

# 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  $\theta$ . 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  $\theta$ , y después encontrar el valor de  $\theta$  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  $L(\theta)$ , es entonces conveniente encontrar 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  $\theta$ , o funciones monótonas de  $\theta$  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 de coeficiente de confianza  $(1 - \alpha)$  para  $\theta$  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  $\theta$  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  $\theta$  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  $(t_i, \delta_i)$ . En esta notación  $\delta_i$  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 (??), 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  $\lambda$  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  $\lambda$  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  $\mu$  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

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

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  $\lambda$  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.