces*

i. $h(t) \geq 0$, $\forall t > 0$,

ii. $H(t) = \int_0^t h(x)dx < \infty$, *para* $t > 0$,

iii. $\int_0^\infty h(x)dx = \infty$.

Demostración: Note que (i) se sigue de la igualdad (1.3), puesto que dicho cociente está dado por cantidades positivas. Por otra parte, para ver (ii), dado $t > 0$,

$$\int_0^t h(x)dx = -\log(S(t)) < \infty,$$

de esto último también se sigue que:

$$h(x) = \frac{-d\ \log(s(x))}{dx}.$$

Para ver que se cumple (iii), nótese,

$$\int_0^\infty h(x)dx = \lim_{t\to\infty}\int_0^t h(x)dx = \lim_{t\to\infty} -\log S(t) = -\log(0) = \infty$$

la penúltima igualdad es debido a la propiedad 3 de la *f.s.*. ■

Puede ver en la Tabla 1.1 las relaciones entre cada una de las funciones de probabilidad descritas anteriormente.

Características de la función de riesgo

En Análisis de Supervivencia, la función de riesgo juega un papel importante no sólo porque esta nos muestra como es la probabilidad de que el individuo experimente el evento de interés a lo largo del tiempo, sino porque también nos ayuda a ajustar un modelo de probabilidad a una muestra de datos concerniente a tiempos de falla.

Algunas de las principales formas que puede tomar esta función se describen a continuación.

i. **Constante:** una función de riesgo constante o aproximadamente constante, indica que la probabilidad de falla en cualquier momento es la misma, comúnmente esto suele darse en entornos estables y son causados por fenómenos aleatorios. Por ejemplo, el tiempo de falla de choques automovilísticos. Un modelo con esta característica es la Exponencial, ya que la función de riesgo asociada a una variable aleatoria T con esta distribución es constante (esto puede verlo en la Figura 1.2 y en 1.3 la función de supervivencia asociada a T).

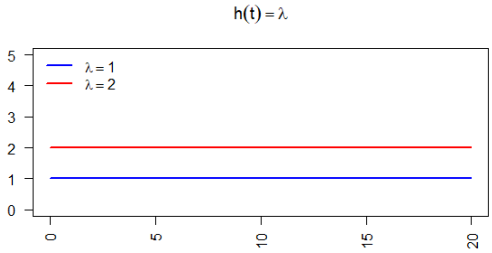


Figura 1.2: *f.r.* para una *v.a* T tipo Exponencial.

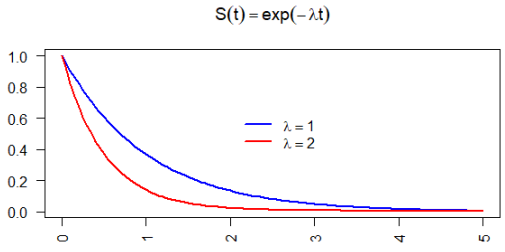


Figura 1.3: *f.s.* para una *v.a* T tipo Exponencial.

ii. **Creciente:** este tipo de función nos dice que la probabilidad de ocurrencia del evento de interés va en aumento con el paso del tiempo. Ejemplos de esto, puede ser cuando individuos están expuestos a esfuerzo constante, o bien sufren algún desgate a lo largo del tiempo.
Se puede ver en la Figura 1.4 que la función de riesgo de una *v.a.* T con distribución de Weibull (que se define en la sección 1.3) de parámetros α y β positivos es creciente cuando $\beta > 1$ y en la Figura 1.5 el comportamiento de su función de supervivencia.

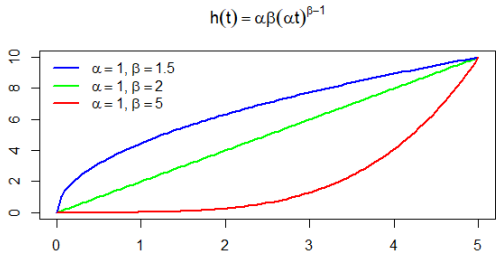


Figura 1.4: *f.r.* para una *v.a.* T tipo Weibull.

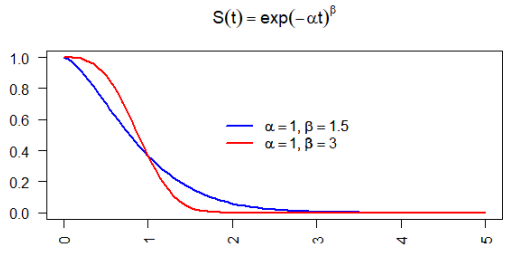


Figura 1.5: *f.s.* para una *v.a.* T tipo Weibull.

Nótese que cuando $\alpha = 1$ y $\beta < 2$ la función de riesgo que se obtiene además de ser creciente es convexa, sin embargo cuando $\alpha = 1$ y $\beta > 2$ la función de riesgo es cóncava y para $\beta = 2$ se obtiene una recta como se puede ver en la figura anterior.

iii. **Decreciente:** la función de riesgo decreciente indica que las fallas son más probables de ocurrir de manera temprana, por ejemplo, en el uso de productos metálicos que con el paso del tiempo se van reforzando.
Ejemplos de funciones de riegos decrecientes también se pueden obtener de la distribución de Weibull de una *v.a.* T con parámetro $\beta < 1$, esto puede verlo en la Figura 1.6 y su función de supervivencia en la Figura 1.7.

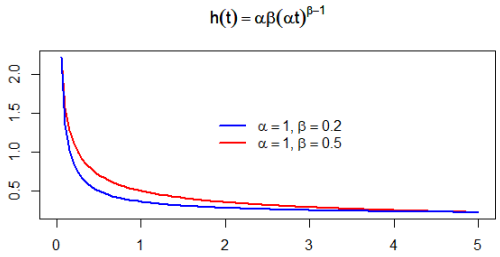


Figura 1.6: *f.r.* para una *v.a.* T con distribución Weibull.

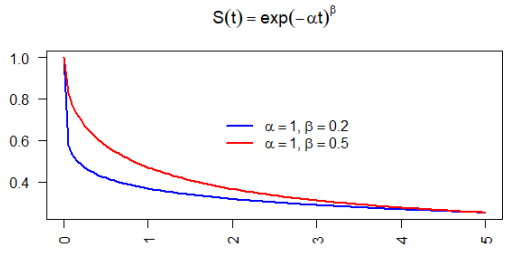


Figura 1.7: *f.s.* para una *v.a.* T con distribución Weibull.

iv. **Forma de bañera:** una función de riesgo se dice que tiene forma bañera debido a las tres etapas que presenta, la primera etapa se caracteriza por tener una elevada tasa de falla que desciende rápidamente con el paso del tiempo. En la segunda etapa, la tasa de falla es mucho menor y constante (o casi constante), lo cual es debido a causas aleatorias externas, por ejemplo, que haya una mala operación, que se haya producido algún accidente, por condiciones inapropiadas u otros. Por último, en la última etapa igualmente se habla de una tasa de falla creciente que normalmente es producido por el desgaste natural a causa del paso del tiempo, esto puede verlo en la Figura 1.8.

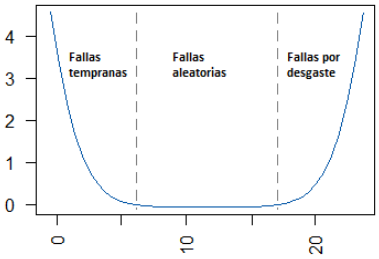


Figura 1.8: Función de Supervivencia en forma de Curva bañera.

	$f(t)$	$S(t)$	$h(t)$	$H(t)$
$F(t)$	$\int_0^t f(x)dx$	1-S(t)	$1-\exp(-\int_0^t h(x)dx)$	$1-\exp(-H(t))$
$f(t)$		$-S'(t)$	$h(t)\exp(-H(t))$	
$S(t)$	$1-\int_0^t f(x)dx$		$\exp(-\int_0^t h(x)dx)$	$\exp(-H(t))$
$h(t)$		$-\frac{d}{dt}\log S(t)$		$H'(t)$
	$\frac{f(t)}{S(t)}$			

Tabla 1.1: Relaciones entre las funciones *f.d.p.* $f(\cdot)$, *f.d.d.* $F(\cdot)$ con la función de supervivencia $S(\cdot)$ y función de riesgo $h(\cdot)$, de la *v.a.* continua T y positiva.

1.1.3. Distribución discreta y su función de supervivencia

Considere ahora que T es una *v.a.* discreta que toma valores $0 = t_0 < t_1 < \cdots$ y sea $f(\cdot)$ la función de densidad de T con función de distribución,

$$F(t) = P(T \leq t) = \sum_{i|t_i \leq t} f(t_i).$$

La función de supervivencia de T está dada por

$$\begin{aligned} S(t) = P(T > t) &= \sum_{i|t_i > t} f(t_i) \\ &= 1 - F(t). \end{aligned} \tag{1.6}$$

Esta función es continua a la derecha, escalonada, no creciente, con $S(t_0) = S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$. Además, se encuentra relacionada también con $F(\cdot)$ y $f(\cdot)$ de la siguiente forma:

$$f(t_j) = F(t_j) - F(t_{j-1}) = S(t_{j-1}) - S(t_j), \quad j = 1, 2, \ldots$$

Observación 1.2. Nótese que, de $S(t_0) = 1$ y de (1.6) se obtiene que $F(0) = 0$, de modo que $f(0) = F(0) = 0$ y $h(t_0) = 0$.

Por otro lado, para ver la expresión de su función de riesgo, considere la siguiente probabilidad condicional,

$$\begin{aligned} P(T = t_i | T \geq t_i) &= \frac{P(\{T = t_i\} \cap \{T \geq t_i\})}{P(T \geq t_i)} = \frac{P(T = t_i)}{P(T \geq t_i)} \\ &= \frac{f(t_i)}{S(t_{i-1})} \quad j = 1, 2, \ldots, \end{aligned}$$

de modo que la función de riesgo es

$$h(t_i) = \frac{f(t_i)}{S(t_{i-1})} \quad i = 1, 2, \ldots, \tag{1.7}$$

o bien, dado que $f(t_i) = S(t_{i-1}) - S(t_i)$, se obtiene que:

$$h(t_i) = \frac{S(t_{i-1}) - S(t_i)}{S(t_{i-1})} = 1 - \frac{S(t_i)}{S(t_{i-1})} \quad i = 1, 2, \ldots. \tag{1.8}$$

Con esta última igualdad podemos ver la relación entre la función de supervivencia y la función de riesgo.

Lema 1.2.1. *Sea T una v.a. discreta con $S(\cdot)$ su función de supervivencia. Para $k \in \mathbb{N}$, $k \leq n$, $t \in [t_k, t_{k+1})$, la f.s. se puede expresar como*

$$S(t) = \prod_{j=1}^{k} \frac{S(t_j)}{S(t_{j-1})} = \prod_{j=1}^{k} (1 - h(t_j)). \tag{1.9}$$

Demostración: Sea $t \in [t_k, t_{k+1})$, que cumplen las condiciones del Lema, se tiene que

$$\begin{aligned} \prod_{j=1}^{k} (1 - h(t_j)) &= \prod_{j=1}^{k} \frac{S(t_j)}{S(t_{j-1})} \\ &= \frac{S(t_1)}{S(t_0)} \frac{S(t_2)}{S(t_1)} \cdots \frac{S(t_k)}{S(t_{k-1})} \\ &= \frac{S(t_k)}{S(t_0)} = S(t_k), \end{aligned}$$

esta última igualdad debido a que $S(t_0) = 1$. Por otro lado, dado que la función $S(\cdot)$ es escalonada y continua a la derecha se tiene que $S(t) = S(t_k)$, con lo cual se obtiene el resultado. ■

Como consecuencia de este lema se obtiene el siguiente resultado.

Teorema 1.2.2. *Sea T una v.a. discreta que toma valores $0 = t_0 < t_1 < \cdots$, con $f(\cdot)$ su función de densidad y $h(\cdot)$ su función de riesgo. Para $i = 0$, $i = 1$, se cumple que $f(t_0) = 0$ y $f(t_1) = h(t_1)$ y para $i \geq 2$ se tiene que,*

$$f(t_i) = h(t_i) \prod_{j=1}^{i-1} (1 - h(t_j)).$$

Demostración: De la Observación 1.2, $f(t_0) = 0$. Luego, del lema anterior se tiene que, $S(t_1) = 1 - h(t_1)$, con lo cual

$$f(t_1) = S(t_1) - S(t_0) = 1 - h(t_1) - 1 = h(t_1),$$

finalmente, de (1.7) se puede ver que

$$f(t_i) = h(t_i)S(t_{i-1}) = h(t_i)\prod_{j=1}^{i-1}(1 - h(t_j)).$$

■

A continuación puede ver en la Tabla 1.2, las relaciones entre las funciones de probabilidad descritas anteriormente, considerando una variable aleatoria discreta.

	$F(t)$	$f(t)$	$S(t)$	$h(t)$
$F(t)$		$\sum_{j\|t_j\leq t} f(t_j)$	$1 - S(t)$	$1 - \prod_{j\|t_j\leq t} (1 - h(t_j))$
$f(t_j)$	$F(t_j) - F(t_{j-1})$		$S(t_{j-1}) - S(t_j)$	$h(t_j) \prod_{i\|t_i\leq t_{j-1}} (1 - h(t_i))$
$S(t)$	$1 - F(t)$	$\sum_{j\|t_j>t} f(t_j)$		$\prod_{j\|t_j\leq t} (1 - h(t_j))$
$h(t_j)$	$\frac{F(t_j)-F(t_{j-1})}{1-F(t_j)}$	$\frac{f(t_i)}{\sum_{i\|t_i\leq t_{j-1}} f(t_i)}$	$1 - \frac{S(t_j)}{S(t_{j-1})}$	
		$\frac{f(t_j)}{S(t_{j-1})}$		

Tabla 1.2: Relaciones entre las funciones *f.d.p.* $f(\cdot)$, *f.d.d.* $F(\cdot)$ con la función de supervivencia $S(\cdot)$ y función de riesgo $h(\cdot)$ de la *v.a.* discreta T.

1.3. Modelos paramétricos

En Análisis de Supervivencia las familias paramétricas se usan para modelar datos de tiempo de supervivencia, entre estos modelos univariados algunos sobresalen más que otros por su gran flexibilidad. A continuación se mencionan algunas de estas familias que son las más usadas para modelar la distribución del tiempo de supervivencia.

1.3.1. Distribución Exponencial

Una de las distribuciones continuas más importantes es la Exponencial ya que tiene una gran utilidad práctica, se utiliza como modelo para representar el tiempo de funcionamiento o el tiempo transcurrido entre eventos que se contabilizan por medio de la distribución de Poisson. Por ejemplo, el tiempo que transcurre entre dos llamadas por teléfono, número de peatones que llegan a un semáforo, ect.

Definición 1.3.1. Diremos que una variable aleatoria continua T sigue una distribución Exponencial de parámetro $\lambda > 0$, lo cual se denota como $T \sim \exp(\lambda)$, si su función de densidad está dada por

$$f(t) = \lambda \exp(-\lambda t), \;\; t \geq 0. \tag{1.10}$$

Con función de distribución dada por,

$$F(t) = 1 - \exp(-\lambda t), \;\; t \geq 0.$$

Así, su función de supervivencia es:

$$S(t) = \exp(-\lambda t), \;\; t \geq 0. \tag{1.11}$$

De (1.10) y (1.11) se obtiene que la función de riesgo es de la forma:

$$h(t) = \lambda, \;\; t \geq 0.$$

Una propiedad que caracteriza a la distribución Exponencial se observa en la función de riesgo, esto es, el riesgo permanece constante a lo largo del tiempo (la probabilidad de ocurrencia del evento de interés siempre es la misma en cualquier momento), a esta propiedad se le conoce como pérdida de memoria.

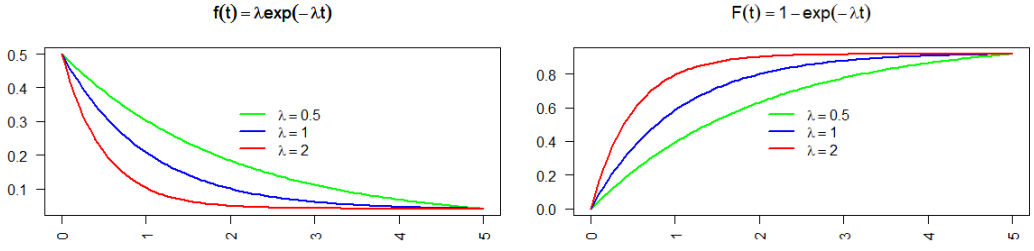
Por otra parte,

$$E(T) = \frac{1}{\lambda}$$

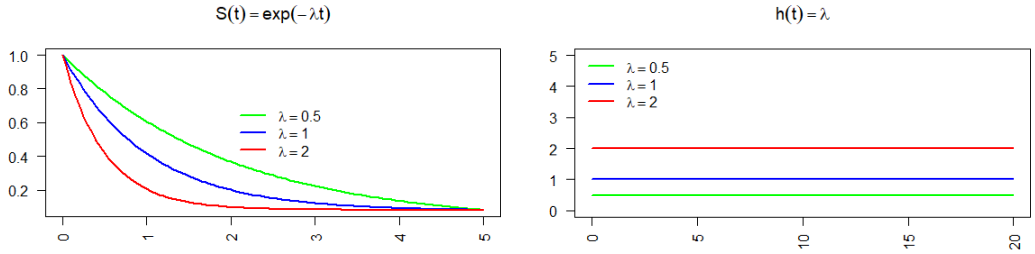
y

$$Var(T) = \frac{1}{\lambda^2}.$$

A continuación se muestra la gráfica de la función de densidad, la $f.d.d.$, la $f.s.$ y la $f.r.$ en la Figura 1.9, para T una variable aleatoria tipo Exponencial con distintos valores del parámetro λ.



(a) (b)



(c) (d)

Figura 1.9: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo de la distribución Exponencial.

1.3.2. Distribución Gamma

El uso de esta distribución es común si se está interesado en la ocurrencia de un evento de interés generado por un proceso de Poisson, ya que la variable que mide el tiempo hasta obtener n ocurrencias del evento de interés sigue una distribución tipo Gamma. Entre sus características principales es que sirve para modelar el comportamiento de variables continuas con asimetría positiva, es decir, para aquellas variables que presentan mayor densidad de sucesos a la izquierda de la media que a la derecha.

Antes de presentar a las *v.a.'s* con distribución Gamma, es necesario recordar como se define la función Gamma, ya que cumple un papel muy importante en muchas ramas de la matemática.

Definición 1.3.2. Dado $k > 0$, se define la función Gamma como:

$$\Gamma(k) = \int_0^\infty x^{k-1} \exp(-x) dx, \quad k > 0.$$

Algunas las propiedades de esta función son:

1. si $k > 1$ entonces $\Gamma(k) = (k-1)\Gamma(k-1)$,
2. si $k \in \mathbb{N}$ entonces $\Gamma(k) = (k-1)!$,
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Otra función importante es la función Gamma incompleta que es una variación de la función Gamma y se define como

$$\gamma(x, y) = \int_0^y u^{x-1} \exp(-u) du, \quad x > 0, y > 0.$$

Definición 1.3.3. Sea T una variable aleatoria continua, se dice que T tiene una distribución Gamma con parámetros k (forma) y λ (escala) positivos ($T \sim G(k, \lambda)$) si su función de densidad está dada por

$$f(t) = \frac{\lambda(\lambda t)^{k-1} \exp(-\lambda t)}{\Gamma(k)}, \quad t \geq 0$$

y la $f.d.d.$ es:

$$\begin{aligned} F(t) &= \frac{1}{\Gamma(k)} \int_0^t \lambda(\lambda x)^{k-1} \exp(-\lambda x) dx \\ &= \frac{1}{\Gamma(k)} \int_0^{\lambda t} u^{k-1} \exp(-u) du \\ &= \frac{\gamma(k, \lambda t)}{\Gamma(k)}, \quad t \geq 0. \end{aligned}$$

Otras propiedades de esta distribución son:

$$E(T) = \frac{k}{\lambda},$$

y

$$Var(T) = \frac{k}{\lambda^2}.$$

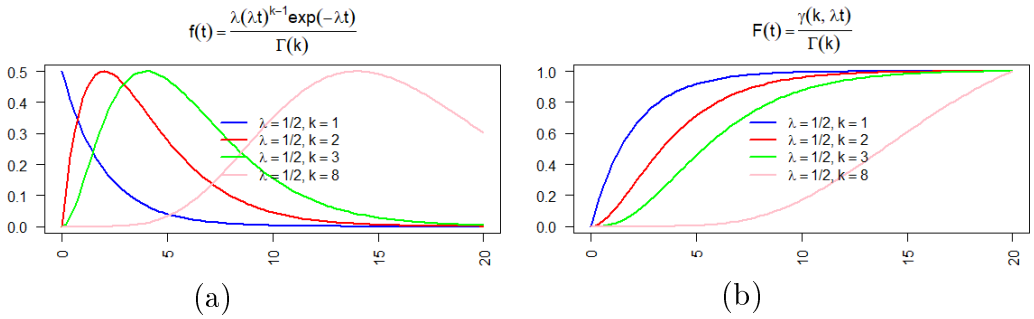
Además,

$$S(t) = 1 - \frac{\gamma(k, \lambda t)}{\Gamma(k)}, \quad t \geq 0,$$

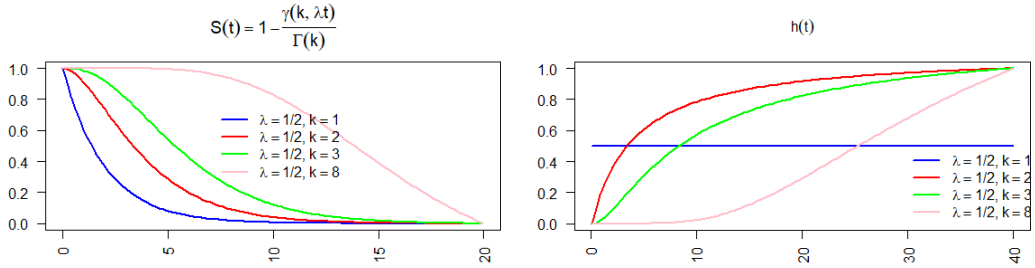
de donde se obtiene la función de riesgo:

$$h(t) = \frac{\lambda(\lambda t)^{k-1} \exp(-\lambda t)}{\Gamma(k) - \gamma(k, \lambda t)}, \quad t \geq 0.$$

Nótese que para $k = 1$, la distribución se reduce a una de tipo Exponencial, y cuando k es un número entero, la distribución es conocida como distribución de Erlang. Esta familia paramétrica es también de gran uso debido a las diferentes formas que puede tomar la función de densidad, puede ver esto en la Figura 1.10, así como el comportamiento de funciones que se derivan de esta.



(a) (b)



(c) (d)

Figura 1.10: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo de la distribución Gamma.

1.3.3. Distribución Weibull

Dada la gran versatilidad de la distribución Weibull, su uso es muy frecuente en estudios de fiabilidad de componentes, ya que una característica importante de esta distribución es la gran variedad de formas que puede tomar el modelo dependiendo de los valores que tomen sus parámetros característicos.

Definición 1.3.4. Diremos que una *v.a.* continua y positiva T sigue una distribución Weibull con parámetros α y β positivos ($T \sim W(\alpha, \beta)$), si tiene función de densidad dada por:

$$f(t) = \alpha\beta(\alpha t)^{\beta-1} \exp\left(-(\alpha t)^{\beta}\right), \quad t \geq 0, \tag{1.12}$$

con función de distribución:

$$F(t) = 1 - \exp\left(-(\alpha t)^{\beta}\right), \quad t \geq 0, \tag{1.13}$$

Al igual que la distribución Gamma, para $\beta = 1$, la distribución es también de tipo Exponencial con parámetro α. Por otro lado, se tiene que:

$$E(T) = \frac{\Gamma\left(1 + \frac{1}{\beta}\right)}{\alpha},$$

$$Var(T) = \frac{\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)}{\alpha^2}.$$

De (1.13) se obtiene que,

$$S(t) = \exp\left(-(\alpha t)^{\beta}\right), \quad t \geq 0, \tag{1.14}$$

y de (1.12) y (1.14) que,

$$h(t) = \alpha\beta(\alpha t)^{\beta-1}, \quad t \geq 0.$$

Esta función de riesgo cumple que para $\beta < 1$, $h(t)$ es decreciente, mientras que para $\beta > 1$ la función es creciente. Para ver los distintos comportamientos que pueden tomar cada una de estas funciones descritas, vea la Figura 1.11.

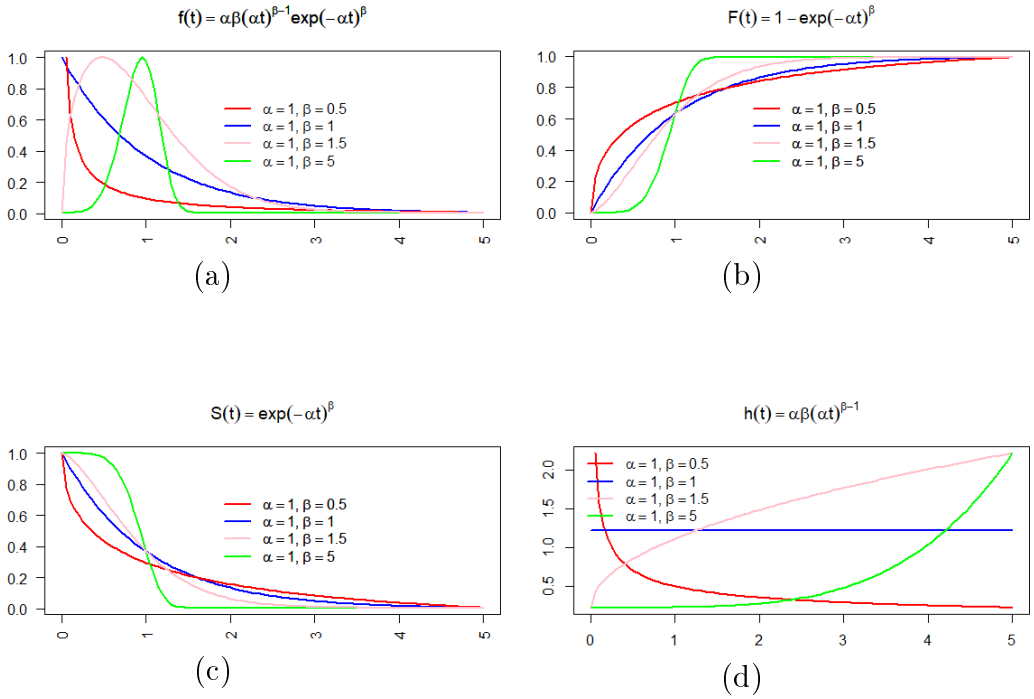


Figura 1.11: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo de la distribución Weibull.

1.3.4. Distribución log-Normal

Esta distribución puede resultar particularmente útil para modelar datos que sean aproximadamente simétricos o asimétricos a la derecha. Además al igual que otras distribuciones, esta puede tener diversas formas dependiendo de su parámetro de escala.

Definición 1.3.5. Sea T una *v.a.* continua, se dice que T tiene distribución log-Normal con parámetros $\mu \in \mathbb{R}$ y $\sigma^2 \in \mathbb{R}^+$ ($T \sim \log\ N(\mu, \sigma^2)$), si $\log(T)$ se distribuye normalmente con media μ y varianza σ^2 ($\log(T) \sim N(\mu, \sigma^2)$), es decir, si su función de densidad está dada por

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\}, \quad t \geq 0 \tag{1.15}$$

con *f.d.d.*,

$$F(t) = \int_0^t \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right\} dx, \quad t \geq 0.$$

Mediante un cambio de variable, esta función también puede ser expresada como

$$F(t) = \int_{-\infty}^{\frac{\log t - \mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}u^2\right\} du = \Phi\left(\frac{\log t - \mu}{\sigma}\right), \quad t \geq 0, \tag{1.16}$$

en donde $\Phi(\cdot)$ es la función de distribución normal.

Se puede ver también que

$$E(T) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad \mathbf{y}$$

$$Var(T) = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right).$$

De (1.16) se tiene la $f.s.$,

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \quad t \geq 0.$$

Por tanto su función de riesgo es

$$h(t) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right)}{t\sigma\sqrt{2\pi}\left(1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)\right)}.$$

Pueden verse las formas que pueden tomar: la $f.d.p.$, la $f.d.d.$, la $f.s.$ y la $f.r.$, de la distribución log-Normal con distintos parámetros en la Figura 1.12.

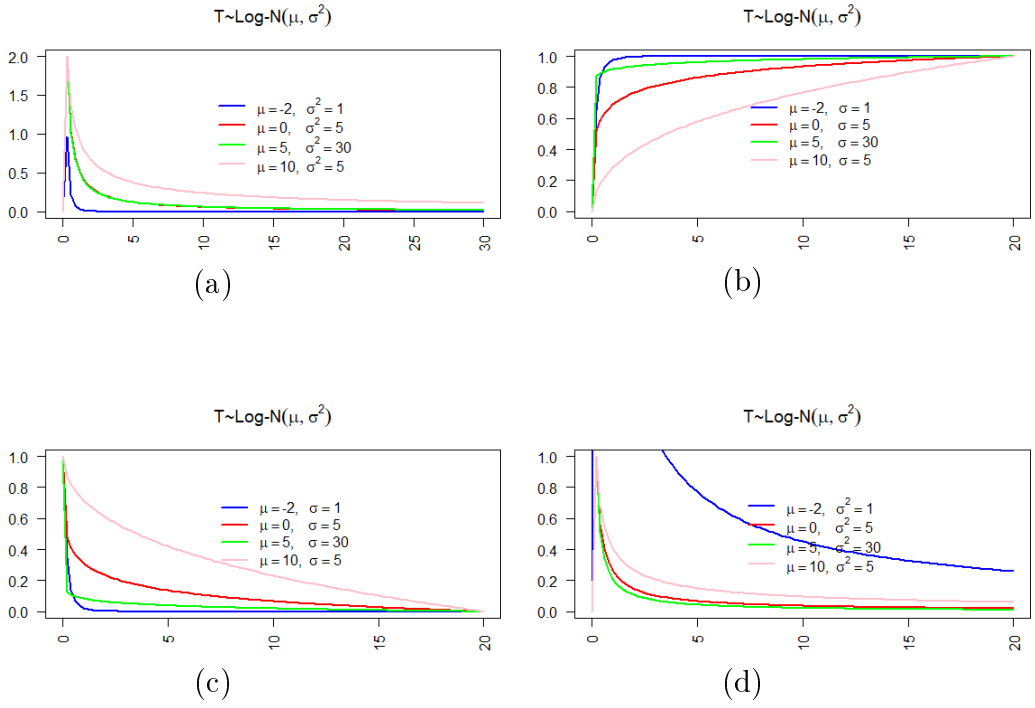


Figura 1.12: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo de la distribución log-Normal.

1.3.5. Distribución Gamma Generalizada

El modelo de Gamma Generalizada (GG) fue propuesto por Stacy en 1962, este modelo se caracteriza porque incluye varios modelos como casos particulares, algunos de ellos son la distribución Exponencial, Weibull y Gamma.

Definición 1.3.6. Sea T una variable aleatoria continua, se dice que T tiene una distribución Gamma Generalizada ($T \sim GG(\alpha, k, \theta)$) si su función de densidad está dada por,

$$f(t) = \frac{\theta}{\alpha\Gamma(k)} \left(\frac{t}{\alpha}\right)^{\theta k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}, \quad t \geq 0. \tag{1.17}$$

con parámetros de forma $\alpha > 0$, $k > 0$ y parámetro de escala $\theta > 0$. Además, la *f.d.d.* es,

$$\begin{aligned} F(t) &= \frac{1}{\Gamma(k)} \int_0^t \frac{\theta}{\alpha} \left(\frac{x}{\alpha}\right)^{\theta k-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^{\theta}\right\} dx \\ &= \frac{1}{\Gamma(k)} \int_0^{\left(\frac{t}{\alpha}\right)^{\theta}} u^{k-1} \exp(-u) du \\ &= \frac{1}{\Gamma(k)} \gamma\left(k, \left(\frac{t}{\alpha}\right)^{\theta}\right). \end{aligned}$$

Con esto se puede probar que

$$E(T) = \frac{\alpha\Gamma(k + \frac{1}{\theta})}{\Gamma(k)}$$

y

$$Var(T) = \frac{\alpha^2}{\Gamma^2(k)} \left\{\Gamma(k)\Gamma\left(k + \frac{2}{\theta}\right) - \Gamma^2\left(k + \frac{1}{\theta}\right)\right\}.$$

Con función de supervivencia

$$S(t) = 1 - \frac{1}{\Gamma(k)} \gamma\left(k, \left(\frac{t}{\alpha}\right)^{\theta}\right), \quad t \geq 0, \tag{1.18}$$

y función de riesgo,

$$h(t) = \frac{\theta}{\alpha} \left(\frac{t}{\alpha}\right)^{\theta k-1} \frac{\exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}}{\Gamma(k) - \gamma\left(k, \left(\frac{t}{\alpha}\right)^{\theta}\right)}, \quad t \geq 0.$$

Observación 1.4. La distribución Gamma Generalizada es una subfamilia de distribuciones que contiene a la distribución Exponencial cuando $k = \theta = 1$, a la distribución Weibull cuando sólo $k = 1$ y a la distribución Gamma si $\theta = 1$, más aún, en el límite, cuando $k \to \infty$ se puede ver que la distribución log-Normal también pertenece a esta subfamilia. Así, en las descripciones anteriores se tienen ejemplos de su comportamiento.

1.4.1. Distribución log-Logístico

La distribución log-logístico se usa cuando el logaritmo de la variable está distribuido logísticamente. Es común que esta distribución se use en modelos de crecimiento y para modelar respuestas binarias en campos como la Bioestadística y la Economía.

Definición 1.4.1. Diremos que una *v.a.* continua T sigue una distribución log-Logístico con parámetros λ y k positivos ($T \sim LLogist(\lambda, k)$) si su función densidad es,

$$f(t) = \frac{\lambda k(\lambda t)^{k-1}}{(1+(\lambda t)^k)^2}, \quad t \geq 0,$$

y su función de distribución,

$$F(t) = \frac{(\lambda t)^k}{1+(\lambda t)^k}, \quad t \geq 0.$$

Se puede probar que,

$$E(T) = \lambda.$$

Por otro lado, las funciones de supervivencia y de riesgo son

$$S(t) = \frac{1}{1+(\lambda t)^k}, \quad t \geq 0,$$

$$h(t) = \frac{\lambda k(\lambda t)^{k-1}}{1+(\lambda t)^k}, \quad t \geq 0.$$

En la Figura 1.13, pueden verse las posibles formas que pueden tomar estas funciones.

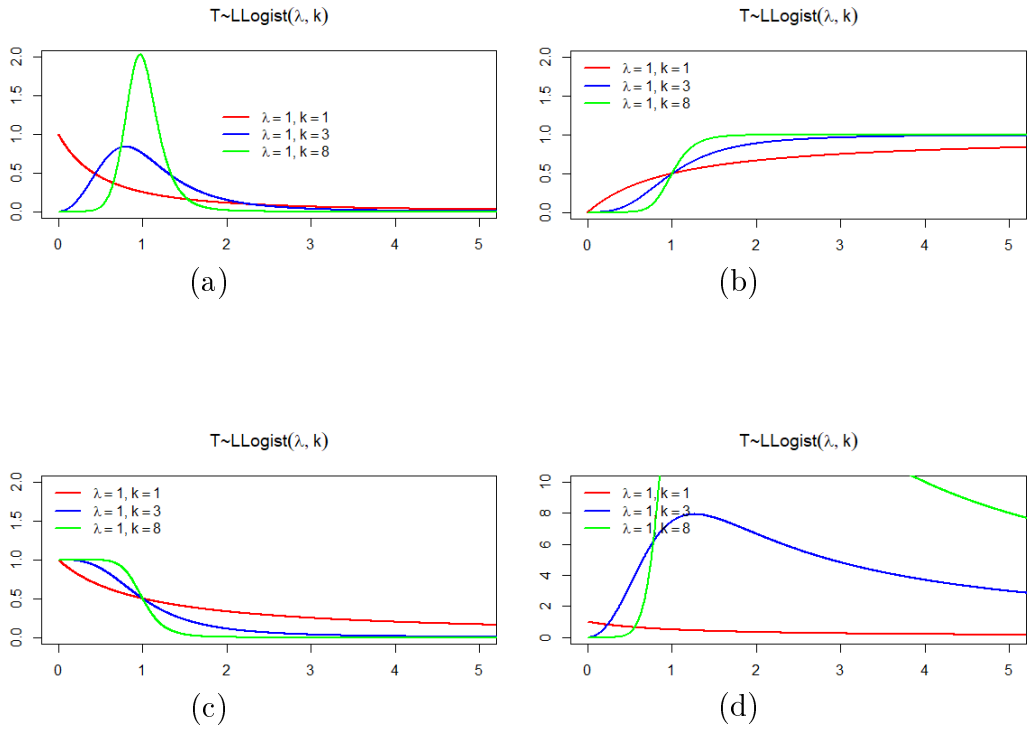


Figura 1.13: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo, de un modelo *log*-logístico.

1.4.2. Distribución Gompertz

Benjamin Gompertz (1825) introdujo la distribución de Gompertz en relación con la mortalidad. Esta distribución es una generalización de la distribución Exponencial y tiene muchas aplicaciones de la vida real, especialmente en estudios médicos. Otra característica importante de la distribución de Gompertz es que tiene una función de riesgo que aumenta exponencialmente durante la vida útil de los sistemas, por lo cual se usa para modelar datos altamente sesgados negativamente en el Análisis de Supervivencia.

Definición 1.4.2. Diremos que T es una variable aleatoria que tiene una distribución de Gompertz con parámetros c y λ positivos $(GM(c,\lambda))$ si su función de densidad es:

$$f(t) = \lambda \exp(ct) \exp\left(-\frac{\lambda}{c}(\exp(cx) - 1)\right), \quad t > 0 \tag{1.19}$$

y su correspondiente función de distribución es:

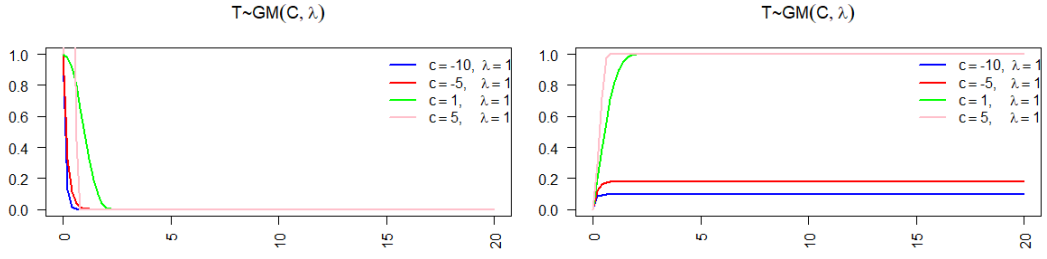
$$F(t) = 1 - \exp\left(-\frac{\lambda}{c}(\exp(ct) - 1)\right), \quad t > 0, c, \lambda > 0.$$

De donde claramente,

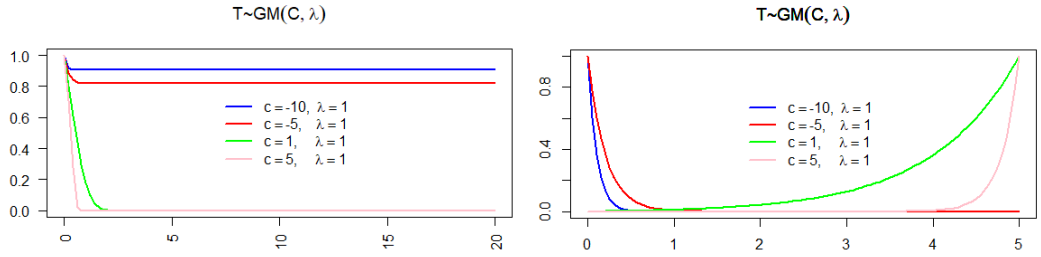
$$S(t) = \exp\left(-\frac{\lambda}{c}(\exp(ct) - 1)\right), \quad t > 0,$$

$$h(t) = \lambda \exp(ct), \quad t > 0,$$

$h(t)$ es creciente para $c > 0$ y decreciente para $c < 0$ y cuando $c = 0$ en (1.19) se obtiene la $f.d.p.$ de una variable con distribución Exponencial de parámetro λ. En la Figura 1.14 puede verse el comportamiento que toman las diferentes funciones con distintos parámetros.



(a) (b)



(c) (d)

Figura 1.14: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo de la distribución de Gompertz.

1.4.3. Distribución de Gumbel

Esta distribución es conocida como una distribución de valores extremos, el cual es un modelo que está enfocado en describir el comportamiento estadístico de $M_n = Max\{X_1, X_2, \ldots, X_n\}$, donde X_i es una *v.a. i.i.d.*.

Definición 1.4.3. Diremos que una *v.a.* continua y positiva T sigue una distribución de Gumbel con parámetros $\alpha > 0$ (de localización) y $\beta \in \mathbb{R}$ (escala) ($T \sim Gumbel(\alpha, \beta)$) si su función de densidad es dada por

$$f(t) = \alpha \exp\Big\{-\alpha(t-\beta) - \exp\{-\alpha(t-\beta)\}\Big\}, \quad t \in \mathbb{R},$$

con función de distribución

$$F(t) = \exp\Big\{-\exp\{-\alpha(t-\beta)\}\Big\}, \quad t \in \mathbb{R},$$

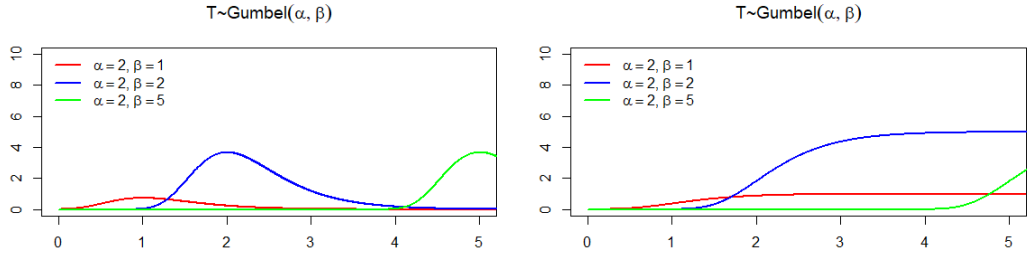
de donde

$$S(t) = 1 - \exp\Big\{-\exp\{-\alpha(t-\beta)\}\Big\}, \quad t \in \mathbb{R},$$

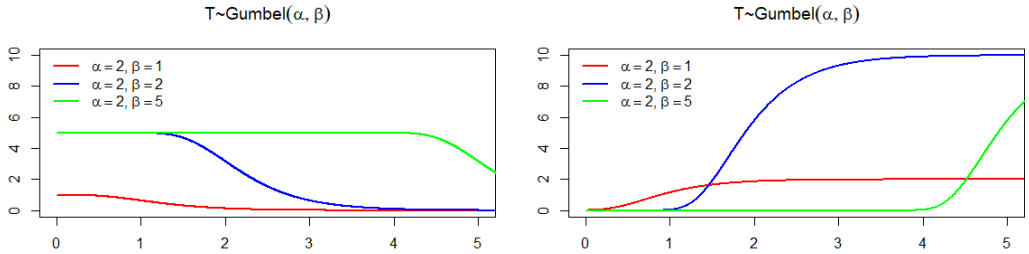
y

$$h(t) = \frac{\alpha \exp\Big\{-\alpha(t-\beta) - \exp\{-\alpha(t-\beta)\}\Big\}}{1 - \exp\Big\{-\exp\{-\alpha(t-\beta)\}\Big\}}, \quad t \in \mathbb{R}.$$

La Figura 1.15, se muestra algunas posibles formas de las funciones anteriores.



(a) (b)



(c) (d)

Figura 1.15: (a) Función de densidad; (b) de distribución; (c) de supervivencia y (d) de riesgo de la distribución de Gumbel.

En la Tabla 1.3 puede ver cada una las funciones de probabilidad para cada uno de los modelos.

Distribución	Notación $T\sim$	*f.d.p.* $f(t)$	*f.d.* $F(t)$	*f.s.* $S(t)$	*f.r.* $h(t)$	Media $E(T)$	Varianza $Var(T)$
Exponencial	$\exp(\lambda)$	$\lambda\exp(-\lambda t)$	$1-\exp(-\lambda t)$	$\exp(-\lambda t)$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$G(k,\lambda)$	$\frac{\lambda(\lambda t)^{k-1}\exp(-\lambda t)}{\Gamma(k)}$	$\frac{\gamma(k,\lambda t)}{\Gamma(k)}$	$1-\frac{\gamma(k,\lambda t)}{\Gamma(k)}$	$\frac{\lambda(\lambda t)^{k-1}\exp(-\lambda t)}{\Gamma(k)-\gamma(k,\lambda t)}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$
Weibull	$W(\alpha,\beta)$	$\alpha\beta(\alpha t)^{\beta-1}\exp(-(\alpha t)^{\beta})$	$1-\exp(-(\alpha t)^{\beta})$	$\exp(-(\alpha t)^{\beta})$	$\alpha\beta(\alpha t)^{\beta-1}$	$\frac{\Gamma(1+\frac{1}{\beta})}{\alpha}$	$\frac{\Gamma(1+\frac{2}{\beta})-\Gamma^2(1+\frac{1}{\beta})}{\alpha^2}$
Log-Normal	$log\text{-}N(\mu,\sigma^2)$	$\frac{1}{t\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right\}$	$\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{\exp\left(-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right)}{t\sigma\sqrt{2\pi}\left(1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right)}$	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$	$\exp(2\mu+\sigma^2)\left(\exp(\sigma^2)-1\right)$
Gamma generalizada	$GG(\alpha,k,\theta)$	$\frac{\theta}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\theta k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}$	$\gamma\left(k,\left(\frac{t}{\alpha}\right)^{\theta}\right)$	$1-\gamma\left(k,\left(\frac{t}{\alpha}\right)^{\theta}\right)$	$\frac{\theta\left(\frac{t}{\alpha}\right)^{\theta k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}}{\alpha\Gamma(k)\left[1-\gamma\left(k,\left(\frac{t}{\alpha}\right)^{\theta}\right)\right]}$	$\frac{\alpha\Gamma(k+\frac{1}{\theta})}{\Gamma(k)}$	$\frac{\alpha^2}{\Gamma^2(k)}\left[\Gamma(k)\Gamma\left(k+\frac{2}{\theta}\right)-\Gamma^2\left(k+\frac{1}{\theta}\right)\right]$
Log-Logístico	$LLogist(\lambda,k)$	$\frac{\lambda k(\lambda t)^{k-1}}{(1+(\lambda t)^k)^2}$	$\frac{(\lambda t)^k}{1+(\lambda t)^k}$	$\frac{1}{1+(\lambda t)^k}$	$\frac{\lambda k(\lambda t)^{k-1}}{1+(\lambda t)^k}$	λ	
Gompertz	$T\sim GM(c,\lambda)$	$\lambda\exp(ct)\exp\left(-\frac{\lambda}{c}(\exp(ct)-1)\right)$	$1-\exp\left(-\frac{\lambda}{c}(\exp(ct)-1)\right)$	$\exp\left(-\frac{\lambda}{c}(\exp(ct)-1)\right)$	$\lambda\exp(ct)$		
Gumbel	$T\sim Gumbel(\alpha,\beta)$	$\alpha\exp\Big\{-\alpha(t-\beta)-\exp\{-\alpha(t-\beta)\}\Big\}$	$\exp\Big\{-\exp\{-\alpha(t-\beta)\}\Big\}$	$1-\exp\Big\{-\exp\{-\alpha(t-\beta)\}\Big\}$	$\frac{\alpha\exp\Big\{-\alpha(t-\beta)-\exp\{-\alpha(t-\beta)\}\Big\}}{1-\exp\Big\{-\exp\{-\alpha(t-\beta)\}\Big\}}$		

Tabla 1.3: *f.d.p.*, *f.d.d.*, *f.s.*, *f.r.*, media $E(T)$ y varianza $Var(T)$ de la variable aleatoria T.

1.5. Método de máxima verosimilitud para datos censurados

En el caso de datos de supervivencia con datos censurados, la función de verosimilitud juega un papel muy importante al permitirnos hacer inferencia acerca del valor del parámetro, por lo cual es necesario llegar a conocer la expresión para esta función bajo censura. En este caso, para se habla de censura por la derecha.

1.5.1. Datos con censura tipo I

Considere un modelo de datos de supervivencia con censura por la derecha tipo I. Sea $T_1, T_2, \ldots T_n$ una muestra aleatoria independientes e idénticamente distribuidos (*i.i.d.*), con función de densidad $f(\cdot)$ y $C_1, C_2, \ldots C_n$ constantes de censura asociadas a estas *v.a.'s* respectivamente. Se tienen las observaciones:

$$(T_1^*, \delta_1), (T_2^*, \delta_2), \ldots, (T_{n-1}^*, \delta_{n-1}), (T_n^*, \delta_n),$$

donde

$$T_i^* = \text{mín}\{T_i, C_i\} \qquad y \qquad \delta_i = I(T_i \leq C_i). \tag{1.20}$$

Considere $t_1, t_2, \ldots, t_n$ una realización con cesura presente de la muestra $T_1^*, T_2^*, \ldots, T_n^*$. El siguiente resultado nos muestra la expresión para la función de verosimilitud para datos con censura por la derecha tipo I.

Teorema 1.5.1. *Bajo censura tipo I, con tiempos de censura fijos, la función de verosimilitud $L(\theta, \underline{t})$ de los datos observados (t_i, δ_i), $i = 1, 2, \ldots, n$ está dada por:*

$$L(\theta, \underline{t}) = \prod_{i=1}^{n} f(t_i, \theta)^{\delta_i} S(t_i, \theta)^{1-\delta_i}.$$

Demostración: Analicemos los casos que se presentan para la unidad muestral (T_i^*, δ_i), para obtener la función de probabilidad conjunta de dicha muestra.
Caso 1: Si un individuo tiene un tiempo de supervivencia censurada, es decir, $\delta_i = 0$ en el tiempo C_i (aquí las C_i son constantes fijas) entonces su tiempo de supervivencia es mayor que C $(T_i > C_i)$, así cuando $\Delta t \to 0^+$ tenemos:

$$\begin{aligned} P(t \leq T_i^* < t + \Delta t, \delta_i = 0) &= P(t \leq T_i^* < t + \Delta t | \delta_i = 0) P(\delta_i = 0) \\ &= P(t \leq \text{mín}\{T_i, C_i\} < t + \Delta t | \delta_i = 0) P(\delta_i = 0) \\ &= P(t \leq C_i < t + \Delta t) P(T_i > C_i), \end{aligned}$$

de donde, si $t < C_i$ entonces,

$$\lim_{\Delta t \to 0^+} P(t \leq C_i < t + \Delta t) = P(C_i = t) = 0,$$

y si $t = C_i$,

$$\lim_{\Delta t \to 0^+} P(t \leq C_i < t + \Delta t) = P(C_i = t) = 1,$$

al tomar el límite se obtiene:

$$\lim_{\Delta t \to 0^+} P(t \leq T_i^* < t + \Delta t, \delta_i = 0) = \begin{cases} P(C_i < T_i), & \text{si } t = C_i, \\ 0, & \text{si } t < C_i. \end{cases} \tag{1.21}$$

Caso 2: $\delta_i = 1$, es decir,cuando no hay censura se tiene lo siguiente:

$$\begin{aligned} P(t \leq T_i^* < t + \Delta t, \delta_i = 1) &= P(t \leq T_i^* < t + \Delta t | \delta_i = 1) P(\delta_i = 1) \\ &= P(t \leq T_i^* < t + \Delta t | T_i \leq C_i) P(T_i \leq C_i) \\ &= \frac{P(t \leq T_i < t + \Delta t)}{F(C_i)} F(C_i), \end{aligned}$$

luego,

$$\lim_{\Delta \to 0^+} P(t \leq T_i^* < t + \Delta t, \delta_i = 1) = \lim_{\Delta \to 0^+} P(t \leq T_i < t + \Delta t) = f(t). \tag{1.22}$$

Por tanto, de (1.21) y (1.22) la función de densidad conjunta de (T_i^*, δ_i) es:

$$f_{(T_i^*, \delta_i)}(t, \delta_i) = \begin{cases} f(t)^{\delta} S(t)^{1-\delta}, & \text{si } \{(t,1) : t \leq C_i\} \cup \{(C_i, 0)\}, \\ 0, & \text{si } \{(t,0) : t < C_i\}, \end{cases}$$

y finalmente la función de verosimilitud para la muestra dada es

$$L(\theta, \underline{t}) = \prod_{i=1}^{n} f(t_i; \theta)^{\delta_i} S(t_i; \theta)^{1-\delta_i}.$$

■

1.5.2. Datos con censura tipo II

Sea $T_1, \ldots, T_n$ una muestra aleatoria *i.i.d.* en el cual deben ocurrir $k \leq n$ fallas, esto es, se observan los $T_{(1)}, \ldots, T_{(k)}$ estadísticos de orden, de modo que los $n - k$ individuos restantes son censurados y $T_{(k)}$ es su tiempo de censura (recordemos que en este caso el tiempo de censura es controlado por el investigador). En este caso la función de verosimilitud está dada por la función de densidad conjunta de los estadísticos de orden, de modo que si $t_1, t_2, \ldots, t_k$ es una realización de esta muestra, entonces se tiene que

$$\begin{aligned}
L(\theta, \underline{t}) &= f_{T_{(1)},\ldots,T_{(k)}}(t_1, t_2, \ldots, t_k) \\
&= \frac{n!}{(n-k)!}\left(\prod_{i=1}^{k} f(t_i;\theta)\right)(1 - F(t_k;\theta))^{n-k} \\
&= \frac{n!}{(n-k)!}\left(\prod_{i=1}^{k} f(t_i;\theta)\right) S(t_k;\theta)^{n-k} \\
&= \frac{n!}{(n-k)!}\left(\prod_{i=1}^{k} f(t_i;\theta)\right) S(t_k;\theta)^{n-\sum_{l=1}^{k}\delta_i} \\
&= \frac{n!}{(n-k)!}\left(\prod_{i=1}^{n} f(t_i;\theta)^{\delta_i}\right)\left(\prod_{i=1}^{n} S(t_k;\theta)^{1-\delta_i}\right) \\
&= \frac{n!}{(n-k)!}\prod_{i=1}^{n}\Big(f(t_i;\theta)^{\delta_i} S(t_k;\theta)^{1-\delta_i}\Big),
\end{aligned}$$

en donde $\delta_1 = \delta_2 = \cdots = \delta_k = 1$ y $\delta_{k+1} = \delta_{k+2} = \cdots = \delta_n = 0$.

1.5.3. Caso general

Considere $T_1, \ldots, T_n$ una muestra aleatoria *i.i.d.* con función de densidad $f(\cdot)$, función de distribución $F(\cdot)$ y función de supervivencia $S(\cdot)$. Sea

$$T_i^* = \text{mín}\{T_i, C_i\},$$

y

$$\delta_i = \begin{cases} 1, & \text{si } T_i^* = T_i, \\ 0, & \text{si } T_i^* = C_i. \end{cases}$$

Suponga además que $C_1, C_2, \ldots, C_n$ son los tiempos de censura (aleatorios) asociadas a las variables T_i, $i = 1, \ldots, n$ respectivamente, con función de densidad $g(\cdot)$ y *f.d.d.* $G(\cdot)$. Bajo mecanismo de censura por la derecha se asume que T_i y C_i son independientes y que además $G(\cdot)$ no depende de ninguno de los parámetros de $S(\cdot)$.

Se busca encontrar la función de densidad conjunta de la unidad muestral (T_i^*, δ_i), buscando obtener la función de verosimilitud para una muestra de tamaño n.

Considere también el caso en que la observación presenta censura ($T_i > C_i$). Denotamos por $H_1(\cdot)$ a la función de distribución para dicha observación con censura en el tiempo $C_i = u$, es decir:

$$\begin{aligned} H_1(t) = P(T_i^* \leq t, \delta_i = 0) &= P(\text{mín}\{T_i, C_i\} \leq t, C_i < T_i) \\ &= P(C_i \leq t, C_i < T_i) \\ &= P(C_i \leq t, u < T_i) \\ &= \int_0^t \left[\int_u^\infty f(v)dv\right] g(u)du \\ &= \int_0^t [1 - F(u)]\, g(u)du, \end{aligned}$$

derivando $H_1(\cdot)$ obtenemos que:

$$h_1(t) = \frac{dH_1}{dt} = [1 - F(t)]g(t).$$

Ahora, considere el caso en el que la observación es no censurada, así, para cuando es observado en el tiempo $T_i = s$ se obtiene que $T_i < C_i$ y

$$\begin{aligned} H_2(t) = P(T_i^* \leq t, \delta_i = 1) &= P(\text{mín}\{T_i, C_i\} \leq t, \delta_i = 1) \\ &= P(T_i \leq t, s = T_i \leq C_i) \\ &= \int_0^t \left[\int_s^\infty g(v)dv\right] f(s)ds \\ &= \int_0^t [1 - G(s)]\, f(s)ds, \end{aligned}$$

derivando $H_2(\cdot)$ se tiene que:

$$h_2(t) = \frac{dH_2}{dt} = [1 - G(t)]f(t),$$

luego,

$$\begin{aligned} f_{(T_i^*, \delta_i)}(t, \delta_i) &= \Big[(1 - F(t))g(t)\Big]^{1-\delta_i} \Big[(1 - G(t))f(t)\Big]^{\delta_i} \\ &= \Big[(1 - F(t))^{1-\delta_i} f(t)^{\delta_i}\Big]\Big[(1 - G(t))^{\delta_i} g(t)^{1-\delta_i}\Big]. \end{aligned}$$

Finalmente, considerando la realización $t_1, t_2, \ldots, t_n$ de la muestra $T_1^*, T_2^*, \ldots, T_n^*$, se obtiene la función de verosimilitud:

$$L(\theta, \underline{t}) = \left(\prod_{i=1}^n f(t_i)^{\delta_i}(1 - F(t_i))^{1-\delta_i}\right)\left(\prod_{i=1}^n g(t_i)^{1-\delta_i}(1 - G(t_i))^{\delta_i}\right). \quad (1.23)$$

Sin embargo, dado que $g(\cdot)$ y $G(\cdot)$ son funciones que no involucran a los parámetros de $f(\cdot)$, podemos considerar a la función de verosimilitud como:

$$L(\theta, \underline{t}) = \prod_{i=1}^n f(t_i)^{\delta_i}(1 - F(t_i))^{1-\delta_i}.$$

1.5.4. Datos con censura aleatoria

Considere $T_1, \ldots, T_n$ una muestra aleatoria *i.i.d.* con función de densidad $f(\cdot)$, función de distribución $F(\cdot)$ y *f.s.* $S(\cdot)$. Considere la muestra:

$$(T_1^*, \delta_1), (T_2^*, \delta_2), \ldots, (T_n^*, \delta_n).$$

Denotamos a C_i como los tiempos de censura (aleatorios) asociadas a las variables T_i, $i = 1, \ldots, n$ respectivamente, con función de densidad $g(\cdot)$ y *f.d.* $G(\cdot)$. Bajo mecanismo de censura aleatoria (no informativa) se asume que T_i y C_i son independientes y que además $G(\cdot)$ no depende de ninguno de los parámetros de la función de supervivencia.

Teorema 1.5.2. *En un modelo con censura aleatoria por la derecha, la función de verosimilitud, de los datos observados*

$$(t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n).$$

con t_i, una realización de T_i^*, $i = 1, 2, \ldots, n$, está dado por:

$$L(\theta, \underline{t}) = \left(\prod_{i=1}^{n} f(t_i)^{\delta_i} (1 - F(t_i))^{1-\delta_i} \right) \left(\prod_{i=1}^{n} g(t_i)^{1-\delta_i} (1 - G(t_i))^{\delta_i} \right).$$

Observación 1.6. La prueba del Teorema 1.5.2 se sigue de manera análoga al caso general por lo cual se omite dicha prueba.

Con los resultados obtenidos anteriormente bajo un mecanismo de censura por la derecha se muestran a continuación la función de verosimilitud para el modelo Exponencial y Weibull.

Ejemplo 1.7. Sea $T_1, \ldots, T_n$ una muestra aleatoria *i.i.d.*, con distribución Exponencial con parámetro $\lambda > 0$ y sea $t_1, \ldots, t_n$ una realización que puede presentar censura por la derecha de la muestra $T_1^*, \ldots, T_n^*$. Para las observaciones siguientes:

$$(T_1^*, \delta_1), \ldots, (T_{n-1}^*, \delta_{n-1}), (T_n^*, \delta_n),$$

la función de verosimilitud está dada por,

$$\begin{aligned} L(\lambda, \underline{t}) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} &= \prod_{i=1}^{n} \Big[\lambda \exp(-\lambda t_i) \Big]^{\delta_i} \Big[\exp(-\lambda t_i) \Big]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda t_i) \\ &= \lambda^{\sum_{i=1}^{n} \delta_i} \exp\left(-\lambda \sum_{i=1}^{n} t_i \right), \end{aligned}$$

luego, la función log-verosimilitud es,

$$l(\lambda) = \log L(\lambda, \underline{t}) = \sum_{i=1}^{n} \delta_i \log(\lambda) - \lambda \sum_{i=1}^{n} t_i.$$

Observe que $r = \sum_{i=1}^{n} \delta_i$ es el número de sujetos no censurados, mientras que $\sum_{i=1}^{n} t_i$ es la suma de todos los tiempos de la realización dada.

Buscamos el *e.m.v.*, así, calculando

$$\frac{dl(\lambda)}{d\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i,$$

igualando a cero, obtenemos el único punto crítico en:

$$\lambda = \frac{r}{\sum_{i=1}^{n} t_i},$$

y como

$$\frac{d^2 l(\lambda)}{d\lambda^2} = -\frac{r}{\lambda^2},$$

se tiene que

$$\frac{d^2 l(\lambda)}{d\lambda^2} = -\frac{r}{\left(\frac{r}{\sum_{i=1}^{n} t_i}\right)} < 0,$$

así, el *e.m.v.* para λ está dado por $\widehat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i}$.

Ejemplo 1.8. Sea $T_1, \ldots, T_n$ una muestra aleatoria *i.i.d.*, con distribución Weibull de parámetros α y β. Sea $t_1, \ldots, t_n$ una realización de la muestra $T_1^*, \ldots, T_n^*$ con censura presente. Considere las observaciones,

$$(T_1^*, \delta_1), \ldots, (T_{n-1}^*, \delta_{n-1}), (T_n^*, \delta_n),$$

de (1.12) y (1.14) la función de verosimilitud está dada por:

$$\begin{aligned}
L(\alpha, \beta, \underline{t}) &= \prod_{i=1}^{n} \Big(\alpha\beta(\alpha t_i)^{\beta-1} \exp\left(-(\alpha t_i)^{\beta}\right)\Big)^{\delta_i} \Big(\exp\left(-(\alpha t_i)^{\beta}\right)\Big)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \Big(\alpha\beta(\alpha t_i)^{\beta-1}\Big)^{\delta_i} \exp\left(-(\alpha t_i)^{\beta}\right) \\
&= (\alpha\beta)^{\sum_{i=1}^{n}\delta_i} \alpha^{(\beta-1)\sum_{i=1}^{n}\delta_i} \left(\prod_{i=1}^{n} t_i^{(\beta-1)\delta_i}\right) \exp\left(-\sum_{i=1}^{n}(\alpha t_i)^{\beta}\right) \\
&= \alpha^{\beta\sum_{i=1}^{n}\delta_i} \beta^{\sum_{i=1}^{n}\delta_i} \left(\prod_{i=1}^{n} t_i^{(\beta-1)\delta_i}\right) \exp\left(-\sum_{i=1}^{n}(\alpha t_i)^{\beta}\right) \\
&= \alpha^{\beta r} \beta^{r} \left(\prod_{i=1}^{n} t_i^{(\beta-1)\delta_i}\right) \exp\left(-\sum_{i=1}^{n}(\alpha t_i)^{\beta}\right),
\end{aligned}$$

en donde $r = \sum_{i=1}^{n} \delta_i$ es el número de individuos no censurados. Luego, la función log-verosimilitud es:

$$l(\alpha, \beta) = L(\alpha, \beta, \underline{t}) = r\log(\beta) + r\beta\log(\alpha) + (\beta - 1)\sum_{i=1}^{n} \delta_i \log(t_i) - \sum_{i=1}^{n}(\alpha t_i)^{\beta}.$$

Calculando:

$$\frac{\partial l(\alpha,\beta)}{\partial \alpha} = \frac{r\beta}{\alpha} - \beta\alpha^{\beta-1}\sum_{i=1}^{n} t_i^{\beta}, \quad \text{y} \tag{1.24}$$

$$\frac{\partial l(\alpha,\beta)}{\partial \beta} = \frac{r}{\beta} + r\log(\alpha) + \sum_{i=1}^{n}\delta_i \log(t_i) - \sum_{i=1}^{n}(\alpha t_i)^{\beta}\log(\alpha t_i). \tag{1.25}$$

Así, para calcular el *e.m.v.*, fijando β en (1.24) e igualando a cero se obtiene que

$$\widehat{\alpha} = \left(\frac{r}{\sum_{i=1}^{n} t_i^{\beta}}\right)^{\frac{1}{\beta}}.$$

Sin embargo de (1.25), nótese que obtener el valor de β no es tan sencillo, por lo cual se requiere de algún método numérico para obtener dicho valor y poder verificar que en efecto es el estimador de máxima verosimilitud. Para ver una forma de obtener los estimadores puntuales adecuados para este modelo puede ver [10].

Capítulo 2

ESTIMACIÓN NO PARAMÉTRICA

Cuando se desean estudiar ciertas características de una muestra aleatoria proveniente de una población, la selección del modelo de inferencia (paramétrico o semi-paramétrico) es muy importante en la práctica ya que una mala elección puede conducir a obtener resultados erróneos.

Comúnmente, para la selección del modelo de inferencia se toma en cuenta la distribución de los datos, lo cual conduce a verificar que en efecto la muestra cumpla con supuestos, las cuales no siempre se llegan a verificar. Así, correctamente optaríamos por un modelo paramétrico si estamos seguros del comportamiento de los datos, es decir, si el modelo se ajusta correctamente al conjunto de datos, sin embargo, aún cuando se puede verificar la bondad de ajuste del modelo, siempre existiría un grado de incertidumbre. Todo esto motiva a hacer uso de métodos de inferencia no paramétrica, los cuales no hacen ningún supuesto acerca de la forma funcional de la función de supervivencia, permitiendo que cada observación en estudio contribuya a la construcción de su propio modelo. Entre las principales ventajas de la estimación no paramétrica se encuentran:

i. primera aproximación real de la curva de supervivencia,

ii. los intervalos de confianza permiten contrastar la validez de uno o más modelos teóricos y comparar subpoblaciones,

iii. paso intermedio para proponer un modelo funcional.

A pesar de las ventajas que presenta, también se puede ver que algunas de sus desventajas son:

i. no existe una relación funcional de la curva de supervivencia,

ii. sólo es posible estimar la forma de la función de supervivencia dentro del periodo de observación, es decir, hay imposibilidad de extrapolar.

Por tanto, en este capítulo se aborda el tema de métodos de inferencia no paramétrica que proporcionan un estimador para la función de supervivencia de una muestra aleatoria *i.i.d.* de tiempos de falla con censura por la derecha.

2.1. Distribución empírica

Para estimar la función de supervivencia de un conjunto de datos, el método más simple es el uso de la función de supervivencia empírica, el cual se define acontinuación [2].

Definición 2.1.1. Sea $T_1, T_2, \ldots, T_n$ una muestra aleatoria, considere $t_1, t_2, \ldots,$ t_n una realización de tal muestra, tales que ninguno de ellos presenta censura (no se hace alguna suposición acerca de su distribución). La función de supervivencia empírica (*f.s.e.*) se define como

$$\begin{aligned} \widehat{S}(t) &= \frac{\#\{i : t_i > t\}}{n} \\ &= \frac{\{\text{No. de individuos que sobreviven al instante t}\}}{n}. \end{aligned} \tag{2.1}$$

Algunas de las características de esta función son:

1. es una función escalonada decreciente,
2. cuando todas las observaciones son distintas los saltos son de $1/n$,
3. si hay d empates, hay saltos de tamaño d/n,
4. si $t_{(1)}$ es la observación mas pequeña, entonces $S(t) = 1$, para $t < t_{(1)}$,
5. si $t_{(n)}$ es la observación más grande, entonces $S(t) = 0$, para $t > t(n)$.

Sin embargo, nótese que esta función en realidad desperdicia información cuando hay sujetos censurados, ya que no toma en cuenta si la observación t_i tal que $t_i \leq t$ fue falla o censura.

2.1.1. Distribución empírica como estimador de máxima verosimilitud de una función de distribución

Una característica importante de la función de supervivencia empírica es que $\widehat{S}(\cdot)$ definida en (2.1) es el estimador de máxima verosimilitud de la función de supervivencia $S(\cdot)$, esto se puede probar por medio de la función de distribución empírica, ya que esta es el *e.m.v.* de una función de distribución, además de que puede ver fácilmente que $\widehat{S}(t) = 1 - \widehat{F}(t)$, donde $\widehat{F}(\cdot)$ es la función de distribución empírica que se define a continuación.

Definición 2.1.2. Sea $T_1, T_2, \ldots, T_n$ una muestra aleatoria de una función de distribución $F \in \mathcal{F} = \{\text{Clase de todas las distribuciones sobre } \mathbb{R}\}$, considere $t_1, t_2, \ldots, t_n$, donde $t_i \neq t_j$ para $i \neq j$ una realización de tal muestra. La función de distribución empírica (*f.d.e.*) se define como:

$$\widehat{F}(t) = \frac{\#\{i : t_i \leq t\}}{n} = \frac{\sum_{i=1}^{n} I\{t_i \leq t\}}{n}. \tag{2.2}$$

Si para esta realización buscamos la función de distribución F que haga más factible dicha observación. Entonces F maximiza el valor de

$$P(t_1 < T_1 \leq t_1 + \Delta t_1, t_2 < T_2 \leq t_2 + \Delta t_2, \ldots, t_n < T_n \leq t_n + \Delta t_n) =$$
$$P_F(t_1 < T_1 \leq t_1 + \Delta t_1) P_F(t_2 < T_2 \leq t_2 + \Delta t_2) \ldots P_F(t_n < T_n \leq t_n + \Delta t_n),$$

además dado que F maximiza esta verosimilitud entonces debe de cumplir que

$$\sum_{i=1}^{n} \lim_{\Delta t_i \to 0} P_F(t_i < T_i \leq t_i + \Delta t_i) = 1,$$

o bien

$$f_F(x_1) + f_F(x_2) + \cdots + f_F(x_n) = 1,$$

donde f_F es la función de densidad, con respecto a F. De aquí que, el producto siguiente,

$$f_F(x_1) f_F(x_2) \cdots f_F(x_n) = \prod_{i=1}^{n} \lim_{\Delta t_i \to 0} P_F(t_i < T_i \leq t_i + \Delta t_i) = q_1 q_2 \cdots q_n$$

debe ser maximizado bajo la condición de que $\sum_{i=1}^{n} q_i = 1$. Así,

$$L(q_1, q_2, \ldots, q_n; t_1, t_2, \ldots, t_n) = q_1 q_2, \ldots, q_n = q_1 q_2, \ldots, q_{n-1} \left(1 - \sum_{i=1}^{n-1} q_i\right)$$

y la función log-verosimilitud,

$$\log(L(q_1, q_2, \ldots, q_n; t_1, t_2, \ldots, t_n)) = \sum_{i=1}^{n-1} \log q_i + \log\left(1 - \sum_{i=1}^{n-1} q_i\right),$$

calculando las derivadas parciales respecto a q_j con $j = 1, 2, \ldots, n-1$ e igualando a cero

$$\frac{\partial \log(L(q_1, q_2, \ldots, q_n; t_1, t_2, \ldots, t_n))}{\partial q_j} = \frac{1}{q_j} - \frac{1}{1 - \sum_{i=1}^{n-1} q_i} = 0$$

se obtiene que $q_j = q_n$ para $j = 1, 2, \ldots, n-1$, además de esto, $1 = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} q_n = nq_n$, es decir, $q_j = q_n = \frac{1}{n}$, $j = 1, 2, \ldots, n-1$.

Por otro lado, para cualquier valor de q_j se cumple

$$\frac{\partial^2 \log(L(q_1, q_2, \ldots, q_n; t_1, t_2, \ldots, t_n))}{\partial q_j^2} = -\frac{1}{q_j^2} - \frac{1}{\left(1 - \sum_{i=1}^{n-1} q_i\right)^2} < 0,$$

en particular para $q_j = \frac{1}{n}$, con lo cual podemos concluir que $\widehat{q}_1 = \widehat{q}_2 = \cdots = \widehat{q}_n = \frac{1}{n}$ es el estimador de máxima verosimilitud (e.m.v.) de F, la cual es la función de distribución empírica definida en (2.2). Más aún, con este resultado y el principio de invarianza para el *e.m.v.* se tiene que $1 - \widehat{F}(t) = \widehat{S}(t)$ es el *e.m.v.* de $S(t)$.

Observación 2.2. Si fijamos t, la variable aleatoria $I\{T_i \leq t\}$ tiene una distribución tipo Bernoulli (sólo toma el valor de 0 o 1) con parámetro $p = F(x)$, por lo cual $n\widehat{F}(t)$ es una*v.a.* binomial con parámetros n y $F(t)$, así $E(n\widehat{F}(t)) = nF(t)$, es decir $E(\widehat{F}(t)) = F(t)$, o bien, el estimador $\widehat{F}(t)$ es insesgado, más aún, podemos probar con la relación $S(t) = 1 - F(t)$, que $\widehat{S}(t)$ es también insesgado.

2.3. Estimador de Kaplan-Meier

Considere una muestra aleatoria de n individuos con tiempo de supervivencia observados en $t_1, t_2, \ldots, t_n$, algunos de ellos con censura por la derecha, además, puede haber más de un individuo con el mismo tiempo de supervivencia observado. Así, suponga que hay r $(r < n)$ individuos con tiempo de falla sin censura, de modo que los $n-r$ sujetos restantes presentan censura. Organizando estos tiempos de falla de forma creciente, sean $0 < t_{(1)} < \ldots < t_{(n)} < \infty$ esos distintos tiempos de falla observadas y ordenadas, con los cuales generamos los siguientes intervalos:

$$I_0 = [0, t_{(1)}),\ I_1 = [t_{(1)}, t_{(2)}), \ldots,\ I_{r-1} = [t_{(r-1)}, t_{(r)}), \ldots,\ I_r = [t_{(r)}, \infty).$$

Denotamos por:

1. n_j: número de individuos que se encuentran vivos justo antes del tiempo $t_{(j)}$, incluso los que están a punto de morir justo en ese momento. Nótese también que si hay l_j individuos que están censurados en el intervalo I_j al tiempo $t_{(j)1}, t_{(j)2}, \ldots, t_{(j)l_j}$, entonces sus tiempos de falla también están incluidos.

2. Para $j = 1, 2, \ldots, r$, d_j: es el número de personas que mueren en el momento $t_{(j)}$.

3. p_j: es la probabilidad condicional de sobrevivir a lo largo del intervalo I_j dado que se ha sobrevivido al inicio de este, es decir, $p_j = P(T > t_{(j)} | T > t_{(j-1)})$.

Buscamos el estimador para $S(\cdot)$ mediante los estimadores de p_j. De hecho, se probará más adelante que el estimador no paramétrico que se presenta aquí es el estimador de máxima verosimilitud de la función de supervivencia.
Sea $\delta > 0$ un tiempo infinitesimal y considere el j-ésimo intervalo I_j, nótese que en el intervalo $[t_{(j)} - \delta, t_{(j)}]$ se presenta un único tiempo de falla (el que ocurre en $t_{(j)}$), mientras que el intervalo $(t_{(j)}, t_{(j+1)} - \delta]$ no tiene ningún tiempo de falla, ya que esta se presenta hasta en el tiempo $t_{(j+1)}$, con esto, dado que n_j representa el número de individuos vivos justo antes del tiempo $t_{(j)}$ y d_j el número de sujetos muertos justo en $t_{(j)}$, la probabilidad de que un individuo presente falla en el intervalo $[t_{(j)} - \delta, t_{(j)}]$ puede ser estimado por d_j/n_j, por lo cual la probabilidad condicional de que el individuo en estudio sobreviva más allá de $t_{(j)}$ dado que ha sobrevivido hasta el tiempo $t_{(j)} - \delta$ es $\frac{n_j - d_j}{n_j}$, es decir,

$$P(T > t_{(j)} | T > t_{(j)} - \delta) = \frac{n_j - d_j}{n_j}$$

y la probabilidad de sobrevivir en el intervalo $(t_{(j)}, t_{(j+1)} - \delta]$ es 1. Luego, la probabilidad conjunta de sobrevivir a lo largo de $[t_{(j)} - \delta, t_{(j)}]$ y de $(t_{(j)}, t_{(j+1)} - \delta]$ puede ser estimado por $\frac{n_j - d_j}{n_j}$. Por lo tanto, al tomar el límite cuando δ tiende a infinito, se tiene que $\frac{n_j - d_j}{n_j} = \widehat{p}_j$ es un estimador de p_j, es decir, de la probabilidad condicional correspondiente al intervalo $I_j = [t_{(j)}, t_{(j+1)})$.

Con todo lo anterior, para t fijo tal que $t \in [t_{(i)}, t_{(i+1)})$ y $i \in \{1, 2, \ldots, r\}$, el producto de estos estimadores $\widehat{p}_j$ para $j = 1, 2, \ldots, i$ da una estimación de la función de supervivencia al tiempo t, esto es,

$$\widehat{S}(t) = \prod_{j=1}^{i} \widehat{p}_j = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j} = \prod_{j | t_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

Definición 2.3.1. Sean $T_1, T_2, \ldots, T_n$ una muestra aleatoria de tamaño n y $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ los r distintos tiempos de falla observados en una muestra, que posiblemente presentan censurada por la derecha. La posibilidad de que se dé más de una falla en el tiempo $t_{(j)}$ está permitido, por lo que denotaremos al número de fallas en ese tiempo por d_j ($d_j \geq 1$). Para t fijo, el estimador Kaplan-Meier (KM) o producto-límite $\widehat{S}(\cdot)$ de $S(\cdot)$ se define como:

$$\widehat{S}(t) = \prod_{j | t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{2.3}$$

Las principales características de este estimador son:

1. La probabilidad de supervivencia se calcula cada vez que un paciente experimenta el evento, generando las probabilidades en cada momento.

2. Las probabilidades de supervivencia se calculan a partir del número de sujetos en riesgo justo antes del momento de producirse el evento en interés.

3. Los valores de supervivencia se calculan sólo para aquellos momentos en los que algún sujeto sufre el evento. En el resto de momentos se asume que la probabilidad de supervivencia es la misma que en el momento inmediatamente anterior.

Observación 2.4. Observe que el estimador puede ser calculado de forma recursiva, es decir, si $t_{(j-1)}$ y $t_{(j)}$ son dos observaciones adyacentes no censuradas, se tiene que

$$P(\{T > t_{(j)}\} \cap \{T > t_{(j-1)}\}) = S(t_{(j)}).$$

Más aún,

$$\begin{aligned} P(\{T > t_{(j)}\} \cap \{T > t_{(j-1)}\}) &= P(T > t_{(j)} | T > t_{(j-1)}) P(T > t_{(j-1)}) \\ &= P(T > t_{(j)} | T > t_{(j-1)}) S(t_{(j-1)}), \end{aligned}$$

y dado que $\frac{d_j}{n_j}$ es un estimador para p_j, $j = 1, 2, \ldots, r$, se obtiene que

$$\widehat{S}(t_{(j)}) = \widehat{S}(t_{(j-1)}) \left(1 - \frac{d_j}{n_j}\right).$$

Por otro lado, nótese que en ausencia de censura el estimador KM se reduce a la función de supervivencia empírica definida en (2.1), de modo que si se tiene una muestra de tamaño n con $t \in [t_{(j)}, t_{(j+1)}]$, se tiene que:

$$\begin{aligned}
\widehat{S}(t) &= \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)\ldots\left(1 - \frac{d_j}{n_j}\right) \\
&= \left(1 - \frac{d_1}{n}\right)\left(1 - \frac{d_2}{n - d_1}\right)\ldots\left(1 - \frac{d_j}{n - d_1 - d_2 - \ldots - d_{j-1}}\right) \\
&= \left(\frac{n - d_1}{n}\right)\left(\frac{n - d_1 - d_2}{n - d_1}\right)\ldots\left(\frac{n - d_1 - \ldots - d_j}{n - d_1 - \ldots - d_{j-1}}\right) \\
&= \frac{n - d_1 - \ldots - d_j}{n} \\
&= \frac{\#\{i|t_i > t\}}{n}.
\end{aligned}$$

Ejemplo 2.5. A continuación se muestra la función de supervivencia en la Figura 2.1, con el estimador no paramétrico Kaplan-Meier para un grupo de 64 pacientes con insuficiencia renal a los cuales se les realizó un trasplante. Para cada paciente se registró el tiempo postrasplante (en meses) hasta el momento en el que se presentó rechazo del injerto o bien, el tiempo durante el cual el sujeto estuvo bajo observación sin haber rechazado el injerto.

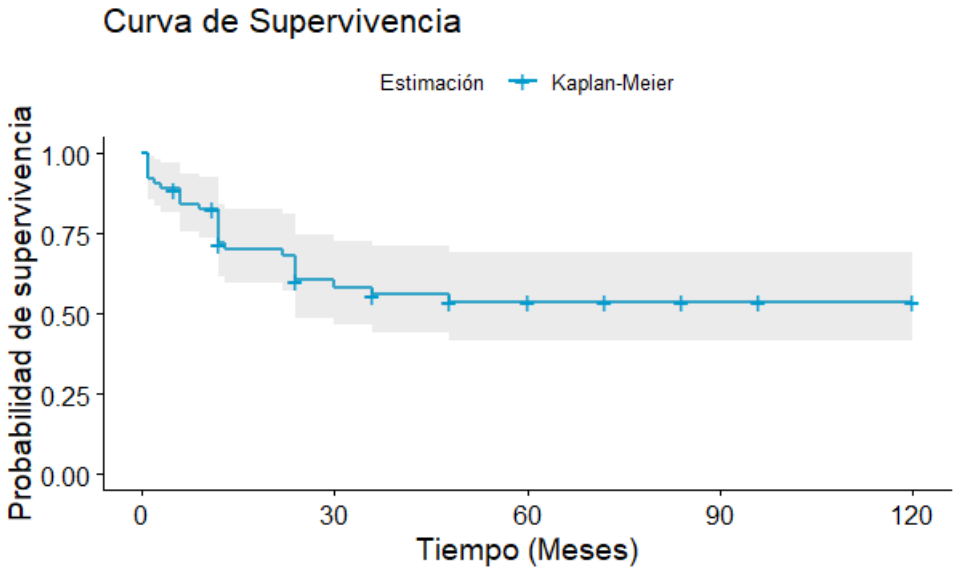


Figura 2.1: Estimador KM de la función de supervivencia, para un grupo de 64 pacientes.

De donde podemos ver que la función de supervivencia decrece de forma pausada dentro del intervalo $(0, 48]$ mientras que dentro del intervalo $(48, 120]$ permanece constante. Esto ocurre debido a que los pacientes que rechazan el injerto por lo general lo hacen dentro de los primeros meses después de haberse realizado la operación, para este conjunto de datos esto ocurre dentro de los

primeros 48 meses, mientras que para los individuos restantes solo les registra el tiempo que han estado en estudio sin haber experimentado el rechazo, el tiempo máximo de observación para este caso es de 120 meses.

En la Figura 2.2 puede ver las curvas de KM para hombres contra mujeres de este mismo conjunto de datos.

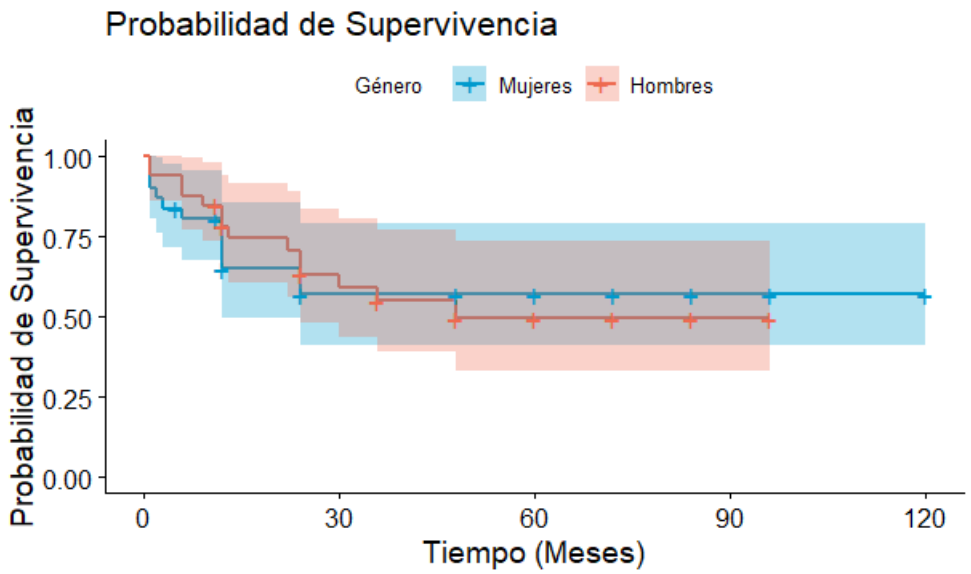


Figura 2.2: Estimador KM de la función de supervivencia, para hombres y mujeres.

Ahora veamos que el estimador Kaplan-Meier definido en (2.3) es el estimador no paramétrico de máxima verosimilitud de $S(\cdot)$. Sean $t_1 < t_2 < \cdots < t_r$ tiempos distintos observados sin censura de una muestra aleatoria de tamaño n. Definímos como d_j al número de fallas en el tiempo t_j, m_j el número de individuos censurados en el intervalo $I_j = [t_j, t_{j+1})$, los cuales se presentan en los tiempos $t_{j1}, t_{j2}, \ldots, t_{jm_j}$, con $j = 1, 2, \ldots, r$, donde $t_0 = 0$ y $t_{r+1} = \infty$, y a n_j como el número de individuos en riesgo justo antes del tiempo t_j, es decir $n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + \cdots + (m_r + d_r)$. Suponga además un modelo de supervivencia a tiempo discreto. Note que la aportación a la verosimilitud de una observación no censurada es:

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

y la contribución verosimilitud de una observación con censurada en el tiempo t_{jl}, con $j \in \{1, 2, \ldots, r\}$ y $l \in \{1, 2, \ldots, m_j\}$ está dada por

$$P(T > t_{jl}) = S(t_{jl}).$$

Además dado que $t_j < t_{jl}$, el *e.m.v.* $\widehat{S}(\cdot)$ de $S(\cdot)$ cumple que $\widehat{S}(t_j) \geq \widehat{S}(t_{jl})$, por lo cual se maximizará la función de verosimilitud tomando $\widehat{S}(t_j) = \widehat{S}(t_{jl})$. Luego, bajo censura no-informativa la función de verosimilitud formulada en los tiempos de falla observadas es

$$L = \prod_{j=1}^{r} \left(\big(S(t_{j-1}) - S(t_j)\big) \prod_{l=1}^{m_j} S(t_{jl}) \right). \tag{2.4}$$

Por otro lado, se tiene que

$$S(t_{j-1}) = \prod_{l=1}^{j-1} (1 - h(t_l)),$$

$$S(t_j) = \prod_{l=1}^{j} (1 - h(t_l)),$$

sustituyendo en (2.4) se obtiene

$$L = \prod_{j=1}^{r} \left\{ \left(\prod_{l=1}^{j-1} (1 - h(t_l)) - \prod_{l=1}^{j} (1 - h(t_l)) \right) \prod_{l=1}^{m_j} S(t_{jl}) \right\}. \tag{2.5}$$

Al simplificar cada uno de sus factores se puede ver,

$$\begin{aligned} \prod_{l=1}^{j-1} (1 - h(t_l)) - \prod_{l=1}^{j} (1 - h(t_l)) &= \left(\prod_{l=1}^{j-1} (1 - h(t_l)) \right) (1 - (1 - h(t_j))) \\ &= \left(\prod_{l=1}^{j-1} (1 - h(t_l)) \right) h(t_j) \end{aligned}$$

y

$$\prod_{l=1}^{m_j} S(t_{jl}) = \prod_{l=1}^{m_j} S(t_j) = S(t_j)^{m_j} = \prod_{l=1}^{j} (1 - h(t_l))^{m_j}.$$

Nótese también que,

$$n_1 = \sum_{j=1}^{r} (d_j + m_j), \;\; n_2 = \sum_{j=2}^{r} (d_j + m_j), \ldots, \;\; n_r = d_r + m_r.$$

Sustituyendo en (2.5),

$$
\begin{aligned}
L &= \prod_{j=1}^{r}\left\{\left(\prod_{l=1}^{j-1}(1-h(t_l))^{d_j}h(t_j)^{d_j}\right)\left(\prod_{l=1}^{j}(1-h(t_l)^{m_j})\right)\right\} \\
&= \prod_{j=1}^{r}\left\{\left(\prod_{l=1}^{j-1}(1-h(t_l))^{d_j}h(t_j)^{d_j}(1-h(t_j))^{d_j}(1-h(t_j))^{-d_j}\right)\right. \\
&\quad \left.\left(\prod_{l=1}^{j}(1-h(t_l)^{m_j})\right)\right\} \\
&= \prod_{j=1}^{r}\left\{\left(\prod_{l=1}^{j}(1-h(t_l))^{d_j}h(t_j)^{d_j}(1-h(t_j))^{-d_j}\right)\left(\prod_{l=1}^{j}(1-h(t_l)^{m_j})\right)\right\} \\
&= \prod_{j=1}^{r}\left\{h(t_j)^{d_j}(1-h(t_j))^{-d_j}\left(\prod_{l=1}^{j}(1-h(t_l))^{d_j}\right)\left(\prod_{l=1}^{j}(1-h(t_l)^{m_j})\right)\right\} \\
&= \prod_{j=1}^{r}\left\{h(t_j)^{d_j}(1-h(t_j))^{-d_j}\left(\prod_{l=1}^{j}(1-h(t_l))^{d_j+m_j}\right)\right\} \\
&= \left(\prod_{j=1}^{r}h(t_j)^{d_j}(1-h(t_j))^{-d_j}\right)\left(\prod_{j=1}^{r}\left[\prod_{l=1}^{j}(1-h(t_l))^{d_j+m_j}\right]\right) \\
&= \left(\prod_{j=1}^{r}h(t_j)^{d_j}(1-h(t_j))^{-d_j}\right)(1-h(t_1))^{n_1}(1-h(t_2))^{n_2}\ldots(1-h(t_r))^{n_r} \\
&= \left(\prod_{j=1}^{r}h(t_j)^{d_j}(1-h(t_j))^{-d_j}\right)\prod_{j=1}^{r}(1-h(t_j))^{n_j} \\
&= \prod_{j=1}^{r}h(t_j)^{d_j}(1-h(t_j))^{-d_j}(1-h(t_j))^{n_j} \\
&= \prod_{j=1}^{r}h(t_j)^{d_j}(1-h(t_j))^{n_j-d_j}
\end{aligned}
$$

así, la función log-verosimilitud está dada por:

$$l(h(t_1),h(t_2),\ldots,h(t_r);\underline{t}) = \sum_{j=1}^{r}\{d_j\log(h(t_j)) + (n_j-d_j)\log(1-h(t_j))\}. \quad (2.6)$$

Para encontrar el *e.m.v.*, se calcula para cada $j = 1, 2, \ldots, r$,

$$\frac{\partial l(h(t_1),h(t_2),\ldots,h(t_r);\underline{t})}{\partial h(t_j)} = \frac{d_j}{h(t_j)} - \frac{n_j-d_j}{1-h(t_j)},$$

igualando a cero se obtiene el punto crítico en $\widehat{h}(t_j) = \frac{d_j}{n_j}$, $j \in \{1, 2, \ldots, r\}$. Además como, para cualquier valor de $h(t_j)$ se cumple

$$\frac{\partial^2 l(h(t_1),h(t_2),\ldots,h(t_r);\underline{t})}{\partial h(t_j)^2} = -\frac{d_i}{h(t_j)^2} - \frac{n_j-d_j}{(1-h(t_j))^2} < 0, \quad \forall j = 1, 2, \ldots, r,$$

en particular para $\widehat{h}(t_j)$ y como

$$\frac{\partial^2 l(h(t_1), h(t_2), \ldots, h(t_r); \underline{t})}{\partial h(t_i) \partial h(t_j)} = 0, \quad i \neq j.$$

Por lo tanto, $\widehat{h}(t_j) = \frac{d_j}{n_j}$ es el *e.m.v.* de $h(t_j)$ y usando el principio de invarianza de los estimadores de máxima verosimilitud se obtiene que en efecto el estimador de máxima verosimilitud de $S(\cdot)$ es $\widehat{S}(\cdot)$ definido en (2.3).

Observación 2.6. De la Observación 2.2 y de la relación $S(t) = 1 - \prod_{j|t_j \leq t} h(t_j)$ se puede probar que el estimador $\widehat{h}(t_j) = \frac{d_j}{n_j}$ es un estimador insesgado. Este hecho es muy importante ya que nos ayuda a obtener la fórmula de Greenwood, la cual se presenta en la siguiente subsección.

2.6.1. Fórmula de Greenwood

Conociendo el estimador de Kaplan-Meier $\widehat{S}(t)$ de la función de supervivencia, muchas veces es deseable conocer la varianza de dicho estimador, esta fórmula es conocida como fórmula de Greenwood, por lo cual se determinará su expresión, además de la determinación de un intervalo de confianza para $S(t)$ del $100(1-\alpha)\,\%$.

Para un modelo a tiempo discreto, sabemos que $\widehat{S}(\cdot)$ puede ser expresado en términos de $\widehat{h}(\cdot)$ como:

$$\widehat{S}(t) = \prod_{j|t_j \leq t} \left(1 - \widehat{h}(t_j)\right),$$

donde $\widehat{h}(t_j) = \frac{d_j}{n_j}$ es el estimador de máxima verosimilitud de $h(t_j)$, con esto se tiene que

$$\log \widehat{S}(t) = \sum_{j|t_j \leq t} \log\left(1 - \widehat{h}(t_j)\right),$$

en consecuencia,

$$Var\left(\log \widehat{S}(t)\right) = \sum_{j|t_j \leq t} Var\left(\log\left(1 - \widehat{h}(t_j)\right)\right). \tag{2.7}$$

Ahora, para aproximar a la $Var\left(\log\left(1 - \widehat{h}(t_j)\right)\right)$ desarrollaremos la expresión: $\log\left(1 - \widehat{h}(t_j)\right)$, en series de Taylor alrededor de $h(t_j) = E(\widehat{h}(t_j))$,

$$\log\left(1 - \widehat{h}(t_j)\right) \approx \log(1 - h(t_j)) - \frac{\widehat{h}(t_j) - h(t_j)}{1 - h(t_j)} + O(n^{-1}),$$

si $O(n^{-1})$ es pequeña. Despejando y elevando al cuadrado de ambos lados

$$\left\{\log\left(1 - \widehat{h}(t_j)\right) - \log(1 - h(t_j))\right\}^2 \approx \left\{\frac{\widehat{h}(t_j) - h(t_j)}{1 - h(t_j)}\right\}^2,$$

así al tomar la esperanza de ambos lados,

$$Var\left(\log\left(1-\widehat{h}(t_j)\right)\right) \approx \left(\frac{1}{1-\widehat{h}(t_j)}\right)^2 Var\left(\widehat{h}(t_j)\right). \tag{2.8}$$

Nuevamente expresando en series de Taylor a $\log \widehat{S}(t)$, alrededor de su media $\log S(t)$ se tiene:

$$\log \widehat{S}(t) \approx \log S(t) + \frac{\widehat{S}(t) - S(t)}{S(t)} + O\left(n^{-1}\right),$$

de donde

$$\left\{\log \widehat{S}(t) - \log S(t)\right\}^2 = \left\{\frac{\widehat{S}(t) - S(t)}{S(t)}\right\}^2,$$

tomando la esperanza de ambos lados

$$Var\left(\log \widehat{S}(t)\right) \approx \frac{1}{\widehat{S}(t)^2} Var\left(\widehat{S}(t)\right). \tag{2.9}$$

De (2.7) y (2.9)

$$Var\left(\widehat{S}(t)\right) \approx \widehat{S}(t)^2 \sum_{j|t_j \leq t} Var\left(\log\left(1-\widehat{h}(t_j)\right)\right). \tag{2.10}$$

Ahora sólo resta determinar la expresión para la $Var\left(\widehat{h}(t_j)\right)$. Considere el vector $\widehat{w} = \left(\widehat{h}(t_1), \widehat{h}(t_2), \ldots, \widehat{h}(t_m)\right)^T$ que es el estimador de máxima verosimilitud de $w = (h(t_1), h(t_2), \ldots, h(t_m))^T$. Una propiedad importante de $\widehat{w}$ es ser asintóticamente normal, es decir, $\widehat{w} \dot{\sim} N(w, I^{-1}(w))$, o bien $\sqrt{n}(\widehat{w} - w) \dot{\sim} N(0, nI^{-1}(w))$, donde

$$I(w) = E\left(-\frac{\partial^2 l(w;\underline{t})}{\partial w \partial w^T}\right).$$

es conocida como la matriz de información de Fisher o matriz de información esperada y la función $l(\cdot)$ es definida en (2.6). Esta matriz puede ser estimada con la matriz de observación $R = (r_{ij})$, donde:

$$r_{ij} = -\left.\frac{\partial^2 l(w,\underline{t})}{\partial h(t_i)\partial h(t_j)}\right|_{w=\widehat{w}}.$$

Así,

$$\begin{aligned} -\left.\frac{\partial^2 l(w,\underline{t})}{\partial h(t_i)\partial h(t_j)}\right|_{w=\widehat{w}} &= \begin{cases} 0, & \text{si } i \neq j, \\ -\left.\frac{\partial^2 l(w,\underline{t})}{\partial h(t_j)^2}\right|_{w=\widehat{w}}, & \text{si } i = j, \end{cases} \\ &= \begin{cases} 0, & \text{si } i \neq j, \\ \frac{d_j}{\widehat{h}(t_j)^2} + \frac{n_j - d_j}{(1-\widehat{h}(t_j))^2} & \text{si } i = j, \end{cases} \end{aligned}$$

luego, la varianza asintótica de $\widehat{h}(t_j)$ es

$$Var\left(\widehat{h}(t_j)\right) = \left(\frac{d_j}{\widehat{h}(t_j)^2} + \frac{n_j - d_j}{(1-\widehat{h}(t_j))^2}\right)^{-1} = \frac{d_j(n_j - d_j)}{n_j^3}.$$

Sustituyendo en (2.8) se obtiene:

$$Var\left(\log(1-\widehat{h}(t_j))\right) \approx \frac{d_j}{n_j(n_j - d_j)},$$

finalmente de (2.10) se obtiene la fórmula

$$Var\left(\widehat{S}(t)\right) \approx \widehat{S}(t)^2 \sum_{j|t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

La cual es conocida como fórmula de Greenwood.

Observación 2.7. En ausencia de censura la fórmula de Greenwood se reduce a $\frac{\widehat{S}(t)\left(1-\widehat{S}(t)\right)}{n}$. Esto debido a que $n_{j+1} = n_j - d_j$ para los tiempos de supervivencia observados $t_{(j)}$, $j = 1, 2, \ldots, r$, es decir, el número de sujetos vivos justo antes de $t_{(j+1)}$, es igual a la diferencia del número de sujetos vivos antes del tiempo $t_{(j)}$ menos los que fallecen justo en $t_{(j)}$. Más aún, con esta relación se puede ver que $\widehat{S}(t) = \frac{n_{k+1}}{n_1}$, con $t \in [t_{(k)}, t_{(k+1)})$. Consecuentemente

$$\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} = \sum_{j=1}^{k} \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \frac{1 - \widehat{S}(t)}{n_1 \widehat{S}(t)}.$$

2.7.1. Intervalo de confianza para la función de supervivencia

Otra manera deseable de obtener estimaciones respecto a un parámetro es la estimación a través de un intervalo de confianza, así, considerando el estimador de Kaplan-Meier $\widehat{S}(\cdot)$ como el estimador de máxima verosimilitud para $S(\cdot)$, se tiene que para muestras grandes, el estimador es asintóticamente normal, esto es,

$$\widehat{S}(t) \dot{\sim} N(S(t), I^{-1}(S(t))).$$

Por lo cual el intervalo de confianza del $100(1-\alpha)\,\%$ para $S(t)$, tomando a $Z = \frac{\widehat{S}(t) - S(t)}{\sqrt{Var\left(\widehat{S}(t)\right)}}$ como cantidad pivotal, es

$$\left(\widehat{S}(t) - z_{\alpha/2}\sqrt{Var\left(\widehat{S}(t)\right)}, \widehat{S}(t) + z_{\alpha/2}\sqrt{Var\left(\widehat{S}(t)\right)}\right),$$

donde $z_{\alpha/2}$ es el $\alpha/2$ cuantil de la distribución normal estándar.

2.8. Estimador de Nelson-Aalen

Ahora se presentan dos estimadores no paramétricos para la función de riesgo acumulativo. Uno de ellos se obtiene a partir del estimador de Kaplan-Meier y de la expresión $H(t) = -\log(S(t))$, resultando

$$\widehat{H}_1(t) = -\log \widehat{S}(t) = -\sum_{j|t_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right). \tag{2.11}$$

El segundo estimador se basa en la expresión $H(t) = \sum_{j|t_j \leq t} h(t_j)$ para el caso de variables aleatorias discretas, el cual es conocido como función de riesgo empírico acumulativo o bien estimador de Nelson-Aalen que se define como

$$\widehat{H}_2(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j}$$

con t_j, $j = 1, 2, \ldots, r$ los distintos tiempos de fallo observados. Este estimador se obtiene al expandir la función $\log(1-x)$ en serie de Maclaurin, es decir, como $\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$, se tiene que

$$\log\left(1 - \frac{d_j}{n_j}\right) \approx -\frac{d_j}{n_j}.$$

Nótese además que de esta expresión se puede obtener nuevamente un estimador para la función de supervivencia:

$$\widehat{S}(t) = \exp\left(-\widehat{H}_2(t)\right),$$

por otro lado, la varianza de este estimador está dada por

$$\begin{aligned} Var\left(\widehat{H}_2(t)\right) &= Var\left(\sum_{j|t_j \leq t} \widehat{h}(t_j)\right) \\ &= \sum_{j|t_j \leq t} Var\left(\widehat{h}(t_j)\right) \\ &= \sum_{j|t_j \leq t} \frac{d_j(n_j - d_j)}{n_j^3}. \end{aligned}$$

En la Figura 2.3 puede ver el comportamiento del estimador de Nelson-Aalen para el riesgo acumulativo, de los pacientes con insuficiencia renal del Ejemplo 2.5.

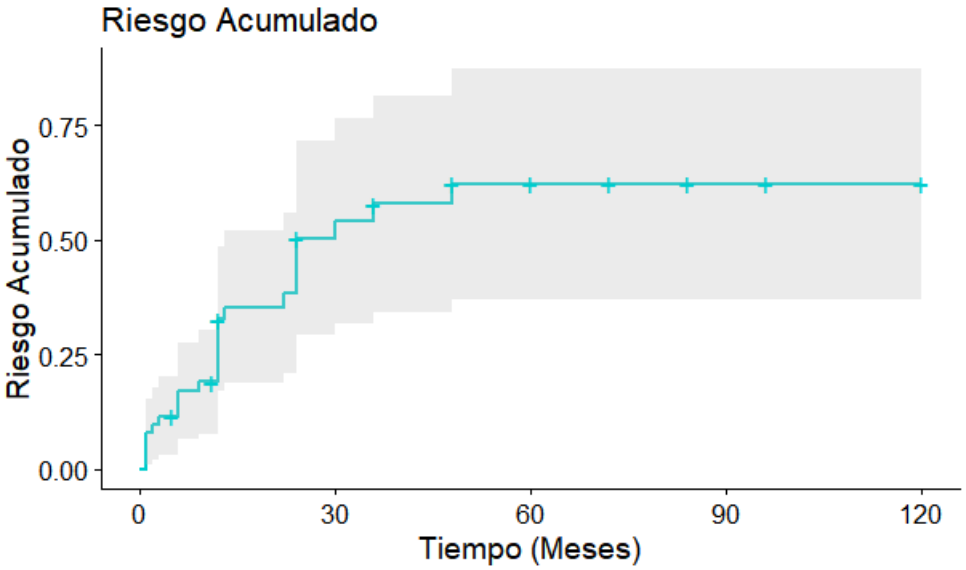


Figura 2.3: Curva del riesgo acumulado con el estimador Nelson Aalen con un intervalo de confianza del 95 %.

Además puede ver la curvas de Nelson Aalen para hombres y mujeres en la Figura 2.4.

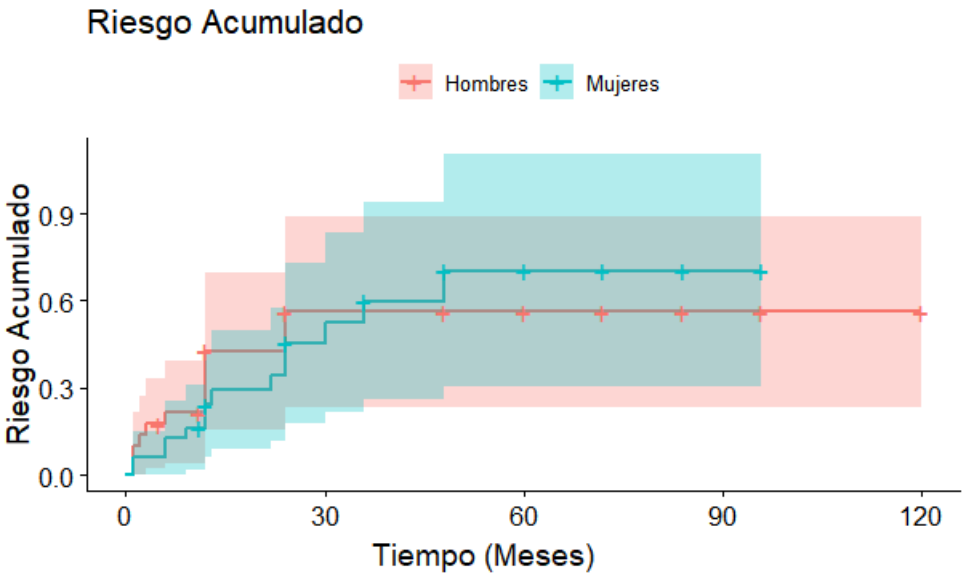


Figura 2.4: Riesgo acumulado con el estimador Nelson Aalen de Hombres Vs Mujeres con un intervalo de confianza del 95 %.

2.9. Criterios de información AIC y BIC para la selección del modelo

En Estadística, la selección del modelo es de primordial importancia, esto es, ¿cómo elegir el modelo más apropiado para el conjunto de datos disponibles dentro de un conjunto de modelos candidatos?. En tal caso es necesario contar con un estadístico que permita seleccionar entre un modelo u otro. El criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC) son criterios de uso frecuente, ambos se basan en la máxima verosimilitud como forma de medida de la bondad de ajuste ([2] y [9]).

Ambos criterios de información se derivan de una minimización insesgada del valor esperado del criterio de Kullback-Leibler (KL) entre un modelo candidato y el modelo verdadero, el cual permite determinar con qué eficiencia los modelos se ajustan a una base de datos. Sin embargo, a pesar de que la información de KL es bastante razonable para evaluar que tan adecuado es un modelo, en la práctica suele ser bastante limitada ya que casi siempre se desconoce la verdadera distribución que genera los datos, así, lo habitual es asumir un modelo paramétrico y luego estimar los parámetros de máxima verosimilitud, más aún, dado que en la práctica es frecuente que no se tenga bien identificado el modelo, se proponen varios modelos paramétricos para el mismo problema, lo cual hace el AIC y el BIC. Así, Akaike propone una nueva forma de evaluar la bondad de ajuste de un modelo paramétrico $f(x;\theta)$ dentro de la familia paramétrica $\{f(x,\theta) : \theta \in \overline{\ominus} \subset \mathbb{R}^p\}$ con el modelo verdadero $g(x;\theta_0)$ y en el cual además considera la situación en el que $g(x;\theta_0) = f(x;\theta_1)$ para algún $\theta_1 \in \overline{\ominus}$, es decir, la función de densidad verdadera está incluida dentro de la familia paramétrica.

La comparación de modelos candidatos se puede realizar con

$$AIC = 2k - 2\log(L(\underline{x};\widehat{\theta})),$$

donde k es el número de paramétros del modelo y $\log(L(\underline{x};\widehat{\theta}))$ el valor máximo en la función de verosimilitud. De esta forma AIC mide la bondad de ajuste con la verosimilitud y la complejidad con el valor de k penalizando la utilización de muchos paramétros. La función $-2\log(L(\underline{x};\widehat{\theta}))$ es decreciente, de modo que el AIC decrece cuando la bondad de ajuste aumenta, es decir, cuanto más alto sea el valor de máxima verosimilitud mejor ajustado estará el modelos de los datos, pues AIC será pequeño.

Note que AIC tiene una forma menos restrictiva que otros indicadores de medir la complejidad del modelo, ya que sólo suma al indicador un número entero k que corresponde al número de parámetros, sin embargo, el problema que resulta de usar solo k como medida de complejidad es que su efecto puede ser pequeño en comparación con el valor que puede llegar a tomar $-2\log(L(\underline{x};\widehat{\theta}))$, que además depende del tamaño de la muestra.

Esto motiva a desarrollar un nuevo criterio de información alternativo, el

cual es conocido como el criterio de información bayesiano (BIC) o criterio de Schwarz (se denomina bayesiano por basarse en argumentos de la llamada Estadística bayesiana), su fórmula está dada por

$$BIC = k\log(n) - 2\log(L(\underline{x};\widehat{\theta}\,)),$$

donde k es el número de parámetros, $\log(L(\underline{x};\widehat{\theta}\,))$ es el valor de máxima verosimilitud y n es el número de datos con los que se cuenta. Nótese que este criterio también se basa en la máxima verosimilitud como método de medida de la bondad de ajuste, en cuanto a la complejidad, esta fórmula incorpora tanto a k como a $\log(n)$, lo cual produce que haya una mayor penalización en la inclusión de paramétros, de hecho el BIC generalmente selecciona el modelo más sencillo y que hace predicciones a menor detalle, mientras que el AIC se inclina por un modelo más complejo pero que hace predicciones con mayor detalle. Sin embargo, el criterio para elegir el mejor modelo es el mismo que el de Akaike, el que tenga el menor valor de BIC.

Capítulo 3

MODELOS DE REGRESIÓN

En problemas más complejos del Análisis de Supervivencia es común que se registre información adicional para cada individuo, la cual puede depender del tiempo. Así, cuando se realiza dicho análisis, generalmente lo que se desea saber es cómo influyen estas características en la variable en estudio.

Por ello, surge la necesidad de considerar generalizaciones de modelos paramétricos para tomar en cuenta ese tipo de información concomitante del sujeto en estudio. Estos modelos son los modelos de regresión, los cuales pueden usarse principalmente por dos motivos, el primero, es cuando el objetivo es predecir lo mejor posible el valor esperado de la variable respuesta T usando un conjunto de variables explicativas (covariables, predictores, factores de riesgo) $\{x_1, x_2, \ldots, x_p\}$, que pueden representar características del individuo tales como tratamientos, propiedades intrínsecas del individuo, por ejemplo, peso, talla, características individuales o otras condiciones. Otro motivo es cuando el interés del investigador es el de estimar el efecto de una o más variables explicativas x_j sobre la variable respuesta. Los modelos de regresión pueden formularse de diversas formas, algunos de ellos se muestran a continuación [11].

3.1. Modelo de regresión Exponencial

Suponga que un individuo tiene un tiempo de vida T y un vector de variables explicativas $\mathbf{X}' = (x_1, x_2, \ldots, x_p)$. Bajo el modelo de distribución exponencial, la función de supervivencia de la *v.a.* T es:

$$S(t|\ \mathbf{X}) = \exp(-\lambda(\ \mathbf{X})t), \quad t \geq 0,$$

en este modelo $\lambda(\mathbf{X})$ es una función especificada, de hecho comúnmente suele utilizarse $\lambda(\mathbf{X}) = \exp(\boldsymbol{\beta}'\mathbf{X})$, donde $\boldsymbol{\beta}$ es el vector de coeficientes de regresión, con la propiedad conveniente de que $\lambda(\mathbf{X}) > 0$ para cualquier valor de $\boldsymbol{\beta}$ y $\mathbf{X}$.

3.2. Modelo de regresión Weibull

En el caso de la distribución de Weibull con parámetros α y θ ($W(\alpha, \theta)$), de la sección 1.3, puede ver la función de supervivencia, de igual manera en

un modelo de regresión, se consideran los parámetros α y $\boldsymbol{\beta}$ que dependen de la covariable $\mathbf{X}$, de modo que la función de supervivencia es:

$$S(t|\mathbf{X}) = \exp\left((-\alpha(\mathbf{X})t)^{\theta(\mathbf{X})}\right), \quad t \geq 0.$$

Sin embargo, en este modelo, es usual suponer que sólo α depende del vector de covariables $\mathbf{X}$.

3.3. Modelo de tiempo de falla acelerada

Este modelo es frecuentemente usado en casos clínicos, ya que supone que las variables explicativas de un individuo actúan multiplicativamente en la escala del tiempo, de modo que la covariable altera la tasa en la que un individuo envejece o rejuvenece a lo largo del tiempo. Esto significa que los modelos pueden interpretarse en términos de la velocidad de progresión de una enfermedad [2].

Antes de presentar el modelo, este es descrito para comparar los tiempos de supervivencia de dos grupos de pacientes. Suponga que a los pacientes de manera aleatoria se les trata con un nuevo tratamiento N o con el tratamiento estándar S. En un modelo de falla acelerada, se considera que el tiempo de supervivencia de un individuo en el nuevo tratamiento es un múltiplo del tiempo de supervivencia para un individuo con el tratamiento estándar, por lo tanto el efecto del nuevo tratamiento es acelerar o ralentizar el paso del tiempo. Así, bajo este supuesto, la probabilidad de que un individuo que es tratado con el nuevo tratamiento sobreviva más allá del tiempo t es la probabilidad de que un individuo con el tratamiento estándar sobreviva más allá del tiempo ϕt, de modo que si $S_N(\cdot)$ y $S_S(\cdot)$ representan la función de supervivencia de los individuos con el nuevo tratamiento y con el tratamiento estándar respectivamente, el modelo de falla acelerada específica que:

$$S_N(t) = S_S(\phi t), \tag{3.1}$$

para cualquier valor de t. Por tanto con esta relación, cuando la preocupación es la muerte del paciente, los valores de $\phi < 1$ producen una desaceleración en el tiempo de un individuo al cual se le ha asignado el nuevo tratamiento, es decir, en este caso, el tratamiento estándar es el más apropiado para promover la longevidad del paciente. En cambio, si el punto final es la recuperación del paciente, para $\phi > 1$ el efecto en el tratamiento nuevo es acelerar el tiempo de recuperación, así, bajo estas circunstancias el tratamiento nuevo es el más apropiado.

Por otra parte, de la relación (3.1), se puede ver que:

$$\begin{aligned} f_N(t) &= \phi f_S(\phi t), \\ h_N(t) &= \phi h_S(\phi t). \end{aligned}$$

Ahora, consideremos a la variable indicadora δ_i, la cual toma el valor de 0 si el i-ésimo individuo recibe el tratamiento estándar y 1 si recibe el tratamiento

nuevo. Bajo el modelo de falla acelerada la función de riesgo para el individuo i-ésimo es,

$$h_i(t) = \phi^{\delta_i} h_0(\phi^{\delta_i} t),$$

donde $h_0(t)$ es conocido como función de riesgo base. Además, dado que el parámetro ϕ no debe ser negativo, por conveniencia se toma $\phi = \exp(\boldsymbol{\beta}\delta_i)$, así la *f.r.* para el individuo con el tratamiento nuevo es

$$h_i(t) = \exp(\boldsymbol{\beta}) h_0(\exp(\boldsymbol{\beta})t).$$

Ahora consideremos que ϕ depende del vector de covariables, es decir, dado un vector de covariables fijas $\mathbf{X}$, existe una función positiva $\phi(\cdot)$ tal que *f.d.p.*, *f.s.* y *f.r.* son tales que

$$\begin{aligned} S(t|\mathbf{X}) &= S_0(t\phi(\mathbf{X})), \\ f(t|\mathbf{X}) &= f_0((t\phi(\mathbf{X})))\phi(\mathbf{X}), \\ h(t|\mathbf{X}) &= h_0(t\phi(\mathbf{X}))\phi(\mathbf{X}), \end{aligned}$$

con $S_0(t)$ es la función de supervivencia base en $\mathbf{X} = 0$ y $\phi(0) = 1$.

Otra forma de definir el modelo de vida acelerada en términos de las variables aleatorias es:

$$T = \frac{T_0}{\phi(\mathbf{X})},$$

T_0 tiene como función de supervivencia a la *f.s.* en $\mathbf{X} = 0$, S_0. Con mayor frecuencia se utiliza $\phi(\mathbf{X}) = \exp(\boldsymbol{\beta}'\mathbf{X})$ ($\boldsymbol{\beta}'$ es el vector de coeficientes de regresión del modelo), puesto que esta función satisface que $\phi(\mathbf{X}) > 0$ y $\phi(0) = 1$. Finalmente, el modelo de falla acelerada con la función de riesgo está dado por:

$$h(t) = \exp(\boldsymbol{\beta}'\mathbf{X}) h_0(\exp(\boldsymbol{\beta}'\mathbf{X})t).$$

3.3.1. Modelo de falla acelerada Weibull

Ahora, si imponemos a la distribución Weibull como distribución de probabilidad en los tiempos de supervivencia en el modelo de falla acelerada. Esto es, la función de riesgo base es $h_0(t) = \lambda\gamma(\lambda t)^{\gamma-1}$, de modo que

$$h(t|\mathbf{X}) = \exp(\boldsymbol{\beta}'\mathbf{X})\lambda\gamma(\lambda \exp(\boldsymbol{\beta}')t)^{\gamma-1},$$

de aquí, se puede ver que el tiempo de supervivencia del sujeto en estudio es nuevamente del tipo Weibull, pero con parámetros $\exp(\boldsymbol{\beta}'\mathbf{X})\lambda$ y γ, es decir, $W(\exp(\boldsymbol{\beta}'\mathbf{X})\lambda, \gamma)$. Con esto, se dice que la distribución de Weibull posee la propiedad del tiempo de falla acelerada.

3.3.2. Modelo de falla acelerada log-Logística

Consideremos ahora la distribución log-Logístico como distribución de probabilidad en los tiempos de supervivencia en el modelo de falla acelerada, así $h_0 = \frac{\lambda k(\lambda t)^{k-1}}{1+(\lambda t)^k}$, luego

$$h(t) = \frac{\exp(\boldsymbol{\beta}'\mathbf{X})\lambda k(\lambda t \exp(\boldsymbol{\beta}'\mathbf{X})^{k-1}}{1 + (\lambda \exp(\boldsymbol{\beta}'\mathbf{X}))^k},$$

de donde la distribución que sigue el tiempo de supervivencia es log-Logístico de parámetros $\lambda \exp(\boldsymbol{\beta}'\mathbf{X})$ y k. Por lo tanto, la distribución log-Logístico al igual que la distribución Weibull también posee la propiedad del tiempo de falla acelerada.

3.4. Modelos de Riesgos Proporcionales de Cox

El Modelo de Regresión de Cox o Modelo de Riesgos Proporcionales de Cox, es el modelo con mayor uso en la investigación clínica dada su facilidad para interpretar los coeficientes de regresión ([3], [11], [17]).

Para la formulación de la función de riesgo del modelo considere T una variable aleatoria de tiempo de vida que depende en la respuesta del vector de covariables fijas (independientes del tiempo) $\mathbf{X}' = (x_1, x_2, \ldots, x_p)$. El modelo de riesgos proporcionales para T es el siguiente:

$$h(t|\mathbf{X}) = h_0(t)\psi(\mathbf{X}), \tag{3.2}$$

donde $h(t|\mathbf{X})$ es la función de riesgo condicional de t dado $\mathbf{X}$, $\psi(\mathbf{X})$ es una función positiva en $\mathbf{X}$ tal que $\psi(0) = 1$ (la selección de ψ dependerá del comportamiento de los datos), y $h_0(t)$ es la función de riesgo base, la cual representa la función de riesgo para una observación cuyo vector de covariables es cero, es decir, $h(t|0) = h_0(t)$. Además, comúnmente se considera que $\psi(\mathbf{X})$ es de la forma $\psi(\mathbf{X}) = \exp(\boldsymbol{\beta}'\mathbf{X})$, donde $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ es un vector de parámetros desconocidos, conocido como vector de coeficientes de regresión, obteniendo de (3.2) que:

$$h(t|\mathbf{X}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{X}), \tag{3.3}$$

esta forma paramétrica de ψ es conocida como forma log-lineal, aunque es frecuente el uso de la forma lineal $\psi(\mathbf{X}) = 1 + \boldsymbol{\beta}'\mathbf{X}$ y la logística $\psi(\mathbf{X}) = 1 + \exp(\boldsymbol{\beta}'\mathbf{X})$.

Nótese que con (3.3) se obtiene un modelo semiparamétrico o parcialmente paramétrico, ya que la parte paramétrica es representada por ψ y no es paramétrico en cuanto a que no se especifica la forma funcional exacta de h_0, lo cual significa a su vez que no supone una forma paramétrica específica de la distribución de supervivencia de los tiempos. Otra de las características del modelo es que las variables explicativas tienen un efecto multiplicativo a través del producto con $h_0(\cdot)$.

3.4.1. Hipótesis de riesgos proporcionales

En el modelo de Cox se busca como primer paso la relación entre los riesgos de muerte de dos individuos expuestos a factores de riesgo diferentes, esto es, si $\mathbf{X}^*$ y $\mathbf{X}$ son dos vectores de covariables observados, de distintos individuos, la razón de sus funciones de riesgo por (3.3) es:

$$HR = \frac{h(t|\mathbf{X})}{h(t|\mathbf{X}^*)} = \frac{\psi(\mathbf{X})h_0(t)}{\psi(\mathbf{X}^*)h_0(t)} = \exp(\boldsymbol{\beta}'(\mathbf{X} - \mathbf{X}^*)),$$

nótese que esta razón no depende del tiempo de falla t, de hecho sólo depende de los valores de las variables explicativas y los coeficientes de regresión. Esta razón es conocido como hipótesis de riesgos proporcionales. Esta hipótesis significa que la razón de riesgos es constante a lo largo del tiempo, de modo que si θ es constante, es decir:

$$h(t|\mathbf{X}) = \theta h(t|\mathbf{X}^*),$$

por lo cual, una vez estimando los coeficientes de regresión por máxima verosimilitud parcial, tenemos que la razón de proporcionalidad es:

$$\hat{\theta} = \exp\left(\sum_{j=1}^{p} \hat{\beta}_j(x_j - x_j^*)\right).$$

Por otra parte, la suposición de riesgos proporcionales requiere una cuidadosa verificación. En [10] se explican 3 formas de cómo se puede realizar dicho análisis, las cuales son: por método gráfico, pruebas de bondad de ajuste o por variables dependientes del tiempo.

3.4.2. Función de supervivencia bajo riesgos proporcionales

Consideremos a T una variable aleatoria continua, con vector de variables explicativas $\mathbf{X} \in \mathbb{R}^p$. Definimos a $S(t|\mathbf{X})$ como la función de supervivencia condicional de t, cuya función de riesgo condicional es $h(t|\mathbf{X})$, así por (1.5) se tiene que la función de supervivencia está dado por:

$$\begin{aligned} S(t|\mathbf{X}) &= \exp\left(-\int_0^t h(s|\mathbf{X})ds\right) = \exp\left(-\int_0^t h_0(s)\exp(\boldsymbol{\beta}'\mathbf{X})ds\right) \\ &= \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{X})\int_0^t h_0(s)ds\right) = \left(\exp\left(-\int_0^t h_0(s)ds\right)\right)^{\exp(\boldsymbol{\beta}'\mathbf{X})} \\ &= S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X})}, \end{aligned}$$

donde $S_0(t) = S(t|\mathbf{X} = 0) = \exp\left(-\int_0^t h_0(s)ds\right)$ es la función base de supervivencia. De igual manera podemos expresar la igualdad anterior como:

$$S(t|\mathbf{X}) = \exp(-H_0(t))^{\exp(\boldsymbol{\beta}'\mathbf{X})},$$

H_0 es la función de riesgo acumulativo. A partir de esto, se puede encontrar la función de densidad $f(t|\mathbf{X})$, mediante la igualdad (1.25) obteniendo:

$$f(t|\mathbf{X}) = h(t|\mathbf{X})S(t|\mathbf{X}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{X})S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X})}. \tag{3.4}$$

3.4.3. Estimación de los parámetros

En 1972 Cox introduce un método para estimar el vector de coeficientes de regresión $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$, sin hacer uso de la especificación exacta de la función de riesgo base $h_0(t)$, mediante el uso de verosimilitud parcial; se denomina parcial debido a que parte de la idea de que los tiempos de datos con censura no aportan información sobre el tiempo de fallo.

A continuación describiremos como obtener las estimaciones de los parámetros del modelo de Cox mediante verosimilitud parcial. Para esto, primero nótese por que considerando sólo la función de verosimilitud, no es posible determinar los coeficiente de regresión. Considere que las fallas son obtenidas en los distintos tiempos $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ con sus covariables respectivas asociadas $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \ldots, \mathbf{X}_{(r)}$, así, los $n-r$ individuos restantes están sujetos a censura, considere también el caso en el que no se presentan empates en estos tiempos de falla, luego por (3.4) la función verosimilitud es:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} h_0(t_{(j)})^{\delta_i}\exp(\boldsymbol{\beta}'\mathbf{X}_{(j)})S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X}_{(j)})},$$

de modo que, al no ser $h_0(t)$ conocida, no es posible estimar el valor de $\boldsymbol{\beta}$.

Para evitar la presencia de $h_0(t)$ considere las condiciones anteriores y el conjunto $R_j = R(t_{(j)})$ formado por los individuos en situación de riesgo en el instante del j-ésimo fallo, inclusive los datos con censura mayor que $t_{(j)}$, y considere los siguientes eventos:

1. V_j: sobrevivir hasta el tiempo $t_{(j)}$.
2. M_j: en el conjunto R_j hay un único fallo en $t_{(j)}$.
3. A_j: un individuo con covariable $\mathbf{X}_{(j)}$ falla en el instante $t_{(j)}$.

Con esto podemos deducir que $A_j \cap M_j = A_j$, M_j y V_j son independientes y que la verosimilitud parcial del j-ésimo individuo que falla es:

$$\begin{aligned}
L^*_{(j)} &= P(A_j|V_j \cap M_j) = \frac{P(A_j \cap V_j \cap M_j)}{P(V_j \cap M_j)} \\
&= \frac{P(A_j \cap V_j)}{P(V_j \cap M_j)} = \frac{P(A_j|V_j)P(V_j)}{P(M_j)P(V_j)} \\
&= \frac{P(A_j|V_j)}{P(M_j)} = \frac{h(t_{(j)}|X_{(j)})}{\sum\limits_{i\in R_j} h(t_{(j)}|X_{(i)})} \\
&= \frac{h_0(t_{(j)})\exp(\boldsymbol{\beta}'X_{(j)})}{\sum\limits_{i\in R_j} h_0(t_{(j)})\exp(\boldsymbol{\beta}'X_{(i)})} = \frac{\exp(\boldsymbol{\beta}'X_{(j)})}{\sum\limits_{i\in R_j}\exp(\boldsymbol{\beta}'X_{(i)})}.
\end{aligned}$$

Por tanto, la verosimilitud parcial conjunta de la muestra es:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} L^*_{(j)} = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_{(j)})}{\sum\limits_{i \in R_j} \exp(\boldsymbol{\beta}'\mathbf{X}_{(i)})},$$

esta expresión ya no depende de $h_0(\cdot)$, además, la expresión del numerador únicamente depende los individuos que fallan, mientras el denominador, hace uso de la información de aquellos individuos que están en riesgo, incluyendo los datos que posiblemente presenten censura.

La función log-verosimilitud es:

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{r} \left(\boldsymbol{\beta}'\mathbf{X}_{(j)} - \log \left(\sum_{i \in R_j} \boldsymbol{\beta}'\mathbf{X}_{(i)} \right) \right).$$

Por otra parte, cuando se presentan empates en los datos, existen varias propuestas para construir la verosimilitud parcial. Algunos de estos métodos son: El método de Breslow (1974), método de Efron (1977) y el método de probabilidad parcial exacta [17].

Para el modelo de Breslow, se considera $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ los r distintos tiempos de falla observadas y ordenadas y denotemos por:

1. d_j: número de fallas ocurridas en el instante $t_{(j)}$.
2. $D_j \equiv D(t_{(j)}) = \{j_1, j_2, \ldots, j_{d_j}\}$: el conjunto de las etiquetas de los individuos que fallan en el tiempo $t_{(j)}$.
3. $S_j = \sum_{l \in D_j} \mathbf{X}_l$.
4. $R_j = R(t_{(j)})$.

La verosimilitud parcial sugerida por Breslow considera que las d_j fallas al tiempo $t_{(j)}$ son distintas y ocurren de manera secuencial. Además, el modelo se considera muy bueno si se cuenta con poco datos [13]. Este es el siguiente:

$$L_B(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' S_j)}{\left[\sum\limits_{l \in R_j} \exp(\boldsymbol{\beta}'\mathbf{X}_l) \right]^{d_j}},$$

de donde la función log-verosimilitud es

$$l^*(\boldsymbol{\beta}) = \sum_{j=1}^{r} \left[\boldsymbol{\beta}' S_j - d_j \log \left(\sum_{l \in R_j} \exp(\boldsymbol{\beta}'\mathbf{X}_l) \right) \right].$$

Otra alternativa a la verosimilitud parcial, es planteada por Efron en 1977, la cual está dada por:

$$L_E(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' S_j)}{\prod\limits_{k=1}^{d_j} \left[\sum\limits_{l \in R_j} \exp(\boldsymbol{\beta}'\mathbf{X}_l) - \frac{k-1}{d_k} \sum\limits_{l \in D_j} \exp(\boldsymbol{\beta}'\mathbf{X}_l) \right]}.$$

3.5. Evaluación de la hipótesis de riesgos proporcionales

En está sección presentamos dos métodos para la evaluación de la hipótesis de riesgos proporcionales en el modelo de Cox. Estos son: el método gráfico y el estudio de la bondad de ajuste.

3.5.1. Método gráfico

El método más popular, es el uso de gráficas llamadas log-log. Para ver si se cumple la hipótesis de riesgos proporcionales, las gráficas para las diferentes clases de las covariables del modelo por separado deben ser paralelas. Otra forma gráfica alternativa, es el método de comparar las curvas de supervivencia "observadas" y las "esperadas" para las distintas clases de las covariables por separado, las cuales deberían ser parecidas visualmente [9].

Veamos primero como usar el método gráfico log-log para evaluar la hipótesis de riesgos proporcionales. Note que $\hat{S}(t)$ está entre 0 y 1, de modo que $\log(\hat{S}(t))$ es negativo, luego $-\log(-\log(\hat{S}(t)))$ llega a tomar valores que oscilan entre $-\infty$ e ∞. De la expresión $\hat{S}(t|\mathbf{X}) = \hat{S}_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X})}$, se tiene que:

$$\log\left(\hat{S}(t|\mathbf{X})\right) = \exp(\boldsymbol{\beta}'\mathbf{X}) \log\left(\hat{S}_0(t)\right),$$

de donde

$$-\log\left(-\log\left(\hat{S}(t|\mathbf{X})\right)\right) = -\boldsymbol{\beta}'\mathbf{X} - \log\left(-\log\left(\hat{S}_0(t)\right)\right).$$

Ahora consideremos dos vectores de covariables $\mathbf{X}_1, \mathbf{X}_2$ de sujetos distintos, la diferencia para un momento de tiempo dado t,

$$\log\left(-\log\left(\hat{S}(t|\mathbf{X}_1)\right)\right) - \log\left(-\log\left(\hat{S}(t|\mathbf{X}_2)\right)\right) \;=\; \boldsymbol{\beta}'(\mathbf{X}_1 - \mathbf{X}_2).$$

La interpretación con respecto a la hipótesis de riesgos proporcionales es que, si al dibujar las gráficas log-log obtenemos que las curvas son paralelas, entonces se cumple dicha hipótesis, ya que en este caso la diferencia debería de ser $\boldsymbol{\beta}'(\mathbf{X}_1 - \mathbf{X}_2)$. Esto es, el método consiste en comparar las curvas sobre diferentes categorías (clases) de variables explicativas que se están investigando.

Por otra parte, una alternativa gráfica es comparar lo "observado" contra lo "esperado". Aquí se trata de validar la hipótesis de riesgos proporcionales para las variables de una en una, por lo cual hacemos uso de las curvas de KM para obtener las gráficas "observadas". Para obtener dichas gráficas debemos separar los datos que se tienen disponibles por categorías del predictor a evaluar, luego obtenemos las curvas de KM por separado para cada categoría o variables en estudio.

Para obtener las gráficas "esperadas" ajustamos un modelo de riesgos proporcionales de Cox que contiene el predictor o covariable que se esta evaluando,

dibujamos las curvas resultantes para cada uno de los sujetos de la clase, así, si las curvas "observadas" y las "esperadas" son similares, entonces podemos aceptar la hipótesis de riesgos proporcionales, en caso contrario la rechazamos.

3.5.2. Estudio de la bondad de ajuste (Goodness-of-fit, GOF)

Aquí se realiza un contraste de hipótesis, donde la hipótesis nula es la proporcionalidad de los riesgos, contra la hipótesis alternativa de no proporcionalidad. Así, si el p-valor es "grande" aceptamos la hipótesis nula y si es pequeño lo rechazamos. De hecho, lo que nos interesa es aceptar dicha hipótesis, por lo cual buscamos p-valores elevados. El contraste se realiza para cada covariable del modelo por separado y por tanto se obtiene un p-valor por cada covariable.

Este método se basa en el uso de los residuos de Schoenfeld los cuales se calculan para cada covariable y para cada individuo. Por ejemplo si tenemos k covariables en el modelo, entonces para cada sujeto con tiempo de muerte se obtienen k residuos de Schoenfeld, una para cada una de las k covariables. Estos residuos están definidos sólo para sujetos no censurados, y se definen como:

$$R_{ij} = \mathbf{X}_{ij} - \frac{\sum_{l \in R_i} \mathbf{X}_{lj} \exp(\boldsymbol{\beta}' \mathbf{X}_l)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}' \mathbf{X}_l)}.$$

La idea estadística es que el supuesto de riesgo proporcional se cumple para una covariable en particular si los residuos de Schoenfeld de dicha covariable no están relacionadas con los tiempo de supervivencia. Esta prueba se basa en 3 pasos: el primero, es obtener los residuos de Schoenfeld del modelo de Cox para cada una de las covariables, el segundo es ordenar los tiempos de fallo etiquetando el orden de cada tiempo. Por último, verificamos la correlación entre las variables creadas en el primer y segundo paso.

La hipótesis nula es que la correlación entre los residuos y los tiempos de fallo es cero, de modo que si se rechaza la hipótesis nula entonces se llega a la conclusión de que el supuesto de riesgos proporcionales no se satisface.

Por otro lado, un punto importante de resaltar es que a pesar de que este enfoque es mucho más objetivo que el método gráfico, lo único que indica es que no hay suficientes pruebas para rechazar la hipótesis nula, además de que el p-valor depende del tamaño de la muestra, por lo cual una violación grave del supuesto nulo puede ser no estadísticamente significativa si la muestra es pequeña, o de manera inversa, una ligera violación de la hipótesis nula puede ser altamente significativa si la muestra es grande.

Capítulo 4

CASO DE ESTUDIO PRIMERA PARTE

En la actualidad cerca de 9 millones de personas sufren de insuficiencia renal en México, de los cuales aproximadamente 200, 000 requieren un trasplante de riñón y cada vez esta cifra va en aumento, por lo cual, México enfrenta una crisis en materia de salud, ya que además de no contar con hospitales suficientes, tampoco cuenta con los recursos económicos necesarios para atender estos problemas. Dentro de las causas más frecuentes de la insuficiencia renal son: diabetes Mellitus, hipertensión arterial, nefritis, estrés, una inadecuada alimentación, entre otros.

Los riñones son órganos vitales que realizan muchas funciones en nuestro cuerpo con el fin de mantener la sangre limpia y químicamente balanceada. Tienen la función tanto de eliminación como de regulación de líquidos internos, excretan agua, pero también la conservan. De hecho, eliminan por medio de la orina todos los productos del metabolismo que se obtienen de los alimentos y que pueden llegar a ser dañinos para el cuerpo devolviendo a la sangre sustancias vitales en cantidades adecuadas. Sin embargo, si los riñones se encuentran dañados entonces no funcionan correctamente, lo cual llega a provocar que puedan acumularse desechos en el organismo, elevando con ello la presión arterial y a retener el exceso de líquidos provocando lo que se conoce como insuficiencia renal, es decir, la insuficiencia renal es el resultado que se produce cuando los riñones no son capaces de filtrar adecuadamente toxinas y sustancias de desecho en la sangre, consulte [5] y [6]).

Por otra parte, si los riñones fallan entonces se llega a requerir de tratamiento especial para reemplazar las funciones que estos realizan. Las opciones de tratamiento son; diálisis y el trasplante renal. La diálisis es un proceso artificial de eliminación de los desechos y exceso de agua en la sangre. Existen dos tipos de diálisis; la hemodiálisis y la diálisis peritoneal.

El trasplante renal es una operación importante, comúnmente es el tratamiento elegido por los pacientes, pues de otra manera la diálisis sería un tratamiento de por vida y aunque no siempre suelen haber complicaciones asociadas, una de las complicaciones más temidas es la reacción al rechazo del

riñón. Además, en comparación con la diálisis, el trasplante de riñón suele asociarse con una mejor calidad de vida, menor riesgo de muerte, menos restricciones en la dieta y un menor costo en el tratamiento del paciente. El fallo de riñón, es también conocido como fallo renal, es un término utilizado para describir una situación en la que los riñones ya no pueden funcionar eficazmente, provocando con ello que se llegue a requerir un trasplante, sin embargo suele ocurrir que la persona a quien se le ha realizado dicha operación rechace el injerto [12].

En este capítulo se presenta una aplicación del Análisis de Supervivencia, de un conjunto de datos formado por registros de 64 pacientes con problemas de insuficiencia renal, del estado de Puebla, datos que fueron obtenidos del Instituto Mexicano del Seguro Social (IMSS). Mediante un análisis detallado se encontró que 26 pacientes rechazaron el injerto trasplantado y 38 pacientes son censurados mediante un mecanismo de censura aleatoria. Para este grupo de pacientes se estima la función de supervivencia haciendo uso de algunos de los modelos paramétricos y no paramétricos estudiados, además de realizar dicho análisis bajo el modelo de riesgos proporcionales.

4.1. Análisis exploratorio

Las principales variables a considerar en el caso de estudio son, la edad, talla en cm, nivel de creatina sérica y el tiempo de isquemia caliente (estas variables se describen con mayor detalle en la sección 4.4). En esta sección se calcularán medidas descriptivas sobre el número de observaciones censuradas de acuerdo a algunas variables con las que se cuenta.

Mediante las funciones *table* y *with* del software R para analizar la base de datos se obtiene la siguiente información.

Censura		Sexo	Censura		Intervalo	Censura	
0	1		0	1		0	1
38	26	Hombre	19	14	$(0, 15]$	17	16
		Mujer	19	12	$(15, 40]$	21	10

Tabla 4.1: Información obtenida del registro de los pacientes.

De la Tabla 4.1, puede ver que existe un mayor número de pacientes que no rechazan el trasplante renal (38 pacientes) durante cierto tiempo de estudio, de estos, los de mayor edad $(15, 40]$ son en su gran mayoría los que no rechazan. En la Figura 4.1, puede ver que hay un gran número pacientes con insuficiencia renal cuyas edades están en el intervalo $(10, 20]$.

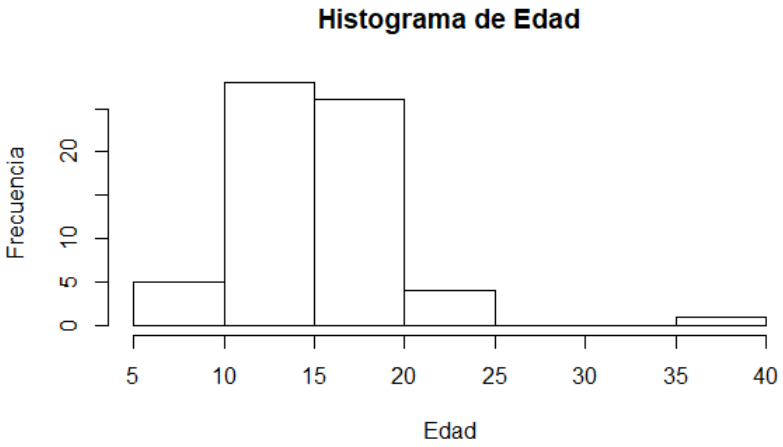


Figura 4.1: Histograma de pacientes con insuficiencia renal.

Ahora, se estimará la curva de supervivencia para algunas de las covariables con la finalidad de encontrar alguna diferencia notable entre estas. La Figura 4.2 muestra la curva de supervivencia de hombres y mujeres, en la cual puede notar que existe una ligera diferencia entre las curvas. Inicialmente la probabilidad de que un individuo sobreviva más allá de cierto tiempo t dado que el sujeto es hombre, es más alta que la probabilidad de supervivencia de una mujer dentro de los primeros 36 meses (3 años), después de dicho tiempo, la probabilidad de supervivencia de los hombres es más baja en comparación con la probabilidad de supervivencia de las mujeres, pero permaneciendo constante después de estos 3 años, esto es debido a que es más común que un paciente rechace el injerto dentro de los primeros meses de haberse realizado la operación.

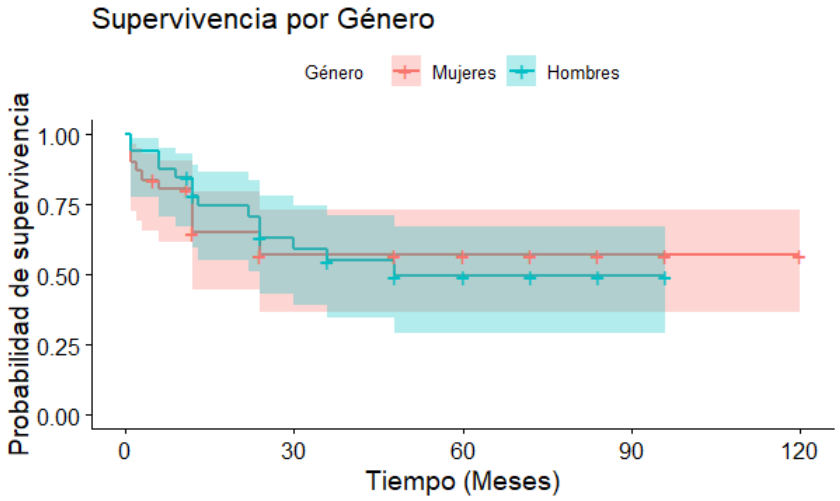


Figura 4.2: Curva de supervivencia Hombres Vs Mujeres con un intervalo de confianza del 95 %.

En la Figura 4.3 puede ver que los grupos de edad presentan una clara diferencia entre sus curvas de supervivencia (en comparación con las curvas

que se generan considerando sólo el género), de hecho, las personas con menor edad (entre $(0, 15]$) presentan un mayor riesgo, ya que su curva de supervivencia está por debajo de la curva de supervivencia de las personas con mayor edad (entre $(15, 40]$).

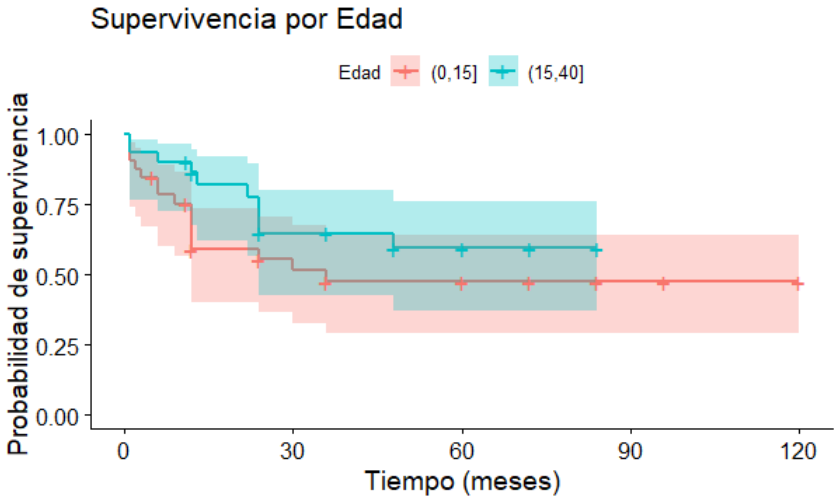


Figura 4.3: Supervivencia por subgrupos de la variable edad con un intervalo de confianza del 95 %.

Ahora, puede ver en la Figura 4.4 las distintas curvas de supervivencia por covariables, en este caso considerando género y edad, y utilizando subgrupos de la covariable edad. Puede ver la diferencia que presentan dichas curvas, ya que pacientes mujeres con menor edad presentan un mayor riesgo que los hombres de menor edad, mientras que los pacientes de mayor edad que son hombres presentan mayor riesgo que las mujeres de mayor edad.

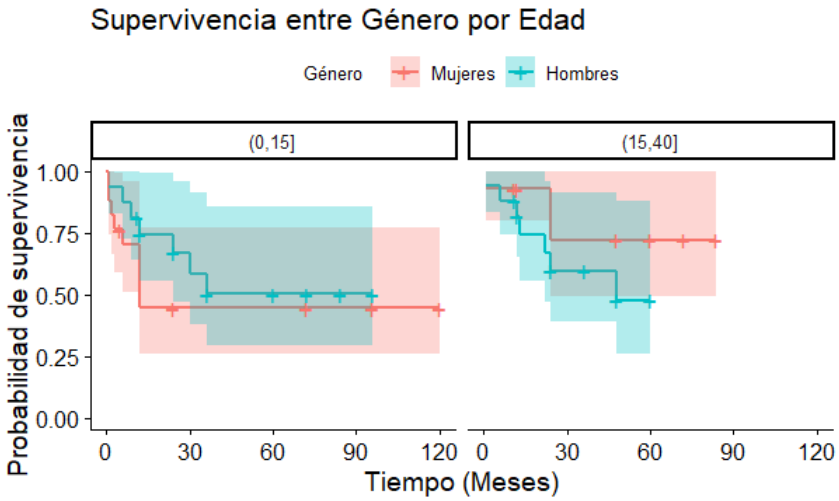


Figura 4.4: Supervivencia por covariables con un intervalo de confianza del 95 %.

4.2. Comparación de modelos paramétricos

La comparación entre los modelos (sin covariables) que se han planteado anteriormente se puede realizar visualmente entre la curva de supervivencia Kaplan-Meier con algunas de las curvas paramétricas dadas en la sección 1.3 o mediante los criterios de información de Akaike (AIC) y/o el criterio de información bayesiano (BIC). Consideramos la base de datos de la Tabla A.1 de registros de pacientes a los que se les ha realizado un trasplante renal, para este conjunto de datos se muestran las distintas curvas de supervivencia generadas por modelos paramétricos, así las curvas para el riesgo acumulado considerando dichos modelos y los valores respectivos para los criterios de información AIC y BIC.

Comparación gráfica

La comparación gráfica es una forma muy útil para verificar qué modelos se ajustan más al conjunto de datos, sin embargo suele ser un tanto subjetivo, de modo que muchas veces además de realizar un análisis gráfico también se acompaña de un análisis analítico.

En la Figura 4.5 puede ver las funciones de supervivencia considerando el modelo Exponencial, Weibull, log-normal y Gamma, junto con el estimador Kaplan-Meier, además de entre estos modelos puede notar que los modelos que mejor se ajustan a la curva KM son el modelo Weibull, log-normal y Gamma.

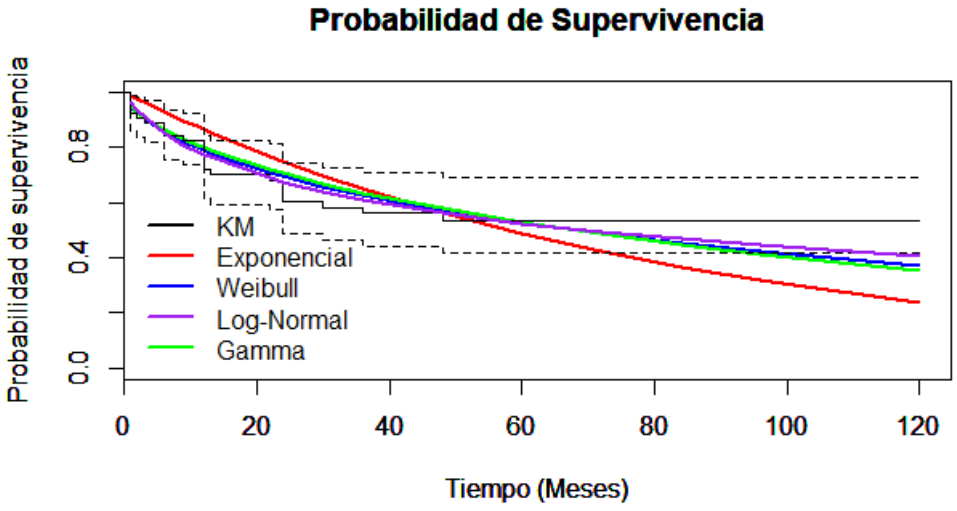


Figura 4.5: Probabilidad de supervivencia de modelos paramétricos.

En la Figura 4.6 puede ver otros modelos paramétricos, tales como: el modelo log-Logístico y el modelo Gompertz, de hecho, puede notar que el modelo de Gompertz es el modelo que mejor se ajusta al conjunto de datos disponible.

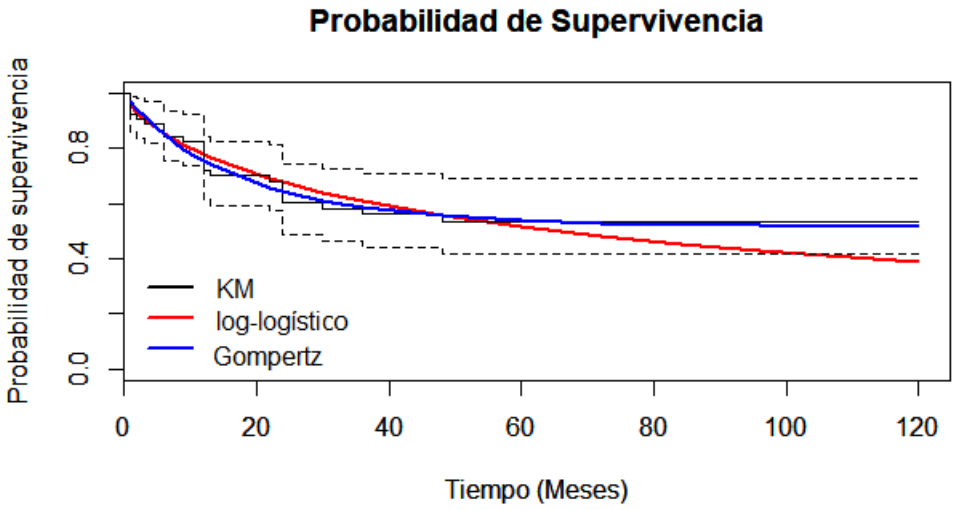


Figura 4.6: Probabilidad de supervivencia de modelos paramétricos.

Finalmente, de las figuras anteriores podemos ver que los modelos que mejor se ajustan al conjunto de datos, son: el modelo Gamma, Weibull, Gompertz y el modelo log-normal, sin embargo, de entre estos modelos se puede resaltar el modelo de Gompertz ya que a lo largo del tiempo es la curva que mejor se acerca a la curva KM.

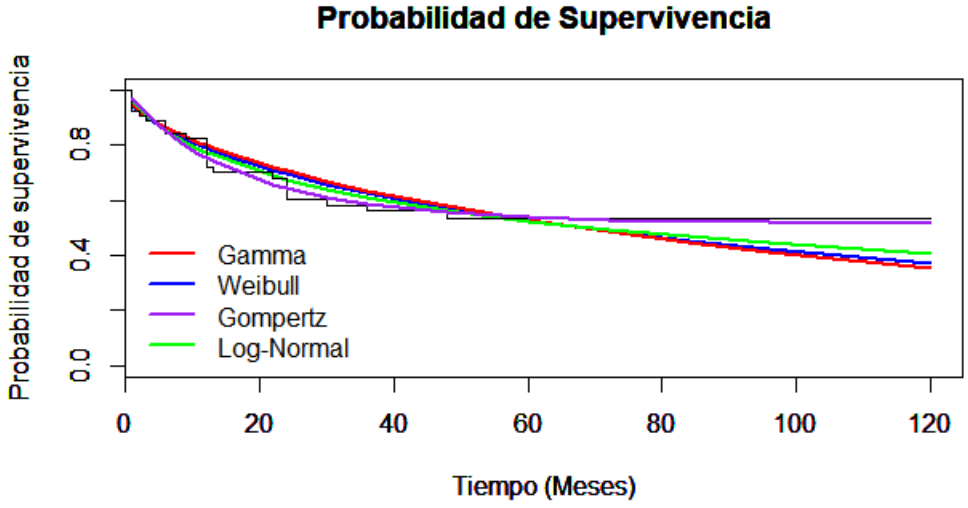


Figura 4.7: Probabilidad de supervivencia para los mejores modelos paramétricos.

Por otra parte, podemos realizar dicho análisis mediante el riesgo acumulativo de estos modelos paramétricos. Así, considerando el estimador de Nelson Aalen se obtienen las Figuras 4.8 y 4.9.

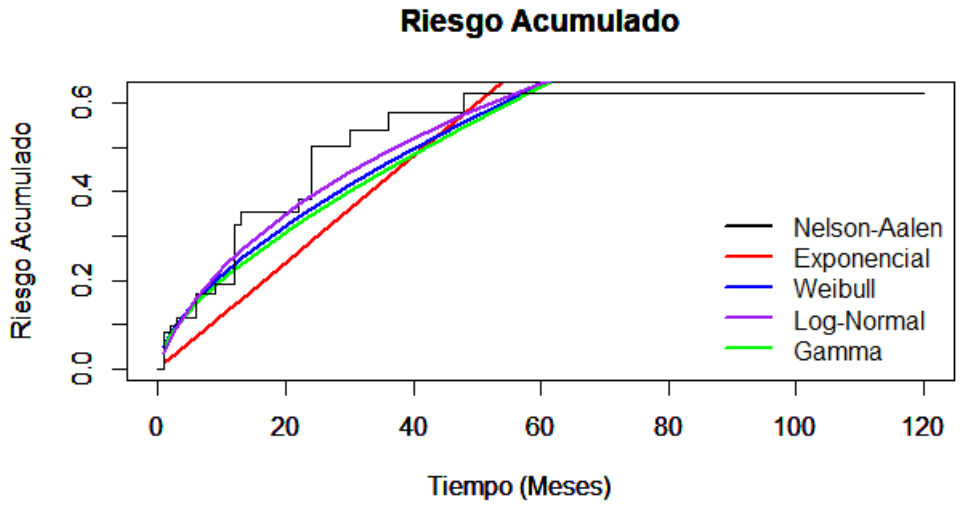


Figura 4.8: Riesgo acumulativo obtenido por el estimador de Nelson-Aalen.

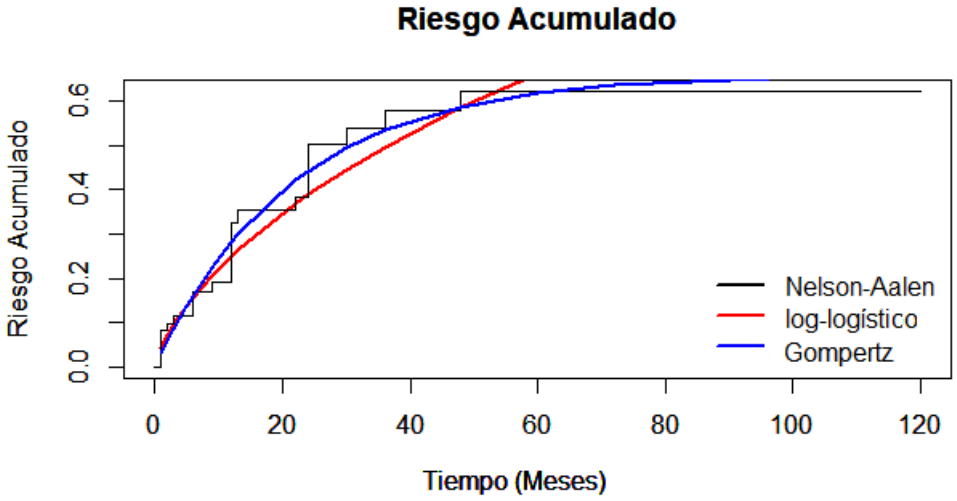


Figura 4.9: Riesgo acumulativo obtenido por el estimador de Nelson-Aalen.

De igual manera, en el caso del riesgo acumulativo las curvas que mejor se ajustan a la curva de Nelson Aalen son los siguientes.

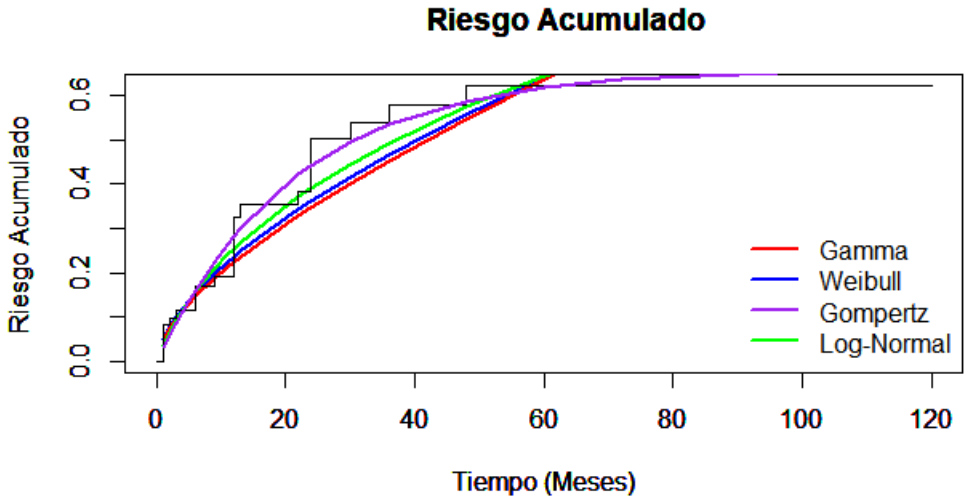


Figura 4.10: Riesgo acumulativo de los modelos mejores ajustados.

Comparación mediante los estadísticos AIC y BIC

Cuando se realiza una comparación entre modelos los criterios de información nos ayudan a escoger un buen modelo para la función de supervivencia de forma más objetiva que el método gráfico, por lo cual se muestra a continuación en la Tabla 4.2 los resultados obtenidos del AIC y el BIC para cada modelo, así como los valores de los estimadores de los mismos.

		AIC	BIC
Estimación obtenida con el modelo Exponencial.	$\widehat{\lambda} = 0.0120$	283.9721	286.131
Estimación obtenida con el modelo Gamma.	$\widehat{\lambda} = 0.0043$ $\widehat{k} = 0.5760$	277.8239	282.1417
Estimación obtenida con el modelo Weibull.	$\widehat{\beta} = 0.6220$ $\widehat{\alpha} = 0.0081$	276.598	280.9158
Estimación obtenida con el modelo log-Normal.	$\widehat{\sigma} = 2.3200$ $\widehat{\mu} = 4.241$	272.9386	277.2564
Estimación obtenida con el modelo log-Logística.	$\widehat{\lambda} = 0.0153$ $\widehat{k} = 0.7450$	274.5733	278.8911
Estimación obtenida con el modelo Gompertz.	$\widehat{\lambda} = 0.0305$ $\widehat{c} = -0.0463$	270.2371	274.5548

Tabla 4.2: Valores del AIC y BIC para distintos modelos paramétricos.

De esta manera, puede ver que aunque los valores AIC y BIC son relativamente semejantes para cada uno de estos modelos, podemos asegurar ya no sólo de manera gráfica sino también de forma analítica que el modelo de la distribución de Gompertz de parámetros $\widehat{\lambda} = 0.0305$ y $\widehat{c} = -0.0463$ es el modelo que mejor se ajusta al conjunto de datos con los que se cuenta, puesto

que los valores de los estadísticos AIC y BIC toman los menores valores de entre todos estos modelos.

Observación 4.3. Note que en este caso, se considera el modelo de Gompertz con $c < 0$, por lo cual; como se vio en la sección 1.3, la función de riesgo es decreciente. Esto refuerza la suposición que se tenía inicialmente: los pacientes que presentaban el rechazo del injerto con mayor frecuencia lo hacen dentro de los primeros meses, estos es, un paciente que conservara el injerto por mayor tiempo tenía menor riesgo de presentar el rechazo en comparación de alguien a quien apenas se le realizara dicha operación. Considerando este valor de c, puede ver la función de riesgo para el modelo de Gompertz en la Figura 4.11 y en Figura 4.12 la función de riesgo con el modelo Weibull.

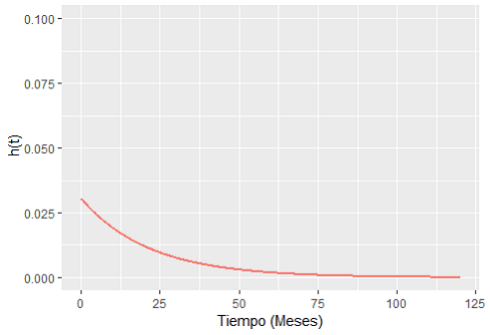


Figura 4.11: Función de Riesgo ajustado a los datos.

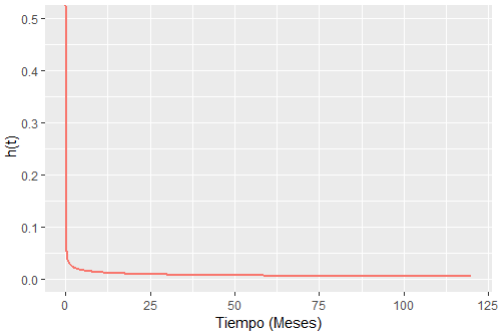


Figura 4.12: Función de Riesgo ajustado a los datos.

Por otra parte, también puede ver la función de riesgo empírico en la Figura 4.13. Aquí puede notar que aún cuando el riesgo asociado al modelo de Gompertz $h(\cdot)$ es el modelo mejor ajustado al conjunto de datos, el riesgo empírico es mucho más alto que $h(\cdot)$, a lo largo de casi todo el tiempo de estudio.

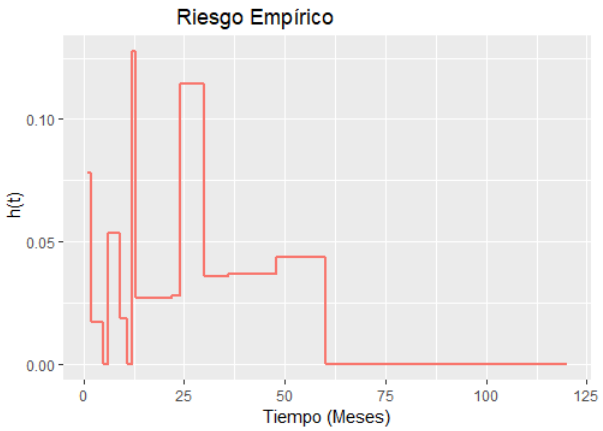


Figura 4.13: Riesgo Empírico de los pacientes en estudio.

4.4. Análisis con el modelo de Cox

Ahora analizaremos el modelo de riesgos proporcionales, en donde además de considerar el tiempo en el que cada paciente estuvo en estudio y si ha sido o no censurado también se consideran covariables para estos pacientes (es decir, aquí se pierde la homogeneidad de los pacientes con insuficiencia renal), estas covariables se describen a continuación.

1. Edad de los pacientes en años cumplidos, en el momento en que entran al estudio.

2. Sexo del paciente, 1 indica que el individuo es hombre y 0 indica que es mujer.

3. Talla del paciente en cm.

4. Creatina Sérica: es un residuo de la masa y actividad muscular. Su nivel en sangre, es el dato más objetivo y fiable para conocer cómo funcionan los riñones, un valor normal es de 0.96 a 1.4 mg/dL, sin embargo los rangos de los valores pueden variar ligeramente entre diferentes laboratorios y niveles superiores a lo normal pueden ser indicio de insuficiencia renal.

5. Tiempo de isquemia caliente: es el intervalo transcurrido, en minutos (o segundos), entre el clampaje de los vasos renales y el enfriamiento del injerto con el líquido de preservación a 4^{o} C. Lo ideal es que ambas cosas se realicen conjuntamente y sea de 0 minutos.

Considerando todas estas covariables para analizar el modelo de Riesgos Proporcionales de Cox, se obtienen los resultados, expuestos en la Tabla 4.3 para los valores de los regresores del modelo.

Covariables	coef	$\exp(coef)$
Edad	-0.1352	0.8735
Sexo	-0.0595	0.9421
Talla	-0.0028	0.9972
Creatina Sérica	0.0159	1.0161
Tiempo de isquemia caliente	-0.0080	0.9920

Tabla 4.3: Valores estimados de los regresores del modelo de Cox.

Con estas covariables y mediante los criterios de información, tenemos que AIC es de 199.5588 y BIC=205.8493. Puede notar que estos valores son mucho menores que los obtenidos con el modelo de Gompertz, donde $AIC = 270.2371$ y $BIC = 274.5548$, por lo cual, se puede ver que el modelo de Riesgos Proporcionales se ajusta aun mejor al conjunto de datos.

Por otra parte, se puede realizar un análisis mediante los valores del estadístico AIC podemos ver que es posible eliminar algunas de las variables del modelo para obtener un modelo más simple, para ello puede ver que si consideramos a cada una de las variables como única covariable del modelo de Cox

se obtienen los siguientes resultados; valor de los regresores, valor AIC y BIC, lo cual se muestra en la Tabla 4.4.

Covariables	coef	exp($coef$)	AIC	BIC
Edad	−0.1024	0.9026	194.8766	196.1346
Sexo	−0.0070	0.9929	199.0692	200.3273
Talla	−0.0213	0.9789	196.3593	197.6174
Creatina Sérica	−0.0030	0.9969	199.0585	200.3165
Tiempo de isquemia caliente	−0.0030	0.9969	199.0585	198.3667

Tabla 4.4: Valor del AIC y BIC del modelo de Cox considerando una sola covariable.

Con lo cual podemos ver que el modelo con la única covariable edad tiene el menor AIC y BIC, indicando que la edad es indispensable en el modelo. Luego, además de considerar la edad del individuo, incluimos otra covariable, obteniendo con ello que un buen modelo con dos covariables es agregando también el tiempo de isquemia ya que obtenemos que AIC=193.8005, el cual es más pequeño que el AIC obtenido considerando sólo la variable edad, además de que el valor BIC=196.3167 también es pequeño, como puede ver en Tabla 4.5.

Covariables	AIC	BIC
Edad+Sexo	196.8689	199.3851
Edad+Talla	196.7268	199.2430
Edad+Creatina Sérica	196.5770	199.0932
Edad+Tiempo de isquemia caliente	193.8005	196.3167

Tabla 4.5: Valores del AIC y BIC con el modelo de Cox considerando dos covariables.

Nuevamente, agregamos una covariable más, obteniendo con ello que un buen modelo con 3 covariables incluye la creatina sérica, pues con ello el valor del AIC es de 195.6117 y BIC=199.3860, los cuales son los valores más pequeños que cualquier otro modelo con 3 covariables. Sin embargo, nótese que los valores AIC y BIC aumentan si se consideran sólo dos covariables, esto puede verlo en la siguiente Tabla 4.6.

Covariables	AIC	BIC
Edad+Tiempo de isquemia caliente+Sexo	195.7880	199.5623
Edad+Tiempo de isquemia caliente+Talla	195.7768	199.5511
Edad+Tiempo de isquemia caliente+Creatina Sérica	195.6117	199.3860

Tabla 4.6: Valores del AIC con el modelo de Cox considerando 3 covariables.

Finalmente, considerando ahora 4 covariables, se obtienen los siguientes resultados expuestos en la Tabla 4.7. En donde puede notar que los valores de los estadísticos AIC y BIC aumentan no tan significativamente, en comparación de un modelo con 3 covariables.

Covariables	AIC	BIC
Edad+Tiempo de isquemia caliente+Creatina+Sexo	197.5792	202.6116
Edad+Tiempo de isquemia caliente+Creatina+Talla	197.5799	202.6123

Tabla 4.7: Valores del AIC y BIC considerando 4 covariables.

Con todo lo anterior puede ver que obtenemos un buen modelo simple considerando sólo dos covariables: edad y tiempo de isquemia caliente. Sin embargo, cuando se desea ver que tanto influyen las covariables en la función de riesgo base para cierto individuo, podemos considerar otras covariables. En este caso, analizaremos el modelo considerando 3 covariables, esto es, incluyendo también la covariable Creatina Sérica, así, puede ver en la Tabla 4.8 los valores estimados de los regresores del modelo.

Covariables	$coef$	$\exp(coef)$
Edad	-0.1444	0.8655
Tiempo de isquemia caliente	-0.0080	0.9919
Creatina Sérica	0.0149	1.0151

Tabla 4.8: Valores de los regresores del modelo de Cox.

Para expresar el riesgo con el modelo de Cox, se denota por; E a la edad, I al tiempo de isquemia caliente y C a la creatina Sérica. Por tanto, el modelo de Cox con 3 covariables puede ser expresado finalmente como:

$$h(t|X) = h_0(t)\exp(-0.1444\ \mathrm{E} - 0.0080\ \mathrm{I} + 0.0149\ \mathrm{C}), \tag{4.1}$$

en donde $h_0(\cdot)$ es la función de riesgo base para cuando el valor de las covariables es 0. Nótese además que la covariable Creatina Sérica produce que la función de riesgo aumente, puesto que $\exp(coef) > 1$, mientras que la variable edad y el tiempo de isquemia hacen que dicha función de riesgo disminuya.

Con este modelo podemos estimar la función de supervivencia (en este caso mediante el uso de R) para distintos valores de las covariables. Consideremos dos pacientes con vector de covariables diferentes, en las Figuras 4.14 y 4.15 puede ver una gran diferencia en las curvas de supervivencia, ya que el paciente con 15 años de edad tiene un mayor riesgo que un paciente de 24 años.

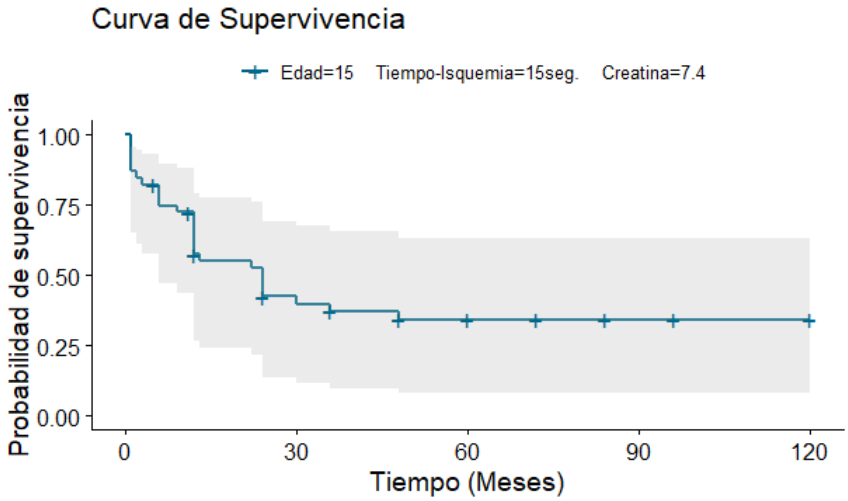


Figura 4.14: Curva estimada del modelo de Cox.

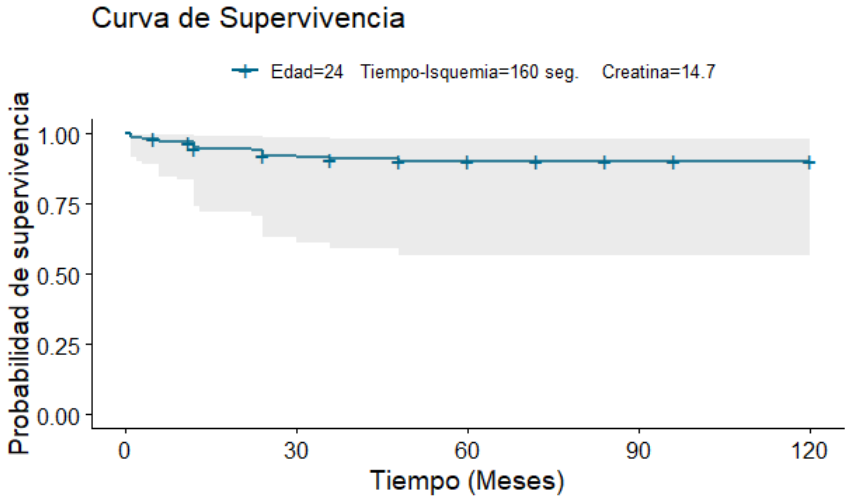


Figura 4.15: Curva estimada del modelo de Cox.

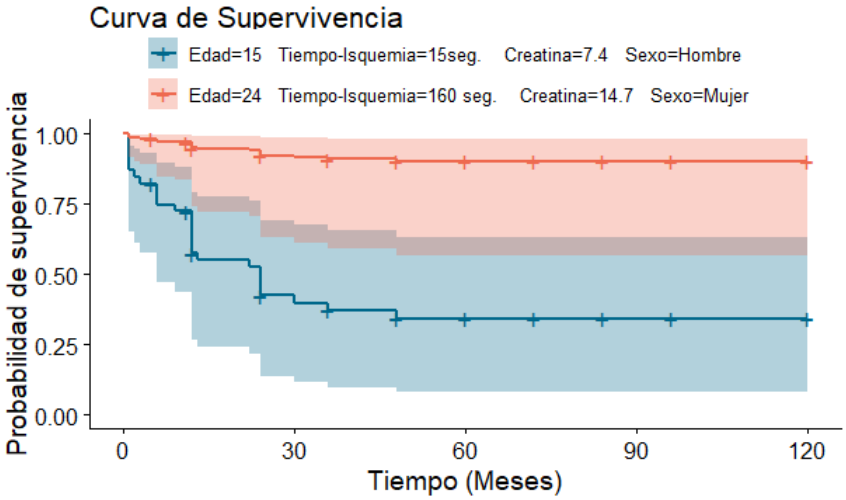


Figura 4.16: Curva estimada del modelo de Cox.

Validación del modelo

A continuación se presenta la validación del modelo para comprobar si las covariables cumplen la hipótesis de riesgos proporcionales, en este caso consideramos el modelo dado en (4.1). Esta validación se hará mediante los residuos de Schoenfeld, para ello haremos uso de la función $cox.zph()$ de R, esta función realiza la prueba de hipótesis de correlación para los residuales Schoenfeld de cada covariable con alguna transformación del tiempo [2]. Para verificar que haya proporcionalidad para una covariable, la curva del estimador de esa covariable en función del tiempo debe ser una línea recta con pendiente cero, es decir, si se asumen riesgos proporcionales, entonces $\beta(t)$ será una línea horizontal. En las Figuras 4.17, 4.18 y 4.19 puede ver las curvas de los estimadores $\beta(\cdot)$ (línea continua) de cada una de las covariables del modelo.

Ahora se comprueba la hipótesis de que los riesgos de dos individuos con covariables diferentes son proporcionales analíticamente mediante la función $cox.zph()$ en R. La salida se muestra en la siguiente Tabla 4.9 en donde para cada covariable se calcula el coeficiente de correlación rho, el estadístico de prueba $chisq$ y su respectivo p-valor.

Covariables	rho	$chisq$	p
Edad	0.0033	0.0002	0.9880
Creatina Sérica	-0.0909	0.0778	0.7800
Tiempo de Isquemia	-0.2039	1.0085	0.3150

Tabla 4.9: Valor del estadístico de prueba y el p-valor.

Nótese que los p-valores obtenidos son mayores que $\alpha = 0.05$ lo cual nos indica es que no hay suficientes pruebas para rechazar la hipótesis nula de proporcionalidad de los riesgos, es decir, podemos asumir que cada una de las covariables consideradas en el modelo cumplen con dicha hipótesis.

Observación 4.5. Aun cuando ya se ha hecho un análisis con el modelo de Cox, es necesario saber si dicho modelo es adecuado para realizar el análisis de los pacientes, por lo cual, en esta subsección se trato de verificar que las covariables cumplieran la proporcionalidad del modelo, así, se realizo un análisis de correlación entre los distintos tiempo de falla y los residuos de Schoenfeld. De este modo, al obtener correlaciones bastante cercanas a cero, podremos asumir que en efecto la proporcionalidad se da y por tanto el modelo de Cox es adecuado para realizar un análisis descriptivo del caso de estudio.

Sin embargo, no se estudio a profundidad este método de validación, ya que no se encontró con diversas fuentes de información que permitieran analizar con mayor detalle dicho método.

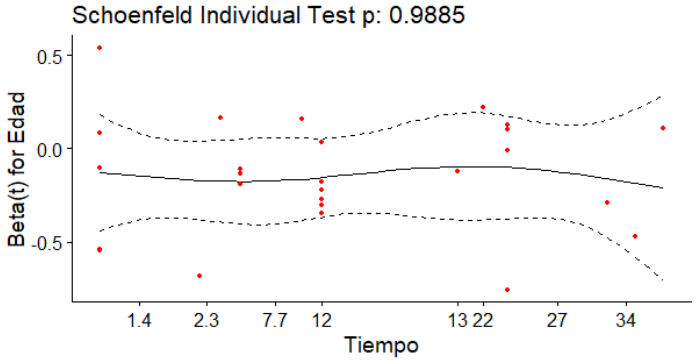


Figura 4.17: Correlación entre los tiempos de falla y los residuos de la covariable Edad.

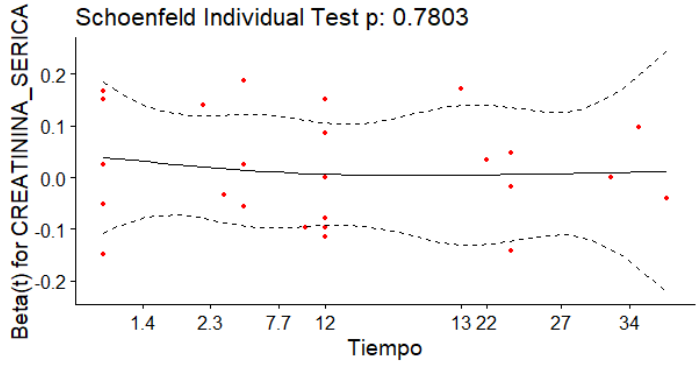


Figura 4.18: Correlación entre los tiempos de falla y los residuos de la covariable Creatina.

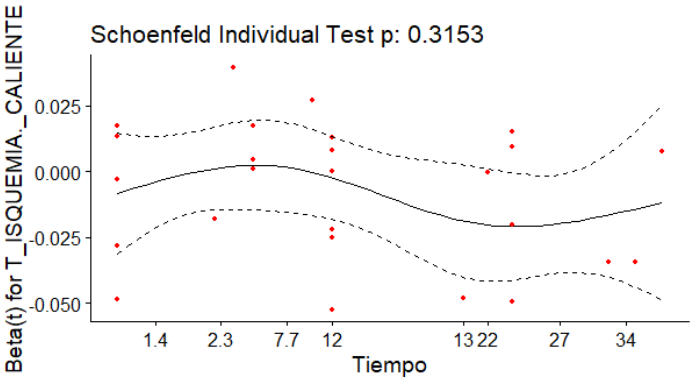


Figura 4.19: Correlación entre los tiempos de falla y los residuos de la covariable Tiempo de Isquemia.

Capítulo 5

ANÁLISIS ESTADÍSTICO DE PUNTO DE CAMBIO

En la actualidad, el análisis estadístico de punto de cambio es de creciente estudio, ya que existen diversos problemas en donde se presenta el cambio, y conocer acerca de estos ayuda a tratar de evitar pérdidas o para aprovechar las transiciones. Algunas de las disciplinas en donde podemos encontrar estos problemas son principalmente; en Control de calidad, Medicina, Economía y Finanzas.

En Estadística, un punto de cambio es el momento o instante de tiempo en el cual las observaciones cambian de distribución, es decir, antes de este momento dichas observaciones siguen una distribución y posteriormente a este tiempo siguen otra distribución. En este análisis existen dos problemas principales, uno es verificar que en efecto dicho punto existe (lo cual se realiza mediante un contrastes de hipótesis), el otro, es cuando se trata de localizar dicho punto cuando se sabe que existe, el cual es analizado como un problema de estimación ([1], [16])

Algunos ejemplos en donde se puede ver la presencia de puntos de cambio se plantean a continuación.

i. Análisis del mercado de valores. El precio de las acciones de cualquier compañía fluctúan a diario, sin embargo, aunque esta fluctuación es normal, de acuerdo con la teoría económica existen cambios anormales que son necesarios de analizar con atención.

ii. Control de calidad. En procesos de producción continuo, por diversos motivos en el proceso se pueden producir productos de menor calidad con el tiempo, por lo cual es conveniente verificar si existe un punto en donde la calidad de los productos empiecen a deteriorarse.

iii. Tasa de mortalidad en el tráfico. Si existe un incremento en el límite de velocidad, ¿se puede producir algún problema en las carreteras?, esto es, ¿puede este cambio de velocidad producir un cambio en el porcentaje de muertes por accidentes de tráfico?.

5.1. Planteamiento del problema

Considere $X_1, X_2, \ldots, X_n$ variables aleatorias independientes con función de distribución $F_1, F_2, \ldots, F_n$ respectivamente. El problema general de punto de cambio es la realización de una prueba de hipótesis, donde la hipótesis nula es:

$$H_0 : \ F_1 = F_2 = \cdots = F_n$$

contra la hipótesis alternativa:

$$H_a : \ F_1 = \cdots = F_k \neq F_{k+1} = \cdots = F_{k_2} \neq F_{k_2+1} = \cdots = F_{k_q} \neq F_{k_q+1} = \cdots = F_n,$$

donde $1 < k_1 < k_2 < \cdots < k_q < n$, q es el número de cambios y $k_1, k_2, \cdots, k_q$ son sus respectivas posiciones que han de ser estimadas.

Por otra parte, nótese que si las distribuciones $F_1, F_2, \cdots, F_n$ son miembros de alguna familia paramétrica $F(\theta)$, $\theta \in \mathbb{R}^p$, entonces el problema se reduce a un contraste de hipótesis que involucra los parámetros del modelo paramétrico, θ_i con $i = 1, 2, \cdots, n$. De modo que:

$$H_0 : \ \theta_1 = \theta_2 = \cdots = \theta_n = \theta, \tag{5.1}$$

con θ desconocido. Contra la hipótesis alternativa,

$$H_a : \ \theta_1 = \cdots = \theta_k \neq \theta_{k+1} = \cdots = \theta_{k_2} \neq \theta_{k_2+1} = \cdots = \theta_{k_q} \neq \theta_{k_q+1} = \cdots = \theta_n,$$

donde q y $k_1, k_2, \cdots k_q$ representan el número y las posiciones de los puntos de cambios respectivamente que han ser estimados. Obsérvese además que este problema así planteado refleja los aspectos más importantes de la inferencia del punto de cambio; la existencia, posición y localización de dichos puntos cuando se presentan.

En la inferencia de puntos de cambio los métodos que comúnmente son utilizados son; la prueba de razón de verosimilitud, prueba bayesiana, pruebas no paramétricas, enfoque de la teoría de información, etc. Aquí abordaremos este tema con pruebas de razón de verosimilitud.

Para la detección del número de puntos de cambio y sus ubicaciones en un proceso aleatorio multidimensional, Vostrikova (1981) propuso un método, llamado procedimiento de segmentación binaria, el cual es frecuentemente utilizada en la detección de múltiples puntos de cambio, ya que presenta la gran ventaja de que con este procedimiento se detecta la cantidad de dichos puntos y sus posiciones simultáneamente, por lo cual se vuelve un método bastante práctico por el ahorro de tiempo al realizar los cálculos. Este procedimiento de segmentación binaria se realiza mediante los siguientes pasos [1].

Paso 1. Se realiza una prueba o tes. Se considera H_0 como la hipótesis nula como en (5.1), es decir, no existe algún punto de cambio, contra la hipótesis alternativa H_a de la existencia de un punto de cambio:

$$H_a : \theta_1 = \theta_2 = \cdots = \theta_k \neq \theta_{k+1} = \cdots = \theta_n. \tag{5.2}$$

en esta etapa, k es la localización del único punto de cambio. Luego, si la hipótesis nula (5.1) no es rechazada entonces el proceso es detenido, si por el contrario (5.1) se rechaza, entonces hay un punto de cambio y se continúa con el siguiente paso.

Paso 2. Realizar un juego de hipótesis para las dos sucesiones que se generan en el paso 1 por separado para verificar la existencia de punto de cambio.

Paso 3. Este proceso continúa hasta que ya no se obtengan subsucesiones donde pueda haber un cambio.

Finalmente la colección de localización de puntos de cambios obtenidas en los pasos descritos anteriormente lo denotamos por $\{\widehat{k}_1, \widehat{k}_2, \ldots, \widehat{k}_q\}$, donde q es el número de puntos de cambio.

5.2. Modelo Normal Univariado

En muchos casos prácticos es común abordar el tema de puntos de cambio con modelos normales univariados y multivariados, sin embargo dicho análisis también se realiza con otros modelos paramétricos como el caso de la Exponencial y el modelo Gamma [1]. A continuación se describe el método de segmentación binaria con el modelo normal univariado, considerando primero el caso de la existencia de punto de cambio en la media y posteriormente en la varianza.

5.2.1. Punto de cambio en la media

Considere $X_1, X_2, \ldots, X_n$ una muestra aleatoria de la distribución normal, con parámetros $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \ldots, (\mu_n, \sigma_n^2)$ respectivamente. Considere primero el caso en que la varianza es común, por lo cual, el interés es verificar la existencia de posibles puntos de cambio en la media μ. Es decir, se tiene el siguiente juego de hipótesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \mu, \tag{5.3}$$

contra la hipótesis alternativa.

$$H_a : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \tag{5.4}$$

k es la posición del único punto de cambio.

El método que se utilizará para este juego de hipótesis es el método de razón de verosimilitud, analizando el caso en el que la varianza es conocida y el caso en el que es desconocida.

Varianza conocida

Sea $X_1, X_2, \ldots, X_n$, una muestra aleatoria de la distribución normal y sea $x_1, x_2, \ldots, x_n$ una realización de esta muestra. Sin pérdida de generalidad supóngase que $\sigma^2 = 1$, de modo que bajo la hipótesis nula (5.3) la función de verosimilitud está dada por:

$$L_0(\mu) = \frac{1}{(\sqrt{2\pi})^n} \exp\left(\frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2} \right).$$

De donde, la función de log-verosimilitud es,

$$l(\mu) = -\frac{n}{2}\log 2\pi - \sum_{i=1}^{n} \frac{(x_i - \mu)}{2}.$$

Así, el estimador de máxima verosimilitud para la media es

$$\widehat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{5.5}$$

Bajo la hipótesis alternativa,

$$L_1(\mu_1, \mu_n) = \frac{1}{(\sqrt{2\pi})^n} \exp\left(- \frac{\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2}{2} \right),$$

luego,

$$l(\mu_1, \mu_2) = \log L_1(\mu_1, \mu_n) = -\frac{n}{2}\log 2\pi - \frac{\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2}{2}$$

de donde los estimadores de máxima verosimilitud son

$$\widehat{\mu}_1 = \overline{X}_k = \frac{1}{k}\sum_{i=1}^{k} x_i \quad y \quad \widehat{\mu}_n = \overline{X}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} x_i$$

respectivamente. Sean,

$$\begin{aligned} S_k &= \sum_{i=1}^{k}\left(x_i - \overline{X}_k\right)^2 + \sum_{i=k+1}^{n}\left(x_i - \overline{X}_{n-k}\right)^2 \\ V_k &= k\left(\overline{X}_k - \overline{X}\right)^2 + (n-k)\left(\overline{X}_{n-k} - \overline{X}\right)^2 \\ S &= \sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \end{aligned}$$

nótese además que:

$$S - S_k = \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 - \sum_{i=1}^{k} \left(x_i - \overline{X}_k\right)^2 - \sum_{i=k+1}^{n} \left(x_i - \overline{X}_{n-k}\right)^2, \qquad (5.6)$$

donde:

$$\begin{aligned}
\sum_{i=1}^{k} \left(x_i - \overline{X}_k\right)^2 &= \sum_{i=1}^{k} \left\{ \left(\overline{X} - \overline{X}_k\right)^2 - 2\left(\overline{X} - \overline{X}_k\right)\left(\overline{X} - x_i\right) + \left(x_i - \overline{X}\right)^2 \right\} \\
&= k\left(\overline{X} - \overline{X}_k\right)^2 - 2k(\overline{X} - \overline{X}_k)^2 + \sum_{i=1}^{k}(x_i - \overline{X})^2 \\
&= -k\left(\overline{X} - \overline{X}_k\right)^2 + \sum_{i=1}^{k}(x_i - \overline{X})^2,
\end{aligned}$$

y

$$\begin{aligned}
\sum_{i=k+1}^{n} \left(x_i - \overline{X}_{n-k}\right)^2 &= \sum_{i=k+1}^{n} \left\{ \left(x_i - \overline{X}\right)^2 - 2\left(\overline{X} - x_i\right)\left(\overline{X} - \overline{X}_{n-k}\right) \right. \\
&\quad \left. + \left(\overline{X} - \overline{X}_{n-k}\right)^2 \right\} \\
&= \sum_{i=k+1}^{n} \left(x_i - \overline{X}\right)^2 - 2(n-k)\left(\overline{X} - \overline{X}_{n-k}\right) \\
&\quad + (n-k)\left(\overline{X} - \overline{X}_{n-k}\right)^2 \\
&= \sum_{i=k+1}^{n} \left(x_i - \overline{X}\right)^2 - (n-k)\left(\overline{X} - \overline{X}_{n-k}\right).
\end{aligned}$$

Así, es claro que $V_k = S - S_k$. Por otra parte, la estadística λ es,

$$\lambda = \frac{L_{\overline{\Theta}_0}}{L_{\overline{\Theta}}} = \frac{\exp\left(-\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{2}\right)}{\exp\left(-\frac{\sum_{i=1}^{k}\left(x_i - \overline{X}_k\right)^2 + \sum_{i=k+1}^{n}\left(x_i - \overline{X}_{n-k}\right)}{2}\right)} = \exp\left(-\frac{1}{2}V_k\right)$$

que es una función decreciente en V_k. Por lo tanto podemos basar la prueba en la estadística,

$$V_{k^*} = \max_{2 \le k \le n-1} V_k. \qquad (5.7)$$

La expresión (5.7) indica que para encontrar la estimación para el punto de cambio, se hace variar k de 2 hasta $n-1$, así, donde se alcanza el valor máximo, se toma como la posición estimada del punto de cambio.

Varianza desconocida

Ahora consideremos el problema de punto de cambio cuando la varianza es desconocida pero común en la muestra. Sea $X_1, X_2, \ldots, X_n$, una muestra de *v.a.'s* con distribución normal, de parámetros $(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_n, \sigma_n)$ y sea $x_1, x_2, \ldots, x_n$ una realización de esta muestra. Además consideramos que $\sigma_1 = \sigma_2 = \cdots = \sigma_n$.

Bajo la hipótesis nula la función de verosimilitud es,

$$L_0(\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(\frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} \right).$$

Así, los *e.m.v.* para μ y σ^2 son

$$\widehat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{y} \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2,$$

respectivamente. Luego, bajo la hipótesis al alternativa obtenemos,

$$L_1(\mu_1, \mu_n, \sigma_1^2) = \frac{1}{(\sqrt{2\pi\sigma_1^2})^n} \exp\left(- \frac{\sum_{i=1}^{k}(x_i - \mu_1)^2 + \sum_{i=k+1}^{n}(x_i - \mu_n)^2}{2\sigma_1^2} \right),$$

de modo que los estimadores de máxima verosimilitud para μ_1 y μ_n son

$$\widehat{\mu}_1 = \overline{X}_k = \frac{1}{k}\sum_{i=1}^{k} x_i \quad y \quad \widehat{\mu}_n = \overline{X}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} x_i$$

respectivamente, mientras que para σ_1^2, se tiene que:

$$\widehat{\sigma_1}^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(x_i - \overline{X}_k)^2 + \sum_{i=k+1}^{n}(x_i - \overline{X}_{n-k})^2 \right).$$

Del procedimiento de la razón de verosimilitud se tiene que el estadístico de prueba está dado por:

$$V = \max_{1 \le k \le n-1} \frac{|T_k|}{S},$$

donde

$$S = \sum_{i=1}^{n}(x_i - \overline{X})^2 \qquad \text{y} \qquad T_k^2 = \frac{k(n-k)}{n}(\overline{X}_k - \overline{X}_{n-k})^2.$$

Este estimador al igual que el estimador definido en (5.7) proporciona un estimador para la posición del punto de cambio. Además, en 1979, la distribución de V fue propuesta por Worsley [19].

5.2.2. Punto de cambio en la varianza

Se analizó anteriormente el problema de punto de cambio en la media considerando el caso en el que la varianza es conocida y cuando es desconocida. Presentamos ahora el análisis de punto de cambio en la varianza cuando la media es conocida.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria, independientes y normalmente distribuidas, con parámetros $(\mu, \sigma^2), (\mu, \sigma_2^2), \ldots, (\mu, \sigma_n^2)$ respectivamente. Además, suponga que μ es conocida. Luego, bajo estas condiciones tenemos el juego de hipótesis siguiente,

$$H_0: \ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2, \tag{5.8}$$

contra la hipótesis alternativa

$$\begin{aligned} H_a: \ \sigma_1^2 \quad &= \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_{k_2}^2 \neq \sigma_{k_2+1} = \cdots \\ &= \sigma_{k_q}^2 \neq \sigma_{k_q+1}^2 = \cdots = \sigma_n^2 \end{aligned}$$

q es número de puntos de cambios y $k_1 < k_2 < \cdots < k_q$ son sus respectivas posiciones desconocidas de cada uno de ellos.

Con el método o procedimiento de segmentación binaria el contraste de hipótesis se reduce a probar (5.8) contra la hipótesis alternativa,

$$H_a: \ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_n^2$$

k representa el punto de cambio. Así, considerando este juego de hipótesis y bajo la hipótesis nula, la función de verosimilitud es,

$$L_0(\sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(\frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} \right),$$

de modo que puede verificar fácilmente que el estimador de máxima verosimilitud para σ^2 es,

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}.$$

Bajo H_a, la función de verosimilitud es,

$$L_1(\sigma_1^2, \sigma_n^2) = \frac{1}{\left(\sqrt{2\pi\sigma_1^2}\right)^k} \frac{1}{\left(\sqrt{2\pi\sigma_n^2}\right)^{n-k}} \exp\left(-\frac{\sum_{i=1}^{k}(x_i - \mu)^2}{2\sigma_1^2} - \frac{\sum_{i=k+1}^{n}(x_i - \mu)^2}{2\sigma_n^2} \right).$$

Así, los *e.m.v.* para σ_1^2 y σ_2^2 son,

$$\widehat{\sigma_1}^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2}{k} \qquad y \qquad \widehat{\sigma_n}^2 = \frac{\sum_{i=k+1}^{n}(x_i - \mu)^2}{n-k},$$

respectivamente. Finalmente obtenemos que el estadístico de prueba basado en el cociente de verosimilitud es,

$$\lambda_n = \left(\max_{1<k<n-1} \left[n - \log \widehat{\sigma}^2 - k \log \sigma_1^2 - (n-k) \log \sigma_n^2 \right] \right)^{1/2}. \qquad (5.9)$$

Con esto puede notar que para obtener los *e.m.v.* es necesario conocer la posición del punto de cambio para $1 < k < n-1$. Además, el estadístico proporciona de forma implícita un estimador para el punto de cambio, el cual es denotado por $\widehat{k}$ y es tal que (5.9) alcanza el máximo en $\widehat{k}$. La distribución asintótica del estimador puede verlo en [1] y [16].

5.2.3. Punto de cambio en la media y varianza

Consideramos ahora abordar el problema de punto de cambio múltiple, esto es, el cambio en la media y en la varianza. Así, sea $X_1, X_2, \ldots, X_n$ una muestra de *v.a.'s* independientes, con distribución normal y de parámetros $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \ldots, (\mu_1, \sigma_n^2)$, respectivamente. Para el contraste de hipótesis consideramos ahora, la hipótesis nula

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_n = \mu \ \ y \ \ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2, \qquad (5.10)$$

donde μ y σ^2 son desconocidos. Contra,

$$\begin{aligned} H_a: \mu_1 \quad & = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_{k_2} \neq \mu_{k_2+1} = \cdots \\ & = \mu_q \neq \mu_{q+1} = \cdots = \mu_n \\ & \quad \text{y} \\ \sigma_1^2 \quad & = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_{k_2}^2 \neq \sigma_{k_2+1} = \cdots \\ & = \sigma_{k_q}^2 \neq \sigma_{k_q+1}^2 = \cdots = \sigma_n^2 \end{aligned}$$

en donde q es el número de cambios. Por otro lado, haciendo uso del procedimiento de segmentación binaria, se tiene tiene el contraste de (5.10), contra la hipótesis alternativa,

$$\begin{aligned} H_a: \ & \sigma_1^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_n^2 \\ & y \\ & \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n. \end{aligned}$$

Bajo la hipótesis nula la función verosimilitud está dada,

$$L_0(\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(\frac{-\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \right),$$

Los estimadores de máxima verosimilitud son entonces,

$$\widehat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{y} \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2,$$

para μ y σ^2 respectivamente. Por otro lado, la función de verosimilitud para la hipótesis alternativa es,

$$L_1(\mu_1, \mu_n, \sigma_1^2, \sigma_n^2) = \frac{1}{\left(\sqrt{2\pi\sigma_1^2}\right)^k} \frac{1}{\left(\sqrt{2\pi\sigma_n^2}\right)^{n-k}} \exp\left(-\sum_{i=1}^{k}\frac{(x_i-\mu_1)^2}{2\sigma_1^2} - \sum_{i=k+1}^{n}\frac{(x_i-\mu_n)^2}{2\sigma_n^2}\right).$$

Los estimadores de verosimilitud para μ_1, μ_n, σ_1^2 y σ_n^2 son,

$$\widehat{\mu}_1 = \overline{X}_k = \frac{1}{k}\sum_{i=1}^{k} x_i, \qquad \widehat{\mu}_n = \overline{X}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n}(x_i - \overline{X}_{n-k}),$$

$$\widehat{\sigma_1}^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{X}_k)^2 \qquad y \qquad \widehat{\sigma_n}^2 = \frac{1}{n-k}\sum_{i=k+1}^{n}(x_i - \overline{X}_{n-k})^2,$$

respectivamente. Por último, del procedimiento de razón de verosimilitud, el estadístico de prueba es,

$$\Delta_n = \max_{1<k<n-1} \frac{(\widehat{\sigma})^n}{(\widehat{\sigma}_1)^k\,(\widehat{\sigma}_n)^{n-k}}. \tag{5.11}$$

Observación 5.3. En 1993 Horváth deriva la distribución asintótica del estadístico definido en (5.11) bajo la hipótesis nula [7].

5.4. Estimación de la función de riesgo con un punto de cambio

El problema de punto de cambio en la función de distribución ha sido muy estudiado en diversos contextos, tales como en fiabilidad y en control de calidad. Sin embargo, este problema es también extendido dentro del Análisis de Supervivencia, aquí, lo que nos interesa es el estudio de punto de cambio en el modelo de riesgo. Por ejemplo, considere $X_1, X_2, \ldots, X_n$ *v.a.'s* de tipo Exponencial, nos interesa entonces, cuando la función de riesgo permanece a una tasa constante, digamos a, para $0 \leq t \leq \tau$ y luego baja a una tasa b para $t > \tau$, de modo que $a > b \geq 0$. Formalmente tratamos de estimar,

$$h(t) = \begin{cases} a, & \text{si } 0 \leq t \leq \tau, \\ b, & \text{si } t > \tau, \end{cases}$$

y donde τ representa el punto de cambio a estimar.

A continuación se muestra cómo obtener una estimación para el punto de cambio en el modelo Exponencial y en el modelo de Gompertz mediante el criterio informacional [1].

5.4.1. Enfoque informacional con el modelo Exponencial

A continuación se muestra una forma de obtener una estimación para τ mediante el uso del criterio de información Bayesiano o criterio de Schwarz (BIC), considerando el modelo Exponencial. Consideremos una *v.a.* T de tiempo de vida, cuya función de riesgo está dado por:

$$h(t) = \begin{cases} a, & \text{si } 0 \leq t \leq \tau, \\ b, & \text{si } t > \tau, \end{cases}$$

donde $a > b \geq 0$. Con esto, la función de densidad está dado por

$$f(t; a, b, \tau) = \begin{cases} ae^{-at}, & \text{si } 0 \leq t \leq \tau, \\ be^{-bt-(a-b)\tau}, & \text{si } t > \tau. \end{cases} \tag{5.12}$$

Ahora, considere una muestra aleatoria $T_1, T_2, \ldots, T_n$ de tiempos de vida y $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ los estadísticos de orden de esta muestra y considere el punto τ_0 tal que $T_{(k)} \leq \tau_0 < T_{(k+1)}$, para algún $k = 1, 2, ..., n-1$. Así, podemos considerar este valor en τ_0 si es posible determinar tal k y usarla para estimar el valor de τ.

Luego, la función de verosimilitud de $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, considerando la realización $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, está dada:

$$\begin{aligned} L(a,b,k) &= L(a,b,k|t_{(1)},t_{(2)},\ldots,t_{(n)}) = f_{T_{(1)},T_{(2)},\ldots,T_{(n)}}(t_{(1)},t_{(2)},\ldots,t_{(n)}) \\ &= n!\left(\prod_{i=1}^{k} f(t_{(i)})\right)\left(\prod_{j=k+1}^{n} f(t_{(j)})\right) \\ &= n!a^k\, e^{-a\sum_{i=1}^{k} t_{(i)}}\, b^{n-k}\, e^{-(n-k)a\tau_0 - b\sum_{j=k+1}^{n}(t_{(j)}-\tau_0)}. \end{aligned}$$

De modo que la función log-verosimilitud es:

$$\begin{aligned} l(a,b,k) &= \log L(a,b,\tau_0) \\ &= \log n! + \log(a^k b^{n-k}) - a\sum_{i=1}^{k} t_{(i)} - (n-k)a\tau_0 - b\sum_{j=k+1}^{n}(t_{(j)}-\tau_0) \\ &= \log n! + k\log a + (n-k)\log b - a\left(\sum_{i=1}^{k} t_{(i)} + (n-k)\tau_0\right) \\ &\quad -b\sum_{j=k+1}^{n}(t_{(j)}-\tau_0). \end{aligned}$$

De esta última igualdad se puede ver fácilmente que los estimadores de máxima verosimilitud de a, b, son:

$$\widehat{a} = \frac{k}{(n-k)\tau_0 + \sum_{i=1}^{k} t_{(i)}},$$

y

$$\widehat{b} = \frac{n-k}{\sum_{j=k+1}^{n} t_{(j)} - (n-k)\tau_0},$$

respectivamente. Ahora definimos el BIC para el modelo de la función de riesgo con un punto de cambio del tipo Exponencial como:

$$\begin{aligned} BIC(k,\tau_0) &= -2l(\widehat{a},\widehat{b},\tau_0) + 3\log n \\ &= -2(n-k)\log\left(\frac{n-k}{\sum_{j=k+1}^{n} t_{(j)} - (n-k)\tau_0}\right) \\ &\quad -2k\log\left(\frac{k}{(n-k)\tau_0 + \sum_{i=1}^{k} t_{(i)}}\right) - 2\log n! + 3\log n + 2n, \end{aligned}$$

considerando una selección empírica de τ_0. Luego, se procede a determinar el valor de $\widehat{k}$ tal que minimice $BIC(k,\tau_0)$, es decir,

$$BIC\left(\widehat{k},\tau_0\right) = \min_{1\le k\le n-1} BIC(k,\tau_0). \tag{5.13}$$

Así, el verdadero punto de cambio τ_0 en la función de riesgo está estimado por:

$$\widehat{\tau} = \frac{\widehat{k} \sum\limits_{j=\widehat{k}+1}^{n} t_{(j)} - \left(n - \widehat{k}\right) \sum\limits_{i=1}^{\widehat{k}} t_{(i)}}{n\left(n - \widehat{k}\right)},$$

el cual se obtiene de resolver

$$\left.\frac{\partial BIC\left(\widehat{k}; \tau_0\right)}{\partial \tau_0}\right|_{\tau_0 = \widehat{\tau}} = 0. \tag{5.14}$$

Observación 5.5. El estimador obtenido en (5.14) es óptimo según el principio de selección del modelo, ya que hay que recordar que bajo este principio el modelo que se ajusta mejor al conjunto de datos es el que tiene menor valor en el BIC.

5.5.1. Enfoque informacional con el modelo Gompertz

De manera análoga al modelo exponencial, se muestra una forma de obtener una estimación para τ mediante el uso del estadístico BIC, ahora considerando el modelo de Gompertz. Consideremos una *v.a.* T de tiempo de vida, con función de riesgo,

$$h(t) = \begin{cases} \lambda_1 \, e^{ct}, & \text{si } 0 \leq t \leq \tau, \\ \lambda_2 \, e^{ct}, & \text{si } t > \tau, \end{cases} \tag{5.15}$$

con $c > 0$ una constante fija y conocida, con $\lambda_2, \lambda_1 > 0$. De modo que la función de densidad y función de distribución considerando esta variable aleatoria son

$$f(t) = \begin{cases} \lambda_1 \, e^{ct} \, e^{\frac{-\lambda_1}{c}(e^{ct}-1)}, & \text{si } 0 \leq t \leq \tau, \\ \lambda_2 \, e^{ct} \, e^{\frac{-\lambda_2}{c}(e^{ct}-1)} \, e^{\frac{\lambda_2-\lambda_1}{c}(e^{c\tau}-1)}, & \text{si } t > \tau, \end{cases}$$

y

$$F(t) = \begin{cases} 1 - e^{\frac{-\lambda_1}{c}(e^{ct}-1)}, & \text{si } 0 \leq t \leq \tau, \\ 1 - e^{\frac{-\lambda_2}{c}(e^{ct}-1)} \, e^{\frac{\lambda_2-\lambda_1}{c}(e^{c\tau}-1)}, & \text{si } t > \tau, \end{cases}$$

respectivamente. Además,

$$S(t) = \begin{cases} e^{\frac{-\lambda_1}{c}(e^{ct}-1)}, & \text{si } 0 \leq t \leq \tau, \\ e^{\frac{-\lambda_2}{c}(e^{ct}-1)} \, e^{\frac{\lambda_2-\lambda_1}{c}(e^{c\tau}-1)}, & \text{si } t > \tau. \end{cases} \tag{5.16}$$

Considere una muestra aleatoria $T_1, T_2, ..., T_n$ de tiempos de vida y $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ los estadísticos de orden de esta muestra, τ_0 un punto tal que $T_{(k)} \leq \tau_0 < T_{(k+1)}$, para algún $k = 1, 2, ..., n-1$. Consideremos τ_0 para ver si es posible determinar tal k y con ello estimar el valor de τ.

La función de verosimilitud de $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, considerando la realización $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, está dada:

$$\begin{aligned}
L(\lambda_1, \lambda_2, k) &= L\left(\lambda_1, \lambda_2, k | t_{(1)}, t_{(2)}, \ldots, t_{(n)}\right) = f_{T_{(1)}, T_{(2)}, \ldots, T_{(n)}}\left(t_{(1)}, t_{(2)}, \ldots, t_{(n)}\right) \\
&= n! \left\{ \prod_{i=1}^{k} f(t_{(i)}) \right\} \left\{ \prod_{j=k+1}^{n} f(t_{(j)}) \right\} \\
&= n! \prod_{i=1}^{k} \lambda_1 \, e^{ct_{(i)}} \, e^{\frac{-\lambda_1}{c} \, (e^{ct_{(i)}} - 1)} \\
&\quad \prod_{j=k+1}^{n} \lambda_2 \, e^{ct_{(j)}} \, e^{\frac{-\lambda_2}{c}(e^{ct_{(j)}} - 1)} \, e^{\frac{\lambda_2 - \lambda_1}{c}(e^{c\tau_0} - 1)} \\
&= n! \left\{ \lambda_1^k \, e^{c \sum_{i=1}^{k} t_{(i)}} \, e^{\frac{-\lambda_1}{c} \sum_{i=1}^{k} (e^{ct_{(i)}} - 1)} \right\} \\
&\quad \left\{ \lambda_2^{n-k} \, e^{c \sum_{j=k+1}^{n} t_{(j)}} \, e^{\frac{-\lambda_2}{c} \sum_{j=k+1}^{n} (e^{ct_{(j)}} - 1)} \, e^{(n-k)\frac{\lambda_2 - \lambda_1}{c}(e^{c\tau_0} - 1)} \right\} \\
&= n! \lambda_1^k \lambda_2^{n-k} \, e^{c \sum_{j=1}^{n} t_{(j)}} \, e^{\frac{-\lambda_1}{c}\left(\sum_{i=1}^{k} \, e^{ct_{(i)}} - k\right)} \, e^{\frac{-\lambda_2}{c}\left(\sum_{i=k+1}^{n} \, e^{ct_{(i)}} - (n-k)\right)} \\
&\quad e^{(n-k)\frac{\lambda_2 - \lambda_1}{c} e^{c\tau_0}} \\
&= n! \lambda_1^k \, \lambda_2^{n-k} \, e^{c \sum_{j=1}^{n} t_{(j)}} \, e^{\frac{n\lambda_1}{c} - \frac{\lambda_1}{c} \sum_{i=1}^{k} \, e^{ct_{(i)}} - \frac{\lambda_2}{c} \sum_{i=k+1}^{n} \, e^{ct_{(i)}}} \\
&\quad e^{(n-k)\frac{\lambda_2 - \lambda_1}{c} \, e^{c\tau_0}} \\
&= n! \lambda_1^k \, \lambda_2^{n-k} \, e^{c \sum_{j=1}^{n} t_{(j)}} \, e^{\frac{\lambda_1}{c}\left(n - \sum_{i=1}^{k} e^{ct_{(i)}} - (n-k)e^{c\tau_0}\right)} \\
&\quad e^{\frac{\lambda_2}{c}\left(-\sum_{i=k+1}^{n} e^{ct_{(i)}} + (n-k)e^{c\tau_0}\right)},
\end{aligned}$$

de donde la función log-verosimilitud es

$$\begin{aligned}
l(\lambda_1, \lambda_2, k) &= L(\lambda_1, \lambda_2, k) \\
&= \log n! + k \log \lambda_1 + (n-k) \log \lambda_2 + c \sum_{j=1}^{n} t_{(j)} \\
&\quad + \frac{\lambda_1}{c} \left(n - \sum_{i=1}^{k} e^{ct_{(i)}} - (n-k)e^{c\tau_0} \right) \\
&\quad + \frac{\lambda_2}{c} \left(- \sum_{i=k+1}^{n} e^{ct_{(i)}} + (n-k)e^{c\tau_0} \right).
\end{aligned}$$

De esto, se sigue que

$$\frac{\partial l(\lambda_1, \lambda_2, k)}{\partial \lambda_1} = \frac{k}{\lambda_1} + \frac{1}{c} \left(n - \sum_{i=1}^{k} e^{ct_{(i)}} - (n-k)e^{c\tau_0} \right),$$

y

$$\frac{\partial l(\lambda_1, \lambda_2, k)}{\partial \lambda_2} = \frac{n-k}{\lambda_2} + \frac{1}{c} \left(- \sum_{i=k+1}^{n} e^{ct_{(i)}} + (n-k)e^{c\tau_0} \right),$$

por tanto, de esto y haciendo uso del criterio del hessiano podemos ver fácilmente que los *e.m.v.* para λ_1 y λ_2 son

$$\widehat{\lambda_1} = \frac{-kc}{n - \sum_{i=1}^{k} e^{ct_{(i)}} - (n-k)e^{c\tau_0}}, \tag{5.17}$$

y

$$\widehat{\lambda_2} = \frac{(n-k)c}{\sum_{j=k+1}^{n} e^{ct_{(j)}} - (n-k)e^{c\tau_0}}, \tag{5.18}$$

respectivamente. Por otra parte, el BIC para el modelo de la función de riesgo con punto de cambio está dada por:

$$\begin{aligned} BIC(k,\tau_0) &= -2l(\widehat{\lambda_1},\widehat{\lambda_2},\tau_0) + 3\log n \qquad (5.19)\\ &= -2\left[\log n! + k\log\widehat{\lambda_1} + (n-k)\log\widehat{\lambda_2} + c\sum_{i=1}^{n} t_{(i)} - \frac{1}{c}kc \right.\\ &\quad \left. -\frac{1}{c}(n-k)c\right] + 3\log n \\ &= -2\left[\log n! + k\log\left(\frac{-kc}{n - \sum_{i=1}^{k} e^{ct_{(i)}} - (n-k)e^{c\tau_0}}\right)\right.\\ &\quad \left. +(n-k)\log\left(\frac{(n-k)c}{\sum_{j=k+1}^{n} e^{ct_{(j)}} - (n-k)e^{c\tau_0}}\right) + c\sum_{i=1}^{n} t_{(i)} - n\right]\\ &\quad +3\log n, \end{aligned}$$

para una selección empírica de τ_0. Se procede a determinar el valor de $\widehat{k}$ tal que,

$$BIC(\widehat{k},\tau_0) = \min_{1\le k\le n-1} BIC(k,\tau_0). \tag{5.20}$$

Luego, para obtener el estimador del punto de cambio τ_0 en la función de riesgo procedemos a calcular,

$$\frac{\partial BIC(\widehat{k},\tau_0)}{\partial\tau_0} = \frac{2kc(n-k)e^{c\tau_0}}{\sum_{i=1}^{k} e^{ct_{(i)}} + (n-k)e^{c\tau_0} - n} - \frac{2c(n-k)^2 e^{c\tau_0}}{\sum_{j=k+1}^{n} e^{ct_{(j)}} - (n-k)e^{c\tau_0}},$$

finalmente, de esta última igualdad obtenemos que el estimador para τ_0 es

$$\widehat{\tau_0} = \frac{1}{c}\log\left(\frac{\widehat{k}\sum_{i=1}^{n} e^{ct_{(i)}} - n\sum_{i=1}^{\widehat{k}} e^{ct_{(i)}} + n(n-\widehat{k})}{n(n-\widehat{k})}\right). \tag{5.21}$$

5.5.2. Enfoque informacional con el modelo Weibull

Se considera ahora el modelo de Weibull, con el cual se busca obtener una estimación para τ mediante el uso del estadístico BIC. Sea T de tiempo de vida, con función de riesgo,

$$h(t) = \begin{cases} \alpha_1\beta\,(\alpha_1 t)^{\beta-1}, & \text{si } 0 \le t \le \tau, \\ \alpha_2\beta\,(\alpha_2 t)^{\beta-1}, & \text{si } t > \tau. \end{cases} \tag{5.22}$$

Donde, por

$$f(t) = \begin{cases} \alpha_1\beta\,(\alpha_1 t)^{\beta-1} e^{-(\alpha_1 t)^\beta}, & \text{si } 0 \le t \le \tau, \\ \alpha_2\beta\,(\alpha_2 t)^{\beta-1} e^{-(\alpha_1\tau)^\beta + (\alpha_2\tau)^\beta - (\alpha_2 t)^\beta}, & \text{si } t > \tau, \end{cases}$$

y

$$F(t) = \begin{cases} 1 - e^{(\alpha_1 t)^\beta}, & \text{si } 0 \le t \le \tau, \\ 1 - e^{-(\alpha_1\tau)^\beta + (\alpha_2\tau)^\beta - (\alpha_2 t)^\beta}, & \text{si } t > \tau. \end{cases}$$

Haciendo un análisis similar, considere una muestra aleatoria $T_1, T_2, ..., T_n$ de tiempos de vida y $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ los estadísticos de orden de esta muestra, τ_0 tal que $T_{(k)} \le \tau_0 < T_{(k+1)}$, para algún $k = 1, 2, ..., n-1$. Se considera τ_0 para ver si es posible determinar tal k y llegar a estimar el valor de τ.

La función de verosimilitud de $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, considerando la realización $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, está dado:

$$\begin{aligned} L(\alpha_1, \alpha_2, k) &= n! \prod_{i=1}^{k} \left\{ \alpha_1\beta\,(\alpha_1 t_{(i)})^{\beta-1} e^{-(\alpha_1 t_{(i)})^\beta} \right\} \prod_{i=k+1}^{n} \left\{ \alpha_2\beta\,(\alpha_2 t_{(i)})^{\beta-1} \right. \\ &\quad \left. e^{-(\alpha_1\tau)^\beta + (\alpha_2\tau)^\beta - (\alpha_2 t_{(i)})^\beta} \right\} \\ &= n!(\beta\alpha_1^\beta)^k e^{-\alpha_1^\beta \sum_{i=1}^{k} t_{(i)}^\beta} \prod_{i=1}^{k} t_{(i)}^{\beta-1} (\beta\alpha_2^\beta)^{n-k} e^{-(n-k)((\alpha_1\tau_0)^\beta - (\alpha_2\tau_0)^\beta)} \\ &\quad e^{-\alpha_2^\beta \sum_{i=k+1}^{n} t_{(i)}^\beta} \prod_{i=k+1}^{n} t_{(i)}^{\beta-1} \\ &= n!\beta^n \alpha_1^{\beta k} \alpha_2^{\beta(n-k)} e^{-(n-k)\left((\alpha_1\tau_0)^\beta - (\alpha_2\tau_0)^\beta\right) - \alpha_2^\beta \sum_{i=k+1}^{n} t_{(i)}^\beta - \alpha_1^\beta \sum_{i=1}^{k} t_{(i)}^\beta} \\ &\quad \prod_{i=1}^{n} t_{(i)}^\beta. \end{aligned}$$

La función log-verosimilitud es entonces,

$$\begin{aligned} l(\alpha_1, \alpha_2, \tau_0) &= \log(n!) + n\log(\beta) + \beta k \log(\alpha_1) + \beta(n-k)\log(\alpha_2) \\ &\quad + (\beta - 1)\sum_{i=1}^{n} \log(t_{(i)}) - (n-k)\left((\alpha_1\tau_0)^\beta - (\alpha_2\tau_0)^\beta\right) \\ &\quad - \alpha_2^\beta \sum_{i=k+1}^{n} t_{(i)}^\beta - \alpha_1^\beta \sum_{i=1}^{k} t_{(i)}^\beta. \end{aligned}$$

Así, se tiene que,

$$\frac{\partial l(\alpha_1,\alpha_2,\tau_0)}{\partial \alpha_1} = \frac{\beta k}{\alpha_1} - (n-k)\beta\alpha_1^{\beta-1}\tau_0^{\beta} - \beta\alpha_1^{\beta-1}\sum_{i=1}^{k} t_{(i)}^{\beta}$$

$$\frac{\partial l(\alpha_1,\alpha_2,\tau_0)}{\partial \alpha_2} = \frac{\beta(n-k)}{\alpha_2} + (n-k)\beta\alpha_2^{\beta-1}\tau_0^{\beta} - \beta\alpha_2^{\beta-1}\sum_{i=k+1}^{n} t_{(i)}^{\beta}.$$

De donde los estimadores de máxima verosimilitud son

$$\widehat{\alpha_1}^{\beta} = \frac{k}{(n-k)\tau_0^{\beta} + \sum_{i=1}^{k} t_{(i)}^{\beta}}, \tag{5.23}$$

$$\widehat{\alpha_2}^{\beta} = \frac{n-k}{\sum_{i=k+1}^{n} t_{(i)}^{\beta} - (n-k)\tau_0^{\beta}}, \tag{5.24}$$

para un valor empírico de τ_0. Por otro lado, se denota por:

$$BIC(k,\tau_0) = BIC(\widehat{\alpha_1},\widehat{\alpha_2},\tau_0).$$

Luego,

$$\begin{aligned} BIC(k,\tau_0) &= -2\left\{\log n! + n\log\beta + k\log\left(\frac{k}{(n-k)\tau_0^{\beta} + \sum_{i=1}^{k} t_{(i)}^{\beta}}\right)\right. \\ &\quad +(n-k)\log\left(\frac{n-k}{\sum_{i=k+1}^{n} t_{(i)}^{\beta} - (n-k)\tau_0^{\beta}}\right) + (\beta-1)\sum_{i=1}^{n}\log t_{(i)} \\ &\quad -(n-k)\tau_0^{\beta}\left(\frac{k}{(n-k)\tau_0^{\beta} + \sum_{i=1}^{k} t_{(i)}^{\beta}} - \frac{n-k}{\sum_{i=k+1}^{n} t_{(i)}^{\beta} - (n-k)\tau_0^{\beta}}\right) \\ &\quad \left. -\frac{(n-k)\sum_{i=k+1}^{n} t_{(i)}^{\beta}}{\sum_{i=k+1}^{n} t_{(i)}^{\beta} - (n-k)\tau_0^{\beta}} - \frac{k\sum_{i=1}^{k} t_{(i)}}{(n-k)\tau_0^{\beta} + \sum_{i=1}^{k} t_{(i)}^{\beta}}\right\} + 3\log n \\ &= -2\left\{\log n! + n\log\beta + k\log\left(\frac{k}{(n-k)\tau_0^{\beta} + \sum_{i=1}^{k} t_{(i)}^{\beta}}\right)\right. \\ &\quad \left. +(n-k)\log\left(\frac{n-k}{\sum_{i=k+1}^{n} t_{(i)}^{\beta} - (n-k)\tau_0^{\beta}}\right) + \sum_{i=1}^{n}\log t_{(i)}^{\beta-1}\right\} \\ &\quad +3\log n. \end{aligned} \tag{5.25}$$

Luego, se procede a determinar el valor de $\widehat{k}$ tal que

$$BIC(\widehat{k},\tau_0) = \min_{1\le k\le n-1} BIC(\widehat{\alpha_1},\widehat{\alpha_2},\tau_0).$$

Así, se obtiene el estimador para τ,

$$\widehat{\tau_0}^{\beta} = \frac{k\left(k\sum_{i=k+1}^{n} t_i^{\beta} - (n-k)\sum_{i=1}^{k} t_{(i)}^{\beta}\right)}{n(n-k)}.$$

Capítulo 6

CASO DE ESTUDIO, SEGUNDA PARTE, PUNTO DE CAMBIO

En esta sección, abordamos el problema de punto de cambio en la función de riesgo para el caso de estudio con el enfoque informacional y considerando el riesgo con el modelo de Gompertz dado en (5.15) y el riesgo con el modelo Weibull dado en (5.22).

Para empezar el análisis de punto de cambio en el caso de estudio, se hará una prueba de hipótesis (basada en el enfoque infomacional) para verificar si en efecto existe tal punto de cambio. Así, dado que el enfoque informacional hace el uso del BIC, este se basa en un contexto de la selección de un "mejor" modelo dentro de un conjunto de modelos. Por tanto, con este enfoque el objetivo es la selección del mejor modelo dentro de los modelos que se generan en la hipótesis nula de la no existencia de punto de cambio y en la hipótesis alternativa de existencia de un único punto de cambio, posteriormente, se llega a comparar el valor del estadístico $BIC(n)$ (obtenido con la hipótesis nula) y el $BIC(k, \tau_0)$, (obtenido en la hipótesis alternativa) para $1 \leq k \leq 63$ y se selecciona el modelo con menor valor. Es decir, de acuerdo con el principio del criterio de información, los datos proporcionan suficiente información para aceptar H_0 si,

$$BIC(n) < \min_{1 \leq k \leq n-1} BIC(k, \tau),$$

o de otra forma, aceptamos H_a si para algún k ocurre:

$$BIC(n) > BIC(k, \tau).$$

Así, este análisis con datos reales se muestra a continuación.

6.0.1. Inferencia en el modelo de Gompertz

Sea $T_1, T_2, \ldots, T_n$ una muestra aleatoria con función de distribución $F_1, F_2, \ldots, F_n$ respectivamente. Luego, el problema de punto de cambio con el modelo de riesgo de Gompertz (5.15), es probar la hipótesis nula, de no existencia de punto de cambio, contra la hipótesis alternativa de un único punto de cambio, esto es en (5.15),

$$H_0: \quad \lambda_1 = \lambda_2, \tag{6.1}$$

contra la hipótesis alternativa,

$$H_a: \quad \lambda_1 \neq \lambda_2. \tag{6.2}$$

Considere $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ los estadísticos de orden de esta muestra. La función de verosimilitud de $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, considerando la realización $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, está dada:

$$L_0(\lambda) = n!\lambda^n e^{c\sum_{i=1}^n t_{(i)}} e^{-\frac{\lambda}{c}\left(\sum_{i=1}^n e^{ct_{(i)}} - n\right)}$$

y la función log-verosimilitud,

$$l(\lambda) = \log n! + n\log\lambda + c\sum_{i=1}^n t_{(i)} - \frac{\lambda}{c}\left(\sum_{i=1}^n e^{ct_{(i)}} - n\right).$$

De esta última igualdad, y dado que $\frac{\partial^2 l(\lambda)}{\partial \lambda^2} < 0$ puede ver que

$$\widehat{\lambda} = \frac{nc}{\sum_{i=1}^n e^{ct_{(i)}} - n}$$

es el estimador de *e.m.v.* de λ. Luego, denotamos por $BIC(n)$ al BIC del modelo bajo la hipótesis nula (6.1) y dado que este modelo (5.15) sólo cuenta con un parámetro, se tiene entonces que,

$$\begin{aligned} BIC(n) &= -2l\left(\widehat{\lambda}\right) + \log(n) \qquad (6.3)\\ &= -2\left(\log n! + n\log\left(\frac{nc}{\sum_{i=1}^n e^{ct_{(i)}} - n}\right) + c\sum_{i=1}^n t_{(i)} - n\right) + \log n. \end{aligned}$$

Por otra parte, considérese el estadístico $BIC(k, \tau_0)$ definido en (5.19), el cual se da bajo la hipótesis alternativa (6.2), nótese además que para poder obtener los *e.m.v.*, sólo podemos detectar cambios que se encuentran entre la primera y $n-1$ posiciones. Así, por medio de simulación en R y haciendo uso del conjunto de datos de los pacientes en estudio se obtiene el valor de (6.3),

$$BIC(64) = 253.756, \tag{6.4}$$

mientras que los valores del los estadísticos $BIC(k, \tau_0)$ dado en (5.19), para $1 \leq k \leq 63$, se muestran en la Tabla 6.1.

k	BIC	k	BIC	k	BIC	k	BIC
1	253.3879	17	252.0028	33	248.9403	49	253.8600
2	255.8144	18	251.6838	34	249.0820	50	254.1937
3	257.1622	19	251.3538	35	249.2015	51	254.5086
4	257.7976	20	251.0128	36	249.3003	52	254.8060
5	257.8943	21	250.6603	37	249.5744	53	255.1268
6	257.5949	22	250.2959	38	249.9643	54	255.4307
7	257.0503	23	249.9192	39	250.3222	55	255.7190
8	256.4473	24	249.5298	40	250.6512	56	255.9930
9	255.8180	25	249.1270	41	250.9541	57	256.2536
10	255.1204	26	248.7102	42	251.3995	58	256.5017
11	254.3641	27	248.2789	43	251.8134	59	256.7576
12	253.8535	28	247.9133	44	252.1989	60	257.0018
13	253.5030	29	248.1056	45	252.5586	61	257.2351
14	253.1430	30	248.3604	46	252.8951	62	257.4685
15	252.7732	31	248.5821	47	253.2101	63	257.6919
16	252.3933	32	248.7745	48	253.5057		

Tabla 6.1: Tabla de valores obtenidos en el estadístico $BIC(k, \tau_0)$.

De donde puede ver,

$$BIC(64) = 253.756 > BIC(1, \tau_0) = 253.3879.$$

Más aún, existen otros valores de k para los cuales esta desigualdad se cumple. Por tanto, el conjunto de datos muestran suficiente información para rechazar H_0, es decir, se afirma la existencia de un punto de cambio en la función de riesgo considerando el modelo de Gompertz y el cual se estima en la siguiente sección.

6.0.2. Inferencia en el Modelo de Weibull

Ahora se analiza la función de riesgo dada en (5.22) del modelo de Weibull. Sea $T_1, T_2, \ldots, T_n$ una muestra de *v.a.'s*. Luego, el problema de punto de cambio con en el modelo de riesgo, es probar la hipótesis nula, de no existencia de punto de cambio, contra la hipótesis alternativa de un único punto de cambio, esto es,

$$H_0 : \quad \alpha_1 = \alpha_2,$$

contra la hipótesis alternativa,

$$H_a : \quad \alpha_1 \neq \alpha_2.$$

en (5.22). Considere $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ los estadísticos de orden de esta muestra. La función de verosimilitud de $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, está dada por:

$$\begin{aligned} L_0(\alpha) &= n! \prod_{i=1}^{n} \alpha\beta(\alpha t_{(i)})^{\beta-1} \exp(-(\alpha t_{(i)})^{\beta}) \\ &= n!\alpha^{n\beta}\beta^n \left(\prod_{i=1}^{n} t_{(i)}^{\beta-1} \right) \exp\left(-\alpha^{\beta} \sum_{i=1}^{n} t_{(i)}^{\beta} \right). \end{aligned}$$

De donde la función log-verosimilitud está dada,

$$\log(L_0(\alpha)) = \log n! + n\log(\alpha^{\beta}\beta) + \sum_{i=1}^{n} \log t_{(i)}^{\beta-1} - \alpha^{\beta} \sum_{i=1}^{n} t_i^{\beta}.$$

Luego, puede ver que el *e.m.v.* es

$$\widehat{\alpha}^{\beta} = \frac{n}{\sum_{i=1}^{n} t_{(i)}^{\beta}}.$$

Con esto también se obtiene que, el BIC bajo la hipótesis nula está dado por

$$BIC(n) = -2\left(\log n! + n\log\left(\frac{n\beta}{\sum_{i=1}^{n} t_{(i)}^{\beta}} \right) + (\beta - 1)\sum_{i=1}^{n} \log t_{(i)} - n \right) + \log n.$$

Finalmente, con el conjunto de datos en estudio se obtiene que

$$BIC(64) = 193.4614, \tag{6.5}$$

y considerando al $BIC(k, \tau_0)$ dado en (5.25). De los valores que puede ver en la Tabla 6.2, puede notar que,

$$BIC(1, \tau_0) = 191.2477. \tag{6.6}$$

Por tanto, de (6.5) y (6.6), se tiene que,

$$BIC(64) > BIC(1, \tau_0),$$

por lo cual los datos proporcionan suficiente información para aceptar H_a, esto es, se acepta la existencia del punto de cambio.

k	BIC	k	BIC	k	BIC	k	BIC
1	191.2477	17	198.9855	33	191.2554	49	193.338
2	194.832	18	198.5081	34	190.9991	50	193.789
3	197.3376	19	198.0041	35	190.7254	51	194.2072
4	199.1307	20	197.4743	36	190.4337	52	194.5942
5	200.3852	21	196.9192	37	190.3627	53	195.2014
6	201.1656	22	196.3389	38	190.4888	54	195.7706
7	201.597	23	195.7336	39	190.5892	55	196.3042
8	201.7643	24	195.1032	40	190.6649	56	196.8045
9	201.7557	25	194.4475	41	190.7167	57	197.2729
10	201.6066	26	193.7661	42	191.1175	58	197.7098
11	201.3345	27	193.0586	43	191.4827	59	198.3263
12	201.0238	28	192.3792	44	191.815	60	198.9065
13	200.6967	29	192.1165	45	192.1166	61	199.4507
14	200.3259	30	191.9252	46	192.3892	62	200.1511
15	199.9152	31	191.7181	47	192.6341	63	200.8076
16	199.4676	32	191.4949	48	192.852		

Tabla 6.2: Tabla de valores obtenidos en el estadístico $BIC(k, \tau_0)$.

6.1. Estimación del Punto de cambio

En la sección anterior se pudo verificar la existencia del punto de cambio en la función de riesgo mediante un juego de hipótesis considerando dos modelos. En esta sección se busca determinar los valores del punto de cambio en dichos modelos modelos, por lo cual, se ha trabajado con el uso del programa estadístico R, para implementar del algoritmo y poder obtener dichos valores.

Estimación con el modelo de Gompertz

Mediante la implementación de un algoritmo en R, se obtienen los valores de los estimadores $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ de λ_1 y λ_2 respectivamente, dados en (5.17) y (5.18). Además, también se obtiene el valor de k que minimiza el valor del BIC (5.20), al cual denotamos por $\widehat{k}$ y finalmente, considerando un valor empírico de τ_0 se obtiene el valor del estimador del punto de cambio $\widehat{\tau_0}$ (5.21) (para ver el algoritmo vea Apéndice C).

Se considera entonces el conjunto de datos de pacientes con insuficiencia renal con el cual se ha trabajado anteriormente y se propone el valor empírico $\tau_0 = 1$ y $c = -0.04$ (valor que se considera en este caso, a partir de los resultados obtenidos en la subsección 4.2, y el cual se muestra en la Tabla 4.2). Por otra parte la función Mgompertz() (Apéndice C), no recibe parámetros y se considera que $datos\$Tiempo$ son los tiempos de vida registrados de cada uno de los pacientes (los cuales están guardados en el dataframe *datos*). Por otra parte, la función imprime los valores obtenidos del BIC para cada $1 \leq k \leq n-1$ y

finalmente imprime el valor de k que minimiza el BIC, con la variable posición.

Analizando el valor del BIC que se obtiene para $1 \leq k \leq 63$, se tiene que el valor de k que minimiza dicho valor, es en $k = 28$. Posteriormente, considerando este valor y sustituyendo en (5.17) y (5.18) obtenemos los valores de: $\widehat{\lambda}_1 = 0.1288$ y $\widehat{\lambda}_2 = 0.05671$. Además, también podemos obtener el valor del punto de cambio $\widehat{\tau}_0$ dado en (5.21), concluyendo que $\widehat{\tau}_0 = 6.0775$. Esto lo podemos obtener mediante la función descrita en Apéndice C.

Finalmente, el modelo descrito en (5.15) obtenido es:

$$h(t) = \begin{cases} 0.1288\, e^{-0.0463\, t}, & \text{si } 0 \leq t \leq 6.0775, \\ 0.05671\, e^{-0.0463\, t}, & \text{si } t > 6.0775. \end{cases}$$

Con estos parámetros se puede ver gráficamente la función de riesgo dada en (5.15) en la Figura 6.1, junto con la función de riesgo empírico.

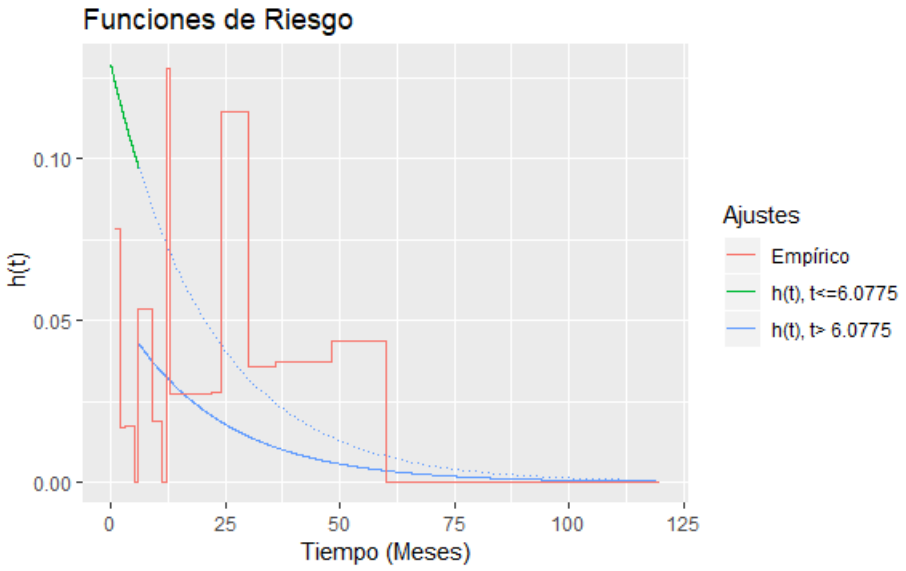


Figura 6.1: Función de Riesgo del modelo Gompertz.

Estimación con el modelo de Weibull

Consideremos ahora el riesgo con el modelo de Weibull descrito en (5.22). Se considera la constante fija $\beta = 0.6220$, obteniendo con este valor que $\widehat{\alpha}_1 = 0.0363$ y $\widehat{\alpha}_2 = 0.0328$, los valores de los estimadores dados en (5.23) y (5.24) respectivamente y son obtenidos de forma similar al modelo de Gompertz mediante un algoritmo en R (vea Apéndice C).

Así, el riesgo (5.22) se describe como,

$$h(t) = \begin{cases} 0.0714\beta\, (0.0714t)^{-0.378}, & \text{si } 0 \leq t \leq 13.2617, \\ 0.0182\beta\, (0.0182t)^{-0.378}, & \text{si } t > 13.2617. \end{cases} \tag{6.7}$$

Mientras, que gráficamente puede verse en Figura la 6.2(la función de riesgo es representada con la línea continua).

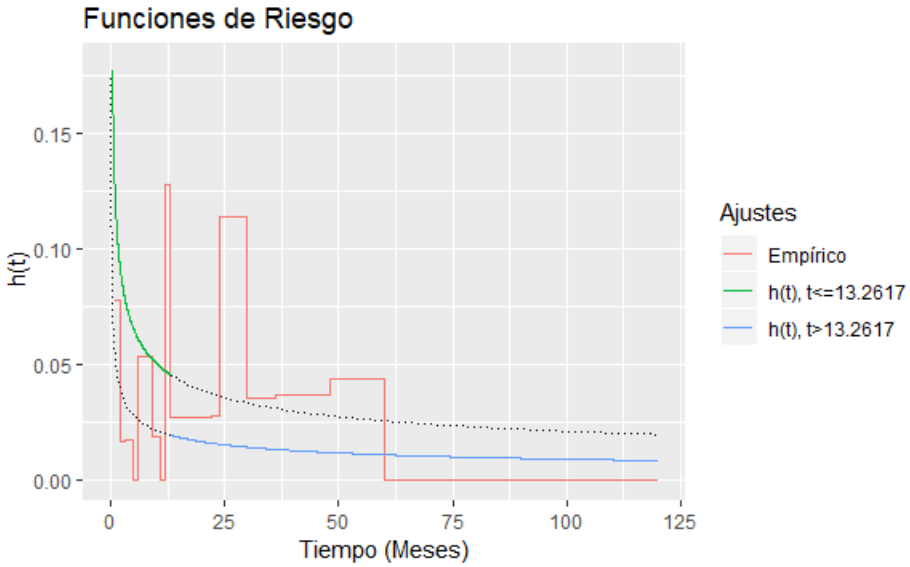


Figura 6.2: Función de Riesgo del modelo Weibull.

Capítulo 7

CONCLUSIÓN

En este trabajo de tesis, habló de cómo generar modelos de tiempo de vida, en particular, aplicado a un caso de trasplante renal de pacientes del estado de Puebla con la finalidad de estudiar su función de supervivencia considerando que los datos pueden estar o no censurados por la derecha. Sin embargo, también se hace la observación de que cada uno de los resultados obtenidos, son sólo considerando los datos en estudio ya que en el análisis de otros problemas clínicos quizá sea conveniente estudiar e incluir otras variables, obteniendo posiblemente otros resultados.

Por otra lado, esto motivó a realizar un estudio de modelos paramétricos, no paramétricos y semiparamétricos, obteniéndose, para el caso de estudio, los siguientes resultados.

Con los métodos no paramétricos, considerando el género de los pacientes, se pudo ver que existe una mayor probabilidad de supervivencia en los hombres que en las mujeres durante los primeros 3 años. Además, se encontraron cambios en la probabilidad de supervivencia durante este periodo de tiempo, debido a que un paciente que rechaza el injerto lo hace con mayor probabilidad en este lapso de tiempo, esto es, entre más tiempo permanezca con el injerto renal, menor es el riesgo que corre para rechazarlo. Por otra parte también se obtuvo que los pacientes más jóvenes son en gran mayoría los que corren un mayor riesgo (posiblemente porque en el conjunto de datos hay más personas jóvenes con dicho problema).

En los modelos paramétricos, haciendo un análisis gráfico se llegó a suponer que el modelo que mejor se ajustaba al conjunto de datos era el modelo de Gompertz, lo cual se pudo corroborar de forma analítica por medio de los estadísticos AIC y BIC, los cuales permiten evaluar la bondad de cada uno de los modelos propuestos. Más aún, analizando la función de Riesgo, se pudo ver que es decreciente, lo cual nos ayudaba a verificar que en efecto, los pacientes, tienen un mayor riesgo al rechazo del trasplante durante los primeros 3 años, esto es, pacientes que no rechazan el injerto durante este periodo tienen menos riesgo de rechazarlo posteriormente.

Considerando el modelo semiparamétrico de riesgos proporcionales o de

Cox, el cual fue utilizado con la finalidad de determinar qué variables de los pacientes en estudio influían significativamente en la función de riesgo base, se pudo ver que la variable Creatina Sérica produce un aumento en la función de riesgo basal.

Finalmente, en el análisis de punto de cambio, se verificó de forma analítica y por medio del enfoque informacional la existencia de un punto de cambio en la función de riesgo considerando dos modelos, el modelo de Gompertz y el modelo de Weibull. Por lo cual, se tuvo que implementar un algoritmo en R para poder obtener el valor de los estimadores del modelo, así como el valor del estimador para el punto de cambio. Así, se encontraron dos puntos de cambio y dado que el conjunto de datos es pequeño, se decidió sólo considerar dichos puntos.

Como se puede ver, el Análisis de Supervivencia es una técnica estadística de gran importancia al permitirnos tener una mejor visión acerca del riesgo y la probabilidad de supervivencia de los pacientes en estudio considerando diversas covariables, por lo cual, para trabajos futuros se podría considerar la inclusión de una mayor cantidad de variables explicativas en el modelo semi-parametrico de Cox, pudiendo así determinar otras covariables que producen un aumento o disminución del riesgo en dichos pacientes, y con esto tener una mejor idea de qué covariables son más influyentes en estos casos. Con todo esto, también se pretendería motivar a expertos en materia de salud a realizar estudios meticulosos de las covariables de los individuos de forma completa, es decir, tratando de no dejar que se produzcan datos faltantes dentro de la información, además de considerar variables relevantes, por ejemplo; edad del donador y si el paciente ha recibido o no, anteriormente un trasplante, ya que esto ayudaría mucho a realizar un mejor análisis al problema, porcentaje de compatibilidad entre el donante y donador, etc., pues se ha podido ver que muchas veces los datos que se proporcionan son incompletos y muchos de ellos son variables descriptivas no cuantitativas.

Por otra parte, en cuanto a modelos, en trabajos posteriores también se podría realizar un mayor estudio con la familia paramétrica Weibull o bien con la Gamma Generalizada, dado que esta es una familia bastante amplia, lo cual ayudaría a considerar posiblemente un modelo que explique mejor el comportamiento de los datos y del fenómeno mismo.

Bibliografía

[1] Chen, J., & Gupta, A. K. (2014). Parametric statistical change point analysis: With applications to genetics, medicine, and finance. Birkhauser Boston.

[2] Collett, D. (2014), Modelling Survival Data in Medical Research. Chapman & Hall/CRC

[3] Cox, D. R. & Oakes, D. (1984). Analysis of Survival Data. London, New York: Chapman & Hall/CRC.

[4] Domínguez, A. M. H. (2010). Análisis Estadístico de Datos de Tiempos de Fallo en R. Tesis de maestría, Departamento de Estadística e I. O., Universidad de Granada.

[5] Estrada, A. V., López, J. A. M., Alvarado, M. R., & Cervantes, M. L. (2009). Insuficiencia renal crónica. Unidad de Proyectos Especiales, Universidad Nacional autónoma de México.

[10] Fernández, L. M. (2011). Métodos de Inferencia para la Distribución Weibull, Aplicación en Fiabilidad Industrial. Tesis de Maestría, Universidad de Vigo.

[6] Gomà, A. S., Peris, P. A., & Alcario, A. B. R. (2010). Calidad de vida en pacientes con insuficiencia renal crónica en tratamiento con diálisis. Revista de la Sociedad Española de Enfermería nefrologica.

[7] Horváth, L. (1996). The maximum likelihood method for testing changes in the parameters of normal observations. The Annals of statistics.

[8] i Massons, J. M. D. (1992). Una aplicación del análisis de la supervivencia en ciencias de la salud. Anuario de psicología/The UB Journal of psychology.

[9] Klein, J. P., & Moeschberger, M. L. (2010). Survival Analysis: Techniques for Censored and Truncated Data. Springer Science & Business Media.

[10] Kleinbaum, D. G., & Klein, M. (2010). Survival analysis. New York: Springer.

[11] Lawless, J. F. (2011). Statistical Models and Methods for Lifetime Data. Nueva Jersey, Wiley-Interscience.

[12] Martín, P., & Errasti, P. (2006). Trasplante renal. In Anales del sistema sanitario de Navarra. Gobierno de Navarra. Departamento de Salud.

[13] Montoya, J. R., Regino, E., & Gómez, S. Y. G. (2017). Comparación de métodos de estimación en regresión de Cox. Comunicaciones en Estadística, Universidad Santo Tomas.

[14] Peña, R. E. B. (2005). Análisis de Sobrevivencia utilizando el lenguaje R. Simposio de Estadística, Paipa, Boyacá, Colombia.

[15] Pérez, K. G. T. (2015). Conceptos del análisis de supervivencia y una aplicación para pacientes con diabetes tipo II. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla.

[16] Rojas, O. P. (2008). Estimación de la Función de Riesgo en un Modelo de Regresión con un Punto de Cambio. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla.

[17] Segura, M. A. M. (2010). Inferencia para Modelos de Supervivencia de un solo Evento y Extensiones para Modelos de Riesgos Competitivos. Tesis de maestría, Departamento de matemáticas, Universidad Autónoma Metropolitana.

[18] Vargas, B. X. M. (2017). Análisis de la Deserción en las Licenciaturas de la FCFM-BUAP Mediante el Modelo de Riesgo Proporcional Semiparamétrico. Tesis de maestría, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla.

[19] Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. Journal of the American Statistical Association.

Apéndice A

Base de Datos

A continuación se muestran los datos de los pacientes con insuficiencia renal, obtenido del IMSS del estado de Puebla. Cada individuo es etiquetado con un número, y en las columnas, se describen cada una de las covariables que se consideraron.

N. individuo	Edad	Sexo	Tiempo (meses)	Censura	Talla (cm)	Creatina Sérica	Tiempo de Isquemia (Seg.)
1	15	Hombre	12	1	164	7.4	15
2	14	Mujer	72	0	142	10	80
3	12	Mujer	96	0	137	8.9	150
4	12	Hombre	96	0	158	7.1	90
5	16	Hombre	24	1	126	7.5	60
6	16	Mujer	84	0	149	9.5	90
7	13	Mujer	12	1	135	10	100
8	13	Hombre	84	0	126	10.3	125
9	12	Mujer	12	1	133	5.6	68
10	12	Mujer	6	1	146	6.9	130
11	14	Hombre	84	0	155	14.5	60
12	11	Mujer	12	1	136	5.5	120
13	15	Hombre	9	1	147	8	130
14	14	Mujer	72	0	147	6.6	60
15	15	Mujer	72	0	145	9.4	60
16	14	Hombre	72	0	140	10	60
17	13	Mujer	1	1	143	8.1	120
18	18	Hombre	60	0	163	10	60
19	19	Mujer	72	0	150	2.6	360
20	15	Hombre	60	0	153	13.6	150
21	6	Hombre	24	1	113	3.5	156
22	16	Mujer	60	0	138	11.4	60
23	10	Mujer	2	1	145	12.2	90
24	10	Hombre	1	1	125	13.4	140

N. individuo	Edad	Sexo	Tiempo (meses)	Censura	Talla (cm)	Creatina Sérica	Tiempo de Isquemia (Seg.)
25	36	Mujer	60	0	153	9.2	90
26	11	Mujer	1	1	120	9	70
27	20	Hombre	1	1	174	10.6	65
28	18	Mujer	48	0	160	9	70
29	16	Hombre	12	1	146	15.7	105
30	24	Hombre	48	0	170	14.7	160
31	15	Mujer	3	1	130	10.3	150
32	16	Hombre	48	0	162	10.3	114
33	18	Mujer	24	1	159	15.8	10
34	19	Mujer	48	0	153	4.1	90
35	20	Hombre	48	0	161	50	90
36	14	Hombre	6	1	160	11.5	103
37	16	Mujer	24	1	150	12	103
38	14	Mujer	120	0	146	13.2	80
39	21	Mujer	48	0	157	15.7	103
40	13	Hombre	36	1	138	13.4	60
41	10	Hombre	36	0	167	9.2	240
42	16	Hombre	48	1	156	11.2	105
43	15	Mujer	12	1	147	17	60
44	19	Hombre	36	0	170	11.1	85
45	22	Hombre	36	0	162	12.1	85
46	20	Mujer	1	1	150	21.1	1
47	18	Hombre	13	1	161	20.4	10
48	15	Mujer	5	0	145	8.8	77
49	16	Hombre	24	0	160	13.3	58
50	12	Hombre	24	0	122	9.2	50
51	14	Mujer	24	0	148	16.6	105
52	14	Hombre	30	1	136	10.8	53
53	15	Hombre	12	0	160	15	50
54	8	Mujer	72	0	113	9.9	124
55	20	Hombre	12	0	178	18.7	120
56	18	Hombre	22	1	165	15.5	80
57	19	Mujer	12	0	162	8.4	84
58	18	Hombre	12	0	168	9.5	120
59	16	Hombre	6	1	160	19	90
60	21	Hombre	11	0	156	12.6	270
61	17	Mujer	11	0	153	15.4	110
62	13	Hombre	11	0	142	10	68
63	16	Mujer	11	0	152	12.8	138
64	17	Mujer	11	0	148	3.7	60

Tabla A.1: Tabla de información de los pacientes en estudio.

Apéndice B

Análisis de Supervivencia en R

En este apéndice exponemos las principales herramientas para llevar a cabo un análisis de tiempos de vida mediante el uso del lenguaje de programación R. Para ver estas funciones con mayor detalle puede consultar [4] y [14]. El Análisis de Supervivencia en el entorno de programación estadística R puede hacerse a través de diferentes librerías especializadas, sin embargo aquí nos centramos aún más en el estudio de la librería survival, ya que ésta nos permite realizar un análisis de tiempo de vida para datos que presentan censura y truncamiento. Posteriormente también se describen: survival, KMsurv, survminer, flexsurv y dplyr.

B.1. Estimación paramétrica

La función de supervivencia puede ser estimada por medio de distribuciones paramétricas con la función flexsurvreg() de la paquetería flexsurv. Los parámetros son estimados por máxima verosimilitud utilizando los algoritmos disponibles en R.

Ejemplo B.2. Se ajustará la distribución exponencial, del conjunto de datos que se tienen en el Apéndice A. Para lo cual se ejecuta la siguiente línea de código:

flex - flexsurvreg(Surv(duration, delta) $\sim$ 1, data = bfeed, dist = "exp")

Las salidas de esta función son las siguientes:

1. est: La estimación máxima verosimilitud.
2. L y U: Intervalos de confianza inferior y superior.
3. se: La desviación estándar de la estimación.
4. N: Número de datos.
5. Events: Número de datos observados.
6. Censored: Número de datos censurados.

7. Total time at risk: Tiempo total en riesgo.
8. Log-likelihood: Log-verosimilitud.
9. df: Grados de libertad.
10. AIC: Índice AIC.

Para graficar la curva estimada, puede ejecutar:

plot(flex,col="blue",main="Curva de Supervivencia", xlab ="Tiempo (meses)", ylab ="Probabilidad de supervivencia", legend.labs ="Modelo Exponencial")

Obteniendo con ello, la Figura B.1:

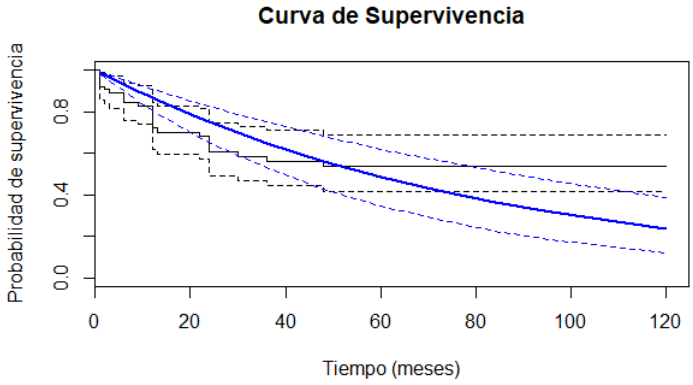


Figura B.1: Función de Supervivencia del modelo Exponencial.

B.3. Estimación no paramétrica

Función Surv

Esta función permite crear objetos de tipo survival, su sintaxis es la siguiente:

Surv(time, time2, event, type=c("right", "left", "interval", "counting", "interval2"))

Donde

1. time: representa el tiempo el tiempo del inicio de la observación.
2. time2: tiempo de finalización de la observación.
3. event: representa el indicador de estado, 1 indica falla (no censurado) y 0 es vivo (censurado).
4. type: especifica el tipo de censura, por defecto, suele ser censura por la derecha.

Función Survfit

Esta función permite crear curvas de supervivencia, como el estimador de Kaplan- Meier (es opción por defecto) o de Fleming y Harrington aunque también permite predecir la función de supervivencia para modelos de Cox, o de un modelo de tiempo de vida acelerada. Esta función en su forma más sencilla sólo requiere un objeto de supervivencia creado por la función Surv(). Los argumentos de esta función son:

1. fórmula: objeto para la fórmula.
2. data: objeto data frame donde están los datos
3. type: tipo de estimador: kaplan-meier o fleming-harrington.

Ejemplo B.4. Considere la base de datos llamada DATOS, éste contiene los datos de 64 pacientes con insuficiencia renal (datos que se han estudiado para el presente trabajo). Algunas de las covariables de cada paciente son:

- Tiempo: tiempo de observación del paciente.
- Censura: indicador del rechazo del trasplante renal (1=rechaza, 0=no rechaza).
- Sexo o género.
- Talla en cm.

Para generar la función de supervivencia con el estimador KM se ejecutan los siguientes comandos:

tongue<- Surv(DATOS$Tiempo, DATOS$Censura)
tongueKM<- survfit(tongue $\sim$ 1, data =Datos, type =“kaplan-meier”)

Sin embargo, para ver la estimación de supervivencia puede usar la función summary(), por ejemplo, en este ejemplo, summary(tongueKM), el cual devuelve los siguientes valores:

1. time: tiempo de la observación
2. n.risk: el número de sujetos en riesgo.
3. n.evento: el número de sujetos que presentaron el evento.
4. survival: la estimación de la función de supervivencia.
5. std.err: la desviación estándar de la estimación.
6. lower y upper CI*: los intervalos de confianza para la estimación.

Así, con el Ejemplo B.4, puede ver en la Figura B.2 los valores que devuelve esta función.

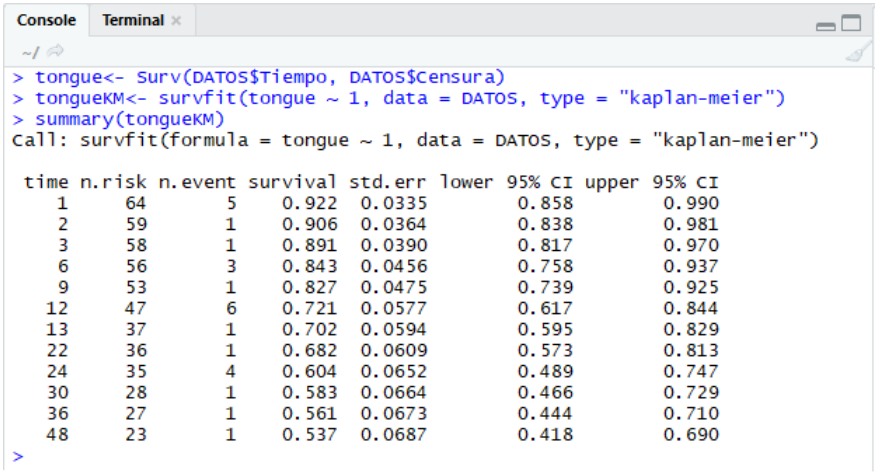


Figura B.2: Estimación de la *f.s.* con KM.

Graficación de la curva de supervivencia

La curva de supervivencia estimada se gráfica con la función ggsurvplot(), el cual requiere de la paquetería survminer y la librería ggplot2. Esta función requiere de muchos parámetros, por lo que solamente ilustraremos los más importantes.

1. fit: objeto tipo survfit.
2. data: un conjunto de datos utilizado para ajustar curvas de supervivencia (Data frame),
3. fun: transformación de la curva de supervivencia, (opcional), algunas posibles opciones son: event para los eventos acumulados, cumhaz para el riesgo acumulado.
4. conf.int: indicador para graficar los intervalos de confianza.
5. title: título de la gráfica.
6. xlab: cadena para el eje x.
7. ylab: cadena para el eje y.
8. legends.lab: vector de nombres para identificar las curvas.
9. legend.title: título de la leyenda.

Ejemplo B.5. Considerando el conjunto de datos como en el ejemplo B.4, para graficar la curva de supervivencia, se tiene el siguiente comando:

ggsurvplot(fit = tongueKM, data =DATOS, conf.int = T, title =“Curva de

Supervivencia", xlab ="Tiempo", ylab = "Probabilidad de supervivencia", legend.title ="Estimación", legend.labs ="Kaplan-Meier")

Lo cual genera la Figura B.3.

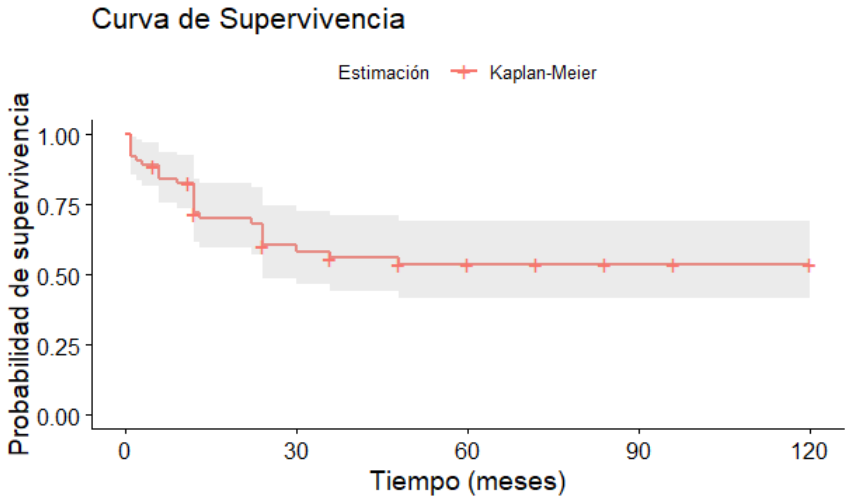


Figura B.3: Estimación no paramétrica de la *f.s.*

Estimación del Riesgo Acumulado

Para el calculo y graficación del riesgo acumulado por el estimador de Nelson Aalen se hace uso de la función fortify sobre un objeto Surv. Además, usamos las funciones %>% y mutate, las cuales pertenecen a la librería dplyr, una librería para la gramática de manipulación de datos. Por tanto, del ejemplo que se ha venido manejando, podemos ejecutar las siguientes líneas de comando y obtener la curva de Nelson Aalen para dicho conjunto de datos.

tongueKM<- survfit(Surv(Tiempo,Censura) ~ 1,DATOS)
R <- tongueKM %>% fortify %>% mutate(CumHaz = cumsum(n.event/n.risk))

B.6. Estimación del Modelo de Riesgo Proporcional

El modelo semiparamétrico de riesgo proporcionales de Cox es realizado con la función coxph() de la paquetería survival, los argumentos de esta función son los siguientes:

1. Fórmula: un objeto fórmula $y \sim x$ que debe tener un objeto *Surv* como variable respuesta y a la izquierda del $\sim$, y del lado derecho las covariables del modelo.

2. Data: objeto data frame que contiene los datos y covariables nombradas en la fórmula.

3. Method: nombre del método usado para los datos repetidos (efron, breslow y exact). El método efron está por defecto.

La salida de la función es una tabla con la siguiente información:

1. coef: coeficiente estimado de beta,
2. exp(coef): exponencial elevado al coeficiente beta estimado,
3. se(coef): la desviación estándar de la estimación,
4. z: valor del estadístico para prueba Wald de beta igual a cero,
5. p: p-valor de la prueba Wald, beta igual a cero.

Ejemplo B.7. Consideremos el conjunto de datos llamado DATOS. Para determinar el valor de los regresores del modelo de Cox para las covariables: sexo y edad, se ejecuta lo siguiente:

coxph(Surv(Tiempo, Censura) ~ Sexo+Edad, data =DATOS)

Lo cual devuelve los valores dados en la Figura B.4.

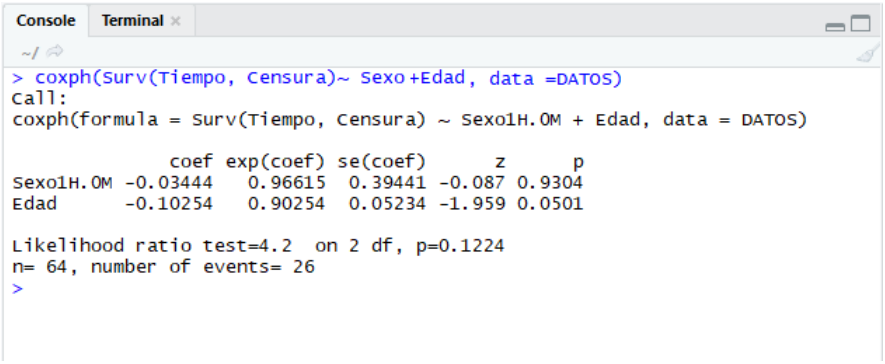


Figura B.4: Valores estimados de los regresores del modelo de Cox.

Por otra parte, también podemos calcular la función de supervivencia pronosticada para un modelo de riesgo proporcional, mediante la función survfit, cuyos argumentos son:

1. formula: un modelo calculado con la función coxph.
2. newdata: un data frame con los mismo nombres que aparecen en el modelo ajustado. La curva producida representara la cohorte con los valores proporcionados.
3. conf.int: nivel del intervalo de confianza (0.95 es fijado por defecto).
4. se.fit: indicador si se deben calcular el error estándar.
5. conf.type: tipo de calculado para los intervalos de confianza:none, plain, log(por defecto) y log-log.

Apéndice C

Punto de cambio en R

Se presenta a continuación el algoritmo para el cálculo del punto de cambio con el modelo de Gompertz y Weibull, así como una forma de como obtener los estimadores de cada uno de estos modelos. Dichos algoritmos son elaboración propia.

Como se ha comentado en la sección 6, se considera el hecho de que se cuenta con una base de datos de 64 pacientes con insuficiencia renal a los cuales se les ha hecho un trasplante, por lo cual en el algoritmo se considera el valor de $n = 64$ y el punto empírico en $t_0 = 1$. Con esto el conjunto de datos se encuentra guardado en un DataFrame denominado *datos*, el cual contiene una columna denominada *Tiempo*. Esta función Mgompertz() regresa el valor de $\widehat{k} = 28$ como resultado del valor que minimiza el valor del BIC.

```
    Mgompertz <-function()
{
     n = 64
     ti=sort(datos$Tiempo)
     t0 = 12
     gg=log(factorial(n))
     s=sum(ti)
     c = 0.0463
     aux = 50000

    for (k in 1:(n-1))
    {
       s1=sum(exp(c*ti[1:k]))
       s2=sum(exp(c*ti[(k+1):n]))
       term1=log(k*c)-log(-n+s1+(n-k)*exp(c*t0))
       term2=log(n*c-k*c)-log( s2-(n-k)*exp(c*t0))
       res=-2*(gg+k*term1+(n-k)*term2+c*s-n)+3*log(n)
       print(res)

       if(res<=aux)
        {
          aux=res
```

```
            posicion=k
          }
    }
      cat("Posicion de k:",posicion)
}
```

Habiendo obtenido el valor de $\widehat{k}$, procedemos a obtener el valor de $\widehat{\tau}_0$, éste valor lo podemos obtener mediante la siguiente función Estimadores(), el cual también devuelve el valor de los estimadores $\widehat{\lambda}_1$ y $\widehat{\lambda}_2$.

```
    Estimadores<-function()
{       n=64
        k=28
        t0=1
        c=-0.0463
        ti=sort(datos$Tiempo)
        term1=log(k*sum(exp(c*ti))-n*sum(exp(c*ti[1:k]))+n*(n-k))
        tau=(1/c)*(term1-log(n*(n-k)))
        print(tau)

        L1=pos*c/(-n+sum(exp(c*ti[1:k]))+(n-k)*exp(c*t0))
        L2=(n-k)*c/(sum(exp(c*ti[(k+1):n]))-(n-k)*exp(c*t0))

        print(L1)
        print(L2)
}
```

Por otro lado, para obtener el valor del estadístico $BIC(n)$ del modelo de riesgo de Gompertz sin puntos de cambio, puede obtenerlo mediante la siguiente linea,

```
sn=-2*(gg+n*log(-n*c/(n-sum(exp(c*ti))))+c*s-n)+log(n)
```

donde cada una de las variables, son definidas como en la función Mgompertz().

Se describe ahora el algoritmo para obtener el valor de $\widehat{k}$, considerando el modelo de Weibull.

```
    Mweibull <-function()
{       n=64
        k=37
        t0=1
        b=0.6220
        gg=log(factorial(n))
        s=(b-1)*sum(log(ti))
```

```
    aux=500000

  for (k in 1:(n-1))
    {
     d1=(n-k)*t0^b+sum(ti[1:k]^b)
     d2=sum(ti[(k+1):n]^b)-(n-k)*t0^b
     s1=k/d1
     s2=(n-k)/d2
     bicc=-2*(gg+n*log(b)+k*log(s1)+(n-k)*log(s2)+s-n)+3*log(n)
     print(bicc)

  if(bicc<=aux)
     {
     aux=bicc
     posicion=k
     print(k)
     }
  }
}
cat("Posicion de K:",posicion)
```

En este caso se obtiene que el valor de $\widehat{k}$ es 37, así, para tener el valor de τ y los estimadores se obtiene por medio del siguiente código.

```
    b=0.622
    t0=1
    k=37
    Pto=(k*sum(ti[(k+1):n]^b)-(n-k)*sum(ti[1:k]^b))/(64*(64-k))
    P=Pto^(1/b)
    print(P)

    l1=k/((64-k)*t0^b+sum(ti[1:k]^b))
    r1=l1^(1/b)
    l2=(64-k)/(sum(ti[(k+1):64]^b)-(64-k)*t0^b)
    r2=l2^(1/b)
    print(r1)
    print(r2)
```

También para obtener el valor del estadístico BIC del modelo de riesgo de Weibull sin puntos de cambio, se puede ejecutar la siguiente instrucción.

```
sn=-2*(gg+n*log(n*b/(sum(ti^b)))+s-n)+log(n)
```