Tratamiento estadístico de datos de tiempos de vida. Análisis paramétrico

January 16, 2009

1 Introducción

Cuando se tiene un conjunto de datos de tiempos de vida en supervivencia o cualquier otro campo, hay que tener en cuenta la naturaleza y comportamiento de los mismos.

En un test experimental puede no ser factible continuar el experimento hasta que fallen todos los items en estudio. Los items que no han fallado aportan información parcial en el estudio. Esta situación se produce con normalidad pues en un ensayo de fiabilidad se consideran n items y se observan anotando los tiempos de fallo, limitándose en ocasiones el tiempo de observación o se observan sólo parte de los items, no fallando todos a la conclusión del mismo. Estos casos son frecuentes en ingeniería y medicina siendo los motivos múltiples; por ejemplo económicos, hacer un experimento con distintos mecanismos puede ser muy costoso; por la imposibilidad de un seguimiento continuado de todos los pacientes en un estudio médico, etc.

Formalmente se dice que una observación está censurada a la derecha en L si el valor exacto de la observación es desconocida pero se sabe que es mayor o igual que L. Se observa un individuo hasta un tiempo L en el que se pierde el contacto con él por distintos motivos.

De forma análoga, una observación está censurada a la izquierda en L si se sabe que la observación es menor o igual a L. Lo más común es que los datos presenten censura a la derecha.

Finalmente, se habla de censura aleatoria cuando cada individuo tiene un tiempo de vida y uno de censura aleatorios.

En todos estos casos se debe determinar la distribución y su correspondiente función de verosimilitud para hacer inferencia con los datos.

2 Tipos de censura

Dependiendo de la naturaleza del estudio se puede hablar de distintos tipos de datos censurados. Se presenta en esta sección la censura tipo II, tipo I,

multicensura, censura aleatoria y se comentará lo que se entiende por datos perdidos.

2.1 Censura tipo I: en tiempo

En este caso se observan n items terminando el estudio en un tiempo L prefijado de antemano. Este tipo de censura se conoce con el nombre de *en tiempo*. En este caso el número de fallos hasta el tiempo L prefijado es aleatorio, no fijo como en el caso de censura tipo II.

Si $T_1, T_2, \ldots, T_n$ son variables aleatorias independientes e idénticamente distribuidas y tienen una distribución continua de función de densidad $f(t)$ y función de supervivencia $S(t)$, la función de densidad conjunta de $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ es

$$\frac{n!}{(n-r)!} f(t_{(1)}) \ldots f(t_{(r)}) \left[S(L)\right]^{n-r}.$$

Recalcar que las verosimilitudes aunque aparentemente sean similares, no es así, puesto que en el caso de censura tipo I el número de fallos es aleatorio y fijo el tiempo de duración del estudio, mientras que en el muestreo censurado tipo II el tiempo de duración del estudio es aleatorio y fijo el número de fallos.

2.2 Censura tipo II: en número de fallos

Una muestra censurada tipo II es aquélla en la que sólo se consideran las r primeras observaciones habiendo n items en estudio, con $1 \leq r \leq n$. En este caso r es un valor fijo, determinado previamente por la naturaleza del estudio u otros motivos.

Formalmente se tienen n items y se consideran los r primeros tiempos de fallos $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ de una muestra aleatoria de n tiempos de vida $T_1, T_2, \ldots, T_n$. En este caso se ha notado por $T_{(i)}$ a tiempo de fallo que tiene lugar en i-ésima posición.

Si $T_1, T_2, \ldots, T_n$ son variables aleatorias independientes e idénticamente distribuidas y tienen una distribución continua de función de densidad $f(t)$ y función de supervivencia $S(t)$, la función de densidad conjunta de $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ es

$$\frac{n!}{(n-r)!} f(t_{(1)}) \ldots f(t_{(r)}) \left[S(t_{(r)})\right]^{n-r}.$$

En un modelo paramétrico se puede hacer inferencia con esta verosimilitud.

2.3 Multicensura

Es frecuente en algunos ensayos de fiabilidad asignar a cada ítem i tiempo de censura fijo y tiempo de vida o fallo aleatorio. A estos tiempos los notamos por L_i y T_i respectivamente, para $1 \leq i \leq n$.

Suponemos que $T_1, T_2, \ldots, T_n$ son variables aleatorias independientes e idénticamente distribuidas y tienen una distribución continua de función de densidad $f(t)$ y función de supervivencia $S(t)$.

El tiempo exacto de vida del item i se observa sólo si $T_i \leq L_i$, en otro caso es un dato censurado (el individuo o ítem sobrevive a su tiempo prefijado de observación). A cada ítem i se asigna el par (t_i, τ_i), siendo $t_i = \min\{T_i, L_i\}$ y $\tau_i = 1$ si $T_i \leq L_i$ y $\tau_i = 0$ en otro caso. De esta manera τ_i indica si el tiempo t_i del individuo i es de censura ($\tau_i = 0$) o de fallo ($\tau_i = 1$).

La función de verosimilitud en este caso es

$$L = \prod_{i=1}^{n} [f(t_i)]^{\tau_i} \cdot [S(t_i)]^{1-\tau_i} = \prod_{u} f(t_i) \cdot \prod_{c} S(L_i),$$

donde u representa el conjunto de tiempos no censurados y c el de censurados.

2.4 Censura aleatoria

Si avanzamos un poco en el caso anterior y se considera que la censura es aleatoria estamos en el caso de muestreo con censura aleatoria. Suponemos que el individuo i tiene un tiempo de vida T_i y un tiempo de censura L_i, variables aleatorias independientes entre ellas y pata todo i con funciones de supervivencia $S(t)$ y $G(t)$ respectivamente.

Los datos consisten en n observaciones (t_i, τ_i),$i = 1, \ldots, n$, siendo $t_i = \min\{T_i, L_i\}$ y $\tau_i = 1$ si $T_i \leq L_i$ y $\tau_i = 0$ en otro caso. La función de densidad de (t_i, τ_i) se obtiene fácilmente. Si la función de densidad de la variable T_i es $f(t)$ y la de L_i es $g(t)$ entonces

$$\begin{aligned} P(t_i = t, \tau_i = 0) &= g(t) \cdot S(t) \\ P(t_i = t, \tau_i = 1) &= f(t) \cdot G(t), \end{aligned}$$

lo que se puede expresar como

$$P(t_i = t, \tau_i) = [f(t) \cdot G(t)]^{\tau_i} \cdot [g(t) \cdot S(t)]^{1-\tau_i}.$$

Dada esta densidad, la función de verosimilitud para una muestra censurada aleatoriamente de n items o individuos es igual a

$$\begin{aligned} L &= \prod_{i=1}^{n} [f(t_i) \cdot G(t_i)]^{\tau_i} \cdot [g(t_i) \cdot S(t_i)]^{1-\tau_i}, \\ L &= \left[\prod_{i=1}^{n} G(t_i)^{\tau_i} \cdot g(t_i)^{1-\tau_i}\right] \cdot \left[\prod_{i=1}^{n} f(t_i)^{\tau_i} \cdot S(t_i)^{1-\tau_i}\right], \end{aligned}$$

lo cual es una generalización del caso censura tipo I considerando las distribuciones de los tiempos de censura como variables degeneradas.

2.5 Datos perdidos

Cuando se realiza un estudio de supervivencia haciendo un seguimiento de n individuos durante un periodo de tiempo determinado, es posible perder la información de alguno de ellos antes de finalizar el período. En un estudio con censura tipo I es sabido que el tiempo de supervivencia supera al de censura, pero en datos perdidos no se sabe lo que ha ocurrido.

Las soluciones más comunes para hacer un estudio con datos perdidos son muy diversos según distintos autores:

- Suprimir los individuos con datos perdidos, lo que conlleva a quedarse para hacer el estudio con una subpoblación determinada y con un tamaño de muestra reducido
- Se reemplazan los valores perdidos por la media de las observaciones presentes. Este método da buenos resultados y es el más aconsejable
- Otra solución es considerar los datos perdidos como casos censurados

3 Truncamiento por la izquierda

En un esquema muestral con truncamiento, sólo aquellos individuos que verifican cierta condición definida de antemano son observados por el investigador. En estudios de supervivencia, el tipo más común de truncamiento por la izquierda ocurre cuando los sujetos comienzan a ser observados a edades aleatorias, esto es, el origen del tiempo de vida precede al origen del estudio. En tal caso, aquellos sujetos en los que el fallo o la muerte tiene lugar antes del inicio del estudio generalmente son ignorados por el investigador. Como consecuencia, si T es el tiempo de vida de un sujeto y X es el tiempo en que éste se incorpora al estudio, un sujeto formará parte de la muestra únicamente si $T \geq X$. De este modo, bajo un esquema muestral con truncamiento a la izquierda, no se realiza inferencia sobre la variable T, sino sobre la variable T *condicionada a la ocurrencia del suceso* $T \geq X$.

Ejemplo. Channing House es un centro de retiro de la localidad Palo Alto, California. Se registraron datos correspondientes a las edades en el momento de fallecimiento de 462 individuos que vivían en la residencia durante el periodo de enero de 1964 a julio de 1975. Dado que un individuo debe sobrevivir un tiempo suficiente (65 años) para entrar en un centro de estas características, todos los individuos que fallecieron a edades tempranas no entrarán en el centro y, por lo tanto, quedan fuera del alcance de las investigaciones, es decir, tales individuos son excluídos del estudio dado que no viven el tiempo suficiente para entrar en el centro. De modo que si la información muestral se restringe a los habitantes de la residencia, no estamos registrando información sobre la variable T = tiempo de vida de un individuo de Palo Alto, sino de la variable T *condicionada al suceso* $T \geq 65$.

Una muestra con truncamiento por la izquierda y posible censura a la derecha viene dada por (x_i, t_i, δ_i), con $i = 1, 2, \ldots, n$, donde $t_i \geq x_i$, siendo t_i el tiempo

de vida observado y $x - i$ el tiempo de truncamiento. El valor de δ indica si el dato es censurado a la derecha o no. La función de verosimilitud se obtiene ahora como

$$L = \prod_{i=1}^{n} \frac{[f(t_i)]^{\tau_i} \cdot [S(t_i)]^{1-\tau_i}}{S(x_i)}$$

Esta figura muestral no debe confundirse con la censura por la izquierda. Para distinguir entre los dos fenómenos, supongamos en el ejemplo anterior que tenemos acceso a datos del resto de la población de esta localidad. Supongamos que conocemos el dato de un determinado individuo y que sabemos que éste falleció antes de ingresar en la residencia. En este caso contamos con un dato censurado a la izquierda de 65, y podemos incorporarlo a nuestro estudio.

4 Inferencia paramétrica con datos de tiempos de vida: estimación máximo verosimilitud

La técnica más usada para hacer estimación en modelos de tiempos de vida es el método de máxima verosimilitud.

Sea Ω el conjunto de valores de β. El estimador $\hat{\beta}$ de β que maximiza la función de verosimilitud $L(\beta)$ se llama estimador máximo verosímil. En la mayoría de los casos, el estimador máximo verosímil se obtiene resolviendo las ecuaciones de verosimilitud. Si suponemos que la dimensión del vector β es k, entonces se tiene

$$\frac{\partial \lg L(\beta)}{\beta_i} = 0 \quad ; \quad i = 1, \ldots, k.$$

Notamos la solución como $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$.

4.0.1 Matriz de información de Fisher

Se denomina matriz de información de Fisher a la matriz $I(\beta)$ cuyos elementos son $I_{ij}(\beta)$, $i, j = 1, \ldots, k$ siendo

$$I_{ij}(\beta) = E\left[-\frac{\partial^2 \lg L(\beta)}{\partial \beta_i \beta_j}\right].$$

4.0.2 Matriz de información

Se denomina matriz de información a la matriz

$$I_0(\hat{\beta}) = \left|-\frac{\partial^2 \lg L(\beta)}{\partial \beta_i \beta_j}\right|_{\beta=\hat{\beta}}.$$

4.1 Distribuciones asociadas al estimador

Se puede ver que el estimador máximo verosímil $\hat{\beta}$ es una variable aleatoria que se distribuye asintóticamente como

$$\hat{\beta} \to N\left(\beta, I(\beta)\right),$$

siendo β el verdadero valor del parámetro β.

El elemento i-ésimo de $\hat{\beta}$, $\hat{\beta}_i$, se distribuye entonces como

$$\hat{\beta}_i \to N(\beta_i, (I(\beta)_{ii}),$$

siendo $(I(\beta)_{ii}$ el elemento i-ésimo de la diagonal principal de la matriz de información de Fisher en el verdadero valor del parámetro.

En la práctica, cuando se hace inferencia, se ha de tener en cuenta que el verdadero valor del parámetro β no es conocido en la matriz de información de Fisher. En este caso se puede reemplazar $I(\beta)$ por $I_0(\hat{\beta})$, obteniendo asintóticamente que

$$\hat{\beta}_i \to N(\beta_i, (I_0(\hat{\beta})_{ii}).$$

Desde este estadístico se pueden construir intervalos de confianza para β_i y realizar tests de hipótesis.

Por otro lado también se tiene que si β_0 es el verdadero valor del parámetro β entonces asintóticamente se tiene que

$$\hat{\Lambda} = -2 \lg \frac{L(\beta_0)}{L(\hat{\beta})} \to \chi^2_k.$$

donde k es la dimensión del parámetro β.

Para muestras de gran tamaño, de la normalidad asintótica de $\hat{\beta}$ se obtiene

$$\left(\hat{\beta} - \beta_0\right)' I_0(\hat{\beta})^{-1} \left(\hat{\beta} - \beta_0\right) \to \chi^2_k,$$

pudiendo ser utilizado para contrastar la hipótesis $\beta = \beta_0$.

4.2 El estadístico Score

Suponemos que se tienen n observaciones independientes $x_1, \ldots, x_n$ de una variable aleatoria cuya ley de probabilidad depende del parámetro $\sigma' = (\sigma_1, \ldots, \sigma_p)$. La verosimilitud de σ es

$$L(\sigma) = \prod_{i=1}^{n} L_i(\sigma).$$

En el caso de muestras censuradas se tiene que $x_i = (t_i, \tau_i)$, siendo t_i el tiempo de fallo o censura y τ_i el identificador de ésta. El vector score se define como

$$U(\sigma) = \sum_{i=1}^{n} U_i(\sigma),$$

donde

$$U_i(\sigma) = \frac{\partial}{\partial \sigma} \lg L_i(\sigma) = \left[\frac{\partial}{\partial \sigma_j} \lg L_i(\sigma) \right]_{j=1,\ldots,p} \qquad i = 1, \ldots, n.$$

Si se pueden intercambiar las operaciones de integración respecto x_i y derivación respecto σ, se puede demostrar que $U_i(\sigma)$ tiene esperanza cero y matriz de covarianzas

$$I_i(\sigma) = E\left[U_i(\sigma) U_i'(\sigma)\right] = -\left[E\left(\frac{\partial^2 \lg L_i}{\partial \sigma_j \partial \sigma_k} \right) \right]_{j,k=1,\ldots,p}$$

En este estudio no se consideran modelos con parámetro umbral.

Dado que los x_i son independientes, entonces los U_i también lo son y aplicando el teorema central del límite se tiene que $U(\sigma)$ se distribuye asintóticamente normal con media cero y matriz de covarianzas

$$I(\sigma) = \sum_{i=1}^{n} I_i(\sigma).$$

El estadístico score se puede interpretar como el gradiente del logaritmo de la verosimilitud. La distribución asintótica de $U(\sigma)$ se puede usar para hacer inferencia sobre σ. Si $H_0 \equiv \sigma = \sigma_0$, bajo H_0, $U(\sigma_0)$ es asintóticamente normal con media el vector cero y varianza la matriz $I(\sigma_0)$. Si $I(\sigma_0)$ es no singular se tiene que

$$U'(\sigma_0) I^{-1}(\sigma_0) U(\sigma_0) \to \chi^2_p.$$