

Modelos de Supervivencia
Gabriel Escarela

Capítulo 1

Introducción

El análisis de supervivencia es la frase usada para describir el análisis de datos que corresponden a la duración de un fenómeno, el cual comienza en un *tiempo origen* bien determinado y concluye con la ocurrencia de un evento. En investigación medica, por ejemplo, el tiempo origen generalmente se define como el momento en que se recluta un individuo para ser observado en un estudio experimental; el evento entonces puede definirse como la muerte de un individuo, los datos son literalmente *tiempos de supervivencia*. Sin embargo, en muchas disciplinas es posible encontrar fenómenos que no necesariamente son de naturaleza fatal, tales como la duración de huelgas o periodos de desempleo. Los métodos para analizar datos de supervivencia en general no se restringen a tiempos de supervivencia en su forma a datos que se refieren al tiempo de ocurrencia de un evento.

1.1. Características Especiales de los Datos de Supervivencia

Es importante considerar las razones por las que los datos de supervivencia no son compatibles con procedimientos estadísticos estándares. En primer lugar se tiene que los datos de supervivencia no son - en general - distribuidos simétricamente. Típicamente, un histograma construido de los datos de supervivencia de un grupo de individuos similares tiende a ser de “cola larga” a la derecha del intervalo que contiene la mayoría de las observaciones. En consecuencia, no es razonable suponer que los datos provienen de una población con distribución normal. Esta dificultad podría ser resuelta al transformar los datos para obtener una distribución simétrica, e.g. aplicando el logaritmo natural. Sin embargo, una metodología más satisfactoria es la de adoptar un modelo alternativo cuya distribución es más apropiada para los datos en términos de bondad de ajuste.

Una segunda característica de los datos de supervivencia la cual hace a los métodos estándares poco apropiados es que los tiempos de supervivencia frecuentemente están *censurados*. El tiempo de supervivencia de un individuo se dice censurado cuando el evento de interés no ha sido observado. Esto puede ocurrir cuando al final del periodo de estudio, varios individuos siguen vivos. Otra forma de censura es cuando el estatus de supervivencia de un individuo en el estudio no es bien sabido porque se ha perdido de vista al individuo. Por ejemplo, suponga que cierto individuo, después de ser reclutado para el estudio, se muda a otro país, o simplemente no se le puede encontrar. La única información disponible acerca de la supervivencia del individuo censurado es la última fecha en que se le vio con vida.

Un individuo que entra a un estudio en el tiempo t_0 muere en tiempo $t_0 + t$. Sin embargo, si el tiempo de supervivencia correspondiente es censurado, t es desconocido, ya sea porque el individuo sigue vivo o porque se la ha perdido de vista.

Si el individuo fue visto con vida por última vez en el tiempo $t_0 + c$, el tiempo c es conocido como el tiempo de supervivencia censurado por la derecha.

Para comenzar con la introducción al modelado estadístico de datos de supervivencia, es pertinente considerar algunas características relevantes de las distribuciones de probabilidad para el análisis; para esto, se partirá de la suposición de que la población es homogénea. A continuación examinan algunas especificaciones de la variable aleatoria positiva T , la cual está asociada con el tiempo para que ocurra el evento, y después se consideran varias distribuciones especiales que son de utilidad para el ajuste de los datos correspondientes.

Supóngase que la variable aleatoria T tiene una distribución de probabilidad cuya *función de densidad* se denota con $f(t)$. La *función de distribución* de T está dada por

$$F(t) = \Pr\{T < t\} = \int_0^t f(u)du,$$

y representa la probabilidad de que el evento de interés ocurra antes de del tiempo t .

La función de supervivencia, definida por $S(T) = \Pr\{T \geq t\}$, es la probabilidad de que el tiempo para que ocurra el evento sea mayor que el tiempo t , y entonces

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

La función de supervivencia entonces representa la probabilidad de que un individuo sobreviva desde el punto origen hasta algún punto mayor que t . Otra interpretación puede ser la proporción de individuos que sobreviven al tiempo t .

La *fuerza de mortalidad*, denotada con $h(t)$, (también conocida en inglés como *hazard function*) es la probabilidad de que un individuo muera en el tiempo t dado que ha sobrevivido hasta ese tiempo. La fuerza de mortalidad representa la tasa instantánea de mortalidad para un individuo que sobrevive al tiempo t . De esta

forma,

$$h(t) = \lim_{\delta t \rightarrow 0} \left[\frac{\Pr\{t \leq T < t + \delta t \mid T \geq t\}}{\delta t} \right].$$

Usando la función de distribución de T y la definición de probabilidad condicional, se puede demostrar que

$$h(t) = \lim_{\delta t \rightarrow 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}.$$

Aquí, es posible observar que

$$\lim_{\delta t \rightarrow 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\}$$

es la definición de la derivada de $F(t)$ con respecto a t , la cual es la función de densidad $f(t)$, y por lo tanto se tiene que

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

además

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

y entonces una expresión de la función de supervivencia en términos de la fuerza de mortalidad es

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

donde

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

es la *fuerza de mortalidad integrada*.

Una función $h(x)$ es una fuerza de mortalidad si y solo si satisface las siguientes propiedades:

1. $h(x) \geq 0$, para toda x .
2. $\int_0^\infty h(x) dx = \infty$.

Estas propiedades son necesarias ya que

$$h(x) = \frac{f(x)}{S(x)} \geq 0$$

y

$$\begin{aligned} \int_0^{\infty} h(x)dx &= \int_0^{\infty} -d[\log S(x)] \\ &= -\log S(x) \Big|_0^{\infty} \\ &= \infty. \end{aligned}$$

Estas propiedades son suficientes ya que la función de distribución resultante $F(x)$ es válida; que es, en términos de la fuerza de mortalidad $h(x)$:

$$F(-\infty) = F(0) = 1 - \exp \left[- \int_0^0 h(t)dt \right] = 0$$

y

$$F(\infty) = 1 - \exp \left[- \int_0^{\infty} h(t)dt \right] = 1,$$

y $F(x)$ es una función creciente de x ya que $\int_0^x h(t)dt$ es una función creciente de x .

En el análisis de datos de supervivencia, la función de supervivencia y la fuerza de mortalidad son las cantidades más relevantes para estimar.

1.1.1. El Modelo Exponencial

La distribución exponencial de parámetro λ y media $1/\lambda$ se caracteriza por

$$S(t) = \exp\{-\lambda t\}, \quad f(t) = \lambda \exp\{-\lambda t\}, \quad h(t) = \lambda, \quad H(t) = \lambda t,$$

donde $\lambda > 0$. La fuerza de mortalidad constante refleja la propiedad de la distribución exponencial conocida como la falta de memoria; esto es, si la variable

aleatoria T se distribuye exponencial con media $1/\lambda$, $T \sim \text{EXP}(\lambda)$, entonces $\Pr\{T > a + t | T > a\} = \Pr\{T > t\}$ para toda $a > 0$ y $t > 0$.

1.1.2. El Modelo Weibull

Supóngase que T se distribuye Weibull con parámetro de escala λ y parámetro de forma α , i.e. $T \sim \text{WEI}(1/\lambda, \alpha)$, cuya función de densidad es

$$f(t) = \alpha \lambda^\alpha t^{\alpha-1} \exp\{-(\lambda t)^\alpha\}, \quad \alpha, \lambda > 0.$$

la fuerza de mortalidad condicional asociada es

$$h(t) = \alpha \lambda (\lambda t)^{\alpha-1}. \quad (1.1)$$

$$S(t) = \exp\{-(\lambda t)^\alpha\}. \quad (1.2)$$

El parámetro λ en el modelo Weibull es aproximadamente inversamente proporcional a la mediana de los tiempos de supervivencia, ya que

$$\text{mediana de } T = \frac{(\log 2)^{1/\alpha}}{\lambda}.$$

Es interesante observar que esta cantidad depende hasta cierto punto de α pero el determinante principal es λ . La conveniencia del modelo Weibull para trabajo aplicado se debe a la simplicidad de las funciones $S(t)$ y $h(t)$. Nótese que cuando $\alpha = 1$, entonces se obtiene el modelo exponencial.

1.1.3. El Modelo del Valor Extremo

Si $T \sim \text{WEI}(1/\lambda, \alpha)$, entonces la transformación $Y = \log T$ tiene la distribución de valor extremo con función de densidad

$$f(y) = \frac{1}{\sigma} \exp\left[\frac{(y - \eta)}{\sigma} - \exp\left\{\frac{(y - \eta)}{\sigma}\right\}\right],$$

1.1. CARACTERÍSTICAS ESPECIALES DE LOS DATOS DE SUPERVIVENCIA 9

donde $\eta = -\lambda$ y $\sigma = 1/\alpha$. La función de supervivencia correspondiente es

$$S(y) = \exp[-\exp\{(y - \eta)/\sigma\}]$$

y la fuerza de mortalidad es

$$h(y) = \frac{1}{\sigma} \exp[(y - \eta)/\sigma].$$

Cuando $\eta = 0$ y $\sigma = 1$, la distribución de Y es el valor extremo estándar.

El Modelo Gompertz-Makeham

Una forma conveniente de la fuerza de mortalidad es

$$h(t) = \rho_0 + \rho_1 \exp\{\rho_2 t\}, \quad \rho_1, \rho_2 > 0, \quad \rho_0 \geq -\rho_1,$$

cuya función de supervivencia es

$$S(t) = \exp\left\{-\rho_0 t - \frac{\rho_1}{\rho_2} [\exp\{\rho_2 t\} - 1]\right\}.$$

Esta especificación es conocida como el modelo Gompertz. Cuando $\rho_0 = 0$ como caso especial, el modelo es conocido como Gompertz.

1.1.4. El Modelo Lognormal

La distribución lognormal está definida por la función de densidad

$$f(t) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\}}{t\sqrt{2\pi\sigma^2}} = \phi\left(\frac{\log t - \mu}{\sigma}\right) / t,$$

donde $t > 0$, $-\infty < \mu < \infty$ y $\sigma > 0$, y su función de supervivencia está dada por

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

La variable aleatoria correspondiente, la cual se denota con $T \sim \text{LOGN}(\mu, \sigma^2)$, está relacionada con la distribución normal ya que $T \sim \text{LOGN}(\mu, \sigma^2)$ si y solo si $X = \log T \sim N(\mu, \sigma^2)$.

La distribución exponencial no es un caso especial de la lognormal. La fuerza de mortalidad asociada con la lognormal no es monótona. Una desventaja de este modelo es que el ajuste es muy sensitivo a observaciones de tiempos cortos.

1.1.5. El Modelo de Pedazos Exponenciales

En muchas ocasiones, el patrón de dependencia de tiempo en los modelos especificados en forma completamente paramétrica, tales como el Weibull o el Gompertz, no siempre ajustan bien los datos. En tales casos, un *modelo de pedazos exponenciales* (en inglés *piece-wise exponential model*) bien puede constituir un esquema mas apropiado. Este modelo es especificado con la fuerza de mortalidad

$$h(t) = \exp(\kappa_m) \quad \text{si } t \in A_m,$$

donde $A_m = [a_{m-1}, a_m)$, $m = 1, \dots, M$, son intervalos mutuamente excluyentes (previamente especificados) que cubren en forma exhaustiva la recta real positiva, i.e. $a_0 = 0$, $a_m > a_{m-1}$ para $m > 0$ y $a_M = \infty$. De esta forma, $h(t)$ es constante en cada uno de los intervalos y κ_m representa la fuerza de mortalidad en cada uno de ellos. La fuerza de mortalidad integrada correspondiente puede expresarse como

$$H(t) = \begin{cases} te^{\kappa_1} & : \text{ si } t \in A_1 \\ \sum_{l=1}^{m-1} (a_l - a_{l-1})e^{\kappa_l} + (t - a_{m-1})e^{\kappa_m} & : \text{ si } t \in A_m, m \geq 2. \end{cases}$$

Capítulo 2

Procedimientos No Paramétricos

Un paso inicial en el análisis de datos de supervivencia es presentar resúmenes gráficos o numéricos de los tiempos de supervivencia para unidades experimentales en cierto grupo. Tales resúmenes pueden dar pauta a un análisis más detallado de los datos. Los datos de supervivencia pueden ser resumidos convenientemente a través de estimadores de la función de supervivencia y de la fuerza de mortalidad. Los métodos para encontrar estos estimadores son llamados *no paramétricos* o *libres de distribución*, ya que no necesitan que se hagan suposiciones específicas sobre la distribución de los tiempos de supervivencia.

2.1. Estimación de la Función de Supervivencia

Supóngase que se tiene una muestra de tiempos de supervivencia donde ninguna de las observaciones tiene censura. La función de supervivencia es la probabilidad de que un individuo sobreviva por un tiempo mayor o igual a t . Esta función se

puede estimar por medio de la *función de supervivencia empírica*:

$$\tilde{S}(t) = \frac{\text{Número de individuos con tiempos de supervivencia } \geq t}{\text{Número total de individuos en los datos}}.$$

En forma equivalente, $\tilde{S}(t) = 1 - \tilde{F}(t)$, donde $\tilde{F}(t)$ es la función de distribución empírica., que es el cociente del número total de individuos vivos en el tiempo t entre el número total de individuos en el estudio. Nótese que $\tilde{S}(t) = 1$ para $t < t_{(1)}$, donde $t_{(1)}$ representa la observación mas chica; además, $\tilde{S}(t) = 0$ para $t \geq t_{(n)}$, donde $t_{(n)}$ es la observación mas grande.

El método para estimar la función de supervivencia usando el cociente no se puede usar cuando hay observaciones con censura. A continuación se describen algunos métodos no paramétricos que permiten estimar $S(t)$ en presencia de datos censurados.

2.1.1. El Estimador Actuarial

El *estimador actuarial*, o la *tabla de vida*, de una función de supervivencia se obtiene al dividir el periodo de observación en una serie de intervalos de tiempo. Estos intervalos no deben de ser de igual magnitud aunque, en general, lo son. Supóngase que el j -ésimo intervalo de un total de m intervalos, $j = 1, 2, \dots, m$, abarca de t'_j a t'_{j+1} , y sean d_j y c_j el número de muertes y el número de tiempos de supervivencia censurados respectivamente, en este intervalo. Sea n_j el número de individuos que estan vivos, y por lo tanto en riesgo de morir, al principio del j -ésimo intervalo. Partiendo de la suposición de que el proceso de censura es tal que los tiempos de supervivencia ocurren uniformemente durante el j -ésimo intervalo, de manera tal que el número promedio de individuos que estan en riesgo durante este intervalo es

$$n'_j = n_j - c_j/2,$$

el cual representa el número ajustado de individuos en riesgo.

En el j -ésimo intervalo la probabilidad de muerte se puede estimar con d_j/n'_j , y entonces la probabilidad de supervivencia correspondiente es $(n'_j - d_j)/n'_j$. Ahora, la probabilidad de que un individuo sobreviva después del tiempo t'_k es el producto de las probabilidades de que un individuo sobreviva los $k - 1$ intervalos anteriores, y entonces el estimador actuarial es

$$S^*(t) = \prod_{j=1}^k \left(\frac{n'_j - d_j}{n'_j} \right), \quad t \in [t'_k, t'_{k+1}),$$

para $k = 1, \dots, m$. Aquí se puede comprobar que $S^*(t) = 1$ para $t < t_{(1)}$, y que $S^*(t) = 0$ para $t \geq t_{(n)}$.

El estimador actuarial es sensible a la elección de los intervalos usados, exactamente en la misma forma en que la gráfica de un histograma depende de la elección de los intervalos de clase. El estimador actuarial es adecuado en situaciones en las que los tiempos de las muestras son desconocidos, y la única información disponible es el número de muertes y el número de observaciones censuradas que ocurren en una serie de intervalos de tiempo.

Cuando los tiempos de supervivencia son conocidos, el estimador actuarial se puede usar pero la agrupación de los intervalos conlleva a la pérdida de información.

Ejemplo 2.1. Los datos en la Tabla 2.1, obtenidos de Krall, Uthoff y Harley (1975), corresponden a 48 pacientes con edades de 50 a 80 años de edad. Varios de estos pacientes no habían muerto al final del estudio, por lo que las observaciones correspondientes son tiempos censurados por la derecha. Por el momento solo será necesario concentrarse en las columnas **time**, los tiempos de supervivencia (en meses), y **status**, el estatus de censura (1=sin censura, 0=con censura).

Los tiempos de supervivencia se agrupan para obtener el número de pacientes que mueren d_j , y el número que está censurado c_j , en cada uno de los cinco años del estudio, y después en los tres restantes. El número de individuos en riesgo al

Tabla 2.1: Tiempos de supervivencia (en meses) de pacientes en un estudio sobre mieloma múltiple.

patient	time	status	age	sex	bun	ca	hb	pcells	protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11	16	1
34	5	1	77	1	23	8	9	29	0
35	10	1	61	1	13	10	14	19	0
36	91	1	58	2	27	11	11	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	17	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

Tabla 2.2: Estimador Actuarial de la función de supervivencia para los datos de mieloma múltiple.

Intervalo	Periodo	d_j	c_j	n_j	n'_j	$(n'_j - d_j)/n'_j$	$S^*(t)$
1	[0, 12)	16	4	48	46.0	0.6521	0.6522
2	[12, 24)	10	4	28	26.0	0.6154	0.4013
3	[24, 36)	1	0	14	14.0	0.9286	0.3727
4	[36, 48)	3	1	13	12.5	0.7600	0.2832
5	[48, 60)	2	2	12	11.0	0.8181	0.2317
6	[60, ∞)	4	1	6	5.5	0.2727	0.0632

principio de cada intervalo n_j se calcula junto con el número ajustado de individuos en riesgo n'_j . Finalmente, se estima la probabilidad de muerte en cada intervalo, y estas cantidades se usan para obtener el estimador de $S(t)$. Los cálculos se muestran en la Tabla 2.2.

Cuando se usa el lenguaje de programación R para obtener la Tabla... es posible usar la función `survfit`. El procedimiento se muestra a continuación:

```
# Invoca la libreria de supervivencia
library(survival)

# Pon en un objeto los datos
mye <- read.table("C:/Mis documentos/multmye.dat",header=T)

# Pon en un objeto cantidades sobre los datos
mye.surv <- survfit(Surv(time,status),data=mye)

# los diferentes tiempos de supervivencia sin repeticion
tiempos <- mye.surv$time

# el numero de individuos en riesgo en cada tiempo:
n.riesgo <- mye.surv$n.risk

# el numero de eventos (muertes) en cada tiempo de supervivencia
n.eventos <- mye.surv$n.event

# Encuentra las d_j
```

```
limites <- c(12, 24, 36, 48, 60, Inf)
M <- length(limites)
sum.n.vent <- 1:M
  <- 1:M
for(i in 1:M){
sum.n.vent[i] <- sum(n.eventos[tiempos < limites[i]])}

lag.sum <- 1:M
lag.sum[1] <- 0
lag.sum[2:M] <- sum.n.vent[1:(M-1)]

d.j <- sum.n.vent - lag.sum

# Calcula n_j
n.j <- 1:M
n.j[1] <- n.riesgo[1]
for(i in 1:(M-1)){
n.j[1+i] <- rev(n.riesgo[tiempos <= limites[i]])[1]
}

# Calcula c_j

cum.c.j <- 1:M
for (i in 1:M){
cum.c.j[i] <- sum(1 - mye$status[mye$time < limites[i]])
}

lag.cum.c.j <- 1:M
lag.cum.c.j[1] <- 0
lag.cum.c.j[2:M] <- cum.c.j[1:(M-1)]

c.j <- cum.c.j - lag.cum.c.j

n.j.prim <- n.j - c.j/2

p.j <- (n.j.prim - d.j)/n.j.prim

S.j <- cumprod(p.j)
```

El Estimador Kaplan-Meier

La determinación del estimador Kaplan-Meier de la función de supervivencia de una muestra que incluye tiempos censurados es parecida a la del estimador actuarial. Sin embargo, cada uno de los intervalos se forma de manera tal que cada tiempo de muerte determina los límites.

Supóngase que hay n individuos y sus tiempos de supervivencia correspondientes son t_1, t_2, \dots, t_n . Algunas de estas observaciones pueden estar censuradas por la derecha, y es posible que existan observaciones con el mismo tiempo de supervivencia observado. Supóngase que hay r tiempos de muertes observadas (observaciones sin censura), con $r \leq n$. Después de ordenar las observaciones en orden ascendente, el j -ésimo tiempo observado se denota $t_{(j)}$, para $j = 1, \dots, r$; de esta forma, los r tiempos de muerte ordenados son $t_{(1)} < t_{(2)} < \dots < t_{(r)}$. El número de individuos que se encuentran con vida antes del tiempo $t_{(j)}$, incluyendo los que están por morir en este tiempo, se denotan con n_j , y d_j denota el número que muere en este tiempo. El intervalo que va de $t_{(j)} - \delta$ a $t_{(j)}$, donde δ es un tiempo infinitesimal, incluye un tiempo de una muerte. Como hay n_j individuos que sobreviven poco antes de $t_{(j)}$, y hay d_j muertes en $t_{(j)}$, la probabilidad de que un individuo muera durante el intervalo que va de $t_{(j)} - \delta$ a $t_{(j)}$ se estima con d_j/n_j . La probabilidad de supervivencia estimada en el intervalo es entonces $(n_j - d_j)/n_j$.

Es posible que varios tiempos de supervivencia censurados ocurran al mismo tiempo que uno o más muertes, de manera tal que el tiempo de muerte y los tiempos censurados ocurren simultáneamente. En este caso, cuando se calcula n_j , los tiempos de supervivencia censurados se toman como si ocurrieran inmediatamente después de del tiempo de la muerte.

En la forma en que se construyen los intervalos de tiempo, el intervalo que va de $t_{(j)}$ a $t_{(j+1)} - \delta$, que es el tiempo inmediatamente anterior al siguiente tiempo

de muerte, no contiene ninguna muerte. Por lo tanto, la probabilidad de sobrevivir de $t_{(j)}$ a $t_{(j+1)} - \delta$ es uno, y la probabilidad conjunta de sobrevivir de $t_{(j)} - \delta$ a $t_{(j)}$ y de $t_{(j)}$ a $t_{(j+1)} - \delta$ se puede estimar con $(n_j - d_j)/n_j$. Tomando el límite $\delta \rightarrow 0$, $(n_j - d_j)/n_j$ se convierte en un estimador de la probabilidad de supervivencia de $t_{(j)}$ a $t_{(j+1)}$.

Suponiendo que la ocurrencia de los eventos en la muestra es independiente de individuo a individuo, el estimador de la función de supervivencia en cualquier intervalo que va de $t_{(k)}$ a $t_{(k+1)}$, $k = 1, \dots, r$, donde $t_{(r+1)} = \infty$, se calcula como la probabilidad de supervivencia después de $t_{(k)}$. Esta es la probabilidad de sobrevivir en el intervalo $t_{(k)}$ a $t_{(k+1)}$ y todos los intervalos anteriores. Este es el *estimador Kaplan-Meier* de la función de supervivencia, el cual está dado por

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

para $k = 1, \dots, r$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$. Si la observación mas grande es un tiempo censurado t^* , entonces $\hat{S}(t)$ no está definido para $t > t^*$. Por otra parte, si la observación más grande es un tiempo sin censura, i.e. $t_{(r)}$, entonces $n_r = d_r$, y por lo tanto $\hat{S}(t)$ es cero para $t \geq t_{(r)}$.

La gráfica del estimador Kaplan-Meier de la función de supervivencia es una función escalonada, en la cual las probabilidades de supervivencia son constantes entre los tiempos de muerte adyacentes y decrece en cada tiempo de de muerte. El estimador Kaplan-Meier es también conocido como el *estimador de producto límite* (en inglés, *product limit estimator*).

Nótese que si no hay tiempos con censura en los datos, entonces $n_j - d_j = n_{j+1}$, $j = 1, 2, \dots, k$, y entonces el estimador Kaplan-Meier se expresa como

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times \dots \times \frac{n_{k+1}}{n_k}.$$

Esto se reduce a $\hat{S}(t) = n_{k+1}/n_1$, para $k = 1, 2, \dots, r - 1$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$

Tabla 2.3: Tiempo en semanas para la discontinuidad del DIU.

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

Tabla 2.4: Estimador Kaplan-Meier para los datos DIU

Intervalo	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0-	18	0	1.0000	1.0000
10-	18	1	0.9444	0.9444
19-	15	1	0.9333	0.8815
30-	13	1	0.9231	0.8137
36-	12	1	0.9167	0.7459
59-	8	1	0.8750	0.6526
75-	7	1	0.8571	0.5594
93-	6	1	0.8333	0.4662
97-	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

y $\hat{S}(t) = 0$ para $t \geq t_{(r)}$. Como n_1 es el número de individuos en riesgo antes de la primera muerte, i.e. $n_1 = n$, y $n_{k+1} = 0$ es el número de individuos con tiempos de supervivencia mayores o iguales a t_{k+1} , entonces $\hat{S}(t)$ es simplemente la función de supervivencia empírica.

Ejemplo 2.2. Los datos en la Tabla 2.3 se refieren al número de semanas desde el comienzo de un dispositivo intrauterino (DIU) hasta su discontinuidad (WHO, 1987). Los datos están dados para 18 mujeres quienes tienen edades de 18 a 35 años y han tenido dos embarazos. Los tiempos de discontinuidad que son censuradas tienen un asterisco. El estimador Kaplan-Meier se muestra en la Tabla 2.4. La gráfica de $\hat{S}(t)$ se observa en la Figura 2.1. Nótese que como el tiempo mas largo tiene censura, $\hat{S}(t)$ no está definido después de $t = 107$.

El language R incluye en su librería `survival` funciones listas para tabular y graficar el estimador Kaplan-Meier. El procedimiento se describe a continuación:

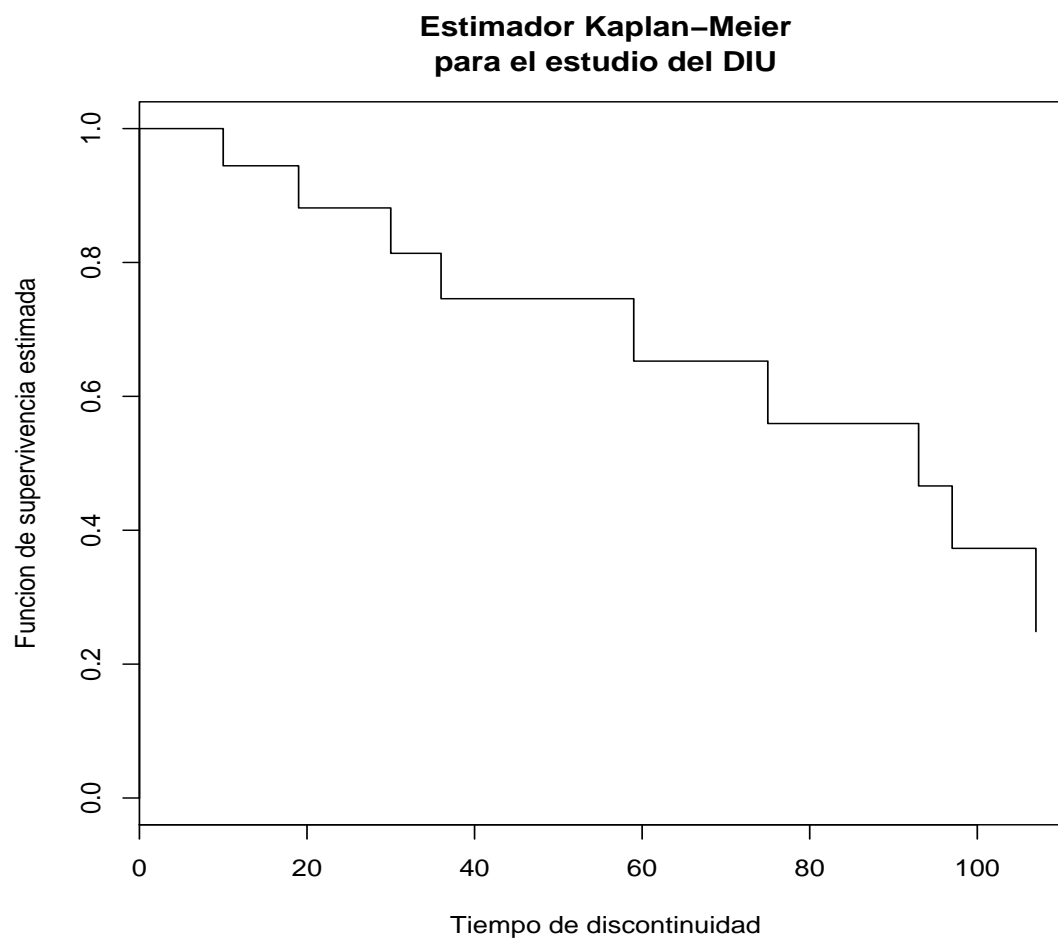


Figura 2.1: Estimador Kaplan-Meier de la función de supervivencia para los datos del DIU.

```
# Pon los datos en el formato data.frame
diu <- data.frame(tiempo=c(10, 13, 18, 19, 23, 30, 36, 38, 54,
                          56, 59, 75, 93, 97, 104, 107, 107, 107),
                  estatus=c(1, 0, 0, 1, 0, 1, 1, 0, 0,
                            0, 1, 1, 1, 1, 0, 1, 0, 0))

library(survival)
```

```
# Encuentra el estimador Kaplan-Meier
diu.surv <- survfit(Surv(tiempo, estatus), data=diu)

# Muestra el estimador y otras cantidades importantes
summary(diu.surv)

# Obten una grafica del estimador
plot(diu.surv, conf.int=FALSE, mark.time=FALSE)
title(main="Estimador Kaplan-Meier\npara el estudio del DIU",
      xlab="Tiempo de discontinuidad",
      ylab="Funcion de supervivencia estimada")
```

2.1.2. Error estándar para el estimador Kaplan-Meier

Como Kaplan-Meier es el estimador de la función de supervivencia más importante y ampliamente usado, es importante derivar el error estándar de $\hat{S}(t)$, se deriva en esta sección.

El estimador Kaplan-Meier de la función de supervivencia evaluada en t , donde t se encuentra en intervalo que va de $t_{(k)}$ a $t_{(k+1)}$, se puede expresar como

$$\hat{S}(t) = \prod_{j=1}^k \hat{p}_j, \quad k = 1, \dots, r,$$

donde $\hat{p}_j = (n_j - d_j)/n_j$ es la probabilidad estimada de que un individuo sobreviva en el intervalo de tiempo que comienza en $t_{(j)}$, $j = 1, \dots, r$. Si se toma el logaritmo natural, se tiene que

$$\log \hat{S}(t) = \prod_{j=1}^k \log \hat{p}_j,$$

y, como los elementos en la muestra son independientes, la varianza de $\log \hat{S}(t)$ está dada por

$$\text{var} \left\{ \log \hat{S}(t) \right\} = \sum_{j=1}^k \text{var} \{ \log \hat{p}_j \}.$$

El número de individuos que sobreviven al intervalo que comienza en $t_{(j)}$ puede tomarse como una variable aleatoria binomial con parámetros n_j y p_j ; i.e. n_j es el número de realizaciones Bernoulli, y p_j es la probabilidad de éxito (de sobrevivir). De esta forma, la varianza del número de individuos que sobrevive $S_j = n_j - d_j$, $S_j \sim \text{BIN}(n_j, p_j)$, está dada por

$$\text{Var}[S_j] = \text{Var}[n_j - d_j] = n_j p_j (1 - p_j).$$

Como $\hat{p}_j = (n_j - d_j)/n_j$, la varianza de \hat{p}_j es $\text{Var}[S_j]/n_j^2$, que es $p_j(1 - p_j)/n_j$. Por lo tanto, la varianza de \hat{p}_j se puede estimar con

$$\hat{p}_j(1 - \hat{p}_j)/n_j.$$

Para obtener la varianza de $\log \hat{p}_j$, se hace uso del resultado general para aproximar la varianza de una función de una variable aleatoria. De acuerdo a este resultado, la varianza de la función $g(X)$ de la variable aleatoria X está dada por

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

Este resultado es conocido como la *aproximación de la serie de Taylor* de la varianza de una función de una variable aleatoria. Usando este resultado, la varianza aproximada de $\log \hat{p}_j$ es $\text{Var}[\hat{p}_j]/\hat{p}_j^2$; por lo que un estimador aproximado de la varianza de $\log \hat{p}_j$ es $(1 - \hat{p}_j)/(n_j \hat{p}_j)$, el cual en sustitución de \hat{p}_j se reduce a

$$\frac{d_j}{n_j(n_j - d_j)}.$$

Se tiene entonces que

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

y usando de nueva cuenta la aproximación de la serie de Taylor, se tiene que

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

y de esta forma

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

Finalmente, el error estándar del estimador Kaplan Meier de la función de supervivencia, el cual se define como la raíz cuadrada de la varianza estimada del estimador, está dado por

$$\text{s.e.} \left[\hat{S}(t) \right] \approx \hat{S}(t) \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}}, \quad \text{para } t_{(k)} \leq t < t_{(k+1)}.$$

A este resultado se le conoce como la *fórmula de Greenwood*.

Si no hay tiempos de supervivencia con censura, entonces $n_j - d_j = n_{j+1}$ y el estimador de $\text{Var}[\log \hat{p}_j]$ es $(n_j - n_{j+1}) / (n_j n_{j+1})$. El valor aproximado de $\text{Var}[\log \hat{S}(t)]$ es

$$\sum_{j=1}^k \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^k \left(\frac{1}{n_{j+1}} - \frac{1}{n_j} \right) = \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

el cual se puede expresar como

$$\frac{1 - \hat{S}(t)}{n_1 \hat{S}(t)},$$

ya que $\hat{S}(t) = n_{k+1}/n_1$ para $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \dots, r - 1$, cuando no hay censura. Usando la aproximación de la serie de Taylor, el estimador de la varianza de $\hat{S}(t)$ es $\hat{S}(t)[1 - \hat{S}(t)]/n_1$.

2.1.3. Intervalos de confianza para valores de la función de supervivencia

Una vez que se ha calculado el error estándar de $\hat{S}(t)$ es posible encontrar un *intervalo de confianza* para el valor correspondiente a $\hat{S}(t)$. Un intervalo de confianza es un estimador por intervalos de la función de supervivencia, y el intervalo se forma de manera tal que el valor verdadero de la función de supervivencia está dentro de sus límites dada una probabilidad de que la contenga, la cual es determinada con anterioridad.

Un intervalo de confianza para el valor verdadero de la función de supervivencia en el tiempo t se puede obtener al suponer que el valor estimado de la función de supervivencia en t se distribuye normal con media $S(t)$ y varianza estimada dada por el cuadrado de la fórmula de Greenwood. El intervalo se calcula con los puntos porcentuales de una distribución normal estándar. De esta forma, si $z_{1-\alpha/2}$ denota el percentil de una distribución normal estándar al nivel $1 - \alpha/2$, i.e. $\Pr\{Z < z_{1-\alpha/2}\} = 1 - \alpha/2$ con $Z \sim N(0, 1)$, entonces un intervalo de confianza al coeficiente

de confianza $1 - \alpha$ para $S(t)$ es el intervalo con límites

$$\hat{S}(t) \pm z_{1-\alpha/2} \text{s.e.}[\hat{S}(t)].$$

Un inconveniente de este procedimiento es que los intervalos de confianza son simétricos. Cuando se obtienen límites para valores de $\hat{S}(t)$ cercanos a cero o uno, los intervalos simétricos no son adecuados pues sus valores se encuentran fuera del intervalo $[0, 1]$. Una forma pragmática de solucionar el problema es reemplazar los límites que exceden uno por 1, los que son inferiores a cero por 0.

Una procedimiento alternativo es transformar $\hat{S}(t)$ para encontrar un valor que se encuentre en el rango $(-\infty, \infty)$, y entonces encontrar un intervalo de confianza para el valor transformado. El intervalo de confianza resultante es entonces transformado de nuevo para obtener el intervalo de confianza para $S(t)$. Transformaciones posibles incluyen la logística, dada por $\log\{S(t)/[1 - S(t)]\}$, y la log-log, dada por $\log\{-\log \hat{S}(t)\}$. Para la transformación log-log, se puede usar el resultado general

$$\text{Var}[\log(-X)] \approx \frac{1}{X^2} \text{Var}[X],$$

y al hacer la sustitución $X = \log \hat{S}(t)$ se obtiene que

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

El error estándar (s.e.) de $\log\{-\log \hat{S}(t)\}$ es la raíz cuadrada de esta cantidad. La aproximación

$$\frac{\log\{-\log \hat{S}(t)\} - \log\{-\log S(t)\}}{\text{s.e.} \left[\log\{-\log \hat{S}(t)\} \right]} \sim N(0, 1)$$

permite determinar límites de $100(1 - \alpha)\%$, que son

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

Ejemplo 2.3. El error estándar y los límites de confianza de correspondiente a los datos del DIU en la Tabla 2.3 se muestran en la Tabla 2.5. En estas tabla, los

Tabla 2.5: Errores estándares e intervalos de confianza de $\hat{S}(t)$ para los datos del DIU.

Intervalo	$\hat{S}(t)$	s.e. $[\hat{S}(t)]$	I. de C. al 95 %
0-	1.0000	0.0000	
10-	0.9444	0.0540	(0.666, 0.992)
19-	0.8815	0.0790	(0.602, 0.969)
30-	0.8137	0.0978	(0.524, 0.936)
36-	0.7459	0.1107	(0.454, 0.897)
59-	0.6526	0.1303	(0.344, 0.843)
75-	0.5594	0.1412	(0.256, 0.780)
93-	0.4662	0.1452	(0.183, 0.710)
97-	0.3729	0.1430	(0.121, 0.631)
107	0.2486	0.1392	(0.047, 0.531)

errores estándares han sido calculados con la fórmula de Greenwood y los intervalos de confianza se han calculado usando la transformación log-log.

En esta tabla se puede observar que en general el error estándar del estimado de la función de supervivencia se incrementa con el tiempo. La razón de esto es que los estimadores en los últimos tiempos se basan en menos observaciones. Una gráfica de la función de supervivencia junto con los límites de confianza se muestra en la Figura 2.2.

El procedimiento para encontrar la Tabla 2.5 y la Figura 2.2 en el lenguaje R se da a continuación:

```
# Pon los datos en el formato data.frame
diu <- data.frame(tiempo=c(10, 13, 18, 19, 23, 30, 36, 38, 54,
                          56, 59, 75, 93, 97, 104, 107, 107, 107),
                 estatus=c(1, 0, 0, 1, 0, 1, 1, 0, 0,
                          0, 1, 1, 1, 1, 0, 1, 0, 0))

# Invoca la libreria de supervivencia
library(survival)

# Obten el estimador Kaplan-Meier como un objeto de tipo survfit
```

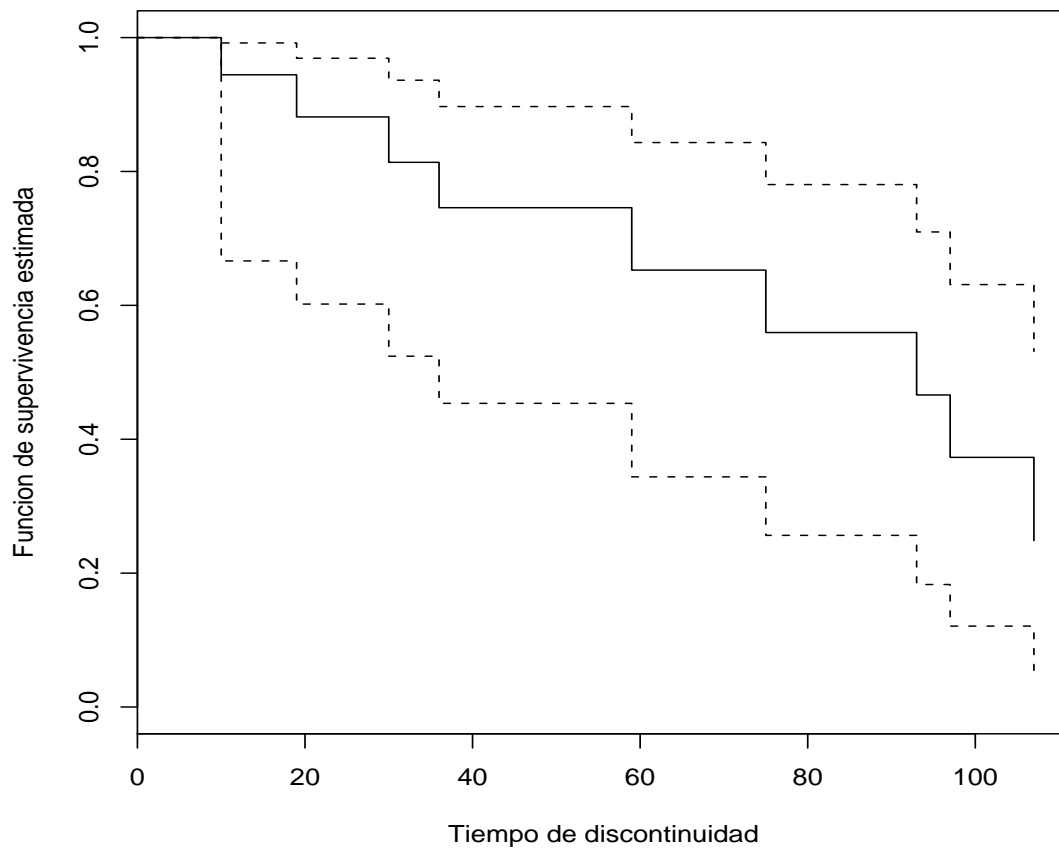


Figura 2.2: Estimador Kaplan-Meier y bandas de confianza al 95 % de la función de supervivencia para los datos del DIU.

```
# usando la formula de Greenwood para el error estandar de la curva
# y bandas de confianza del tipo "log-log".
diu.surv <- survfit(Surv(tiempo, estatus), data=diu,
error="greenwood", conf.type="log-log")

# Muestra los resultados
summary(diu.surv)

# Grafica el estimador Kaplan-Meier con sus bandas de confianza:
plot(diu.surv, conf.int=TRUE, mark.time=FALSE)
title(xlab="Tiempo de discontinuidad",
      ylab="Funcion de supervivencia estimada")
```

2.2. Comparación de tiempos de supervivencia para dos grupos

La forma más simple de comparar los tiempos de supervivencia de dos grupos de individuos es graficar los estimadores de las dos funciones de supervivencia correspondientes. Es posible que exista una diferencia real entre las dos curvas, lo que indica que un grupo tiene una supervivencia diferente a la del otro. Por otra parte, es posible ver pocas diferencias reales entre los grupos y que estas pequeñas diferencias sean resultado de variaciones sin mayor explicación.

Para ayudar a distinguir si existen en verdad diferencias significativas en los dos grupos, es posible llevar a cabo una prueba de hipótesis. En esta sección es pertinente concentrarse en la *prueba log-rank*. Para construir este procedimiento, se debe de comenzar por considerar los tiempos de supervivencia de los dos grupos por separado. Los grupos se etiquetarán como Grupo I y Grupo II.

Supóngase que hay r tiempos de muertes observadas en los dos grupos, $t_{(1)} < t_{(2)} < \dots < t_{(r)}$, y que en el tiempo $t_{(j)}$ hay d_{1j} muertes en el grupo I y d_{2j} muertes en el grupo II, $j = 1, 2, \dots, r$. Supóngase además que hay n_{1j} individuos en el primer grupo en riesgo de morir poco antes del tiempo $t_{(j)}$, y que hay n_{2j} en el segundo grupo. De esta forma, hay $d_j = d_{1j} + d_{2j}$ muertes en $t_{(j)}$ de un total de $n_j = n_{1j} + n_{2j}$ individuos en riesgo. Este escenario está resumido en la Tabla 2.6.

Considérese la hipótesis nula de que no hay diferencia entre los tiempos de supervivencia de los dos grupos. Una forma de evaluar la validez de esta hipótesis es considerar las desviaciones entre el número observado de individuos en los dos grupos que mueren en cada tiempo de muerte, y el número esperado bajo la hipótesis

Tabla 2.6: Número de muertes en el j -ésimo tiempo de muerte en cada uno de los grupos de individuos.

Grupo	Número de muertes en $t_{(j)}$	Número de sobrevivientes después de $t_{(j)}$	Número en riesgo poco antes de $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

nula. La información de estas desviaciones pueden entonces combinarse con todas los tiempos de muerte.

Si los totales marginales en la Tabla 2.6 se toman como fijos, y si la hipótesis nula es verdadera, entonces las cuatro entradas de la tabla son determinadas solamente por el número de muertes en el grupo I en $t_{(j)}$, d_{1j} . De esta forma es posible tomar a d_{1j} como una variable aleatoria, la cual puede tomar valores entre cero y el mínimo de d_j y n_{1j} . Bajo estas circunstancias, d_{1j} se distribuye *hipergeométrica*, $d_{1j} \sim \text{HIP}(n_{1j}, d_j, n_j)$, y entonces la función de probabilidad correspondiente es la probabilidad de que el número de muertes en el primer grupo tome el valor d_{1j} de un total de n_j individuos en riesgo cuando hay $n_j - d_j$ individuos que sobreviven y n_{1j} individuos en el primer grupo que están en riesgo de morir, que es

$$\frac{\binom{d_j}{d_{1j}} \binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}.$$

La media de la variable aleatoria hipergeométrica d_{1j} está dada por

$$e_{1j} = n_{1j}d_j/n_j,$$

que es el número esperado de individuos que mueren en el grupo I en $t_{(j)}$. Ahora, bajo la hipótesis nula de que la probabilidad de muerte en $t_{(j)}$ no depende del

grupo al que un individuo pertenece, la probabilidad de muerte en $t_{(j)}$ es d_j/n_j , y multiplicando esta cantidad por n_{1j} se obtiene e_{1j} .

El siguiente paso es obtener un estadístico que resuma las desviaciones entre lo observado y lo esperado. La forma más simple de hacerlo es calcular la suma de las diferencias $d_{1j} - e_{1j}$ para $r = 1, 2, \dots, r$. El estadístico resultante es

$$U_L = \sum_{j=1}^r (d_{1j} - e_{1j}).$$

Este estadístico tiene esperanza cero ya que $E[d_{1j}] = e_{1j}$. La varianza del estadístico U_L es

$$V_L = \text{Var}[U_L] = \sum_{j=1}^r v_{1j},$$

donde v_{1j} es la varianza de d_{1j} dada por

$$v_{1j} = \text{Var}[d_{1j}] = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

Cuando el número de muertes es relativamente grande U_L se distribuye aproximadamente normal y entonces $U_L/\sqrt{V_L}$ se distribuye aproximadamente normal estándar; esto es

$$\frac{U_L}{\sqrt{V_L}} \sim N(0, 1) \quad \text{cuando } n \rightarrow \infty.$$

El cuadrado de esta variable aleatoria se distribuye asintóticamente ji cuadrada con un grado de libertad:

$$W_L = \frac{U_L^2}{V_L} \sim \chi^2(1).$$

Por lo tanto, W_L resume las desviaciones que hay entre los tiempos de supervivencia observados en los dos grupos y los esperados bajo la hipótesis nula de no diferencias entre los dos grupos. Entre más grandes sean los valores del estadístico W_L mayor será la evidencia en contra de la hipótesis nula. La hipótesis nula se rechaza si la W_L observada con los datos excede el valor crítico $\chi_{1-\alpha}$, donde α es el nivel de significancia y $\chi_{1-\alpha}$ es el percentil $1 - \alpha$ de una ji cuadrada con un grado de libertad.

Tabla 2.7: Tiempos de supervivencia de mujeres con log.rank tumores que registraron marcas positivas y negativas de HPA.

Marca negativa	Marca positiva	
23	5	68
47	8	71
69	10	76*
70*	13	105*
71*	18	107*
100*	24	109*
101*	26	113
148	26	116*
181	31	118
198*	35	143
208*	40	154*
212*	41	162*
224*	48	188*
	50	212*
	59	217*
	61	225*

El valor p correspondiente se calcula como

$$\text{valor } p = \Pr\{\chi^2 > W_L\}, \quad \text{donde } \chi^2 \sim \chi^2(1).$$

Ejemplo 2.4. La Tabla 2.7 muestra tiempos de supervivencia (en meses) de mujeres que recibieron una mastectomía para tratar un tumor de grado II, III, IV, entre enero de 1969 y diciembre de 1971. Los tiempos de supervivencia se encuentran clasificados de acuerdo a si el cáncer fué o no marcado con *Helix pomatia agglutinin* (HPA), una marca que indica si el cáncer mamario primario tiene metastásis o no.

El interés principal en este estudio es el de determinar si hay una diferencia significativa entre los tiempos de supervivencia de los dos grupos. La Figura 2.3 muestra los estimadores Kaplan-Meier de las funciones de supervivencia para cada grupo. Es posible notar que existen ciertas diferencias, las mujeres con una marca negativa de HPA tienden a sobrevivir más que las que tienen una marca positiva.

Las cantidades relevantes para construir el estadístico de prueba se muestran en la Tabla 2.8. En este caso se obtiene $\sum d_{1j} = 5$, $\sum e_{1j} = 9,565$, $V_L = 5,929$. Por lo tanto, el valor observado del estadístico W_L es de 3.515. El valor p correspondiente es de 0.0608, lo que inspira duda sobre la hipótesis nula de que no hay diferencia entre

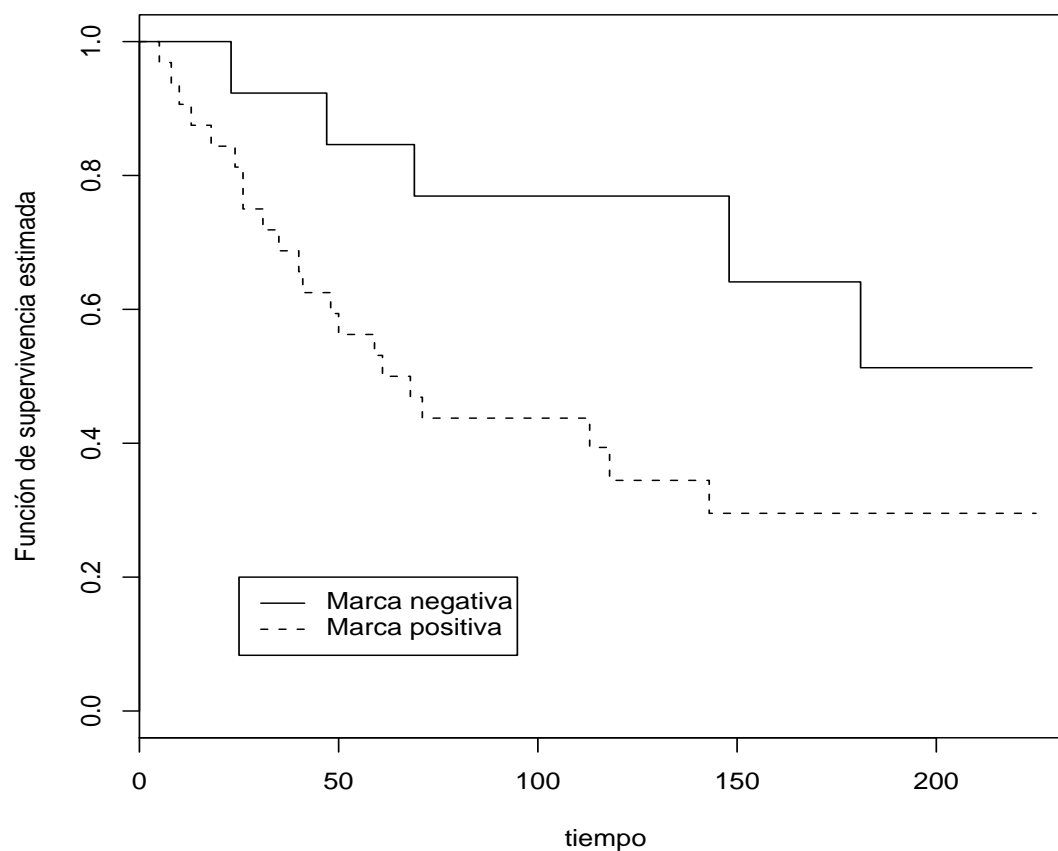


Figura 2.3: Estimador Kaplan-Meier de la función de supervivencia para los datos de cáncer de mujeres clasificado por marca de HPA.

las funciones de supervivencia de los dos grupos de mujeres. Se puede, entonces, concluir que hay cierta evidencia de que el diagnóstico de una paciente de cáncer mamario depende del resultado del estatus de la marca de HPA.

En el lenguaje R de programación es posible llevar a cabo este análisis usando la función `survdiff`. A continuación se ilustra el ejemplo de los datos en la Tabla 2.7.

```
#Escribe los datos en un objeto data.frame:
hpa <- data.frame(tiempo=c(23, 47, 69, 70, 71, 100, 101, 148, 181,
                          198, 208, 212, 224,
                          5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50,
                          59, 61, 68, 71, 76, 105, 107, 109, 113, 116, 118,
```

2.2. COMPARACIÓN DE TIEMPOS DE SUPERVIVENCIA PARA DOS GRUPOS³³

Tabla 2.8: El cálculo del estadístico log-rank para los datos de cancer con marcas positivas o negativas de HPA.

Tiempo de muerte	d_{1j}	n_{1j}	d_{2j}	n_{2j}	d_j	n_j	e_{1j}	v_{1j}
5	0	13	1	32	1	45	0.2889	0.2054
8	0	13	1	31	1	44	0.2955	0.2082
10	0	13	1	30	1	43	0.3023	0.2109
13	0	13	1	29	1	42	0.3095	0.2137
18	0	13	1	28	1	41	0.3171	0.2165
23	1	13	0	27	1	40	0.3250	0.2194
24	0	12	1	27	1	39	0.3077	0.2130
26	0	12	2	26	2	38	0.6316	0.4205
31	0	12	1	24	1	36	0.3333	0.2222
35	0	12	1	23	1	35	0.3429	0.2253
40	0	12	1	22	1	34	0.3529	0.2284
41	0	12	1	21	1	33	0.3636	0.2314
47	1	12	0	20	1	32	0.3750	0.2314
48	0	11	1	20	1	31	0.3548	0.2289
50	0	11	1	19	1	30	0.3667	0.2322
59	0	11	1	18	1	29	0.3793	0.2354
61	0	11	1	17	1	28	0.3929	0.2385
68	0	11	1	16	1	27	0.4074	0.2414
69	1	11	0	15	1	26	0.4231	0.2441
71	0	9	1	15	1	24	0.3750	0.2344
113	0	6	1	10	1	16	0.3750	0.2344
118	0	6	1	8	1	14	0.4286	0.2449
143	0	6	1	7	1	13	0.4615	0.2485
148	1	6	0	6	1	12	0.5000	0.2500
181	1	5	0	4	1	9	0.5556	0.2469
Total	5						9.5652	5.9289

```

143, 154, 162, 188, 212, 217, 225),
censura = c(1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
            rep(1, 16), 1, 1, 0, 0, 0, 0, 1, 0, 1, 1,
            rep(0, 6)),
marca=c(rep(-1,13), rep(1, (16*2))))

```

```
library(survival)
```

```
#Obten los estimadores Kaplan-Meier para los dos grupos
```

```
hpa.survfit <- survfit(Surv(tiempo, censura) ~ marca, data= hpa)
```

```
#Dibuja la grafica de las dos curvas
```

```
plot(hpa.survfit, mark.time=FALSE, lty=1:2)
```

```
title(xlab="tiempo", ylab="Funcin de supervivencia estimada")
```

```
legend(25, 0.2, legend=c("Marca negativa", "Marca positiva"), lty=1:2)
```

```
#Realiza la prueba log-rank  
hpa.survdiff <- survdiff(Surv(tiempo, censura) ~ marca, data= hpa, rho=0)
```

Capítulo 3

Ajuste de modelos completamente paramétricos

3.1. El modelo exponencial

3.1.1. Muestras completas

Cuando se tiene una muestra aleatoria X_1, \dots, X_n sin censura de una población exponencial, $X_i \sim \text{EXP}(\theta)$, es posible usar el método de máxima verosimilitud para encontrar un estimador del parámetro desconocido θ . El procedimiento consiste en expresar la función de verosimilitud, la cual se define como la función de densidad conjunta de $\mathbf{X} = (X_1, \dots, X_n)$ en términos de θ , y después encontrar el valor de θ el cual maximiza la función de verosimilitud.

Para muestras completas, en general, la *función de verosimilitud* se puede definir como

$$L(\theta) = f_{\mathbf{X}}(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta),$$

donde $f_{X_i}(x_i)$ es la función de densidad marginal de X_i . Cuando $X_i \sim \text{EXP}(\theta)$ la

función de verosimilitud es

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{1}{\theta} \exp \left\{ -\frac{x_i}{\theta} \right\} \\ &= \theta^{-n} \exp \left\{ -\frac{1}{\theta} \sum_{i=1}^n x_i \right\}. \end{aligned}$$

Como el punto que maximiza el logaritmo natural de $L(\theta)$ es el mismo que maximiza $L(\theta)$, el *estimador de máxima verosimilitud* (EMV) se puede encontrar al obtener el punto óptimo de la función *log-verosimilitud* $l(\theta) = \log L(\theta)$ a través de la solución de la ecuación

$$\frac{dl(\theta)}{d\theta} = 0.$$

Para el caso exponencial se tiene que

$$l(\theta) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^n x_i,$$

y entonces

$$\frac{dl(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i.$$

Cuando se iguala a cero esta ecuación, se obtiene que el EMV es

$$\hat{\theta} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n},$$

el cual es un estimador insesgado de varianza mínima.

Cuando se trata de estimar la función de supervivencia, se puede usar la propiedad de invariancia de la técnica de máxima verosimilitud, y entonces el EMV de $S(t; \theta)$ es

$$\hat{S}(t) = \exp \left\{ -\frac{t}{\bar{x}} \right\};$$

sin embargo, este estimador no es insesgado. Es posible verificar que un estimador insesgado de varianza mínima es

$$\tilde{S}(t) = \begin{cases} \left[1 - \frac{t}{n\bar{x}} \right]^{n-1} & : n\bar{x} > t \\ 0 & : \text{de otra forma.} \end{cases}$$

Intervalos de confianza y pruebas de hipótesis sobre θ , o funciones monótonas de θ tales como la función de supervivencia $S(t)$, se pueden establecer basándose en la propiedad:

$$\frac{2n\bar{X}}{\theta} \sim \chi^2(2n).$$

Por ejemplo, un intervalo de confianza del coeficiente de confianza $(1 - \alpha)$ para θ es

$$\left(\frac{2n\bar{X}}{\chi_{1-\alpha/2}^2}, \frac{2n\bar{X}}{\chi_{\alpha/2}^2} \right),$$

donde $\chi_{1-\alpha/2}^2$ y $\chi_{\alpha/2}^2$ son percentiles de una distribución ji cuadrada con $2n$ grados de libertad.

Ejemplo 3.1. Considérense los siguientes tiempos de fallas en horas de vuelo de aparatos de aire acondicionado para aviones:

23, 261, 87, 7, 120, 14, 62, 47, 3, 95, 225, 71, 246, 21, 42,
20, 5, 12, 120, 11, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52.

El EMV de θ es $\hat{\theta} = \bar{x} = 59.6$. El EMV de la fuerza de mortalidad es $\hat{h}(t) = 1/\hat{\theta} = 0.017$, y el EMV de la función de supervivencia en el tiempo $t = 20$ es $\hat{S}(t) = \exp\{-20/59.6\} = 0.715$. Un intervalo de confianza de 95 % para θ es:

$$\left(\frac{2n\bar{x}}{\chi_{0.975}^2}, \frac{2n\bar{x}}{\chi_{0.025}^2} \right), \quad \text{que es} \quad \left(\frac{60(59.6)}{83.30}, \frac{60(59.6)}{40.48} \right);$$

por lo que los límites de confianza son 42.93 y 88.34. En este caso, la media de los datos se calculó con la función `mean` del lenguaje R, mientras que los percentiles se calcularon con la función `qchisq`.

3.1.2. Muestras con censura

Supóngase que los datos son n parejas de observaciones, donde el par correspondiente a la i -ésima observación, $i = 1, \dots, n$, es (t_i, δ_i) . En esta notación δ_i es una variable indicadora la cual toma el valor cero cuando el tiempo de supervivencia t_i está censurada y el valor uno cuando t_i es un tiempo de supervivencia sin censura. Una observación con muerte en t contribuye a la verosimilitud con $f(t)$, la densidad

evaluada en t . La contribución de una observación cuyo tiempo de supervivencia tiene censura en c es $S(c)$, la probabilidad de supervivencia después de c . La función de supervivencia completa de n observaciones independientes es entonces

$$L = \prod_{i=1}^n \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}. \quad (3.1)$$

Esta función puede ser maximizada con respecto a los parámetros desconocidos en las funciones de densidad y de probabilidad.

Supóngase que los tiempos de supervivencia siguen una distribución exponencial con media $1/\lambda$. Además, supóngase que hay r observaciones con tiempos de muertes y que los $n - r$ tiempos de supervivencia restantes están censuradas.

Para la distribución exponencial se tiene que

$$f(t) = \lambda e^{-\lambda t}, \quad S(t) = e^{-\lambda t},$$

y haciendo la sustitución en la Ecuación (3.1), la función de verosimilitud está dada por

$$L(\lambda) = \prod_{i=1}^n (\lambda e^{-\lambda t_i})^{\delta_i} (e^{-\lambda t_i})^{1-\delta_i}.$$

Al simplificar esta expresión se obtiene que

$$L(\lambda) = \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda t_i},$$

y entonces la función log-verosimilitud es

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

Como los datos contienen r muertes, entonces $\sum_{i=1}^n \delta_i = r$ y

$$l(\lambda) = r \log \lambda - \lambda \sum_{i=1}^n t_i.$$

Ahora es necesario identificar el valor $\hat{\lambda}$ que maximiza la función log-verosimilitud. La diferenciación con respecto a λ es

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

y al igualar la derivada a cero se obtiene que el EMV de λ es

$$\hat{\lambda} = r / \sum_{i=1}^n t_i.$$

La media de la distribución exponencial es $\mu = \lambda^{-1}$, y entonces el EMV de μ es

$$\hat{\mu} = \hat{\lambda}^{-1} = \frac{1}{r} \sum_{i=1}^n t_i.$$

Este estimador es el tiempo total sobrevivido por los n individuos en los datos dividido por el número de muertes observadas.

El error estándar de $\hat{\lambda}$ se puede obtener a partir de la segunda derivada de la función log-verosimilitud, usando los resultados asintóticos de la teoría de máxima verosimilitud. La segunda derivada de $l(\theta)$ es

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

y por lo tanto la varianza asintótica de $\hat{\lambda}$ es

$$\text{Var}[\hat{\lambda}] = \left(-E \left[\frac{d^2l(\lambda)}{d\lambda^2} \right] \right)^{-1} = \frac{\lambda^2}{r}.$$

En forma consecuente, el error estándar de $\hat{\lambda}$ está dado por

$$\text{s.e.}(\hat{\lambda}) = \frac{\hat{\lambda}}{\sqrt{r}}.$$

Este resultado puede usarse para obtener intervalos de confianza de la esperanza de supervivencia. En particular, los límites de un intervalo de confianza de $100(1-\alpha)\%$ para λ son $\hat{\lambda} \pm z_{1-\alpha/2} \text{s.e.}(\hat{\lambda})$, donde $z_{1-\alpha/2}$ es el percentil de una distribución normal estándar al punto $1 - \alpha/2$.

Al presentar los resultados de un análisis de supervivencia, los estimadores de cantidades tales como la función de supervivencia, la fuerza de mortalidad, la mediana y otros percentiles son útiles para su presentación. En particular, bajo el modelo exponencial, el estimador de la fuerza de mortalidad es $\hat{h}(t) = \hat{\lambda}$ y la función de supervivencia estimada es $\hat{S}(t) = \exp\{-\hat{\lambda}t\}$. Además, el tiempo mediano de supervivencia estimado es

$$t_{0,5} = \hat{t}(50) = \hat{\lambda}^{-1} \log 2,$$

y el p -ésimo percentil está dado por

$$\hat{t}(p) = \frac{1}{\hat{\lambda}} \log \left(\frac{100}{100-p} \right).$$

El error estándar del estimador del tiempo mediano de supervivencia se puede encontrar usando el resultado para aproximar la varianza de una función de una variable aleatoria. De acuerdo a este resultado, una aproximación de la varianza de $g(\hat{\lambda})$ es

$$\text{Var}[g(\hat{\lambda})] \approx \left[\frac{dg(\hat{\lambda})}{d\hat{\lambda}} \right]^2 \text{Var}[\hat{\lambda}]. \quad (3.2)$$

Usando este resultado, la varianza aproximada del p -ésimo percentil estimado está dado por

$$\text{Var}[\hat{t}(p)] \approx \left[-\frac{1}{\hat{\lambda}^2} \log \left(\frac{100}{100-p} \right) \right]^2 \text{Var}[\hat{\lambda}].$$

Simplificando esta expresión y tomando la raíz cuadrada, se obtiene que

$$\begin{aligned} \text{s.e.}[\hat{t}(p)] &= \frac{1}{\hat{\lambda}^2} \log \left(\frac{100}{100-p} \right) \text{s.e.}[\hat{\lambda}] \\ &= \hat{t}(p) / \sqrt{r}. \end{aligned}$$

En particular, el error estándar del tiempo mediano de supervivencia es

$$\text{s.e.}[\hat{t}(50)] = \hat{t}(50) / \sqrt{r}.$$

Los intervalos de confianza de un percentil son obtenidos al aplicar la función exponencial de los límites de confianza del percentil. Este procedimiento garantiza que los límites de confianza del percentil sean positivos. De nueva cuenta, usando el resultado en la Ecuación (3.2), el error estándar de $\log \hat{t}(p)$ está dado por

$$\begin{aligned} \text{s.e.}[\log \hat{t}(p)] &= \hat{t}(p)^{-1} \text{s.e.}[\hat{t}(p)] \\ &= 1/\sqrt{r}. \end{aligned}$$

Usando este resultado, los límites de confianza de $100(1 - \alpha)\%$ del p -ésimo percentil son $\hat{t}(p) \exp\{\pm z_{1-\alpha/2}/\sqrt{r}\}$.

Ejemplo 3.2. Considérense las duraciones de discontinuación del DIU en la Tabla 2.3. Se ajusta entonces el modelo que supone una fuerza de mortalidad constante. Para estos datos, el total de las duraciones con y sin censura es de 1046 días, y el número de de observaciones sin censura es de 9. De esta forma, se tiene que $\hat{\lambda} = 9/1046 = 0.0086$ y el error estándar de $\hat{\lambda}$ es $\text{s.e.}[\hat{\lambda}] = 0.0086/\sqrt{9} = 0.0029$. La función de supervivencia estimada es $\hat{h}(t) = 0.0086$, la función de supervivencia estimada es $\hat{S}(t) = \exp\{-0.0086t\}$, y la mediana del tiempo de discontinuidad es de 81 días.

Un estimador del 90-vo percentil de la distribución de los tiempos de discontinuación es $\hat{t}(90) = \log(10)/0.0086 = 267.61$. Esto significa que bajo la suposición de que el riesgo de discontinuidad del DIU es independiente del tiempo, 90% de las mujeres tienen tiempo de discontinuidad de menos de 268 días.

El error estándar del tiempo de discontinuidad mediano estimado es de $80.56/\sqrt{9} = 26.85$ días. Los límites de un intervalo de confianza de 95% para el tiempo de discontinuación mediano son

$$80.56 \exp\{\pm 1.96/\sqrt{9}\},$$

y entonces el intervalo va de 42 días a 155 días.

3.2. El modelo Weibull

3.2.1. Muestras completas

Si se tiene una muestra aleatoria X_1, \dots, X_n sin censura de una población Weibull, $X_i \sim \text{WEI}(\theta, \beta)$, la función de verosimilitud se puede expresar como:

$$L(\theta, \beta) = \prod_{i=1}^n \frac{\beta}{\theta^\beta} x_i^{\beta-1} \exp \left\{ - \left(\frac{x_i}{\theta} \right)^\beta \right\},$$

y la función log-verosimilitud es:

$$l(\theta, \beta) = n \log \beta - \beta n \log \theta + (\beta - 1) \sum_{i=1}^n \log x_i - \sum_{i=1}^n \left(\frac{x_i}{\theta} \right)^\beta.$$

En este caso, los EMV de θ y β se obtienen al resolver el siguiente sistema de ecuaciones:

$$\begin{cases} \frac{\partial l(\theta, \beta)}{\partial \theta} = 0 \\ \frac{\partial l(\theta, \beta)}{\partial \beta} = 0 \end{cases},$$

donde

$$\frac{\partial l(\theta, \beta)}{\partial \theta} = -\frac{n\beta}{\theta} + \frac{\beta}{\theta^{\beta+1}} \sum_{i=1}^n x_i^\beta$$

y

$$\frac{\partial l(\theta, \beta)}{\partial \beta} = \frac{n}{\beta} - n \log \theta + \sum_{i=1}^n \log x_i - \sum_{i=1}^n \left(\frac{x_i}{\theta} \right)^\beta \log \left(\frac{x_i}{\theta} \right),$$

Este sistema de ecuaciones se reduce a

$$\frac{\sum_{i=1}^n x_i^{\hat{\beta}} \log x_i}{\sum_{i=1}^n x_i^{\hat{\beta}}} - \frac{1}{\hat{\beta}} = \frac{\sum_{i=1}^n \log x_i}{n}$$

y

$$\hat{\theta} = \left(\frac{\sum_{i=1}^n x_i^{\hat{\beta}}}{n} \right)^{1/\hat{\beta}},$$

donde $\hat{\theta}$ y $\hat{\beta}$ son los EMV.

El sistema de ecuaciones no se puede resolver en forma cerrada. Como en la solución es única, es posible usar un método numérico para aproximar a $\hat{\theta}$ y $\hat{\beta}$. El procedimiento Newton-Raphson, por ejemplo, puede usarse para resolver una ecuación $g(\hat{\beta}) = 0$ usando aproximaciones sucesivas $\hat{\beta}_j$, donde $\hat{\beta}_{j+1} = \hat{\beta}_j - g(\hat{\beta}_j)/g'(\hat{\beta}_j)$. Muchas otras técnicas pueden ser implementadas usando una computadora.

Para obtener la matriz información es necesario encontrar el Hessiano de $l(\theta, \beta)$, para lo cual

$$\frac{\partial^2 l(\theta, \beta)}{\partial \theta^2} = \frac{n\beta}{\theta^2} - \frac{\beta(\beta+1)}{\theta^{\beta+2}} \sum_{i=1}^n X_i^\beta,$$

$$\frac{\partial^2 l(\theta, \beta)}{\partial \beta^2} = -\frac{n}{\beta^2} - \sum_{i=1}^n \left(\frac{X_i}{\theta}\right)^\beta \log\left(\frac{X_i}{\theta}\right)^2,$$

y

$$\frac{\partial^2 l(\theta, \beta)}{\partial \theta \beta} = -\frac{n}{\theta} + \frac{1}{\theta} \sum_{i=1}^n \left(\frac{X_i}{\theta}\right)^\beta + \frac{\beta}{\theta} \sum_{i=1}^n \left(\frac{X_i}{\theta}\right)^\beta \log\left(\frac{X_i}{\theta}\right).$$

La matriz información para θ y β es entonces

$$\mathbf{I}(\theta, \beta) = \mathbf{E} \begin{bmatrix} -\frac{\partial^2 l(\theta, \beta)}{\partial \theta^2} & -\frac{\partial^2 l(\theta, \beta)}{\partial \theta \beta} \\ -\frac{\partial^2 l(\theta, \beta)}{\partial \theta \beta} & -\frac{\partial^2 l(\theta, \beta)}{\partial \beta^2} \end{bmatrix}$$

Como esta matriz está en términos de los parámetros desconocidos θ y β , es posible estimarla usando los EMV con

$$\tilde{\mathbf{I}}(\hat{\theta}, \hat{\beta}) = \begin{bmatrix} -\frac{\partial^2 l}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} & -\frac{\partial^2 l}{\partial \theta \beta} \Big|_{\theta=\hat{\theta}, \beta=\hat{\beta}} \\ -\frac{\partial^2 l}{\partial \theta \beta} \Big|_{\theta=\hat{\theta}, \beta=\hat{\beta}} & -\frac{\partial^2 l}{\partial \beta^2} \Big|_{\beta=\hat{\beta}} \end{bmatrix}.$$

De esta forma, la inversa de esta matriz es un estimador de la matriz de covarianzas

de $(\hat{\theta}, \hat{\beta})^T$:

$$\tilde{\mathbf{I}}^{-1}(\hat{\theta}, \hat{\beta}) = \begin{bmatrix} \widetilde{\text{Var}}[\hat{\theta}] & \widetilde{\text{Cov}}[\hat{\theta}, \hat{\beta}] \\ \widetilde{\text{Cov}}[\hat{\theta}, \hat{\beta}] & \widetilde{\text{Var}}[\hat{\beta}] \end{bmatrix}.$$

Usando los resultados asintóticos de máxima verosimilitud se tiene la siguiente aproximación

$$(\hat{\theta}, \hat{\beta})^T \sim \text{NBV} \left((\theta, \beta)^T, \tilde{\mathbf{I}}^{-1}(\hat{\theta}, \hat{\beta}) \right),$$

donde NBV denota la distribución normal bivariada. por lo que

$$\hat{\theta} \sim \text{N}(\theta, \widetilde{\text{Var}}[\hat{\theta}]) \quad \text{aproximadamente,}$$

y

$$\hat{\beta} \sim \text{N}(\beta, \widetilde{\text{Var}}[\hat{\beta}]) \quad \text{aproximadamente.}$$

3.2.2. Muestras con censura

Supóngase que se tiene una muestra de tamaño n tomada de una distribución Weibull con parámetro de escala $\lambda^{1/\gamma}$ y parámetro de forma γ , $\text{WEI}(1/[\lambda^{1/\gamma}], \gamma)$. Supóngase también que hay r muertes de los n individuos y $n - r$ tiempos de supervivencia con censura. La función de densidad y la función de supervivencia están dadas por

$$f(t) = \lambda \gamma t^{\gamma-1} \exp(-\lambda t^\gamma), \quad S(t) = \exp(-\lambda t^\gamma).$$

Esta es una reparametrización de la Weibull; sin embargo, los resultados en esta sección son equivalentes.

La función de verosimilitud se puede expresar como:

$$\begin{aligned} L(\lambda, \gamma) &= \prod_{i=1}^n \{f(t_i; \lambda, \gamma)\}^{\delta_i} \{S(t_i; \lambda, \gamma)\}^{1-\delta_i} \\ &= \prod_{i=1}^n \{\lambda \gamma t_i^{\gamma-1} \exp(-\lambda t_i^\gamma)\}^{\delta_i} \{\exp(-\lambda t_i^\gamma)\}^{1-\delta_i}. \end{aligned}$$

La función log-verosimilitud correspondiente está dada por

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

y como $\sum_{i=1}^n \delta_i = r$, la función verosimilitud es

$$l(\lambda, \gamma) = r \log(\lambda\gamma) + (\gamma - 1) \sum_{i=1}^n \delta_i \log t_i - \lambda \sum_{i=1}^n t_i^\gamma.$$

Los estimadores de máxima verosimilitud de λ y γ se encuentran al diferenciar $l(\lambda, \gamma)$ con respecto a λ y γ a la vez, al igualar a cero las dos derivadas, y al encontrar la solución del sistema de ecuaciones $\hat{\lambda}$ y $\hat{\gamma}$. Las ecuaciones del sistema resultante son

$$\frac{r}{\hat{\lambda}} - \sum_{i=1}^n t_i^{\hat{\gamma}} = 0,$$

y

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^n \delta_i \log t_i - \hat{\lambda} \sum_{i=1}^n t_i^{\hat{\gamma}} \log t_i = 0.$$

Por lo tanto

$$\hat{\lambda} = r / \sum_{i=1}^n t_i^{\hat{\gamma}},$$

y al sustituir $\hat{\lambda}$ en la primera ecuación del sistema se tiene que

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^n \delta_i \log t_i - \frac{r}{\sum_{i=1}^n t_i^{\hat{\gamma}}} \sum_{i=1}^n t_i^{\hat{\gamma}} \log t_i = 0.$$

Esta es una ecuación no lineal la cual solo puede resolverse usando un procedimiento iterativo tal como Newton-Raphson.

Una vez que se encuentran los estimadores de λ y γ , se pueden estimar percentiles del tiempo de supervivencia. El p -ésimo percentil de la distribución es

$$\hat{t}(p) = \left[\frac{1}{\hat{\lambda}} \log \left(\frac{100}{100 - p} \right) \right]^{1/\hat{\gamma}},$$

y entonces el estimador de la mediana del tiempo de supervivencia está dado por

$$\hat{t}(50) = \left[\frac{1}{\hat{\lambda}} \log 2 \right]^{1/\hat{\gamma}}.$$

El error estándar del p -ésimo percentil estimado se puede obtener mas fácilmente al encontrar primero la varianza de $\log \hat{t}(p)$ y al usar el siguiente resultado:

$$\text{Var}[g(\hat{\theta}_1, \hat{\theta}_2)] \approx \left(\frac{\partial g}{\partial \hat{\theta}_1} \right)^2 \text{Var}[\hat{\theta}_1] + \left(\frac{\partial g}{\partial \hat{\theta}_2} \right)^2 \text{Var}[\hat{\theta}_2] + 2 \left(\frac{\partial g}{\partial \hat{\theta}_1} \frac{\partial g}{\partial \hat{\theta}_2} \right) \text{Cov}[\hat{\theta}_1, \hat{\theta}_2]. \quad (3.3)$$

Aquí se tiene que,

$$\log \hat{t}(p) = \frac{1}{\hat{\gamma}} \log \left\{ \hat{\gamma}^{-1} \log \left(\frac{100}{100-p} \right) \right\},$$

y entonces

$$\log \hat{t}(p) = \frac{1}{\hat{\gamma}} \left\{ c_p - \log \hat{\lambda} \right\},$$

donde

$$c_p = \log \log \left(\frac{100}{100-p} \right).$$

Usando el resultado de la Ecuación (3.3), se tiene que

$$\begin{aligned} \text{Var}[\log \hat{t}(p)] &\approx \left(\frac{\partial \log \hat{t}(p)}{\partial \hat{\lambda}} \right)^2 \text{Var}[\hat{\lambda}] + \left(\frac{\partial \log \hat{t}(p)}{\partial \hat{\gamma}} \right)^2 \text{Var}[\hat{\gamma}] \\ &\quad + 2 \frac{\partial \log \hat{t}(p)}{\partial \hat{\lambda}} \frac{\partial \log \hat{t}(p)}{\partial \hat{\gamma}} \text{Cov}[\hat{\lambda}, \hat{\gamma}]. \end{aligned}$$

La derivadas de $\log \hat{t}(p)$ con respecto a λ y γ son

$$\begin{aligned} \frac{\partial \log \hat{t}(p)}{\partial \hat{\lambda}} &= -\frac{1}{\hat{\lambda} \hat{\gamma}}, \\ \frac{\partial \log \hat{t}(p)}{\partial \hat{\gamma}} &= -\frac{c_p - \log \hat{\lambda}}{\hat{\gamma}^2}, \end{aligned}$$

y entonces

$$\text{Var}[\log \hat{t}(p)] \approx \frac{1}{\hat{\lambda}^2 \hat{\gamma}^2} \text{Var}[\hat{\lambda}] + \frac{(c_p - \log \hat{\lambda})^2}{\hat{\gamma}^4} \text{Var}[\hat{\gamma}] + \frac{2(c_p - \log \hat{\lambda})}{\hat{\lambda} \hat{\gamma}^3} \text{Cov}[\hat{\lambda}, \hat{\gamma}].$$

La varianza de $\hat{t}(p)$ se encuentra con la aproximación

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

por lo que

$$\text{Var}[\hat{t}(p)] \approx [\hat{t}(p)]^2 \text{Var}[\log \hat{t}(p)].$$

y entonces

$$\begin{aligned} \text{s.e.}[\hat{t}(p)] &= \frac{\hat{t}(p)}{\hat{\lambda}\hat{\gamma}^2} \left\{ \hat{\gamma}^2 \text{Var}[\hat{\lambda}] + \hat{\lambda}^2 (c_p - \log \hat{\lambda})^2 \text{Var}[\hat{\gamma}] \right. \\ &\quad \left. + 2\hat{\lambda}\hat{\gamma} (c_p - \log \hat{\lambda}) \text{Cov}[\hat{\lambda}, \hat{\gamma}] \right\}^{1/2}. \end{aligned}$$

Un intervalo de confianza de $100(1 - \alpha)\%$ para el p -ésimo percentil se puede encontrar al manipular con operaciones válidas los siguientes límites de confianza para $\log t(p)$

$$\log \hat{t}(p) \pm z_{1-\alpha/2} \text{s.e.}[\log \hat{t}(p)],$$

por lo cual los límites son

$$\hat{t}(p) \exp\{\pm z_{1-\alpha/2} \text{s.e.}[\log \hat{t}(p)]\},$$

Ejemplo 3.3. En el lenguaje de programación R, la función `survreg` ajusta al modelo Weibull $\text{WEI}(\exp \beta_0, 1/\alpha)$, de manera tal que la función de supervivencia es

$$S(t) = \exp \left\{ - \left(\frac{t}{\exp \beta_0} \right)^{1/\alpha} \right\}.$$

A continuación se ilustra el ajuste de los datos DIU presentado en la Tabla 2.3

```
diu <- data.frame(tiempo=c(10, 13, 18, 19, 23, 30, 36, 38, 54,
                          56, 59, 75, 93, 97, 104, 107, 107, 107),
                 estatus=c(1, 0, 0, 1, 0, 1, 1, 0, 0,
                          0, 1, 1, 1, 1, 0, 1, 0, 0))
```

```
library(survival)
```

```
# Encuentra el Modelo Weibull para los datos en diu
diu.wei <- survreg(Surv(tiempo, estatus)~1, data=diu,
                  dist='weibull')
```

En este caso, el signo `~1` en la fórmula indica a R que la población es homogénea. Algunos de los atributos del objeto `diu.wei` se pueden ver a continuación:

```

> #los EMV de BETA0 y log(alfa) son:
> diu.wei$icoef
(Intercept)  Log(scale)
  4.5915178  -0.5166504
> #la matriz de covarianzas correspondiente es:
> diu.wei$var
                (Intercept) Log(scale)
(Intercept)  0.04176813  0.01297235
Log(scale)   0.01297235  0.07541491

```

De esta forma, $\hat{\beta}_0 = 4.5915178$, $\text{s.e.}[\hat{\beta}_0] = \sqrt{0,04176813} = 0,2043725$, $\log \hat{\alpha} = -0,5166504$ y $\text{s.e.}[\log \hat{\alpha}] = \sqrt{0,07541491} = 0,2746178$. Un intervalo de 95% de confianza para $\log \alpha$ tiene límites $\log \hat{\alpha} \pm 1,96\text{s.e.}[\log \hat{\alpha}]$, por lo que el intervalo es $(-1,054901, 0,02160041)$. Como el intervalo incluye a cero, es pausable que el modelo exponencial se ajuste tan bien como el Weibull.

Capítulo 4

El uso de información concomitante

4.1. Datos de supervivencia con variables explicativas

En muchas situaciones prácticas, los datos de supervivencia vienen con información asociada o *concomitante* de la cual se cree que dependen los tiempos de supervivencia. En datos de medicina, por ejemplo, esta información puede ser por diseño, en el sentido de que uno o mas tratamientos se comparan con el efecto de tiempos de supervivencia de un grupo de “control”; o puede ocurrir al clasificar pacientes de alguna forma natural, como es género, grupo de edad, etc.

En la mayoría de las situaciones en que se observan tiempos de supervivencia habrá alguna, o a veces mucha, información auxiliar. Esta información puede venir en la forma de una observación de una variable continua, o bien en la forma de una variable categórica. Para entender cómo incluir esta información en los modelos de supervivencia, a continuación se da un breve resumen de su uso en el contexto de los modelos lineales generalizados.

4.2. El Modelo lineal generalizado

Supóngase que se tiene una variable respuesta Y , que es la variable aleatoria de interés, y varias variables explicativas. El conocimiento del contexto en el cual se obtuvieron los datos - tales como relaciones teóricas entre las variables, el diseño del estudio y los resultados de un análisis exploratorio de los datos - pueden hacerse para formular el modelo. El modelo lineal tiene dos componentes:

1. La función de distribución de Y , e.g. $Y \sim N(\mu, \sigma^2)$, $Y \sim \text{EXP}(\theta)$.
2. Una ecuación que *liga* el valor de Y con una combinación lineal de las variables explicativas, e.g. $E[Y] = \beta_0 + \beta_1 x$ o $\log(E(Y)) = \beta_0 + \beta_1 \text{sen}(\alpha x)$.

Los modelos lineales generalizados son aquellos cuyas funciones de densidad pertenecen a la familia exponencial de distribuciones, las cuales incluyen la normal, binomial, Poisson etc. La ecuación en el segundo componente tiene la forma general:

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

donde la parte $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ se llama el componente lineal.

Para respuestas Y_1, \dots, Y_n , i.e. para una muestra aleatoria, esta forma general se puede escribir en notación de matrices como

$$g(E[Y]) = \mathbf{X}\boldsymbol{\beta}$$

donde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

es el vector de respuestas,

$$g(\mathbf{E}[\mathbf{Y}]) = \begin{pmatrix} g[\mathbf{E}(Y_1)] \\ \vdots \\ g[\mathbf{E}(Y_n)] \end{pmatrix}$$

es el vector de funciones de los términos $E[Y_i]$ (con la misma g para cada elemento),

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

es el vector de parámetros, y \mathbf{X} es la matriz cuyos elementos son constantes que representan niveles de variables explicativas categóricas o variables explicativas de medidas continuas.

Para una variable explicativa continua x (tal como edad) el modelo contiene un término $\beta_1 x$ donde el parámetro β representa el cambio de la respuesta correspondiente al cambio de una unidad en x .

Para una variable explicativa categórica hay parámetros para los diferentes niveles de un *factor*. Los elementos correspondientes en \mathbf{X} se eligen para excluir o incluir los parámetros apropiados de cada observación. Estas variables son llamadas *variables ficticias* o *variables indicadoras* (en inglés *dummy variable*); si se componen de ceros y unos, se usa el término *variable indicador*.

Si hay $(p + 1)$ parámetros en el modelo y N observaciones, entonces Y es un vector aleatorio de $n + 1$, $\boldsymbol{\beta}$ es un vector de $p + 1$ parámetros y \mathbf{X} es una matriz de constantes conocidas. A \mathbf{X} se le llama la *matriz diseño* y $\mathbf{X}\boldsymbol{\beta}$ es conocido como el componente lineal.

Ejemplo 4.1. Considérese el modelo de regresión lineal simple para dos grupos (e.g. por sexo) con el mismo número de observaciones en cada grupo

$$g(\mathbf{E}[Y_{jk}]) = g(\mu_{jk}) = \alpha_j + \beta_j X_{jk},$$

donde $j = 1, 2$ y $k = 1, \dots, K$. Aquí el vector de respuestas y el vector de coeficientes toman la forma

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K} \\ Y_{21} \\ \vdots \\ Y_{2K} \end{pmatrix}, \quad \text{y} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

y la matriz diseño correspondiente es

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & X_{11} & 0 \\ 1 & 0 & X_{12} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & X_{1K} & 0 \\ 0 & 1 & 0 & X_{21} \\ 0 & 1 & 0 & X_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & X_{2K} \end{pmatrix}.$$

Ejemplo 4.2. Considérense formulaciones alternativas para comparar los efectos de dos grupos. La muestra aleatoria tiene la forma $Y_{11}, \dots, Y_{1K_1}, Y_{21}, \dots, Y_{2K_2}$.

a) Si $g(E[Y_{1j}]) = \beta_1$ y $g(E[Y_{2k}]) = \beta_2$, entonces

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2K_2} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \text{y} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

b) Si $g(E[Y_{1j}]) = \mu + \alpha_1$ y $g(E[Y_{2k}]) = \mu + \alpha_2$, μ representa la media general y α_1

y α_2 son las diferencias a partir de μ . De esta forma:

$$\beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} \quad y \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}.$$

Esta formulación, sin embargo, no se recomienda. El problema es que se tienen muchos parámetros dentro del modelo. Es, entonces, necesario hacer una modificación.

- c) Si $g(E[Y_{1K}]) = \mu$ y $g(E[Y_{2K}]) = \mu + \alpha$, el grupo 1 se trata como el grupo de referencia y α representa el efecto adicional del grupo 2. Entonces

$$\beta = \begin{pmatrix} \mu \\ \alpha \end{pmatrix} \quad y \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$$

- d) Si $g(E[Y_{1K}]) = \mu + \alpha$ y $g(E[Y_{2K}]) = \mu - \alpha$, entonces los grupos se tratan simétricamente; μ es el efecto promedio general y α representa las diferencias del grupo. Aquí,

$$\beta = \begin{pmatrix} \mu \\ \alpha \end{pmatrix} \quad y \quad \mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}$$

Ejemplo 4.3. Variables explicativas ordinales. Supóngase que los datos se obtienen para tres grupos de pacientes con una enfermedad poco ayuda, moderada o severa. Los grupos pueden describirse por niveles de una variable ordinal. Esta se

puede describir como:

$$\begin{aligned}
 g(\mathbb{E}[Y_{1j}]) &= \mu, \\
 g(\mathbb{E}[Y_{2k}]) &= \mu + \alpha, \\
 g(\mathbb{E}[Y_{3l}]) &= \mu + \alpha_1 + \alpha_2,
 \end{aligned}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{y} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \end{pmatrix}$$

Así, α_1 representa el efecto del grupo 2 relativo al grupo 1 y α_2 representa el efecto del grupo 3 relativo al grupo 2.

Inclusión de un factor

Un factor es, entonces, una variable explicativa asociada con una variable categórica con J niveles o formas de agrupar las observaciones. Un método conveniente para incluir un factor en un modelo es considerar los efectos principales del factor $\alpha_1, \alpha_2, \dots, \alpha_J$, de manera tal que

$$g(\mathbb{E}[Y_{jk}]) = g(\mu_j) = \mu + \alpha_j, \quad j = 1, \dots, J,$$

con la restricción $\alpha_1 = 0$.

Bajo esta estructura, los términos $\mu, \alpha_2, \dots, \alpha_J$ pueden modelarse al definir J variables indicadoras $X_0, X_2, X_3, \dots, X_J$ para obtener el modelo lineal

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}.$$

La matriz diseño puede entonces construirse de acuerdo a la siguiente descripción dependiendo del nivel a que pertenezca cada observación:

Nivel de A	X_0	X_2	X_3	\cdots	X_J
1	1	0	0		0
2	1	1	0	\cdots	0
3	1	0	1	\cdots	\vdots
\vdots	\vdots			\ddots	\vdots
J	1	0	\cdots		1

De esta forma, el modelo queda como

$$g(\mathbb{E}[Y]) = \mu x_0 + \alpha_2 x_2 + \cdots + \alpha_J x_J$$

donde x_j es el valor de X_j para un individuo que en particular. Cuando se tiene la restricción $\alpha_1 = 0$, se tienen contrastes de tratamientos, en el lenguaje S-PLUS, es posible invocar este formato con el comando:

```
options(contrasts = c("contr.treatment", "contr.poly"))
```

El lenguaje R tiene esta estructura siempre.

Inclusión de dos factores

Considérense el factor A con J niveles y el factor B con K niveles. Es posible clasificar los datos en forma *cruzada* con los JK subgrupos que forman todas las combinaciones de los niveles de A y B.

Aquí las respuestas son del estilo Y_{jkl} , la l -ésima observación del k -ésimo nivel del factor B del j -ésimo nivel del factor A. Las siguientes estructuras forman todas las posibilidades de incluir los factores en el modelo:

i) El *modelo nulo*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu,$$

que es el modelo más simple.

ii) El *modelo A*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu_j, \quad j = 1, \dots, J,$$

o en forma equivalente

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \alpha_j,$$

con alguna de las siguientes restricciones

a) $\alpha_1 = 0$, ó

b) $\sum \alpha_j = 0$.

iii) El *modelo B*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \beta_k, \quad k = 1, \dots, K,$$

con alguna de las siguientes restricciones

a) $\beta_1 = 0$ ó

b) $\sum \beta_k = 0$

iv) El *modelo A + B*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \alpha_j + \beta_k \quad j = 1, \dots, J, \quad k = 1, \dots, K,$$

donde la restricción es

i) $\alpha_1 = \beta_1 = 0$ ó

ii) $\sum \alpha_j = \sum \beta_k = 0$.

v) El *modelo saturado A * B con interacción*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \quad j = 1, \dots, J, \quad k = 1, \dots, K;$$

Tabla 4.1: Modelos anidados para los factores A y B.

Fórmula	Modelo	número de parámetros en β
$A * B$	$\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$	JK
$A + B$	$\mu + \alpha_j + \beta_k$	$JK - (J - 1)(K - 1)$
A	$\mu + \alpha_j$	J
B	$\mu + \beta_k$	K
nulo	μ	1

aquí, $(\alpha\beta)_{jk}$ corresponde a los efectos de interacción, y α_j y β_k corresponden a los efectos principales de A y B respectivamente.

La Tabla 4.1 muestra un resumen de los modelos generados con los factores A y B. Es posible observar que los modelos están relacionados con ciertos parámetros, lo cual es útil para llevar a cabo pruebas de significancia.

Las interacciones son términos en el modelo que corresponden a efectos individuales para cada combinación de niveles de los factores. Para incluir el término $(\alpha\beta)_{jk}$ en el modelo, se calculan los productos de las variables indicadoras incluidos en los efectos principales. De esta forma, existen $(J - 1)(K - 1)$ parámetros asociados con la interacción. Si, por ejemplo, $J = 2$ y $K = 2$, entonces el vector de coeficientes del modelo saturado es

$$\beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ (\alpha\beta)_{22} \end{pmatrix}.$$

Inclusión de un factor y una variable continua

Hay situaciones en las que se puede estar interesado en comparar medias de subgrupos definidos por los niveles de un factor pero reconociendo que variables

contínuas pueden afectar las respuestas. En este caso, $g(E[Y_{jk}])$ se puede modelar como una línea recta de una variable continua x_{jk} con la posibilidad de asignar pendientes y ordenadas al origen a cada subgrupo del factor A.

A continuación se describen los modelos que se pueden formar con una factor A y una variable continua c.

i) El modelo nulo:

$$g(E[Y_{jk}]) = \mu,$$

el modelo más simple.

ii) El modelo A:

$$E[Y_{jk}] = \mu + \alpha_j \quad j = 1, \dots, J,$$

con la restricción $\alpha_1 = 0$.

iii) El modelo c:

$$g(E[Y_{jk}]) = \mu + \gamma x_{jk},$$

el ajuste de una recta con la variable explicativa x_{jk} para $g(E[Y_{jk}])$.

iv) El modelo A+c:

$$E[Y_{jk}] = \mu + \alpha_j + \gamma X_{jk},$$

donde $\alpha_1 = 0$, el modelo con diferentes ordenadas al origen y pendiente común para todos los niveles del factor.

v) El modelo A*c:

$$E[Y_{jk}] = \mu + \alpha_j + \gamma X_{jk} + (\alpha\gamma)_j X_{jk},$$

donde $\alpha_1 = 0$, el modelo que considera diferentes pendientes y ordenadas al origen para cada subgrupo en A. Aquí, $(\alpha\gamma)_j$ representa el coeficiente correspondiente a la interacción entre A y c.

En general, es posible incluir los factores y variables continuas que se deseen, siempre y cuando el número de coeficientes en β no supere el tamaño de la muestra. La interpretación de los modelos son generalizaciones de los resultados aquí mostrados. Cuando se trate de pronosticar $g(E[Y])$ para un individuo en particular, sólo se necesita sustituir los valores correspondientes en cada una de las variables explicativas del modelo estimado.

4.3. Inferencia para Modelos de regresión paramétrica

4.3.1. Las distribuciones Exponencial y Weibull

Para modelar una distribución exponencial, la dependencia del tiempo de supervivencia T , $T \sim \text{EXP}(\theta)$, de las variables explicativas se puede modelar con

$$\log(E[T]) = \log(\theta) = \mathbf{x}^T \boldsymbol{\beta},$$

donde $\mathbf{x}^T = (1, x_1, \dots, x_p)$ es el vector de variables explicativas y $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$ es el vector de coeficientes correspondiente. En este caso, la función liga g es el logaritmo natural. Esta liga es conveniente pues asegura que el parámetro θ permanezca siempre positivo. Nótese que β_0 es el coeficiente correspondiente a la ordenada al origen.

La distribución Weibull de dos parámetros, a diferencia de la Exponencial, no pertenece a la familia de distribuciones exponencial; por lo cual, no es posible aplicar una liga directamente a la media de los tiempos de supervivencia como se ha descrito en la sección anterior. Sin embargo, es posible tomar algunas ideas de los modelos lineales generalizados y entonces incluir un componente lineal en alguno de los parámetros asociados con la distribución.

Se ha visto que una de las parametrizaciones de la función de supervivencia del

modelo Weibull es:

$$S_T(t) = \exp(-\lambda t^\alpha),$$

y la fuerza de mortalidad correspondiente se expresa como

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

Si se toma la transformación $Y = \log T$, la función de supervivencia correspondiente de Y es

$$S_Y(y) = \exp(-\lambda e^{\alpha y}).$$

Si se redefinen los parámetros con $\lambda = \exp\{-\mu/\sigma\}$ y $\sigma = 1/\alpha$, entonces Y toma la forma de un modelo log lineal con

$$Y = \log T = \mu + \sigma W,$$

donde W es la distribución de valor extremo con función de densidad dada por

$$f_W(w) = \exp\{w - e^w\}$$

y función de supervivencia,

$$S_W(w) = \exp(-e^w).$$

De esta forma, las funciones de densidad y de supervivencia de Y están dadas por

$$f_Y(y) = (1/\sigma) \exp[(y - \mu)/\sigma - e^{(y-\mu)/\sigma}]$$

y

$$S_Y(y) = \exp(-e^{(y-\mu)/\sigma})$$

respectivamente. Bajo esta estructura, se modelan tiempos de supervivencia T , donde $T \sim \text{WEI}(\exp\{\mathbf{x}^T \boldsymbol{\beta}\}, 1/\sigma)$. Cuando $\alpha = 1$, o en forma equivalente, cuando $\sigma = 1$, entonces la distribución Weibull se reduce a la distribución exponencial.

La función de verosimilitud para datos con censura por la derecha está dada por

$$\begin{aligned} L &= \prod_{j=1}^n [f_Y(y_j)]^{\delta_j} [S_Y(y_j)]^{(1-\delta_j)} \\ &= \prod_{j=1}^n \left[f_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{\delta_j} \left[S_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{(1-\delta_j)}. \end{aligned}$$

Una vez que se han encontrado los estimadores de máxima verosimilitud para los parámetros μ y σ , entonces los de λ y α , se pueden obtener estimadores para las funciones de supervivencia y de fuerza de mortalidad para T o Y .

Los estimadores de μ y α se encuentran numéricamente con rutinas ya disponibles. La matriz de covarianzas de los parámetros μ y α están disponibles en los paquetes estadísticos también. Usando la propiedad de invarianza de los estimadores de máxima verosimilitud, se tiene que los estimadores de máxima verosimilitud de λ y α son

$$\hat{\lambda} = \exp(-\hat{\mu}/\hat{\sigma}) \quad \text{y} \quad \hat{\alpha} = 1/\hat{\sigma}.$$

Para obtener las varianzas y covarianzas de $\hat{\lambda}$ y $\hat{\alpha}$ es posible usar las aproximaciones de series de Taylor; estas son:

$$\text{Var}[g(\hat{\theta}_1, \hat{\theta}_2)] \approx \left(\frac{\partial g}{\partial \hat{\theta}_1} \right)^2 \text{Var}[\hat{\theta}_1] + \left(\frac{\partial g}{\partial \hat{\theta}_2} \right)^2 \text{Var}[\hat{\theta}_2] + 2 \left(\frac{\partial g}{\partial \hat{\theta}_1} \frac{\partial g}{\partial \hat{\theta}_2} \right) \text{Cov}[\hat{\theta}_1, \hat{\theta}_2].$$

y

$$\begin{aligned} \text{Cov}[g_1(\hat{\theta}_1, \hat{\theta}_2), g_2(\hat{\theta}_1, \hat{\theta}_2)] &\approx \left(\frac{\partial g_1}{\partial \hat{\theta}_1} \frac{\partial g_2}{\partial \hat{\theta}_1} \right) \text{Var}[\hat{\theta}_1] + \left(\frac{\partial g_1}{\partial \hat{\theta}_2} \frac{\partial g_2}{\partial \hat{\theta}_2} \right) \text{Var}[\hat{\theta}_2] + \\ &\quad \left(\frac{\partial g_1}{\partial \hat{\theta}_1} \frac{\partial g_2}{\partial \hat{\theta}_2} + \frac{\partial g_1}{\partial \hat{\theta}_2} \frac{\partial g_2}{\partial \hat{\theta}_1} \right) \text{Cov}[\hat{\theta}_1, \hat{\theta}_2]. \end{aligned}$$

De esta forma,

$$\text{Var}(\hat{\lambda}) = \exp\{-2\hat{\mu}/\hat{\sigma}\} [\text{Var}(\hat{\mu})/\hat{\sigma}^2 + \hat{\mu}^2 \text{Var}(\hat{\sigma})/\hat{\sigma}^4 - 2\hat{\mu} \text{Cov}(\hat{\mu}, \hat{\sigma})/\hat{\sigma}^3],$$

$$\text{Var}(\hat{\alpha}) = \text{Var}(\hat{\sigma})/\hat{\sigma}^4,$$

Tabla 4.2: Estimadores del modelo log lineal Weibull con el factor estado y la variable edad para los pacientes de cáncer de la laringe

Variable	Estimador del Parámetro	Error Estándar	Ji cuadrada de Wald	Valor p
Ordenada $\hat{\beta}_0$	3.53	0.90		
Escala $\hat{\sigma}$	0.88	0.11		
x_1 : Etapa II ($\hat{\beta}_1$)	-0.15	0.41	0.13	0.717
x_2 : Etapa III ($\hat{\beta}_2$)	-0.59	0.32	3.36	0.067
x_3 : Etapa IV ($\hat{\beta}_3$)	-1.54	0.36	18.07	< 0.0001
x_4 : Edad ($\hat{\beta}_4$)	-0.02	0.01	1.87	0.172

y

$$\text{Cov}(\hat{\lambda}, \hat{\alpha}) = \exp(-\hat{\mu}/\hat{\sigma})[\text{Cov}(\hat{\mu}, \hat{\sigma})/\hat{\sigma}^3 - \hat{\mu}\text{Var}(\hat{\sigma})/\hat{\sigma}^4].$$

Para incorporar variables auxiliares al modelo Weibull, es posible usar el formato lineal del logaritmo natural del tiempo, de manera tal que:

$$Y = \mathbf{x}^T \boldsymbol{\beta} + \sigma W,$$

donde $\mathbf{x}^T = (1, x_1, \dots, x_p)$ es el vector de variables explicativas y $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$ es el vector de coeficientes.

Ejemplo 4.4. Considere el estudio de 90 personas del sexo masculino con diagnóstico de cáncer en la laringe (Kardaun, 1983). Se registraron los intervalos (en años) entre el primer tratamiento y la muerte o el término del estudio. También se registró la edad de cada individuo en el momento del diagnóstico, y se usó un criterio médico para clasificar cuatro etapas del cáncer. Los datos se encuentran en el archivo `larynx.dat`. Se desea ajustar el modelo Weibull con la especificación

$$Y = \log T = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \sigma W,$$

donde x_i , $i = 1, \dots, 3$ son los indicadores de las etapas II, III y IV del cáncer, y x_4 es la edad del paciente. Los estimadores de los parámetros, los errores estándares, los estadísticos ji-cuadrada de Wald, y los valores p para probar que $\beta_i = 0$ están dados en la Tabla 4.2. A continuación se muestra la rutina en el lenguaje R para obtener la regresión de los datos de cáncer de laringe:

```
# Invoca la libreria de supervivencia
library(survival)

# Pon en un objeto los datos
larynx <- read.table("C:/Mis documentos/larynx.dat",
                    col.names=c("estado", "tiempo", "edad", "fecha", "estatus"))
# Crea el factor ETAPA en los datos larynx:
larynx$ETAPA <- factor(larynx$estado, labels = c("I", "II", "III", "IV"))

# Ajusta el modelo Weibull:
larynx.regwei <- survreg(Surv(tiempo, estatus)~ ETAPA + edad,
                        data=larynx, dist='weibull')

#Los coeficientes y el estimador de log(sigma) son:
coeficientes <- c(larynx.regwei$coefficients, log(larynx.regwei$scale))

# los errores estandares son:
errores <- sqrt(diag(larynx.regwei$var))

# el estadistico Wald de Ji cuadrada es:
ji.wald <- (coeficientes/errores)^2

# obten el valor p para cada parametro:
valor.p <- 1 - pchisq(ji.wald, 1)
```

4.4. El modelo Cox de riesgos proporcionales

4.4.1. Un modelo para la comparación de dos grupos

Supóngase que dos grupos de individuos - o unidades experimentales - son clasificados en dos grupos independientes A y B, y defina $h_A(t)$ y $h_B(t)$ las fuerzas de mortalidad en el tiempo t correspondientes a cada grupo. De acuerdo a un modelo simple de supervivencia, la fuerza de mortalidad en el tiempo t para un individuo del grupo B es proporcional a la fuerza de mortalidad en el mismo tiempo t para un individuo en el grupo A. Este *modelo de riesgos proporcionales* (en inglés,

proportional hazards model) se puede expresar en la forma

$$h_B(t) = \psi h_A(t),$$

para cualquier valor no negativo de t , donde ψ es una constante. Una implicación de esta estructura es que la función de supervivencia de los dos grupos no se cruzan.

El valor de ψ es el cociente de las fuerzas de supervivencia en cualquier tiempo t para un individuo en el grupo B relativo a uno en el grupo A. Si $\psi < 1$, la fuerza de mortalidad en t es menor para un individuo en B, relativo a un individuo en A; por otra parte, si $\psi > 1$, la fuerza de mortalidad en t es mayor para un individuo en B.

Es posible reparametrizar el modelo de riesgos proporcionales descrito aquí y generalizarlo para cualquier individuo i , $i = 1, \dots, n$, en la muestra combinada al considerar un factor con dos niveles. En este caso, se tiene una variable explicativa x_i la cual toma el valor uno si el i -ésimo individuo pertenece al grupo B y cero si pertenece al A. De esta forma, la fuerza de mortalidad para el i -ésimo individuo es

$$h_i(t) = \exp\{\beta x_i\} h_A(t);$$

aquí, $\psi = \exp \beta$, donde β puede tomar cualquier valor en los reales. Este es el *modelo de riesgos proporcionales* para la comparación de dos grupos.

4.4.2. El modelo de riesgos proporcionales generalizado

Cuando se considera un conjunto de variables explicativas, las cuales se representan con el vector $\mathbf{z}^T = (x_1, \dots, x_n)$, el modelo de riesgos proporcionales, se usa una fuerza de mortalidad de referencia $h_0(t)$ que corresponde a la de un individuo cuya forma funcional de las variables explicativas ligadas a la constante ψ dan el cero. La función $h_0(t)$ es llamada la *fuerza de mortalidad inicial* (en inglés, *baseline hazard function*). La fuerza de mortalidad para el i -ésimo individuo se puede

expresar como

$$h_i(t; \mathbf{z}_i) = \psi(\mathbf{z}_i)h_0(t),$$

donde $\psi(\mathbf{z}_i)$ es la función que liga las variables explicativas del i -ésimo individuo con el cociente de las fuerzas de supervivencia.

Como el riesgo relativo $\psi(\mathbf{z}_i)$ no puede ser negativo, es conveniente definirlo como $\psi(\mathbf{z}_i) = \exp\{\mathbf{z}^T \boldsymbol{\gamma}\}$, donde $\boldsymbol{\gamma}^T = (\gamma_1, \dots, \gamma_p)$ es el vector de coeficientes de las variables explicativas. De esta forma, el modelo de riesgos proporcionales generalizado es

$$h_i(t; \mathbf{z}_i) = \exp\{\mathbf{z}_i^T \boldsymbol{\gamma}\}h_0(t).$$

Como este modelo se puede re-expresar en la forma

$$\log\left(\frac{h_i(t; \mathbf{z}_i)}{h_0(t)}\right) = \mathbf{z}^T \boldsymbol{\gamma}$$

el modelo de riesgos proporcionales puede también ser visto como un modelo lineal para el logaritmo natural del cociente de los riesgos.

Nótese que el vector de coeficientes $\boldsymbol{\gamma}$ no contiene γ_0 , el coeficiente correspondiente a la ordenada; i.e. $\mathbf{z} = \mathbf{x}_{-1}$. Si se agregara el término γ_0 , la fuerza de mortalidad inicial sería re-escalada al dividir $h_0(t)$ por $\exp\{\gamma_0\}$, y el término constante se cancelaría, por lo que γ_0 es redundante.

Cuando se procede a estimar el modelo generalizado, sería apropiado estimar el vector de coeficientes $\boldsymbol{\gamma}$ y la fuerza de mortalidad inicial $h_0(t)$. Sin embargo, es posible demostrar que estos dos componentes se pueden estimar en forma donde $\mathbf{z}_{(j)}$ es el vector de variables explicativas para el individuo que muere en el j -ésimo tiempo de muerte ordenado $t_{(j)}$. La suma del denominador de la función verosimilitud corresponde a los valores de $\exp\{\mathbf{z}^T \boldsymbol{\gamma}\}$ de todos los individuos que se encuentran en riesgo en el tiempo $t_{(j)}$. Es posible notar que el producto se toma con las etiquetas de los individuos para los cuales se observó una muerte. Los individuos con tiempos censurados no tienen contribución en el numerador, su contribución se encuentra

en la suma de los individuos en riesgo del denominador, a separada. El vector γ se estima primero y luego estos estimadores se usan para construir un estimador de la fuerza de mortalidad origen. Este es un resultado importante sobre esta estructura pues implica que si se desean hacer inferencias sobre los efectos de las p variables explicativas en \mathbf{z} en presencia del riesgo relativo $h_i(t)/h_0(t)$, no es necesario estimar $h_0(t)$.

La estimación de γ se pueden estimar usando el método de máxima verosimilitud. Supóngase que los datos están compuestos por n individuos, de los cuales hay r tiempos de muertes diferentes y $n - r$ tiempos de supervivencia con censura. En este punto se supondrá que solo un individuo muere en cada tiempo de muerte en cada uno de los tiempos de muertes, i.e. se supondrá que no hay tiempos *empataados*. Los r tiempos de muertes ordenados se denotan con $t_{(1)} < t_{(2)} < \dots < t_{(r)}$, de manera tal que $t_{(j)}$ es el j -ésimo tiempo de muerte ordenado. El conjunto de individuos que se encuentran en riesgo en el tiempo $t_{(j)}$ se denotará con $R(t_{(j)})$, así que $R(t_{(j)})$ es el conjunto de individuos que se encuentran con vida y sin censura antes de $t_{(j)}$. La cantidad $R(t_{(j)})$ es conocida como *el conjunto en riesgo*.

Cox (1972) demostró que la función de verosimilitud relevante para el modelo de riesgos proporcionales está dado por

$$L(\gamma) = \prod_{j=1}^r \frac{\exp\{\mathbf{z}_{(j)}^T \gamma\}}{\sum_{l \in R(t_{(j)})} \exp\{\mathbf{z}_l^T \gamma\}},$$

donde $\mathbf{z}_{(j)}$ es el vector de variables explicativas para el individuo que muere el el j -ésimo tiempo de muerte ordenado $t_{(j)}$. La suma del denominador de la función verosimilitud corresponde a los valores de $\exp\{\mathbf{z}^T \gamma\}$ de todos los individuos que se encuentran en riesgo en el tiempo $t_{(j)}$. Es posible notar que el producto se tomcon las etiquetas de los individuos para los cuales se observó una muerte. Los individuos con tiempos censurados no tienen contribución en el numerador, su contribución se encuentra en la suma de los individuos en riesgo del denominador. De esta forma,

las inferencias sobre las variables explicativas en el modelo de supervivencia solo dependen de rango de orden de los tiempos de supervivencia.

Existen varias aproximaciones de la función de verosimilitud para cuando hay empates, estas pueden consultarse en Kalbleisch y Prentice (1980). La función `coxph` del paquete estadístico R ajusta el modelo semiparamétrico de Cox y además incluye algunas de las aproximaciones para cuando hay empates; en general, las variaciones de estas aproximaciones son muy pequeñas.