UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Modelos Paramétricos de Regresión en Análisis de Supervivencia

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

PRESENTA:

ALAN SILVA TORRES

DIRECTOR DE TESIS:

MAT. MARGARITA ELVIRA CHÁVEZ CANO

2017

1. Datos del alumno
Silva
Torres
Alan
68110942
Universidad Nacional Autónoma de México
Facultad de Ciencias
Licenciado en Actuaría
410017719

2. Datos del tutor
Mat.
Margarita Elvira
Chávez
Cano

3. Datos del sinodal 1
Act.
Jaime
Vázquez
Alamilla

4. Datos del sinodal 2
Dra.
Lizbeth
Naranjo
Albarrán

5. Datos del sinodal 3
Act.
Ángel Manuel
Godoy
Aguilar

6. Datos del sinodal 4
Act.
Francisco
Sánchez
Villarreal

7. Datos del trabajo escrito
Modelos Paramétricos de Regresión en
Análisis de Supervivencia
193 p.
2017

Para mi hermano y mis padres
Por su amor y apoyo incondicional
Por estar presentes cuando nadie más está.

A la profesora Margarita
Por su tiempo, dedicación y compromiso con este trabajo
Gracias por su confianza y valiosas enseñanzas.

A mis amigos de la facultad y a unos cuantos más
Por ser el pilar que me ayudó a no dejarme derrotar.

A los profesores que me heredaron
su valioso conocimiento y el gusto por aprender más.

A los familiares y amigos que me
acompañaron el día de mi examen profesional.

A mis sinodales
Por sus sugerencias y provechosos comentarios.

A los familiares y personas que contribuyeron
con su apoyo, directa o indirectamente, en esta tesis.

¡Este logro es gracias a ustedes!

Contenido

Intoducción

El análisis de supervivencia es un conjunto de procedimientos estadísticos en el que el principal objetivo es estudiar el tiempo de ocurrencia de algún evento específico de interés. Una característica que distingue a los datos en este análisis es la presencia de observaciones incompletas, en las cuales, se desconoce el tiempo exacto de ocurrencia del evento de interés, a este tipo de observaciones se les conoce como datos censurados.

Generalmente los datos de supervivencia son estudiados con un enfonque paramétrico o semiparamétrico, en la mayoría de los casos se prefiere utilizar estos últimos debido a que son más sencillos de abordar en la práctica. Habitualmente los modelos paramétricos de regresión requieren de una revisión más detallada, sin embargo, si se encuentra un modelo de probabilidad teórico, con el cual se pueda ajustar de manera adecuada a los datos de supervivencia, es posible obtener resultados más precisos con respecto a la descripción del tiempo de ocurrencia del evento de interés.

El objetivo principal de esta tesis, es estudiar a los modelos paramétricos de riesgos proporcionales y comparar la eficiencia de estos modelos con la de los modelos semiparamétricos de riesgos proporcionales, mejor conocidos como modelos de Cox.

En el capítulo uno se introducen conceptos y resultados esenciales del análisis de supervivencia; se revisa con detalle el concepto de falla, se identifican los distintos tipos de censura y se exponen algunos resultados derivados de las funciones de riesgo y supervivencia. Como una primera aproximación a los modelos paramétricos de regresión, en este capítulo se presentan las funciones de supervivencia, de densidad, de riesgo y de riesgo acumulado de los modelos con distribución exponencial, Weibull, gamma, log-normal, entre otros.

En el capítulo 2, se estudia un enfoque general de los modelos de regresión en el análisis de supervivencia, aún considerando aquellos modelos que no se definen de manera paramétrica, tal es el caso de los modelos de Cox, el cual se define como un modelo semiparamétrico. No obstante, como tema principal de este capítulo se exponen los modelos paramétricos de riesgos proporcionales, en particular se presentan los modelos de riesgo exponencial, Weibull y Gompertz.

Finalmente en el capítulo tres se ajusta un modelo paramétrico de riesgos proporcionales a los datos de supervivencia de un grupo de pacientes con leucemia aguda.

Con el objetivo de conocer las principales características de la población en estudio, se presenta un análisis descriptivo de la muestra, asimismo, se realiza un ajuste no paramétrico de los datos utilizando el estimador de Kaplan-Meier y un ajuste semiparamétrico con el modelo de riesgos proporcionales de Cox, este último con el propósito de comparar la eficiencia de los modelos paramétricos y semiparamétricos de riesgos proporcionales.

En los capítulos uno y dos se pueden encontar ejemplos prácticos, los cuales han sido seleccionados pensando en aquellos problemas que aparecen de manera cotidiana en los textos clásicos de análisis de datos de supervivencia. El desarrollo de algunos de los ejemplos, así como los resultados gráficos que se exponen en esta tesis, fueron realizados con ayuda del software estadístico **R**, si se desea reproducir estos resultados, se pueden consultar los códigos correspondientes a cada gráfica y cálculo numérico realizado en el Apéndice A, mientras que las bases de datos utilizadas se pueden revisar en el Apéndice B.

Capítulo 1

Introducción al Análisis de Supervivencia

Algunos ensayos clínicos llevan a cabo el seguimiento de pacientes con el propósito de conocer el tiempo de ocurrencia de algún evento específico de interés, ejemplos de estos eventos son: la muerte, recaída, curación o el alta clínica de un paciente. El tiempo de ocurrencia de estos eventos puede variar desde unas semanas hasta varios años, esta situación presenta una de las primeras complicaciones al momento de analizar el conjunto de datos recopilados en el ensayo, debido a que el tiempo de estudio puede ser tan corto, que no permite al investigador observar el evento de interés durante ese periodo, o bien el tiempo puede ser tan extenso que se pierde el seguimiento del paciente, ya sea por abandono del mismo o por causas ajenas al evento de interés definido. La ausencia de observaciones completas en estos ensayos no permite analizar los datos con técnicas estadísticas convencionales, sin embargo, es posible realizar inferencias de una población en estudio mediante una metodología estadística conocida como *análisis de supervivencia*.

El **análisis de supervivencia** es una rama de la estadística que se concentra en analizar los datos de uno o varios grupos de individuos en los cuales la variable de estudio es el tiempo que transcurre desde un origen hasta que se produce un evento específico de interés, conocido como falla. Al tiempo trascurrido desde el origen hasta el estado final o al momento en el que ocurre la falla se le conoce como *tiempo de supervivencia*, usualmente, a este tiempo también se le denomina *tiempo de falla*.

Actualmente en el análisis de supervivencia se tienen los siguientes objetivos:

- Estimar, a partir de un conjunto de datos, un modelo de distribución de los tiempos de supervivencia.
- Comparar los resultados obtenidos entre dos o más grupos de estudio.
- Establecer una relación entre la supervivencia de un individuo y el conjunto de *variables explicativas* que lo determinan, ejemplos de estas variables son: el tipo de sangre, el peso, la estatura o el género de una persona.

- Ajustar un modelo de regresión para las funciones de distribución del tiempo de supervivenca en función de un conjunto de variables explicativas.

El término supervivencia se atribuye a los ensayos médicos, ya que las primeras aplicaciones de este análisis se utilizaban para determinar el tiempo de muerte de un paciente. No obstante, el análisis de supervivencia es utilizado en diversas áreas de estudio como en la ingeniería, la economía, la física y la biología.

La variable tiempo puede ser considerada como una variable aleatoria continua por lo tanto, podría pensarse en ajustar una distribución normal a los datos de supervivencia, sin embargo, la pérdida de información y la presencia de datos incompletos, conocidos como *observaciones censuradas*, dan lugar a un comportamiento asimétrico en la distribución del tiempo. Debido a las complicaciones expuestas con anterioridad, el análisis de supervivencia requiere una metodología propia para llevar a cabo de manera exitosa el análisis de un conjunto datos. Por esta razón conviene definir cada uno de los elementos que componen el análisis de supervivencia antes de intentar ajustar un modelo de regresión o cualquier otro modelo estadístico a estos grupos de datos.

1.1. Tiempo de Falla

Se define al tiempo de falla T como una variable aleatoria no negativa, que se caracteriza por medir el tiempo que transcurre a partir de una fecha determinada hasta la ocurrencia de un evento de interés conocido como falla. En algunas ocasiones la variable aleatoria T puede estar asociada a una medida distinta del tiempo, como: el área, el volúmen o una carga, sin embargo, la condición de que $T \geq 0$ es estrictamente necesaria en cada uno de los casos. A continuación se presentan algunos ejemplos de tiempos de falla.

a) La duración de un matrimonio.

b) El tiempo en el que un producto es rentable en el mercado.

c) El tiempo de diseminación por metástasis de algún tipo de cáncer.

d) La longitud de onda de una frecuencia estereofónica.

Las fallas definidas en los ejemplos anteriores son: a) la ruptura matrimonial (divorcio, viudez, desaparición conyugal), b) rentabilidad de un producto en el mercado, en este ejemplo será necesario definir cuándo un producto se considera rentable, c) la diseminación por metástasis y en d) la longitud medida hasta presentar la primera interferencia sonora. En particular la falla puede ocurrir a lo más una vez en cada uno de los individuos en el experimento[1], tal y como se especifíca en los ejemplos anteriores.

[1]No necesariamente todos los individuos tienen que presentar la falla, generalmente en los estudios de supervivencia existe más de una observación que se registra sin haber reportado la falla, ya sea por que el individuo sobrevive a lo largo del estudio, o bien, por que el sujeto abandona el experimento en curso, por lo que se desconoce si este sujeto presenta la falla en algún momento determinado.

Para determinar el tiempo de falla de forma precisa es necesario:

1. Definir de manera concisa el significado de la falla.

2. Establecer un tiempo de origen que determine cuando inicia el estudio.

3. Fijar una escala para medir el tiempo (o alguna otra medida de interés) preferentemente, congruente con el diseño del experimento.

Los experimentos relacionados con la muerte de un paciente que sufre alguna determinada enfermedad ejemplifícan la importancia de definir de manera precisa el tiempo de falla. Supóngase que se quiere llevar a cabo un análisis del tiempo hasta la muerte de un grupo de pacientes con VIH, un análisis inapropiado podría recopilar información errónea con relación a la falla, en este caso la muerte, pues los pacientes pueden morir por causas ajenas a la enfermedad que padecen y entonces, se obtendrían inferencias poblacionales poco significativas con respecto a la proporción de individuos con VIH, por lo tanto, la falla deberá ser definida de manera adecuada de acuerdo con los objetivos que se persiguen en el estudio. Esto último no significa que únicamente se debe recopilar información de los sujetos que presentan la falla definida en el ensayo, pues el complemento del grupo de individuos para los cuales se presenta la falla también es parte del estudio del análisis de supervivencia.

Existen otros experimentos en los cuales existe más de una manera de que ocurra la falla, tal es el caso de la duración de un matrimonio, pues la disolución de una relación conyugal se puede dar por varios motivos, sin embargo, esto no excluye el hecho de que la falla sea definida de manera concisa, más aún, la falla podría definirse para un caso particular de ruptura matrimonial, por ejemplo el divorcio.

La *falla* se debe definir de la misma forma para cada uno de los individuos que constituyen el estudio y no se podrá cambiar su significado una vez que el estudio haya iniciado, asimismo, la falla se debe precisar de manera que ésta ocurra a lo más una vez en cada sujeto del experimento en curso.

Para determinar de manera exacta el tiempo en el que ocurre la falla es necesario definir un *tiempo de origen*. Usualmente el tiempo de origen es diferente para cada uno de los sujetos en estudio (*véase la* Figura 1.1). El tiempo de origen determina el momento en el cual un individuo presenta la posibilidad de presentar una falla.

Ejemplos de experimentos que tienen la misma fecha calendario de inicio son:

a) El tiempo en que tarda en presentarse la primera lluvia del año en algunas ciudades del país.

b) El tiempo que tarda un maratonista en abondonar una carrera.

c) El tiempo que transucurre en un partido de futbol hasta que ocurre la primera anotación.

El tiempo de origen en a) es el inicio del año en curso y es el mismo para cada una de las ciudades en el estudio, mientras que en b) y c) el tiempo de origen está sujeto al inicio de una exhibición en particular, la carrera y el comienzo del partido.

Algunos ensayos en los cuales se exhibe un punto de entrada distinto en el estudio son:

a) El tiempo que transcurre hasta que un exconvicto regresa a prisión.

b) El tiempo que tarda un determinado medicamento en hacer efecto sobre un paciente.

c) El tiempo de espera de un paciente hasta recibir un trasplante de córnea.

El tiempo de origen en a) es distinto para cada exconvicto ya que los prisioneros son liberados de prisión en fechas generalmente distintas, en este caso la fecha de liberación es el tiempo de origen del sujeto en estudio. El inciso b) representa un caso muy común en el análisis de supervivencia de ensayos clínicos, en este ejemplo se define el tiempo de origen para cada individuo como el momento en el cual el paciente recibe el medicamento, usualmente los pacientes no reciben la medicina en la misma fecha, lo que provoca un tiempo de origen distinto en el análisis de sus observaciones. Evidentemente en el inciso c) los pacientes no entran en la lista de espera de transplantes en la misma fecha, sin embargo, es posible observar el evento de interés en cada uno de ellos a partir de que ingresan a esta lista.

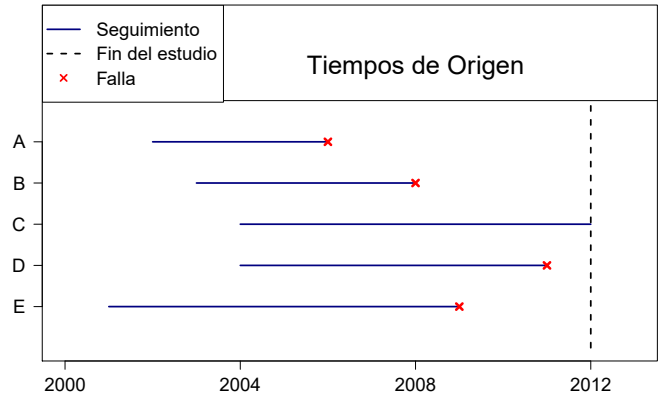


Figura 1.1: Seguimiento de un grupo de pacientes que se caracterizan por tener una fecha de entrada distinta en el estudio.

Una medida como el tiempo necesita de una escala para llevar a cabo el seguimiento de un estudio de manera ordenada. La escala para medir el tiempo en el que suceden los acontecimientos de interés puede variar desde un par de segundos hasta unos cuantos años y generalmente siempre está ubicada en una fecha del tiempo calendario.

El tiempo de supervivencia debe ser medido con una escala adecuada al problema, es decir, si se desea medir el tiempo que tarda un estudiante en concluir su carrera universitaria, sería muy poco práctico medir este proceso en horas, días o semanas, pues es más conveniente, al momento de interpretar la información, medir este tiempo en semestres o en años. Análogamente, si se desea estudiar el tiempo que tarda una partícula subatómica proveniente del Sol en penetrar la atmósfera de la Tierra, definitivamente, se optaría por trabajar con unidades de tiempo menores o iguales a los minutos.

El tiempo no es la única medida con la que se trabaja en el análisis de supervivencia, por lo que medidas asociadas a las áreas, vólumenes o longitudes deben especificar una escala de medición en el estudio, estas medidas pueden ser hectáreas, litros, yardas, etc. Ejemplos de estos ensayos son:

a) Las yardas por pase de un mariscal de campo hasta que se presenta la primera intercepción en el juego.

b) La cantidad de kilómetros que recorre un automóvil hasta que ocurre un accidente.

c) Los litros de agua que se consumen hasta perfeccionar un proceso industrial.

1.2. Censura y Truncamiento

Una de las caracteristicas principales del análisis de supervivencia es la presencia de observaciones incompletas. Estas observaciones puede ser *censuradas* o *truncadas* según sea el tipo de información que se disponga para el estudio.

El conocimiento de las observaciones censuradas y truncadas es escencial en el análisis de supervivencia, cualquier estudio que no diferencie estos registros dentro del conjunto de sus observaciones proporcionará resultados sesgados.

Censura: Se dice que una observación es censurada si el tiempo de falla ocurre en un intervalo de tiempo, siendo el tiempo de supervivencia exacto desconocido.

Una observación censurada representa información parcial en la muestra y se dice que es parcial porque es posible obtener valores incompletos con relación al tiempo de falla T.

El seguimiento de los sujetos en estudio puede verse interrumpido por alguna de las siguientes causas de censura.

- El estudio termina antes de que la falla se presente en algunos individuos.
- Los individuos presentan una falla distinta al evento de interés definido.
- Algunos sujetos deciden abandonar el estudio o se pierden antes de terminarlo.

Los ejemplos que se muestran con anterioridad presentan distintos tipos de censura, generalmente la censura es clasificada en tres categorías que más adelante darán lugar a una función de verosimilitud correspondiente a cada tipo de censura.

Las observaciones censuradas se clasifican como:

1. Censura por la derecha:
    - Tipo I
    - Tipo II
    - Tipo III ó aleatoria.
2. Censura por la izquierda.
3. Censura por intervalo.

En particular el número de observaciones censuradas está directamente relacionado con la probabilidad de ocurrencia del evento de interés en el periodo de estudio, por lo tanto, mientras más largo sea el periodo de estudio se espera un mayor número de observaciones censuradas.

Truncamiento: Se dice que una observación está truncada si ésta decide no registrase debido a que no satisface ciertas condiciones preestablecidas para pertenecer al estudio.

Por ejemplo, supóngase que se quiere determinar el tiempo que tarda un alumno en abandonar la preparatoria, sin embargo, se establece que si el alumno tiene más de 20 años cumplidos no se considerará si éste presentó o no la falla y por lo tanto, su tiempo de supervivencia está truncado.

Dado que las observaciones truncadas no son consideradas como parte de la investigación, a diferencia de los datos censurados, los registros truncados no representan ningún tipo de información de interés en el estudio.

Análogo a las observaciones censuradas, el truncamiento se clasifica en dos categorías para su estudio. Estas categorías son:

1. Truncamiento por la izquierda.
2. Truncamiento por la derecha.

Observación no censurada o completa: Decimos que el tiempo de supervivencia no está censurado si se conce el tiempo de supervivencia de un individuo y éste ocurre durante el período de estudio del experimento.

Existen observaciones que pueden ser censuradas y truncadas al mismo tiempo, asimismo, existen observaciones que presentan una combinación entre los tipos de registros anteriormente mencionados. En el análisis de supervivencia lo más común es obtener observaciones que son censuradas por la derecha y truncadas por la izquierda.

1.2.1. Censura por la derecha

Es el tipo de censura más común en el análisis de supervivencia y se divide en tres categorías.

Censura tipo I

Este tipo de censura está directamente relacionado con la duración del estudio, por lo general, por cuestiones económicas, el tiempo de estudio es finito, es decir, se establece una fecha determinada a partir de la cual ya no se aceptan más registros con relación a la falla de los individuos, por lo que se desconoce el tiempo exacto de ocurrencia del evento definido como falla.

Este tipo de censura se presenta en numerosos ensayos clínicos. Por ejemplo, para aprobar una nueva cepa vacunal los científicos realizan pruebas en algunos animales para determinar la eficencia de la vacuna. En diversas ocasiones los investigadores tienen un límite de tiempo para observar los efectos de esta cepa, debido a esta condición temporal algunos de los animales en estudio podrían no presentar el evento de interés antes de concluir el experimento, por lo cual, el tiempo exacto de supervivencia de los animales no es conocido y por lo tanto, estos registros se consideran censurados.

En la *censura tipo I* las observaciones que presentan la falla antes del tiempo de finalización del estudio se consideran como observaciones no censuradas, por el contrario, si el individuo no presenta la falla durante este período de tiempo, la observación se considera censurada. Si no se pierde el seguimiento de ningún individuo a lo largo del estudio, el *tiempo censurado* de los individuos es igual al tiempo de realización del estudio.

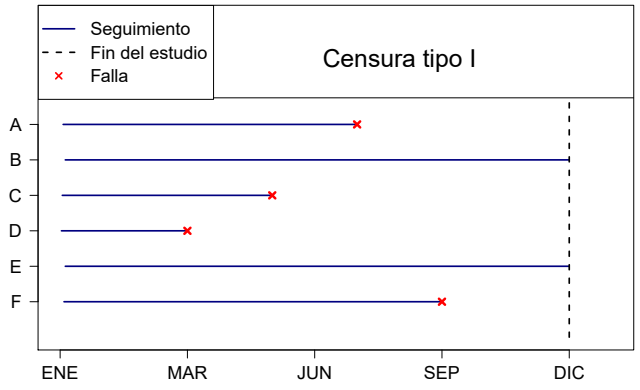


Figura 1.2: Los sujetos B y E no presentaron la falla antes de antes de la fecha de terminación del estudio por lo que su registro queda censurado.

Para una variable aleatoria T y C_r una observación censurada por la derecha con T_i y C_{r_i} el tiempo de falla y de censura del i-ésimo individuo respectivamente. Decimos que T_i es observación censurada por la derecha si:

$$T_i > C_{r_i}$$

En el caso de la censura tipo I, el tiempo de falla T_i se observa si y sólo si $T_i \leq C_{r_i}$ por lo que tendremos los datos del estudio representados por la pareja (t_i, δ_i) donde:

$$t_i = min(T_i, C_{r_i})$$

y δ_i es una función indicadora definida como:

$$\delta_i = \begin{cases} 1 & \text{si } \ T_i \leq C_{r_i} \\ \\ 0 & \text{si } \ T_i > C_{r_i} \end{cases}$$

Censura tipo II

Supóngase que tenemos el mismo experimento presentado para la censura tipo I y supóngase ahora que el científico decide terminar el experimento en el momento en el que r de los n idividuos en estudio $(r < n)$ presenten la falla, evidentemente el tiempo de supervivencia de los $(n - r)$ individuos restantes no es conocido con exactitud, dado que el sujeto pudo perderse durante el seguimiento o bien, pudo haber sobrevivido hasta el momento en el que se dió por terminado el estudio y por lo tanto, se registran $(n - r)$ observaciones censuradas.

La *censura tipo II* ocurre cuando n individuos comienzan un estudio al mismo tiempo y éste termina cuando se presentan las primeras r fallas. Las r-fallas son parte del conjunto de observaciones no censuradas, mientras que las $(n - r)$ observaciones restantes son observaciones con censura tipo II.

A diferencia de la censura tipo I, las observaciones con censura tipo II están relacionadas directamente con el tamaño de la muestra, si no se pierde el seguimiento de un individuo hasta antes del momento de que se observe la r-ésima falla, los *tiempos censurados* igualan al tiempo de falla de la r-ésima observación.

En el caso de la censura tipo II, decimos que el tiempo de falla del i-ésimo individuo T_i esta censurado por la derecha si:

$$T_i > C_{r_i} = T_{(r)}$$

donde $T_{(r)}$ representa la r-ésima estadística de orden. Es claro que el tiempo de falla T_i se observa solamente si $T_i \leq T_{(r)}$. Análogo a la censura tipo I, podemos representar a los datos como una pareja (t_i, δ_i) donde:

$$t_i = min(T_i, T_{(r)})$$

y δ_i es una función indicadora definida como:

$$\delta_i = \begin{cases} 1 & \text{si} \quad T_i \leq T_{(r)} \\ \\ 0 & \text{si} \quad T_i > T_{(r)} \end{cases}$$

Para efectos de reducir costos en un experimento, es conveniente definir el problema en términos de la censura. Si los objetos de estudio tienen poca probabilidad de presentar una falla, los experimentos con censura tipo II podrían resultar más costosos que los definidos bajo la censura tipo I. Por el contrario, si los sujetos de estudio presentan la falla en repetidas ocasiones, un experimento con un tiempo de duración largo tendría costos más elevados.

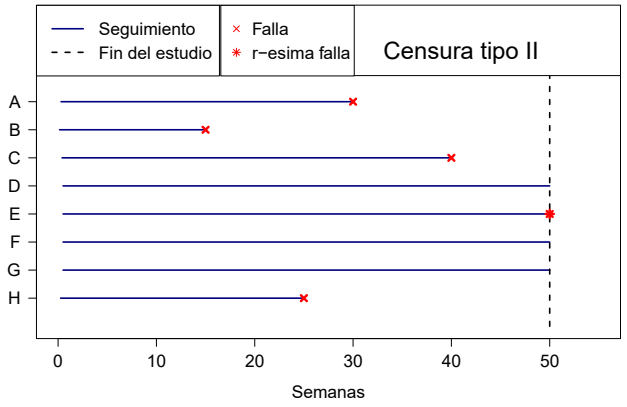


Figura 1.3: Este experimento termina en la semana 50, dado que en este tiempo ocurre la r-ésima falla, de modo que los sujetos D, F y G quedan censurados.

Censura tipo III o aleatoria

La censura tipo III también conocida como *censura aleatoria*, se presenta cuando los individuos no tienen una salida controlada en el estudio, ejemplos de este tipo de censura son:

- Los individuos abandonan el estudio antes de presentar la falla.
- El sujeto muere o presenta una falla distinta a la que se definió, lo cual le impide continuar en el seguimiento.
- La pérdida del expediente médico de algunos pacientes.

Evidentemente los motivos por los cuales se pierde el seguimiento de un sujeto en los ejemplos anteriores son de carácter aleatorio, por lo que este tipo de censura surge cuando los tiempos de censura para cada uno de los individuos son considerados como una variable aleatoria C_i independiente del tiempo de falla T_i.

Siguiendo la misma metodología utilizada para la censura tipo I y tipo II, se representan los datos obtenidos como una pareja (t_i, δ_i) con $t_i = min(T_i, C_i)$ y

$$\delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i \\ \\ 0 & \text{si } T_i > C_i \end{cases}$$

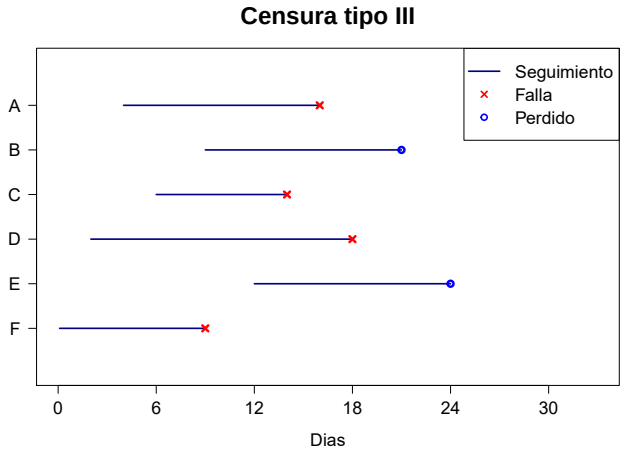


Figura 1.4: Salida aleatoria del estudio.

1.2.2. Censura por la izquierda

Este tipo de censura se presenta cuando el tiempo de falla de un individio es menor al tiempo en el que se presentó la censura, por lo general este tipo de censuras registran el evento de intéres momentos antes de que se inicie el estudio. Por ejemplo, si se desea determinar el tiempo que tarda en aprender a leer y escribir un niño durante el primer año de primaria, muy probablemente existan alumnos que cursen ese grado escolar que ya tengan la capacidad de leer y escribir antes de que ingresen a la primaria, por lo tanto, estos niños serán parte del conjunto de observaciones censuradas por la izquierda.

Análogo a la definición de censura por la derecha, se define a la cesura por la izquierda C_l en términos de la variable aleatoria del tiempo de falla T.

Sean T_i y C_{l_i} el tiempo de falla y de censura del i-ésimo individuo respectivamente. Decimos que T_i es una observación censurada por la izquierda si:

$$T_i < C_{l_i}$$

Esta definición se satisface siempre y cuando C_{l_i} sea menor que un tiempo límite del estudio, o bien, $C_{l_i} < T_{(r)}$ en el caso de que el experimento finalice al momento de observar r fallas en un conjunto de n individuos. Bajo estas condiciones podemos representar a los datos como una pareja (t_i, δ_i), donde

$$t_i = max(T_i, C_i)$$

y δ_i una función indicadora definida como:

$$\delta_i = \begin{cases} 1 & \text{si} \quad T_i \geq C_{l_i} \\ \\ 0 & \text{si} \quad T_i < C_{l_i} \end{cases}$$

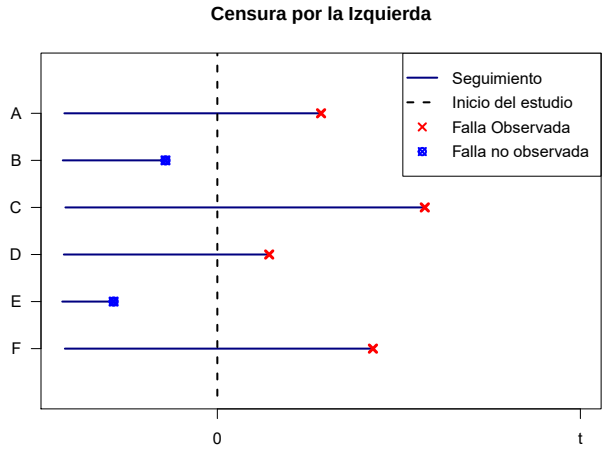


Figura 1.5: El evento de interés se presenta anticipadamente en los sujetos B y E.

1.2.3. Censura doble

En algunos ensayos es posible encontrarse con una observación *doblemente censurada*, es decir, que está censurada por la izquierda y por la derecha, ejemplos de estas observaciones se presentan de manera cotidiana en las encuestas. Supóngase que se le pregunta a un conjunto de individuos si alguna vez han jugado al "melate", algunos individuos podrán constestar afirmativamente esta pregunta, sin embargo, uno de los participantes podría no recordar la fecha exacta en la que participó en el concurso y por lo tanto, se tendría una observación censurada por la izquierda, por el contrario, si un encuestado presume jamás haber jugado al "melate" se tendría una observación censurada por la derecha, pues aunque en un futuro el individuo decida ser partícipe del juego, el tiempo en el que el individuo entra al juego será desconocido para el encuestador.

En particular el tiempo de falla T_i para el i-ésimo individuo será observado si:

$$C_{l_i} \leq T_i \leq C_{r_i}$$

En el caso de que exista *doble censura*, los datos pueden ser representados por la pareja (t_i, δ_i) donde:

$$t_i = max(min(T_i, C_{r_i}), C_{l_i})$$

y δ_i definida como:

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

1.2.4. Censura por intervalo

La censura por intervalo ocurre cuando se determina que la falla ocurrió en un intervalo de tiempo específico sin tener el conocimiento exacto del tiempo en el que sucedió el evento de interés. Como ejemplo ilustrativo se podría pensar en el momento en el cual ocurrió la primera lluvia del año pasado, muy probablemente se recuerde entre que meses o incluso entre que días comenzó a llover pero difícilmente se podrá recordar la fecha exacta en la que ocurrió este acontecimiento.

Sea el intervalo (L_i, R_i) para el i-ésimo individuo, decimos que T_i es una observación censurada si:

$$T_i \in (L_i, R_i)$$

siempre y cuando el tiempo exacto de supervivencia T_i sea desconocido.

1.2.5. Truncamiento por la izquierda

También conocido como *tiempo de retraso de entrada al estudio*, el truncamiento por la izquierda es el más común en el análisis de supervivencia. Cuando un individuo presenta la falla antes de incorporarse al estudio y éste es excluído del ensayo debido a esa situación, se dice que el sujeto tiene una observación *truncada por la izquierda*.

El concepto de truncamiento por la izquierda no debe confundirse con el de la censura por la izquierda debido a que este último proporciona información parcial de la muestra, mientras que un dato truncado es totalmente excluído del estudio.

Para ejemplificar estas observaciones supóngase que en una comunidad se quiere determinar el tiempo de supervivencia de aquellas personas que pertenecen a la tercera edad (mayor a 60 años), si un individuo muere después de los 60 años será considerado como parte del estudio, sin embargo, si la edad de muerte de un individuo es menor a la edad señalada este sujeto será excluído del análisis y por lo tanto, considerado como una observación truncada por la izquierda.

Otro ejemplo es aquel experimento que tiene la tarea de proporcionar el diámetro de partículas microscópicas. En particular el investigador sólo va a recopilar información de aquellas partículas que sean lo suficientemente grandes a la vista de un microscopio dejando a un lado aquellas partículas que no pueden ser percibidas por este instrumento de laboratorio, lo que provoca que estas últimas observaciones se

registren como datos truncados por la izquierda.

De manera general para el truncamiento (por izquierda y por derecha) el tiempo de falla T_i para el i-ésimo individuo ocurre o es censurado si:

$$T_i \in (L_i, R_i)$$

donde L_i y R_i representan un intervalo cuyos extremos son los eventos de truncamiento del experimento.

Si $T_i > R_i$ o bien $T_i < L_i$ entonces T_i es una observación truncada.

Se dice que una observación está truncada por la izquierda si:

$$T_i < L_i$$

y se dice que es observada dentro del intervalo (L_i, ∞) si:

$$L_i \leq T_i$$

es decir, el tiempo de supervivencia excede al tiempo de truncamiento L_i.

1.2.6. Truncamiento por la derecha

Decimos que una observación está truncada por la derecha si en el estudio no se incluye esta observación debido a que el sujeto no presentó la falla o evento de interés.

Un ejemplo de truncamiento por la derecha surge cuando se estudia el tiempo que transcurre entre la infección por el virus VIH hasta el desarrollo de la enfermedad, lo que se conoce como periodo de latencia del virus del SIDA. Si el paciente enferma de SIDA durante el tiempo de estudio del ensayo clínico su observación quedará registrada, sin embargo, si el individuo no llegara a presentar la enfermedad hasta antes de la fecha de finalización del estudio su observación quedará truncada por la derecha, dado que el sujeto, no presentó el evento de interés.

Un segundo ejemplo ocurre al estimar la distancia que separa la Tierra de las estrellas. Si la estrella se encuentra cercana a la Tierra ésta será considerada para el estudio, estrellas muy distantes no serán consideradas por lo tanto, las estrellas lejanas son observaciones truncadas por la derecha.

Sea el intervalo (L_i, R_i) para el i-ésimo individuo, con $L_i = 0$, se dice que el tiempo de supervivencia T_i es observado si:

$$T_i \leq R_i$$

de lo contrario si:

$$T_i > R_i$$

entonces se dice que el i-ésimo individuo representa una observación truncada por la derecha.

1.3. Funciones de Riesgo y Supervivencia

En esta sección se presentan las funciones más importantes en el análisis de supervivencia. El objetivo de estas distribuciones es obtener información que nos permita obtener características de interés de una población. En el caso de los modelos paramétricos el problema se reduce a hacer inferencia sobre uno o varios parámetros con base en el valor proporcionado de los tiempos de falla.

La distribución de los tiempos de supervivencia también puede describirse a partir de su función de densidad, por lo que en esta sección se incluye el estudio de ésta y otras funciones que nos ayudan a estudiar las características de la variable aleatoria T.

1.3.1. Función de Supervivencia

Sin importar que se utilice un modelo paramétrico o un modelo no paramétrico para analizar datos de supervivencia, es posible asociar a la variable aleatoria T una función que nos permita determinar la probabilidad de que un individuo sobreviva (no falle) más allá del tiempo t. Esta función es conocida como *función de supervivencia* y se define de la siguiente manera.

Definición 1.3.1 (Función de Supervivencia)**.** Sea T una variable aleatoria que toma valores no negativos, la función de supervivencia de la variable aleatoria T es la función $S(t) : \mathbb{R}^+ \cup \{0\} \to [0, 1]$, que satisface

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

En el caso *no paramétrico*, si no existe censura, la función de supervivencia coincide con la proporción de individuos supervivientes en el tiempo t.

La siguiente relación, nos permite expresar a la función de supervivencia $S(t)$, en términos de la función de distribución:

$$S(t) = 1 - P(T \leq t) = 1 - F(t)$$

en donde $F(t)$ representa la probabilidad de que un individuo presente la falla antes del tiempo t. En particular la función de distribución de una variable aleatoria, es una función que se caracteriza por ser continua por la derecha, monotona no decreciente y con límites 1 y 0 cuando $t \to +\infty$ y $t \to -\infty$ respectivamente[2]. De la relación anterior es posible deducir las siguientes propiedades para la función de supervivencia.

Proposición 1.3.2. Sea T una variable aleatoria y $S(t)$ su función de supervivencia, entonces:

1. $S(0) = 1$
2. $\lim_{t\to\infty} S(t) = 0$

[2]Debido a que el tiempo es una variable aleatoria no negativa, en el análisis de supervivencia la expresión $t \to -\infty$ es remplazada por la expresión $t \to 0$ o simplemente por $t = 0$.

3. $S(t)$ es una función continua por la derecha.

4. Si $t_1 \leq t_2$, entonces $S(t_1) \geq S(t_2)$, es decir, $S(t)$ es una función monotona no creciente.

Una forma de describir los datos provenientes del análisis de datos de supervivencia es con una representación gráfica de $S(t)$. La gráfica de $S(t)$ es conocida como *curva de supervivencia* y en muchas ocasiones permite identificar la distribución que mejor se aproxime a las distribución de la variable aleatoria T.

A continuación se presentan algunos ejemplos de *curvas de supervivencia*.

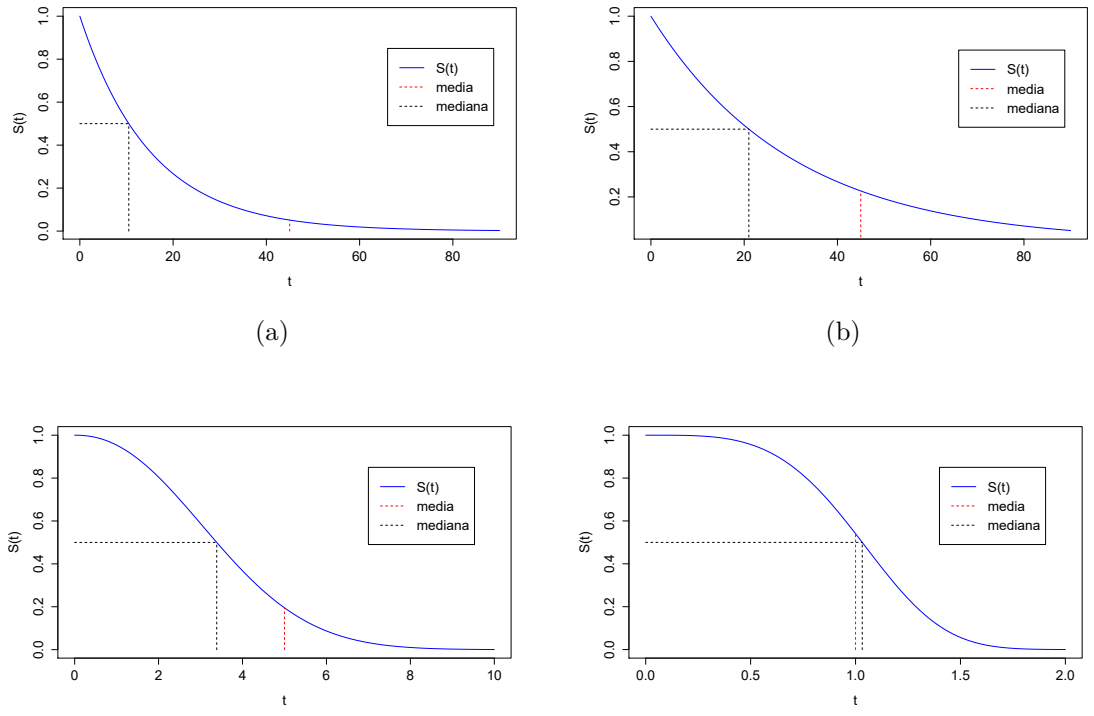


Figura 1.6: Curvas de Supervivencia

A partir de las curvas de supervivencia se pueden obtener características numéricas de la variable aleatoria T, como la media, la mediana y otros percentiles de interés. En las gráficas anteriores se observa una alta concentración de eventos cerca del origen y una baja densidad de sucesos en la cola derecha para cada una de las curvas de supervivencia. El sesgo de estas distribuciones sugiere utilizar a la *mediana* como la medida de tendecia central más representativa, debido a que la *media* se ve afectada por los valores cercanos a las colas de las curvas. Esta situación se puede ver reflejada en las gráficas a), b) y c) de la Figura 1.6, sin embargo, existen ocasiones en las que la *media* funciona como un buen estimador para la tendencia central tal y como se puede observar en el cuadrante d) de la figura anterior.

Las curvas de supervivencia también se utilizan para comparar, de manera ambigua, dos o más grupos de estudios. La curva de supervivencia a) de la Figura 1.6 representa una tasa de supervivencia más baja que la curva expuesta en la gráfica b), de manera general las curvas que tienden a aproximarse más rápido a cero tiene un tiempo de supervivencia más corto que las que se aproximan de manera menos anticipada. Las curvas de supervivencia como las que aparecen en las gráficas c) y d) ejemplifican tiempos más largos de supervivencia.

Los modelos paramétricos permiten identificar la distribución que mejor se aproxima a la verdadera distribución de las observaciones, esta idendificación se puede lograr realizando gráficas de diagnóstico como histogramas y qq plots, o bien se consigue utilizando un contraste estadístico como la prueba de la ji-cuadrada, la prueba de Kolmogorv-Smirnov o la prueba modificada de Lilliefors[3] con el propósito de verificar si se satisface una hipótesis con respecto a la distribución de los datos proporcionados.

En [7] y [8] se detallan algunas técnicas que ayudan a identificar a la función de distribución de T, no obstante, será conveniente conocer otras funciones relacionadas con $S(t)$ que nos proporcionen más información relacionada con el tiempo de falla.

1.3.2. Función de Densidad

Como cualquier otra variable aleatoria, el tiempo de falla T tiene una función de densidad asociada, dependiendo de las características de su función de distribución. Generalmente el tiempo de falla se mide de manera continua, sin embargo, existen ocasiones en la que el tiempo conviene representarse de manera discreta.

En esta sección se presenta la *función de densidad*, también conocida como *tasa de falla incondicional*, en términos de la función de supervivencia.

Caso Continuo

Haciendo uso de las propiedades de la función de distribución se tiene lo siguiente.

Definición 1.3.3. La función de de densidad, es una función no negativa f que satisface

$$\int_0^\infty f(t)\,dt = 1$$

en particular si f es una función de densidad, entonces la función de supervivencia $S(t)$ se define como:

$$S(t) = \int_t^\infty f(x)\,dx$$

[3]La prueba modificada de Lilliefors considera la suma de todas las diferencias absolutas entre la función de distribución exponencial y la función de distribución empírica de la muestra. Esta prueba es una extención de la prueba de Lilliefors y se utiliza en los modelos de duración y en el análisis de supervivencia.

De la ecuación anterior es posible inferir las siguientes propiedades:

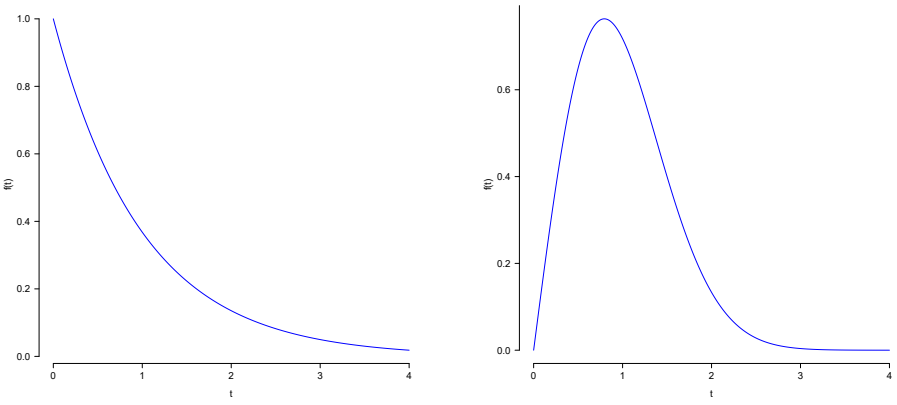
$$\begin{aligned} F(t) &= 1 - S(t), \quad \forall\, t \geq 0 \\ &= 1 - \int_t^{\infty} f(x)\, dx \end{aligned}$$

finalmente, por el teorema fundamental del cálculo, tenemos que:

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

Una aplicación inmediata que tiene la función de densidad es proporcionar la probabilidad asignada a un intervalo de valores del tiempo de falla, lo que nos permite conocer la proporción de individuos que pertenecen a este intervalo, asimismo, la función de densidad se puede utilizar para encontrar la *moda* del tiempo de falla, es decir, aquel punto en donde la función de densidad alcanza un máximo local.

Gráficamente la función de densidad ayuda al estadístico a sugerir un modelo de distribución teórico para la variable aleatoria T, la manera más sencilla de establecer este modelo es a tráves del suavizamiento del histograma que se obtiene al registrar los datos procedentes de los tiempos de falla. La gráfica de la función de densidad nos permite identificar la forma en que se separan y acumulan los valores, es decir, nos proporciona información relacionada con sus coeficientes de asimetría y curtósis.



(a) (b)

Figura 1.7: Funciones de densidad

La gráfica a) de la figura anterior presume de tener una alta tasa de fallas al principio del estudio y evidentemente, la probabilidad de sobrevivir se considera demasiado baja. La Figura 1.7 b) presenta una curva que es asimétricamente positiva, por lo que los valores se tienden a reunir en la parte izquierda de la media, en la cima de esta gráfica se encuentra la máxima frecuencia de fallas observadas.

Caso discreto

Los casos en los que la variable aleatoria T es discreta no son muy comunes en la práctica, sin embargo, se presentan cuando los tiempos de falla son agrupados en intervalos o se decide trabajar con unidades de tiempo discretas como días, semanas o años.

Definición 1.3.4. Sea T una variable aleatoria discreta que toma valores $\{t_i\}_{i\in\mathbb{N}}$ donde $0 \leq t_1 < t_2 < \ldots$, se define la función de densidad o de *masa de probabilidad* como:

$$f(t_i) = P(T = t_i)$$

entonces la función de supervivencia está dada por:

$$S(t) = P(T > t) = \sum_{t<t_i} f(t_i)$$

1.3.3. Función de Riesgo

La *función de riesgo* (hazard[4] function) es quizás la función más importante del análisis de supervivencia y la más utilizada en los modelos de regresión de los tiempos de falla. Esta función, también conocida como la *tasa instantánea de mortalidad*, se define como la probabilidad de que un individuo presente la falla en un instante de tiempo t siempre y cuando el individuo haya sobrevivido hasta ese instante. El hecho de que un individuo presente una falla en un instante significa que la falla ocurre en un intervalo de tiempo $[t, t + \Delta t)$ infinitamente pequeño, por lo tanto es posible definir la función de riesgo como el límite de la probabilidad de observar la falla en el intervalo $[t, t + \Delta t)$ dado que el individuo ha sobrevivido hasta el tiempo t.

Definición 1.3.5. Sea T una variable aleatoria continua y no negativa, se define a la función de riesgo del tiempo de supervivencia como:

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} \tag{1.1}$$

La función anterior tiene la característica de ser una función no negativa, esto es, que para todo $t \in T$ se tiene que $h(T) \geq 0$ y al igual que la función de supervivencia y la función de densidad, la función $h(t)$ proporciona información que puede ser significativa al momento de seleccionar un modelo probabilístico para los tiempos de falla.

Entre otras cosas, la función de riesgo $h(t)$ es utilizada para determinar el riesgo de falla por unidad de tiempo, en particular la cantidad $h(t)\Delta t$, derivada de la ecuación (1.1), refleja la proporción de individuos de edad t que se espera que fallen en el intervalo de tiempo $[t, t + \Delta t)$. De manera general la función $h(t)$ describe la variación de la tasa instantánea de que falle un individuo a lo largo del tiempo.

[4]En particular la palabra "hazard", era el nombre de un antiguo juego de dados que se jugaba en Europa, sin embargo, la raíz etimológica del término proviene de la palabra árabe "al-azar" que significa dado.

La función de riesgo puede reducirse a una expresión más sencilla llevando a cabo el cálculo del límite que aparece en la Definición 1.3.5. Recordando que T es una variable aleatoria continua se tiene que:

$$\begin{aligned} h(x) &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t \, P(T \geq t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \\ &= -\frac{d}{dt} log\, S(t) \end{aligned}$$

por lo tanto, la función de riesgo se puede escribir como:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} log\, S(t) \tag{1.2}$$

de la igualdad anterior es posible obtener a la función de supervivencia en términos de la función de riesgo, por lo tanto, despejando $S(t)$ de la igualdad (1.2) tenemos:

$$\begin{aligned} -log(S(t)) &= \int_0^t h(x)\, dx \\ S(t) &= exp\Big(- \int_0^t h(x)\, dx \Big) \end{aligned}$$

sea

$$H(t) = \int_0^t h(x)\, dx \tag{1.3}$$

entonces

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

La expresión $H(t)$ obtenida en (1.3) es conocida como la *función acumulada de riesgo* y se caracteriza por presentar mayor o igual incidencia de riesgo conforme avanza el tiempo, por lo tanto, es siempre una función monótona no decreciente.

Las gráficas que describen las funciones de riesgo son útiles para comparar el riesgo al que están sometidos dos o más grupos de estudios. Algunos ejemplos de estas gráficas se presentan en la Figura 1.8.

Una gráfica constante como la que aparece en la Figura 1.8 a) considera el mismo riesgo en toda la historia del experimento, estas gráficas representan el riesgo de una distribución exponencial.

Las funciones de riesgo crecientes, como en la figura 1.8 b), indican que los individuos presentan la falla conforme avanza el tiempo, este tipo de riesgo es común en la vida útil de algunos electrodomésticos, máquinas industriales y productos perecederos.

Un comportamiento decreciente como el que se muestra en la Figura 1.8 c) de la función $h(t)$, señala que la probabilidad de caer en el estado de falla es más alta en las etapas tempranas del ensayo, ejemplos de esta situación son: el tiempo de vida de un recién nacido desde el parto hasta sus primeros años de vida, el reclamo de una garantía o los errrores que puede tener un nuevo software estadístico.

La curva de tubo de baño Figura 1.8 d) es una gráfica muy representativa de las funciones de riesgo en el análisis de supervivencia y por lo general describen el riesgo de muerte de un indivuduo a lo largo de su vida. Naturalmente estas gráficas están compuestas de tres etapas, en la primera el riesgo es decreciente, en la segunda el riesgo permance constante y cuando se aproximan los últimos años esta gráfica tiene un comportamiento creciente.

Por último una función de riesgo en forma de montaña Figura 1.8 e) podría representar el riesgo de muerte por una enfermedad después de haber recibido un tratamiento especial.

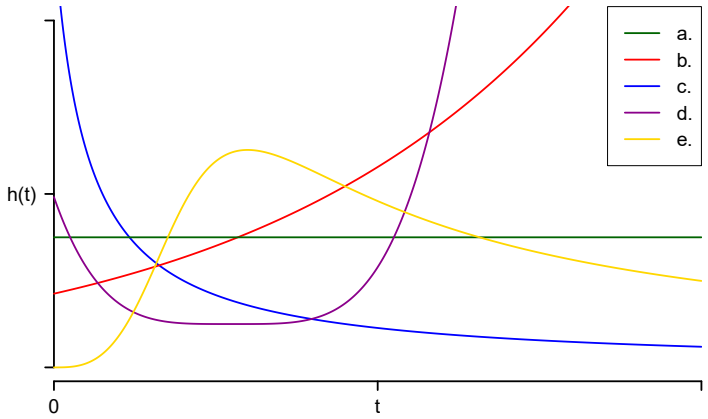


Figura 1.8: Funciones de riesgo.

Caso discreto

Cuando T es una variable aleatoria discreta que toma valores $\{t_i\}_{i\in\mathbb{N}}$, la función de riesgo se define como la probabilidad de presentar la falla a tiempo $t = t_i$, dado que el individuo sobrevive antes de t_i, es decir

$$h(t_i) = P(T = t_i \,|\, T \geq t_i)$$

resolviendo la probabilidad condicional se tiene

$$\begin{aligned} h(t_i) &= \frac{P(T = t_i)}{P(T \geq t_i)} \\ &= \frac{f(t_i)}{S(t_{i-1})} \end{aligned}$$

obsérvese que cuando T es una v.a discreta la función $S(t_i) \neq S(t_{i-1})$ pues

$$P(T \geq t_i) = 1 - P(T < t_i) = S(t_{i-1}) \neq S(t_i)$$

Al igual que el caso continuo la *función de riesgo acumulado* $H(t)$, acumula los valores de la función de riesgo $h(t_i)$ hasta el momento t de la siguiente manera

$$H(t) = \sum_{t_i \leq t} h(t_i) \tag{1.5}$$

No obstante, este resultado no satisface la igualdad (1.4), por lo cual, existe una forma alternativa de aproximar a $H(t)$ en términos de $S(t)$, por el momento se limitará a presentar esta función, sin embargo, más adelante se realizará el procedimiento que explica este resultado. La función de riesgo acumulado queda definida como:

$$H(t) = -\sum_{t_i \leq t} log[1 - h(t_i)] \tag{1.6}$$

Las definiciones (1.5) y (1.6) preservan las propiedades que caracterizan a la función de riesgo acumulado, es decir, son funciones no decrecientes y no negativas.

1.3.4. Relaciones de las funciones de supervivencia

Las funciones de supervivencia presentadas con anterioridad guardan una relación con las demás y es posible definir cada una de ellas en términos de las otras funciones.

Caso continuo

De la definición de función de supervivencia sabemos que:

$$S(t) = 1 - F(t)$$

derivando la expresión anterior

$$\begin{aligned} \frac{d}{dt}S(t) &= \frac{d}{dt}[1 - F(t)] \\ S'(t) &= -f(t) \end{aligned}$$

multiplicando ambos lados de la última igualdad por $-1/(1 - F(t))$, y haciendo uso de la igualdad (1.2)

$$-\frac{S'(t)}{1 - F(t)} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = h(t) \tag{1.7}$$

sustituyendo a $f(t)$ por $-S'(t)$ en la igualdad anterior

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

integrando $h(\cdot)$ de 0 a t se tiene que

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

utilizando la igualdad (1.4) obtenemos

$$H(t) = -log(S(t))$$

multiplicando por un -1 y aplicando la exponencial en ambos lados se llega a:

$$S(t) = exp(-H(t)) = exp\Big(-\int_0^t h(x)\,dx\Big) \tag{1.8}$$

de (1.7) podemos escribir a $S(t)$ como

$$S(t) = f(t)/h(t)$$

despejando a $f(t)$ y sustituyendo $S(t)$ por el resultado en (1.8) se concluye

$$f(t) = h(t)\,exp(-H(t))$$

Evidentemente solo se requiere conocer una de las funciones anteriores para encontrar las otras, lo que es de gran utilidad en la práctica, pues en ocasiones, la función de riesgo tiene un aspecto mas amigable que el de cualquier función de densidad.

Caso discreto

Partiendo nuevamente de la definición de la función de supervivencia $S(t)$, se tiene que

$$S(t) = 1 - F(t) = 1 - \sum_{t_i \leq t} f(t_i) = \sum_{t < t_i} f(t_i)$$

con base en la igualdad anterior se obtiene

$$\begin{aligned} S(t_{i-1}) &= \sum_{t_{i-1}<t_j} f(t_j) = f(t_i) + f(t_{i+1}) + f(t_{i+2}) + \dots \\ S(t_i) &= \sum_{t_i<t_j} f(t_j) = f(t_{i+1}) + f(t_{i+2}) + f(t_{i+3}) + \dots \end{aligned}$$

la siguiente suma telescópica $S(t_{i-1}) - S(t_i)$ da como resultado

$$f(t_i) = S(t_{i-1}) - S(t_i)$$

dividiendo entre $S(t_{i-1})$ tenemos:

$$h(t_i) = \frac{f(t_i)}{S(t_{i-1})} = \frac{S(t_{i-1}) - S(t_i)}{S(t_{i-1})} = 1 - \frac{S(t_i)}{S(t_{i-1})} \tag{1.9}$$

despejando a $S(t_i)$ de la ecuación (1.9)

$$S(t_i) = [1 - h(t_i)]\, S(t_{i-1})$$

obsérverse que $S(t_i)$ es una función de carácter recursivo

$$S(t_1) = [1 - h(t_1)]\, S(0) = [1 - h(t_1)]$$

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

$$S(t_3) = [1 - h(t_3)]\, S(2) = [1 - h(t_3)][1 - h(t_2)][1 - h(t_1)]$$

por lo que:

$$S(t) = \prod_{t_i<t} [1 - h(t_i)] = \prod_{t_i<t} \frac{S(t_i)}{S(t_{i-1})}$$

calculando el logaritmo del primer producto que aparece en la expresión de arriba se obtiene

$$log(S(t)) = \sum_{t_i<t} log[1 - h(t_i)]$$

si multiplicamos por -1 a la igualdad anterior, se obtiene la expresión (1.6), es decir, la función de riesgo acumulado

$$H(t) = -\sum_{t_i<t} log[1 - h(t_i)]$$

Aunque es posible obtener una equivalencia directa entre las funciones de distribución de los tiempos de falla, a diferencia del caso continuo, los cálculos de estas funciones pueden llegar a ser muy complicados, por lo tanto, es conveniente utilizar algún software que facilite los cálculos.

Ejemplo 1.3.1.

Sea $h(t) = a+bt$ una función de riesgo continua con $a, b > 0$, se obtienen las funciones $H(t)$, $S(t)$ y $f(t)$ como:

$$\begin{aligned} H(T) &= \int_0^t h(x)\,dx \\ &= \int_0^t (a+bx)\,dx \\ &= at + b\frac{t^2}{2} \end{aligned}$$

en consecuencia

$$\begin{aligned} S(t) &= exp(-H(t)) \\ &= exp\Big(-\Big(at + \frac{bt^2}{2}\Big)\Big) \end{aligned}$$

y por último

$$\begin{aligned} f(t) &= h(t)S(t) \\ &= (a+bt)\,exp\Big(-\Big(at + \frac{bt^2}{2}\Big)\Big) \end{aligned}$$

1.4. Modelos Paramétricos

Los modelos paramétricos de distribución se utilizan cuando puede suponerse un modelo probabilístico para la población a partir de las observaciones que constituyen la muestra.

Los modelos paramétricos más comunes dentro del análisis de supervivencia son:

- Modelo Exponencial
- Modelo Weibull
- Modelo Log-normal
- Modelo Gamma

1.4.1. Modelo Exponencial

La distribución exponencial es un caso particular de la familia de distribuciones gamma y es utilizada, entre otras cosas, para modelar experimentos en los que se quiere conocer el tiempo que transcurre hasta que se produce un evento de interés,

generalmente en cortos periodos de tiempo. Esta distribución tiene muchas aplicaciones en el análisis de supervivencia y su importancia es tan relevante como la de la distribución normal en la inferencia tradicional, sin embargo, existe una propiedad conocida como *pérdida de la memoria* que caracteriza a esta distribución y que limita el uso de los modelos exponenciales en el análisis de supervivencia.

Algunos de los problemas que pueden ser modelados con la distribución exponencial en el análisis de supervivencia son:

- Determinar el tiempo de falla de componentes de equipos electrónicos.
- Estimar el tiempo que tarda en desintegrarse una partícula radioactiva (útil para la datación de fósiles mediante la técnica del carbono-14 ^{14}C).
- Modelar el tiempo que tarda un paciente en recibir atención médica.

Definición 1.4.1. Se dice que una variable aleatoria T sigue una distribución exponencial con parámetro $\lambda > 0$ si su función de densidad está dada por:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t > 0 \\ \\ 0 & \text{si } t \leq 0 \end{cases}$$

La función de distribución de T es:

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

Utilizando las relaciones vistas en la *sección* 1.3.4 podemos obtener a su función de supervivencia, de riesgo y de riesgo acumulado de la manera siguiente.

Función de supervivencia:

$$\begin{aligned} S(t) &= 1 - F(t) \\ \\ &= e^{-\lambda t} \end{aligned}$$

Función de riesgo:

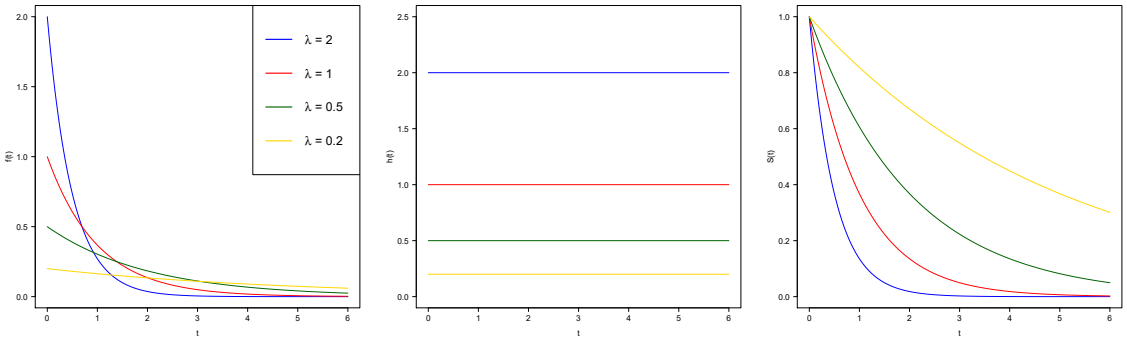
$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\ \\ &= \lambda \end{aligned}$$

en particular, se puede observar que el riesgo es constante y por lo tanto la falla no depende del tiempo t. Valores grandes en λ representan un riesgo más elevado.

Función de riesgo acumulado:

$$\begin{aligned} H(t) &= -log(S(t)) \\ &= \lambda t \end{aligned}$$

En la Figura 1.9, se pueden observar las funciones de densidad, de riesgo y de supervivencia para diferentes valores del parámetro λ.



(a) Función de Densidad (b) Función de Riesgo (c) Función de Supervivencia

Figura 1.9: Modelo Exponencial

El conocimiento de la distribución de la variable aleatoria T permite obtener características númericas de interés como; la esperanza, la varianza o algún momento específico de la variable aleatoria.

Esperanza:

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

Segundo momento:

$$E(T^2) = \frac{2}{\lambda^2}$$

***n*-ésimo momento:**

$$E(T^n) = \frac{n!}{\lambda^n} \qquad n \in \mathbb{N}$$

Varianza:

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

Como se puede observar en la Figura 1.6, en el análisis de supervivencia, cuantiles como la mediana suelen proporcionar información más relevante que la media, por lo que conviene presentar su cálculo.

Dado que $S(t) = exp(-\lambda t)$ entonces:

$$-\lambda t = log(S(t))$$

por lo tanto, despejando el valor de t se tiene que el ***p*-ésimo cuantil** de la distribución exponencial está dado por:

$$t_p = -\lambda^{-1} log(1-p)$$

Pérdida de memoria

Se dice que la distribución de una variable aleatoria T satisface la *propiedad de pérdida de la memoria* si; para cualesquiera valores $s, t > 0$ se verifica:

$$P(T > t + s \,|\, T > s) = P(T > t) \tag{1.10}$$

Esta propiedad refleja que el tiempo de ocurrencia de un evento no depende de los acontecimientos ocurridos en el pasado, por ejemplo, la propiedad de pérdida de memoria indica que el tiempo de vida restante de un componente eléctrico es independiente de su antigüedad.

La propiedad de pérdida de memoria es una característica importante de la distribución exponencial, dado que esta propiedad es la responsable de mantener una *tasa de riesgo* $h(t)$ constante en este modelo.

Utilizando la ecuación (1.10) se verifica la propiedad de pérdida de memoria cuando $T \sim exp(\lambda)$

$$\begin{aligned} P(T > t + s \,|\, T > s) &= \frac{P(T > t + s, T > s)}{P(T > s)} = \frac{S(t+s)}{S(s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = S(t) \\ &= P(T > t) \end{aligned}$$

gráficamente este resultado se ilustra en la Figura 1.10:

En particular la distribución exponencial es la única distribución absolutamente continua que satisface la propiedad de pérdida de memoria.

Una tasa instantánea de falla constante significa que el riesgo al que están sometidos los individuos en estudio no evoluciona con el paso del tiempo, por lo tanto, esta característica restringe al modelo exponencial en algunas aplicaciones médicas e industriales en el análisis de supervivencia.

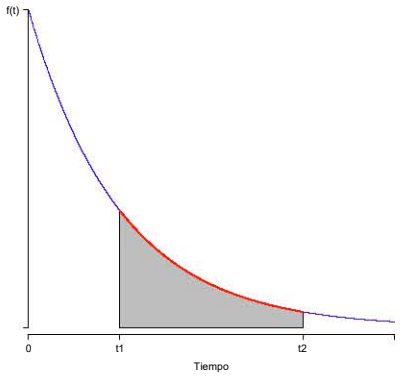


Figura 1.10: Pérdida de memoria del modelo exponencial.

1.4.2. Modelo Weibull

El modelo Weibull es una generalización del modelo exponencial y a menudo se utiliza para determinar el tiempo que transcurre hasta presentarse una falla. A diferencia del modelo exponencial, en el modelo Weibull se supone una razón de falla monótona, es decir, el riesgo de presentar una falla aumenta o disminuye con el paso del tiempo.

Algunas aplicaciones del modelo Weibull en el análisis de supervivencia son:

- Determinar el tiempo de vida de los componentes de un automóvil.
- Establecer la ocurrencia con la que se presenta un tumor cancerígeno en una población.
- Modelar el tiempo de corrosión de algún metal en particular.
- Estudios relacionados con la mortalidad infantil.
- Conocer el tiempo de degradación de algunos artículos.

La **función de riesgo** $h(t)$ del modelo Weibull proporciona el número de fallas observadas por unidad de tiempo. Esta función se describe en término de dos parámetros y se define de la siguiente manera

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \quad \text{con } \lambda,\, \gamma,\, t > 0 \tag{1.11}$$

En particular, el valor del parámetro γ nos indica la relación de monotonía entre el tiempo y la tasa de mortalidad, por lo que:

- $h(t)$ es monótona decreciente si $\gamma < 1$, por lo tanto, la probabilidad de fallo disminuye conforme avanza el tiempo.
- $h(t)$ es monótona creciente si $\gamma > 1$, por lo tanto, la probabilidad de que ocurra una falla aumenta con el paso del tiempo.

- Si $\gamma = 1$ el riesgo permanece constante, es decir, el modelo Weibull se reduce al modelo exponencial.

Por otra parte, el parámetro λ de la función de riesgo $h(t)$ define la razón de falla, generalmente su inverso λ^{-1} es conocido como el parámetro de escala y es aproximadamente el cuantil .632 de los datos, es decir, determina que el 63.2 % de los individuos son propensos a fallar en las primeras λ^{-1} unidades de tiempo. En ocasiones se utiliza el parámetro $\lambda^{-1} = \theta$ para representar la distribución y el riesgo definidos por esta variable.

A partir de la función de riesgo (1.11) se obtienen las funciones de supervivencia, densidad y riesgo acumulado.

Función de riesgo:

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \quad \text{con } \lambda,\, \gamma,\, t > 0$$

Función de riesgo acumulado:

$$\begin{aligned} H(t) &= \int_0^t \lambda\gamma(\lambda x)^{\gamma-1}\, dx \\ &= \frac{\lambda\gamma\lambda^{\gamma-1}x^{\gamma}}{\gamma}\Big|_0^t \\ &= (\lambda t)^{\gamma} \end{aligned}$$

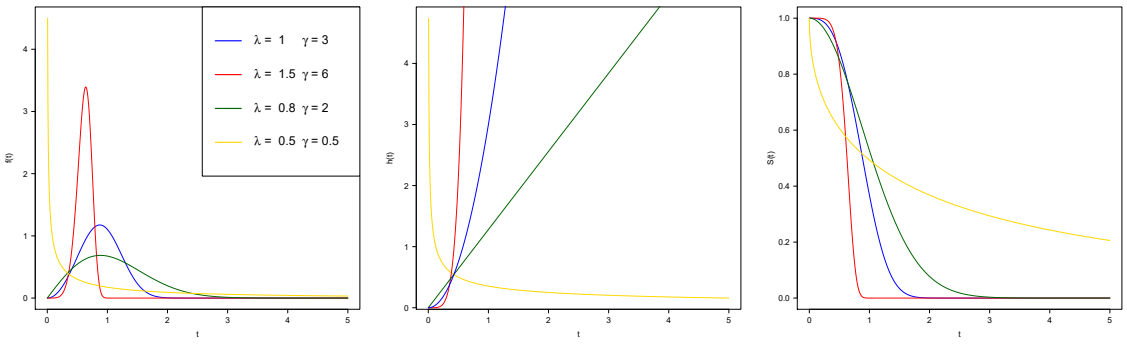
Función de supervivencia:

$$\begin{aligned} S(t) &= exp(-H(t)) \\ &= e^{-(\lambda t)^{\gamma}} \end{aligned}$$

Función de densidad:

$$\begin{aligned} f(t) &= h(t)\, S(t) \\ &= \lambda\gamma(\lambda t)^{\gamma-1}\, e^{-(\lambda t)^{\gamma}} \end{aligned}$$

La gráfica de la función de densidad del modelo Weibull es muy flexible, pues de acuerdo con el valor de sus parámetros, su imagen resulta similar a las de otras distribuciones, por lo que este modelo es muy útil al momento de ajustar diferentes tipos de datos.



(a) Función de Densidad (b) Función de Riesgo (c) Función de Supervivencia

Figura 1.11: Modelo Weibull

La Figura 1.11 muestra la función de densidad, de supervivencia y de riesgo, para diferentes valores de los parámetros λ y γ.

A continuación se resumen las principales características numéricas de los modelos que siguen una distribución Weibull.

Esperanza:

$$E(T) = \lambda^{-1}\,\Gamma\left(\frac{1}{\gamma}+1\right)$$

Varianza:

$$Var(T) = \lambda^{-2}\left[\Gamma\left(\frac{2}{\gamma}+1\right) - \Gamma\left(\frac{1}{\gamma}+1\right)^2\right]$$

el término $\Gamma(n)$ corresponde a la *función gamma* definida como sigue

$$\Gamma(n) = \int_0^\infty x^{n-1}\,e^{-x}\,dx$$

***p*-ésimo cuantil:**

$$t_p = \lambda^{-1}[-log(1-p)]^{1/\gamma}$$

como se había mencionado con anterioridad el inverso del parámetro λ refleja el cuantil .632 de la muestra, este resultado se demuestra haciendo $p = 1 - e^{-1}$ en la expresión anterior.

1.4.3. Modelo Log-normal

Generalmente el modelo log-normal es utilizado para modelar el comportamiento de observaciones que, en su mayoría, presentan la falla o evento de interés en intervalos cercanos al origen.

Ejemplos de aplicaciones del modelo log-normal en el análisis de supervivencia son:

- Estimar el tiempo de vida de aislantes eléctricos.
- Modelar el tiempo de duración de un procedimeinto quirúrgico.
- Modelar el tiempo que tarda en fallar una máquina después de haber recibido un mal mantenimiento.

Se dice que una variable aleatoria T tiene una distribución *log-normal* si la variable $Y = log(T)$ sigue una distribución normal. Por consiguiente la función de densidad de T se obtiene de la distribución normal mediante la trasformación de su ecuación.

Considerando que $Y = log(T) \sim N(\mu, \sigma^2)$, etonces la función densidad de Y es:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\Big[-\frac{1}{2}\Big(\frac{y-\mu}{\sigma} \Big)^2 \Big]$$

por lo tanto, se dice que T tiene una distribución *log-normal* con parámetros μ, σ^2 si su función de densidad de está dada por:

Función de densidad:

$$f(t) = \begin{cases} \frac{1}{t\,\sigma\sqrt{2\pi}}\, exp\Big[-\frac{1}{2}\Big(\frac{log(t)-\mu}{\sigma} \Big)^2 \Big] & \text{si } t > 0 \\ \\ 0 & \text{si } t \leq 0 \end{cases}$$

A partir del conocimiento de la función de densidad de la variable aleatoria T, es posible obtener; la función de supervivencia, la de riesgo y la de riesgo acumulado.

Función de supervivencia:

$$\begin{aligned} S(t) &= 1 - F(t) = 1 - \int_0^t f(x)\, dx \\ &= 1 - \int_0^t \frac{1}{x\,\sigma\sqrt{2\pi}}\, exp\Big[-\frac{1}{2}\Big(\frac{log(x)-\mu}{\sigma} \Big)^2 \Big]\, dx \\ &= 1 - \Phi\Big(\frac{log(t)-\mu}{\sigma} \Big) \end{aligned}$$

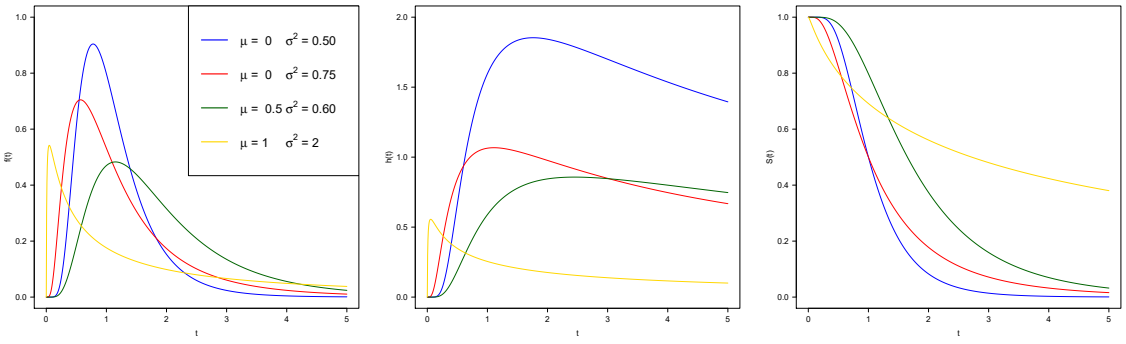
Función de riesgo:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\frac{1}{t\sigma\sqrt{2\pi}}\, exp\Big[-\frac{1}{2}\Big(\frac{log(t)-\mu}{\sigma} \Big)^2 \Big]}{1 - \Phi(\, \frac{log(t)-\mu}{\sigma}\,)} \end{aligned}$$

Función de riesgo acumulado:

$$\begin{aligned} H(t) &= -log(S(t)) \\ &= -log\left[(1-\Phi\Big(\frac{log(t)-\mu}{\sigma}\Big)\right] \end{aligned}$$

En la Figura 1.12 se observa el comportamiento de la función de densidad, de supervivencia y de riesgo de la distribución *log-normal* con distintos parámetros.



(a) Función de Densidad (b) Función de Riesgo (c) Función de Supervivencia

Figura 1.12: Modelo Log-normal

La esperanza, la varianza y el n-ésimo momento del modelo log-normal son:

Esperanza:

$$E(T) = exp\Big(\mu + \frac{\sigma^2}{2}\Big)$$

Varianza:

$$Var(T) = exp\Big(2\mu + 2\sigma^2\Big) - exp\Big(2\mu + \sigma^2\Big)$$

***n*-ésimo momento:**

$$E(X^n) = exp\Big(\mu n + \frac{1}{2}\sigma^2 n^2\Big)$$

***p*-ésimo cuantil:**

$$t_p = exp(\mu + \sigma z_p)$$

donde z_p es el cuantil de orden p de una variable aleatoria con distribución normal estándar.

1.4.4. Modelo Log-logístico

La distribución logística es una distribución de probabilidad continua. Esta distribución se caracteriza por ser simétrica y porque la gráfica de su función de densidad se parece a la de la distribución normal, con la diferencia de que la distribución lógistica tiene colas más pesadas. Las distribuciones logísticas a menudo se utilizan en modelos de crecimiento y su función de densidad está dada por:

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

Análogo a la distribución log-normal, la distribución *log-logistic* de una variable aleatoria T surge cuando $Y = log(T)$ tiene una distribución logística. Este modelo representa una buena alternativa para los modelos que siguen una distribución Weibull, dado que presentan una alta flexibilidad en sus funciones de riesgo y densidad.

El modelo *log-logístico* presenta las siguientes aplicaciones en el análisis de supervivencia.

- Modelar el tiempo de vida de un servicio.
- Determinar el tiempo de vida de un organismo.
- Establecer la supervivencia de individuos que padecen una enfermedad del corazón.
- Aplicaciones similares a las del modelo Weibull.

A continuación se presenta la función de supervivencia del modelo log-logístico, la cual permite obtener de una manera más sencilla el resto de las funciones de interés.

Sea T una distribución *log-logistic* con parámetros λ, $\kappa > 0$, se define la función de supervivencia de T como:

Función de supervivencia:

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

con $\lambda = exp(-\mu)$ y $\kappa = \frac{1}{\sigma}$, donde μ y σ son los parámetros correspondientes de la distribución logística.

Función de densidad:

$$\begin{aligned} f(t) &= -S'(t) \\ &= \frac{1}{[\,1 + (\lambda t)^\kappa\,]^2}\,\frac{d}{dt}\,(1 + (\lambda t)^\kappa) \\ &= \frac{\kappa\,\lambda\,(\lambda t)^{\kappa-1}}{[\,1 + (\lambda t)^\kappa\,]^2} \end{aligned}$$

Función de riesgo:

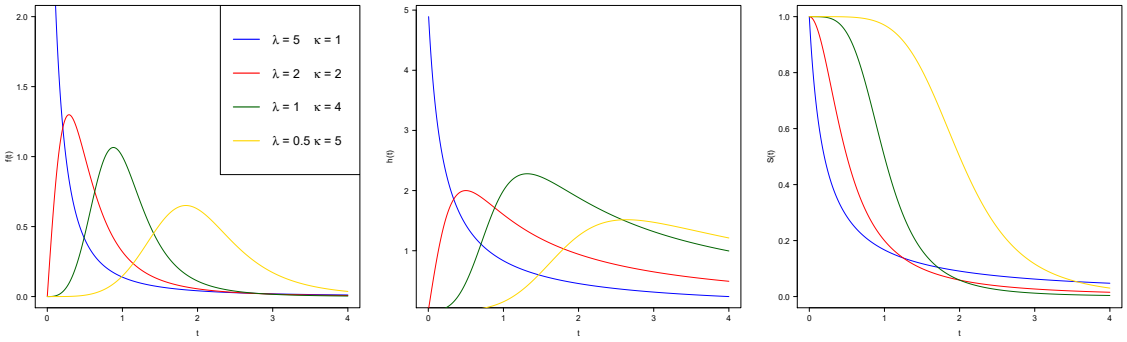
$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\kappa(\lambda t)^{\kappa-1}}{1+(\lambda t)^{\kappa}} \end{aligned}$$

La función de riesgo se caracteriza por se monónota decreciente cuando $\kappa \leq 1$, sin embargo, cuando $\kappa > 1$ la posibilidad de que ocurra una falla es creciente hasta que la función $h(t)$ alcanza su máximo, después de este punto el riesgo vuelve a ser decreciente.

Función de riesgo acumulado:

$$\begin{aligned} H(t) &= -log\left(\frac{1}{1+(\lambda t)^{\kappa}}\right) \\ &= log\left[1+(\lambda t)^{\kappa}\right] \end{aligned}$$

El comportamiento gráfico de las funciones de densidad, supervivencia y riesgo se observa en la siguiente figura.



(a) Función de Densidad (b) Función de Riesgo (c) Función de Supervivencia

Figura 1.13: Modelo Log-logístico

La esperanza, la varianza y el p-ésimo cuantil para el modelo log-logístico son:

Esperanza:

$$E(T) = \frac{\pi}{\kappa\lambda\left(\, sen\left(\frac{\pi}{\kappa}\right)\right)}$$

Varianza:

$$Var(t) = \frac{\pi\left(2\kappa\left(1-cos\left(\frac{\pi}{\kappa}\right)^2\right)+\pi sen\left(\frac{\pi(\kappa+2)}{\kappa}\right)\right)}{\left(sen\left(\frac{\pi(\kappa+2)}{\kappa}\right)\right)\left(\left(cos\left(\frac{\pi}{\kappa}\right)\right)^2-1\right)(\lambda\kappa^2)}$$

p-ésimo cuantil:

$$t_p = \frac{1}{\lambda}\Big(\frac{p}{1-p}\Big)^{1/\kappa}$$

1.4.5. Modelo Gamma

La distribución gamma incluye a las distribuciones exponencial y ji-cuadrada como casos especiales y tiene propiedades parecidas a las del modelo Weibull, sin embargo, el *modelo gamma* no es tan utilizado en el análisis de supervivencia, como los modelos anteriormente presentados, debido a que no tiene una expresión cerrada en sus funciones de riesgo y supervivencia. No obstante, la distribución gamma ajusta algunos modelos de manera adecuada.

Aplicaciones por las que la distribución gamma ha sido utilizada en el análisis de supervivencia son:

- Problemas relacionados con la confiabilidad industrial.
- Establecer el tiempo promedio de supervivencia de plaquetas.
- Modelar el desgaste de algunos materiales.

La variable aleatoria $T \sim gamma(\alpha\,,\,\beta)$ con $\alpha\,,\beta\,>0$ si su función de densidad es:

Función de densidad:

$$f(t) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-t\beta} & \text{si } t > 0 \\ \\ 0 & \text{si } t \leq 0 \end{cases}$$

De acuerdo con el valor de sus parámetros la función de densidad del modelo gamma se reduce a los siguientes casos:

- Si $\alpha\,,\beta\,=1$ entonces $T \sim exp(1)$
- Para $\alpha = 1$ la variable aleatoria T sigue una distribución Weibull con parámetros $\gamma = 1$ y $\lambda = \beta$
- Cuando $\alpha \to \infty$ entonces la distribución de T se aproxima a una distribución normal.
- Sea $\alpha \in \mathbb{Z}^+$, si $\nu = 2\alpha$ y $\beta^{-1} = 2$, entonces $T \sim \chi^2_\nu$

Función de supervivencia:

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

en particular a la función $I(\beta t, \alpha)$ se le conoce como la *función gamma incompleta*

Función de riesgo:

$$h(t) = \frac{\beta^{\alpha} t^{\alpha-1} e^{-t\beta}}{\Gamma(\alpha)[1 - I(\beta t, \alpha)]}$$

dada la relación que existe con la distribución Weibull, en el modelo gamma el valor del parámetro α nos indica la relación de monotonía entre el tiempo y la tasa de mortalidad, por lo que:

- $h(t)$ es una función monótona decreciente si $\alpha < 1$, y
- $h(t)$ es monótona creciente si $\alpha > 1$
- $h(t) = \beta$, es decir, una constante siempre y cuando $\alpha = 1$.

Función de riesgo acumulado:

$$H(t) = -log[1 - I(\beta t, \alpha)]$$

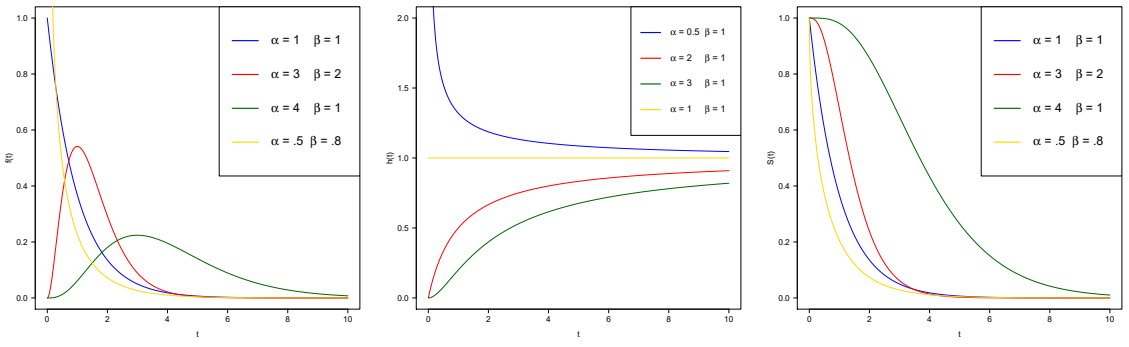
La esperanza y varianza del modelo gamma son:

Esperanza:

$$E(T) = \frac{\alpha}{\beta}$$

Varianza:

$$Var(T) = \frac{\alpha}{\beta^2}$$



(a) Función de Densidad (b) Función de Riesgo (c) Función de Supervivencia

Figura 1.14: Modelo Gamma

1.4.6. Modelo Gompertz

La distribución *Gompertz* se ha caracterizado por tener importantes aplicaciones en la matemática actuarial debido a la capacidad que tiene para modelar el tiempo de vida hasta la muerte de una persona en edad adulta.

Aunque originalmente el modelo desarrollado por el actuario Benjamin Gompertz fue desarrollado como un modelo demográfico, la distribución de Gompertz es utilizada en el análisis de supervivencia en los siguientes estudios.

- Modelar el tiempo de vida de personas adultas.
- Analizar el tiempo de vida de algunos procesos biológicos.
- Estimar las fallas que ocurren al ejecutar un código computacional.
- Establecer el tiempo de duración de una persona como cliente de algún determinado servicio.

Función de riesgo:

$$h(t) = \alpha e^{\beta t} \quad \alpha,\, \beta > 0$$

donde α representa la fuerza de mortalidad de una persona a edad 0 y β funciona como la tasa de envejecimiento.

Partiendo de las mismas relaciones utilizadas hasta ahora, se obtiene la función de densidad, supervivencia y de riesgo acumulado como:

Función de riesgo acumulado:

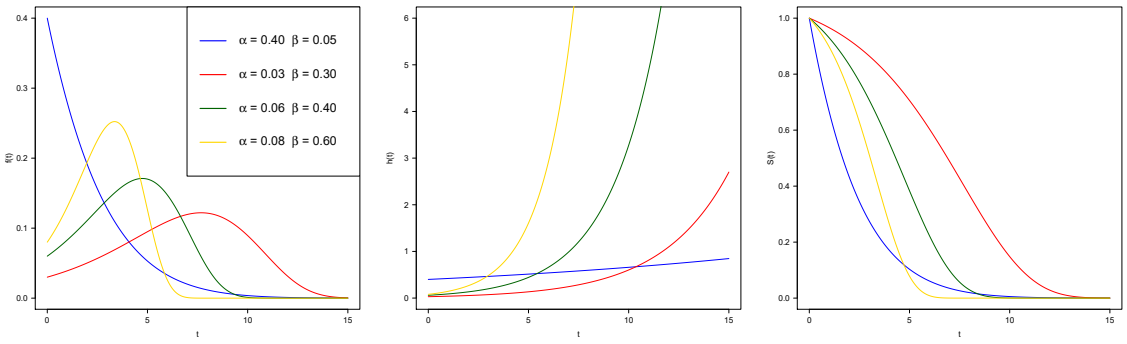
$$\begin{aligned} H(t) &= \int_0^t \alpha e^{\beta x}\, dx \\ &= \frac{\alpha e^{\beta x}}{\beta}\Big|_0^t \\ &= \frac{\alpha}{\beta}\,(\, e^{\beta t} - 1\,) \end{aligned}$$

Función de supervivencia:

$$\begin{aligned} S(t) &= exp(-H(t)) \\ &= exp\Big(-\frac{\alpha}{\beta}\,(\, e^{\beta t} - 1\,)\Big) \end{aligned}$$

Función de densidad:

$$f(t) = \alpha e^{\beta t}\, exp\Big(-\frac{\alpha}{\beta}\,(\, e^{\beta t} - 1\,)\Big)$$



(a) Función de Densidad (b) Función de Riesgo (c) Función de Supervivencia

Figura 1.15: Modelo Gompertz

Este modelo puede generalizarse incluyendo una constante c en la función de riesgo, de tal manera que:

$$h(t) = \alpha e^{\beta t} + c$$

La función anterior es conocida como la función de riesgo del modelo generalizado de Gompertz, el modelo *Gompertz-Makeham* y el resto de las funciones de interés se obtiene de manera análoga al modelo Gompertz.

$$\begin{aligned} H(t) &= \int_0^t (\, \alpha e^{\beta x} + c \,)\, dx \\ &= \frac{\alpha}{\beta} (\, e^{\beta t} - 1 \,) + ct \end{aligned}$$

en consecuencia

$$S(t) = exp\left(-\frac{\alpha}{\beta} (\, e^{\beta t} - 1 \,) - ct \right)$$

por lo que finalmente se obtiene

$$f(t) = (\alpha e^{\beta t} + c)\left(exp\left(-\frac{\alpha}{\beta} (\, e^{\beta t} - 1 \,) - ct \right)\right)$$

1.4.7. Tabla de Modelos Paramétricos

La Tabla 1.1 presenta un resumen de las funciones de riesgo, de supervivencia y de densidad, así como el valor de la esperanza de cada uno de los modelos paramétricos expuestos con anterioridad y algunos más. De acuerdo con la notación utilizada hasta el momento en el documento; $S(t)$ denota la función de supervivencia, $h(t)$ es la función de riesgo, $f(t)$ es la función de densidad y $E(T)$ es el valor esperado de la variable aleatoria T.

Distribución	$h(t)$	$S(t)$	$f(t)$	$E(T)$
Exponencial $\lambda > 0$ $t \geq 0$	λ	$exp(-\lambda t)$	$\lambda\, exp(-\lambda t)$	$1/\lambda$
Weibull $\lambda,\ \gamma > 0$ $t \geq 0$	$\lambda\gamma(\lambda t)^{\gamma-1}$	$e^{(-(\lambda t)^{\gamma})}$	$\lambda\gamma(\lambda t)^{\gamma-1}\, e^{-(\lambda t)^{\gamma}}$	$\lambda^{-1}\,\Gamma\Big(\frac{1}{\gamma}+1\Big)$
Log-normal $\sigma > 0$ $t > 0$	$f(t)/S(t)$	$1-\Phi\Big(\frac{log(t)-\mu}{\sigma}\Big)$	$\frac{exp\Big[-\frac{1}{2}\Big(\frac{log(t)-\mu}{\sigma}\Big)^2\Big]}{t\,\sigma\sqrt{2\pi}}$	$exp\Big(\mu+\frac{\sigma^2}{2}\Big)$
Log-logística $\lambda,\ \kappa > 0$ $t \geq 0$	$\frac{\kappa(\lambda t)^{\kappa-1}}{1+(\lambda t)^{\kappa}}$	$\frac{1}{1+(\lambda t)^{\kappa}}$	$\frac{\kappa\,\lambda\,(\lambda t)^{\kappa-1}}{[1+(\lambda t)^{\kappa}]^2}$	$\frac{\pi}{\kappa\lambda\Big(sen\Big(\frac{\pi}{\kappa}\Big)\Big)}$
Gamma $\alpha,\ \beta > 0$ $t > 0$	$f(t)/S(t)$	$1-I(\beta t,\alpha)$	$\frac{\beta^{\alpha}}{\Gamma(\alpha)}t^{\alpha-1}e^{-t\beta}$	α/β
Gompertz $\alpha,\ \beta > 0$ $t \geq 0$	$\alpha e^{\beta t}$	$exp(-\frac{\alpha}{\beta}\,(\,e^{\beta t}-1\,))$	$\alpha e^{\beta t}\, exp(-\frac{\alpha}{\beta}\,(\,e^{\beta t}-1\,))$	$\int_0^{\infty} t\, f(t)\, dt$
Normal $\sigma > 0$ $t \geq 0$	$f(t)/S(t)$	$1-\Phi\Big(\frac{t-\mu}{\sigma}\Big)$	$\frac{exp\Big[-\frac{1}{2}\Big(\frac{t-\mu}{\sigma}\Big)^2\Big]}{\sigma\sqrt{2\pi}}$	μ
Pareto $\theta,\ \lambda > 0$ $t \geq \lambda$	θ/t	$\Big(\frac{\lambda}{t}\Big)^{\theta}$	$\frac{\theta}{t}\Big(\frac{\lambda}{t}\Big)^{\theta}$	$\frac{\theta\lambda}{\theta-1}$ si $\theta > 1$

Tabla 1.1: Modelos paramétricos

Las funciones $I(\beta t,\, \alpha)$ y $\Gamma(\alpha)$ que aparecen en la tabla anterior se definen como:

$$I(\beta t,\, \alpha) = \int_0^{\beta t} \frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}\, du \qquad \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}\, dt$$

1.5. La Función de Verosimilitud

Con el propósito de hacer inferencias, en el análisis de supervivencia se busca obtener el valor de los parámetros que maximicen a la función de verosimilitud.

La función de verosimilitud para los datos de supervivencia requiere una revisión especial debido a que el conjunto de observaciones incompletas, que a menudo aparecen en este análisis, impiden trabajar con esta función de la misma forma que en la estadística inferencial cuando la información es completa.

La presencia de observaciones incompletas involucra el calcular una función de verosimilitud en la cual se incluyen a los datos censurados y/o truncados (*véase sección* 1.2). Una hipótesis importante que se hace en este cálculo es considerar como independientes a las variables T_i y C_i para cualesquiera individuos en el estudio.

En esta sección se muestra la forma de la función de verosimilitud para los distintos casos de censura por la derecha (tipo I, tipo II y tipo III), asimismo, se presentará la forma proporcional de la función de verosimilitud cuando un estudio comprende distintos tipos de observaciones censuradas y truncadas.

1.5.1. Verosimilitud: cuando se presenta censura tipo I

Considere que se tiene una muestra de n individuos con tiempos de vida T_i, $i = 1, 2, \ldots, n$ representados por la pareja (t_i, δ_i) donde, de acuerdo con la *sección* 1.2.1, $t_i = min(T_i, C_{r_i})$ y

$$\delta_i = \begin{cases} 1 & \text{si } \ t_i \leq T_i \\ \\ 0 & \text{si } \ T_i > C_{r_i} \end{cases}$$

Si $\delta = 0$, es decir, cuando la observación está censurada se tiene que[5];

$$\begin{aligned} P(\, t_i = t \,,\, \delta_i = 0 \,) &= P(\, t_i = t \,|\, \delta_i = 0) \, P(\delta_i = 0) \\ \\ &= P(\delta_i = 0) = P(T_i > t) \\ \\ &= S(t) \end{aligned}$$

[5]Obsérvese que por la forma en la que se define a t_i, dicha t_i representa una variable aleatoria y no una observación del experimento.

Si $\delta = 1$, entonces;

$$\begin{aligned} P(\, t_i = t \,,\, \delta_i = 1 \,) &= P(\, t_i = t \,|\, \delta_i = 1) \, P(\delta_i = 1) \\ &= P(\, T_i = t \,|\, T_i \leq C_{r_i}) \, P(T_i \leq C_{r_i}) \\ &= \frac{f(t)}{1 - S(C_r)}(1 - S(C_r)) \\ &= f(t) \end{aligned}$$

Las expresiones para $\delta = 0$ y $\delta = 1$ pueden ser representadas como una única expresión mediante:

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

por lo que la función de verosimilitud L para los datos con censura tipo I tiene la forma:

$$L = \prod_{i=1}^{n} (f(t_i))^{\delta_i}(S(t_i))^{1-\delta_i} \tag{1.12}$$

$$= \prod_{i=1}^{n} (f(t_i))^{\delta_i} \frac{S(t_i)}{(S(t_i))^{\delta_i}}$$

$$= \prod_{i=1}^{n} \left(\frac{f(t_i)}{S(t_i)}\right)^{\delta_i} S(t_i)$$

$$= \prod_{i=1}^{n} (h(t_i))^{\delta_i} S(t_i) \tag{1.13}$$

1.5.2. Verosimilitud: cuando se presenta la cesura tipo II

En particular este tipo de censura se caracteriza por tener r sujetos, de n individuos en estudio, los cuales presentan la falla del siguiente modo

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

Análogo a la verosimilitud cuando se presenta la censura tipo I y utilizando las propiedades de las estadísticas de orden, se tiene que la función de distribución conjunta de $T_{(1)}$, $T_{(2)}$, $\ldots$, $T_{(r)}$ está dada por:

$$\frac{n!}{(n-r)!} \prod_{i=1}^{r} f(t_{(i)}) \, (S(t_{(r+)}))^{n-r}$$

donde $S(t_{(r+)})$ representa la supervivencia de los $(n - r)$ individuos que no presentaron la falla en el estudio, utilizando la función δ_i para la censura tipo II

$$\delta_i = \begin{cases} 1 & \text{si} \;\; T_i \leq T_{(r)} \\ \\ 0 & \text{si} \;\; T_i > T_{(r)} \end{cases}$$

se construye la función de verosimilitud para este tipo de censura como:

$$\frac{n!}{(n-r)!}\prod_{i=1}^{r}(f(t_{(i)}))^{\delta_i}\,(S(t_{(r)}))^{1-\delta_i} \tag{1.14}$$

obsérvese que el término $n!/(n-r)!$ no involucra ningún parámetro de interés de $f(t)$, por lo que la función de verosimilitud (1.14) es proporcional a:

$$L=\prod_{i=1}^{r}(f(t_{(i)}))^{\delta_i}\,(S(t_{(r)}))^{1-\delta_i}$$

asimismo, note como la función de verosimilitud para la censura tipo II coincide con la función de verosimilitud de la censura tipo I que aparece en (1.12), de este modo la verosimilitud para la censura tipo II coincide con la expresión (1.13)

$$L=\prod_{i=1}^{n}(h(t_i))^{\delta_i}S(t_i)$$

1.5.3. Verosimilitud: cuando se presenta la censura tipo III

En este caso se supone que cada individuo tiene un tiempo de falla T_i y tiempo de censura C_i, ambos tiempos representan variables aleatorias independientes con funciones de supervivencia $S(t)$ y $G(t)$ respectivamente y funciones de densidad $f(t)$ y $g(t)$ correspondientes.

Sea $G(t)$ una función que no depende de ninguno de los parámetros de $S(t)$ y sean $(\,t_i\,,\,\delta_i\,), i=1,\,2\ldots,\,n$, las parejas de observaciones independientes de los n individuos, con $t_i=min(T_i,C_i)$ y

$$\delta_i=\begin{cases}1 & \text{si}\;\; t_i\leq T_i\\ \\ 0 & \text{si}\;\; T_i>C_i\end{cases}$$

entonces, la función de distribución conjunta se construye como;

Si $\delta_i=0$

$$\begin{aligned}P(\,t_i=t\,,\,\delta_i=0\,) &= P(\,C_i=t\,,\,T_i>C_i\,)\\ \\ &= P(\,C_i=t\,)P(\,T_i>C_i\,)\\ \\ &= g(t)S(t)\end{aligned}$$

Si $\delta_i=1$

$$\begin{aligned}P(\,t_i=t\,,\,\delta_i=1\,) &= P(\,T_i=t\,,\,T_i\leq C_i\,)\\ \\ &= P(\,T_i=t\,)P(\,T_i\leq C_i\,)\\ \\ &= f(t)G(t)\end{aligned}$$

La expresiones para $\delta_i = 0$ y $\delta_i = 1$ se pueden representar a través de una sola expresión de la siguiente manera

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

entonces la función de verosimilitud L de la muestra con censura tipo III es:

$$\prod_{i=1}^{n}[\,g(t_i)\,]^{\delta_i}[\,G(t_i)\,]^{1-\delta_i}[\,f(t_i)\,]^{\delta_i}[\,S(t_i)\,]^{1-\delta_i} \tag{1.15}$$

Las funciones $G(t)$ y $g(t)$ no involucran a los parámetros de interés de $f(t)$, entonces la función de verosimilitud (1.15) es proporcional a:

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

Obsérvese como la expresión obtenida es la misma que la que resultó en la censura tipo I y tipo II, por lo que la función de verosimilitud para la censura tipo III puede reducirse a:

$$L = \prod_{i=1}^{n}(h(t_i))^{\delta_i}S(t_i)$$

Evidentemende la función de verosimilitud resulta ser la misma para cualquiera de los tipos de censura expuestos con anterioridad, por lo tanto la metodología con la que se desarrollan los cálculos es la misma para los tres casos.

1.5.4. Verosimilitud

Como se ha mostrado con anterioridad, la función de verosimilitud se construye a partir de las observaciones obtenidas de t_i y el modelo de distribución $f(t_i)$ asociado, por lo que en el análisis de supervivencia solo se requiere del conocimiento de alguna de las funciones de supervivencia ($S(t)$, $h(t)$, $H(t)$ o $f(t)$), para poder obtener el valor del estimador máximo verosímil.

El siguiente esquema proporciona las probabilidades de los distintos tipos de observaciones (censuradas o truncadas), con el propósito de obtener la función de verosimilitud general cuando se involucran más de un tipo de observaciones incompletas en el estudio.

Observaciones completas	$f(t_i)$
Censura por la derecha	$S(C_{r_i})$
Censura por la izquierda	$S(C_{l_i})$
Censura por intervalo	$S(L_i) - S(R_i)$
Truncamiento por la izquierda	$f(t_i)/S(L_i)$
Truncamiento por la derecha	$f(t_i)/[1 - S(R_i)]$

Bajo el supuesto de que cada una de las observaciones anteriores son independientes, la función de verosimilitud de la manera siguiente:

$$L \propto \prod_{i \in D} f(t_i) \prod_{i \in R} S(C_{r_i}) \prod_{i \in L} S(C_{l_i}) \prod_{i \in I} [\, S(L_i) - S(R_i) \,]$$

Donde D es el conjunto de observaciones completas, R la proporción de observaciones censuradas por la derecha, L señalan el grupo de observaciones censuradas por la izquierda y por último I asume el control de las observaciones de las censuras por intervalo.

Para observaciones truncadas por la izquierda, las observaciones exactas $f(t_i)$ se remplazan por $f(t_i)/S(L_i)$, mientras que las observaciones censuradas $S(C_{r_i})$ se remplazan por $S(C_{r_i})/S(L_i)$, de manera que L_i representa la i-ésima observación truncada por la izquierda.

En el caso del truncamiento por la derecha solo se observan fallas, es decir, que cuando existen observaciones truncadas por la derecha no se cuenta con registros censurados, por lo tanto la función de verosimilitud es de la forma:

$$L \propto \prod_{i} \frac{f(R_i)}{1 - S(R_i)}$$

En el caso de que cada individuo tenga una distribución de supervivencia diferente, como podría ser el caso cuando se utilizan técnicas de regresión, la función de verosimilitud es de la forma:

$$L \propto \prod_{i \in D} f_i(t_i) \prod_{i \in R} S_i(C_{r_i}) \prod_{i \in L} S_i(C_{l_i}) \prod_{i \in I} [\, S_i(L_i) - S_i(R_i) \,]$$

Ejemplo 1.5.1.

Supóngase que se tiene una muestra aleatoria de observaciones[6] de T

$$0.2,\ 0.4,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.7,\ 0.8,\ 0.9+,\ 0.9+$$

Sea T una variable aleatoria con función de densidad

$$f_T(t) = \frac{(\,0.1\,)^\theta\, \theta}{t^{\theta+1}} \quad ; \quad \theta > 0.1$$

entonces el valor del estimador máximo verosímil de θ para dicha muestra se obtiene de:

$$L = \prod_{i=1}^{n} (h(t_i))^{\delta_i}\, S(t_i)$$

en este caso $n = 10$, además $S(t)$ y $h(t)$ pueden obtenerse a partir del conocimiento de $f(t)$.

[6]El símbolo "+", significa que la observación está censurada, asimismo, puede utilizarse la notación (0.9,0) y (0.4,1) para indicar que la observación se encuentra censurada y no censurada respectivamente.

$$S(t) = 1 - F(t) = 1 - \int_{0.1}^{t} \frac{(0.1)^{\theta}\,\theta}{x^{\theta+1}}\,dx \;= 1 - \left(-\frac{(\,0.1\,)^{\theta}}{x^{\theta}} \Big|_{0.1}^{t} \right) \;=\; \frac{(\,0.1\,)^{\theta}}{t^{\theta}}$$

en consecuencia:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta}{t}$$

La función de verosimilitud y su proporcional función *log-verosimilitud* son:

$$\begin{aligned} L &= \prod_{i=1}^{n} (h(t_i))^{\delta_i}\, S(t_i) \\ &\propto log\left(\prod_{i=1}^{n} (h(t_i))^{\delta_i}\, S(t_i) \right) \\ &= \sum_{i=1}^{n} log\left[(h(t_i))^{\delta_i}\, S(t_i) \right] \\ &= \sum_{i=1}^{n} \Big(\delta_i\, log\left(h(t_i)\right) + log\left(S(t_i)\right) \Big) \end{aligned}$$

sustituyendo los valores correspondientes de $h(t)$ y $S(t)$ se tiene que:

$$\begin{aligned} log\,(L) &= \sum_{i=1}^{n} \left(\delta_i\, log\left(\frac{\theta}{t_i}\right) + log\left(\left(\frac{0.1}{t_i} \right)^{\theta} \right) \right) \\ &= \sum_{i=1}^{n} \Big(\delta_i\, log\,(\theta) - \delta_i\, log(t_i) + \theta\, log\,(0.1) - \theta\, log\,(t_i) \Big) \end{aligned}$$

derivando la expresión anterior con respecto al parámetro θ se obtiene:

$$\frac{\partial}{\partial\theta}\, log\,(L) = \sum_{i=1}^{n} \Big(\frac{\delta_i}{\theta} + log\,(0.1) - log\,(t_i) \Big)$$

se iguala a cero la expresión anterior y se despeja θ, para obtener el estimador máximo verosímil $\hat{\theta}$ del parámetro θ.

$$\begin{aligned} 0 = \;& \sum_{i=1}^{n} \Big(\frac{\delta_i}{\theta} + log\,(0.1) - log\,(t_i) \Big) = \sum_{i=1}^{n} \frac{\delta_i}{\theta} \; + n\, log(0.1) - \sum_{i=1}^{n} log\,(t_i) \\ \Rightarrow \quad & \frac{\sum_{i=1}^{n} \delta_i}{\theta} = \sum_{i=1}^{n} log\,(t_i) - n\, log\,(0.1) \end{aligned}$$

dando como resultado:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} log\,(t_i) - n\, log\,(0.1)}$$

Una vez que se ha obtenido el estimador del parámetro por máxima verosimilitud, se obtiene el resultado para la muestra proporcionada en el ejemplo haciendo una simple sustitución de datos en el resultado anterior.

$$\widehat{\theta} = \frac{8}{\sum_{i=1}^{10} log\,(t_i) - 10\,log(0.1)}$$

Nota: En particular $\sum_{i=1}^{10} \delta_i = 8$ *debido a que existen 8 observaciones completas.*

Finalmente, realizando las sumas correspondientes en el denominador se obtiene que el estimador máximo verosímil para θ, con los datos de esta muestra, es:

$$\hat{\theta} = 0.4642$$

Gráficamente se puede observar en la Figura 1.16, que las funciones de verosimilitud y log-verosimilitud, alcanzan su máximo en el valor estimado $\hat{\theta} = 0.4642$.

El metódo de máxima verosimilitud puede ser un procedimiento muy complicado al momento de obtener el estimador máximo verosímil, por lo que en la práctica es conveniente utilizar algoritmos numéricos para encontrar los valores de los parámetros que maximizan a la función de verosimilitud, algunos de los algoritmos más utilizados son: Newton-Raphson y EM (Expectation Maximization).

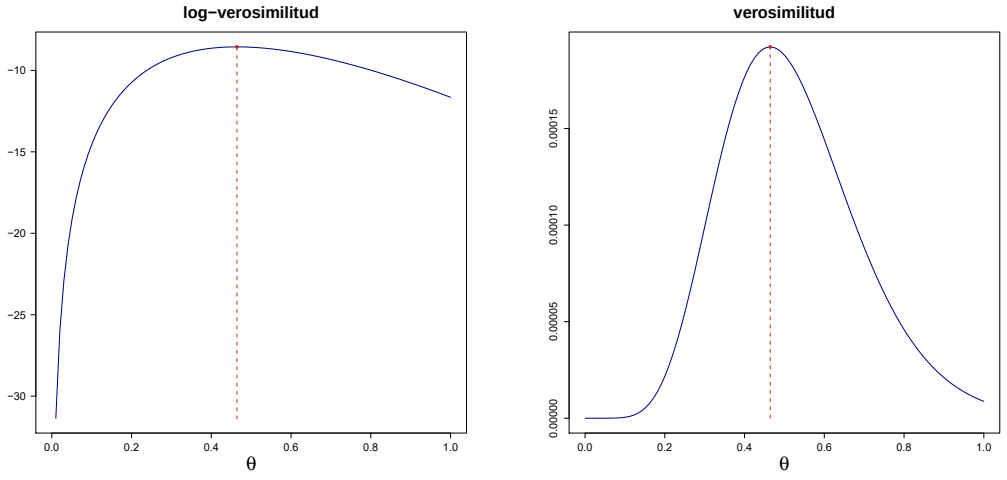


Figura 1.16:

1.6. Estimación No-paramétrica de la Función de Supervivencia

En esta sección se introduce el método de estimación límite-producto de la función de supervivencia desarrollado por Edward L.Kaplan y Paul Meier en 1958. El estimador límite-producto es un estimador no paramétrico de la función de supervivencia, es decir, no se supone el conocimiento de alguna distribución teórica, sin embargo, los estimadores que se obtienen por este método (particularmente la presentación gráfica de la curva de supervivencia estimada) pueden ser útiles si se pretende elegir un modelo de distribución conocido.

Para una muestra aleatoria de tamaño n sin censura, es decir, los tiempos de supervivencia de los individuos de la muestra son exactos y conocidos, la función de supervivencia puede ser estimada a partir de la *función de supervivencia empírica*:

$$\widehat{S}(t) = \frac{\text{número de individuos que viven más allá de } t}{\text{número total de individuos}}$$

Sean t_1, $t_2, \ldots,$ t_n los tiempos exactos de supervivenacia y $t_{(i)}$ el i-ésimo tiempo de supervivencia ordenado de tal forma que $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$, entonces la función de supervivencia en el tiempo $t_{(i)}$ es estimada como:

$$\widehat{S}(t_{(i)}) = \frac{n-i}{n}$$

donde el término $(n-i)$ corresponde al número total de individuos en la muestra que sobreviven más allá de $t_{(i)}$ y n = tamaño de la muestra.

Evidentemente $\widehat{S}(t_{(0)}) = 1$ pues al principio del estudio todos los individuos de la muestra se encuentran con vida. Análogamente $\widehat{S}(t_{(n)}) = 0$ debido a que se considera que ningún individuo está con vida más allá del tiempo $t_{(n)}$.

Si ocurren dos o más fallas al mismo tiempo de tal forma que $t_{(i)} = t_{(j)} = t_{(k)}$ con $i \neq j \neq k$, se usa el valor más grande de estos índices para estimar a la función de supervivencia:

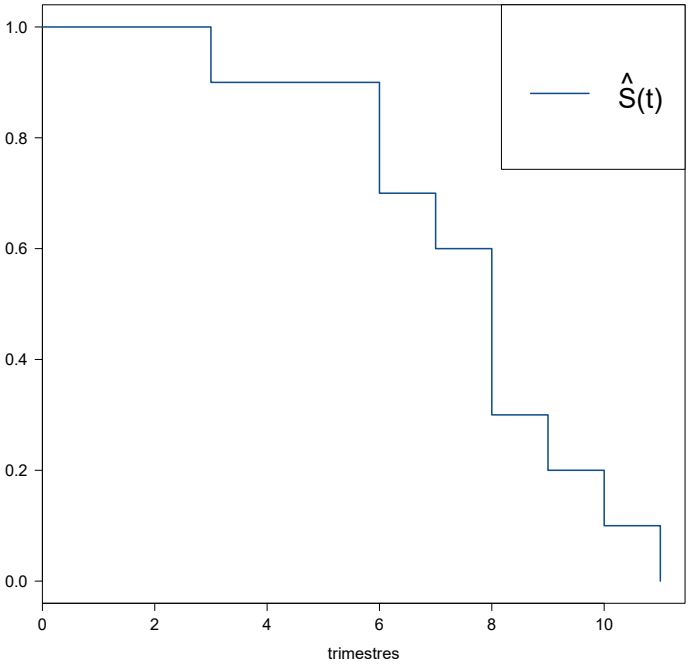
$$\widehat{S}(t_{(i)}) = \widehat{S}(t_{(j)}) = \widehat{S}(t_{(k)}) = \frac{n - max\,(i,j,k)}{n}$$

El comportamiento gráfico de la función $\widehat{S}(t)$ (*también conocida como función de supervivencia empírica*) corresponde al de una función escalonada que empieza en uno y decrece hacia cero, asimismo, la gráfica de esta función permanece constante en los intervalos de tiempo $[\,i, i+1)$, $i = 0,\ 1,\ 2, \ldots,\ n-1$.

Ejemplo 1.6.1. La siguiente tabla muestra los tiempos de supervivencia[7] en trimestres de un grupo de pacientes en edad adulta con distrofia muscular Duchenne.

$t_{(i)}$	i	$\widehat{S}(t_{(i)})$
3	1	$\frac{10-1}{10} = 0.9$
6	2	$\frac{10-3}{10} = 0.7$
6	3	$\frac{10-3}{10} = 0.7$
7	4	$\frac{10-4}{10} = 0.6$
8	5	$\frac{10-7}{10} = 0.3$
8	6	$\frac{10-7}{10} = 0.3$
8	7	$\frac{10-7}{10} = 0.3$
9	8	$\frac{10-8}{10} = 0.2$
10	9	$\frac{10-9}{10} = 0.1$
11	10	$\frac{10-10}{10} = 0$

(a) Tabla A



(b) Figura A

En la Tabla A, además de los tiempos de supervivencia, puede observarse el cálculo de la función de supervivencia estimada $\widehat{S}(t_{(i)})$, mientras que la Figura A representa el comportamiento de la función de supervivencia estimada $\widehat{S}(t)$.

Cuando en una muestra hay presencia de observaciones censuradas, la función de supervivencia es estimada a partir del estimador límite-producto desarrollado por Kaplan y Meier, mejor conocido como **estimador de Kaplan-Meier**.

Definición 1.6.1 (Estimador de Kaplan-Meier). Sean $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$, los tiempos de supervivencia exactos y ordenados, se define el estimador de Kaplan-Meier como:

$$\widehat{S}(t) = \prod_{t_{(i)}<t} \left(\frac{n_i - d_i}{n_i} \right)$$

d_i denota el número de individuos que fallan en $t_{(i)}$ y
n_i es el número de individuos vivos o sin presentar la falla justo antes de $t_{(i)}$.

[7] Aunque el tiempo es una variable aleatoria continua, se pueden permitir observaciones repetidas, pues el tiempo de supervivencia generalmente es medido en una escala discreta.

El término n_i también es conocido como el conjunto en riesgo en el tiempo $t_{(i)}$. Usualmente el estimador de Kaplan-Meier se escribe como:

$$\widehat{S}(t) = \prod_{j=1}^{k} \widehat{p_j} = \left(\frac{n_j - d_j}{n_j}\right) \quad t \in [\, t_k, t_{k+1}\,)$$

donde $\widehat{p_j}$ es la probabilidad estimada de que un individuo sobreviva a lo largo del intervalo que empieza en $t_{(j)}$. Calculando el logaritmo de $\widehat{S}(t)$ se tiene que:

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

Supóngase que el número de individuos que sobreviven a lo largo del intervalo que comienza en $t_{(j)}$ tiene una distribución Binomial con parámetros n_j y p_j, es decir, $(n_j - p_j) \sim Binomial(n, p)$, donde p_j es la verdadera probabilidad de supervivencia a lo largo del intervalo, entonces, haciendo uso del método Delta

$$Var\,(\,g(X)\,) \approx \left(\frac{d\,g(X)}{d\,X}\right)^2 Var(X)$$

se obtiene que, la varianza de $log(\widehat{S}(t))$ está dada por:

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

de modo que despejando la $Var\,(\,\widehat{S}(t)\,)$ de la expresión anterior se tiene que:

$$Var\,(\,\widehat{S}(t)\,) \approx \left(\widehat{S}(t)\right)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

entonces, el error estándar del estimador de Kaplan-Meier es:

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

Esta última expresión es conocida como la *fórmula de Greenwood* y se utiliza principalmente para obtener intervalos de confianza para $S(t)$.

Para un valor dado τ, un intervalo del $(1-\alpha)\times 100\,\%$ para $S(t)$ es:

$$\widehat{S}(\tau) \pm z_{1-\alpha/2}\, s.e\left(\widehat{S}(\tau)\right)$$

donde $z_{1-\alpha/2}$ es el cuantil $(1-\alpha/2)$ de una distribución normal estándar.

Algunos de los resultados obtenidos en esta sección, así como la construcción del estimador de Kaplan-Meier y su correspondiente varianza e intervalos de confianza, pueden consultarse en [1].

Ejemplo 1.6.2. Los siguientes datos corresponden al tiempo de supervivencia en meses de un grupo de 12 camiones que son seguidos hasta presentar un siniestro.

$$6 \quad 7+ \quad 9 \quad 10+ \quad 12 \quad 12$$

$$14 \quad 15 \quad 15+ \quad 16 \quad 18+ \quad 20$$

En la Tabla 1.2, se pueden encontrar los cálculos correpondientes a la función de supervivencia estimada, los errores estándar y el intervalo de confianza para los distintos valores de t.

t	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\widehat{S}(t)$	$s.e\left(\widehat{S}(t)\right)$	95 % IC
0	12	0	1	1	0	-
6	12	1	0.9166	0.9166	0.0798	(0.7729,1.000)
9	10	1	0.9000	0.8249	0.1128	(0.6311,1.000)
12	8	2	0.7500	0.6186	0.1520	(0.3823,1.000)
14	6	1	0.8333	0.5155	0.1578	(0.2830,0.939)
15	5	1	0.8000	0.4124	0.1564	(0.1962,0.867)
16	3	1	0.6666	0.2749	0.1532	(0.0923,0.819)
20	1	1	0	0	-	-

Tabla 1.2:

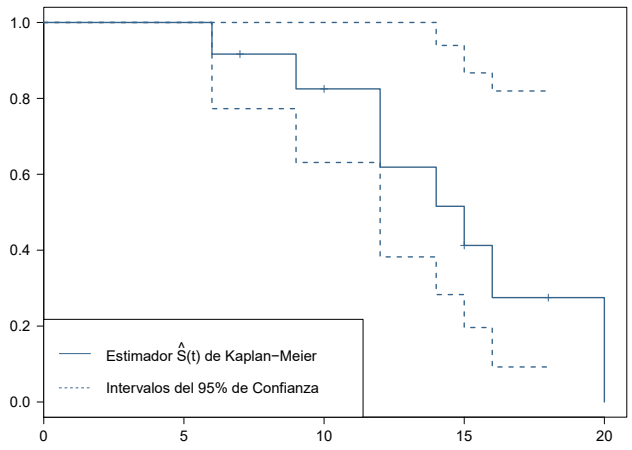


Figura 1.17: Gráfica de la función de supervivencia estimada.

El estimador de Kaplan-Meier tiene diversas aplicaciones, principalmente en las investigaciones médicas, por ejemplo, se utiliza para estimar la proporción de individuos que se mantienen con vida después de haber recibido un tratamiento, o bien, para determinar, a partir de dos o más curvas de supervivencia estimadas, que tratamiento es más efectivo cuando dos o varios grupos de pacientes son sometidos a tratamientos distintos. Los métodos más comunes para comparar distintas curvas de Kaplan-Meier son: la prueba *log-rank* y el modelo de regresión de Cox.

Actualmente, existe una variedad de paquetes estadísticos que implementan funciones relacionadas al cálculo del estimador de Kaplan-Meier y que permiten repesentar a este estimador gráficamente. En el Apéndice A de este documento, se incluyen los códigos en R que se ejecutaron para conseguir los resultados gráficos obtenidos en esta sección, sin embargo, a continuación se incluye un breve tutorial en el que se explica como obtener algunos de los resultados relacionados con el estimador de Kaplan-Meier utilizando las funciones disponibles en el software estadístico R.

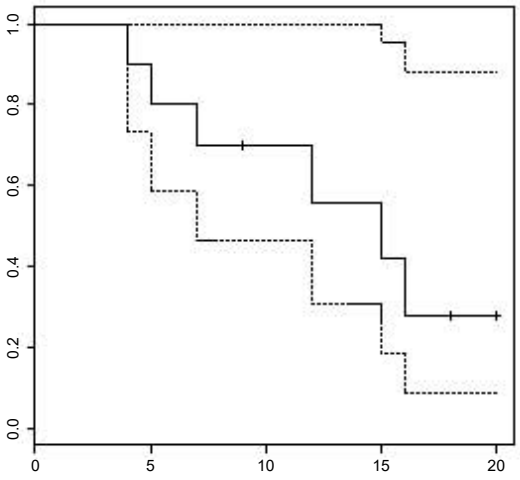
Considere que se tiene los siguientes datos, los cuales representan los tiempos de supervivencia de un conjunto de individuos.

$$7 \quad 4 \quad 9+ \quad 12 \quad 9+ \quad 15 \quad 18+ \quad 5 \quad 20+ \quad 16$$

el signo + significa que la observación $t+$ está censurada por la derecha en el tiempo señalado. La manera más sencilla de indicar a R cuáles son los tiempos que están censurados , es mediante la creación de un vector de ceros y unos, donde cero indica que la observación está censurada y uno indica que la observación es completa.

En el siguiente programa 1.- se crea un objeto de supervivencia para muestras que, precisamente, contienen individuos censurados por la derecha, 2.- se obtiene la estimación de la función de supervivencia por el método de Kaplan-Meier y 3.- se grafica la función de supervivencia estimada.

```
# Para llevar a cabo un análisis de supervivencia para un conjunto, es
# necesario cargar la biblioteca survival con la siguiente instrucción:
library(survival)

# Los tiempos de supervivencia utilizados en este ejemplo son:
tiempos <- c(7,4,9,12,9,15,18,5,20,16)

# El indicador de censura (respetando el orden del vector de tiempos) es:
censura <- c(1,1,0,1,0,1,0,1,1,0)

# En la siguiente instrucción se crea el objeto de supervivencia deseado:
datos <- Surv(tiempos,censura)
datos

# Con el comando survfit se proporciona el estimador de la función de KM
# Con summary(·) se genera la tabla que se exhibe en la Figura 1.18
km <- survfit(datos~1)
summary(km)

# Finalmente con el comando plot(·) se genera la gráfica de la función Ŝ(t)
plot(km)
```



(a) Resultado del comando **plot(km)**.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
4	10	1	0.90	0.0949	0.7320	1.000
5	9	1	0.80	0.1265	0.5868	1.000
7	8	1	0.70	0.1449	0.4665	1.000
12	5	1	0.56	0.1706	0.3082	1.000
15	4	1	0.42	0.1763	0.1845	0.956
16	3	1	0.28	0.1640	0.0889	0.882

(b) Resultado del comando **summary(km)**.

Figura 1.18: Gráfica y tabla proporcionadas por comandos de R

La columna time de la tabla que aparece en la Figura 1.17 contiene las observaciones completas (no censuradas) que aparecen en la muestra, la columna n.risk es el conjunto en riesgo, n.event es una columna en la que se muestra la cantidad de individuos que fallan en el tiempo indicado, survival es la función de supervivencia estimada en el tiempo t (para este ejemplo, en $t = 4, 5, 7, ..., 16$) de manera que std.err es el error estándar de la correspondiente función de supervivencia estimada, por último, las columnas lower 95 % CI y upper 95 % CI representan el límite inferior y el límite superior de un intervalo al 95 % de confianza para $S(t)$.

Con el objetivo de crear objetos de supervivencia más completos, la estructuras de los comandos ***Surv*** y ***survfit*** se pueden escribir de una forma más amplia. Estas estructuras pueden consultarse en R con la ejecución de los siguientes comandos:

- **help(Surv)**
- **help(survfit)**

Para finalizar esta sección se presenta un ejemplo en el cual se comparan gráficamente dos curvas de supervivencia estimadas por el método de Kaplan-Meier, el código en R se encuentra en el Apéndice A.

Ejemplo 1.6.3. Considere los siguientes tiempos de supervivencia correspondientes a dos grupos de individuos (grupo I y grupo II) los cuales han recibido un Tratamiento A y un Tratamiento B respectivamente.

Tratamiento A	1+	4	5	5	9+	10	13	16	16+	18
	21	22+	23	24+	27	30+	33	35	35+	40+
Tratamiento B	1	1+	2+	3+	4	4	6+	8	10+	12
	14	14+	17	20	21+	23	23+	26	33	40+

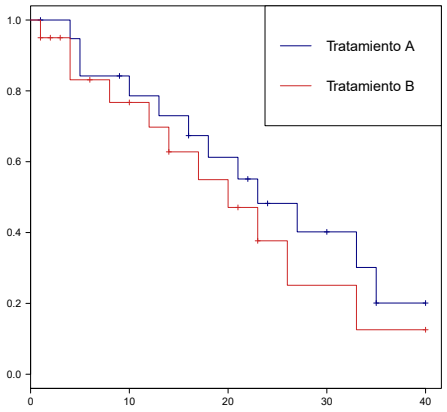


Figura 1.19: Gráficamente se puede decir que el tratamiento A es más efectivo que el tratamiento B, sin embargo, es recomendable realizar otro tipo de comparación.

Tratamiento A

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
4	19	1	0.947	0.0512	0.8521	1.000
5	18	2	0.842	0.0837	0.6931	1.000
10	15	1	0.786	0.0951	0.6201	0.996
13	14	1	0.730	0.1035	0.5527	0.964
16	13	1	0.674	0.1097	0.4896	0.927
18	11	1	0.612	0.1156	0.4231	0.887
21	10	1	0.551	0.1192	0.3608	0.842
23	8	1	0.482	0.1226	0.2931	0.794
27	6	1	0.402	0.1258	0.2177	0.742
33	4	1	0.301	0.1283	0.1309	0.694
35	3	1	0.201	0.1185	0.0632	0.639

Tratamiento B

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	20	1	0.950	0.0487	0.8591	1.000
4	16	2	0.831	0.0894	0.6733	1.000
8	13	1	0.767	0.1029	0.5900	0.998
12	11	1	0.698	0.1147	0.5053	0.963
14	10	1	0.628	0.1227	0.4281	0.921
17	8	1	0.549	0.1300	0.3454	0.874
20	7	1	0.471	0.1330	0.2706	0.819
23	5	1	0.377	0.1357	0.1859	0.763
26	3	1	0.251	0.1367	0.0864	0.730
33	2	1	0.126	0.1121	0.0218	0.722

Capítulo 2

Modelos Paramétricos de Regresión

En el presente capítulo se estudia un enfoque paramétrico de los modelos de regresión, sin embargo, se ha decidido incluir, debido a la relevancia que tiene en el análisis de supervivencia, el *Modelo de Riesgos Proporcionales de Cox*, a pesar de que éste se define como un modelo semiparamétrico en los modelos de riesgos proporcionales.

2.1. Modelos de Regresión en Supervivencia

En muchos ensayos, la falla que se determina para los sujetos en estudio no sólo depende del tiempo que transcurre en el experimento, sino también de un conjunto de características que determinan al individuo que pertenece a la muestra y lo distingue de los demás. De este modo el tiempo de supervivencia puede describirse a partir de una relación funcional, en la cual se involucran todas aquellas características que se cree que influyen en el comportamiento de este tiempo, estas características que describen en algún sentido el tiempo de supervivencia serán llamadas *variables explicativas o covariables*.

Los ensayos clínicos funcionan de manera adecuada para ejemplificar la situación anterior. Supóngase que se desea inferir acerca del tiempo de supervivencia de un grupo de pacientes con algún tipo de cáncer en particular, la muerte de estos individuos generalmente está relacionada con su edad, el género, el peso, la etapa o estadio del cáncer, el conteo de glóbulos blancos, el número de plaquetas en la sangre y el tipo de tratamiento con el que se decide tratar la afección, esta última variable explicativa resulta ser muy útil al momento de comparar dos o más distribuciones de supervivencia en las investigaciones biomédicas.

Evidentemente las variables explicativas que describen el tiempo que transcurre desde la detección del cáncer hasta el fallecimiento del paciente se pueden identificar como datos cuantitativos y/o cualitativos. Por ejemplo el número de visitas al hospital se registra como un dato discreto, sin embargo, el peso o el logaritmo del conteo de glóbulos blancos en la sangre, estarán asociados a un valor continuo. Por

otro lado, las variables relacionadas con el género o el tipo de tratamiento pueden definirse de manera nominal, mientras que el tipo de cáncer puede ser jerarquizado para determinar los órdenes; primero, segundo, tercero y cuarto.

Una de las características más importantes del conjunto de variables explicativas que describen el tiempo de supervivencia en un experimento es que estas variables pueden variar su registro según su ubicación temporal, lo que provoca fluctuaciones aleatorias en la *variable de respuesta o tiempo de falla*. Indiscutiblemente variables como el número de plaquetas en la sangre, el peso y el conteo de góbulos blancos en un paciente con cáncer podrían no ser los mismos el día de hoy que en la próxima evaluación médica, sin embargo, características como el género, la etnia y el tratamiento se mantienen fijas a lo largo del experimento.

En los *modelos de regresión* se estudia el tiempo de supervivencia y el efecto que tienen las covariables en la falla que presentan los individuos en el estudio. Al conjunto de individuos en estudio se les conoce como *población heterogénea*, debido a la influencia que tienen las covariables en el tiempo de falla de cada individuo.

Los modelos de regresión más utilizados en el análisis de supervivencia son:

- **PH** (Proportional hazard models): Modelos de riesgos proporcionales.
- **AFT** (Accelerated Failure Time): Modelos de tiempo de vida acelerada.

Los modelos de riesgos proporcionales pueden ser paramétricos, semiparamétricos o no paramétricos. Generalmente las pruebas paramétricas proporcionan mejores resultados cuando se selecciona de manera adecuada la distribución que describe el fenómeno en estudio, sin embargo, las pruebas no paramétricas son atractivas debido a su facilidad y simplicidad al momento de conjeturar suposiciones acerca de los datos, lo que les permite ser más relevantes que las pruebas paramétricas en algunas aplicaciones en las que la población no sigue una distribución teórica.

En los modelos de vida acelerada los sujetos en estudio son sometetidos a condiciones significativamente más altas que las que se perciben normalmente con el propósito de acelerar una falla temprana en el individuo, debido a que la falla en condiciones normales necesita un intervalo prolongado de tiempo para poder ser observada. Por ejemplo, el desempeño de un sistema de refrigeración industrial depende de factores como la temperatura, la humedad y el voltaje, por lo tanto, si se pretende acelerar la falla de este sistema, los investigadores podrían alterar significativamente los factores asociados al rendimiento del equipo.

Generalmente los modelos de vida acelerada se trabajan con un enfoque paramétrico, sin embargo, también existe una variante semiparamétrica de estos modelos.

Aunque los modelos de regresión pueden describirse a partir de muchas variables explicativas, el análisis de supervivencia realizado sobre el conjunto de datos es considerado como una técnica univariada, pues existe una única variable de respuesta,

el tiempo de falla.

Los modelos de regresión son adecuados para modelar no sólo la relación entre la tasa de supervivencia y el tiempo, sino también permiten estimar el efecto de una o más variables explicativas sobre la variable de respuesta o tiempo de falla.

Análogo a los modelos clásicos de regresión, el número $\boldsymbol{p}$ de variables que intervienen en el modelo determina la dificultad con la que se ha de trabajar, en respuesta a esta situación se propone plantear el modelo en función de aquellas variables cuya influencia sea más significativa que las otras.

Como se revisó en el capítulo anterior, el tiempo de supervivencia T puede estar censurado al tiempo C por cualesquiera de los tipos de censura estudiados hasta el momento, aunque generalmente se trabaja con datos censurados por la derecha, de tal modo, que una muestra observada en los modelos de regresión consiste de vectores de la forma:

$$(T_i, \delta_i, \boldsymbol{X_i})$$

donde, si C_i es una observación censurada por la derecha[1] entonces $T_i = \text{mín}(t_i, C_i)$, δ_i es la función indicadora de censura (0 si la observación está censurada y 1 si no está censurada) y $\boldsymbol{X_i}$ es el vector de variables explicativas de orden $\boldsymbol{p}$ correspondiente a la i-ésima observación, también conocido como vector de covariables. Usualmente estos datos son acomodados en una tabla en la cual se exhibe el valor de cada una de las entradas del vector $(T_i, \delta_i, \boldsymbol{X_i})$.

i	$T_i = t_i$	δ_i	X_i' (Variables explicativas)					
1	t_1	δ_1	$x_{1\,1}$	$x_{1\,2}$	$\ldots$	$x_{1\,j}$	$\ldots$	$x_{1\,p}$
2	t_2	δ_2	$x_{2\,1}$	$x_{2\,2}$	$\ldots$	$x_{2\,j}$	$\ldots$	$x_{2\,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
k	t_k	δ_k	$x_{k\,1}$	$x_{k\,2}$	$\ldots$	$x_{k\,j}$	$\ldots$	$x_{k\,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	t_n	δ_n	$x_{n\,1}$	$x_{n\,2}$	$\ldots$	$x_{n\,j}$	$\ldots$	$x_{n\,p}$

[1] A partir de este capítulo, cuando se hable de una observación censurada, ésta será considerada como una observación con censura por la derecha, a menos que se especifique alguno de los otros tipos de censura anteriormente expuestos.

Estadísticamente, la especificación de un modelo de regresión requiere de una elección apropiada de las variables explicativas y de un componente *error* que está determinado por las covariables cuya influencia se considera despreciable en el modelo. Análogo a la regresión clásica, en la elección de las covariables ($X_i's$), se involucran constantes desconocidas llamadas parámetros. Estos parámetros permiten controlar el comportamiento del modelo y son estimados a partir de las observaciones.

Dado el interés que se tiene en conocer la influencia que tienen las covariables en el tiempo de falla y la flexibilidad que se les da a las X_i's para representar tanto variables aleatorias continuas como dicretas, uno de los modelos de regresión más comúnmente utilizado en los datos de supervivencia es [2]:

$$T_i = exp\,(\boldsymbol{\beta}^{'}\boldsymbol{X_i}) \times \varepsilon \qquad \forall\ i = 1,\, 2, \ldots,\, \mathrm{n}. \tag{2.1}$$

donde:

- $\boldsymbol{\beta}^{'} = (\beta_1,\, \beta_2,\, \ldots,\, \beta_p)$
- $\boldsymbol{X_i} = (x_{i\,1},\, x_{i\,2},\, \ldots,\, x_{i\,p})^{'}$
- $\varepsilon = \varepsilon rror$

como T_i es una variable aleatoria no negativa, se puede aplicar el *logaritmo* en la ecuación (2.1) para linealizar el modelo de regresión de la siguiente manera:

$$\begin{aligned} log\,(T_i) &= log(\, exp\,(\boldsymbol{\beta}^{'}\boldsymbol{X_i}) \times \varepsilon\,) \\ Y_i &= \boldsymbol{\beta}^{'}\boldsymbol{X_i} + \varepsilon^{*} \\ Y_i &= \beta_1\, x_{i\,1} + \beta_2\, x_{i\,2} + \cdots + \beta_p\, x_{i\,p} + \varepsilon^{*} \end{aligned}$$

donde:

- $Y_i = log\,(T_i)$
- $\varepsilon^{*} = log\,(\varepsilon)$

El modelo de regresión será más preciso si se conoce una importante cantidad de información relacionada al proceso con el que ocurre la falla. Análogamente, se tendrá un mayor umbral de incertidumbre en el modelo si se dispone de poca información o información poco significativa relacionada con el tiempo de supervivencia.

Existen otras formas de representar al tiempo de falla utilizando un modelo de regresión, la selección del modelo dependerá de la influencia considerada por parte de las variables explicativas, así como de la experiencia o el conocimiento que se tenga con relación al fenómeno estudiado.

[2] *véase* [3] pag 87.

Otros ejemplos de modelos de regresión en el análisis de supervivencia son;

- $T_i = exp\,(1 + \boldsymbol{\beta}^{'}\boldsymbol{X_i}) \times \varepsilon$
- $T_i = log\,[1 + exp\,(\boldsymbol{\beta}^{'}\boldsymbol{X_i})] \times \varepsilon$

Como ya se había mencionado con anterioridad, al igual que en los modelos de regresión clásica, las variables explicativas pueden mantener un valor fijo o bien un valor aleatorio en el modelo. Estrictamente hablando, si el valor de las covariables es conocido previo a la realización del experimento o si éstas admiten un error despreciable en la descripción de la falla, se puede considerar a $\boldsymbol{X}$ como un vector no aleatorio. De lo contrario, se identificará a $\boldsymbol{x}$ como una realización del vector aleatorio $\boldsymbol{X}$. En ambos casos se utiliza la siguiente notación para describir a la función de supervivencia del tiempo de falla: La función de supervivencia de T dada la covariable $\boldsymbol{x}$ se denota como $S(\cdot\,|\boldsymbol{x})$ y se define de la siguiente manera.

Definición 2.1.1 (Función de supervivencia dado $\boldsymbol{X}$)**.** Sean T una variable aleatoria y $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^{'}$ un vector $\boldsymbol{p}$-dimensional de variables explicativas asociado a un vector de parámetros $\boldsymbol{\beta'} = (\beta_1,\, \beta_2, \ldots,\, \beta_p)$ la función de supervivencia dado el vector $\boldsymbol{X}$, se define como:

$$S\,(\,t\,|\,\boldsymbol{x}\,) = P(T > t\,|\,\boldsymbol{X} = \boldsymbol{x})$$

Análogamente se definen las funciones de densidad $f(\,\cdot\,|\,\boldsymbol{x})$, de riesgo $h(\,\cdot\,|\,\boldsymbol{x})$ y de riesgo acumulado $H(\,\cdot\,|\,\boldsymbol{x})$.

$$\begin{aligned} f(\,t\,|\,\boldsymbol{x}) &= -\,S'(t\,|\,\boldsymbol{x}) \\ h(t\,|\,\boldsymbol{x}) &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t\,|\,T > t,\, \boldsymbol{X} = \boldsymbol{x})}{\Delta t} = \frac{f(\,t\,|\,\boldsymbol{x})}{S(\,t\,|\,\boldsymbol{x})} \\ H(t\,|\,\boldsymbol{x}) &= -\,log\,[S(t\,|\,\boldsymbol{x})] \end{aligned}$$

La elección del enfoque (*paramétrico* y *semiparmétrico*) en el modelo riesgos proporcionales, dependerá si se decide especificar la forma de la llamada ***función de riesgo inicial*** $h_0(t)$, misma que se introduce en la siguiente sección.

2.2. Modelos de Riesgos Proporcionales

En algunas enfermedades tales como como el cáncer o la granulomatosa crónica, los pacientes suelen responder de manera distinta (en beneficio o perjucio de su afección) al tratamiento que reciben para combatir la enfermedad. Uno de los objetivos de las investigaciones biomédicas consiste en determinar que tratamiento resulta ser más eficiente comparando los resultados proporcionados, digamos de un grupo I y un grupo II de individuos a los cuales se les suministró un tratamiento A y un tratamiento B de manera correspondiente.

Evidentemente el tiempo de remisión de un paciente con una determinada enfermedad podría estar directamente relacionado con un conjunto de variables explicativas que influyen en el tiempo de supervivencia del individuo, por lo que los modelos de regresión resultan adecuados para determinar el tratamiento ideal[3].

El modelo de riesgos proporcionales, también conocido como *modelo de riesgo multiplicativo* o *modelo logarítmico de riesgo relativo*, es un modelo de regresión en el análisis de supervivencia en el que se supone que el riesgo al que está sometido un individuo en un grupo es proporcional al riesgo al que está sometido un individuo en otro grupo al mismo tiempo t. Esto último puede ser expresado como:

$$h_2(t) = h_1(t)\,\psi\,, \qquad \forall\ t\ \in\ T \tag{2.2}$$

Donde:

- $h_1(t)$:el riesgo de que un individuo del grupo I presente la falla en el tiempo t
- $h_2(t)$:el riesgo de que un individuo del grupo II presente la falla en el tiempo t
- ψ: (riesgo relativo) es una constante independiente de t.

El valor de ψ, también conocido como *razón de riesgo*, se define como el cociente de los riesgos de falla para un individuo del grupo II relativo a un individuo en el grupo I. Como la función de riesgo $h(t)$ es no negativa, entonces $\psi > 0$ y satisface lo siguiente:

- Si $\psi < 1$ entonces, el riesgo de falla en el tiempo t es menor para un individuo del grupo II.
- Si $\psi > 1$ entonces, el riesgo de falla en el tiempo t es mayor para un individuo del grupo II.
- Si $\psi = 1$ el riesgo de falla en el tiempo t es el mismo en ambos grupos.

[3]En caso de que la remisión del paciente no esté en función de un conjunto de variables explicativas y si las funciones de supervivencia de los grupos de estudio siguen un modelo conocido, se propone utilizar una prueba paramétrica de comparación de funciones de supervivencia tales como la *prueba de razón de verosimilitudes* o la prueba *F de Cox*, para seleccionar el tratamiento adecuado.

Habitualmente, en la práctica estadística en el grupo I se registran individuos de los cuales ya se tiene un conocimiento preliminar en relación al estudio, mientras que en el grupo II los integrantes tienen una variante en el ensayo que los distingue de los individuos del grupo I. Por ejemplo, supóngase que un grupo de científicos desarrolla una nueva vacuna contra la influenza A-HINI. Con el propósito de inferir acerca de la eficiencia de esta nueva vacuna, los investigadores suministrarán al grupo I la vacuna convencional y al grupo II la nueva cepa vacunal. Así pues, de acuerdo con el valor del riesgo relativo ψ se podrá concluir cuál de estas dos vacunas resulta ser la más eficiente.

Dado que en la ecuación (2.2) el riesgo relativo es mayor que cero ($\psi > 0$), resulta conveniente expresar este término como:

$$\psi = exp\,(\beta) \tag{2.3}$$

En particular β representa un parámetro cuyo valor es el logaritmo natural de la *razón de riesgos* ψ, por lo que β puede tomar cualquier valor en la recta real, es decir, $\beta \in \mathbb{R}$. Si $\beta < 0$ el riesgo de falla en el tiempo t es menor para un individuo del grupo II relativo a un individuo en el grupo I, valores de $\beta > 0$ indican un riesgo de falla mayor para un individuo del grupo II, mientras $\beta = 0$ significa que se tiene la misma exposición de falla en el instante t.

La igualdad (2.3), permite expresar al modelo (2.2) de una forma más sencilla y generalizada al momento de incorporar a las covariables que intervienen en el modelo. Supóngase que se dispone de los datos de supervivencia de n individuos, sea $h_i(t)$, i =1, 2,. . . , n, la función de riesgo para el i-ésimo individuo y $h_0(t)$ la función de riesgo para un individuo en el grupo I, entonces, la función de riesgo para un individuo con riesgo proporcional al grupo I está dada por:

$$h_i(t) = h_0(t)\,\psi \tag{2.4}$$

Regresando al ejemplo de la vacuna, sea X una variable indicadora, es decir, que toma el valor de cero si un individuo es tratado con la vieja vacuna y uno si es tratado con el nuevo suministro, utilizando la expresión (2.4) y la igualdad (2.3), entonces, la función de riesgo para el i-ésimo individuo puede ser escrita como:

$$h_i(t) = h_0(t)\,exp\,(\beta\,x_i) \tag{2.5}$$

donde x_i es el valor de X para el i-ésmio individuo en el estudio.

El modelo (2.5) es conocido como el ***modelo de riesgos proporcionales para la comparación de dos grupos***. La forma en la que está diseñado este modelo permite generalizar la situación donde el tiempo de falla en un determinado tiempo t, depende de los valores x_{i1},x_{i2},. . . ,x_{ip} de p variables explicativas X_1,X_2,. . . ,X_p.

En el **modelo general de riesgos proporcionales** o modelo de **PH** por sus iniciales en inglés, se considera un conjunto de valores para cada individuo x_{i1}, $x_{i2}, \ldots, x_{ip}$ correspondientes a las $\boldsymbol{p}$ variables explicativas $X_1, X_2, \ldots, X_p$, las cuales involucran a los parámetros β's que intervienen en el modelo. El modelo general de riesgos proporcionales para el i-ésimo individuo se escribe como:

$$h_i(t \mid \boldsymbol{X_i}) = h_0(t)\, \psi(\boldsymbol{X_i}\,;\boldsymbol{\beta}) \tag{2.6}$$

Existen varios métodos de verificación de este supuesto, entre los que destacan los métodos de *residuales de Cox-Snell* y los *residuales martingala*. Si un modelo de regresión satisface la igualdad anterior, se dice que el modelo verifica la condición de riesgos proporcionales.

Proposición 2.2.1. Sea $\psi = \psi(\boldsymbol{X}\,;\boldsymbol{\beta})$ una función independiente de t, entonces la función de supervivencia del *modelo de riesgos proporcionales* satisface:

$$S(t \mid \boldsymbol{X}) = [\, S_0(t)\,]^{\psi}$$

donde $S_0(t)$ denota la *función de supervivencia inicial*.

Demostración. Obteniendo el logaritmo natural de ambos lados de la igualdad, se tiene que

$$\begin{aligned} log\,(\, S(t \mid \boldsymbol{X})\,) &= \psi\, log\,(\, S_0(t)\,) \\ -H(t \mid \boldsymbol{X}) &= -\psi\, H_0(t) \\ -\int_0^t h(u \mid \boldsymbol{X})\, du &= -\int_0^t \psi\, h_0(u)\, du \\ \int_0^t h(u \mid \boldsymbol{X})\, du &= \int_0^t \psi\, h_0(u)\, du \end{aligned}$$

por del teorema fundamental del cálculo, derivando ambos lados de la igualdad con respecto al tiempo de supervivencia y utilizando la hipótesis de que ψ es independiente de t

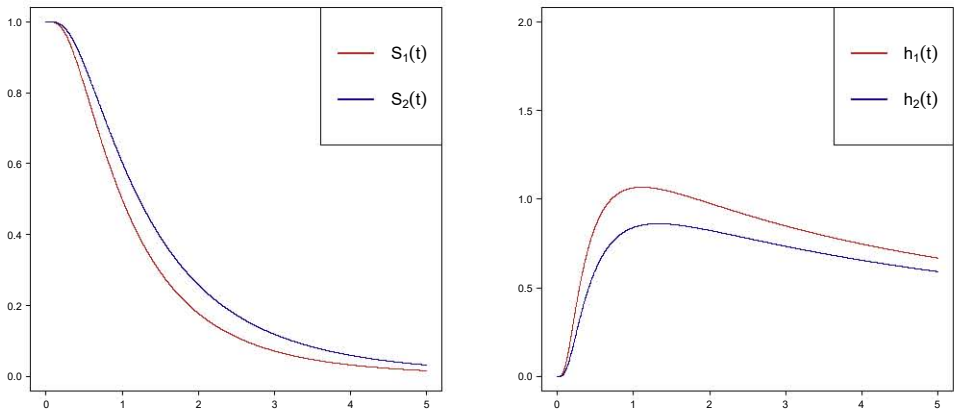
$$\frac{d}{dt}\int_0^t h(u \mid \boldsymbol{X})\, du = \frac{d}{dt}\int_0^t \psi\, h_0(u)\, du$$

obteniendo como resultado

$$h(t \mid \boldsymbol{X}) = h_0(t)\, \psi$$

por lo tanto, la función de supervivencia satisface el supuesto de riesgos proporcionales y tiene la forma especificada, de este modo la proposición queda demostrada. □

Una verificación gráfica del supuesto de proporcionalidad entre los riesgos de dos individuos es que las funciones de riesgo y supervivencia no se intersectan. Este resultado se puede observar en la siguiente gráfica.



(a) (b)

Figura 2.1: Obsérvese como la gráfica de supervivencia del individuo 2 queda completamente por encima de la gráfica del primer idividuo, análogamente, la gráfica del riesgo del individuo 1 se mantiene por arriba de la del otro individuo.

Regresando a la expresión (2.6) y análogo a la igualdad (2.2), se observa que $\psi(\boldsymbol{X_i}\,;\beta)$ es una función no negativa y que por el momento supondremos independiente de t, en el modelo más comúnmente utilizado en el análisis de supervivencia se supone la siguiente expresión para el riesgo relativo:

$$\psi(\boldsymbol{X_i}\,;\boldsymbol{\beta}) = exp\,(\eta_i) \tag{2.7}$$

donde η_i es una combinación lineal de p variables explicativas:

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \forall \;\; i = 1,\, 2,\, 3,\ldots,\, \text{n.}$$

usualmente, en la literatura estadística el modelo general de riesgos proporcionales (2.6) se escribe como:

$$h_i(t \mid \boldsymbol{X_i}) = h_0(t)\, exp\,(\boldsymbol{\beta'}\, \boldsymbol{X_i}) \tag{2.8}$$

donde:

- $\boldsymbol{X_i} = (x_{i1},\, x_{i2}, \ldots,\, x_{ip})'$: es el vector de covariables para el i-ésimo individuo en estudio.
- $\boldsymbol{\beta'} = (\beta_1,\, \beta_2, \ldots,\, \beta_p)$: es el vector de parámetros desconocidos asociados a las p covariables en el modelo, también conocidos como coeficientes de regresión.
- $exp\,(\boldsymbol{\beta'}\, \boldsymbol{X_i})$: es una función no negativa y que por construcción no depende de t, es decir, supone que las covariables involucradas en el modelo no cambian en el tiempo, y que los valores de estas variables han sido registrados en el tiempo

origen del estudio. En principio, cualquier función $\psi(\boldsymbol{X_i}\,;\boldsymbol{\beta})$ no negativa podría utilizarse para resumir los efectos de las variables explicativas en la falla del individuo, usualmente se utiliza una representación exponencial.

- $h_0(t)$: conocida como la ***función de riesgo inicial***, esta función representa un riesgo que se supone que es común en todas las personas en el estudio, por lo que $h_0(t)$ es una función que no depende de los valores de $\boldsymbol{X}$.

Como el modelo de riesgos proporcionales en (2.8) puede ser expresado de la forma:

$$log\Big(\frac{h_i(t)}{h_0(t)}\Big) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\,,$$

éste se considera como un modelo lineal para el logaritmo natural de la razón de riesgos. Una segunda observación que puede hacerse con respecto a la expresión (2.8) es que no hay un término constante en el componente lineal, dicho término puede ser incluido de la siguiente manera:

$$h_i(t\,|\,\boldsymbol{X_i}) = h_0(t)\,exp\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\,),$$

con el proposito de eliminar el término constante, la función de riesgo inicial suele ser cambiada de escala dividiendo $h_0(t)$ entre $exp(\beta_0)$.

$$\begin{aligned} h_i(t\,|\,\boldsymbol{X_i}) &= \frac{h_0(t)}{exp\,(\beta_0)}\,exp\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\,) \\ &= h_0^*(t)\,exp\,(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\,) \end{aligned}$$

Evidentemente la característica principal del modelo de **PH**, misma por la que lleva su nombre, es que las tasas de riesgo de dos individuos con valores diferentes en las covariables son proporcionales, es decir, se supone que existe una relación proporcional entre las funciones de riesgo correspondientes a diferentes individuos con distinto vector de covariables, esto último se interpreta matemáticamente del siguiente modo.

Sean $\boldsymbol{X_i} \neq \boldsymbol{X_j}$ los vectores de covariables correspondientes al i-ésimo y j-ésimo inviduo, entonces la relación del riesgo para cualquier tiempo t es:

$$\frac{h_i(t\,|\,\boldsymbol{X_i})}{h_j(t\,|\,\boldsymbol{X_j})} = \frac{h_0(t)\,\psi(\boldsymbol{X_i}\,;\boldsymbol{\beta}\,)}{h_0(t)\,\psi(\boldsymbol{X_j}\,;\boldsymbol{\beta}\,)}$$

si se interpreta a $\psi(\boldsymbol{X}\,;\boldsymbol{\beta})$ como la expresión (2.7), la relación de riesgo anterior queda como:

$$\begin{aligned} \frac{h_i(t\,|\,\boldsymbol{X_i})}{h_j(t\,|\,\boldsymbol{X_j})} &= exp\,(\,\boldsymbol{\beta'}\,(\boldsymbol{X_i} - \boldsymbol{X_j}\,)\,) \\ &= exp\,(\,\beta_1(x_{i1} - x_{ji}) + \beta_2(x_{i2} - x_{j2}) + \cdots + \beta_p(x_{ip} - x_{jp})\,) \end{aligned}$$

El estimador puntual de esta razón de riesgos es:

$$\frac{\widehat{h_i(t \mid \boldsymbol{X}_i)}}{h_j(t \mid \boldsymbol{X_j})} = exp\,(\,\widehat{\boldsymbol{\beta'}}\,(\boldsymbol{X_i} - \boldsymbol{X_j}\,)\,)$$

donde $\widehat{\boldsymbol{\beta}}$ es el estimador máximo verosímil de $\boldsymbol{\beta}$.

Usualmente la estimación de los parámetros β's se hace con un *software estadístico* y dependiendo de los datos de supervivencia que se dispongan, conviene trabajar con los coeficientes $\hat{\beta}_k$ o $\exp(\hat{\beta}_k)$.

Hasta el momento solo se ha considerado el caso en el que las variables explicativas del modelo de regresión mantienen un valor fijo en el tiempo. Supóngase ahora que

$$h(t \mid \boldsymbol{X_t}) = h_0(t)\,\psi(\boldsymbol{X_t}\,;\boldsymbol{\beta}) \tag{2.9}$$

es decir, el vector de covariables para un individuo en el tiempo t_1 es distinto para el mismo individuo en el tiempo t_2 ($\boldsymbol{X_{t_1}} \neq \boldsymbol{X_{t_2}}$), por lo que el factor de proporcionalidad varía con el tiempo de falla en lugar de ser constante como en los casos estudiados previamente. Evidentemente la función de supervivencia y las otras funciones relacionadas con ésta, tendrán una forma más complicada, suponiendo que $\psi(\boldsymbol{X_t}\,;\boldsymbol{\beta})$ puede representarse como

$$\psi(\boldsymbol{X_t}\,;\boldsymbol{\beta}) = exp\,(\boldsymbol{\beta'}\,\boldsymbol{X_t})$$

entonces, la función de supervivencia para el modelo de regresión presentado en (2.9) tiene la expresión

$$S(t \mid \boldsymbol{X_t}) = exp\Big[-\int_0^t h(u \mid \boldsymbol{X_u})\,du\Big] = exp\Big[-\int_0^t h_0(u)\,exp(\boldsymbol{\beta'}\,\boldsymbol{X_u})\,du\Big]$$

debido a que $\boldsymbol{X}$ tiene una relación de variabilidad temporal, la forma de simplificar la expresión de la función de supervivencia resulta un poco más compleja que cuando se considera a $\boldsymbol{X}$ como un vector independiente de t, sin embargo, es posible llegar a una expresión más sencilla suponiendo que cada covariable es constante en un intervalo de tiempo t.

Supóngase que una covariable puede tomar dos valores dependiendo si el tiempo de falla ocurre antes o después de una fecha.

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{X_1} \quad \text{si} \quad t < s \\ \boldsymbol{X} &= \boldsymbol{X_2} \quad \text{si} \quad t \geq s \end{aligned}$$

entonces la función de supervivencia puede escribirse como:

$$\begin{aligned} S(t \mid \boldsymbol{X_t}) &= exp\Big[- \int_0^s h_0(u)\, exp(\boldsymbol{\beta'}\, \boldsymbol{X_1})\, du \, - \int_s^t h_0(u)\, exp(\boldsymbol{\beta'}\, \boldsymbol{X_2})\, du \Big] \\ &= exp\Big[- \psi_1 \int_0^s h_0(u)\, du - \psi_2 \int_s^t h_0(u)\, du \Big] \\ &= exp\Big[- \psi_1 \int_0^s h_0(u)\, du \Big] exp\Big[- \psi_2 \int_s^t h_0(u)\, du \Big] \\ &= [S_0(s)]^{\psi_1} \frac{[S_0(t)]^{\psi_2}}{[S_0(s)]^{\psi_2}} \end{aligned}$$

Este resultado significa, que la probabilidad de que un individuo con vector de covariables $\boldsymbol{X_t}$ sobreviva al tiempo t es proporcional a la probabilidad de que cualquier individuo sobreviva al tiempo s, multiplicada por la probabilidad de que éste sobreviva al tiempo t condicionado a que sobrevive al tiempo s.

En las siguientes secciones se estudiarán los enfoques paramétricos y semiparamétricos del modelo de *riesgos proporcionales* (**PH**), en particular, debido al objetivo de esta tesis, la mayor parte del material expuesto estará dedicado a los modelos paramétricos de los riesgos proporcionales.

Como ya se había mencionado con anterioridad, los modelos de riesgos proporcionales pueden ser paramétricos o semiparamétricos, dependiendo de la elección de la función de riesgo inicial $h_0(t)$ en el modelo subyacente.

- **Modelos paramétricos de regresión**: Se caracterizan por especificar a la función de riesgo inicial $h_0(\cdot;\alpha)$ y $\psi(\cdot;\beta)$ paramétricamente, es decir, $h_0(\cdot;\alpha)$ sigue un modelo de probabilidad teórico, mientras que $\psi(\cdot;\beta)$ puede ser representado paramétricamente como se indica en la expresión (2.8).
- **Modelos semiparamétricos de regresión**: Estos modelos difieren de los anteriores en el sentido de que la función $h_0(\cdot)$ no está asociada con una distribución, asimismo, $\psi(\cdot;\beta)$ mantiene una expresión paramétrica, por lo tanto, se dice que éste es un modelo semiparamétrico.

Los objetivos pirncipales de estos modelos son:

- Estimar los parámetros β's, para lo cual usualmente se utiliza el método de máxima verosimilitud, aunque en ocasiones, es conveniente obtener estos estimadores a través de métodos numéricos.
- Probar hipótesis acerca de los parámetros β's.
- Construir un modelo de regresión para interpretar de manera adecuada la variable de interés en el estudio.
- Estimar a la función de riesgo inicial.

2.2.1. Modelo de Riesgos Proporcionales de Cox

El modelo de Cox es un modelo de regresión que se desarrolla bajo el supuesto de los riesgos proporcionales, es decir, los riesgos correspondientes a dos individuos distintos (con diferentes valores de covariables) son proporcionales a lo largo del tiempo. Para estimar el riesgo con el que ocurre la falla en el modelo, la regresión de Cox busca interpretar a la razón de riesgos como una función lineal de aquellas covariables que describen el evento de interés.

El modelo clásico de riesgos proporcionales de Cox está dado por:

$$h(t \,|\, \boldsymbol{X}) = h_0(t)\, exp\,(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

no obstante, la función $exp\,(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$ puede ser intercambiada por cualquier otra función $\psi(\boldsymbol{X}\,;\boldsymbol{\beta})$ no negativa y que, preferentemente, pueda ser reducida a una expresión lineal.

El modelo de regresión de Cox es el más utilizado en los modelos de la familia de **PH** y se caracteriza por ser un modelo semiparamétrico, es decir:

- La función de riesgo inical $h_0(t)$ no es conocida y se identifica como la parte no paramétrica del modelo.
- Mientras que la función $\psi(\boldsymbol{X}\,;\boldsymbol{\beta})$ asume una forma paramétrica en el modelo.

El objetivo principal del modelo de regresión de Cox consiste en estimar los parámetros que intervienen en el modelo, por otro lado, aunque la función de riesgo inicial no es conocida, el modelo de Cox no necesita un estimador para $h_0(t)$.

Los parámetros desconocidos β's en el modelo de regresión pueden ser estimados por el método de máxima verosimilitud, la construcción de esta verosimilitud, conocida como ***verosimilitud parcial***, se basa en el supuesto de que los intervalos sucesivos de tiempo, en los cuales ocurre la falla, no contienen información de la incidencia de las covariables sobre el tiempo de falla. La función de verosimilitud parcial se puede obtener del siguiente modo:

Supóngase que se dispone de los tiempos de falla t_1, t_2, $\ldots$, t_n correspondientes a n inividuos, los cuales son ordenados de menor a mayor $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ (por el momento no admitiremos repeticiones) de tal manera que $t_{(j)}$ representa la j-ésima estadística de orden. Sea $R(t_{(j)}) = \{i : t_i \geq t_{(j)}\}$ el grupo de individuos que están vivos y no censurados justo antes del tiempo $t_{(j)}$, es decir, el conjunto de individuos en riesgo al tiempo $t_{(j)}$. Entonces la probabilidad de que el i-ésimo individuo presente la falla al tiempo $t_{(j)}$ dado que un individuo del conjunto en riesgo $R(t_{(j)})$ presenta la falla en el tiempo $t_{(j)}$, se denota y desarrolla como:

$$P(\, i \text{ falle en } t_{(j)} \,|\, \text{un individuo de } R(t_{(j)}) \text{ falla en } t_{(j)})$$

$$= \frac{P(\, i \text{ falla en } t_{(j)}\,)}{P(\, \text{un individuo falla en } t_{(j)}\,)}$$

como la falla en el denominador la puede presentar cualquier individuo perteneciente al conjunto en riesgo, la expresión anterior puede ser escrita como:

$$
\begin{aligned}
&= \frac{P(\, i \ \text{falla en } t_{(j)}\,)}{\sum\limits_{k \in R(t_{(j)})} P(\,\text{el individuo } k \text{ falla en } t_{(j)}\,)} \\
&\simeq \frac{P[\, i \ \text{falla en } (\, t_{(j)}\,,\, t_{(j)} + \Delta t\,)\,]\,/\Delta t}{\sum\limits_{k \in R(t_{(j)})} P[\,\text{el individuo } k \text{ falla en } (\, t_{(j)}\,,\, t_{(j)} + \Delta t\,)\,]\,/\Delta t} \\
&= \frac{lim_{\Delta t_{(j)} \to 0}\, P[\, i \ \text{falla en } (\, t_{(j)}\,,\, t_{(j)} + \Delta t\,)\,]\,/\Delta t}{lim_{\Delta t_{(j)} \to 0} \sum\limits_{k \in R(t_{(j)})} P[\,\text{el individuo } k \text{ falla en } (\, t_{(j)}\,,\, t_{(j)} + \Delta t\,)\,]\,/\Delta t}
\end{aligned}
$$

desarrollando los límites, en el numerador se obtiene la función de riesgo para el i-ésimo individuo, mientras que en el denominador, si se incorpora el límite en la suma, se tiene la suma de las funciones de riesgo de los k individuos expuestos al riesgo al tiempo $t_{(j)}$ este resultado se muestra en la siguiente fracción:

$$
\begin{aligned}
&= \frac{h_i(t_{(j)})}{\sum_{k \in R(t_{(j)})} h_k(t_{(j)})} \\
&= \frac{h_0(t_{(j)})\, \psi(\boldsymbol{X_i}(t_{(j)})\,;\boldsymbol{\beta}\,)}{\sum_{k \in R(t_{(j)})} h_0(t_{(j)})\, \psi(\boldsymbol{X_k}(t_{(j)})\,;\boldsymbol{\beta}\,)}
\end{aligned}
$$

como en el denominador la función $h_0(t_{(j)})$ no depende de k, la expresión anterior queda como:

$$\frac{\psi(\boldsymbol{X_i}(t_{(j)})\,;\boldsymbol{\beta}\,)}{\sum_{k \in R(t_{(j)})} \psi(\boldsymbol{X_k}(t_{(j)})\,;\boldsymbol{\beta}\,)}$$

Este último resultado nos permite obtener la representación de Cox de la función de verosimilitud parcial $L(\boldsymbol{\beta})$, como:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\psi(\boldsymbol{X_i}(t_{(j)})\,;\boldsymbol{\beta}\,)}{\sum_{k \in R(t_{(j)})} \psi(\boldsymbol{X_k}(t_{(j)})\,;\boldsymbol{\beta}\,)} \tag{2.10}$$

donde r es el total de individuos que pertencen al conjunto en riesgo y $\psi(\boldsymbol{X_i}(t_{(j)})\,;\boldsymbol{\beta})$ es una función que admite una representación paramétrica con $X_i(t_{(j)})$ el vector de covariables del i-ésmio individuo cuya falla se presenta al tiempo $t_{(j)}$.

Obsérvese que la función de verosimilitud parcial (2.10) solamente es válida para aquellos individuos que no registran observaciones censuradas. Sean t_1, t_2, ..., t_n los tiempos de falla de n individuos y δ_i una función indicadora de censura que toma el valor cero si el i-ésimo tiempo de falla es censurado y uno si no lo es, entonces, la función de verosimilitud parcial puede ser expresada en la forma siguiente:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\psi(\boldsymbol{X_j}(t_{(i)})\,;\boldsymbol{\beta}\,)}{\sum_{k \in R(t_{(i)})} \psi(\boldsymbol{X_k}(t_{(i)})\,;\boldsymbol{\beta}\,)} \right]^{\delta_i} \tag{2.11}$$

Sea:

$$\psi\left(\boldsymbol{X}(t_{(i)})\,;\boldsymbol{\beta}\,\right) = exp\left(\boldsymbol{\beta'}\,\boldsymbol{X}(t_{(i)})\right)$$

entonces, la expresión (2.11) queda como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\frac{exp\left(\boldsymbol{\beta'}\,\boldsymbol{X_j}(t_{(i)})\,\right)}{\sum_{k\,\in\,R(t_{(i)})}exp\left(\boldsymbol{\beta'}\,\boldsymbol{X_k}(t_{(i)})\,\right)}\right]^{\delta_i}$$

La correspondiente función *log-verosimilitud* es entonces:

$$log\left(\,L(\boldsymbol{\beta})\,\right) = \sum_{i=1}^{n}\delta_i\left[\boldsymbol{\beta'}\,\boldsymbol{X_j}(t_{(i)}) - log\left(\sum_{k\,\in\,R(t_{(i)})}exp\left(\,\boldsymbol{\beta'}\,\boldsymbol{X_k}(t_{(i)})\,\right)\right)\right] \qquad (2.12)$$

Habitualmente los estimadores máximo verosímiles de los parámetros β's se obtienen maximizando la función de *verosimilitud parcial* o de forma equivalente maximizando la función *log-verosimilitud parcial* para $\beta_1,\,\beta_2,\ldots,\,\beta_p$. Dada la dificultad que puede generar este cálculo en el proceso, es conveniente utilizar algún *software estadístico* que nos ayude a obtener estos parámetros.

A partir de la relación (2.3) y teniendo los parámetros estimados, es posible construir un intervalo del $(1-\alpha)\times 100\,\%$ de confianza para ψ de la forma:

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

esto debido a que los estimadores obtenidos por máxima verosimilitud son asintóticamente normales[4] (en este caso el estimador para la razón de riesgo).

Dado que se supone $\psi\left(\,\cdot,\,\boldsymbol{\beta}\right) = exp(\boldsymbol{\beta})$, entonces:

$$\hat{\psi} = exp\left(\,\widehat{\boldsymbol{\beta}}\,\right)$$

luego por el método Delta[5] sabemos que para X variable aleatoria se satisface:

$$Var\left(g(X)\right) \approx \left(\,\frac{d\,g(X)}{d\,X}\,\right)^2 Var\left(X\right)$$

lo que implica que:

$$\begin{aligned} Var\left(\hat{\psi}\right) &= \left(\,exp\left(\widehat{\boldsymbol{\beta}}\right)\right)^2 Var\left(\widehat{\boldsymbol{\beta}}\right) \\ \\ s.e\left(\hat{\psi}\right) &= exp(\widehat{\boldsymbol{\beta}})\,s.e\left(\widehat{\boldsymbol{\beta}}\right) \end{aligned}$$

Dando como resultado en términos de $\hat{\beta}'$s, el siguiente intervalo de confianza para el riesgo relativo ψ.

$$\left(\,exp\left(\widehat{\boldsymbol{\beta}}\,\right) - z_{1-\alpha/2}\,exp(\widehat{\boldsymbol{\beta}})\,s.e\left(\widehat{\boldsymbol{\beta}}\right),\,exp\left(\widehat{\boldsymbol{\beta}}\,\right) + z_{1-\alpha/2}\,exp(\widehat{\boldsymbol{\beta}})\,s.e\left(\widehat{\boldsymbol{\beta}}\right)\,\right)$$

[4]Esta propiedad se demuestra más adelante en la *sección* 2.2.2.

[5]El método Delta es una técnica estadística que permite aproximar una función de una variable aleatoria para un estimador estadístico asintóticamente normal, a partir del conocimiento de la varianza del estimador mismo.

Los errores estándar también pueden ser calculados a partir de un *software estadístico.*

En particular la verosimilitud parcial de Cox definida en (2.11) no considera el caso en el cual se admiten repeticiones en el tiempo de supervivencia, es decir, que dos o más individuos presentan la falla al mismo tiempo. Afortunadamente existen modificaciones de la verosimilitud parcial que consideran esta situación.

Sean $t_{(1)} < t_{(2)} < \cdots < t_{(s)}$ los s distintos tiempos de falla ordenados, d_i el número de fallas que ocurren en $t_{(i)}$ y $\mathbb{D}_i$ el conjunto de individuos cuya falla ocurre en $t_{(i)}$, si se define $\boldsymbol{v_i}$ como la suma de las p covariables del vector $\boldsymbol{X_k}(t_j)$, correspondiente al k-ésimo individuo, sobre todo los individuos que mueren en $t_{(i)}$, esto es:

$$\boldsymbol{v_i} = \sum_{j \in \mathbb{D}_i} \boldsymbol{X_k}(t_j)$$

Entonces, la función de ***verosimilitud parcial con empates en el tiempo de falla*** desarrollada por Norman Breslow (1974) se expresa como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{s} \frac{exp\,(\,\boldsymbol{\beta'}\,\boldsymbol{v_i}\,)}{\left[\sum_{k \in R(t_{(j)})} exp\left(\,\boldsymbol{\beta'}\,\boldsymbol{X_k}(t_{(j)})\,\right)\right]^{d_i}} \tag{2.13}$$

Evidentemente los estimadores del vector $\boldsymbol{\beta}$ se obtienen maximizando la función *log-verosimilitud parcial con empates* y los intervalos de confianza para el riesgo se construyen de manera análoga a los obtenidos a partir de (2.12).

El método de Efron (1977) expuesto en [6] sugiere otra forma de obtener la función de verosimilitud parcial cuando existen empates en los tiempos de falla.

A continuación se presentan algunos ejemplos relacionados con el cálculo de la función de verosimilitud considerando la presencia y la no presencia de empates en el tiempo de falla.

Ejemplo 2.2.2. El siguiente ejemplo tiene como propósito exhibir la función de verosimilitud para los individuos cuyo tiempo de falla se presenta en la Figura 2.2, una revisión más detallada de este ejemplo puede encontrarse en [1] pag.66.

Dado que dos individuos de la muestra están censurados y no se tienen tiempos de falla repetidos, la función de verosimilitud parcial sobre los tiempos de supervivencia puede ser expresada de la forma (2.11). En este caso los individuos en riesgo para los correspondientes tiempos de falla ordenados son:

$$R(t_{(1)}) = \{1,\, 2,\, 3,\, 4,\, 5\}, \quad R(t_{(2)}) = \{1,\, 2,\, 4\} \quad \text{y} \quad R(t_{(3)}) = \{4\}$$

Recordemos que $\psi(\boldsymbol{X_j}(t_{(i)})\,;\boldsymbol{\beta}\,)$, es una función de p covariables para el j-ésimo individuo que presenta la falla en el tiempo $t_{(i)}$. Para facilitar la notación suponemos que:

$$\psi\,(\boldsymbol{X_j}(t_{(i)})\,;\boldsymbol{\beta}\,) = \psi(j)$$

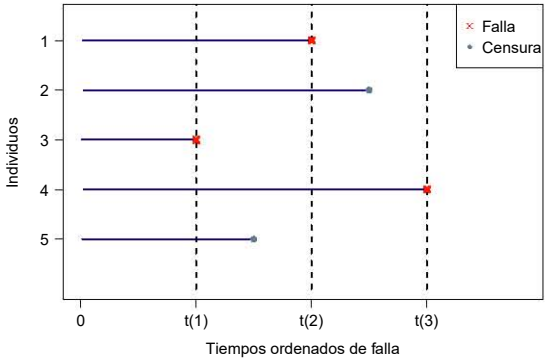


Figura 2.2: Tiempos de falla y censura.

Debido a que hay un total de 3 fallas observadas, la función de verosimilitud parcial estará conformada de la siguiente forma:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\psi(\boldsymbol{X_j}(t_{(i)})\,;\boldsymbol{\beta}\,)}{\sum_{k\,\in\, R(t_{(i)})} \psi(\boldsymbol{X_k}(t_{(i)})\,;\boldsymbol{\beta}\,)} \right]^{\delta_i} = \prod_{i=1}^{n} \left[\frac{\psi(j)}{\sum_{k\,\in\, R(t_{(i)})} \psi(k)} \right]^{\delta_i} =$$

$$\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \times \frac{\psi(1)}{\psi(1)+\psi(2)+\psi(4)} \times \frac{\psi(4)}{\psi(4)}$$

Ejemplo 2.2.3. Considere la siguiente tabla con datos de supervivencia para 6 individuos (*este ejemplo puede encontrarse de forma ilustrada en* [2]).

j	t_j	δ_j	X_j
1	51	1	50
2	51	1	47
3	322	1	48
4	828	0	42
5	339	0	54
6	551	1	50

Como existen empates en los tiempos de falla, es conveniente clasificar los datos y escribir los conjuntos en riesgo $R(t_{(i)})$ para finalmente poder expresar la verosimilitud parcial.

individuo j	falla de orden i	$t_{(i)}$	$\delta_{(j)}$	d_i	$X_{(j)}$	$R(t_{(i)})$
$1, 2$	1	51	$1, 1$	2	$50, 47$	{1, 2, 3, 4, 5, 6}
3	2	322	1	1	48	{3, 4, 5, 6}
5	3	339	0	1	54	
6	4	551	1	1	50	{4, 6}
4	5	838	0	1	42	

La tabla anterior muestra que se tienen 5 tiempos de falla distintos, de los cuales 3 son conocidos y dos son censurados, lo que implica que tendremos 3 factores para la función de verosimilitud parcial.

Utilizando la expresión (2.13), se tiene que el primer factor de la verosimilitud parcial es:

$$\frac{e^{(50+47)\,\beta}}{(\,e^{50\,\beta}+e^{47\,\beta}+e^{48\,\beta}+e^{54\,\beta}+e^{50\,\beta}+e^{42\,\beta}\,)^2}$$

Mientras que el segundo está dado por:

$$\frac{e^{(48)\,\beta}}{e^{48\,\beta}+e^{54\,\beta}+e^{50\,\beta}+e^{42\,\beta}}$$

y finalmente el último factor es:

$$\frac{e^{(50)\,\beta}}{e^{50\,\beta}+e^{42\,\beta}}$$

Por lo tanto, la función de verosimilitud parcial $L(\boldsymbol{\beta})$ es igual al producto de los tres factores obtenidos con anterioridad.

Ejemplo 2.2.4. A continuación se presenta un ejemplo que muestra como obtener los estimadores β's del modelo clásico de riesgos proporcionales de Cox, utilizando el *software estadístico* **R**.

Considere que se tiene la información de los tiempos de supervivencia disponibles en la base de datos **larynx** (*véase* el Apéndice B), la cual muestra el tiempo de remisión de un grupo de pacientes diagnosticados con cáncer de laringe.

Supóngase que el tiempo de supervivencia del paciente puede modelarse en función de la edad y de las distintas etapas en las que se presenta el cáncer, en este caso *etapa I*, *etapa II*, *etapa III* y *etapa IV* de manera que podemos representar estas covariables como:

$X_1 =$ edad del paciente.

$X_2 = 1$ si el paciente está en la etapa II, 0 en otro caso.

$X_3 = 1$ si el paciente está en la etapa III, 0 en otro caso.

$X_4 = 1$ si el paciente está en la etapa IV, 0 en otro caso.

De modo que el riesgo al tiempo t para cualquier individuo puede representarse como

$$h(t \mid \boldsymbol{X}) = h_0(t)\, exp\,(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$$

La manera en la que se ha construido este modelo coloca a los pacientes con cáncer en etapa I en el grupo de referencia, es decir, los riesgos estimados de muerte para

los pacientes diagnosticados con cáncer en las etapas II, III y IV relativos a los diagnosticados con cáncer en la etapa I son $exp(\hat{\beta}_2)$, $exp(\hat{\beta}_3)$ y $exp(\hat{\beta}_3)$ respectivamente.

Los parámetros β's e información adicional relacionada con el experimento puede obtenerse a partir de los siguientes códigos.

```
library(KMsurv)
data(larynx)
larynx

# La fórmula general para ajustar el modelo clásico de Cox en R es:
# coxph(Surv(tiempo, delta)) ~ X_1 + ... + X_p, data=nombre, method="nombre")
# en el argumento data se escribe el nombre de la base de datos
# mientras que en el argumento method puede escribirse: breslow, efron o exact.

# Particularmente, en este ejemplo, algunas de las covariables del modelo
# provienen de la misma columna por lo que se debe comunicar a R de esta
# situación con la siguiente línea de código:

larynx$stage <- factor(larynx$stage)

# La función que representa la fórmula de Cox para este modelo es:

vesper_0 <- coxph(Surv(time,delta) ~ age+stage,data=larynx,method="breslow")
vesper_0
```

Dando como resultado la siguiente salida:

```
          coef    exp(coef)   se(coef)    z       p
age      0.0189     1.0191     0.0143   1.33   0.185
stage2   0.1386     1.1486     0.4623   0.30   0.764
stage3   0.6383     1.8934     0.3561   1.79   0.073
stage4   1.6931     5.4361     0.4222   4.01   6.1e-05

Likelihood ratio test=18.1  on 4 df, p=0.0012
n= 90, number of events= 50
```

Los parámetros estimados correspondientes a las covariables del modelo son;

$$\hat{\beta}_1 = 0.0189 \quad \hat{\beta}_2 = 0.1386 \quad \hat{\beta}_3 = 0.6383 \quad \hat{\beta}_4 = 1.6931$$

Uno de los muchos resultados que podemos inferir, es que el riesgo relativo de un individuo (en la etapa IV del cáncer) de 60 años de edad comparado con el de un individuo (en la misma etapa de cáncer) de 40 años es; $exp[(60-20)\hat{\beta}_1] = 1.45$, equivalente a decir, que la probabilidad que tienen de morir un individuo de 60 años es aproximadamente 1.45 veces más elevada que la probabilidad de morir de un individuo de 40 años.

Con el objetivo de comparar el riesgo al cual están sometidos dos o más grupos de individuos, los resultados que se obtienen del estimador puntual de la razón de riesgos

$$\frac{\widehat{h_i(t \mid \boldsymbol{X_i})}}{h_j(t \mid \boldsymbol{X_j})}$$

se interpretan conforme a los *tipos de datos* (continuos, categóricos, etc. . .) que están involucrados en el modelo de regresión.

Debido a la variedad de modelos de regresión que involucran estos tipos de datos se ha decidido incluir en esta sección, en forma de resumen, los modelos más comunes de los modelos de Cox así como la interpretación de los resultados obtenidos para el estimador puntual de la razón de riesgos.

Modelo con una variable continua

$$h(t \mid x) = h_0(t)\, exp(\beta\, x) \qquad x,\, \beta \,\in\, \mathbb{R}$$

En este modelo podemos interpretar a β como un valor que representa el aumento de riesgo por unidad de incremento x, es decir, un valor de β grande implica un riesgo más elevado.

Modelo con dos o más covariables

$$h(t \mid \boldsymbol{x}) = h_0(t)\, exp(\beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_p\, x_p) \quad (x_1,\, x_2, \ldots,\, x_p);\, (\beta_1,\, \beta_2, \ldots,\, \beta_p) \,\in\, \mathbb{R}^p$$

Bajo este modelo interpretamos a β_1 como el aumento de riesgo por la diferencia del valor de x_1 con relación a x_i $(i \neq 1)$, manteniendo los valores de los restantes x_j fijos $(j \neq 1,\, i)$. La razón de riesgos $\boldsymbol{r} = exp(\hat{\beta}_1(x_1 - x_i))$ se puede interpretar como; La probabilidad de falla de un individuo de covariable x_1 es r-veces más elevada (si $\boldsymbol{r}$ es mayor que la unidad) o más baja (si $\boldsymbol{r}$ es menor que la unidad) que la del individuo de covariable x_i. La interpretación de $\beta_2, \beta_3, \ldots, \beta_p$ es análoga a la de β_1.

Por otro lado, si se tiene un modelo de dos covariables expresado como:

$$h(t \mid \boldsymbol{x}) = h_0(t)\, exp(\beta_1\, x_1 + \beta_2\, x_2 + \beta_{12}\, x_1\, x_2) \quad (x_1,\, x_2) \,\in\, \mathbb{R}^2, \quad (\beta_1,\, \beta_2,\, \beta_{12}) \,\in\, \mathbb{R}^3$$

obsérvese que en el modelo anterior, los valores x_1 y x_2 se encuentran interactuando entre si, esta interacción puede generar dificultades al momento de interpretar los parámetros β's involucrados en la descripción de la falla. En esta interacción se considera que hay un efecto distinto en x_1 sobre el riesgo de fallar a consecuencia de x_2.

Modelos con variables categóricas. El caso más sencillo de este tipo de modelos es aquel que consta de una sola variable dicotómica, es decir, una variable que toma solo dos valores, generalmente uno y cero.

$$h(t \mid x) = h_0(t)\, exp(\beta\, x), \qquad x \,\in\, \{0, 1\};\, \beta \,\in\, \mathbb{R}$$

Sustituyendo los valores $x = 0$ y $x = 1$ en el modelo, se tiene que:

$$h(t \mid 0) = h_0(t) \qquad h(t \mid 1) = h_0(t)\, exp(\beta)$$

Por lo tanto, la razón de riesgos para este modelo es constante e igual a $exp(\beta)$.

En el caso general, es decir, cuando x pertenece a una de las p categorías, el modelo se expresa como:

$$h(t \mid \boldsymbol{x}) = h_0(t)\, exp(\beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_{p-1} c_{p-1}), \qquad x \in \{c_0, c_1, \ldots, c_{p-1}\}$$

Para interpretar los resultados obtenidos de este modelo conviene construirse una variable dicotómica $C_{i\,k} = 1$ si $x_i = c_k$ y *cero* en otro caso, donde $C_{i\,0} = 1 - \sum_{k=1}^{p-1} C_{i\,k}$, entonces, la función de riesgo para el p-ésimo grupo de x's está dada por:

$$\begin{aligned} h(t \mid c_0) &= h_0(t) \\ h(t \mid c_1) &= h_0(t)\, exp(\beta_1) \\ &\vdots \\ h(t \mid c_{p-1}) &= h_0(t)\, exp(\beta_{p-1}) \end{aligned}$$

Por lo tanto, la razón de riesgos para el grupo $c_k \neq 0$ relativa al grupo c_0 es $exp(\beta_k)$. Análogamente la razón de riesgos para los grupos $c_i \neq 0$ y $c_j \neq 0$ es igual al $exp(\beta_i - \beta_j)$.

Modelo con una variable categórica y una continua. Sean: $x_1 \in \{0, 1\}$ y $x_2 \in \mathbb{R}$ el modelo de riesgos proporcionales se expresa como:

$$h(t \mid \boldsymbol{x}) = h_0(t) exp(\beta_1\, x_1 + \beta_2\, x_2) \quad (\beta_1,\, \beta_2) \in \mathbb{R}^2$$

Sustituyendo el valor $x_1 = 0$ en el modelo, se tiene que:

$$h(t \mid x_1 = 0,\, x_2) = h_0(t)\, exp(\beta_2\, x_2)$$

$$h(t \mid x_1 = 1,\, x_2) = h_0(t)\, exp(\beta_1)\, exp(\beta_2\, x_2)$$

Entonces, la razón de riesgos para cuando $x_1 = 0$ es $exp(\beta_2)$ y para $x_1 = 1$ la razón de riesgos es $exp(\beta_1)\, exp(\beta_2\, x_2)$.

La manera en que los modelos anteriormente expuestos nos permiten comparar el riesgo entre dos o más grupos de individuos es esencial en la mayoría de los estudios clínicos, por este y otros motivos los modelos semiparamétricos son los más utilizados en el análisis de regresión de registros de supervivencia. No obstante, existen modelos más precisos y que nos permiten hacer predicciones con relación al tiempo de falla.

2.2.2. Modelos Paramétricos de Regresión

En la seccción anterior, se expuso como el modelo de riesgos proporcionales de Cox resulta útil cuando se desea comparar el riesgo al que están sometidos dos o más grupos *distintos* de individuos, sin embargo, debido a que el modelo no cuenta con una especificación paramétrica de la función de riesgo inicial $h_0(t)$, éste presenta las siguientes desventajas.

- El cálculo de la función de riesgo $h(t\,|\boldsymbol{X})$, requiere estimar la función de riesgo inicial $h_0(t)$.
- Los parámetros son altamente sensibles ante cualquier error de especificación en el modelo.
- No facilitan la predicción debido a que el modelo no supone una distribución teórica para la función $h_0(t)$.
- Los parámetros del modelo son estimados a partir de una verosimilitud parcial, es decir, no considera la influencia de los paramétros involucrados en la función de riesgo inicial.

El enfoque paramétrico de riesgos proporcionales no presenta las desventajas anteriormente mencionadas, sin embargo, la estructura del modelo requiere de una revisión más detallada. Bajo el supuesto de riesgos proporcionales los modelos paramétricos de regresión tienen la siguiente estructura para su función de riesgo.

$$h(t\,|\,\boldsymbol{X}) = h(t\,|\,\boldsymbol{X};\,\boldsymbol{\alpha},\,\boldsymbol{\beta}) = h_0(\,t;\boldsymbol{\alpha})\,\psi(\boldsymbol{X};\,\boldsymbol{\beta}) \tag{2.14}$$

- $\psi(\boldsymbol{X};\,\boldsymbol{\beta})$ es una función no negativa que representa a la función de riesgo relativo y al igual que en el enfoque semiparamétrico de los modelos de riesgos proporcionales, se supone que ésta tiene una forma paramétrica conocida en el modelo.
- La función de riesgo inicial $h_0(t\,|\,\boldsymbol{\alpha})$ se encuentra especificada por una distribución teórica de parámetros $\alpha_1,\,\alpha_2,\,\ldots,\,\alpha_m$, evidentemente el tamaño del vector de parámetros $\boldsymbol{\alpha}$ quedará determinado por la cantidad de parámetros asociados a la distribución especificada.
- Dado que en $h_0(t\,|\,\boldsymbol{\alpha})$ y $\psi(\boldsymbol{X};\,\boldsymbol{\beta})$ se supone una forma paramétrica, decimos que el modelo de regresión se identifica como un modelo paramétrico.

Obsérvese que a diferencia de los modelos semiparamétricos, la función de riesgo inicial en (2.14) si especifica el vector de parámetros $\boldsymbol{\alpha}$ que aparece en la descripción del modelo de regresión, por lo tanto, los modelos paramétricos de regresión requieren estimar tanto a los parámetros β's como a los parámetros α's.

Los paramétros α's y β's desconocidos en el modelo son estimados utilizando el método de máxima verosimilitud, los modelos paramétricos permiten estimar todos los parámetros involucrados en la descripción de la falla, lo que los hace ser más precisos que los modelos semiparamétricos al momento de inferir algún resultado con relación a la supervivencia de uno o varios individuos.

Usualmente los modelos paramétricos de regresión se presentan en términos de la función de supervivencia dado el vector $\boldsymbol{X}$ la cual escribiremos de la siguiente manera:

$$S(t \mid \boldsymbol{X}) = S(t \mid \boldsymbol{X};\, \boldsymbol{\alpha},\, \boldsymbol{\beta}) = [\, S_0(t;\, \boldsymbol{\alpha})]^{\psi(\boldsymbol{X};\, \boldsymbol{\beta})}$$

donde, análogo a la función de riesgo (2.14) se define:

- $S(t \mid \boldsymbol{X})$: la función de supervivencia en el tiempo t para los individuos con vector de covariables $\boldsymbol{X}$ asociada a los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ correspondientes a las funciones de riesgo inicial y riesgo relativo respectivamente.

- $S_0(t\,; \boldsymbol{\alpha})$; la función de supervivencia inicial en el tiempo t, la cual sigue una distribución teórica de parámetros $\alpha_1,\, \alpha_2, \ldots,\, \alpha_m$.

Asimismo, presentamos las expresiones para las funciones de densidad y de riesgo acumulado dado el vector $\boldsymbol{X}$, pues en ocasiones conviene representar a la función de verosimilitud en términos de alguna de estas funciones relacionadas con la supervivencia.

$$f(t \mid \boldsymbol{X}) \;=\; S(t \mid \boldsymbol{X})\, h(t \mid \boldsymbol{X})$$

$$H(t \mid \boldsymbol{X}) \;=\; -log(S(t \mid \boldsymbol{X}))$$

En los modelos paramétricos de regresión la función de verosimilitud se construye de la misma manera que la verosimilitud expuesta en la *sección* 1.5 del capítulo anterior, con la diferencia de que en esta verosimilitud, las funciones de riesgo y supervivencia están condicionadas por el vector de covariables $\boldsymbol{X}$ y los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ involucrados en el modelo, por lo que en ocasiones se dice que es una función de verosimilitud multivariada.

Función de verosimilitud

Supóngase que la función de supervivencia $S(t \mid \boldsymbol{X})$ tiene una forma paramétrica $S(\cdot \mid \cdot\,; \boldsymbol{\alpha},\, \boldsymbol{\beta})$, donde $\boldsymbol{\alpha'} = (\alpha_1,\, \alpha_2, \ldots,\, \alpha_m)$ representa los $\boldsymbol{m}$ parámetros de la distribución y $\boldsymbol{\beta'} = (\beta_1,\, \beta_2, \ldots,\, \beta_p)$ la influencia de las $\boldsymbol{p}$ covariables que intervienen en el modelo, sea $\boldsymbol{m + p = k}$ entonces la función de verosimilitud en los modelos paramétricos de regresión está dada por:

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} h(t_i \mid \boldsymbol{X_i}\,; \boldsymbol{\vartheta})^{\delta_i}\, S(t_i \mid \boldsymbol{X_i}\,; \boldsymbol{\vartheta}) \qquad \boldsymbol{\vartheta'} = (\boldsymbol{\alpha'},\, \boldsymbol{\beta'}) = (\vartheta_1,\, \vartheta_2, \ldots, \vartheta_k) \quad (2.15)$$

El método de máxima verosimilitud consiste en obtener el vector $\boldsymbol{\vartheta}$ donde la función de verosimilitud alcanza el máximo o supremo, esto con el propósito de obtener los parámetros que maximicen la verosimilitud de observar un resultado en particular.

Análogo a la función de verosimilitud expuesta en la *sección* 1.5, el máximo de la función (2.15) se obtiene de manera más sencilla al maximizar la función *log-verosimilitud* dada por:

$$\begin{aligned} log\,(L(\boldsymbol{\vartheta})) &= \sum_{i=1}^{n} log\Big(\, h(t_i \,|\, \boldsymbol{X_i}\,;\, \boldsymbol{\vartheta})^{\delta_i}\, S(t_i \,|\, \boldsymbol{X_i}\,;\, \boldsymbol{\vartheta})\Big) \\ &= \sum_{i=1}^{n} \Big(log\,(\,h(t_i \,|\, \boldsymbol{X_i}\,;\, \boldsymbol{\vartheta})\,)^{\delta_i} + log\,(S(t_i \,|\, \boldsymbol{X_i}\,;\, \boldsymbol{\vartheta})\,)\Big) \\ &= \sum_{i=1}^{n} \Big(\delta_i\, log\,(\,h(t_i \,|\, \boldsymbol{X_i}\,;\, \boldsymbol{\vartheta})\,) - H(t_i \,|\, \boldsymbol{X_i}\,;\, \boldsymbol{\vartheta})\Big) \end{aligned} \tag{2.16}$$

Obsérvese que la expresión (2.16) queda en términos de las funciones de riesgo y riesgo acumulado.

Supóngase que $L(\cdot)$ es dos veces diferenciable, entonces el estimador máximo verosímil $\widehat{\boldsymbol{\vartheta}}$ es:

$$\widehat{\boldsymbol{\vartheta}} = argmax_{\vartheta}\, L(\boldsymbol{\vartheta})$$

el cual es proporcional a:

$$\widehat{\boldsymbol{\vartheta}} = argmax_{\vartheta}\, log\,(L(\boldsymbol{\vartheta})\,) \tag{2.17}$$

Generalmente, la expresión (2.17) es la solución de la ecuación $U(\boldsymbol{\vartheta}) = \mathbf{0}$, en la literatura estadística $U(\boldsymbol{\vartheta}) = (U_1(\boldsymbol{\vartheta}),\, U_2(\boldsymbol{\vartheta}), \ldots,\, U_k(\boldsymbol{\vartheta}))'$ es conocido como el vector de puntajes y es una función asociada a la verosimilitud definida como[6]:

$$U_r(\boldsymbol{\vartheta}) = \frac{\partial\, log(L(\boldsymbol{\vartheta}))}{\partial \vartheta_r} \qquad r = 1, \ldots,\, m + p = k$$

es decir, $U_r(\boldsymbol{\vartheta})$ representa la r-ésima componente del vector $U(\boldsymbol{\vartheta})$.

Por lo tanto, al resolver el siguiente sistema de ecuaciones de verosimilitud en $\vartheta_1,\, \vartheta_2, \ldots,\, \vartheta_k$.

$$\begin{aligned} \frac{\partial}{\partial \vartheta_1}\, log(L(\boldsymbol{\vartheta})) &= 0 \\ \frac{\partial}{\partial \vartheta_2}\, log(L(\boldsymbol{\vartheta})) &= 0 \\ &\vdots \\ \frac{\partial}{\partial \vartheta_k}\, log(L(\boldsymbol{\vartheta})) &= 0 \end{aligned}$$

se obtiene como resultado los estimadores máximo verosímiles $\hat{\vartheta}_1,\, \hat{\vartheta}_2, \ldots,\, \hat{\vartheta}_k$ de los correspondientes parámetros $\vartheta_1,\, \vartheta_2, \ldots,\, \vartheta_k$.

[6]Obsérvese que el **cero** que aparece como solución del vector $U(\boldsymbol{\vartheta})$ es un vector de ceros, es decir, $\mathbf{0} \in \mathbb{R}^k$, donde k representa el tamaño del vector $\boldsymbol{\vartheta}$.

Propiedades asintóticas del estimador máximo verosímil

En particular el vector de puntajes $U(\boldsymbol{\vartheta})$ satisface las siguientes propiedades estadísticas: Sea $\boldsymbol{\vartheta}$ el verdadero valor del parámetro tal que $U(\boldsymbol{\vartheta}) = \mathbf{0}$ entonces el vector de puntajes tiene media ***cero***

$$E\,[\,U(\boldsymbol{\vartheta})\,] = \mathbf{0} \tag{2.18}$$

y *matriz de varianzas y covarianzas* dada por:

$$\boldsymbol{V}\,(\,U(\boldsymbol{\vartheta})\,) = E\,[\,(\,U(\boldsymbol{\vartheta})\,)\,(\,U(\boldsymbol{\vartheta})\,)'\,] = \boldsymbol{\Sigma}\,(\boldsymbol{\vartheta}) \tag{2.19}$$

en donde bajo *condiciones de regularidad*[7] la matriz de varianzas y covarianzas puede obtenerse como menos la esperanza de las segundas derivadas de la *log-verosimilitud*:

$$\boldsymbol{\Sigma}\,(\boldsymbol{\vartheta}) = -E\left[\frac{\partial^2\, log(\,L(\boldsymbol{\vartheta})\,)}{\partial\boldsymbol{\vartheta}\,\partial\boldsymbol{\vartheta'}}\right]_{k\times k} \tag{2.20}$$

En particular $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ es una matriz de $k \times k$, debido a que $U(\boldsymbol{\vartheta})$ y $U(\boldsymbol{\vartheta})'$ son vectores de $k \times 1$ y $1 \times k$ respectivamente. La matriz $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ se conoce como la ***matriz de información de Fisher*** y se caracteriza por ser una matriz definida positiva, esto último significa que para cualquier vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k) \in R^k$ se verifica la desigualdad $\langle \boldsymbol{\Sigma}(\boldsymbol{\vartheta})\,\mathbf{Y},\ \mathbf{Y}\rangle \geq 0$, en done $\langle\cdot,\cdot\rangle$ denota el producto interior de $\mathbb{R}^k$.

Nuevamente, bajo condiciones de regularidad se puede verificar que el vector de puntajes $U(\boldsymbol{\vartheta})$ tiene una distribución normal k-variada con esperanza (2.18) y matriz de varianzas y covarianzas (2.19).

$$U(\boldsymbol{\vartheta}) \sim N_k(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\vartheta})) \tag{2.21}$$

[7]Las condiciones de regularidad de una función de densidad $f(x;\vartheta)$ son:

- $\dfrac{\partial}{\partial\vartheta}\, log\, f(x;\vartheta)$ existe $\forall\, x$ y $\forall\,\vartheta$
- $\dfrac{\partial}{\partial\vartheta}\displaystyle\int\cdots\int\prod_{i=1}^{n} f(x_i;\,\vartheta)\,dx_1\cdots dx_n = \int\cdots\int\frac{\partial}{\partial\vartheta}\prod_{i=1}^{n} f(x_i;\,\vartheta)\,dx_1\cdots dx_n$
- $\dfrac{\partial}{\partial\vartheta}\displaystyle\int\cdots\int t(x_1,..,x_n)\prod_{i=1}^{n} f(x_i;\vartheta)\,dx_1\cdot\cdot dx_n = \int\cdots\int t(x_1,..,x_n)\,\frac{\partial}{\partial\vartheta}\prod_{i=1}^{n} f(x_i;\vartheta)\,dx_1\cdot\cdot dx_n$
- $0 < E\left\{\left(\dfrac{\partial}{\partial\vartheta}\, log\, f(X;\vartheta)\right)^2\right\} < \infty \quad \forall\,\vartheta$
- La función $f(x;\vartheta)$ es tres veces diferenciable como función de ϑ, más aún, para toda ϑ existe una constante c y una función $G(x)$ tal que:

  $$\left|\frac{\partial^3 log\, f(x;\,\vartheta)}{\partial\vartheta^3}\right| \leq G(x)$$

  con $E_{\vartheta_0}[G(X)] < \infty$ para toda $|\vartheta - \vartheta_0| < c$ y para toda x.

  Esta última condición permite concluir, que a partir del tercer sumando, la serie de Taylor de segundo orden alrededor de ϑ_0 para $f(x;\vartheta)$, está acotada en probabilidad y por lo tanto no presenta ningún problema, asintóticamente hablando, truncar la serie desde el tercer término.

Al igual que en los modelos estadísticos para variables aleatorias independientes e idénticamente distribuidas se puede demostrar que, en condiciones de regularidad, el estimador $\widehat{\boldsymbol{\vartheta}}$ es asintóticamente normal, es decir:

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}\right)\xrightarrow{d} N_k(0,\,(n\boldsymbol{\Sigma}(\boldsymbol{\vartheta})\,)^{-1}) \tag{2.22}$$

la expresión $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})^{-1}$ denota la matriz inversa de la matriz definida en (2.20). La demostración de la afirmación (2.22) se basa en los siguientes pasos[8].

Demostración. *Paso 1:* Supóngase que $\widehat{\boldsymbol{\vartheta}}$ es el estimador máximo verosímil de $\boldsymbol{\vartheta}$, esto implica que:

$$U(\widehat{\boldsymbol{\vartheta}})=\frac{\partial\, log(L(\boldsymbol{\vartheta}))}{\partial\boldsymbol{\vartheta}}=\sum_{i=1}^{n}\frac{\partial\, log(\,L_i(\boldsymbol{\vartheta},T_i,\boldsymbol{\Delta_i},\boldsymbol{X_i})\,)}{\partial\boldsymbol{\vartheta}}=\mathbf{0} \tag{2.23}$$

debido a que $\hat{\boldsymbol{\vartheta}}$ es la solución de la ecuación

$$\frac{\partial\, log(L(\boldsymbol{\vartheta}))}{\partial\boldsymbol{\vartheta}}=\mathbf{0}$$

Nota: Recuérdese que $U(\boldsymbol{\vartheta})$ es un vector de dimensión $k\times 1$, por lo que la expresiones $\frac{\partial}{\partial\boldsymbol{\vartheta}}\,log(L(\boldsymbol{\vartheta}))$ y la suma que aparecen en (2.23) también son vectores con la misma dimensión, particularmente, esta suma es la expresión desarrollada de la derivada con respecto a $\boldsymbol{\vartheta}$ de la función *log-verosimilitud*.

Suponiendo que la función de verosimilitud $L(\boldsymbol{\vartheta})$ es tres veces diferenciable, es posible expresar al vector de puntajes $U(\hat{\boldsymbol{\vartheta}})$ como una serie de Taylor expandida alrededor de $\hat{\boldsymbol{\vartheta}}=\boldsymbol{\vartheta}$

$$\frac{\partial\, log(L(\widehat{\boldsymbol{\vartheta}}))}{\partial\boldsymbol{\vartheta}}=\frac{\partial\, log(L(\boldsymbol{\vartheta}))}{\partial\boldsymbol{\vartheta}}+\frac{\partial^2\, log(L(\boldsymbol{\vartheta}))}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta'}}(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})+\cdots=\mathbf{0}$$

o en términos del vector de puntajes (truncando la serie desde el tercer sumando) como[9]

$$\mathbf{0}=U(\widehat{\boldsymbol{\vartheta}})\doteq U(\boldsymbol{\vartheta})-M(\boldsymbol{\vartheta})(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}) \tag{2.24}$$

donde $M(\boldsymbol{\vartheta})$ es la matriz $k\times k$ con elementos

$$M_{sr}(\boldsymbol{\vartheta})=-\frac{\partial^2 log(\,L(\boldsymbol{\vartheta})\,)}{\partial\vartheta_s\,\partial\vartheta_r},\quad s,r=1,2,\ldots,k.$$

Resolviedo la ecuación (2.24) para $(\hat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})$ y multiplicando ambos lados por $\sqrt{n}$

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}\right)\doteq\sqrt{n}\,M(\boldsymbol{\vartheta})^{-1}U(\boldsymbol{\vartheta})=\left(\frac{M(\boldsymbol{\vartheta})}{n}\right)^{-1}\frac{U(\boldsymbol{\vartheta})}{\sqrt{n}} \tag{2.25}$$

[8]Parte de esta demostración se puede consultar en [2] pag. 11

[9]El símbolo $\doteq$ significa aproximadamente igual.

Paso 2: Aplicando el teorema del límite central para vectores aleatorios independientes pero no necesariamente idénticamente distribuidos a

$$U(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \frac{\partial\, log(\, L_i(\boldsymbol{\vartheta}, T_i, \boldsymbol{\Delta_i}, \boldsymbol{X_i})\,)}{\partial \boldsymbol{\vartheta}}$$

se obtiene que:

$$\frac{1}{n}\sum_{i=1}^{n} Var\left(\frac{\partial\, log(\, L_i(\boldsymbol{\vartheta}, T_i, \boldsymbol{\Delta_i}, \boldsymbol{X_i})\,)}{\partial \boldsymbol{\vartheta}}\Big|\, \boldsymbol{X_i}\right) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}_i(\boldsymbol{\vartheta} \mid \boldsymbol{X_i}) \to \boldsymbol{\Sigma}^*(\boldsymbol{\vartheta})$$

El resultado anterior[10] implica que:

$$\frac{U(\boldsymbol{\vartheta})}{\sqrt{n}} \xrightarrow{d} N_k(\mathbf{0}, \boldsymbol{\Sigma}^*(\boldsymbol{\vartheta})) \tag{2.26}$$

donde $\boldsymbol{\Sigma}^*(\boldsymbol{\vartheta})$ es una matriz definida positiva que satisface $\boldsymbol{\Sigma}(\boldsymbol{\vartheta}) \stackrel{d}{=} n\boldsymbol{\Sigma}^*(\boldsymbol{\vartheta})$. Obsérvese que el resultado obtenido en (2.26) demuestra la afirmación (2.21).

Paso 3: Por la ley débil de los grandes números (*véase* la Figura 2.3), para vectores aleatorios independientes, pero no necesariamente idénticamente distribuidos se tiene que:

$$\left(\frac{M(\boldsymbol{\vartheta})}{n} - \frac{E\,[M(\boldsymbol{\vartheta})]}{n}\right) \xrightarrow{p} 0$$

entonces, bajo condiciones usuales de regularidad

$$\frac{E\,[M(\boldsymbol{\vartheta})]}{n} = \sum_{i=1}^{n} \boldsymbol{V}_i(\boldsymbol{\vartheta},\ \boldsymbol{X_i})$$

y por lo tanto

$$\frac{M(\boldsymbol{\vartheta})}{n} \xrightarrow{p} \boldsymbol{\Sigma}^*(\boldsymbol{\vartheta}) \tag{2.27}$$

Paso 4: Partiendo de la expresión (2.25) y utilizando los resultados obtenidos en (2.26) y (2.27) se tiene que:

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\right) \doteq \sqrt{n}\, M(\boldsymbol{\vartheta})^{-1} U(\boldsymbol{\vartheta}) \to (\boldsymbol{\Sigma}^*(\boldsymbol{\vartheta}))^{-1}\, N_k(\mathbf{0}, \boldsymbol{\Sigma}^*(\boldsymbol{\vartheta})) \to N_k\,(\,\mathbf{0},\, (\boldsymbol{\Sigma}^*(\boldsymbol{\vartheta}))^{-1}\,)$$

por lo tanto

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\right) \xrightarrow{d} N_k(\,\mathbf{0},\, (\,\boldsymbol{\Sigma}^*(\boldsymbol{\vartheta})\,)^{-1})\,) \qquad \text{y} \qquad \sqrt{n}\left(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\right) \xrightarrow{d} N_k(\,\mathbf{0},\, (\,n\boldsymbol{\Sigma}(\boldsymbol{\vartheta})\,)^{-1})\,)$$

con lo que se demuestra la afirmación (2.22), es decir, el estimador $\widehat{\boldsymbol{\vartheta}}$ es asintóticamente normal. □

[10] La expresión

$$\left(\frac{\partial\, log(\, L_i(\boldsymbol{\vartheta}, T_i, \boldsymbol{\Delta_i}, \boldsymbol{X_i})\,)}{\partial \boldsymbol{\vartheta}}\Big|\, \boldsymbol{X_i}\right)$$

se refiere al vector de puntajes para un individuo perteneciente a la muestra.

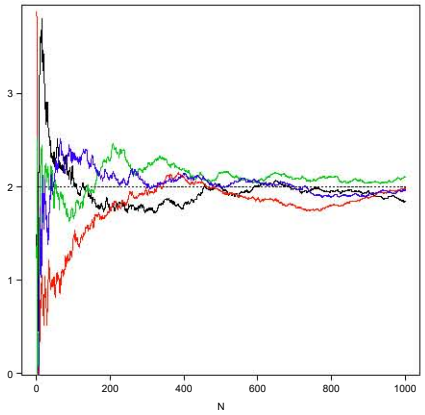


Figura 2.3: (Ley de los grandes números) Comportamiento del cociente S_N/N de cuatro funciones con distribución normal con media dos y varianza cuatro. Donde $S_N = X_1 + X_2 + \cdots + X_N$ y X_i una variable aleatoria tal que $X_i \sim N(2, 4)$ para toda $i = 1, 2, \ldots, N$. Gráficamente puede observarse como el cociente S_N/N converge a la media en valores grandes de N.

Estrictamente hablando se tiene que[11]

$$\widehat{\boldsymbol{\vartheta}} \dot{\sim} N_k(\boldsymbol{\vartheta},\, (\,\boldsymbol{\Sigma}(\boldsymbol{\vartheta})\,)^{-1}) \tag{2.28}$$

debido a que el cálculo de la esperanza en $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ puede llegar a ser muy complicado, es conveniente utilizar un estimador consistente de $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$. Este estimador es conocido como la información observada y lo denotaremos por $\mathbf{I}(\widehat{\boldsymbol{\vartheta}})$, donde el (i, j)-ésimo elemento de esta información está dado por:

$$\mathbf{I}_{i,j}(\boldsymbol{\vartheta}) = -\frac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_i\,\partial\vartheta_j} \quad i, j = 1, 2, \ldots, k$$

Usualmente el estimador anterior suele escribirse en su expresión matricial, la cual está dada por:

$$\mathbf{I}\,(\boldsymbol{\vartheta}) = -\begin{pmatrix} \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_1^2} & \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_1\,\partial\vartheta_2} & \cdots & \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_1\,\partial\vartheta_k} \\ \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_2\,\partial\vartheta_1} & \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_2^2} & \cdots & \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_2\,\partial\vartheta_k} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_k\,\partial\vartheta_1} & \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_k\,\partial\vartheta_2} & \cdots & \dfrac{\partial^2 log\,(L(\boldsymbol{\vartheta}))}{\partial\vartheta_k^k} \end{pmatrix}_{k\times k}$$

[11] El símbolo $\dot{\sim}$ significa "se distribuye aproximadamente como".

La matriz de información $\mathbf{I}(\widehat{\boldsymbol{\vartheta}})$ tiene la ventaja de ser computacionalmente más sencilla de evaluar que la matriz $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ y debido a que $\mathbf{I}(\boldsymbol{\vartheta})$ es un estimador consistente de $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$, entonces, se puede remplazar a la matriz de información de Fisher $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ en (2.28) por $\mathbf{I}(\boldsymbol{\vartheta})$ y por lo tanto se tiene que:

$$\widehat{\boldsymbol{\vartheta}} \dot{\sim} N_k(\boldsymbol{\vartheta},\, (\,\mathbf{I}(\boldsymbol{\vartheta})\,)^{-1}) \tag{2.29}$$

En particular, algunos de los elementos de la inversa de la matriz de información $(\,\mathbf{I}(\widehat{\boldsymbol{\vartheta}})\,)^{-1})$ se utilizan para calcular intervalos de confianza para $\boldsymbol{\vartheta'}$, los elementos de esta matriz pueden escribirse como:

$$(\,\mathbf{I}(\widehat{\boldsymbol{\vartheta}})\,)^{-1} = \begin{pmatrix} \mathfrak{S}_{11} & \mathfrak{S}_{12} & \dots & \mathfrak{S}_{1k} \\ \mathfrak{S}_{21} & \mathfrak{S}_{22} & \dots & \mathfrak{S}_{2k} \\ \vdots & \vdots & & \vdots \\ \mathfrak{S}_{k1} & \mathfrak{S}_{k2} & \dots & \mathfrak{S}_{kk} \end{pmatrix}_{k\times k}$$

Errores estándar e intervalos de confianza

Los errores estándar de $\widehat{\boldsymbol{\vartheta}}' = (\hat{\vartheta}_1,\, \hat{\vartheta}_2, \dots,\, \hat{\vartheta}_k)$ y los intervalos de confianza para $\boldsymbol{\vartheta}' = (\vartheta_1,\, \vartheta_2, \dots,\, \vartheta_k)$ se basan en la aproximación normal.

Los errores estándar de $\hat{\vartheta}_r$, $r = 1, 2, \dots, k$ se obtienen de la diagonal de la matriz $(\,\mathbf{I}(\widehat{\boldsymbol{\vartheta}})\,)^{-1}$ y se denotan por $se(\hat{\vartheta}_r)$, los cuales corresponden a:

$$se(\hat{\vartheta}_r) = \sqrt{\mathfrak{S}_{rr}} \quad r = 1,\, 2, \dots,\, k. \tag{2.30}$$

El resultado (2.29), permite construir un intevalo del $(1-\alpha)\times 100\,\%$ de confinza para ϑ_r, este intervalo está dado por:

$$\left(\,\hat{\vartheta}_r - z_{1-\alpha/2}\,\sqrt{\mathfrak{S}_{rr}}\,,\, \hat{\vartheta}_r + z_{1-\alpha/2}\,\sqrt{\mathfrak{S}_{rr}}\,\right)$$

donde $z_{1-\alpha/2}$ es el cuantil $(1-\alpha/2)$ de una distribución normal estándar. Habitualmente el intervalo anterior se escribe en términos del error estándar de $\hat{\vartheta}_r$ como se expresa a continuación:

$$\left(\,\hat{\vartheta}_i - z_{1-\alpha/2}\,s.e\,(\hat{\vartheta}_i),\, \hat{\vartheta}_i + z_{1-\alpha/2}\,s.e\,(\hat{\vartheta}_i)\right) \tag{2.31}$$

Asimismo, utilizando las propiedades de la varianza para vectores aleatorios, es posible construir un intervalo de confianza para $(\vartheta_i - \vartheta_j)$, con $i, j = 1, 2, \dots, r.\ i \neq j$ del siguiente modo:

$$(\,\hat{\vartheta}_i - \hat{\vartheta}_j\,) \pm z_{1-\alpha/2}\sqrt{\mathfrak{S}_{ii} + \mathfrak{S}_{jj} - 2\,\mathfrak{S}_{ij}}$$

Contrastes de hipótesis

Una vez que se ha construido el estimador $\widehat{\boldsymbol{\vartheta}}$ y se han obtenido los intervalos de confianza para $\boldsymbol{\vartheta}$, resulta interesante preguntarse cuál de las covariables en el modelo tiene un efecto significativo en la descripción del tiempo de falla.

Hay tres pruebas que se utilizan comúnmente para probar la hipótesis de que una covariable tiene un efecto significativo en el modelo. Estas pruebas son:

- La prueba de razón de verosimilitudes.
- La prueba de Wald.
- La prueba de puntajes.

Estas pruebas son asintóticamente equivalentes, es decir, en la mayoría de los casos basta con utilizar cualquiera de estas pruebas para probar o rechazar la hipótesis.

1.- Prueba de razón de verosimilitudes

Esta prueba se distingue de las demás por presentar una mayor confiabilidad y como su nombre lo indica, esta prueba está basada en el criterio asintótico de la razón de verosimilitudes.

Hipótesis a probar

Se puede considerar que el vector de parámetros $\boldsymbol{\vartheta}$ tiene los valores $\boldsymbol{\vartheta_0}$ si no se rechaza la siguiente hipótesis nula

$$H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta_0} \quad \text{v.s} \quad H_a : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta_0}$$

para un parámetro en particular

$$H_0 : \vartheta_m = \vartheta_0 \quad \text{v.s} \quad H_a : \vartheta_m \neq \vartheta_0$$

o simplemente

$$H_0 : \boldsymbol{\vartheta} = \mathbf{0} \quad \text{v.s} \quad H_a : \boldsymbol{\vartheta} \neq \mathbf{0}$$

en este último caso, si no se rechaza la hipótesis nula, se puede considerar que el parámetro $\boldsymbol{\vartheta}$ no es significativo en el modelo, es decir, el parámetro no proporciona información significativa con respecto a la descripción de la falla.

Generalmente, está y el resto de las pruebas presentan las hipótesis

$$H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta_0} \quad \text{v.s} \quad H_a : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta_0}$$

por lo que comúnmente las estadísticas de prueba se presentan en la literatura estadística bajo las hipótesis anteriormente mencionadas, aunque es claro, que se puede realizar para cualquiera de las hipótesis mencionadas, sin embargo, por comodidad, de aquí en adelante, solo se presentará el caso general.

Estadística de prueba

En particular la estadística de prueba tiene una distrubución *Ji-cuadrada* con k grados de libertad χ^2_k

$$\boldsymbol{\Lambda} = 2\,log\,(\,L(\widehat{\boldsymbol{\vartheta}})\,) - 2\,log\,(\,L(\boldsymbol{\vartheta_0})\,)$$

donde $L(\boldsymbol{\vartheta_0})$ es la función de verosimilitud para el modelo reducido, mientras $L(\widehat{\boldsymbol{\vartheta}})$ es la función de verosimilitud que contiene toda la información de los parámetros estimados.

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$$\boldsymbol{\Lambda} > \chi^2_{k\,;\,1-\alpha}$$

donde $\chi^2_{k\,;\,1-\alpha}$ es el cuantil $(1-\alpha)$ de una distribución *Ji-cuadrada* con k grados de libertad, en particular el número k corresponde a la dimensión del vector $\boldsymbol{\vartheta}$. Por otro lado, si la estadística $\boldsymbol{\Lambda}$ se asocia con un ***p-value*** menor que α, la hipótesis nula sobre $\boldsymbol{\vartheta}$ debe ser rechazada. Este último criterio de rechazo es muy útil al momento llevar a cabo una prueba computacionalmente.

2.- Prueba de Wald

La prueba de Wald se utiliza para poner a prueba el verdadero valor del parámetro basado en la estimación de la muestra.

Hipótesis a probar

$$H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta_0} \quad \text{v.s} \quad H_a : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta_0}$$

Estadística de prueba

Al igual que la prueba anterior, esta estadística tiene una distrubución *Ji-cuadrada* con k grados de libertad χ^2_k para muestras grandes cuando H_0 es cierta

$$\boldsymbol{\mathcal{W}} = (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta_0})'\,\mathbf{I}\,(\widehat{\boldsymbol{\vartheta}})\,(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta_0})$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$$\boldsymbol{\mathcal{W}} > \chi^2_{k\,;\,1-\alpha}$$

donde $\chi^2_{k\,;\,1-\alpha}$ es el cuantil $(1-\alpha)$ de una distribución *Ji-cuadrada* con k grados de libertad. De acuerdo con el valor del ***p-value*** rechazamos la hipótesis nula H_0 si el valor de éste es relativamente pequeño comparado con α.

3.- Prueba de Puntajes

Este contraste de hipótesis utiliza el vector de puntajes y la matriz de información observada en la estadística de prueba.

Hipótesis a probar

$$H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta_0} \quad \text{v.s} \quad H_a : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta_0}$$

Estadística de prueba

Esta estadística también tiene una distrubución *Ji-cuadrada* con k grados de libertad χ^2_k para muestras grandes cuando H_0 es cierta y está dada por:

$$\boldsymbol{\mathcal{S}} = (\, U(\boldsymbol{\vartheta_0})\,)' \,(\, \mathbf{I}\,(\boldsymbol{\vartheta_0})\,)^{-1}\,(\, U(\boldsymbol{\vartheta_0})\,)$$

donde $U(\boldsymbol{\vartheta_0})$ es el vector de puntajes.

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$$\boldsymbol{\mathcal{S}} > \chi^2_{k\,;\,1-\alpha}$$

donde $\chi^2_{k\,;\,1-\alpha}$ es el cuantil $(1-\alpha)$ de una distribución *Ji-cuadrada* con k grados de libertad. Al igual que los otros dos contrastes podemos utilizar el criterio del ***p-value*** para determinar si se rechaza o no se rechaza la hipótesis nula.

Si se pretende llevar a cabo una prueba de hipótesis para un parámetro específico (este es un caso particular de la **prueba de Wald**) ϑ_r, $r = 1, \ldots, k$ es recomendable utilizar el intervalo de cofianza que aparece en la expresión (2.31). En este contraste se tiene la siguiente hipótesis y regla de decisión.

Hipótesis a probar

$$H_0 : \vartheta_r = \vartheta_{0r} \quad \text{v.s} \quad H_a : \vartheta_r \neq \vartheta_{0r}$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si:

$$\frac{|\,\widehat{\vartheta}_r - \vartheta_{0r}\,|}{s.e\,(\widehat{\vartheta}_r)} > z_{1-\alpha/2}$$

donde $z_{1-\alpha/2}$ es el cuantil $(1-\alpha/2)$ de una distribución *normal estándar*. En particular, la mayoría de los paquetes estadísticos llevan a cabo esta prueba de hipótesis bajo el siguiente supuesto $\vartheta_{0r} = 0$, $r = 1, \ldots, k$.

Los resultados obtenidos con anterioridad como los errores estándar, los intervalos de confianza y los contrastes de hipótesis, también son válidos para los modelos semiparamétricos de riesgos proporcionales y la construcción de estos resultados es análoga (con su respectiva función de verosimilitud) a la que se ha realizado.

Modelos paramétricos

La estructura del modelo de riesgos proporcionales (2.14) requiere que se especifique la función de riesgo inicial $h_0(t \mid \boldsymbol{\alpha})$, generalmente esta función se identifica con una distribución teórica. Uno de los criterios de decisión bajo los que se elige esta distribución es el conocimiento empírico del fenómeno de interés, sin embargo, la distribución también puede ser determinada por alguna prueba de bondad, aunque generalmente se utilizan métodos de diagnóstico gráficos. Cuando la distribución teórica es especificada, el modelo adopta el nombre de la distribución bajo la cual está diseñado.

Los modelos paramétricos de riesgos proporcionales más utilizados en el análisis de supervivencia son:

- Modelo de Regresión Exponencial.
- Modelo de Regresión Weibull.
- Modelo de Regresión Gompertz.

Estos modelos serán presentados con detalle en las siguientes secciones.

A continuación se presenta una tabla en la que se clasifican los modelos paramétricos de regresión de acuerdo con su distribución.

AFT	**PH**
Exponencial	Exponencial
Weibull	Weibull
Log-normal	Gompertz
Log-logístico	
Gamma generalizada	

2.3. Modelo de Riesgos Proporcionales Exponencial

La modelación, en términos de regresión estadística, es un proceso que consiste en desarrollar expresiones matemáticas que describen mediante una relación funcional el comportamiento de una variable aleatoria de interés. Por lo tanto, el objetivo de los modelos de regresión exponencial en el análisis de supervivencia es describir el tiempo de falla de un individuo mediante una distribución eponencial.

2.3.1. Modelo de Regresión Exponencial

En el capítulo anterior, se revisó que si una variable aleatoria T se distribuye exponencialmente con parámetro λ se tiene que:

- $f(t) = \lambda\, e^{-\lambda t}$
- $S(t) = e^{-\lambda t}$
- $h(t) = \lambda$
- $H(t) = \lambda t$

Decimos que el modelo de regresión

$$h(t \,|\, \boldsymbol{X}) = h_0(t;\, \boldsymbol{\alpha})\, \psi(\boldsymbol{X};\, \boldsymbol{\beta})$$

es exponencial, si se puede considerar, que el tiempo de falla T sigue una distribución exponencial con función de riesgo:

$$h(t \,|\, \boldsymbol{X}) = \lambda\, \psi(\boldsymbol{X};\, \boldsymbol{\beta})$$

En el modelo de regresión exponencial más comúnmente utilizado se considera que:

$$\psi(\boldsymbol{X};\, \boldsymbol{\beta}) = exp\,(\boldsymbol{\beta'X}) = exp\,(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p),$$

bajo este supuesto, el modelo de regresión exponencial se escribe como:

$$h(t \,|\, \boldsymbol{X}) = \lambda\, exp\,(\boldsymbol{\beta'X}) \tag{2.32}$$

A partir de la función anterior es posible obtener el modelo de regresión exponencial en términos de la función de supervivencia, de densidad y de riesgo acumulado dado el vector $\boldsymbol{X}$, correspondientes a $S(t \,|\, \boldsymbol{X})$, $f(t \,|\, \boldsymbol{X})$ y $H(t \,|\, \boldsymbol{X})$.

Función de supervivencia dado $\boldsymbol{X}$:

$$\begin{aligned} S(t \,|\, \boldsymbol{X}) &= exp\Big(-\int_0^t h(u \,|\, \boldsymbol{X})\, du \Big) \\ &= exp\Big(-\int_0^t \lambda\, exp\,(\boldsymbol{\beta'X})\, du \Big) \\ &= exp(-\lambda\, t\; exp\,(\boldsymbol{\beta'X})\,) \qquad (2.33) \\ &= (S_0(t))^{\psi(\boldsymbol{X};\boldsymbol{\beta})} \end{aligned}$$

Función de densidad dado $\boldsymbol{X}$:

$$f(t \mid \boldsymbol{X}) = h(t \mid \boldsymbol{X})\, S(t \mid \boldsymbol{X}) = \lambda\, exp(\boldsymbol{\beta'}\, \boldsymbol{X})\, exp(-\lambda\, t\ exp\,(\boldsymbol{\beta'} \boldsymbol{X})\,)$$

Función de riesgo acumulado dado $\boldsymbol{X}$:

$$H(t \mid \boldsymbol{X}) = -log(\, S(t \mid \boldsymbol{X})\,) = \lambda\, t\, exp(\boldsymbol{\beta'}\, \boldsymbol{X})$$

Función de verosimilitud:

Sea $\boldsymbol{\vartheta'} = (\lambda,\, \boldsymbol{\beta_1},\, \boldsymbol{\beta_2}, \ldots,\, \boldsymbol{\beta_p})$, T una variable aleatoria con distribución exponencial con parámetro λ, $\psi(\boldsymbol{X};\, \boldsymbol{\beta}) = exp\,(\boldsymbol{\beta'} \boldsymbol{X})$ y δ_i la i-ésima observación de la función indicadora de censura, entonces, la función de verosimilitud para el modelo de regresión exponencial está dada por:

$$\begin{aligned} L(\boldsymbol{\vartheta}) &= \prod_{i=1}^{n} h(t_i \mid \boldsymbol{X_i};\, \boldsymbol{\vartheta})^{\delta_i}\, S(t_i \mid \boldsymbol{X_i};\, \boldsymbol{\vartheta}) \\ L(\boldsymbol{\vartheta}) &= \prod_{i=1}^{n} \Big(\, \lambda\, exp(\boldsymbol{\beta'} \boldsymbol{X_i})\, \Big)^{\delta_i}\, exp(-\lambda\, t_i\, exp\,(\boldsymbol{\beta'} \boldsymbol{X_i})\,) \end{aligned}$$

Para encontrar el estimador máximo verosímil $\hat{\lambda}$ para λ hay que resolver

$$\frac{\partial}{\partial \lambda}\, log\,(\, L(\boldsymbol{\vartheta})\,) = 0$$

El procedimiento para encontrar el estimador máximo verosímil es el siguiente:

De la expresión (2.16) sa sabe que:

$$ln(\, L(\boldsymbol{\vartheta})\,) = \sum_{i=1}^{n} \Big(\delta_i\, log(\, h(t_i \mid \boldsymbol{X_i})\,) - H(t_i \mid \boldsymbol{X_i})\, \Big)$$

Sustituyendo los valores de las funciones de riesgo y riesgo acumulado dado el vector $\boldsymbol{X}$ se obtiene:

$$\begin{aligned} log(\, L(\boldsymbol{\vartheta})\,) &= \sum_{i=1}^{n} \Big(\delta_i\, log\,(\, \lambda\, exp\,(\boldsymbol{\beta'} \boldsymbol{X_i})\,) - \lambda\, t_i\, exp(\boldsymbol{\beta'}\, \boldsymbol{X_i})\, \Big) \\ &= \sum_{i=1}^{n} \delta_i\, log\,(\, \lambda\, exp\,(\boldsymbol{\beta'} \boldsymbol{X_i})\,) - \sum_{i=1}^{n} \lambda\, t_i\, exp(\boldsymbol{\beta'}\, \boldsymbol{X_i}) \\ &= \sum_{i=1}^{n} \delta_i\, log(\lambda) + \sum_{i=1}^{n} \delta_i\, \boldsymbol{\beta'}\, \boldsymbol{X_i} - \sum_{i=1}^{n} \lambda\, t_i\, exp(\boldsymbol{\beta'}\, \boldsymbol{X_i}) \qquad (2.34) \end{aligned}$$

derivando la expresión (2.34) con respecto a λ:

$$\frac{\partial}{\partial \lambda} log(\, L(\boldsymbol{\vartheta})\,) = \sum_{i=1}^{n} \frac{\delta_i}{\lambda} - \sum_{i=1}^{n} t_i\, exp(\boldsymbol{\beta'}\, \boldsymbol{X_i})$$

igualando a cero la derivada y despejando al parámetro λ se obtiene:

$$0 = \sum_{i=1}^{n} \frac{\delta_i}{\lambda} - \sum_{i=1}^{n} t_i \, exp(\boldsymbol{\beta'} \, \boldsymbol{X_i})$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\delta_i}{\lambda} = \sum_{i=1}^{n} t_i \, exp(\boldsymbol{\beta'} \, \boldsymbol{X_i})$$

de donde:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i \, exp(\boldsymbol{\beta'} \, \boldsymbol{X_i})} \tag{2.35}$$

Debido a la complejidad que existe en el cálculo de los estimadores de los parámetros $(\beta_1, \ldots, \beta_p)$ del modelo, se hace la estimación de estos parámetros con un *software estadístico*.

Considerando el resultado (2.35) y despejando t de (2.33) se obtiene el **estimador del p-ésimo cuantil** para un individuo de covariables $\boldsymbol{X_i}$ en el modelo de regresión exponencial, el cual está dado por:

$$\widehat{t_p} = \frac{\widehat{\lambda}^{-1}}{log(1-p) \, exp\,(\widehat{\boldsymbol{\beta}}' \boldsymbol{X_i})}$$

Análogamente, sustituyendo λ por $\widehat{\lambda}$ y los paramétros $(\beta_1, \beta_2, \ldots, \beta_p)$ por los correspondientes valores estimados $(\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$ se obtienen las funciones estimadas de supervivencia $\widehat{S}(t|\boldsymbol{X})$ y de riesgo $\widehat{h}(t|\boldsymbol{X})$.

$$\widehat{S}(t \,|\, \boldsymbol{X}) = exp(-\widehat{\lambda}\, t \; exp\,(\widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \cdots + \widehat{\beta}_p X_p)\,)$$

$$\widehat{h}(t \,|\, \boldsymbol{X}) = exp\,(\widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \cdots + \widehat{\beta}_p X_p)$$

Otros modelos de regresión exponencial que pueden encontrarse en la literatura estadística se obtienen modificando la función de riesgo relativo $\psi(\boldsymbol{X}; \, \boldsymbol{\beta})$, algunos ejemplos de estos son:

a) $h(t \,|\, \boldsymbol{X}) = \lambda \, exp\,(1 + \boldsymbol{\beta'} \boldsymbol{X})$

b) $h(t \,|\, \boldsymbol{X}) = \lambda \, log\,[1 + exp\,(\boldsymbol{\beta'} \boldsymbol{X})]$

c) $h(t \,|\, \boldsymbol{X}) = exp\,(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$

El estimador máximo verosímil $\widehat{\lambda}$ para λ en los modelos a) y b), se obtiene de manera similar al estimador obtenido en (2.35) y están dados por:

a) $\hat{\lambda} = \dfrac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i \, exp\,(1 + \boldsymbol{\beta'} \boldsymbol{X_i})}$ b) $\hat{\lambda} = \dfrac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i \, log\,[1 + exp\,(\boldsymbol{\beta'} \boldsymbol{X_i})]}$

Obsérvese que en el modelo que se describe en c) el parámetro λ, correspondiente a la parte paramétrica del modelo, no aparece de manera "*explícita*" en la expresión de la función de riesgo, no obstante, el diseño de este modelo tiene gran relevancia al momento de implementar un modelo de regresión exponencial en algunos de los softwares estadísticos más conocidos. La importancia de este modelo y la implementación del modelo de regresión exponecial en R se exponen con detalle a continuación.

2.3.2. Implementación del modelo en R

Una de las funciones de **R** que se utilizan para ajustar modelos paramétricos de regresión en el análisis de supervivencia es la función **survreg**, las distribuciones que esta función puede modelar son: la exponencial, la Weibull, la lognormal, la logistic y la log-logistic.

La estructura más sencilla de la función **survreg** para ajustar un modelo de regresión exponencial como el que se muestra en (2.32) es la siguiente:

$$\textbf{survreg(Surv(}\text{tiempos,censura}\textbf{)} \sim X_3 + X_5 + \cdots + X_r, \textbf{dist="exponential")}$$

los argumentos variables de esta función, en el modelo exponencial, son:

- tiempos: una muestra de tiempos de supervivencia.
- censura: indicador de censura de los tiempos de supervivencia.
- X_i: las covariables que el usuario supone que se incluyen en el modelo.

La estructura completa de la función **survreg** puede consultarse en **R** con la ejecución del comando:

- **help(survreg)**

Sin embargo, con la función **survreg** se ajusta un modelo de regresión de vida acelerada **AFT** con una distribución teórica especificada, por lo que los parámetros estimados que se obtienen con la ayuda de esta función deben de reparametrizarse con el objetivo de obtener los parámetros estimados correspondientes al modelo de riesgos proporcionales **PH**. Para ajustar un modelo de regresión **AFT** no es necesario realizar dicha reparametrización.

Sean $\widehat{\vartheta}_0$, $\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_p$ los parámetros estimados por la función **survreg**, se tiene que los parámetros correspondientes al modelo exponencial de riesgos proporcionales están dados por:

$$\begin{aligned} \widehat{\lambda} &= exp\left(-\widehat{\vartheta}_0\right) \\ \widehat{\beta}_i &= -\widehat{\theta}_i \qquad \forall\, i{=}1, 2, \ldots, p. \end{aligned} \tag{2.36}$$

Con fines ilustrativos, considere que la función de riesgo y supervivencia del modelo de regresión exponencial, que se realizan en la función **survreg** de **R** están dadas por:

$$h(t \mid \boldsymbol{X}) = exp\,(-\vartheta_0 - \vartheta_1 X_1 - \vartheta_2 X_2 - \cdots - \vartheta_p X_p) \tag{2.37}$$

$$S(t \mid \boldsymbol{X}) = exp(-exp(-\vartheta_0 - \vartheta_1 X_1 - \vartheta_2 X_2 - \cdots - \vartheta_p X_p))$$

descomponiendo el lado izquierdo de la igualdad (2.37) se tiene:

$$h(t \mid \boldsymbol{X}) = exp\,(-\vartheta_0)\, exp\,(-\vartheta_1 X_1 - \vartheta_2 X_2 - \cdots - \vartheta_p X_p)$$

si se considera que $\lambda = exp(-\vartheta_0)$ y $\vartheta_i = -\beta_i$ para toda i=$0, 1, 2, \ldots, p$, entonces:

$$h(t \mid \boldsymbol{X}) = \lambda\, exp\,(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) = \lambda\, exp\,(\,\boldsymbol{\beta'X}\,)$$

de manera que la expresión anterior coincide con el modelo presentado en (2.32) y por lo tanto, los parámetros estimados del modelo (2.32), pueden reescribirse en términos de los parámetros estimados del modelo (2.37) como:

$$\begin{aligned} \widehat{\lambda} &= exp\,(-\widehat{\vartheta}_0) \\ \widehat{\beta}_i &= -\widehat{\theta}_i \qquad \forall\, i{=}1, 2, \ldots, p. \end{aligned} \tag{2.38}$$

Obsérvese como la reparametrización realizada con anterioridad, coincide con la reparametrización (2.36) que se debe de realizar cuando se desea ajustar un modelo paramétrico de riesgos proporcionales mediante la función **survreg** de **R**.

Con el objetivo de aclarar las ideas expuestas en esta sección, se presenta un ejemplo en el que se desarrolla paso a paso, con ayuda de **R** en los cálculos, el procedimiento para obtener los parámetros estimados del modelo de regresión exponencial, asimismo, se ilustra como ajustar este modelo usando la función **survreg**.

Ejemplo 2.3.1. Supóngase que a los tiempos de supervivencia de los pacientes con *leucemia aguda* considerados en la base de datos **leuk** (*véase* Apéndice B) se les ajusta el siguiente modelo de regresión exponencial.

$$h(t \mid \boldsymbol{X}) = \lambda\, exp\,(\beta_1\, log_{10}(X_1) + \beta_2 X_2) \tag{2.39}$$

donde las covariables X_1 y X_2 para este modelo representan:

$X_1 =$ El conteo de glóbulos blancos en la sangre.

$$X_2 = \begin{cases} 1 & \text{si el paciente tiene características morfológicas} \\ 0 & \text{si el paciente no tiene características morfológicas} \end{cases}$$

Cabe mencionar que la base de datos **leuk** no tiene observaciones censuradas y/o truncadas, es decir, los tiempos de supervivencia de los pacientes son completamente conocidos.

La función de verosimilitud bajo este modelo de regresión exponencial es:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \lambda \, exp\left\{\beta_1 \, log_{10}(x_{i,1}) + \beta_2 x_{i,2}\right\} exp\{ -\lambda \, t_i \, exp(\beta_1 \, log_{10}(x_{i,1}) + \beta_2 x_{i,2})\}$$

la cual es equivalente a:

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} exp\{-\vartheta_0 - \vartheta_1 \, log_{10}(x_{i,1}) - \vartheta_2 x_{i,2}\} exp\{-t_i \, exp(-\vartheta_0 - \vartheta_1 \, log_{10}(x_{i,1}) - \vartheta_2 x_{i,2})\}$$

la log-verosimilitud de esta última expresión es:

$$log(L(\boldsymbol{\vartheta})) = -\sum_{i=1}^{n} (\vartheta_0 + \vartheta_1 \, log_{10}(x_{i,1}) + \vartheta_2 x_{i,2}) - \sum_{i=1}^{n} t_i \, exp(-\vartheta_0 - \vartheta_1 \, log_{10}(x_{i,1}) - \vartheta_2 x_{i,2})$$

y el estimador $\widehat{\boldsymbol{\vartheta}}_{\boldsymbol{i}}$ es la solución de las ecuaciones:

$$-\sum_{i=1}^{n} z_{i,j} + \sum_{i=1}^{n} t_i \, z_{i,j} \, exp\left(-\boldsymbol{\vartheta}' \boldsymbol{Z_i}\right) = 0 \qquad j = 1,\, 2,\, 3.$$

donde el vector $\boldsymbol{Z_i} = (1,\, log_{10}(x_{i,1}),\, x_{i,2})'$ y $z_{i,j}$ es el j-ésimo elemento del vector $\boldsymbol{Z_i}$.

La matriz de información de Fisher es una matriz de (3×3) con elementos:

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{\vartheta})_{rs} &= E\Big[-\sum_{i=1}^{n} z_{i,j} + \sum_{i=1}^{n} t_i \, z_{i,j} \, exp\left(-\boldsymbol{\vartheta}' \boldsymbol{Z_i}\right)\Big] \\ &= \sum_{i=1}^{n} E[T_i] \, z_{i,r} \, z_{i,s} \, exp\left(-\boldsymbol{\vartheta}' \boldsymbol{Z_i}\right) \qquad r, s = 1,\, 2,\, 3. \end{aligned}$$

debido a que la muestra no cuenta con observaciones censuradas se tiene que:

$$E(T_i) = exp\left(\boldsymbol{\vartheta}' \boldsymbol{Z_i}\right)$$

y entonces los elementos de la matriz $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ son:

$$\boldsymbol{\Sigma}(\boldsymbol{\vartheta})_{rs} = \sum_{i=1}^{n} z_{i,r} \, z_{i,s}.$$

mientras que los elementos de la matriz $\mathbf{I}(\boldsymbol{\vartheta})$ tienen la siguiente forma:

$$\mathbf{I}(\boldsymbol{\vartheta})_{r,s} = \sum_{i=1}^{n} t_i \, z_{i,r} \, z_{i,s} \, exp\left(-\boldsymbol{\vartheta}' \boldsymbol{Z_i}\right)$$

los resultados de la matriz de información observada $\mathbf{I}(\widehat{\boldsymbol{\vartheta}})$, así como los valores estimados de los parámetros ϑ's, que se muestran a continación, son el resultado del código **Ejemplo 2.3.1.** disponible en el Apéndice A.

$$\mathbf{I}\left(\widehat{\boldsymbol{\vartheta}}\right) = \begin{pmatrix} 32.98 & 136.41 & 16.99 \\ 136.41 & 577.51 & 67.28 \\ 16.99 & 67.28 & 16.99 \end{pmatrix}$$

la matriz inversa es:

$$\left(\mathbf{I}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)^{-1} = \begin{pmatrix} 1.5945 & -0.3543 & -0.1918 \\ -0.3543 & 0.0819 & 0.0299 \\ -0.1918 & 0.0299 & 0.1322 \end{pmatrix}$$

considerando el resultado de la expresión (2.30) se obtienen los errores estándar para cada uno de los parámetros estimados.

Parámetro	**Estimador**	$s.e\left(\widehat{\vartheta}_i\right)$
ϑ_0	5.8154	1.2674
ϑ_1	-0.7009	0.2862
ϑ_2	1.0176	0.3636

Tabla 2.1:

De acuerdo con las expresiones obtenidas en (2.38) se tiene que los parámetros estimados para el modelo de regresión exponencial de **PH** en términos de λ y β's son:

Parámetro	**Estimador**	$s.e\widehat{(\cdot)}$
λ	0.0029	0.0037
β_1	0.7009	0.2862
β_2	-1.0176	0.3636

Por lo tanto, la función estimada de riesgo para el modelo de regresión exponencial (2.39) está dada por:

$$\widehat{h}(t|\,\boldsymbol{X}) = 0.0029\, exp\left(0.7009\, log_{10}(X_1) - 1.0176 X_2\right)$$

De la igualdad anterior se puede inferir que el riesgo para un individuo con covariables $X_1 = 100,000$ y $X_2 = 0$, es decir, sin características morfológicas, es:

$$\widehat{h}(t|\,\boldsymbol{X}) = 0.0029\, exp\left(0.7009 \times 5\right) = 0.9646$$

A continuación se ajusta el modelo de regresión exponencial utilizando la función **survreg** del software estadístico **R**, antes de continuar con la lectura, se recomienda revisar la base de datos **leuk**.

Los parámetros ϑ's e información adicional relacionada con el experimento puede obtenerse a partir de las siguientes líneas de código.

```
library(survival)
library(MASS)
data(leuk)
leuk

# La fórmula para ajustar este modelo de regresión exponencial en R es:
# survreg(Surv(tiempo) ~ log10(wbc)+ag, data=leuk, dist="exponential")
# Obsérvese que el argumento Surv(tiempo) no es de la forma
# Surv(tiempos,censura)
# Particularmente, para estos datos, el argumento censura puede ser omitido,
# debido a que la muestra utilizada no tiene observaciones censuradas.

# Escriba el siguiente código:

vesper1=survreg(Surv(time)~ log10(wbc)+ag, data=leuk, dist="exponential")
summary(vesper1)
```

Dando como resultado la siguiente salida:

```
Call:
survreg(formula = Surv(time) ~ log10(wbc) + ag, data = leuk,
    dist = "exponential")
             Value     Std.Error      z        p
(Intercept)  5.815       1.263     4.60   4.15e-06
log10(wbc)  -0.701       0.286    -2.45   1.44e-02
agpresent    1.018       0.364     2.80   5.14e-03

Scale fixed at 1

Exponential distribution
Loglik(model)= -146.5   Loglik(intercept only)= -155.5
        Chisq= 17.82 on 2 degrees of freedom, p= 0.00014
Number of Newton-Raphson Iterations: 5
n= 33
```

Las columnas y otros resultados de la salida anterior, proporcionan la siguiente información:

- **Value**: El valor de las parámetros estimados $\widehat{\vartheta}_i$
- **Std.Error**: El error estándar de los correspondientes parámetros estimados.
- $\boldsymbol{z}$: Corresponde al valor del cociente $\dfrac{\widehat{\vartheta}_i}{s.e(\widehat{\vartheta}_i)}$.
- **p**: Es el p-value.
- **Loglik(model)**: La función log-verosimilitud estimada.
- **Loglik(intercept only)**: La función log-verosimilitud estimada en $\widehat{\vartheta}_0$.

Los valores de los parámetros estimados y sus correspondientes errores estándar se presentan en la siguiente tabla:

Parámetro	**Estimador**	$s.e\,(\widehat{\vartheta}_i)$
ϑ_0	5.815	1.263
ϑ_1	-0.701	0.286
ϑ_2	1.018	0.364

Obsérvese como la tabla anterior exhibe prácticamente los mismos resultados que los de la Tabla 2.1 (cuyo procedimeinto fue desarrollado paso a paso). Finalmente, utilizando la reparametrización (2.36) y haciendo uso del método delta para el cálculo de los errores estándar, se tiene que:

Parámetro	**Estimador**	$s.e\widehat{(\cdot)}$
λ	0.0029	0.003
β_1	0.701	0.286
β_2	-1.018	0.364

El cálculo del error estándar de $\widehat{\lambda}$ es el siguiente:

$$Var(\widehat{\lambda}) = Var(exp(-\widehat{\vartheta}_0)) \approx \left(\frac{\partial\, exp(-\widehat{\vartheta}_0)}{\partial\, \widehat{\vartheta}_0}\right)^2 Var(\widehat{\vartheta}_0)$$

$$Var(\widehat{\lambda}) \approx (-exp(-\widehat{\vartheta}_0))^2\, Var(\widehat{\vartheta}_0) = (exp(-\widehat{\vartheta}_0))^2\, Var(\widehat{\vartheta}_0)$$

$$\Rightarrow\ s.e\,(\widehat{\lambda}) = exp(-\widehat{\vartheta}_0)\, s.e\,(\widehat{\vartheta}_0) = 0.003$$

Uno de los supuestos en el ejemplo anterior es que las observaciones, correspondientes a la base de datos **leuk**, siguen una distribución exponencial, sin embargo, en el análisis de los datos de supervivencia, este supuesto debe de ser verificado mediante algún método estadístico.

2.3.3. Diagnóstico de distribución exponencial

Existen una variedad de métodos que permiten identificar si un conjunto de datos siguen una distribución exponencial. Generalmente estas pruebas tienen como hipótesis nula H_0 : *Las observaciones siguen una distribución exponencial* $exp(\lambda)$ y como hipótesis alternativa H_a : *Las observaciones no siguen una distribución exponencial* $exp(\lambda)$, es decir, rechazamos o no el supuesto de distribución exponencial.

Las pruebas de bondad de ajuste e identificación gráfica *(gráficas de probabilidad)* son algunos de los métodos de diagnóstico de distribución más utilizados[12]. En el análisis de supervivencia existe preferencia por los métodos gráficos de diagnóstico distribucional, debido a que las pruebas de bondad de ajuste tienden o bien a ser poco precisas para tamaños de muestras pequeños o rechazar siempre un modelo determinado para muestras grandes.

[12] Algunas de estás pruebas pueden consultarse en [8] .

El método gráfico más utilizado en los modelos con distribución exponencial es la llamada gráfica de probabilidad. La gráfica de probabilidad exponencial se basa en la siguiente transformación:

$$y = -log(S(t)) = H(t) = \lambda\, t$$

Si la gráfica de $y{=}H(t)$ contra $x{=}t$ tiene aproximadamente un comportamiento lineal, entonces, se puede considerar que los datos siguen una distribución exponencial, es decir, no se rechaza la hipótesis nula H_0 : *Las observaciones tienen una distribución exponencial.*

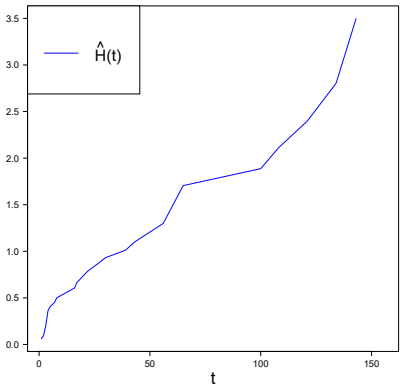


Figura 2.4: Obsérvese que la gráfica de probabilidad exponencial ($H(t)$ *v.s* t), para los datos del **Ejemplo 2.3.1**, tiene aproximadamente un comportamiento lineal, por lo tanto, no se rechaza la hipótesis de distribución exponencial. Cuando los parámetros del modelo son desconocidos se grafica la función $\hat{H}(t)$ *v.s* t.

Otro método gráfico utilizado comúnmente en el análisis de supervivencia es la comparación gráfica entre las curvas de supervivencia $\hat{S}(t)$ y de Kaplan-Meier. Si la curva $\hat{S}(t)$ se ajusta aproximadamente a la curva del estimador de Kaplan-Meier se puede considerar que los datos siguen una distribución exponencial.

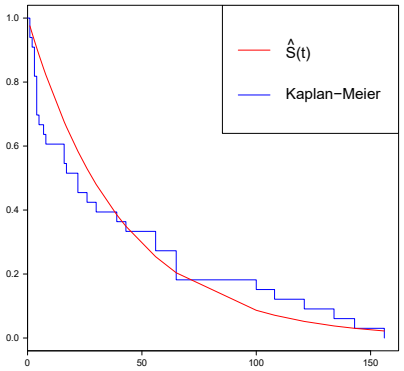


Figura 2.5: Siendo poco estrictos, es plausible considerar a $\hat{S}(t)$ como un buen ajuste de la curva del estimador de Kaplan-Meier.

2.4. Modelo de Riesgos Proporcionales Weibull

El modelo de regresión Weibull es quizás el modelo de riesgos proporcionales más utilizado en los experimentos relacionados con el análisis de los datos de supervivencia, debido a que la imágen de la función de densidad Weibull es similar (de acuerdo con el valor de sus parámetros) a la gráfica de otras distribuciones.

2.4.1. Modelo de Regresión Weibull

En la **sección 1.4.2** se pueden revisar las funciones de supervivencia relacionadas con una variable aleatoria T con distribución Weibull, sin embargo, para fines de los modelos de riesgos proporcionales, considere que T tiene funciones de densidad, supervivencia, riesgo y riesgo acumulado dadas por:

- $f(t) = \lambda\gamma\, t^{\gamma-1}\, exp(-\lambda\, t^{\gamma})$
- $S(t) = exp(-\lambda\, t^{\gamma})$
- $h(t) = \lambda\gamma\, t^{\gamma-1}$
- $H(t) = \lambda\, t^{\gamma}$

Decimos que el modelo de regresión

$$h(\, t \,|\, \boldsymbol{X}) = h_0(t;\, \boldsymbol{\alpha})\, \psi(\boldsymbol{X};\, \boldsymbol{\beta}),$$

es Weibull, si se puede considerar, que el tiempo de falla T sigue una distribución Weibull con parámetros (γ, λ), entonces, el modelo de regresión Weibull se escribe como:

$$h(\, t \,|\, \boldsymbol{X}) = \lambda\gamma\, t^{\gamma-1}\, \psi(\boldsymbol{X};\, \boldsymbol{\beta}) \tag{2.40}$$

Evidentemente el modelo de regresión Weibull se puede presentar en distintas formas, según sea el comportamiento de la función de riesgo relativo $\psi(\boldsymbol{X};\, \boldsymbol{\beta})$ que aparece en la expresión (2.40). La función de riesgo relativo más comúnmente utilizada es:

$$\psi(\boldsymbol{X};\, \boldsymbol{\beta}) = exp(\boldsymbol{\beta'}\, \boldsymbol{X}) = exp(\beta_1\, X_1 + \beta_2\, X_2 + \cdots + \beta_p\, X_p)$$

Sustituyendo la expresión anterior en (2.40), es posible escribir al modelo de regresión Weibull como:

$$h(\, t \,|\, \boldsymbol{X}) = \lambda\gamma\, t^{\gamma-1}\, exp(\boldsymbol{\beta'}\, \boldsymbol{X}) \tag{2.41}$$

Considerando esta última igualdad, a continuación se definen las funciones de supervivencia, densidad y riesgo acumulado dado el vector $\boldsymbol{X}$ $S(t|\, \boldsymbol{X})$, $f(t|\, \boldsymbol{X})$ y $H(t \,|\, \boldsymbol{X})$ respectivamente.

Función de supervivencia dado $\boldsymbol{X}$:

$$S(t|\, \boldsymbol{X}) = exp\Big(- \int_0^t h(u|\, \boldsymbol{X})\, du \Big)$$

$$
\begin{aligned}
&= exp\Big(- \int_0^t \lambda\gamma\, u^{\gamma-1}\, exp(\boldsymbol{\beta'}\, \boldsymbol{X})\, du \Big) \\
&= exp(-\lambda\, t^{\gamma}\, exp(\boldsymbol{\beta'}\, \boldsymbol{X})) \qquad (2.42) \\
&= (S_0(t))^{\psi(\boldsymbol{X};\,\beta)}
\end{aligned}
$$

Función de densidad dado $\boldsymbol{X}$:

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

simplificando se tiene que:

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

Función de riesgo acumulado dado $\boldsymbol{X}$:

$$H(t|\,\boldsymbol{X}) = -log(\,S(t|\,\boldsymbol{X})\,) = \lambda\, t^{\gamma}\, exp(\boldsymbol{\beta'}\, \boldsymbol{X})$$

Función de verosimilitud:

Sean $\boldsymbol{\vartheta'} = (\lambda,\, \gamma,\, \beta_1,\, \beta_2, \ldots,\, \beta_p)$ los parámetros correspondientes del modelo, T una variable aleatoria con distribución Weibull $(\gamma,\, \lambda)$, $\psi(\boldsymbol{X};\boldsymbol{\beta}) = exp(\boldsymbol{\beta'})$ la función de riesgo relativo y δ_i la i-ésima observación de la función indicadora de censura, entonces, la función de verosimilitud se escribe como:

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} h(t_i|\,\boldsymbol{X_i};\boldsymbol{\vartheta})^{\delta_i}\, S(t_i|\,\boldsymbol{X_i};\boldsymbol{\vartheta})$$

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} (\,\lambda\gamma\, t_i^{\gamma-1}\, exp(\boldsymbol{\beta'}\, \boldsymbol{X})\,)^{\delta_i}\, exp(-\lambda\, t_i^{\gamma}\, exp(\boldsymbol{\beta'}\, \boldsymbol{X}))$$

Los parámetros desconocidos del modelo de riesgos proporcionales Weibull se estiman mediante la maximización de la función de verosimilitud, procedimiento que es equivalente a maximizar la función log-verosimilitud con respecto a los parámetros desconocidos, es decir, para estimar el parámetro ϑ_i se resuelve

$$\frac{\partial}{\partial\, \vartheta_i}\, log(\,L(\boldsymbol{\vartheta})\,) = 0$$

La función log-verosimilitud correspondiente al modelo (2.41) está dada por:

$$
\begin{aligned}
log(\,L(\boldsymbol{\vartheta})\,) &= \sum_{i=1}^{n} \Big(\, \delta_i\, log(\,h(t_i|\,\boldsymbol{X_i};\boldsymbol{\vartheta})\,) + log(\,S(t_i|\,\boldsymbol{X_i};\boldsymbol{\vartheta})\,) \Big) \\
&= \sum_{i=1}^{n} \Big(\, \delta_i\, log(\,h(t_i|\,\boldsymbol{X_i};\boldsymbol{\vartheta})\,) - log(\,H(t_i|\,\boldsymbol{X_i};\boldsymbol{\vartheta})\,) \Big)
\end{aligned}
$$

sustituyendo a $h(t_i|\boldsymbol{X_i};\,\boldsymbol{\vartheta})$ y $H(t_i|\boldsymbol{X_i};\,\boldsymbol{\vartheta})$

$$\begin{aligned} log(\,L(\boldsymbol{\vartheta})\,) &= \sum_{i=1}^{n}\Big(\,\delta_i\,log(\,\lambda\gamma\,t_i^{\gamma-1}\,exp(\boldsymbol{\beta'}\,\boldsymbol{X_i})\,)-\lambda\,t_i^{\gamma}\,exp(\boldsymbol{\beta'}\,\boldsymbol{X_i})\,\Big) \\ &= \sum_{i=1}^{n}\Big(\,\delta_i(\,log(\lambda\gamma)+(\gamma-1)\,log(t_i)+\boldsymbol{\beta'}\boldsymbol{X_i}\,)-\lambda\,t_i^{\gamma}\,exp(\boldsymbol{\beta'}\,\boldsymbol{X_i})\,\Big) \\ &= \sum_{i=1}^{n}\Big(\delta_i(log(\lambda\gamma)+\gamma log(t_i)+\boldsymbol{\beta'}\boldsymbol{X_i})-\lambda\,t_i^{\gamma}\,exp(\boldsymbol{\beta'}\,\boldsymbol{X_i})\Big)-\sum_{i=1}^{n}\delta_i log(t_i) \end{aligned}$$

El último término de esta expresión $-\sum_{i=1}^{n}\delta_i log(t_1)$ no involucra ninguno de los parámetros desconocidos, por lo que esta suma puede ser omitida de la función log-verosimilitud. Entonces para el modelo de riesgos proporcionales Weibull (2.41):

$$log(\,L(\boldsymbol{\vartheta})\,)=\sum_{i=1}^{n}\Big(\,\delta_i(\,log(\lambda\gamma)+\gamma\,log(t_i)+\boldsymbol{\beta'}\boldsymbol{X_i}\,)-\lambda\,t_i^{\gamma}\,exp(\boldsymbol{\beta'}\,\boldsymbol{X_i})\,\Big) \qquad (2.43)$$

Generalmente la ecuación (2.43) se resuelve con la ayuda de algún software estadístico, estos softwares también proporcionan resultados relacionados con los parámetros estimados como los son: los intervalos de confianza, los errores estándar y algunos percentiles de interés.

Si se sustituyen los parámetros (γ, λ, β_1, $\beta_2,\ldots,$ β_p) por los correspondientes estimadores ($\widehat{\gamma}$, $\widehat{\lambda}$, $\widehat{\beta}_1$, $\widehat{\beta}_2,\ldots,$ $\widehat{\beta}_p$) en las expresiones (2.41) y (2.42) respectivamente, se obtienen las funciones estimadas de riesgo y supervivencia, las cuales están dadas por:

$$\widehat{h}(\,t\,|\,\boldsymbol{X})=\widehat{\lambda}\,\widehat{\gamma}\,t^{\hat{\gamma}-1}\,exp(\widehat{\beta}_1\,X_1+\widehat{\beta}_2\,X_2+\cdots+\widehat{\beta}_p\,X_p)$$

$$\widehat{S}(\,t\,|\,\boldsymbol{X})=exp(-\widehat{\lambda}\,t^{\hat{\gamma}}\,exp(\widehat{\beta}_1\,X_1+\widehat{\beta}_2\,X_2+\cdots+\widehat{\beta}_p\,X_p)\,)$$

El **estimador del *p*-ésimo cuantil** para un individuo de covariables $\boldsymbol{X_i}$ se obtiene al despejar t de la expresión $\widehat{S}(\,t\,|\,\boldsymbol{X})$=$(1-p)$, este despeje tiene como resultado

$$\hat{t}_p=\left\{\frac{\widehat{\lambda}^{-1}}{exp(\widehat{\boldsymbol{\beta'}}\boldsymbol{X_i})}\,log\Big(\frac{1}{1-p}\Big)\right\}^{1/\hat{\gamma}}$$

El error estándar para $\hat{t}_p$ y el intervalo de confianza para t_p, se obtienen a partir del cálculo del error estándar de $log(\hat{t}_p)$, el cálculo del error estándar de $log(\hat{t}_p)$ puede encontrárse en [1] y el resultado de éste es:

$$s.e\,(\,log(\hat{t}_p)\,)=\widehat{\gamma}^{-1}(\boldsymbol{d_0'}\,\boldsymbol{\Sigma}\,\boldsymbol{d_0})^{1/2}$$

donde $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas de los parámetros estimados, y $\boldsymbol{d_0}$ es un vector de $p+2$ componentes dadas por:

$$(\widehat{\lambda}^{-1},\,\widehat{\gamma}^{-1}\{c_p-log(\widehat{\lambda})-\widehat{\boldsymbol{\beta'}}\boldsymbol{X}\},\boldsymbol{X_1},\,\boldsymbol{X_2},\ldots,\,\boldsymbol{X_p})$$

con

$$c_p = log\Big(log\Big(\frac{1}{1-p}\Big)\Big)$$

El error estándar para $\hat{t}_p$ es entonces:

$$s.e\,(\hat{t}_p) = \hat{t}_p\, s.e(\, log(\hat{t}_p)\,)$$

Otros modelos de regresión Weibull pueden obtenerse modificando la función de riesgo relativo $\psi(\boldsymbol{X};\, \boldsymbol{\beta})$, algunos de estos modelos pueden derivarse de manera análoga de los modelos de regresión exponencial obtenidos en la sección anterior.

2.4.2. Implementación del modelo en R

La mayoría de los softwares estadísticos, incluyendo a **R**, utilizan una forma diferente a la adoptada en este trabajo para ajustar el modelo Weibull de riesgos proporcionales. Esto sucede porque las funciones de los paquetes estadísticos ajustan un modelo de regresión **AFT**[13] y no implementan el modelo de riesgos proporcionales.

Considérese que el modelo de riesgos proporcionales Weibull que interpreta la función **survreg** del paquete estadístico **R** está dada por:

$$S(t|\,\boldsymbol{X}) = exp\Big\{ -exp\Big(\,\frac{log\,(t) - \mu - \boldsymbol{\alpha'}\,\boldsymbol{X}}{\sigma}\,\Big)\Big\} \tag{2.44}$$

En particular el modelo anterior resulta ser un modelo log-lineal para la variable aleatoria T_i correspondiente al i-ésimo individuo de una muestra.

La correspondencia que hay entre los parámetros del modelo (2.42) y el modelo interpretado por **R**, es:

$$\begin{aligned} \gamma &= \sigma^{-1} \\ \lambda &= exp\,(-\mu/\sigma) \\ \beta_j &= -\alpha_j/\sigma \qquad \forall\, j = 1,\, 2, \ldots, p \end{aligned} \tag{2.45}$$

Como ya se había mencionado con anterioridad, la función de **R** que se utiliza para ajustar un modelo de regresión Weibull en el análisis de supervivencia es la función **survreg**. La estructura más sencilla de la función **survreg** para ajustar un modelo Weibull como el que se muestra en (2.41) es:

$$\textbf{survreg(Surv(}\text{tiempos,censura}\textbf{)} \sim X_3 + X_5 + \cdots + X_r, \textbf{dist=“weibull”)}$$

Los argumentos variables de esta función son los mismos que para los del modelo de regresión exponencial, asimismo, la estructura completa de la función **survreg** puede consultarse en **R** mediante la ejecución del comando: **help(survreg)**.

[13] Se les conoce como modelos de regresión **AFT** a los modelos de tiempo de vida o falla acelerada.

Además de proporcionar los parámetros estimados $\widehat{\sigma}$, $\widehat{\mu}$ y $\widehat{\alpha}_j$, en los resultados de la función **survreg** se puede consultar los intervalos de confianza, los errores estándar y algunos percentiles de interés en términos de $log\,(\widehat{\sigma})$, $\widehat{\mu}$ y $\widehat{\alpha}_j$.

La propiedad de invarianza de los estimadores de máxima verosimilitud nos permite encontrar de manera sencilla a los parámetros $\widehat{\gamma}$, $\widehat{\lambda}$ y $\widehat{\beta}_j$ a partir de las expresiones dadas en (2.45), sin embargo, los cálculos correspondientes para los errores estándar de estos parámetros requieren de un procedimiento más elaborado.

Utilizando el método delta, se tiene que:

$$\begin{aligned} Var(\widehat{\gamma}) &= Var(\widehat{\sigma}^{-1}) \approx \left(\frac{\partial\, \widehat{\sigma}^{-1}}{\partial\, \widehat{\sigma}} \right)^2 Var(\widehat{\sigma}) \\ &= \left(\frac{-1}{\widehat{\sigma}^2} \right)^2 Var(\widehat{\sigma}) \\ &= \frac{Var(\widehat{\sigma})}{\widehat{\sigma}^4} \end{aligned}$$

por lo tanto

$$s.e\,(\widehat{\gamma}) = s.e\,(\widehat{\sigma})/\widehat{\sigma}^2 \tag{2.46}$$

Partiendo ahora del método delta bivariado:

$$Var(g(\widehat{\theta}_1, \widehat{\theta}_2)) \approx \left(\frac{\partial\, g}{\partial\, \widehat{\theta}_1} \right)^2 Var(\widehat{\theta}_1) + \left(\frac{\partial\, g}{\partial\, \widehat{\theta}_2} \right)^2 Var(\widehat{\theta}_2) + 2 \left(\frac{\partial\, g}{\partial\, \widehat{\theta}_1} \right) \left(\frac{\partial\, g}{\partial\, \widehat{\theta}_2} \right) Cov\,(\widehat{\theta}_1, \widehat{\theta}_2)$$

se calcula la varianza del estimador $\widehat{\lambda}$ como:

$$Var\,(\widehat{\lambda}) = Var\,(\,g(\widehat{\sigma}, \widehat{\mu})\,) = Var\,(exp\,(-\widehat{\mu}/\widehat{\sigma}))$$

$$\begin{aligned} Var\,(\widehat{\lambda}) \;\approx\; & \left(\frac{\widehat{\mu}\, exp\,(-\widehat{\mu}/\widehat{\sigma})}{\widehat{\sigma}^2} \right)^2 Var(\widehat{\sigma}) \;+\; \left(\frac{-exp\,(-\widehat{\mu}/\widehat{\sigma})}{\widehat{\sigma}} \right)^2 Var(\widehat{\mu}) \;+ \\ & 2 \left(\frac{\widehat{\mu}\, exp(-\widehat{\mu}/\widehat{\sigma})}{\widehat{\sigma}^2} \right) \left(\frac{-exp(-\widehat{\mu}/\widehat{\sigma})}{\widehat{\sigma}} \right) Cov\,(\widehat{\mu},\, \widehat{\sigma}) \end{aligned}$$

de donde:

$$s.e\,(\widehat{\lambda}) = (\,Var\,(\widehat{\lambda})\,)^{1/2} \tag{2.47}$$

Usando una vez más el método delta bivariado, se obtienen la varianza y el error estándar para los parámetros β_j para toda $j = 1,\, 2, \ldots,\, p$.

De la expresión (2.45) se tiene que $Var(\widehat{\beta}_j) = Var(-\widehat{\alpha}_j/\widehat{\sigma})$, entonces por el método delta bivariado se tiene que:

$$Var(\widehat{\beta}_j) \approx \left(\frac{\widehat{\alpha}_j}{\widehat{\sigma}^2}\right)^2 Var(\widehat{\sigma}) + \left(\frac{-1}{\widehat{\sigma}}\right)^2 Var(\widehat{\alpha}_j) + 2\left(\frac{\widehat{\alpha}_j}{\widehat{\sigma}^2}\right)\left(\frac{-1}{\widehat{\sigma}}\right) Cov\left(\widehat{\alpha}_j, \widehat{\sigma}\right)$$

Análogamente el error estándar de $\widehat{\beta}_j$ se obtiene de la raíz cuadrada de la expresión anterior

$$s.e\left(\widehat{\beta}_j\right) = \left(\, Var\left(\widehat{\beta}_j\right)\right)^{1/2} \tag{2.48}$$

A continuación se presenta un ejemplo de como ajustar un modelo de riesgos proporcionales Weibull haciendo uso de la función **survreg** en el software estadístico **R**.

Ejemplo 2.4.1 Considere la base de datos **leuk**, misma que fue utilizada para el Ejemplo 2.3.1, y supóngase ahora que se ajusta, a los tiempos de supervivencia de los pacientes con leucemia aguda, el siguiente modelo de regresión Weibull[14]

$$h(t|\,\boldsymbol{X}) = \lambda\gamma\, t^{\gamma-1}\, exp\left(\beta_1\, log_{10}(X_1) + \beta_2 X_2\right) \tag{2.49}$$

De las expresiones (2.44) y (2.45), el modelo anterior, puede reescribirse en términos de la siguiente función de supervivencia:

$$S(t|\,\boldsymbol{X}) = exp\left\{ - exp\left(\frac{log\,(t) - \mu - (\,\alpha_1\, log_{10}(X_1) + \alpha_2 X_2\,)}{\sigma} \right)\right\}$$

Los parámetros estimados $\widehat{\sigma}$, $\widehat{\mu}$, $\widehat{\alpha}_j$ e información adicional relacionada con el experimento puede obtenerse a partir de las siguientes líneas de código.

```
library(survival)
library(MASS)
data(leuk)
leuk

# La fórmula para ajustar este modelo de regresión Weibulll en R es:
# survreg(Surv(tiempo) ~ log10(wbc)+ag, data=leuk, dist="weibull")
# Obsérvese que el argumento Surv(tiempo) no es de la forma
# Surv(tiempos,censura)
# Particularmente, para estos datos, el argumento censura puede ser omitido,
# debido a que la muestra utilizada no tiene observaciones censuradas.

# Escriba el siguiente código:

vesper2=survreg(Surv(time)~ log10(wbc)+ag, data=leuk, dist="weibull")
summary(vesper2)
```

[14]Las especificaciones de las variables explicativas que intervienen en el modelo, pueden consultarse en el **Ejemplo 2.3.1** o en el Apéndice B.

Dando como resultado la siguiente salida:

```
Call:
survreg(formula = Surv(time) ~ log10(wbc) + ag, data = leuk,
    dist = "weibull")
               Value      Std.Error      z        p
(Intercept)    5.8524       1.323      4.425   9.66e-06
log10(wbc)    -0.7146       0.302     -2.363   1.81e-02
agpresent      1.0206       0.378      2.699   6.95e-03
Log(scale)     0.0399       0.139      0.287   7.74e-01

Scale= 1.04

Weibull distribution
Loglik(model)= -146.5   Loglik(intercept only)= -153.6
        Chisq= 14.18 on 2 degrees of freedom, p= 0.00084
Number of Newton-Raphson Iterations: 6
n= 33
```

Sea $\boldsymbol{\theta} = (\widehat{\sigma},\, \widehat{\mu},\, \widehat{\alpha}_j)$, la matriz inversa $(\mathbf{I}(\widehat{\boldsymbol{\theta}}))^{-1}$ de la matriz de información observada, se muestra al usuario escribiendo el siguiente código:

```
vesper2$var

              (Intercept)    log10(wbc)    agpresent     Log(scale)
(Intercept)    1.74958436  -0.391562138  -0.204571971   0.018232313
log10(wbc)    -0.39156214   0.091468019   0.031505395  -0.006676612
agpresent     -0.20457197   0.031505395   0.142963239   0.001372447
Log(scale)     0.01823231  -0.006676612   0.001372447   0.019376157
```

En particular, la salidas anteriores, no proporcionan el error estándar del estimador $\widehat{\sigma}$, ni la función de covarianza de $\widehat{\mu}$ y $\widehat{\sigma}$, sin embargo, ambos resultados pueden encontrase con ayuda del método delta

$$Var(\, log\,(\widehat{\sigma})\,) \approx \left(\frac{\partial\, log\,(\widehat{\sigma})}{\partial\, \widehat{\sigma}}\right)^2 Var(\widehat{\sigma})$$

$$\Rightarrow Var(\widehat{\sigma}) \approx \widehat{\sigma}^2\, Var(\, log\,(\widehat{\sigma})\,)$$

$$\therefore\ s.e\,(\widehat{\sigma}) = \widehat{\sigma}\, s.e\,(\, log(\widehat{\sigma})\,) \tag{2.50}$$

Sustituyendo los valores correspondientes de nuestro ejemplo en la expresión (2.50) se tiene que :

$$s.e\,(\widehat{\sigma}) = (1.04)\,(0.139) = 0.144$$

La expresión para la covarianza de $\widehat{\mu}$ y $\widehat{\sigma}$ es la siguiente

$$Cov\,(\widehat{\mu},\, \widehat{\sigma}) = \widehat{\sigma}\, Cov\,(\,\widehat{\mu},\, log(\,\widehat{\sigma}\,)) \tag{2.51}$$

Utilizando nuestros datos,

$$Cov\,(\widehat{\mu},\, \widehat{\sigma}) = (1.04)\,(0.01823) = 0.018$$

Los valores de los parámetros estimados y sus correspondientes errores estándar se presentan en la siguiente tabla:

Parámetro	**Estimador**	$s.e\widehat{(\cdot)}$
μ	5.8524	1.323
α_1	-0.7146	0.302
α_2	1.0206	0.378
$log\,(\sigma)$	0.0399	0.139
σ	1.04	0.144

De la reparametrización realizada en (2.45) y de las expresiones (2.46), (2.47), (2.48) y (2.51), se tiene que los valores de los parámetros estimados $\widehat{\gamma}$, $\widehat{\lambda}$, $\widehat{\beta}_j$ y sus errores estándar correspondientes son:

Parámetro	**Estimador**	$s.e\widehat{(\cdot)}$
γ	0.9615	0.133
β_1	0.6871	0.100
β_2	-0.9813	0.150
λ	0.0035	2.3e−05

Por lo tanto, la función estimada de riesgo para el modelo de regresión Weibull (2.49) está dada por:

$$\widehat{h}(t|\,\boldsymbol{X}) = (0.0035)\,(0.9615)\,t^{0.9615-1}\,exp\,(0.6871\,log_{10}(X_1) - 0.9813\,X_2)$$

2.4.3. Diagnóstico de distribución Weibull

Análogo al diagnóstico de distribución exponencial, existen una gran variedad de métodos que permiten identificar si un conjunto de datos siguen una distribución Weibull. En particular, se recomiendan usar métodos gráficos para el diagnóstico de la distribución Weibull.

Los métodos gráficos que se van a utilizar tienen como hipótesis nula H_0 : *Las observaciones tienen una distribución Weibull*(γ, λ) y como hipótesis alternativa H_a : *Las observaciones no siguen una distribución Weibull*(γ, λ).

Los métodos gráficos que se utilizan son; la gráfica de probabilidad Weibull y la comparación gráfica entre las curvas de supervivencia $\widehat{S}(t)$ de Kaplan-Meier.

La gráfica de probabilidad Weibull se basa en las siguientes transformaciones

$$y = log(-log(\,S(t)\,)\,)$$

$$y = log\,(\,H(t)\,)$$

y la transformación:

$$x = log\,(t)$$

Si la gráfica de $y = log\,(H(t))$ contra $x = log\,(t)$ tiene aproximadamente un comportamiento lineal, entonces, se puede considerar que los datos siguen una distribución Weibull, de lo contrario, se rechaza la hipótesis nula H_0 : *Las observaciones tienen una distribución Weibull.*

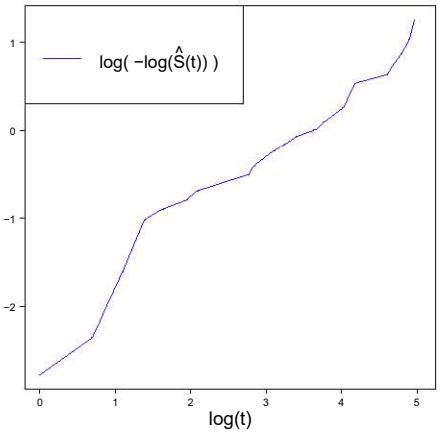


Figura 2.6: La imagen de esta gráfica corresponde aproximadamente al comportamiento de una línea recta, por lo tanto, no se rechaza la hipótesis de distribución Weibull para los datos del **Ejemplo 2.4.1**. Cuando los parámetros del modelo son desconocidos se grafica la función $log\,(\widehat{H}(t))$ $v.s$ $\log\,(t)$.

El segundo método gráfico que se realiza es la comparación gráfica entre las curvas de supervivencia $\widehat{S}(t)$ y de Kaplan-Meier. Si la curva $\widehat{S}(t)$ se ajusta aproximadamente a la curva del estimador de Kaplan-Meier entonces se puede considerar que los datos siguen una distribución Weibull.

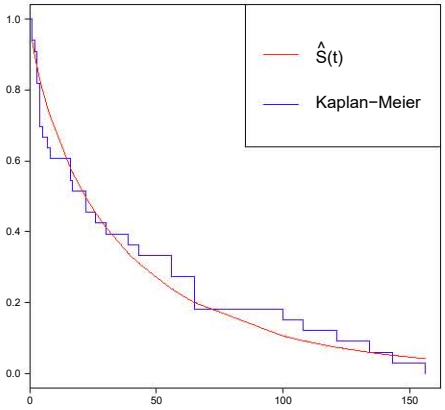


Figura 2.7: Gráficamente la función $\widehat{S}(t)$ funciona como un buen candidato para ajustar a la curva del estimador de Kaplan-Meier, por lo tanto, no se rechaza la hipótesis de distribución Weibull.

2.5. Modelo de Riesgos Proporcionales Gompertz

A diferencia de los modelos Weibull y exponencial, el Gompertz es un modelo exclusivo de los modelos de riesgos proporcionales, es decir, no se puede ajustar una distribución Gompertz a un modelo de vida acelerada.

El modelo de riesgos proporcionales Gompertz ha sido ampliamente utilizado en la demografía, principalmente en estudios relacionados con la mortalidad de los adultos en países desarrollados.

2.5.1. Modelo de Regresión Gompertz

Sean:

$$h(t) = \alpha e^{\beta t} \qquad \text{y} \qquad S(t) = exp\Big(-\frac{\alpha}{\beta}(e^{\beta t} - 1) \Big)$$

las funciones de riesgo y supervivencia de una distribución Gompertz de parámetros (α, β), se define el modelo de riesgos proporcionales Gompertz, en términos de la función de riesgo, como:

$$h(t|\, \boldsymbol{X}) = \alpha e^{\beta t}\, \psi(\boldsymbol{X};\, \boldsymbol{\beta}) \tag{2.52}$$

Si

$$\psi(\boldsymbol{X};\, \boldsymbol{\beta}) = exp\,(\boldsymbol{\beta'X}) = exp\,(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

entonces, el modelo (2.52) se escribe como:

$$\begin{aligned} h(t|\, \boldsymbol{X}) &= \alpha e^{\beta t}\, exp\,(\boldsymbol{\beta'X}) \\ &= \alpha e^{\beta t}\, exp\,(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \end{aligned} \tag{2.53}$$

Otros modelos de riesgos proporcionales Gompertz pueden obtenerse al modificar la función $\psi(\boldsymbol{X};\, \boldsymbol{\beta})$ de la expresión (2.52).

Las funciones de supervivencia, de densidad y de riesgo acumulado, asociadas a la expresión (2.53) se obtienen de la siguiente manera:

Función de supervivencia dado $\boldsymbol{X}$:

$$\begin{aligned} S(t|\, \boldsymbol{X}) &= exp\Big(-\int_0^t h(u|\, \boldsymbol{X})\, du \Big) \\ &= exp\Big(-\int_0^t \alpha e^{\beta u} exp\,(\boldsymbol{\beta'X})\, du \Big) \\ &= exp\Big(-\frac{\alpha e^{\beta t}}{\beta}\, exp\,(\boldsymbol{\beta'X}) \Big|_0^t \Big) \\ &= exp\Big(-\frac{\alpha}{\beta}\, exp\,(\boldsymbol{\beta'X})\, (e^{\beta t} - 1) \Big) \end{aligned} \tag{2.54}$$

Función de densidad dado $\boldsymbol{X}$:

$$\begin{aligned} f(t|\,\boldsymbol{X}) &= h(t|\,\boldsymbol{X})\,S(t|\,\boldsymbol{X}) \\ &= \alpha e^{\beta t}\,exp\,(\boldsymbol{\beta'X})\,exp\Big(-\frac{\alpha}{\beta}\,exp\,(\boldsymbol{\beta'X})\,(e^{\beta\,t}-1)\Big) \end{aligned}$$

Función de riesgo acumulado dado $\boldsymbol{X}$:

$$H(t|\,\boldsymbol{X}) = -log\,(S(t|\,\boldsymbol{X})) = \frac{\alpha}{\beta}\,exp\,(\boldsymbol{\beta'X})\,(e^{\beta\,t}-1)$$

Función de verosimilitud:

Sea $\boldsymbol{\vartheta'} = (\alpha,\,\beta,\,\beta_1,\,\beta_2,\ldots,\,\beta_p)$ entonces, la función de verosimilitud correspondiente al modelo de riesgos proporcionales (2.53) es:

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} h(t_i\,|\,\boldsymbol{X_i};\,\boldsymbol{\vartheta})^{\delta_i}\,S(t_i\,|\,\boldsymbol{X_i};\,\boldsymbol{\vartheta})$$

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^{n}(\,\alpha e^{\beta t}\,exp\,(\boldsymbol{\beta'X_i})\,)^{\delta_i}\,exp\Big(-\frac{\alpha}{\beta}\,exp\,(\boldsymbol{\beta'X_i})\,(e^{\beta\,t}-1)\Big)$$

los parámetros estimados del modelo se obtienen maximizando la función *log-verosimilitud*[15]

$$log\,(\,L(\boldsymbol{\vartheta})\,) = \sum_{i=1}^{n}\Big\{\delta_i\,[\,log\,(\alpha)+\beta\,t_i+\boldsymbol{\beta'X_i}\,] - \frac{\alpha}{\beta}\,exp\,(\boldsymbol{\beta'X_i})\,(e^{\beta\,t_i}-1)\Big\}$$

Las funciones de supervivencia y riesgo estimadas dado $\boldsymbol{X}$ se obtienen al sustituir el vector de parámetros $(\alpha,\,\beta,\,\beta_1,\,\beta_2,\ldots,\,\beta_p)$ por los estimadores $(\widehat{\alpha},\,\widehat{\beta},\,\widehat{\beta}_1,\,\widehat{\beta}_2,\ldots,\,\widehat{\beta}_p)$ en las expresiones (2.54) y (2.53) respectivamente.

$$\widehat{S}(t|\,\boldsymbol{X}) = exp\Big(-\frac{\widehat{\alpha}}{\widehat{\beta}}\,exp\,(\widehat{\beta}_1X_1+\widehat{\beta}_2X_2+\cdots+\widehat{\beta}_pX_p)\,(e^{\widehat{\beta}\,t}-1)\Big)$$

$$\widehat{h}(t|\,\boldsymbol{X}) = \widehat{\alpha}e^{\widehat{\beta}t}\,exp\,(\widehat{\beta_1}X_1+\widehat{\beta}_2X_2+\cdots+\widehat{\beta}_pX_p)$$

***p*-ésimo cuantil:**

Haciendo $\widehat{S}(t|\,\boldsymbol{X}) = 1-p$ y despejando t de esta expresión se obtiene el valor estimado para el p-ésimo cuantil, realizando los cálculos necesarios, se tiene que:

$$\hat{t}_p = \widehat{\beta}^{-1}\,log\Big\{\Big(-\frac{\widehat{\beta}}{\widehat{\alpha}}\Big)\Big(\frac{log\,(1-p)}{exp\,(\widehat{\boldsymbol{\beta}}\boldsymbol{'X})}\Big)+1\Big\}$$

[15] Análogo a los modelos de riesgos proporcionales Weibull y exponencial, los parámetros del modelo de regresión Gompertz son estimados con ayuda de un software estadístico.

2.5.2. Implementación del modelo en R

La función **flexsurvreg** de **R** se utiliza para ajustar modelos de riesgos proporcionales Gompertz.

La estructura más sencilla de la función **flexsurvreg** para ajustar un modelo Gompertz como el que se muestra en (2.53) es:

$$\textbf{flexsurvreg(Surv(}\text{tiempos,censura}\textbf{)} \sim X_3 + X_5 + \cdots + X_r, \textbf{dist=“gompertz”)}$$

Los argumentos variables de esta función, son: **tiempos**; tiempos de supervivencia de una muestra, **censuras**; indicador de censura, $\boldsymbol{X_i}$; las covariables que se involucran en la descripción de la falla y **dist**; la distribución del modelo.

La función **flexsurvreg** también se puede utilizar para ajustar otros modelos paramétricos de riesgos proporcionales como el Weibull y el exponencial.

A diferencia de la función **survreg** con distribución Weibull y/o exponencial, los parámetros estimados que se obtienen al aplicar la función **flexsurvreg** con distribución Gompertz no necesitan ser reparametrizados para propocionar los parámetros en términos del modelo presentado en (2.53).

Ejemplo 2.5.1 Supóngase que se ajusta el siguiente modelo de regresión a los tiempos de supervivencia de la base de datos **leuk** (*véase* Apéndice B).

$$h(t|\,\boldsymbol{X}) = \alpha\, e^{\beta\, t}\, exp\,(\beta_i\, log_{10}(X_1) + \beta_2(X_2)) \tag{2.55}$$

Los estimadores $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\beta}_j$ e información adicional relacionada con el experimento se obtienen al escribir las siguientes líneas de código en **R**.

```
library(survival)
library(MASS)
data(leuk)
leuk

# La fórmula para ajustar este modelo de PH Gompertz en R es:
# flexsurvreg(Surv(tiempo) ~ log10(wbc)+ag, data=leuk, dist="gompertz")
# Obsérvese que el argumento Surv(tiempo) no es de la forma
# Surv(tiempos,censura)
# Particularmente, para estos datos, el argumento censura puede ser omitido,
# debido a que la muestra utilizada no tiene observaciones censuradas.

# Escriba el siguiente código:

vesper3=flexsurvreg(Surv(time)~ log10(wbc)+ag, data=leuk, dist="gompertz")
vesper3
```

Dando como resultado la siguiente salida:

```
Call:
flexsurvreg(formula = Surv(time) ~ log10(wbc) + ag, data = leuk,
    dist = "gompertz")

Estimates:
            data mean      est        L95%      U95%       se      exp(est)
shape              NA  0.008457  -0.00315   0.02007  0.005926        NA
rate               NA  0.001518   0.00010   0.02246  0.002087        NA
log10(wbc)  4.136108  0.819015   0.22949   1.40853  0.300780   2.26826
agpresent   0.515152 -1.268923  -2.09179  -0.44605  0.419838   0.28113

N = 33,  Events: 33,  Censored: 0
Total time at risk: 1349
Log-likelihood = -145.5534, df = 4
AIC = 299.1067
```

Los parámetros estimados del modelo y sus correspondientes errores estándar son:

Parámetro	**Estimador**	$s.e\widehat{(\cdot)}$
α	0.0015	0.0020
β	0.0084	0.0059
β_1	0.8190	0.3007
β_2	-1.2689	0.4198

La matriz $(\mathrm{I}(\boldsymbol{\theta}))^{-1}$ con $\boldsymbol{\theta} = (\beta,\, log\,(\alpha),\, \beta_1,\, \beta_2)$ se obtiene escribiedo:

```
vesper3$cov
                shape          rate      log10(wbc)     agpresent
shape       3.511767e-05 -0.003089457  0.0005469858 -0.001182927
rate       -3.089457e-03  1.890080026 -0.4051858207 -0.111666884
log10(wbc)  5.469858e-04 -0.405185821  0.0904687845  0.016743930
agpresent  -1.182927e-03 -0.111666884  0.0167439302  0.176263829
```

Finalmente, sustituyendo en (2.55) los resultados obtenidos, se tiene que:

$$h(t|\,\boldsymbol{X}) = 0.0015\, e^{0.0084\, t}\, exp\,(0.819\, log_{10}(X_1) - 1.2689\, \beta_2)$$

Como parte de este ejemplo, se incluye un código en el Apéndice A, en el que se programa un procedimiento que conduce al resultado de los parámetros estimados y sus correspondientes errores estándar.

Existe otra función en **R** que se utiliza para ajustar modelos paramétricos de riesgos proporcionales, esta función es **phreg**. A partir de esta función se pueden ajustar los modelos Weibull, Gompertz y exponencial, estos últimos como un caso particular de la distribución Weibull.

phreg(Surv(tiempos,censura**))** $\sim X_1 + X_2 + \cdots + X_r$, **dist="gompertz")**

La expresión anterior corresponde a la estructura más sencilla de la función **phreg** de **R** (*es necesario cargar la paquetería* **eha**), si no se especifica ninguna distribución se ajusta un modelo de riesgos proporcionales Weibull.

2.5.3. Diagnóstico de distribución Gompertz

Uno de los métodos más utilizados para verificar que un conjunto de datos se puede ajustar con una distribución Gompertz es el análisis de la gráfica $log(\widehat{\mathsf{h}}(t))$. En particular $\widehat{\mathsf{h}}(t)$ es conocida como la función de riesgo empírico, definida como:

$$\widehat{\mathsf{h}}(t_i) = \frac{-[\, log\,(\,\widehat{S}(t_i)\,) + log\,(\,\widehat{S}(t_{i-1})\,]}{t_i - t_{i-1}}$$

donde $\widehat{S}(t_i)$ es el estimador de Kaplan-Meier.

Si la gráfica $y = log(\widehat{\mathsf{h}}(t))$ $v.s$ $x = t$ tiene un aspecto semejante al de una línea recta, entonces, el supuesto de distribución de la muestra es razonable, de lo contrario, se rechaza la hipótesis nula H_0 : *Las observaciones tienen una distribución Gompertz con parámetros* (α, β).

Análogo a los diagnósticos de distribución Weibull y exponencial, se puede realizar una comparación gráfica entre las curvas de supervivencia $\widehat{S}(t)$ y de Kaplan-Meier. Se considera que los datos se pueden ajustar a una distribución Gompertz si la curva $\widehat{S}(t)$ es aproximadamente igual a la curva del estimador de Kaplan-Meier.

Capítulo 3

Aplicación del Modelo de Riesgos Proporcionales con R

En este capítulo se ajusta un modelo paramétrico de riesgos proporcionales a los datos **anderson.dat** (*véase* Apéndice B), la base de datos contiene la información del tiempo de remisión por esteroides (en semanas) de 42 pacientes con leucemia aguda[1].

Con el objetivo de comparar la eficiencia de un tratamiento, los pacientes fueron divididos aleatoriamente en dos grupos iguales. El primer grupo recibe un tratamiento 6-mercaptopurina o abreviadamente 6-MP, mientras que al segundo grupo (grupo de referencia) se le suministra un placebo.

Adicionalmente en la base de datos **anderson.dat** se dispone de la siguiente información:

- **Censura:** Indicador de censura (0 = observación censurada, 1 = el paciente presentó una falla, es decir, una recaída). *Nota:* Se considera que las observaciones censuradas, están censuradas por la derecha.
- **Sexo:** El género del paciente (0 = femenino, 1 = masculino).
- **logWBC:** El logaritmo del conteo de los glóbulos blancos. (log White Blood Cell count).
- **Tratamiento:** El tipo de tratamiento que recibe el paciente a lo largo del estudio (0 = tratamiento 6-MP, 1 = Placebo).

La base de datos **anderson.dat**, utilizada en esta tesis, puede ser consultada en: http://statweb.stanford.edu/~olshen/hrp262spring01/spring01Assignments/anderson.txt, asimismo, en el Apéndice A, en la sección correspondiente a los códigos del Capítulo 3, se puede revisar cómo y de dónde se descarga esta base de datos.

[1]Una remisión ocurre cuando el trastorno de la enfermedad aparece como inactivo en aquellas personas que sufren de un mal crónico, por el contrario, se dice que un paciente sale de remisión (presenta una falla) cuando el trastorno de la enfermedad regresa.

Análisis descriptivo de la muestra

Con el objetivo de conocer algunas de las principales características de los pacientes con leucemia aguda, se presentan algunos datos estadísticos haciendo uso de gráficas y tablas, las cuales facilitan la interpretación de estos datos.

En algunas ocasiones, la variable **logWBC** será utilizada como una variable categórica de acuerdo con los niveles de glóbulos blancos registrados para cada individuo en el estudio, siguiendo a autores como Kleinbaum y Klein, *véase* [7], la variable **logWBC** es clasificada en las siguientes categorías:

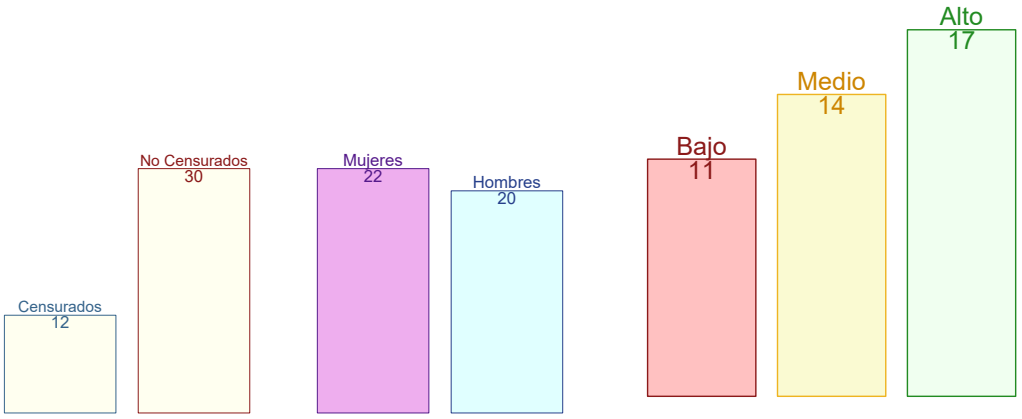
- bajo (0-2.30]
- medio (2.31-3.00]
- alto (>3.00)

Con ayuda del comando **summary()** de **R**, aplicado directamente sobre la base de datos, se obtienen los valores de las siguientes estadísticas para cada una de las columnas que forman parte de la matriz de datos: **Min**: el valor mínimo del registro, **1st Qu.**: el primer cuartil, **Median**: la mediana, **Mean**: la media, **3rd Qu.**: tercer cuartil y el valor máximo denotado por **Max**.

```
summary(and)

tiempo          delta            sex              lwbc            trat
Min.   : 1.00   Min.   :0.000   Min.   :0.000   Min.   :1.450   Min.   :0.0
1st Qu.: 6.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:2.303   1st Qu.:0.0
Median :10.50   Median :1.000   Median :0.000   Median :2.800   Median :0.5
Mean   :12.88   Mean   :0.714   Mean   :0.476   Mean   :2.930   Mean   :0.5
3rd Qu.:18.50   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:3.490   3rd Qu.:1.0
Max.   :35.00   Max.   :1.000   Max.   :1.000   Max.   :5.000   Max.   :1.0
```

En la tabla anterior se puede observar que el tiempo máximo de supervivencia es de 35 semanas, en contraste, el tiempo mínimo es de una semana, mientras que la media y la mediana son de 12.88 y 10.50 semanas respectivamente. Por otro lado, el logWBC (logaritmo del conteo de glóbulos blancos), tiene un máximo de 5 unidades y un mínimo de 1.45, la media es de 2.93, los valores del primer, segundo y tercer cuartil son 1.45, 2.8 y 3.49 correspondientemente. Para las variables dicotómicas como el sexo y el indicador de censura (delta), el valor de la media de la tabla anterior refleja, en el caso del sexo, la proporción de hombres en el estudio y en el indicador de censura, el porcentaje de la población que presentó una falla.

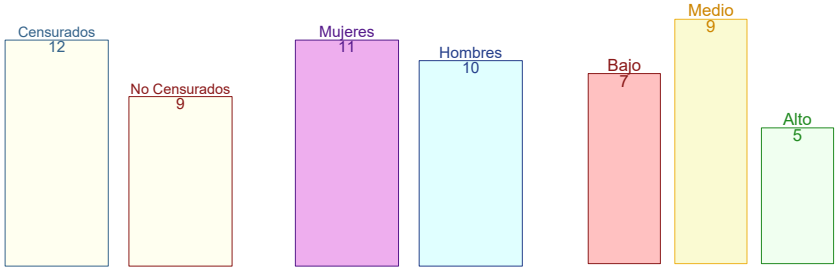
A continuación se presenta un conjunto de gráficas de barras en el cual se refleja, visualmente, el comportamiento de la población en estudio, asimismo, se presentan algunas gráficas de barraras para la población según el tipo de tratamiento suministrado, sus niveles de glóbulos blancos en escala logarítmica y el género.



(a) Indicador de censura (b) Pacientes (c) Niveles de logWBC

Figura 3.1: Estas gráficas de barras sirven para representar visualmente la cantidad total de pacientes en el estudio; a) que se encuentran censurados, b) hombres y mujeres y c) el nivel de glóbulos blancos de los pacientes.

El 71.4 % de la población presentó la falla, es decir, salieron de remisión y por lo tanto, el trastorno de la enfermedad regresa en al menos 30 de las 42 personas en el estudio. La población de mujeres es mayor que la de los hombres, así como la población de pacientes con un registro alto de niveles de glóbulos blancos.

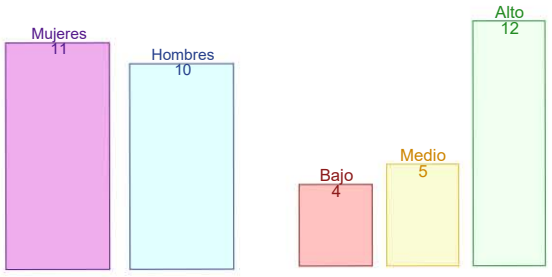


(a) Indicador de censura (b) Pacientes (c) Niveles de logWBC

Figura 3.2: Tratamiento 6-MP.

Una observación interesante es que el todos los pacientes censurados de la muestra fueron tratados con el fármaco 6-MP, esto podría indicar que los pacientes que reciben este suministro tienen una mejora significativa en la enfermedad. Obsérvese también que los pacientes que reciben el tratamiento 6-MP tienen en su mayoría un registro medio de niveles de glóbulos blancos al comienzo del estudio.

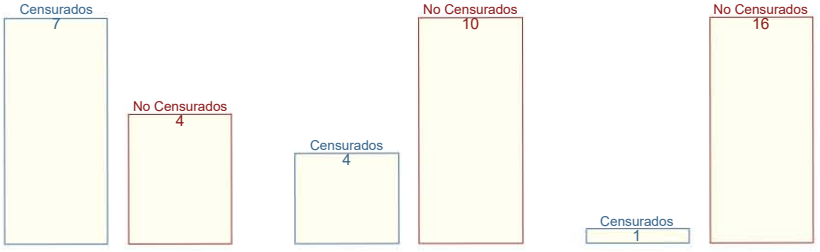
Claramente los pacientes que tomaron el placebo (*véase* Figura 3.3.) tienen un conteo de glóbulos blancos más elevado que los pacientes que recibieron el tratamiento 6-MP (*véase* Figura 3.2). Nótese que la cantidad de mujeres que fueron tratadas con el placebo es igual que la cantidad de mujeres que se tomaron el tratamiento 6-MP, mismo caso para los hombres.



(a) Pacientes (b) Niveles de logWBC

Figura 3.3: Placebo.

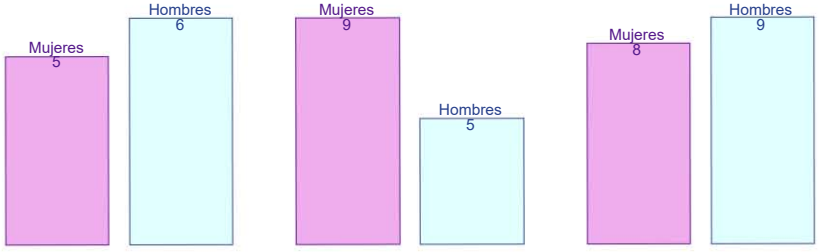
En cuanto a los indicadores de censura se refiere, se puede observar que los pacientes con un nivel alto de logWBC son más propensos a salir del estado de remisión. Por otro lado, las personas con un nivel bajo de logWBC se caraterizan por tener un registro de observaciones censuradas más alto que el de los otros dos niveles de logWBC.



(a) logWBC Bajo (b) logWBC Medio (c) logWBC Alto

Figura 3.4: Indicador de censura por niveles de glóbulos blancos.

En las gráficas de barras de la Figura 3.5, se muestra la proporción de hombres y de mujeres en el estudio según el registro de sus niveles de glóbulos blancos.



(a) logWBC Bajo (b) logWBC Medio (c) logWBC Alto

Figura 3.5: Niveles de glóbulos blancos por género.

En la Figura 3.6 se puede observar que la probabilidad de presentar una falla es más elevada en las primeras 10 semanas del estudio. La curva de distribución ajustada nos permite inferir un modelo de distribución teórico para los datos del ensayo clínico, en este caso podría sugerirse una distribución Weibull, debido a la flexibilidad que presentan las curvas de supervivencia con esta distribución, sin embargo, conviene realizar otro tipo de análisis antes de sugerir este modelo.

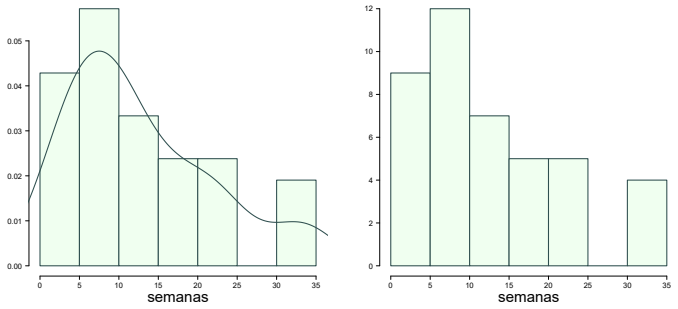
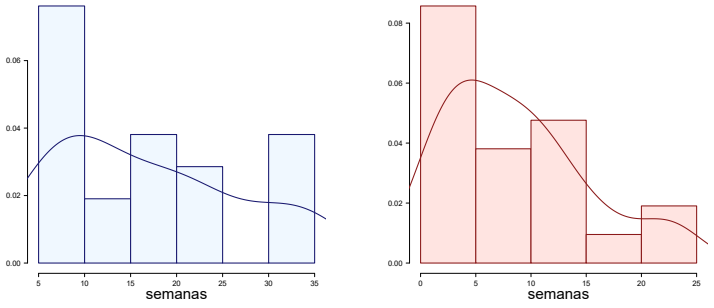
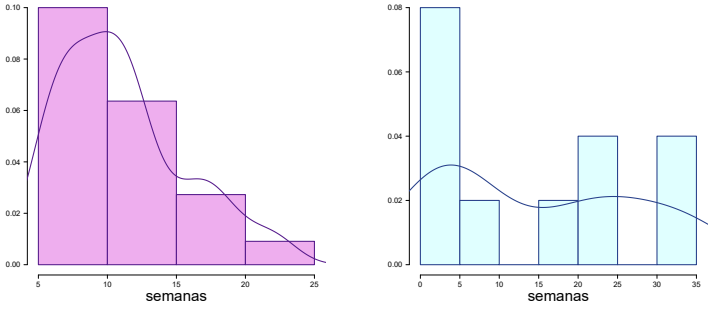


Figura 3.6: Histogramas del tiempo de remisión de pacientes con leucemia aguda.



(a) Tratamiento 6-MP (b) Placebo

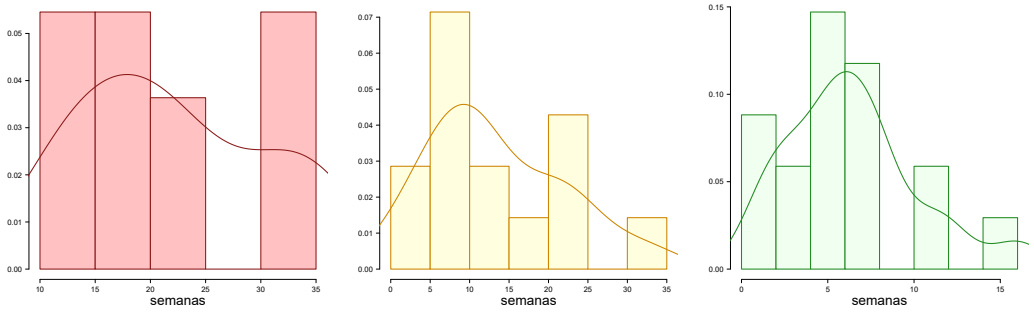


(c) Mujeres (d) Hombres

Figura 3.7: Histogramas

La relación de los tiempos de supervivencia por el tipo de tratamiento y por el género del paciente se pueden observar en la Figura 3.7. Al igual que en el histograma de la Figura 3.6, en estos histogramas se puede observar que la distribución de los tiempos de remisión esta sesgada a la derecha, es decir, es más probable observar ua falla (recaída) en las primeras semanas del ensayo.

Finalmente para concluir este análisis descriptivo se presentan los histogramas del tiempo de remisión según los niveles de glóbulos blancos.



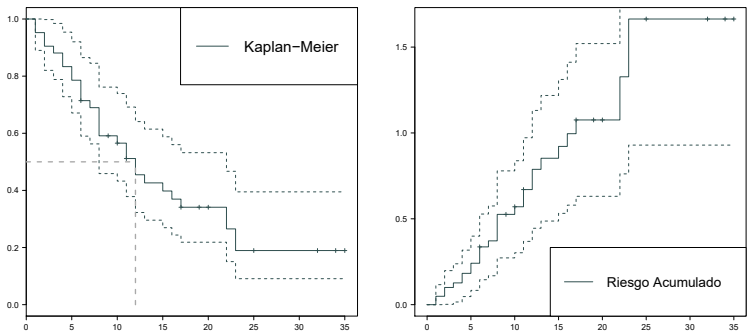
(a) logWBC Bajo (b) logWBC Medio (c) logWBC Alto

Figura 3.8: Histogramas por niveles de glóbulos blancos.

Análisis de Kaplan-Meier y supuestos de distribución y proporcionalidad:

Los valores de $\widehat{S}(t)$, así como las gráficas de supervivencia y riesgo acumulado $\widehat{H}(t)$, utilizando el estimador de Kaplan-Meier, son:

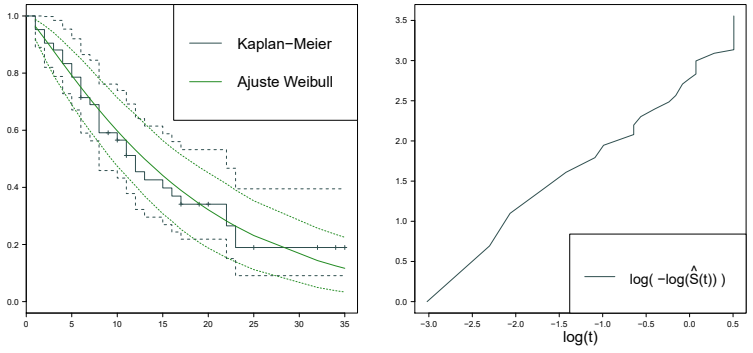
```
km_M <- survfit(Surv(tiempo, delta) ~ 1)
summary(km_M)

Call: survfit(formula = Surv(tiempo, delta) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     42       2    0.952  0.0329        0.8901        1.000
    2     40       2    0.905  0.0453        0.8202        0.998
    3     38       1    0.881  0.0500        0.7883        0.985
    4     37       2    0.833  0.0575        0.7279        0.954
    5     35       2    0.786  0.0633        0.6709        0.920
    6     33       3    0.714  0.0697        0.5899        0.865
    7     29       1    0.690  0.0715        0.5628        0.845
    8     28       4    0.591  0.0764        0.4588        0.762
   10     23       1    0.565  0.0773        0.4325        0.739
   11     21       2    0.512  0.0788        0.3783        0.692
   12     18       2    0.455  0.0796        0.3227        0.641
   13     16       1    0.426  0.0795        0.2958        0.615
   15     15       1    0.398  0.0791        0.2694        0.588
   16     14       1    0.369  0.0784        0.2437        0.560
   17     13       1    0.341  0.0774        0.2186        0.532
   22      9       2    0.265  0.0765        0.1507        0.467
   23      7       2    0.189  0.0710        0.0909        0.395
```



(a) $\widehat{S}(t)$ (b) $\widehat{H}(t)$

Figura 3.9: En $\widehat{S}(t)$ se pueden observar las bandas de confianza del estimador de Kaplan-Meier y el tiempo mediano de supervivencia de los pacientes en el estudio.

Como el objetivo es ajustar un modelo paramétrico, se compara la curva del estimador de Kaplan-Meier con la función de supervivencia de alguna distribución teórica, asimismo, se realiza un método de diagnóstico de distribución utilizando la gráfica de probabilidad de alguna de las distribuciones revisadas con anterioridad.



(a) Curvas de supervivencia (b) Gráfica de probabilidad

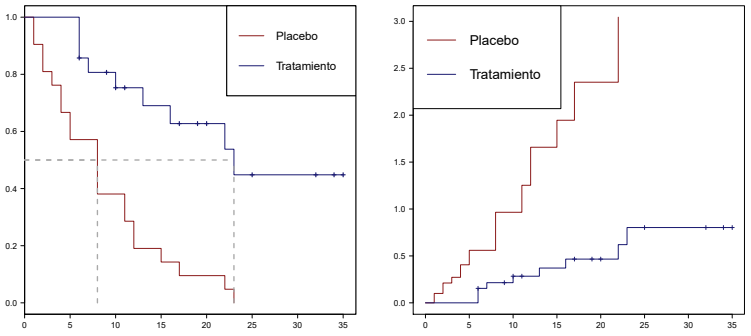
Figura 3.10: Métodos gráficos: Diagnóstico de distribución Weibull.

En la Figura 3.10 se puede observar que las curva de supervivencia con distribución Weibull es un buen candidato para ajustar a la curva del estimador de Kaplan-Meier, asimismo, la gráfica de probabilidad (Figura 3.10 b) tiene un comportamiento aproximadamente lineal, por lo tanto, no se rechaza la hipótesis de distribución Weibull para lo datos de supervivencia de pacientes con leucemia aguda.

La curva de supervivencia de la Figura 3.10 a), es el resultado de ajustar un modelo de distribución Weibull para los datos de supervivencia sin incluir ninguna covariable en la descripción del tiempo de falla. Debido a que el objetivo del estudio es analizar el efecto del tipo de tratamiento aunado con los posibles valores del logWBC y su relación con el género, conviene analizar las curvas de Kaplan-Meier y realizar diagnósticos de distribución para cada una de estas covariables.

Una condición necesaria en el modelo de riesgos proporcionales, es verificar que no se rechaza el supuesto de proporcionalidad para las covariables que se pretenden incluir en el modelo de regresión. Gráficamente el supuesto de proporcionalidad se puede diagnosticar graficando las curvas[2] $\mathbf{log}(-\mathbf{log}(\,\mathbf{S}(\mathbf{t})\,))$ de las distintas categorías en cada una de las covariables, si estas curvas son paralelas entonces no se rechaza el supuesto de proporcionalidad.

En el caso de los modelos paramétricos, se grafican las gráficas de probabilidad de las distintas categorías, si estas gráficas se mantienen paralelas y además tienen un comportamiento aproximadamente lineal, entonces, no se rechaza el supuesto de proporcionalidad ni el supuesto de la distribución subyacente.



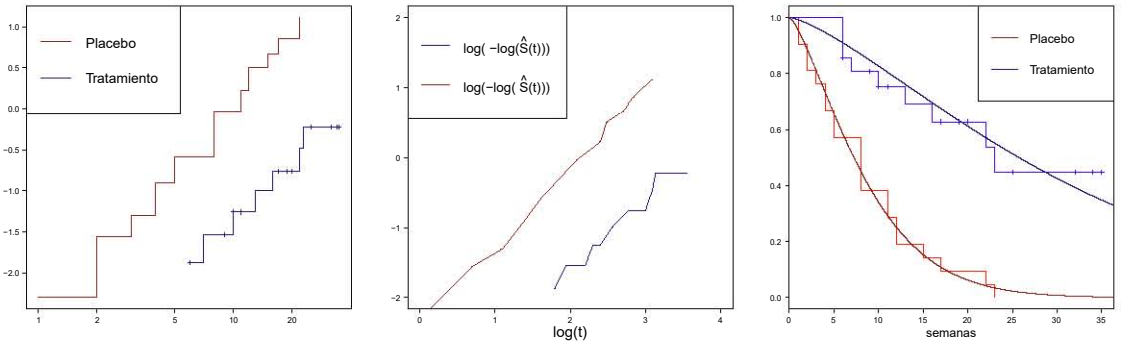
(a) Kaplan-Meier (b) Riesgo Acumulado

Figura 3.11: Análisis por el tipo de tratamiento.

El comportamiento paralelo en las Figuras 3.11, nos ayuda a plantear una hipótesis de proporcionalidad, esta hipótesis se prueba gráficamente en la Figura 3.12. En la Figura 3.11 a) se puede observar que la curva de supervivencia de los pacientes que reciben el tratamiento 6-MP es considerablemente más alta que la curva de supervivencia de aquellos pacientes que toman el placebo, esto indica que el tratamiento 6-MP es más eficiente que el placebo, aunque esto último se verifica más adelante con ayuda de la prueba *log-rank*. Obsérvese también, que a medida de que pasan las semanas, las curvas de supervivencia se encuentran cada vez más separadas, lo que sugiere que los efectos de haber recibido el tratamiento son más notorios en las últimas semanas del estudio.

Las gráficas a) y b) de la Figura 3.12 se mantienen paralelas aproximadamente después de la semana diez, por lo tanto, no se rechaza el supuesto de proporcionalidad, asimismo, en b) las gráficas para el tratamiento y el placebo siguen un comportamiento lineal, por lo tanto, no se rechaza el supuesto de distribución Weibull. En c) el modelo Weibull se ajusta bien a las curvas de Kaplan-Meier para el tipo de tratamiento, por lo tanto, el diagnóstico de distribución Weibull parece razonabe.

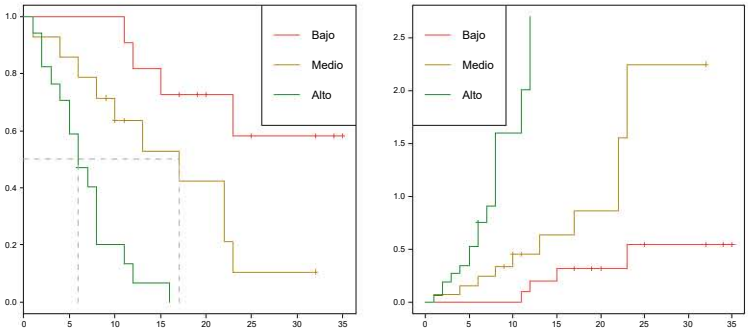
[2] $S(t)$ el estimador de Kaplan-Meier



(a) $log(-log(\,S(t)\,))$ (b) Gráficas de probabilidad (c) Ajuste Weibull

Figura 3.12: Diagnósticos de proporcionalidad y distribución.

En particular la variable logWBC es una variable continua en el modelo, si se desea revisar el efecto de proporcionalidad para este tipo de variables es necesario reescribirlas como variables categóricas, preferentemente en grupos pequeños (2 o 3 categorías). La gráfica del estimador de Kaplan-Meier y del riesgo acumulado se muestran en la Figura 3.13.

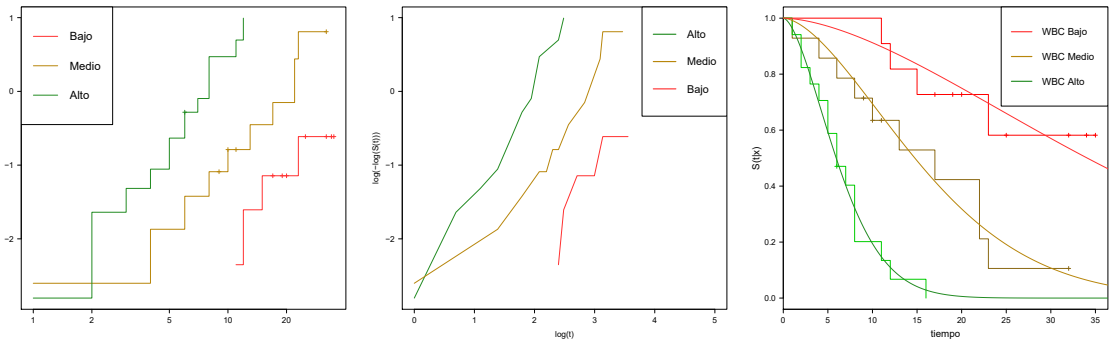


(a) Kaplan-Meier (b) Riesgo Acumulado

Figura 3.13: Análisis por el nivel de glóbulos blancos.

Las gráficas de la Figura 3.14 a) y b) se intersectan en las primeras semanas del estudio, es decir, se viola el supuesto de proporcionalidad, sin embargo, esto ocurre al principio del ensayo y conforme pasa el tiempo las gráficas se mantiene paralelas (a partir de la semana diez), por lo tanto, se decide no rechazar el supuesto de proporcionalidad. Por otro lado, en b) las gráficas para los distintos niveles de glóbulos blancos son aproximadamente lineales[3] y en c) las curvas con distribución Weibull se ajustan de manera adecuada a las curvas de Kaplan-Meier para los distintos niveles de glóbulos blancos, por lo que no se rechaza el supuesto de distribución Weibull para la covariable logWBC.

[3]Si la pendiente de estas gráficas es igual a la unidad, se puede considerar un modelo exponencial para ajustar a los datos de supervivencia.



(a) $log(-log(\,S(t)\,))$ (b) Gráficas de probabilidad (c) Ajuste Weibull

Figura 3.14: Diagnósticos de proporcionalidad y distribución.

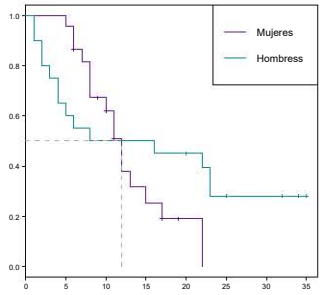
En el caso de la variable sexo, obsérvese que en todas las gráficas de la Figura 3.15, las curvas de supervivencia, riesgo acumulado y las gráficas de diagnóstico de proporcionalidad c) y d), se intersectan en algún punto del tiempo, por lo tanto, no se mantienen paralelas y se rechaza el supuesto de proporcionalidad, no obstante, no existe suficiente evidencia para rechazar la hipótesis de distribución Weibull, pues las gráficas en d) siguen aproximadamente un comportamiento lineal.

Una vez revisados los supuesto de proporcionalidad y de distribución para cada una de las variables en el modelo, nuevamente, se debe de realizar este análisis de manera simultánea para estas variables (particularmente para aquellas variables en las que no se rechaza la hipótesis de riesgos proporcionales), sin embargo, este análisis no es apropiado si se dispone de pocos datos.

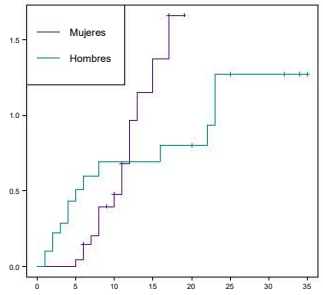
De las gráficas anteriores se puede exhibir que existe una diferencia significativa entre los pacientes que reciben el tratamiento 6-MP y el placebo, así como aquellos pacientes que tienen un nivel de glóbulos blancos alto, medio y bajo, sin embargo, conviene realizar una prueba estadística, como la prueba de *log-rank* o la prueba de *Peto y Peto*, para rechazar la hipótesis nula de igualdad de supervivencias.

En la Tabla A de la página 123, se utiliza la prueba de *log-rank* en **R**, con el propósito de probar la igualdad de supervivencia entre los tipos de tratamiento, así como la igualdad de la supervivencia entre los distintos niveles de góbulos blancos. En ambas pruebas se obtiene un *p-value* menor que el nivel de significancia[4] ($\alpha = 5\,\%$), por lo tanto, se rechaza la hipótesis nula de igualdad de supervivencias, es decir, existe una diferencia significativa entre los tipos tratamiento (*p-value* = 4.17e-05) y entre los distinto niveles de logWBC (*p-value* = 1.86e-06).

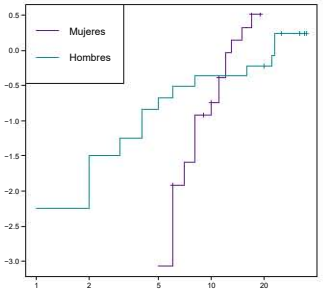
[4]En **R** se utiliza un nivel de significancia del 5 %



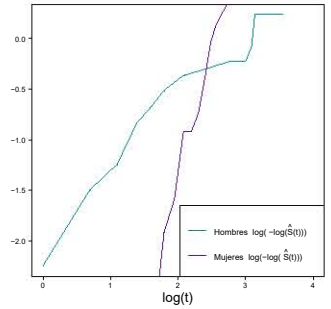
(a) Kaplan-Meier



(b) Riesgo Acumulado



(c) $log(-log(\,S(t)\,))$



(d) Gráficas de probabilidad

Figura 3.15: Análisis por género.

```
# Prueba para la diferencia entre los tipos de tratamiento
survdiff(Surv(tiempo,delta)~trat,rho=0)
Call:
survdiff(formula = Surv(tiempo, delta) ~ trat, rho = 0)

        N Observed Expected (O-E)^2/E (O-E)^2/V
trat=0 21       21     10.7      9.77      16.8
trat=1 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05

 # p-value = 4.17e-05

# Prueba para la diferencia: logWBC
survdiff(Surv(tiempo,delta)~lwbc_f,rho=0)
Call:
survdiff(formula = Surv(tiempo, delta) ~ lwbc_f, rho = 0)

           N Observed Expected (O-E)^2/E (O-E)^2/V
lwbc_f=0 11        4    13.06    6.2880   12.7695
lwbc_f=1 14       10    10.72    0.0489    0.0809
lwbc_f=2 17       16     6.21   15.4173   23.1040

 Chisq= 26.4  on 2 degrees of freedom, p= 1.86e-06
 # p-value = 1.86e-06
```

Tabla A: Pruebas de *Log-Rank*.

Modelos de riesgos proporcionales de Cox:

La manera más sencilla de ajustar un modelo paramétrico de regresión de PH, es a partir del ajuste de un modelo semiparamétrico de riesgos proporcionales, si el modelo semiparamétrico de PH proporciona un buen ajuste, entonces, es posible que este modelo proporcione un buen ajuste para un modelo paramétrico de riesgos proporcionales.

El principal objetivo es utilizar como variable regresora el tipo de tratamiento y determinar el posible efecto de las covariables logWBC y/o el sexo, así como una posible interacción entre estas.

Con ayuda de la función **coxph()** de **R**, se prueban los siguientes modelos de riesgo con el propósito de descartar aquellas variables que no sean significativas[5] en la descripción del tiempo de supervivencia[6]:

- **Modelo 1:** $h(t \mid \boldsymbol{X}) = h_0(t)\, exp(\beta\, Tratamiento)$
- **Modelo 2:** $h(t \mid \boldsymbol{X}) = h_0(t)\, exp(\beta\, Sexo)$
- **Modelo 3:** $h(t \mid \boldsymbol{X}) = h_0(t)\, exp(\beta\, logWBC)$

Modelo 1:

```
coxph(formula = Surv(tiempo, delta) ~ trat)


       coef exp(coef) se(coef)     z       p
trat -1.572     0.208    0.412 -3.81 0.00014

Likelihood ratio test=16.4  on 1 df, p=5.26e-05
n= 42, number of events= 30

# El p-value es: 0.00014
```

En la salida anterior, se puede observar que el tipo de tratamiento tiene un *p-value* menor que el nivel de significancia, esto rechaza la hipótesis nula $H_0 : \beta = 0$ y por lo tanto, el tipo de tratamiento es significativo en el modelo.

En el **Modelo 2**, se obtiene un *p-value* $= 0.44$, por lo tanto, no se rachaza la hipótesis nula $H_0 : \beta = 0$, es decir, la variable sexo no es significativa en el modelo.

La salida del **Modelo 3**, muestra un *p-value* < 0.05, por lo tanto, la variable logWBC también es significativa en el modelo.

[5]Recuérdese que la **prueba de Wald** nos ayuda a determinar que variable o que variables son significativas en el modelo de regresión (*véase* pag 85-86).

[6]Del análisis realizado para las curvas de supervivencia y de los diagnósticos de proporcionalidad, se puede inferir que el sexo no es una variable significativa en el modelo de riesgos proporcionales.

Modelo 2:

```
coxph(formula = Surv(tiempo, delta) ~ sex)


      coef exp(coef) se(coef)     z    p
sex -0.310     0.733    0.404 -0.77 0.44

Likelihood ratio test=0.6  on 1 df, p=0.44
n= 42, number of events= 30

# El p-value  es:  0.44
```

Modelo 3:

```
coxph(formula = Surv(tiempo, delta) ~ lwbc)


      coef exp(coef) se(coef)    z       p
lwbc 1.646     5.188    0.298 5.53 3.3e-08

Likelihood ratio test=34.8  on 1 df, p=3.58e-09
n= 42, number of events= 30

# El p-value es: 3.3e-08
```

Utilizando las variables significativas en el modelo, se prueba el **Modelo 4** y el **Modelo 5**, en particular este último se considera una posible interacción entre el tipo de tratamiento y el logaritmo de los glóbulos blancos, mientras que en el modelo 4 simplemente se considera el efecto de incluir la variable logWBC en el modelo cuando se utiliza como variable regresora el tipo de tratamiento.

- **Modelo 4:** $h(t \mid \boldsymbol{X}) = h_0(t)\, exp(\beta_1\, Tratamiento + \beta_2\, logWBC)$
- **Modelo 5:** $h(t \mid \boldsymbol{X}) = h_0(t)\, exp(\beta_1\, Trat + \beta_2\, logWBC + \beta_{12}\, Trat \times logWBC)$

Modelo 4:

```
coxph(formula = Surv(tiempo, delta) ~ trat + lwbc)

       coef exp(coef) se(coef)     z       p
trat -1.386     0.250    0.425 -3.26  0.0011
lwbc  1.691     5.424    0.336  5.03 4.8e-07

Likelihood ratio test=46.7  on 2 df, p=7.19e-11
n= 42, number of events= 30
```

De acuerdo con los valores del *p-value* se puede decir que el tipo de tratamiento y el *logWBC* siguen siendo varibles significativas en este modelo (**Modelo 4**), sin embargo, en el **Modelo 5**, el tipo de tratamiento y la interacción entre este y la variable *logWBC* no son significativas, por lo tanto, descartamos el **Modelo 5**.

Modelo 5:

```
coxph(formula = Surv(tiempo, delta) ~ trat * lwbc)


              coef exp(coef) se(coef)     z        p
trat        -2.375     0.093    1.706 -1.39     0.16
lwbc         1.555     4.735    0.399  3.90 9.6e-05
trat:lwbc    0.318     1.374    0.526  0.60     0.55

Likelihood ratio test=47.1  on 3 df, p=3.36e-10
n= 42, number of events= 30
```

De los resultados anteriores, se puede concluir que las covariables *tratamiento* y $logWBC$ tienen un efecto significativo en el tiempo de supervivencia. Los Modelos 1, 3 y 4 parecen ajustar de manera adecuada el tiempo de remisión para los pacientes con leucemia aguda, sin embargo, si se busca seleccionar el mejor modelo se debe calcular el **AIC** (Criterio de Información de Akaike[7]) para cada uno de los modelos. En la mayoría de los casos se escoge el modelo con menor **AIC**, aunque no en todos los casos éste proporciona el mejor ajuste.

El valor del **AIC** se puede obtener en **R** con el el siguiente comando:
extractAIC(modelo**)[2]**

```
# Modelo 1:
extractAIC(cphm.t)[2]
[1] 172.0168

# Modelo 3:
extractAIC(cphm.g)[2]
[1] 153.527

# Modelo 4:
extractAIC(cphm.tg)[2]
[1] 143.6562
```

El modelo con el menor AIC, es el **Modelo 4**, por lo tanto, se considera que éste es el mejor modelo.

Los modelos paramétricos y semiparamétricos de riesgos proporcionales deben de verificar el supuesto de proporcionalidad. Anteriormente, se utilizaron métodos gráficos para probar este supuesto, sin embargo, conviene realizar una prueba de bondad de ajuste para comprobar la hipótesis de proporcionalidad.

Usualmente, los residuales de Schoenfeld [8] se utilizan para verificar que el coeficiente de regresión de cada covariable en el modelo cumple el supuesto de proporcionalidad, la salida del comando **cox.zph(**modelo**)** en **R** nos ayuda a rechazar o no

[7] **AIC** (Akaike information criterion) para mayor información consulte [1]

[8] El tema de los residuales de Schoenfeld no se expone en esta tesis, si se desea consultar información relacionada puede consultarse [1].

rechazar la hipótesis nula del supuesto de proporcionalidad.

```
#  Salida del comando: cox.zph()
cox.zph(cphm.tg)

            rho     chisq      p
trat    -0.00451 0.000542 0.981
lwbc     0.02764 0.034455 0.853
GLOBAL        NA 0.034529 0.983
```

En la salida anterior se puede observar un *p-value* > 0.05 para las covariables *tratamiento*, $logWBC$ y el ajuste global, por lo tanto, no se rechaza el supuesto de proporcionalidad en el **Modelo 4**.

Finalmente, decimos que el **Modelo 4**, es el mejor candidato de los modelos semiparamétricos de PH para ajustar el riesgo de remisión de los pacientes con leucemia aguda. De acuerdo con los valores de los parámetros este modelo está dado por las siguientes expresiones:

Placebo:

$$h(t|\,\boldsymbol{X}) = h_0(t)\, exp(1.691\, logWBC)$$

Tratamiento 6-MP:

$$h(t|\,\boldsymbol{X}) = h_0(t)\, exp(-1.386\, Tratamiento + 1.691\, logWBC)$$

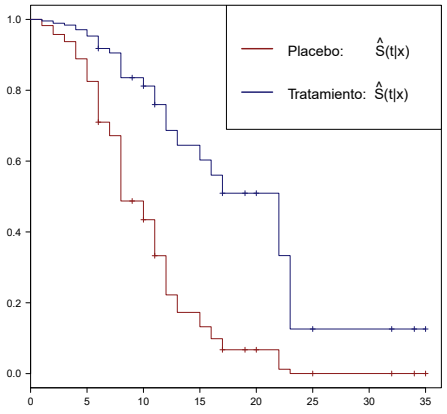


Figura 3.16: Ajuste de Cox.

Ajuste de modelos paramétricos de regresión:

Partiendo del modelo seleccionado para el ajuste semiparamétrico y considerando que se puede ajustar un modelo de distribución Weibull a los datos de supervivencia, supóngase que el modelo de riesgo (3.1) es hasta ahora, el mejor candidato para representar el riesgo de salir de remisión para los pacientes con leucemia aguda.

$$h(t|\,\boldsymbol{X}) = \lambda\,\gamma\,t^{\gamma-1}\,exp(\beta_1\,Tratamiento + \beta_2\,logWBC), \tag{3.1}$$

A continuación, con ayuda de **R**, se calculan los coeficientes de regresión de la expresión (3.1), asimismo, se verifica si estos parámetros son significativos en el modelo paramétrico.

```
# Ajuste con el comando phreg()

phreg(formula = Surv(tiempo, delta) ~ trat + lwbc)

Covariate          W.mean      Coef Exp(Coef)  se(Coef)     Wald p
trat                0.664    -1.456     0.233     0.414     0.000
lwbc                2.482     1.784     5.954     0.338     0.000

log(scale)                    4.740                0.368     0.000
log(shape)                    0.793                0.142     0.000
```

Tabla 3.A: Ajuste paramétrico con el comando phreg(·)

La columna **Wald p** de la tabla anterior, muestra los valores del *p-value* después de llevar a cabo la prueba de Wald, evidentemente, el *p-value* es considerablemente pequeño y menor que el nivel de significancia de la prueba ($\alpha = 5\,\%$) para cada una de las variables en el modelo, por lo tanto; el tratamiento, el logWBC y los parámetros γ y λ en el modelo (los últimos dos en su forma reparametrizada) son significativos en el modelo.

```
survreg(formula = Surv(tiempo, delta) ~ trat + lwbc, dist = "weibull")
             Value Std. Error     z        p
(Intercept)  4.740      0.368 12.87 6.62e-38
trat         0.659      0.189  3.49 4.90e-04
lwbc        -0.807      0.108 -7.45 9.68e-14
Log(scale)  -0.793      0.142 -5.58 2.45e-08

Scale= 0.452

Weibull distribution
Loglik(model)= -90.1   Loglik(intercept only)= -116.4
        Chisq= 52.68 on 2 degrees of freedom, p= 3.6e-12
Number of Newton-Raphson Iterations: 7
n= 42
```

Tabla 3.B: Ajuste paramétrico con el comando survreg(·).

De la Tabla 3.A, se pueden obtener directamente los valores de los coeficientes β_1 y β_2. Los estimadores $\widehat{\gamma}$ y $\widehat{\lambda}$ pueden ser calculados a partir de la Tabla 3.B utilizando la reparametrización (2.45) (*véase* pag 101 de este documento).

En la siguiente tabla se muestra: 1) el valor de los estimadores $\widehat{\gamma}$, $\widehat{\lambda}$, $\widehat{\beta}_1$ y $\widehat{\beta}_2$, 2) el exponente del coeficiente de regresión, 3) el error estándar de cada una las covariables del modelo y 4) los correspondientes intervalos de confianza.

Parámetro	**Estimador**	$\widehat{exp(\cdot)}$	$\widehat{s.e(\cdot)}$	L.95 %	U.95 %
γ	2.210	9.115	0.314	1.594	2.826
β_1	-1.456	0.233	0.414	-2.268	-0.644
β_2	1.784	5.953	0.337	1.122	2.445
λ	2.82e-05	1	4.82e-05	-6.624e-05	1.122e-04

Tabla 3.C: Resultados del modelo Weibull.

Con los resultados anteriores, se comprueba que el modelo paramétrico de la expresión (3.1), es un buen candidato para representar el riesgo de salir de remisión para los pacientes con leucemia aguda, sin embargo, resta verificar, sí, en efecto, éste es el mejor modelo, es decir, aquel con el menor AIC.

Considérese los siguientes modelos paramétricos de regresión Weibull[9]:

- **Modelo I:** $h(t \mid \boldsymbol{X}) = h_0(t) = \lambda\,\gamma\,t^{\gamma-1}$
- **Modelo II:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta\,Tratamiento)$
- **Modelo III:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta\,logWBC)$
- **Modelo IV:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta_1\,Tratamiento + \beta_2\,Sexo)$
- **Modelo V:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta_1\,Tratamiento + \beta_2\,logWBC)$
- **Modelo VI:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta_1\,Tratamiento + \beta_2\,logWBC + \beta_3\,Sexo)$
- **Modelo VII:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta_1\,Trat + \beta_2\,Sexo + \beta_{12}\,Trat\ \times\ Sexo)$
- **Modelo VIII:** $h(t \mid \boldsymbol{X}) = h_0(t)\,exp(\beta_1\,Trat + \beta_2\,logWBC + \beta_{12}\,Trat \times logWBC)$

Los valores de la Tabla 3.D, muestran que el **Modelo V**, mismo que corresponde al de la expresión (3.1), tiene el menor AIC, este resultado sugiere que el **Modelo V** es el más apropiado.

[9]En estos modelos no se calcula la significancia de los coeficientes de regresión, el interés se centra únicamente en cálculo del AIC.

Modelos	**AIC**
Modelo I	236.81
Modelo II	219.15
Modelo III	199.79
Modelo IV	220.74
Modelo V	**188.12**
Modelo VI	189.39
Modelo VII	215.25
Modelo VIII	189.80

Tabla 3.D

Análogamente, si se desea comparar el modelo **Modelo V** con el mismo modelo pero con distribución exponencial, se puede calcular el **AIC**.

Las salidas correspondientes de los coeficientes de regresión del modelo exponencial: $h(t \mid \boldsymbol{X}) = h_0(t)\, exp(\beta_1 Trat + \beta_2 logWBC)$ son:

```
# Ajuste con el comando phreg()
phreg(formula = Surv(tiempo, delta) ~ trat + lwbc, shape = 1)

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
trat                0.664    -1.093     0.335     0.413     0.008
lwbc                2.482     0.884     2.422     0.216     0.000

log(scale)                    4.865                 0.756     0.000

# Valor del AIC
extractAIC(eph.tg)[2]
207.5473
```

Tabla 3.E: Obsérvese que los coeficientes de regresión en el modelo exponencial también son significativos.

En particular el **AIC** del modelo exponencial es: **207.54**. Claramente, el **Modelo V** con distribución exponencial, tiene un AIC mayor que el del **Modelo V** con distribución Weibull (AIC = **188.12**) por lo tanto, se sigue considerando al modelo Weibull como el más apropiado.

Parámetro	**Estimador**	$exp(\cdot)$	$s.e\widehat{(\cdot)}$	L.95 %	U.95 %
λ	0.007	1.007	0.005	-0.003	0.019
β_1	-1.093	0.335	0.413	-1.903	-0.283
β_2	0.884	2.421	0.215	0.461	1.307

Tabla 3.F: Resultados del modelo exponencial[10].

[10]Los cálculos de esta tabla se pueden encontrar en el Apéndice de códigos, en este caso era necesario encontrar el valor de $\widehat{\lambda}$, pues en la Tabla 3.E, se muestra un valor reparametrizado de $\widehat{\lambda}$.

Resultados Gráficos e interpretación del modelo paramétrico:

Los modelos de riesgo para los datos de remisión de los pacientes con leucemia aguda son:

$$Plecebo: \qquad h\left(t \mid \boldsymbol{X}\right) = 2.82\text{e-}05 \times 2.21\, t^{1.21}\, exp\left(1.784\, logWBC\right)$$

$$Tratamiento: h\left(t \mid \boldsymbol{X}\right) = 2.82\text{e-}05 \times 2.21\, t^{1.21}\, exp\left(-1.456 + 1.784\, logWBC\right)$$

Evidentemente, el riesgo para los pacientes en el grupo placebo aumenta más rápido que el riesgo para los pacientes que fueron tratados con el fármaco 6-MP.

De la tercer columna de la **Tabla 3.C** se puede decir que, para valores fijos de la variable *logWBC*, los pacientes que reciben el tratamiento 6-MP tienen **0.233** veces el riesgo de los pacientes que reciben el placebo, es decir, el riesgo de salir de remisión para los pacientes que toman el placebo se incrementa en $\mathbf{(0.233)^{-1} = 3.144}$ unidades. Por otro lado, para pacientes con el mismo tipo de tratamiento, el riesgo es **5.95** más elevado por unidad de cambio en la variable $logWBC$, es decir, cantidades altas de glóbulos blancos implican riesgos más elevados.

Considerando un nivel de glóbulos blancos en escala logarítmica de **2.5** unidades se tiene que, el tiempo mediano de supervivencia estimado para los pacientes que reciben el placebo es de: **12.88** semanas. Mientras que el tiempo mediano de supervivencia estimado para los pacientes que reciben el tratamiento 6-MP, con el mismo nivel de glóbulos blancos, es de: **24.90** semanas.[11]

De acuerdo con los valores que se muestran en las tablas **Modelo 4** y **Tabla 3.C**, correspondientes al modelo de Cox y al modelo Weibull respectivamente, el riesgo de los pacientes que reciben el placebo es más bajo en el modelo paramétrico (***riesgo*** **= 0.233**) que en el modelo de regresión de Cox (***riesgo*** **= 0.25**), contrariamente, el riesgo por incremento de glóbulos blancos es más bajo en el modelo semiparamétrico (***riesgo*** **= 5.424**) que en el modelo Weibull (***riesgo*** **= 5.95**).

Otro resultado que se puede obtener directamente de la **Tabla 3.C** y de las salidas del ajuste semiparamétrico **Modelo 4**, es la longitud del intervalo de confianza para las covariables *tratamiento* y *logWBC*. En el modelo paramétrico la longitud del intervalo de confianza para la covariable *tratamiento* es de **1.623**, mientras que en el modelo semiparamétrico la longitud del intervalo es de **1.665**. Para la covariable *logWBC* la longitud del intervalo es de **1.323** en el modelo paramétrico y de **1.316** en el modelo semiparamétrico.[11]

Aunque se esperaba mayor precisión en el modelo de riesgos proporcionales Weibull con respecto al modelo de Cox, éste último proporcionó resultados muy parecidos a los obtenidos en el modelo paramétrico. Debido a que la longitud del intervalo

[11] Los cálculos de este párrafo fueron realizados con ayuda del software estadístico **R** y pueden ser consultados en la página 177 del Apéndice A.

de confianza para el tipo de tratamiento (la cual es la variable de mayor interés en el estudio) es menor en el modelo Weibull, se decide seleccionar a este modelo como el mejor ajuste para los datos de supervivencia.

A continuación se muestran algunas gráficas relacionadas con las funciones de riesgo y supervivencia del modelo paramétrico de riesgos proporcionales Weibull.

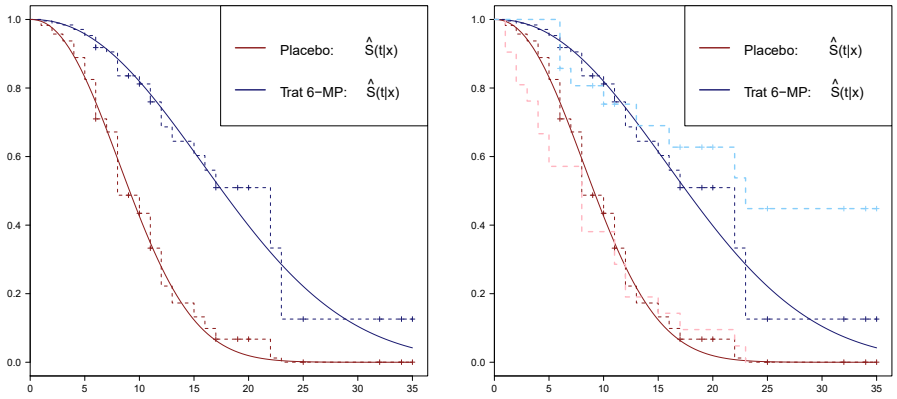


Figura 3.17: En la gráfica de la izquierda se comparan la curvas de supervivencia del modelo Weibull, con las curvas de supervivencia del modelo de Cox. En el lado derecho las líneas punteadas corresponden a las gráficas del estimador de Kaplan-Meier según el tipo de tratamiento suministrado.

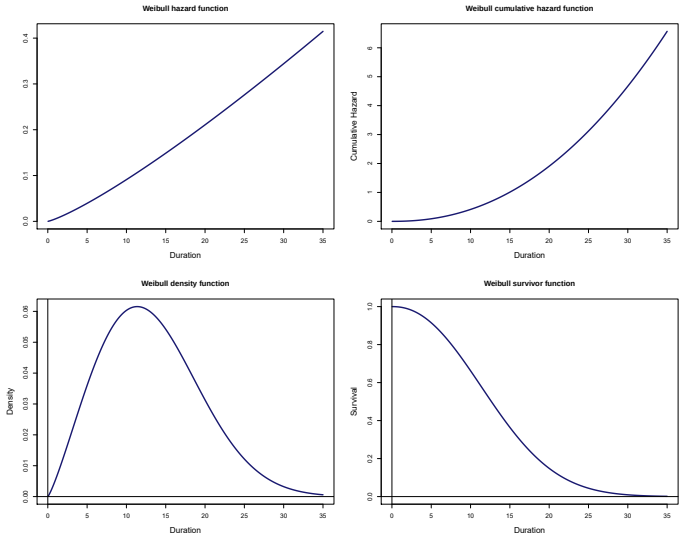


Figura 3.18: Funciones relacionadas con el modelo Weibull. a) Función de riesgo, b) Función de riesgo acumulado, c) Función de densidad, d) Función de supervivencia.

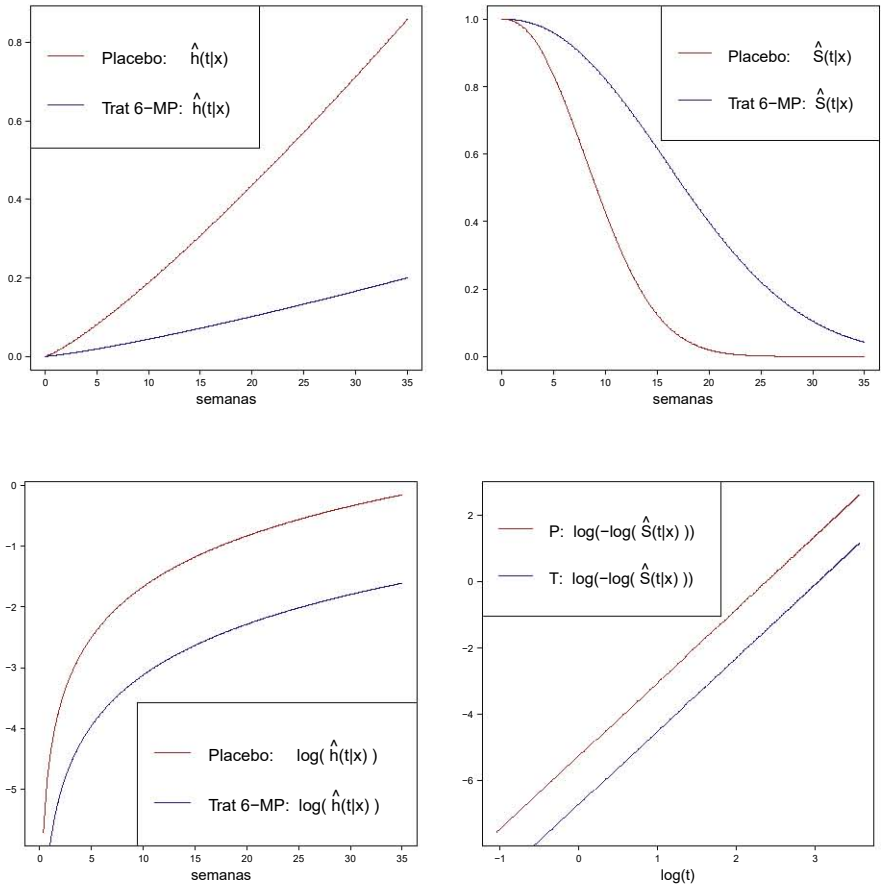


Figura 3.19: Curvas de riesgo y supervivencia del modelo Weibull según el tipo de tratamiento. a) Funciones de riesgo, b) Funciones de supervivencia, c) Funciones de $log(\,riesgo\,)$, d) Funciones de $log\,(\,-log\,(\,supervivencia\,)\,)$.

Comentarios Finales

Los modelos paramétricos de regresión en el análisis de supervivencia tienden a proporcionar resultados más precisos con relación al tiempo de falla, siempre y cuando se haya elegido un modelo paramétrico adecuado, sin embargo, en el capítulo III se puede observar que los resultados que se obtienen de los modelos paramétricos no se alejan mucho de los resultados que se pueden obtener cuando se ajusta un modelo semiparamétrico de regresión a los datos de supervivencia, por lo que estos últimos son más utilizados en la práctica, especialmente cuando se cuenta con muchos datos.

Uno de los objetivos por los que se desarrolló esta tesis, fue para ser utilizada como material de apoyo para aquellos estudiantes interesados en aprender o que cursen una asignatura relacionada con el análisis de datos de supervivencia, motivo por el cual se desarrollaron varios ejemplos a lo largo del documento y se incluyeron los códigos utilizados en el software estadístico R, de manera que si el estudiante desea reproducir algún resultado tenga la disponibilidad de hacerlo.

Asimismo, en el capítulo 3, se desarrolló paso a paso el procedimiento para ajustar un modelo paramétrico y semiparamétrico de riesgos proporcionales, de tal forma que este capítulo sirva de guía para aquellas personas interesadas en ajustar alguno de estos modelos de regresión a los datos de supervivencia.

Apéndice A

Códigos de R

En el siguiente apéndice se muestran los códigos que se generaron en R para producir todas las gráficas, funciones y resultados obtenidos que aparecen en el documento.

A.1. Códigos Capítulo 1

Figura 1.1:

```
gt1 <- seq(2001,2009,length=100)
gt2 <- seq(2004,2011,length=100)
gt3 <- seq(2004,2012,length=100)
gt4 <- seq(2003,2008,length=100)
gt5 <- seq(2002,2006,length=100)
gt6 <- seq(2003,2014,length=100)

plot(gt1,gt1/gt1,type="l",xlab="",ylab="",xlim=c(2000,2013),
     ylim=c(0,8),col="navy",lwd=2,axes=F)
text(2008,6.8,"Tiempos de Origen",cex=1.4)

lines(gt2,2*gt2/gt2,col="navy",lwd=2)
lines(gt3,3*gt3/gt3,col="navy",lwd=2)
lines(gt4,4*gt4/gt4,col="navy",lwd=2)
lines(gt5,5*gt5/gt5,col="navy",lwd=2)
lines(gt6,5.87*gt6/gt6)
lines(c(2012,2012),c(0,5.87),lty=2,lwd=2)

points(2009,1,col="red",pch=4,lwd=3)
points(2011,2,col="red",pch=4,lwd=3)
points(2008,4,col="red",pch=4,lwd=3)
points(2006,5,col="red",pch=4,lwd=3)
legend("topleft",c("Seguimiento","Fin del estudio","Falla"),
       lty=c(1,2,NA),pch=c(NA,NA,4),lwd=2,col=c("navy","black","red"))
axis(side=1,at=c(2000,2004,2008,2012))
axis(side=2,at=c(1,2,3,4,5),labels=c("E","D","C","B","A"))
box()

# Nota: Las proporciones de esta figura podr'i'an variar su aspecto
# en funci'o'n de las dimensiones de la pantalla de su computadora.
```

Figura 1.2:

```
fm1 <- seq(0,9,length=100)
fm2 <- seq(0,12,length=100)
fm3 <- seq(0,3,length=100)
fm4 <- seq(0,5,length=100)
fm5 <- seq(0,12,length=100)
fm6 <- seq(0,7,length=100)
fm7 <- seq(0,14,length=100)

plot(fm1,fm1/fm1,type="l",xlab="",ylab="",xlim=c(0,13),
     las=1,ylim=c(0,9),col="navy",lwd=2,axes=F)

text(8,7.8,"Censura tipo I",cex=1.4)
lines(fm2,2*fm2/fm2,col="navy",lwd=2)
lines(fm3,3*fm3/fm3,col="navy",lwd=2)
lines(fm4,4*fm4/fm4,col="navy",lwd=2)
lines(fm5,5*fm5/fm5,col="navy",lwd=2)
lines(fm6,6*fm6/fm6,col="navy",lwd=2)
lines(fm7,6.75*fm7/fm7)
lines(c(12,12),c(0,6.75),lty=2,lwd=2)
points(9,1,col="red",pch=4,lwd=3)
points(3,3,col="red",pch=4,lwd=3)
points(5,4,col="red",pch=4,lwd=3)
points(7,6,col="red",pch=4,lwd=3)
legend("topleft",c("Seguimiento","Fin del estudio","Falla"),
       lty=c(1,2,NA),pch=c(NA,NA,4),lwd=2,col=c("navy","black","red"))

axis(side=1,at=c(0,3,6,9,12),labels=c("ENE","MAR","JUN","SEP","DIC"))
axis(side=2,at=c(1,2,3,4,5,6),labels=c("F","E","D","C","B","A"),las=1)
box()

# Nota: Las proporciones de esta figura podr'i'an variar su aspecto
# en funci'o'n de las dimensiones de la pantalla de su computadora.
```

Figura 1.3:

```
fn1 <- seq(0,25,length=100)
fn2 <- seq(0,50,length=100)
fn3 <- seq(0,50,length=100)
fn4 <- seq(0,50,length=100)
fn5 <- seq(0,50,length=100)
fn6 <- seq(0,40,length=100)
fn7 <- seq(0,15,length=100)
fn8 <- seq(0,30,length=100)
fn9 <- seq(29.8,60,length=100)

plot(fn1,fn1/fn1,type="l",xlab="Semanas",las=1,
     ylab="",xlim=c(0,55),ylim=c(0,11),col="navy",lwd=2,axes=F)
text(41,9.75,"Censura tipo II",cex=1.4)

lines(fn2,2*fn2/fn2,col="navy",lwd=2)
lines(fn3,3*fn3/fn3,col="navy",lwd=2)
lines(fn4,4*fn4/fn4,col="navy",lwd=2)
# continua pag 139
```

```
lines(fn5,5*fn5/fn5,col="navy",lwd=2)
lines(fn6,6*fn6/fn6,col="navy",lwd=2)
lines(fn7,7*fn7/fn7,col="navy",lwd=2)
lines(fn8,8*fn8/fn8,col="navy",lwd=2)
lines(fn9,8.93*fn7/fn7)
lines(c(50,50),c(0,8.9),lty=2,lwd=2)

points(25,1,col="red",pch=4,lwd=3)
points(50,4,col="red",pch=8,lwd=3)
points(40,6,col="red",pch=4,lwd=3)
points(15,7,col="red",pch=4,lwd=3)
points(30,8,col="red",pch=4,lwd=3)
legend("topleft",c("Seguimiento","Fin del estudio"),
       lty=c(1,2),lwd=2,col=c("navy","black","red","red"))
legend(x=16.55,y=11.44,c("Falla","r-esima falla"),
       pch=c(4,8),col=c("red","red"))
axis(side=1,at=c(0,10,20,30,40,50))
axis(side=2,at=c(1,2,3,4,5,6,7,8),
     labels=c("H","G","F","E","D","C","B","A"),las=1)
box()
# Nota: Las proporciones de esta figura podr'i'an variar su aspecto
# en funci'o'n de las dimensiones de la pantalla de su computadora.
```

Figura 1.4:

```
fo1 <- seq(0,9,length=100)
fo2 <- seq(12,24,length=100)
fo3 <- seq(2,18,length=100)
fo4 <- seq(6,14,length=100)
fo5 <- seq(9,21,length=100)
fo6 <- seq(4,16,length=100)
fo7 <- seq(0,33,length=100)

plot(fo1,fo1/fo1,type="l",xlab="Dias",las=1,
     ylab="",xlim=c(0,33),ylim=c(0,7),col="navy",lwd=2,axes=F,
     main="Censura tipo III", cex.main=1.4)

lines(fo2,2*fo2/fo2,col="navy",lwd=2)
lines(fo3,3*fo3/fo3,col="navy",lwd=2)
lines(fo4,4*fo4/fo4,col="navy",lwd=2)
lines(fo5,5*fo5/fo5,col="navy",lwd=2)
lines(fo6,6*fo6/fo6,col="navy",lwd=2)

points(9,1,col="red",pch=4,lwd=3)
points(24,2,col="blue",pch=1,lwd=3)
points(18,3,col="red",pch=4,lwd=3)
points(14,4,col="red",pch=4,lwd=3)
points(21,5,col="blue",pch=1,lwd=3)
points(16,6,col="red",pch=4,lwd=3)
legend("topright",c("Seguimiento","Falla","Perdido"),
       lty=c(1,NA,NA),pch=c(NA,4,1),lwd=2,col=c("navy","red","blue"))
axis(side=1,at=c(0,6,12,18,24,30))
axis(side=2,at=c(1,2,3,4,5,6),labels=c("F","E","D","C","B","A"),las=1)
box()
```

Figura 1.5:

```
fp1 <- seq(0,6,length=100)
fp2 <- seq(0,1,length=100)
fp3 <- seq(0,4,length=100)
fp4 <- seq(0,7,length=100)
fp5 <- seq(0,2,length=100)
fp6 <- seq(0,5,length=100)

plot(fp1,fp1/fp1,type="l",xlab="",las=1,
     ylab="",xlim=c(0,10),ylim=c(0,7),col="navy",lwd=2,axes=F,
     main="Censura por la Izquierda", cex.main=1.2)

lines(fp2,2*fp2/fp2,col="navy",lwd=2)
lines(fp3,3*fp3/fp3,col="navy",lwd=2)
lines(fp4,4*fp4/fp4,col="navy",lwd=2)
lines(fp5,5*fp5/fp5,col="navy",lwd=2)
lines(fp6,6*fp6/fp6,col="navy",lwd=2)
lines(c(3,3),c(0,9),lty=2,lwd=2)

points(6,1,col="red",pch=4,lwd=3)
points(1,2,col="blue",pch=13,lwd=3)
points(4,3,col="red",pch=4,lwd=3)
points(7,4,col="red",pch=4,lwd=3)
points(2,5,col="blue",pch=13,lwd=3)
points(5,6,col="red",pch=4,lwd=3)

legend("topright",c("Seguimiento","Inicio del estudio",
       "Falla Observada","Falla no observada"),
       lty=c(1,2,NA,NA),pch=c(NA,NA,4,13),lwd=2,
       col=c("navy","black","red","blue"))

axis(side=1,at=c(3,10),labels=c("O","t"))
axis(side=2,at=c(1,2,3,4,5,6),labels=c("F","E","D","C","B","A"),las=1)
box()
```

Figura 1.6:

```
##### Figura A
ex_4 <- seq(0,90,length=900)
l_4 <- .066
med_4 <- log(2)/l_4
e_4 <- 1/l_4
mm_4 <- mean(ex_4)
rem_4 <- exp(-l_4*mm_4)

plot(ex_4,exp(-ex_4*l_4),xlim=c(0,90),type="l",
     xlab="t",ylab="S(t)",col="blue")
lines(c(med_4,med_4), c(0,0.5), lty=2)
lines(c(0,med_4),c(0.5,0.5), lty=2)
lines(c(mm_4,mm_4),c(0,rem_4), lty=2,col="red")
legend(x=66,y=0.85, legend=c("S(t)","media","mediana"),
       lty=c(1,2,2),col=c("blue","red","black"))

# continua pag 141
```

```
##### Figura B

ex_3 <- seq(0,90,length=900)
l_3 <- .033
med_3 <- log(2)/l_3
e_3 <- 1/l_3
mm_3 <- mean(ex_3)
rem_3 <- exp(-l_3*mm_3)

plot(ex_3,exp(-ex_3*l_3),xlim=c(0,90),type="l",xlab="t",ylab="S(t)",col
    ="blue")
lines(c(med_3,med_3), c(0,0.5), lty=2)
lines(c(0,med_3),c(0.5,0.5), lty=2)
lines(c(mm_3,mm_3),c(0,rem_3), lty=2,col="red")
legend(x=66,y=0.85, legend=c("S(t)","media","mediana"),
        lty=c(1,2,2),col=c("blue","red","black"))


##### Figura C

we_2 <- seq(0,10,length=900)
g_2 <- 2.2
l_2 <- .25
med_2 <- ((log(2))^(1/g_2))/l_2
e_2 <- gamma(1+1/g_2)/l_2
mm_2 <- mean(we_2)
rem_2 <- exp(-(mm_2*l_2)^g_2)

plot(we_2,exp(-(we_2*l_2)^g_2),xlim=c(0,10),type="l",
      xlab="t",ylab="S(t)",col="blue")
lines(c(med_2,med_2), c(0,0.5), lty=2)
lines(c(0,med_2),c(0.5,0.5), lty=2)
lines(c(mm_2,mm_2),c(0,rem_2), lty=2,col="red")
legend(x=7,y=0.85, legend=c("S(t)","media","mediana"),
        lty=c(1,2,2),col=c("blue","red","black"))


##### Figura D

we_1 <- seq(0,2,length=900)
g_1 <- 3.8
l_1 <- .88
med_1 <- ((log(2))^(1/g_1))/l_1
e_1 <- gamma(1+1/g_1)/l_1
mm_1 <- mean(we_1)
rem_1 <- exp(-(mm_1*l_1)^g_1)

plot(we_1,exp(-(we_1*l_1)^g_1),xlim=c(0,2),type="l",
      xlab="t",ylab="S(t)",col="blue")
lines(c(med_1,med_1), c(0,0.5), lty=2)
lines(c(0,med_1),c(0.5,0.5), lty=2)
lines(c(mm_1,mm_1),c(0,rem_1), lty=2,col="red")
legend(x=1.4,y=0.85, legend=c("S(t)","media","mediana"),
        lty=c(1,2,2),col=c("blue","red","black"))
```

Figura 1.7:

```
##### Figura A

a_1<-seq(0,4,length=900)
an_1<- -1*a_1

plot(a_1,exp(an_1),xlim=c(0,4),type="l",xlab="t",
     ylab="f(t)",col="blue")


##### Figura B

a_2<-seq(0,4,length=900)

an_21<-.89
an_22<-2
and_1<-an_21*an_22
and_2<-(an_21*a_2)^(an_22-1)
and_3<-exp(-1*(an_21*a_2)^(an_22))

plot(a_2,and_1*and_2*and_3,type="l",xlim=c(0,4),xlab="t",
     ylab="f(t)",col="blue")
```

Figura 1.8:

```
### Nota: Esta figura se logra con algunas de las funciones
### programadas para las figuras: 1.9, 1.11-1.15.

plot(sec3,ris_llogis(sec3,1,5),type="l",col="blue",xlim=c(0,4),
     ylim=c(0,4),axes=F,ylab="",xlab="",xaxs="i",yaxs="r",lwd=2)

axis(side=1,at=c(0,2,4),labels=c(0,"t",""),lwd=2)
axis(side=2,at=c(0,2,4),labels=c("","h(t)",""),lwd=2)

lines(sec3,ris_gompertz(sec3,.5,.85),type="l",col="red",lwd=2)
lines(sec3,ris_exp(sec3,1.5),type="l",col="darkgreen",lwd=2)
lines(sec3,(sec3-1.1)^4+.5,type="l",col="darkmagenta",lwd=2)
lines(sec3,ris_llogis(sec3,4,1.1),type="l",col="gold",lwd=2)

legend("topright",c("a.","b.","c.","d.","e."),lty=c(1,1,1,1,1),
       col=c("darkgreen","red","blue","darkmagenta","gold"),lwd=2)
```

Figura 1.9:

```
########## Exponencial~(lambda)
# R: (t,lambda)
# descargar: package(stats)

# Secuencias
sec1 <- seq(0,6,length=900)

### Supervivencia
sup_exp <- function(u,d){
  sup_exp = 1-pexp(u,d)
  return(sup_exp)
}

### Riesgo
ris_exp <- function(u,d){
  ris_exp = dexp(u,d)/(1-pexp(u,d))
  return(ris_exp)
}


## G_dis: Figura A
plot(sec1,dexp(sec1,2),type="l",col="blue",xlab="t",ylab="f(t)")
lines(sec1,dexp(sec1,1),type="l",col="red")
lines(sec1,dexp(sec1,1/2),type="l",col="darkgreen")
lines(sec1,dexp(sec1,1/5),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=2")),
expression(paste(lambda,"=1")),expression(paste(lambda,"=0.5")),
expression(paste(lambda,"=0.2"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


## G_ris: Figura B
plot(sec1,ris_exp(sec1,2),type="l",col="blue",
xlab="t",ylab="h(t)",ylim=c(0,2.5))
lines(sec1,ris_exp(sec1,1),type="l",col="red")
lines(sec1,ris_exp(sec1,1/2),type="l",col="darkgreen")
lines(sec1,ris_exp(sec1,1/5),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=2")),
expression(paste(lambda,"=1")),expression(paste(lambda,"=0.5")),
expression(paste(lambda,"=0.2"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


## G_sup: Figura C
plot(sec1,sup_exp(sec1,2),type="l",col="blue",xlab="t",ylab="S(t)")
lines(sec1,sup_exp(sec1,1),type="l",col="red")
lines(sec1,sup_exp(sec1,1/2),type="l",col="darkgreen")
lines(sec1,sup_exp(sec1,1/5),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=2")),
expression(paste(lambda,"=1")),expression(paste(lambda,"=0.5")),
expression(paste(lambda,"=0.2"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)
```

Figura 1.10:

```
sec3 <- seq(0,4,length=900)
subsec3 <- seq(1,3,length=900)

plot(sec3,dexp(sec3,1),type="l",col="blue",xaxs="i",yaxs="r",
     axes=F,xlab="Tiempo",ylab="")

axis(side=1,at=c(0,1,3,4),labels=c("0","t1","t2",""))
axis(side=2,at=c(0,1),labels=c("","f(t)"))

x_arf <- c(1,subsec3,3)
y_arf <- c(0,dexp(subsec3,1),0)

polygon(x_arf,y_arf,col="grey")
lines(subsec3,dexp(subsec3,1),type="l",col="red",lwd=3)
```

Figura 1.11:

```
########## Weibul~(lambda,gamma)
# (t,shape=gamma,scale=1/lambda)
# descargar: package(stats)

# Secuencias
sec2 <- seq(0,5,length=900)

### Supervivencia
sup_weibull <- function(u,d,t){
  c = 1/t
  sup_weibull = 1-pweibull(u,d,c)
  return(sup_weibull)
}

### Riesgo
ris_weibull <- function(u,d,t){
  c = 1/t
  ris_weibull = dweibull(u,d,c)/(1-pweibull(u,d,c))
  return(ris_weibull)
}


## G_dis: Figura A
plot(sec2,dweibull(sec2,3,1),type="l",col="blue",
xlab="t",ylab="f(t)",ylim=c(0,4.5))
lines(sec2,dweibull(sec2,6,.66),type="l",col="red")
lines(sec2,dweibull(sec2,2,1.25),type="l",col="darkgreen")
lines(sec2,dweibull(sec2,.5,2),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=1",gamma,"=3")),
expression(paste(lambda,"= 1.5",gamma,"=6")),
expression(paste(lambda,"=0.8",gamma,"=2")),
expression(paste(lambda,"=0.5",gamma,"=0.5"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)

# continua pag 145
```

```
## G_ris: Figura B
plot(sec2,ris_weibull(sec2,.5,.5),type="l",
col="gold",xlab="t",ylab="h(t)")
lines(sec2,ris_weibull(sec2,6,1.5),type="l",col="red")
lines(sec2,ris_weibull(sec2,2,.8),type="l",col="darkgreen")
lines(sec2,ris_weibull(sec2,3,1),type="l",col="blue")
legend("topright",c(expression(paste(lambda,"=1",gamma,"=3")),
expression(paste(lambda,"=1.5",gamma,"=6")),
expression(paste(lambda,"=0.8",gamma,"=2")),
expression(paste(lambda,"=0.5",gamma,"=0.5"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


## G_sup: Figura C
plot(sec2,sup_weibull(sec2,3,1),type="l",col="blue",
xlab="t",ylab="S(t)",ylim=c(0,1))
lines(sec2,sup_weibull(sec2,6,1.5),type="l",col="red")
lines(sec2,sup_weibull(sec2,2,.8),type="l",col="darkgreen")
lines(sec2,sup_weibull(sec2,.5,.5),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=1",gamma,"=3")),
expression(paste(lambda,"=1.5",gamma,"=6")),
expression(paste(lambda,"=0.8",gamma,"=2")),
expression(paste(lambda,"=0.5",gamma,"=0.5"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)
```

Figura 1.12

```
########## Log-Normal~(mu,sigma^2)
# R: (t,mu,sigma^2)
# descargar: package(stats)

# Secuencias
sec2 <- seq(0,5,length=900)

### Supervivencia
sup_lnorm <- function(u,d,t){
  sup_lnorm = 1-plnorm(u,d,t)
  return(sup_lnorm)
}


### Riesgo
ris_lnorm <- function(u,d,t){
  ris_lnorm = dlnorm(u,d,t)/(1-plnorm(u,d,t))
  return(ris_lnorm)
}

# continua pag 146
```

```
## G_dis: Figura A
plot(sec2,dlnorm(sec2,0,.5),type="l",col="blue",
xlab="t",ylab="f(t)",ylim=c(0,1))
lines(sec2,dlnorm(sec2,0,.75),type="l",col="red")
lines(sec2,dlnorm(sec2,.5,0.6),type="l",col="darkgreen")
lines(sec2,dlnorm(sec2,1,2),type="l",col="gold")
legend("topright",c(expression(paste(mu,"=0",sigma^2,"=0.50")),
expression(paste(mu,"=0",sigma^2,"=0.75")),
expression(paste(mu,"=0.5",sigma^2,"=0.60")),
expression(paste(mu,"=1",sigma^2,"=2"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


## G_ris: Figura B
plot(sec2,ris_lnorm(sec2,0,.5),type="l",
col="blue",xlab="t",ylab="h(t)",ylim=c(0,2))
lines(sec2,ris_lnorm(sec2,0,.75),type="l",col="red")
lines(sec2,ris_lnorm(sec2,.5,.6),type="l",col="darkgreen")
lines(sec2,ris_lnorm(sec2,1,2),type="l",col="gold")
legend("topright",c(expression(paste(mu,"=0",sigma^2,"=0.50")),
expression(paste(mu,"=0",sigma^2,"=0.75")),
expression(paste(mu,"=0.5",sigma^2,"=0.60")),
expression(paste(mu,"=1",sigma^2,"=2"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


## G_sup: Figura C
plot(sec2,sup_lnorm(sec2,0,.5),type="l",col="blue",
xlab="t",ylab="S(t)",ylim=c(0,1))
lines(sec2,sup_lnorm(sec2,0,.75),type="l",col="red")
lines(sec2,sup_lnorm(sec2,.5,.6),type="l",col="darkgreen")
lines(sec2,sup_lnorm(sec2,1,2),type="l",col="gold")
legend("topright",c(expression(paste(mu,"=0",sigma^2,"=0.50")),
expression(paste(mu,"=0",sigma^2,"=0.75")),
expression(paste(mu,"=0.5",sigma^2,"=0.60")),
expression(paste(mu,"=1",sigma^2,"=2"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)
```

Figura 1.13:

```
########## Log-Logistic~(lambda,kappa)
# R: (t,shape=kappa,scale=1/lambda)
# descargar: package(flexsurv)

# Secuencias
sec3 <- seq(0,4,length=900)

### Supervivencia
sup_llogis <- function(u,d,t){
  k = 1/t
  sup_llogis = 1-pllogis(u,d,k)
  return(sup_llogis)
}
# continua pag 147
```

```
### Riesgo
ris_llogis <- function(u,d,t){
  k = 1/t
  ris_llogis = dllogis(u,d,k)/(1-pllogis(u,d,k))
  return(ris_llogis)
}


## G_dis: Figura A
plot(sec3,dllogis(sec3,1,.2),type="l",col="blue",
xlab="t",ylab="f(t)",ylim=c(0,2))
lines(sec3,dllogis(sec3,2,.5),type="l",col="red")
lines(sec3,dllogis(sec3,4,1),type="l",col="darkgreen")
lines(sec3,dllogis(sec3,5,2),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=5",kappa,"=1")),
expression(paste(lambda,"=2",kappa,"=2")),
expression(paste(lambda,"=1",kappa,"=4")),
expression(paste(lambda,"=0.5",kappa,"=5"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


# G_ris: Figura B
plot(sec3,ris_llogis(sec3,1,5),type="l",col="blue",
xlab="t",ylab="h(t)")
lines(sec3,ris_llogis(sec3,2,2),type="l",col="red")
lines(sec3,ris_llogis(sec3,4,1),type="l",col="darkgreen")
lines(sec3,ris_llogis(sec3,5,.5),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=5",kappa,"=1")),
expression(paste(lambda,"=2",kappa,"=2")),
expression(paste(lambda,"=1",kappa,"=4")),
expression(paste(lambda,"=0.5",kappa,"=␣5"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)


## G_sup: Figura C
plot(sec3,sup_llogis(sec3,1,5),type="l",col="blue",
xlab="t",ylab="S(t)",ylim=c(0,1))
lines(sec3,sup_llogis(sec3,2,2),type="l",col="red")
lines(sec3,sup_llogis(sec3,4,1),type="l",col="darkgreen")
lines(sec3,sup_llogis(sec3,5,.5),type="l",col="gold")
legend("topright",c(expression(paste(lambda,"=5",kappa,"=1")),
expression(paste(lambda,"=2",kappa,"=2")),
expression(paste(lambda,"=1",kappa,"=4")),
expression(paste(lambda,"=0.5",kappa,"=5"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)
```

Figura 1.14:

```
########## Gamma~(alpha,beta)
# R: (t,alpha,beta)
# descargar: package(stats)

# Secuencias
sec4 <- seq(0,10,length=900)

### Supervivencia
sup_gamma <- function(u,d,t){
  sup_gamma = 1-pgamma(u,d,t)
  return(sup_gamma)
}

### Riesgo
ris_gamma <- function(u,d,t){
  ris_gamma = dgamma(u,d,t)/(1-pgamma(u,d,t))
  return(ris_gamma)
}

## G_dis: Figura A
plot(sec4,dgamma(sec4,1,1),type="l",col="blue",
xlab="t",ylab="f(t)")
lines(sec4,dgamma(sec4,3,2),type="l",col="red")
lines(sec4,dgamma(sec4,4,1),type="l",col="darkgreen")
lines(sec4,dgamma(sec4,.5,.8),type="l",col="gold")
legend("topright",c(expression(paste(alpha,"=1",beta,"=1")),
expression(paste(alpha,"=3",beta,"=2")),
expression(paste(alpha,"=4",beta,"=1")),
expression(paste(alpha,"=.5",beta,"=.8"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)

## G_ris: Figura B
plot(sec4,ris_gamma(sec4,.5,1),type="l",col="blue",
xlab="t",ylab="h(t)",ylim=c(0,2))
lines(sec4,ris_gamma(sec4,1,1),type="l",col="gold")
lines(sec4,ris_gamma(sec4,2,1),type="l",col="red")
lines(sec4,ris_gamma(sec4,3,1),type="l",col="darkgreen")
legend("topright",c(expression(paste(alpha,"=0.5",beta,"=1")),
expression(paste(alpha,"=2",beta,"=1")),
expression(paste(alpha,"=3",beta,"=1")),
expression(paste(alpha,"=1",beta,"=1"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)

## G_sup: Figura C
plot(sec4,sup_gamma(sec4,1,1),type="l",col="blue",
xlab="t",ylab="S(t)")
lines(sec4,sup_gamma(sec4,3,2),type="l",col="red")
lines(sec4,sup_gamma(sec4,4,1),type="l",col="darkgreen")
lines(sec4,sup_gamma(sec4,.5,.8),type="l",col="gold")
legend("topright",c(expression(paste(alpha,"=1",beta,"=1")),
expression(paste(alpha,"=3",beta,"=2")),
expression(paste(alpha,"=4",beta,"=1")),
expression(paste(alpha,"=.5",beta,"=.8"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)
```

Figura 1.15:

```
########## Gompertz~(alpha,beta)
# R: (t,scale=beta,shape=alfa)
# descargar: package(flexsurv)

#Secuencia
sec5 <- seq(0,15,length=900)

### Supervivencia
sup_gompertz <- function(u,d,t){
  sup_gompertz = 1-pgompertz(u,d,t)
  return(sup_gompertz)
}

### Riesgo
ris_gompertz <- function(u,d,t){
  ris_gompertz = dgompertz(u,d,t)/(1-pgompertz(u,d,t))
  return(ris_gompertz)
}

## G_dis: Figura A
plot(sec5,dgompertz(sec5,.05,.4),type="l",
col="blue",xlab="t",ylab="f(t)")
lines(sec5,dgompertz(sec5,.3,.03),type="l",col="red")
lines(sec5,dgompertz(sec5,.4,.06),type="l",col="darkgreen")
lines(sec5,dgompertz(sec5,.6,.08),type="l",col="gold")
legend("topright",c(expression(paste(alpha,"=0.40",beta,"=0.05")),
expression(paste(alpha,"=0.03",beta,"=0.30")),
expression(paste(alpha,"=0.06",beta,"=0.40")),
expression(paste(alpha,"=0.08",beta,"=0.60"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)

## G_ris: Figura B
plot(sec5,ris_gompertz(sec5,.05,.4),type="l",
col="blue",xlab="t",ylab="h(t)",ylim=c(0,6))
lines(sec5,ris_gompertz(sec5,.3,.03),type="l",col="red")
lines(sec5,ris_gompertz(sec5,.4,.06),type="l",col="darkgreen")
lines(sec5,ris_gompertz(sec5,.6,.08),type="l",col="gold")
legend("topright",c(expression(paste(alpha,"=0.40",beta,"=0.05")),
expression(paste(alpha,"=0.03",beta,"=0.30")),
expression(paste(alpha,"=0.06",beta,"=0.40")),
expression(paste(alpha,"=0.08",beta,"=0.60"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)

## G_sup: Figura C
plot(sec5,sup_gompertz(sec5,.05,.4),type="l",
col="blue",xlab="t",ylab="S(t)")
lines(sec5,sup_gompertz(sec5,.3,.03),type="l",col="red")
lines(sec5,sup_gompertz(sec5,.4,.06),type="l",col="darkgreen")
lines(sec5,sup_gompertz(sec5,.6,.08),type="l",col="gold")
legend("topright",c(expression(paste(alpha,"=0.40",beta,"=0.05")),
expression(paste(alpha,"=0.03",beta,"=0.30")),
expression(paste(alpha,"=0.06",beta,"=0.40")),
expression(paste(alpha,"=0.08",beta,"=0.60"))),lty=1,
col=c("blue","red","darkgreen","gold"),lwd=3)
```

Figura 1.16:

```
library(stats)

secp<-seq(0,1,length=100)
op<-c(0.2,0.4,0.4,0.5,0.6,0.7,0.7,0.8,0.9,0.9)
cp<-c(1,1,1,1,1,1,1,1,0,0)
leno<-length(op)
pop<-op*cp
lop<-log(op)
lpop<-sum(log(pop[1:8]))
slop<-sum(log(op))
mop<-prod(op)
smop<-sum(cp)

luc<-function(theta){
  -sum(cp*log(theta)-cp*lop+theta*log(0.1)-theta*lop)
}

lucy<-nlm(luc,c(1),hessian=TRUE)
lucy<-lucy$estimate
ml<- 8*log(lucy)-lpop+10*lucy*log(0.1)-lucy*slop
mlp<-(((lucy^smop)/prod(op[1:8]))*
      ((0.1)^(leno*lucy))/((prod(op))^lucy))

### Figura A
plot(secp,8*log(secp)-lpop+10*secp*log(0.1)-secp*slop,
    type="l",main="log-verosimilitud",col="darkblue",lwd=2,las=1,
    xlab=expression(theta),ylab=expression(paste("log(␣L(",theta,"))")))
lines(c(lucy,lucy),c(-31.4,ml),col="firebrick3",lwd=2,lty=2)
points(lucy,ml,col="firebrick3",pch=16)

### Figura B
plot(secp,((secp^smop)/prod(op[1:8]))*
    ((0.1)^(leno*secp))/((prod(op))^secp),
    type="l",main="verosimilitud",col="darkblue",lwd=2,
    xlab=expression(theta),ylab=expression(paste("L(",theta,")")))
lines(c(lucy,lucy),c(0.0000001,mlp),col="firebrick3",lwd=2,lty=2)
points(lucy,mlp,col="firebrick3",pch=16)
```

Figura A:

```
library(survival)

v_duchenne<-c(3,6,6,7,8,8,8,9,10,11)
v_duchfit<-survfit(Surv(v_duchenne)~1)

plot(v_duchfit,xlab="trimestres",ylab="",
      col="dodgerblue4",conf=F,lwd=2,las=1)
legend("topright",c(expression(paste(hat(S), "(t)"))),
        lty=1,lwd=2,col=c("dodgerblue4"))

summary(v_duchfit)
```

Figura 1.17:

```
library(survival)

v_tim<-c(9,16,6,7,10,12,12,15,15,14,20,18)
v_del<-c(1,1,1,0,0,1,1,1,0,1,1,0)
v_kmfit<-survfit(Surv(v_tim,v_del) ~1)

plot(v_kmfit,ylab="",xlab="",col="steelblue4",las=1,lwd=2)
legend("bottomleft",c(expression(paste("Estimador ",
          hat(S), "(t) de Kaplan-Meier")),
          expression("Intervalos del 95% de Confianza")),
          lty=c(1,2),col=c("steelblue4","steelblue4"))
```

Figura 1.18:

```
library(survival)

tiempos<-c(7,4,9,12,9,15,18,5,20,16)
censura<-c(1,1,0,1,0,1,0,1,0,1)

datos<-Surv(tiempos,censura)
datos

km<-survfit(datos~1,se.fit=F)
summary(km)
plot(km)
```

Figura 1.19

```
library(survival)

tiems<-c(1,4,5,5,9,10,13,16,16,18,21,22,23,24,27,30,33,35,35,40,
            1,1,2,3,4,4,6,8,10,12,14,14,17,20,21,23,23,26,33,40)
cens<-c(0,1,1,1,0,1,1,1,0,1,1,0,1,0,1,0,1,1,0,0,
             1,0,0,0,1,1,0,1,0,1,1,0,1,1,0,1,0,1,1,0)
trats<-factor(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
                2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2))

ajuste<-survfit(Surv(tiems,cens)~trats)

plot(ajuste,col=c("navy","darkred"),las=1,lwd=2)
legend("topright",c("Tratamiento A","Tratamiento B"),lty=c(1,1),
       col=c("navy","darkred"))

summary(ajuste)
```

A.2. Códigos Capítulo 2

Figura 2.1:

```
### Nota: Esta figura se logra con las funciones
### programadas para la Figura 1.12.

sec2 <- seq(0,5,length=900)

## Figura A
plot(sec2,sup_lnorm(sec2,0,.75),type="l",
col="firebrick",xlab="",ylab="",ylim=c(0,1),lwd=2,las=1)
lines(sec2,sup_lnorm(sec2,.2,.76),type="l",col="darkblue",lwd=2)
legend("topright",c(expression(paste(S[1](t))),
expression(paste(S[2](t)))),lty=1,col=c("firebrick","darkblue"),lwd=3)

## Figura B
plot(sec2,ris_lnorm(sec2,0,.75),type="l",
col="firebrick",xlab="",ylab="",ylim=c(0,2),lwd=2,las=1)
lines(sec2,ris_lnorm(sec2,.2,.76),type="l",col="darkblue",lwd=2)
legend("topright",c(expression(paste(h[1](t))),
expression(paste(h[2](t)))),lty=1,col=c("firebrick","darkblue"),lwd=3)
```

Figura 2.2:

```
ves1 <- seq(0,3, length=100)
ves2 <- seq(0,6, length=100)
ves3 <- seq(0,2, length=100)
ves4 <- seq(0,5, length=100)
ves5 <- seq(0,4, length=100)

plot(ves1,ves1/ves1,type="l",xlab="Tiempos ordenados de falla",
ylab="Individuos", xlim=c(0,7.7),ylim=c(0,5.5),col="navy",
lwd=2,las=1,axes=F,main="", cex.main=1.2)

lines(ves2,2*ves2/ves2,col="navy",lwd=2)
lines(ves3,3*ves3/ves3,col="navy",lwd=2)
lines(ves4,4*ves4/ves4,col="navy",lwd=2)
lines(ves5,5*ves5/ves5,col="navy",lwd=2)
lines(c(2,2),c(0,9),lty=2,lwd=2)
lines(c(4,4),c(0,9),lty=2,lwd=2)
lines(c(6,6),c(0,9),lty=2,lwd=2)

points(3,1,col="lightskyblue4",pch=20,lwd=3)
points(6,2,col="red",pch=4,lwd=3)
points(2,3,col="red",pch=4,lwd=3)
points(5,4,col="lightskyblue4",pch=20,lwd=3)
points(4,5,col="red",pch=4,lwd=3)

legend("topright",c("Falla","Censura"),
pch=c(4,20),col=c("red","lightskyblue4"))

axis(side=1,at=c(0,2,4,6),labels=c(0,"t(1)","t(2)","t(3)"))
axis(side=2,at=c(1,2,3,4,5),labels=c("5","4","3","2","1"))
box()
```

Ejemplo 2.2.3:

```
library(KMsurv)
data(larynx)
larynx

larynx$stage <- factor(larynx$stage)
vesper_0 <- coxph(Surv(time, delta)~age+stage,
data=larynx, method=''breslow'')

vesper_0

# Salida:
          coef     exp(coef)   se(coef)     z        p
age       0.0189      1.0191     0.0143   1.33    0.185
stage2    0.1386      1.1486     0.4623   0.30    0.764
stage3    0.6383      1.8934     0.3561   1.79    0.073
stage4    1.6931      5.4361     0.4222   4.01    6.1e-05

Likelihood ratio test=18.1  on 4 df, p=0.0012
n= 90, number of events= 50
```

Figura 2.3:

```
vesper_n<-function(K,N,m,s){
  for(i in 1:K){
    if(i<2){
      a<-c(1:N)
      a[1]=rnorm(1,m,s)
      r=a
      f<-seq(0,N,length=300)
      for(j in 2:N){
        r[j]<-r[j-1]+rnorm(1,m,s)
        a[j]<-r[j]/j
      }
      j=1
      plot(a,type="l",col=i,xlab="N",ylab="",las=1,ylim=c(m-s/2,m+s/2))
      lines(f,m*f/f,lty=2)
    }
    else{
      a<-c(1:N)
      a[1]=rnorm(1,m,s)
      r=a
      for(j in 2:N){
        r[j]<-r[j-1]+rnorm(1,m,s)
        a[j]<-r[j]/j
      }
      j=1
      lines(a,type="l",col=i)
    }
  }
}

vesper_n(4,1000,2,4)
# K=ensayos, N=simulaciones de una normal con media=m y varianza=s.
```

Ejemplo 2.3.1:

```
library(eha)
library(FAdist)
library(survival)
library(flexsurv)
library(graphics)
library(KMsurv)
library(lattice)
library(MASS)
library(muhaz)
library(OIsurv)
library(splines)
library(stats)
library(stats4)


leuk
leucem<-leuk
leucem$wbc=log10(leucem$wbc)
leucem
leucema<-leucem
names(leucema)
leucema$wbc=round(leucema$wbc,2)
leucema_wbc<-leucema$wbc
leucema_t<-leucema$time
leucema_f<-c(rep(1,17),rep(0,16))

# Funci'o'n -logverosimilitud.
vesper_leuk<- function(theta){
  -sum(-theta[1]-theta[2]*leucema_wbc-theta[3]*leucema_f-
  leucema_t*exp(-theta[1]-theta[2]*leucema_wbc-theta[3]*leucema_f))
}

# Resultados:
mt<-nlm(vesper_leuk,c(6,0,1),hessian=TRUE)
mt

# 1.- Estimadores de theta
pt<-mt$estimate
pt

# 2.- Matriz de informaci'o'n observada
mt1<-mt$hessian
mt1

# 3.- Inversa de la matriz
solve(mt1)
mt_in<-solve(mt1)
mt_in

# continua pag 155
```

```
# 4. Errores est'a'ndar de theta estimado
for(i in 1:3){
  se=sqrt(mt_in[i,i])
  print(se)
}

# Par'a'metros: lambda, beta_1 y beta_2
lambda_l=exp(-pt[1])
beta1_l=-pt[2]
beta2_l=-pt[3]

paras<-c(lambda_l,beta1_l,beta2_l)
paras

# Error est'a'ndar de lambda
lam_se=exp(-pt[1])*sqrt(mt_in[1,1])
lam_se


# Ajuste con survreg
vesper1=survreg(Surv(time)~log10(wbc)+ag,
                data=leuk,dist="exponential")

# Resumen de los resultados
summary(vesper1)
```

Figura 2.4 y 2.5:

```
library(survival)
library(MASS)

leuk
vesper_rec <- survfit(Surv(time) ~ 1,data=leuk)

##### Figura 2.4: Recta Ajustada Exponencial
plot(vesper_rec$time,-log(vesper_rec$surv),col="blue",
     type="l",lwd=2,las=1,xlab="t",ylab="",cex.lab=1.8)
legend("topleft",c(expression(paste(hat(H),"(t)"))),
       col="blue",lwd=2,cex=1.7)


##### Figura 2.5: Ajuste Kaplan-Meier Exponencial
plot(vesper_rec,conf.int=F,col="blue",las=1)
lines(flexsurvreg(Surv(time) ~ 1,data=leuk,dist="exponential"),col="red
    ",ci=F)
legend("topright",c(expression(paste(hat(S),"(t)")),
                    "Kaplan-Meier"),lty=c(1,1),
       col=c("red","blue"),lwd=2,cex=1.6)
```

Ejemplo 2.4.1:

```
library(Survival)
library(MASS)
leuk

vesper2=survreg(Surv(time)~log10(wbc)+ag,data=leuk,dist="weibull")
summary(vesper2)

# Matriz de informaci'o'n
vesper2$var

# C'a'lculo de los par'a'metros: gamma, lambda, betas
gamma_l=1/vesper2$scale
gamma_l

lambda_l=exp(-vesper2$coeff[1]/vesper2$scale)
lambda_l

bet1_l= -vesper2$coeff[2]/vesper2$scale
bet1_l

bet2_l= -vesper2$coeff[3]/vesper2$scale
bet2_l

# Errores est'a'ndar: sigma, gamma, lambda, beta1, beta2
ssig_l=(vesper2$scale)*(sqrt(vesper2$var[4,4]))
ssig_l # se(sigma)

covms_l=(vesper2$scale)*(vesper2$var[1,4])
covb1s_l=(vesper2$scale)*(vesper2$var[2,4])
covb2s_l=(vesper2$scale)*(vesper2$var[3,4])

sgam_l=ssig_l/((vesper2$scale)^2)
sgam_l # s.e(gamma)

slam_l=((vesper2$coeff[1]*exp(-vesper2$coeff[1]/vesper2$scale)/
            ((vesper2$scale)^2))^2)*((sgam_l)^2)+((-exp(-vesper2$coeff
                [1]/
            vesper2$scale)/((vesper2$scale)^2))^2)*vesper2$var[1,1]+
            2*(vesper2$coeff[1]*exp(-vesper2$coeff[1]/vesper2$scale)/
            ((vesper2$scale)^2))*(-exp(-vesper2$coeff[1]/vesper2$scale)/
            ((vesper2$scale)^2))*covms_l
slam_l # s.e(lambda)

sbet1_l=((vesper2$coeff[2]/((vesper2$scale)^2))^2)*((sgam_l)^2)+
          ((-1/vesper2$scale)^2)*(vesper2$var[2,2])+2*(vesper2$coeff[2]/
         ((vesper2$scale)^2))*((-1/vesper2$scale)^2)*covb1s_l
sbet1_l #s.e(beta1)

sbet2_l=((vesper2$coeff[3]/((vesper2$scale)^2))^2)*((sgam_l)^2)+
          ((-1/vesper2$scale)^2)*(vesper2$var[3,3])+2*(vesper2$coeff[3]/
          ((vesper2$scale)^2))*((-1/vesper2$scale)^2)*covb2s_l
sbet2_l #s.e(beta2)
```

Figura 2.6 y 2.7:

```
library(survival)
library(MASS)

leuk
vesper_rec <- survfit(Surv(time) ~ 1,data=leuk)

##### Figura 2.6: Recta Ajustada Weibull
plot(log(vesper_rec$time),log(-log(vesper_rec$surv)),col="blue",
     type="l",lwd=2,las=1,xlab="log(t)",ylab="",cex.lab=1.8)
legend("topleft",c(expression(paste("log(-log(",hat(S),"(t))")))),
       col="blue",lwd=2,cex=1.7)


##### Figura 2.7: Ajuste Kaplan-Meier Weibull
plot(vesper_rec,conf.int=F,col="blue",las=1)
lines(flexsurvreg(Surv(time) ~ 1,data=leuk,dist="weibull"),
col="red",ci=F)
legend("topright",c(expression(paste(hat(S),"(t)")),
                    "Kaplan-Meier"),lty=c(1,1),
       col=c("red","blue"),lwd=2,cex=1.6)
```

Ejemplo 2.5.1:

```
library(survival)
library(MASS)

leuk
# Ajuste del modelo gompertz: flexsurvreg
vesper3 <- flexsurvreg(Surv(time) ~ log10(wbc)+ag,
                       data = leuk,dist = "gompertz")
vesper3

# Inversa de la matriz de varianzas y covarianzas
vesper3$cov

## Procedimiento para los estimadores y los errores est'a'ndar
leucem<-leuk
leucem$wbc=log10(leucem$wbc)
leucem
leucema<-leucem
names(leucema)


leucema$wbc=round(leucema$wbc,2)
leucema_wbc<-leucema$wbc
leucema_t<-leucema$time
leucema_f<-c(rep(1,17),rep(0,16))
```

```
# Funci'o'n reparametrizada para la (-log.verosimilitud)
vesper3_leuk<- function(theta){
  -sum(-theta[1]-theta[2]*leucema_t-theta[3]*
       leucema_wbc-theta[4]*leucema_f+(1/theta[2])*
        exp(-theta[1]-theta[3]*leucema_wbc-theta[4]*
         leucema_f)*(exp(-theta[2]*leucema_t)-1))
}


# Resultados:
mtp<-nlm(vesper3_leuk,c(0.01,0.01,1,-1),hessian=TRUE)
mtp

# 1.- Estimadores de theta
ptp<-mtp$estimate
ptp

# 2. Estimadores del modelo: alpha, beta, beta1, beta2
ptp_m=-ptp
ptp_m[1]=exp(-ptp[1])
ptp_m

# 2.- Matriz de informaci'o'n observada
mt1p<-mtp$hessian
mt1p

# 3.- Inversa de la matriz
solve(mt1p)
mt_inp<-solve(mt1p)
mt_inp

# 4. Errores est'a'ndar de theta estimado
for(i in 1:4){
  sep=sqrt(mt_inp[i,i])
  print(sep)
}


## Ajuste del modelo con la funci'o'n phreg:
vesper3_ph <- phreg(Surv(time) ~ log10(wbc)+ag, data = leuk,
                    dist = "gompertz")
vesper3_ph
```

A.3. Códigos Capítulo 3

Estructura de las bases de datos:

```
# Con este comando se descarga la base de datos del portal web citado:

anderson.dat<- read.table("http://web1.sph.emory.edu/dkleinb
/allDatasets/surv2datasets/anderson.dat")

anderson.dat


# Descomposici'o'n de la base:
tcolum<-c("Tiempo","Censura","Sexo","logWBC","Tratamiento")
names(anderson.dat)<-tcolum
anderson.dat

## Las variables de la base de datos
## tienen la siguiente representaci'o'n
# Tiempo: tiempo de supervivencia (en semanas)
# Censura: observaci'o'n (0 = censura, 1 = falla)
# Sexo: g'e'nero (0 = femenino, 1 = masculino)
# logWBC: log WBC (logaritmo de conteo de gl'o'bulos balncos
#           por sus siglas en ingles White Blood Cell)
# Tratamiento: medicamento (0 = tratamiento, 1 = placebo),
# sin embargo, como el grupo de referencia es el placebo
# cambiemos esta estructura asignando el valor 0 al placebo
# y uno al tratamiento 6-MP

attach(anderson.dat)
c_anderson.dat <- anderson.dat
tiempo <- c_anderson.dat$Tiempo
delta <- c_anderson.dat$Censura
sex <- ifelse(Sexo==0, 0, 1)
trat <- ifelse(Tratamiento==1, 0, 1)
lwbc <- c_anderson.dat$logWBC
and <- data.frame(cbind(tiempo,delta,sex,lwbc,trat))
lwbc_f <- lwbc
for(i in 1:42){
  if ( lwbc[i] <= 2.3) {
    lwbc_f[i] = 0
  }
}
for(i in 1:42){
  if ( 2.3<lwbc[i] & lwbc[i] <=3) {
    lwbc_f[i] = 1
  }
}
for(i in 1:42){
  if ( 3<lwbc[i] & lwbc[i] <=5) {
    lwbc_f[i] = 2
  }
}

# continua pag 160
```

```
lwbc_f <- factor(lwbc_f,levels=c("0","1","2"))
and_f <- data.frame(cbind(tiempo,delta,sex,lwbc_f,trat))

## A las variables de base de datos, cuando se
## considera a la variable logWBC como categ'o'rica
## se les asigna la siguiente representaci'o'n
## se les asigna la siguiente representaci'o'n
# tiempo: tiempo de supervivencia (en semanas)
# delta: observaci'o'n (0 = censura, 1 = falla)
# sex: g'e'nero (0 = femenino, 1 = masculino)
# lwbc: conteo de gl'o'bulos blancos (1 = bajo, 2 = normal, 3 = alto)
#         bajo:<=2.3, normal: (2.3,3], alto: (3,5]
# trat: medicamento (1 = tratamiento, 0 = placebo-referencia)

attach(and)
attach(and_f)

### SubDataSets
#   Por el tipo de Tratamiento
and_t1 <- subset(and_f, trat==1,select=c(tiempo,delta,sex,lwbc_f))
and_t2 <- subset(and_f, trat==0,select=c(tiempo,delta,sex,lwbc_f))

# Por el nivel de gl'o'bulos blancos
and_g1 <- subset(and_f, lwbc_f==1,select=c(tiempo,delta,sex))
and_g2 <- subset(and_f, lwbc_f==2,select=c(tiempo,delta,sex))
and_g3 <- subset(and_f, lwbc_f==3,select=c(tiempo,delta,sex))

# Por el sexo
and_s1 <- subset(and_f, sex==0,select=c(tiempo,delta))
and_s2 <- subset(and_f, sex==1,select=c(tiempo,delta))


attach(and_t1)
attach(and_t2)
attach(and_t3)

attach(and_g1)
attach(and_g2)
attach(and_g3)

attach(and_s1)
attach(and_s2)
```

Tabla de Análisis descriptivo de la muestra:

```
# Con el siguiente comando se obtienen las siguientes estad'i'sticas;
# Min=m'i'nimo, 1st Qu=primer cuartil, Median= mediana,
# Mean=media, 3rd Qu= tercer cuartil, Max= m'a'ximo.

summary(and)
```

Figura 3.1:

```
par(lwd=3)

# Figura 3.1 a)
ta_delta <- table(delta)
ba_delta <- barplot(ta_delta,col=c("ivory","ivory1"),
                    axisnames = F, axes=F,
                    border=c("steelblue4","firebrick4"),main="",las=1)
text(x=ba_delta,y=c(ta_delta[1],ta_delta[2]),
     labels=c("Censurados","No Censurados"),pos=3,
     col=c("steelblue4","firebrick4"),cex=1.5,xpd=TRUE)
text(x=ba_delta,y=c(ta_delta[1],ta_delta[2]),
     labels=c(ta_delta[1],ta_delta[2]),pos=1,
      col=c("steelblue4","firebrick4"),cex=1.5,xpd=TRUE)


# Figura 3.1 b)
ta_sex <- table(sex)
ba_sex <- barplot(ta_sex,col=c("plum2","lightcyan"),
                  axisnames = F, axes=F,
                  border=c("purple4","royalblue4"),main="",las=1)
text(x=ba_delta,y=c(ta_sex[1],ta_sex[2]),
     labels=c("Mujeres","Hombres"),pos=3,
     col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)
text(x=ba_delta,y=c(ta_sex[1],ta_sex[2]),
     labels=c(ta_sex[1],ta_sex[2]),pos=1,
     col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)


# Figura 3.1 c)
ta_g <- table(lwbc_f)
ba_g <- barplot(ta_g,
          col=c("rosybrown1","lightgoldenrodyellow","honeydew"),
          axes = F,axisnames = F,
          border=c("firebrick4","goldenrod2","forestgreen"),
          main="",las=1)
text(x=ba_g,y=c(ta_g[1],ta_g[2],ta_g[3]),
     labels=c("Bajo","Medio","Alto"),pos=3,
     col=c("firebrick4","orange3","forestgreen"),cex=1.5,xpd=TRUE)
text(x=ba_g,y=c(ta_g[1],ta_g[2],ta_g[3]),
     labels=c(ta_g[1],ta_g[2],ta_g[3]),pos=1,
     col=c("firebrick4","orange3","forestgreen"),cex=1.5,xpd=TRUE)

par(lwd=1)
```

Figura 3.2:

```
par(lwd=3)

# Figura 3.2 a)
ta_deltat1 <- table(and_t1$delta)
ba_deltat1 <- barplot(ta_deltat1,col=c("ivory","ivory1"),
                axisnames = F, axes=F,
                border=c("steelblue4","firebrick4"),main="",las=1)
text(x=ba_deltat1,y=c(ta_deltat1[1],ta_deltat1[2]),
       labels=c("Censurados","No Censurados"),pos=3,
      col=c("steelblue4","firebrick4"),cex=1.5,xpd=TRUE)
text(x=ba_deltat1,y=c(ta_deltat1[1],ta_deltat1[2]),
       labels=c(ta_deltat1[1],ta_deltat1[2]),pos=1,
      col=c("steelblue4","firebrick4"),cex=1.5,xpd=TRUE)


# Figura 3.2 b)
ta_sext1 <- table(and_t1$sex)
ba_sext1 <- barplot(ta_sext1,col=c("plum2","lightcyan"),
                    axisnames = F, axes=F,
                    border=c("purple4","royalblue4"),
                    main="",las=1)
text(x=ba_sext1,y=c(ta_sext1[1],ta_sext1[2]),
      labels=c("Mujeres","Hombres"),pos=3,
      col=c("purple4","royalblue4"),cex=2.5,xpd=TRUE)
text(x=ba_sext1,y=c(ta_sext1[1],ta_sext1[2]),
      labels=c(ta_sext1[1],ta_sext1[2]),pos=1,
      col=c("purple4","royalblue4"),cex=2.5,xpd=TRUE)


# Figura 3.2 c)
ta_gt1 <- table(and_t1$lwbc_f)
ba_gt1 <- barplot(ta_gt1,
           col=c("rosybrown1","lightgoldenrodyellow","honeydew"),
           axes = F,axisnames = F,
           border=c("firebrick4","goldenrod2","forestgreen"),
                  main="",las=1)
text(x=ba_gt1,y=c(ta_gt1[1],ta_gt1[2],ta_gt1[3]),
      labels=c("Bajo","Medio","Alto"),pos=3,
      col=c("firebrick4","orange3","forestgreen"),cex=1.5,xpd=TRUE)
text(x=ba_gt1,y=c(ta_gt1[1],ta_gt1[2],ta_gt1[3]),
      labels=c(ta_gt1[1],ta_gt1[2],ta_gt1[3]),pos=1,
      col=c("firebrick4","orange3","forestgreen"),cex=1.5,xpd=TRUE)

par(lwd=1)
```

Figura 3.3:

```
par(lwd=3)

# Figura 3.3 a)
ta_sext2 <- table(and_t2$sex)
ba_sext2 <- barplot(ta_sext2,col=c("plum2","lightcyan"),
                    axisnames = F, axes=F,
                    border=c("purple4","royalblue4"),
                    main="",las=1)
text(x=ba_sext2,y=c(ta_sext2[1],ta_sext2[2]),
     labels=c("Mujeres","Hombres"),pos=3,
     col=c("purple4","royalblue4"),cex=2.5,xpd=TRUE)
text(x=ba_sext2,y=c(ta_sext2[1],ta_sext1[2]),
     labels=c(ta_sext2[1],ta_sext2[2]),pos=1,
     col=c("purple4","royalblue4"),cex=2.5,xpd=TRUE)


# Figura 3.3 b)
ta_gt2 <- table(and_t2$lwbc_f)
ba_gt2 <- barplot(ta_gt2,
            col=c("rosybrown1","lightgoldenrodyellow","honeydew"),
            axes = F, axisnames = F,
            border=c("firebrick4","goldenrod2","forestgreen"),
           main="",las=1)
text(x=ba_gt2,y=c(ta_gt2[1],ta_gt2[2],ta_gt2[3]),
     labels=c("Bajo","Medio","Alto"),pos=3,
     col=c("firebrick4","orange3","forestgreen"),cex=1.5,xpd=TRUE)
text(x=ba_gt2,y=c(ta_gt2[1],ta_gt2[2],ta_gt2[3]),
     labels=c(ta_gt2[1],ta_gt2[2],ta_gt2[3]),pos=1,
     col=c("firebrick4","orange3","forestgreen"),cex=1.5,xpd=TRUE)

par(lwd=1)
```

Figura 3.4:

```
par(lwd=3)

# Figura 3.4 a)
ta_deltag1 <- table(and_g1$delta)
ba_deltag1 <- barplot(ta_deltag1,col=c("ivory","ivory1"),
                      axisnames = F, axes=F,
                      border=c("steelblue4","firebrick4"),
                      main="",las=1)
text(x=ba_deltag1,y=c(ta_deltag1[1],ta_deltag1[2]),
      labels=c("Censurados","No Censurados"),pos=3,
     col=c("steelblue4","firebrick4"),cex=2.2,xpd=TRUE)
text(x=ba_deltag1,y=c(ta_deltag1[1],ta_deltag1[2]),
     labels=c(ta_deltag1[1],ta_deltag1[2]),pos=1,
     col=c("steelblue4","firebrick4"),cex=2.5,xpd=TRUE)

# Continua pag 164
```

```
# Figura 3.4 b)
ta_deltag2 <- table(and_g2$delta)
ba_deltag2 <- barplot(ta_deltag2,col=c("ivory","ivory1"),
                      axisnames = F, axes=F,
                      border=c("steelblue4","firebrick4"),
                      main="",las=1)
text(x=ba_deltag2,y=c(ta_deltag2[1],ta_deltag2[2]),
     labels=c("Censurados","No Censurados"),pos=3,
     col=c("steelblue4","firebrick4"),cex=2.2,xpd=TRUE)
text(x=ba_deltag2,y=c(ta_deltag2[1],ta_deltag2[2]),
     labels=c(ta_deltag2[1],ta_deltag2[2]),pos=1,
     col=c("steelblue4","firebrick4"),cex=2.5,xpd=TRUE)


# Figura 3.4 c)
ta_deltag3 <- table(and_g3$delta)
ba_deltag3 <- barplot(ta_deltag3,col=c("ivory","ivory1"),
                      axisnames = F, axes=F,
                      border=c("steelblue4","firebrick4"),
                      main="",las=1)
text(x=ba_deltag3,y=c(ta_deltag3[1],ta_deltag3[2]),
     labels=c("Censurados","No Censurados"),pos=3,
     col=c("steelblue4","firebrick4"),cex=2.2,xpd=TRUE)
text(x=ba_deltag3,y=c(ta_deltag3[1],ta_deltag3[2]),
     labels=c(ta_deltag3[1],ta_deltag3[2]),pos=1,
     col=c("steelblue4","firebrick4"),cex=2.5,xpd=TRUE)

par(lwd=1)
```

Figura 3.5:

```
par(lwd=3)

# Figura 3.5 a)
ta_sexg1 <- table(and_g1$sex)
ba_sexg1 <- barplot(ta_sexg1,col=c("plum2","lightcyan"),
                    axisnames = F, axes=F,
                    border=c("purple4","royalblue4"),
                    main="",las=1)
text(x=ba_sexg1,y=c(ta_sexg1[1],ta_sexg1[2]),labels=c("Mujeres","
    Hombres"),pos=3,
     col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)
text(x=ba_sexg1,y=c(ta_sexg1[1],ta_sexg1[2]),labels=c(ta_sexg1[1],
    ta_sexg1[2]),pos=1,
     col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)

# Continua pag 165
```

```
# Figura 3.5 b)
ta_sexg2 <- table(and_g2$sex)
ba_sexg2 <- barplot(ta_sexg2,col=c("plum2","lightcyan"),
                    axisnames = F, axes=F,
                    border=c("purple4","royalblue4"),
                    main="␣",las=1)
text(x=ba_sexg2,y=c(ta_sexg2[1],ta_sexg2[2]),labels=c("Mujeres","
    Hombres"),pos=3,
      col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)
text(x=ba_sexg2,y=c(ta_sexg2[1],ta_sexg2[2]),labels=c(ta_sexg2[1],
    ta_sexg2[2]),pos=1,
      col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)


# Figura 3.5 c)
ta_sexg3 <- table(and_g3$sex)
ba_sexg3 <- barplot(ta_sexg3,col=c("plum2","lightcyan"),
                    axisnames = F, axes=F,
                    border=c("purple4","royalblue4"),
                    main="␣",las=1)
text(x=ba_sexg3,y=c(ta_sexg3[1],ta_sexg3[2]),labels=c("Mujeres","
    Hombres"),pos=3,
      col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)
text(x=ba_sexg3,y=c(ta_sexg3[1],ta_sexg3[2]),labels=c(ta_sexg3[1],
    ta_sexg3[2]),pos=1,
      col=c("purple4","royalblue4"),cex=1.5,xpd=TRUE)

par(lwd=1)
```

Figura 3.6:

```
par(lwd=2)

# Figura 3.6 a)
hist(and$tiempo,col="honeydew",border="darkslategray",
      xlab="semanas",ylab="␣",main="␣",
      cex.lab=2,las=1,probability=T)
lines(density(and$tiempo),col="darkslategray")


# Figura 3.6 b)
hist(and$tiempo,col="honeydew",border="darkslategray",
      xlab="semanas",ylab="␣",main="␣",cex.lab=2,las=1)


par(lwd=1)
```

Figura 3.7:

```
par(lwd=4)

# Figura 3.7 a)
hist(and_t1$tiempo,col="aliceblue",border="midnightblue",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)
lines(density(and_t1$tiempo),col="midnightblue")

# Figura 3.7 b)
hist(and_t2$tiempo,col="mistyrose",border="firebrick4",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)
lines(density(and_t2$tiempo),col="firebrick4")


# Figura 3.6 c)
hist(and_s1$tiempo,col="plum2",border="purple4",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)
lines(density(and_s1$tiempo),col="purple4")


# Figura 3.6 d)
hist(and_s2$tiempo,col="lightcyan",border="royalblue4",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)

par(lwd=1)
```

Figura 3.8:

```
par(lwd=3)
# Figura 3.8 a)
hist(and_g1$tiempo,col="rosybrown1",border="firebrick4",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)
lines(density(and_g1$tiempo),col="firebrick4")

# Figura 3.8 b)
hist(and_g2$tiempo,col="lightyellow",border="orange3",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)
lines(density(and_g2$tiempo),col="orange3")

# Figura 3.8 c)
hist(and_g3$tiempo,col="honeydew",border="forestgreen",
     xlab="tiempo",ylab="␣",main="␣",
     cex.lab=1.2,las=1,probability=T)
lines(density(and_g3$tiempo),col="forestgreen")

par(lwd=1)
```

Análisis de Kaplan-Meier:

```
# Ajuste de Kaplan-Meier:
# Muestra completa:
km_M <- survfit(Surv(tiempo,delta) ~ 1)
summary(km_M)

# Por el tipo de Tratamiento*:
km_T <- survfit(Surv(tiempo,delta) ~ trat)
summary(km_T)

# Por los niveles de logWBC*:
km_G <- survfit(Surv(tiempo,delta) ~ lwbc_f)
summary(km_G)

# Por el sexo*:
km_S <- survfit(Surv(tiempo,delta) ~ sex)
summary(km_S)

# Nota: Los resultados marcados con * no aparecen
# de manera impresa en esta tesis
```

Figura 3.9

```
km_M <- survfit(Surv(tiempo,delta) ~ 1)
km_M
medi_M <- 12
# El valor 12 se obtiene del valor median
# impreso en la salida del comando km_M

# Figura 3.9 a)
plot(km_M,col="darkslategray",main="",lwd=2,las=1)
lines(c(medi_M,medi_M),c(0,0.5),lty=2,lwd=3.5,col="darkgray")
lines(c(0,medi_M),c(0.5,0.5),lty=2,lwd=3.5,col="darkgray")
legend("topright",c("Kaplan-Meier"),
       col=c("darkslategray"),lwd=2,cex=1.8)


# Figura 3.9 b)
plot(km_M,fun="cumhaz",col=c("darkslategray"),
     main="",lwd=2,las=1)
legend("bottomright",c("Riesgo Acumulado"),
       col=c("darkslategray"),lwd=2,cex=1.5)
```

Figura 3.10:

```
# Figura 3.10 a)
km_M <- survfit(Surv(tiempo,delta) ~ 1)
plot(km_M,col="darkslategray",main="",lwd=2,las=1)
wph <- flexsurvreg(Surv(tiempo,delta)~1,dist="weibull")
lines(wph,col="forestgreen")
legend("topright",c("Kaplan-Meier","Ajuste Weibull"),
       col=c("darkslategray","forestgreen"),lwd=2,cex=1.8)

# Figura 3.10 b)
fit_M <- survfit(Surv(tiempo,delta) ~ 1)
plot(log(fit_M$time),log(-log(fit_M$surv)),
     col="darkslategray", type="l",lwd=2,las=1,
     xlab="log(t)",ylab="",cex.lab=1.8)
legend("bottomright",
       c(expression(paste("log( -log(",hat(S),"(t)) )"))),
       col="darkslategray",lwd=2,cex=1.7)
```

Figura 3.11:

```
# Figura 3.11 a)
km_T <- survfit(Surv(tiempo,delta) ~ trat)
km_T
medi_T1 <- 8
medi_T2 <- 23
plot(km_T,col=c("firebrick4","midnightblue"),
     main="",lwd=2,las=1)
lines(c(medi_T1,medi_T1),c(0,0.5),lty=2,
      lwd=3.5,col="darkgray")
lines(c(0,medi_T1),c(0.5,0.5),lty=2,lwd=3.5,
      col="darkgray")
lines(c(medi_T2,medi_T2),c(0,0.5),lty=2,lwd=3.5,
      col="darkgray")
lines(c(0,medi_T2),c(0.5,0.5),lty=2,lwd=3.5,
      col="darkgray")
legend("topright",c("Placebo","Tratamiento"),
       col=c("firebrick4","midnightblue"),l
       wd=2,cex=1.4)

# Figura 3.11 b)
plot(km_T,fun="cumhaz",
     col=c("firebrick4","midnightblue"),
     main="",lwd=2,las=1)
legend("topleft",c("Placebo","Tratamiento"),
       col=c("firebrick4","midnightblue"),lwd=2)
```

Figura 3.12:

```
# Figura 3.12 a)
plot(km_T,fun="cloglog",log="x",
      col=c("firebrick4","midnightblue"),
      main="",lwd=2,las=1)
legend("topleft",c("Placebo","Tratamiento"),
        col=c("firebrick4","midnightblue"),lwd=2)

# Figura 3.12 b)
fit_t1 <- survfit(Surv(and_t1$tiempo,and_t1$delta) ~1)
fit_t2 <- survfit(Surv(and_t2$tiempo,and_t2$delta) ~1)
plot(log(fit_t2$time),log(-log(fit_t2$surv)),
      col="firebrick4",xlim=c(0,4),
      ylim=c(-2,2),type="l",lwd=2,las=1,
      xlab="log(t)",ylab="",cex.lab=1.8)
lines(log(fit_t1$time),log(-log(fit_t1$surv)),
        col="midnightblue",type="l",lwd=2,las=1,
        xlab="log(t)",ylab="",cex.lab=1.8)
legend("topleft",
         c(expression(paste("log(-log(",hat(S),"(t)))")),
           expression(paste("log(-log(",hat(S),"(t)))"))),
        col=c("midnightblue","firebrick4"),lwd=3,cex=1.4)

# Figura 3.12 c)
wph.t <- flexsurvreg(Surv(tiempo,delta)~as.factor(trat),dist="weibull")
plot(wph.t,xlab="semanas",ylab="",las=1,col=c("midnightblue","
    firebrick4"),cex.lab=1.5)
lines(km_T,col=c("red","blue"))
legend("topright",c("Placebo","Tratamiento"),
        col=c("firebrick4","midnightblue"),lwd=3,cex=1.5)
```

Figura 3.13:

```
# Figura 3.13 a)
km_G <- survfit(Surv(tiempo,delta) ~ lwbc_f)
km_G
medi_G2 <- 17
medi_G3 <-  6
  plot(km_G,col=c("firebrick1","darkgoldenrod","forestgreen"),
      main="",lwd=2,las=1)
lines(c(medi_G2,medi_G2),c(0,0.5),lty=2,lwd=3.5,col="darkgray")
lines(c(0,medi_G2),c(0.5,0.5),lty=2,lwd=3.5,col="darkgray")
lines(c(medi_G3,medi_G3),c(0,0.5),lty=2,lwd=3.5,col="darkgray")
lines(c(0,medi_G3),c(0.5,0.5),lty=2,lwd=3.5,col="darkgray")
legend("topright",c("Bajo","Medio","Alto"),
        col=c("firebrick1","darkgoldenrod","forestgreen"),
        lwd=2,cex=1.4)

# Figura 3.13 b)
plot(km_G,fun="cumhaz",
      col=c("firebrick1","darkgoldenrod","forestgreen"),main="",lwd=2,
          las=1)
legend("topleft",c("Bajo","Medio","Alto"),
        col=c("firebrick1","darkgoldenrod","forestgreen"),lwd=2,cex=1.4)
```

Figura 3.14:

```
# Figura 3.14 a)
plot(km_G, fun="cloglog", log="x",
     col=c("firebrick1","darkgoldenrod","forestgreen"),
     main="", lwd=2, las=1)
legend("topleft", c("Bajo","Medio","Alto"),
       col=c("firebrick1","darkgoldenrod","forestgreen"),
       lwd=2, cex=1.4)


# Figura 3.14 b)
fit_g1 <- survfit(Surv(and_g1$tiempo, and_g1$delta) ~1)
fit_g2 <- survfit(Surv(and_g2$tiempo, and_g2$delta) ~1)
fit_g3 <- survfit(Surv(and_g3$tiempo, and_g3$delta) ~1)
plot(log(fit_g3$time), log(-log(fit_g3$surv)),
     col="forestgreen", xlim=c(0,5), type="l", lwd=2,
     las=1, xlab="log(t)", ylab=c("log(-log(S(t)))"))
lines(log(fit_g2$time), log(-log(fit_g2$surv)),
      col="darkgoldenrod", type="l", lwd=2, las=1,
      xlab="log(t)", ylab="", cex.lab=1.8)
lines(log(fit_g1$time), log(-log(fit_g1$surv)), col="firebrick1",
      type="l", lwd=2, las=1, xlab="log(t)", ylab="", cex.lab=1.8)
legend("topright", c("Alto","Medio","Bajo"),
       col=c("forestgreen","darkgoldenrod","firebrick1"), lwd=3, cex=1.3)


# Figura 3.14 c)
wph.g <- flexsurvreg(Surv(tiempo, delta)~as.factor(lwbc_f),
         dist="weibull")
plot(wph.g, xlab="semanas", ylab="", las=1, cex.lab=1.5,
     col=c("forestgreen","darkorange4","firebrick1"))
lines(km_G, col=c("green","darkorange","red"))
legend("topright", c("Alto","Medio","Bajo"),
       col=c("forestgreen","darkorange4","firebrick1"),
       lwd=3, cex=1.3)
```

Figura 3.15:

```
# Figura 3.15 a)
km_S <- survfit(Surv(tiempo, delta) ~ sex)
km_S
medi_F <- 12
plot(km_S, col=c("purple4","darkcyan"), main="", lwd=2, las=1)
lines(c(medi_F, medi_F), c(0, 0.5), lty=2, lwd=3.5, col="darkgray")
lines(c(0, medi_F), c(0.5, 0.5), lty=2, lwd=3.5, col="darkgray")
legend("topright", c("Mujeres","Hombress"),
       col=c("purple4","darkcyan"), lwd=2, cex=1.4)

# continua pag 171
```

```
# Figura 3.15 b)
plot(km_S,fun="cumhaz",col=c("purple4","darkcyan"),
     main="",lwd=2,las=1)
legend("topleft",c("Mujeres","Hombres"),
       col=c("purple4","darkcyan"),lwd=2,cex=1.4)

# Figura 3.15 c)
plot(km_S,fun="cloglog",log="x",col=c("purple4","darkcyan"),
     main="",lwd=2,las=1)
legend("topleft",c("Mujeres","Hombres"),
       col=c("purple4","darkcyan"),lwd=2,cex=1.4)

# Figura 3.15 d)
fit_s1 <- survfit(Surv(and_s1$tiempo,and_s1$delta) ~1)
fit_s2 <- survfit(Surv(and_s2$tiempo,and_s2$delta) ~1)
plot(log(fit_s2$time),log(-log(fit_s2$surv)),col="darkcyan",
     xlim=c(0,4),type="l",lwd=2,las=1,
     xlab="log(t)",ylab="",cex.lab=1.8)
lines(log(fit_s1$time),log(-log(fit_s1$surv)),col="purple4",
      type="l",lwd=2,las=1,xlab="log(t)",ylab="",cex.lab=1.8)
legend("bottomright",
       c(expression(paste("Hombres  log( -log(",hat(S),"(t)))")),
       expression(paste("Mujeres  log(-log( ",hat(S),"(t)))"))),
       col=c("darkcyan","purple4"),lwd=3,cex=1.12)
```

Prueba de log-rank:

```
## Por el tipo de tratamiento:
# H_0: Trat 6-MP = Placbo  v.s  H_a: Trat 6-MP \= Placebo
survdiff(Surv(tiempo,delta)~trat,rho=0)


## Por los niveles de gl'o'bulos blancos:
# H_0: bajo = medio = alto v.s H_a: Existe uno diferente a los dem'a's
survdiff(Surv(tiempo,delta)~lwbc_f,rho=0)
```

Modelos semiparamétricos de riesgos proporcionales:

```
##### Ajuste semiparam'e'trico
# Modelo 1:
cphm.t <- coxph(Surv(tiempo,delta)~trat)
cphm.t

# Modelo 2:
cphm.s <- coxph(Surv(tiempo,delta)~sex)
cphm.s

# Modelo 3:
cphm.g <- coxph(Surv(tiempo,delta)~lwbc)
cphm.g

# Modelo 4:
chpm.tg <- coxph(Surv(tiempo,delta)~trat+lwbc)
cphm.tg

# Modelo 5:
chpm.tgi <- coxph(Surv(tiempo,delta)~trat*lwbc)
cphm.tgi
```

AIC de los modelos semiparamétricos:

```
#### AIC de los Modelos 1, 3 y 4:
# AIC Modelo 1:
extract(AIC)(cphm.t)[2]

# AIC Modelo 3:
extract(AIC)(cphm.g)[2]

# AIC Modelo 4:
extract(AIC)(cphm.tg)[2]
```

Proporcionalidad:

```
# Residuales de Schoenfeled
cox.zph(cphm.tg)

# Las gr'a'ficas de estos residuales se pueden obtener con el comando
plot(cox.zph(cphm.tg))

# Nota: se obtiene una gr'a'fica para el tipo de tratamiento y
# otra para los niveles de gl'o'bulos blancos.
```

Figura 3.16:

```
champa <- data.frame(trat=c(0,1),lwbc=rep(mean(lwbc),2))
detach()
plot(survfit(cphm.tg,newdata=champa),las=1,
     lwd=c(2,2),col=c("firebrick4","midnightblue"))
legend("topright",c(expression(paste("Placebo:",hat(S),"(t|x)")),
                    expression(paste("Trat 6-MP:",hat(S),"(t|x)"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.4)
```

Modelos paramétricos:

Tabla 3.A:

```
# Ajuste Weibull con la funci'o'n phreg()
library(eha)
wph.and <- phreg(Surv(tiempo,delta)~trat+lwbc)
wph.and
```

Tabla 3.B:

```
# Ajuste Weibull con la funci'o'n survreg()
wph.tg <- survreg(Surv(tiempo,delta)~trat+lwbc,dist="weibull")
summary(wph.tg)
```

Tabla 3.C:

```
# gamma
gam = 1/wph.tg$scale

# lambda
lam = exp(-wph.tg$coeff[1]/wph.tg$scale)

# beta_1
bet1 = -wph.tg$coeff[2]/wph.tg$scale

# beta_2
bet2 = -wph.tg$coeff[3]/wph.tg$scale


## par'a'metros
param <- c(gam,lam,bet1,bet2)
titl <-c("gamma","lambda","beta_I", "beta_II")
names(param)<-titl


# Continua  pag 174
```

```
## C'a'lculo de los errores est'a'ndar
mu = wph.tg$coeff[1]
sig = wph.tg$scale
a1 = wph.tg$coeff[2]
a2 = wph.tg$coeff[3]
se_mu = summary(wph.tg)$table[1,2]
se_a1 = summary(wph.tg)$table[2,2]
se_a2 = summary(wph.tg)$table[3,2]
se_logs = summary(wph.tg)$table[4,2]
cov_mulog = wph.tg$var[1,4]
cov_a1log = wph.tg$var[2,4]
cov_a2log = wph.tg$var[3,4]
se_sig = sig*se_logs
se_gam = se_sig/((sig)^2)
vamu = (se_mu)^2
vasig = (se_sig)^2
cov_musig = sig*cov_mulog

se_lam =(((mu*lam/(sig*sig))^2)*(vasig)+
           ((-lam/sig)^2)*(vamu)+
           2*(mu*lam/(sig*sig))*((-lam/sig))*cov_musig)^(1/2)

cov_a1sig = sig*cov_a1log
cov_a2sig = sig*cov_a2log
se_b1 = (((a1/(sig^2))^2)*((se_sig)^2)+((-1/sig)^2)*((se_a1)^2)+
           2*(a1/(sig^2))*(-1/sig)*(cov_a1sig) )^(1/2)

se_b2 = (((a2/(sig^2))^2)*((se_sig)^2)+((-1/sig)^2)*((se_a2)^2)+
           2*(a2/(sig^2))*(-1/sig)*(cov_a2sig) )^(1/2)


## Errores est'a'ndar
se_wph.tg <- c(se_gam,se_lam,se_b1,se_b2)
setitl <- c("gamma","lambda","beta_T-II", "beta_WBC")
names(se_wph.tg) <- setitl

## C'a'lculo de los intervalos de confianza
# qnorm(0.975) = 1.96
l_gam <- param[1]-1.96*se_wph.tg[1]
l_lam <- param[2]-1.96*se_wph.tg[2]
l_b1 <- param[3]-1.96*se_wph.tg[3]
l_b2 <- param[4]-1.96*se_wph.tg[4]
r_gam <- param[1]+1.96*se_wph.tg[1]
r_lam <- param[2]+1.96*se_wph.tg[2]
r_b1 <- param[3]+1.96*se_wph.tg[3]
r_b2 <- param[4]+1.96*se_wph.tg[4]


# Continua pag 175
```

```
## Intervalos de confianza
L_wph.tg <- c(l_gam,l_lam,l_b1,l_b2)
Ltitl <-c("gamma","lambda","beta_G-B", "beta_G-C")
names(L_wph.tg)<-Ltitl
R_wph.tg <- c(r_gam,r_lam,r_b1,r_b2)
Rtitl <-c("gamma","lambda","beta_G-B", "beta_G-C")
names(R_wph.tg)<-Rtitl
# por la izquierda: L_wph.tg
# por la derecha: R_wph.tg

## Valores en forma de matriz
est <- c(param)
se <- c(se_wph.tg)
L95 <- c(L_wph.tg)
U95 <- c(R_wph.tg)
expos <- exp(est)

# Matriz de datos
Mtx <- cbind(est,expos,se,L95,U95)
Mtx
```

Tabla 3.D:

```
wph <- survreg(Surv(tiempo,delta)~1,dist="weibull")
summary(wph)
extractAIC(wph)[2]

wph.t <- survreg(Surv(tiempo,delta)~trat,dist="weibull")
summary(wph.t)
extractAIC(wph.t)[2]

wph.g <- survreg(Surv(tiempo,delta)~lwbc,dist="weibull")
summary(wph.g)
extractAIC(wph.g)[2]

wph.ts <- survreg(Surv(tiempo,delta)~trat+sex,dist="weibull")
summary(wph.ts)
extractAIC(wph.ts)[2]

wph.tg <- survreg(Surv(tiempo,delta)~trat+lwbc,dist="weibull")
summary(wph.tg)
extractAIC(wph.tg)[2]

wph.tgs <- survreg(Surv(tiempo,delta)~trat+lwbc+sex,dist="weibull")
summary(wph.tgs)
extractAIC(wph.tgs)[2]

wph.tsi <- survreg(Surv(tiempo,delta)~trat*sex,dist="weibull")
summary(wph.tsi)
extractAIC(wph.tsi)[2]

wph.tgi <- survreg(Surv(tiempo,delta)~trat*lwbc,dist="weibull")
summary(wph.tgi)
extractAIC(wph.tgi)[2]
```

Tabla 3.E:

```
## Ajuste Exponencial
peph.tg <- phreg(Surv(tiempo,delta)~trat+lwbc,shape=1)
peph.tg
eph.tg <- survreg(Surv(tiempo,delta)~trat+lwbc,dist="exponential")
extractAIC(eph.tg)[2]
```

Tabla 3.F:

```
# lambda, beta_1 y beta_2
vlam = exp(-eph.tg$coeff[1])
vbet1 = -eph.tg$coeff[2]
vbet2 = -eph.tg$coeff[3]

# par'a'metros
vparam <- c(vlam,vbet1,vbet2)
vtitl <-c("lambda","beta_I", "beta_II")
names(vparam)<-vtitl

# C'a'lculo de los errores estandar
se_vlam = vlam*summary(eph.tg)$table[1,2]
se_vbet1 = summary(eph.tg)$table[2,2]
se_vbet2 = summary(eph.tg)$table[3,2]

# Errores est'a'ndar
se_eph.tg <- c(se_vlam,se_vbet1,se_vbet2)
vsetitl <-c("lambda","beta_T-II", "beta_WBC")
names(se_eph.tg)<-vsetitl

# C'a'lculo de los intervalos de confianza
l_vlam <- vparam[1]-1.96*se_eph.tg[1]
l_vb1 <- vparam[2]-1.96*se_eph.tg[2]
l_vb2 <- vparam[3]-1.96*se_eph.tg[3]
r_vlam <- vparam[1]+1.96*se_eph.tg[1]
r_vb1 <- vparam[2]+1.96*se_eph.tg[2]
r_vb2 <- vparam[3]+1.96*se_eph.tg[3]

# Intervalos de confianza
L_eph.tg <- c(l_vlam,l_vb1,l_vb2)
vLtitl <-c("lambda","beta_G-B", "beta_G-C")
names(L_eph.tg)<-vLtitl
R_eph.tg <- c(r_vlam,r_vb1,r_vb2)
vRtitl <-c("lambda","beta_G-B", "beta_G-C")
names(R_eph.tg)<-vRtitl

# Matriz de datos:
estim <- c(vparam)
s.e <- c(se_eph.tg)
L.95 <- c(L_eph.tg)
U.95 <- c(R_eph.tg)
p_val <- c(p_val1,p_val2,p_val3)
expose <- exp(estim)
Mtx_vex <- cbind(estim,expose,s.e,L.95,U.95)
Mtx_vex
```

Cálculos para el tiempo mediano de supervivencia:

```
# La f'o'rmula del tiempo mediano de supervivencia para el modelo
# param'e'trico de regresi'o'n Weibull se puede encontrar en la
# p'a'gina 100 de este documento.

timed_ebo<-((((2.82e-05)^(-1))/exp(1.784*2.5))*log(2))^(1/2.21)
timed_trat<-((((2.82e-05)^(-1))/exp(-1.145+1.784*2.5))*log(2))^(1/2.21)

# Tiempo mediano para pacientes con el placebo.
timed_ebo

# Tiempo mediano para pacientes con el tratamiento.
timed_trat
```

Cálculos para la longitud del intervalo de confianza:

```
### Modelo param'e'trico
intmat <- Mtx[,5]-Mtx[,4]
logitmat_trat <- intmat[3]
logitmat_lwbc <- intmat[4]

# Longitud del intervalo para el tratamiento.
logitmat_trat

# Longitud del intervalo para la covariable logWBC.
logitmat_lwbc


### Modelo de Cox
tabol <- summary(cphm.tg)
infit_trat <- tabol$coef[1,1]-1.96*(tabol$coef[1,3])
supit_trat <- tabol$coef[1,1]+1.96*(tabol$coef[1,3])
logit_trat <- supit_trat-infit_trat


infit_lwbc <- tabol$coef[2,1]-1.96*(tabol$coef[2,3])
supit_lwbc <- tabol$coef[2,1]+1.96*(tabol$coef[2,3])
logit_lwbc <- supit_lwbc-infit_lwbc

# Longitud del intervalo para el tratamiento.
logit_trat

# Longitud del intervalo para la covariable logWBC.
logit_lwbc




### Nota: Para realizar este programa, es necesario haber realizado
### todos los programas correspondientes al ajuste semiparam'e'trico
### y param'e'trico de los modelos de riesgos proporcionales.
```

Figura 3.17:

```
secv = seq(0,35,length=100)
secvs = seq(0,180,length=1000)
cope = mean(lwbc)
phi2 = exp(cope*param[4])
phi1 = exp(param[3]+cope*param[4])
risini = param[2]*(param[1]*secv^(param[1]-1))
lrisini2 = log(risini*phi2)
lrisini1 = log(risini*phi1)
suv2 = exp(-param[2]*secv^param[1])^(phi2)
suv1 = exp(-param[2]*secv^param[1])^(phi1)
secvl = log(secv)
lsuv2 = log(-log(suv2))
lsuv1 = log(-log(suv1))

bills <- data.frame(trat=c(0,1),lwbc=rep(mean(lwbc),2))
detach()

# Figura 3.17 a)
plot(survfit(cphm.tg,newdata=bills),las=1,lty=2,
     lwd=c(2,2),col=c("firebrick4","midnightblue"))
legend("topright",c(expression(paste("Placebo: ",hat(S),"(t|x)")),
                    expression(paste("Trat 6-MP: ",hat(S),"(t|x)"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.4)
lines(secv,suv1,col="midnightblue",lwd=2,type="l")
lines(secv,suv2,col="firebrick4",lwd=2,type="l")


# Figura 3.17 b)
plot(survfit(cphm.tg,newdata=rega),las=1,lty=2,
     lwd=c(2,2),col=c("firebrick4","midnightblue"))
legend("topright",c(expression(paste("Placebo: ",hat(S),"(t|x)")),
                    expression(paste("Trat 6-MP: ",hat(S),"(t|x)"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.4)
lines(secv,suv1,col="midnightblue",lwd=2,type="l")
lines(secv,suv2,col="firebrick4",lwd=2,type="l")
lines(km_T,col=c("lightpink","lightskyblue"),lty=2,lwd=3)
```

Figura 3.18:

```
wph.and <- phreg(Surv(tiempo,delta)~trat+lwbc)

plot(wph.and,col=c("blue"),lwd=3,cex.lab=1.3)
```

Figura 3.19:

```
secv =seq(0,35,length=100)
secvs = seq(0,180,length=1000)
cope = mean(lwbc)
phi2 = exp(cope*param[4])
phi1 = exp(param[3]+cope*param[4])
risini = param[2]*(param[1]*secv^(param[1]-1))
lrisini2 = log(risini*phi2)
lrisini1 = log(risini*phi1)
suv2 = exp(-param[2]*secv^param[1])^(phi2)
suv1 = exp(-param[2]*secv^param[1])^(phi1)
secvl = log(secv)
lsuv2 = log(-log(suv2))
lsuv1 = log(-log(suv1))

# Figura 3.19 a)
plot(secv,risini*phi2,col="firebrick4",lwd=2,type="l",
     xlab="semanas", cex.lab=1.5,ylab="␣",las=1)
lines(secv,risini*phi1,col="midnightblue",lwd=2,type="l")
legend("topleft",c(expression(paste("Placebo:",hat(h),"(t|x)")),
       expression(paste("Trat␣6-MP:",hat(h),"(t|x)"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.7)

# Figura 3.19 b)
plot(secv,suv1,col="midnightblue",lwd=2,type="l",
     xlab="semanas",ylab="␣", cex.lab=1.5,las=1,ylim=c(0,1))
lines(secv,suv2,col="firebrick4",lwd=2,type="l")
legend("topright",c(expression(paste("Placebo:␣",hat(S),"(t|x)")),
       expression(paste("Trat␣6-MP:",hat(S),"(t|x)"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.6)

# Figura 3.19 c)
plot(secv,lrisini2,col="firebrick4",lwd=2,type="l",
      xlab="semanas",ylab="␣",las=1,cex.lab=1.5)
lines(secv,lrisini1,col="midnightblue",lwd=2,type="l")
legend("bottomright",
       c(expression(paste("Placebo:␣log(␣",hat(h),"(t|x)␣)")),
       expression(paste("Trat␣6-MP:␣log(␣",hat(h),"(t|x)␣)"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.7)

# Figura 3.19 d)
plot(secvl,lsuv2,col="firebrick4",lwd=2,type="l",
     xlab="log(t)",ylab="␣",las=1,cex.lab=1.5)
lines(secvl,lsuv1,col="midnightblue",lwd=2,type="l")
legend("topleft",
       c(expression(paste("P:log(-log(␣",hat(S),"(t|x)␣))")),
       expression(paste("T:␣␣log(-log(␣",hat(S),"(t|x)␣))"))),
       col=c("firebrick4","midnightblue"),lwd=3,cex=1.6)
```

Apéndice B

Bases de Datos

B.1. larynx

La base de datos **larynx** está disponible en el software estadístico **R** en la paquetería **library(KMsurv)**, contiene un total de 90 observaciones de hombres diagnosticados con cáncer de laringe (en sus distintas etapas) durante el periódo de 1970-1978 reportadas por en un hospital holandés.

stage	time	age	diagyr	delta
1	0.6	77	76	1
1	1.3	53	71	1
1	2.4	45	71	1
1	2.5	57	78	0
1	3.2	58	74	1
1	3.2	51	77	0
1	3.3	76	74	1
1	3.3	63	77	0
1	3.5	43	71	1
1	3.5	60	73	1
1	4	52	71	1
1	4	63	76	1
1	4.3	86	74	1
1	4.5	48	76	0
1	4.5	68	76	0
1	5.3	81	72	1
1	5.5	70	75	0
1	5.9	58	75	0
1	5.9	47	75	0
1	6	75	73	1
1	6.1	77	75	0
1	6.2	64	75	0
1	6.4	77	72	1
1	6.5	67	70	1
1	6.5	79	74	0
1	6.7	61	74	0
1	7	66	74	0
1	7.4	68	71	1
1	7.4	73	73	0
1	8.1	56	73	0
1	8.1	73	73	0
1	9.6	58	71	0
1	10.7	68	70	0
2	0.2	86	74	1
2	1.8	64	77	1
2	2	63	75	1
2	2.2	71	78	0
2	2.6	67	78	0
2	3.3	51	77	0
2	3.6	70	77	1
2	3.6	72	77	0
2	4	81	71	1
2	4.3	47	76	0
2	4.3	64	76	0
2	5	66	76	0
2	6.2	74	72	1
2	7	62	73	1
2	7.5	50	73	0
2	7.6	53	73	0
2	9.3	61	71	0
3	0.3	49	72	1
3	0.3	71	76	1
3	0.5	57	74	1
3	0.7	79	77	1
3	0.8	82	74	1
3	1	49	76	1
3	1.3	60	76	1
3	1.6	64	72	1
3	1.8	74	71	1
3	1.9	72	74	1
3	1.9	53	74	1
3	3.2	54	75	1
3	3.5	81	74	1
3	3.7	52	77	0
3	4.5	66	76	0
3	4.8	54	76	0
3	4.8	63	76	0
3	5	59	73	1
3	5	49	76	0
3	5.1	69	76	0
3	6.3	70	72	1
3	6.4	65	72	1
3	6.5	65	74	0
3	7.8	68	72	1
3	8	78	73	0
3	9.3	69	71	0
3	10.1	51	71	0
4	0.1	65	72	1
4	0.3	71	76	1
4	0.4	76	77	1
4	0.8	65	76	1
4	0.8	78	77	1
4	1	41	77	1
4	1.5	68	73	1
4	2	69	76	1
4	2.3	62	71	1
4	2.9	74	78	0
4	3.6	71	75	1
4	3.8	84	74	1
4	4.3	48	76	0

Figura B.1: Reporte de pacientes diagnosticados con cáncer de laringe.

Las columnas de la Tabla B.1 proporcionan la información de los siguientes datos:

- **stage** Etapa de la enfermedad.
- **time** Tiempo de muerte (en meses) o de estudio del paciente.
- **age** La edad que tiene el paciente al ser diagnosticado con cáncer de laringe.
- **diagyr** Año del diagnóstico de cáncer de laringe.
- **delta** Indicador de muerte (0 = vivo, 1 = muerto).

B.2. leuk

Base de datos **leuk** disponible en **library(MASS)**, contiene datos sobre la supervivencia de las víctimas de leucemia aguda considerados por Feigl y Zelen.

wbc	ag	time
2300	present	65
750	present	156
4300	present	100
2600	present	134
6000	present	16
10500	present	108
10000	present	121
17000	present	4
5400	present	39
7000	present	143
9400	present	56
32000	present	26
35000	present	22
100000	present	1
100000	present	1
52000	present	5
100000	present	65

wbc	ag	time
4400	absent	56
3000	absent	65
4000	absent	17
1500	absent	7
9000	absent	16
5300	absent	22
10000	absent	3
19000	absent	4
27000	absent	2
28000	absent	3
31000	absent	8
26000	absent	4
21000	absent	3
79000	absent	30
100000	absent	4
100000	absent	43
-	-	-

Figura B.2: Conteo de glóbulos blancos de pacientes diagnosticados con leucemia.

Información proporcionada por las columnas de la Tabla B.2:

- **time** Tiempo de diagnóstico (total de pacientes = 33).
- **wbc** Conteo de glóbulos blancos en la sangre al tiempo de diagnóstico.
- **ag** Indicador de presencia de características morfológicas.
  present = tiene características morfológicas, **absent** = no las tiene.

B.3. anderson.dat

La siguiente base de datos contiene la información del tiempo de remisión de 42 pacientes con leucemia aguda, los pacientes fueron divididos aleatoriamente en dos grupos según el tipo de tratamiento que recibieron mientras se encontraban en el estado de remisión.

Tiempo	Censura	Sexo	logWBC	Tratamiento
35	0	1	1.45	0
34	0	1	1.47	0
32	0	1	2.20	0
32	0	1	2.53	0
25	0	1	1.78	0
23	1	1	2.57	0
22	1	1	2.32	0
20	0	1	2.01	0
19	0	0	2.05	0
17	0	0	2.16	0
16	1	1	3.60	0
13	1	0	2.88	0
11	0	0	2.60	0
10	0	0	2.70	0
10	1	0	2.96	0
9	0	0	2.80	0
7	1	0	4.43	0
6	0	0	3.20	0
6	1	0	2.31	0
6	1	1	4.06	0
6	1	0	3.28	0

Tiempo	Censura	Sexo	logWBC	Tratamiento
23	1	1	1.97	1
22	1	0	2.73	1
17	1	0	2.95	1
15	1	0	2.30	1
12	1	0	1.50	1
12	1	0	3.06	1
11	1	0	3.49	1
11	1	0	2.12	1
8	1	0	3.52	1
8	1	0	3.05	1
8	1	0	2.32	1
8	1	1	3.26	1
5	1	1	3.49	1
5	1	0	3.97	1
4	1	1	4.36	1
4	1	1	2.42	1
3	1	1	4.01	1
2	1	1	4.91	1
2	1	1	4.48	1
1	1	1	2.80	1
1	1	1	5.00	1

Figura B.3: Remisión de pacientes diagnosticados con leucemia aguda.

En las columnas de la tabla anterior se puede observar la siguiente información:

- **Censura:** Indicador de censura (0 = observación censurada, 1 = el paciente presentó una falla, es decir, una recaída). *Nota:* Se considera que las observaciones censuradas están censuradas por la derecha.
- **Sexo:** El género del paciente (0 = femenino, 1 = masculino).
- **logWBC:** El logaritmo del conteo de los glóbulos blancos. (log White Blood Cell count).
- **Tratamiento:** El tipo de tratamiento que recibe el paciente a lo largo del estudio (0 = tratamiento 6-MP, 1 = Placebo).

Bibliografía

[1] Collet, D.(2003). Modelling Survival Data in Medical Research. Chapman & Hall.

[2] Hannerlore, L. Regression Models for Survival Data. Institute of Mathematics. University of Potsdam.

[3] Hosmer, W. and Lemeshow S.(1999) Applied Survival Analysis. Regression Modeling of Time to Event Data. A Wiley-Interscience Publication.

[4] Jenkins, Stephen P.(2004). Survival Analysis. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, UK.

[5] Kalbfleisch, J.D. and Prentice, R.L.(2002). The Statistical Analysis of Failure Time Data. Wiley series in probability and statistics.

[6] Klein, John P. and Moeschberger, Melvin L.(1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer.

[7] Kleinbaum, David and Mitchel, Klein.(2012). Survival Analysis: A Self-Learning Test. Springer.

[8] Lee, E.T. and Wang Wanyu (2003) Statistical Methods for Survival Data Analysis. Wiley series in probability and statistics.

[9] Murray, A. , Francis, B. and Darnell,R. (2009) Statistical Modelling in R. Oxford Statistical Science Series.

[10] Tableman, Mara (2005). Survival Analysis Using S: Analysis of Time to Event Data. Chapman & Hall/CRC.