

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 de 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 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 se 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 $1/\lambda$, $X \sim \text{EXP}(1/\lambda)$, 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

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

the associated conditional hazard function is

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

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

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 la 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 resúmenes 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 de mortalidad integrada correspondiente puede expresarse como

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