Teoría de Riesgos II
Departamento de Matemáticas
Universidad de Guanajuato

Modelo de Competing Risk usando el algoritmo EM

Roberto Vásquez Martínez

9 de Junio de 2020

Índice general

1. Resumen

Consideremos un sistema el cual está hecho de múltiples componentes en serie. En este caso, cuando el fallo de el sistema está causado por el fallo temprano de algún componente, y esto es comúnmente referido como *competing risk*.

En algunas situaciones, es complicado o costoso revisar la razón específica de el fallo, pero se da el caso que el *tiempo de fallo* se observa, pero su correspondiente causa de fallo no está por completo investigada, es decir, los datos están incompletos en ese sentido.

Además este modelo de *competing risk* se puede tornar más complicado debido a una censura de los datos. En la práctica, la censura de los datos es muy común debido a las consideraciones especiales que pueden tener algunos experimentos.

En este proyecto se divide en dos partes, en primer lugar, describiremos un modelo general de *competing risk* y en segundo lugar, mediante el algoritmo EM, resolveremos un caso de *competing risk* suponiento *tiempos de fallo* distribuidos log-normalmente, esto para ejemplicar algunas consideraciones importantes al intentar implementar este modelo.

2. Modelación de el Competing Risk

En este apartado, nuestra intención es dar las definiciones y las consideraciones previas para modelar matemáticamente un escenario de *competing risk*.

2.1. Consideraciones previas

Para comenzar, daremos algunas definiciones importantes

Definición 2.1.1 (Tiempo de supervivencia) *Sea X el tiempo de supervivencia y F_X su función de distribución. Definimos la* ***función de supervivencia*** *como*

$$S_X(x) = \mathbb{P}[X > x] = 1 - F_X(x).$$

En otras palabras, la función de supervivencia es la probabilidad de sobrevivir más allá de el tiempo x.

Otra definición importante para describir el modelo es la siguiente

Definición 2.1.2 (Función de Riesgo (Hazard)) *Sea X una variable aleatoria con densidad f_X y función de supervivencia Definimos la función de riesgo (hazard function) de la variable aleatoria X como*

$$h_X(x) = \frac{f_X(x)}{S_X(x)}$$

Remark 2.1.3 *La función de riesgo no es una densidad o una función de distribución de probabilidad. Sin embargo, podemos pensar que es la probabilidad de falla en un periodo de tiempo infinitesimal, es decir, es la tasa de ocurrencia de los fallos.*

2.2. Modelación de el Competing Risk

Los modelos de *competing risk* consideran varios alternativas y modos de fallo simultáneas, y la falla se dice que ocurre tan pronto la primer falla tiene lugar, por lo que las siguientes suposiciones son las que se hacen

1. Un conjunto de diferentes modos de fallos se considera con sus respectivas funciones de supervivencia $\{S_i(x) : i \in I\}$
2. El fallo de un sistema se dice que ocurre bajo *el principio de el enlace más débil*, esto es que la función de supervivencia de el sistema queda expresada de la siguiente forma

$$S(x; I) = \prod_{i \in I} S_i(x)$$

Bajo estos supuestos, consideremos un sistema i formado de $\{1, 2, \ldots, J\}$ componentes y que la causa de el fallo puede o no estar identificada, es decir, la causa de el fallo se encuentra en algún subconjunto de $\{1, 2, \ldots, J\}$.
Por ejemplo

- Si sabemos que fallo ocurre por la causa j, el conjunto de posibles causas, denotado por M_i, vendría siendo $M_i = \{j\}$ con $j \in \{1, 2, \ldots, J\}$.
- Si la causa de el fallo es completamente desconocida entonces $M_i = \{1, 2, \ldots, J\}$
- Si la causa de el fallo es identificada pero el conjunto de posibles casos contiene más de un elemento, entonces en este caso la causa es parcialmente conocida y $M_i = \{j_1, \ldots, j_k\} \subsetneq \{1, 2, \ldots, J\}$

Denotemos por $T_i^{(j)}$ los tiempos de vida de el $i-$ésimo sistema debido a la causa j, donde $i = 1, 2, \ldots, n$, es decir, estamos considerando n sistemas.
Haremos las siguientes suposiciones

- $T_i^{(j)}$ son variables aleatorias independientes para todo i, j.
- $T_i^{(j)}$ son idénticamente distribuidas para todo i dado el j.

Las correspondientes funciones de distribución, densidad, supervivencia, y riesgo las denotaremos respectivamente como

$$F^j(\cdot|\theta^j), f^j(\cdot|\theta^j), S^j(\cdot|\theta^j), h^j(\cdot|\theta^j),$$

donde θ^j es el vector de parámetros que caracteriza a la distribución de el componentes j
Considerando lo anterior surge la siguiente definición

Definición 2.2.1 (Tiempo de Vida Observado) *El tiempo de vida observado de el sistema i viene dado por la siguiente variable aleatoria*

$$T_i = \text{mín}\{T_i^{(1)}, T_i^{(2)}, \ldots, T_i^{(J)}\}$$

Comúnmente, en problemas de la vida real, observaciones completas de T_i no son posibles debido a distintas formas de censura inherentes a la colección de los datos. Suponemos así que T_i se puede censurar por los tiempos C_i, los cuales son independientes de los T_i para todo i.

Definición 2.2.2 (Indicador de censura) *Decimos que el sistema i está censurado si* $M_i = \emptyset$.

Además nuestro indicador de censura *lo denotamos por*

$$\Delta_i = \begin{cases} -1 & si\ |M_i| > 1 \\ j & si\ M_i = \{j\} \\ 0 & si\ está\ censurado \end{cases}$$

Sea $X_i = \text{mín}\{T_i, C_i\}$, luego la colección de datos queda determinada por la terna (X_i, Δ_i, M_i), y denotamos una realización de (X_i, Δ_i) como (x_i, δ_i)

Con esta notación, procederemos a describir una forma de estimar θ^j para cada j.

2.2.1. Estimación por Máxima Verosimilitud

Aquí describiremos el método general por máxima verosimilitud

Consideremos $\mathbb{I}[A]$ la función indicadora de el conjunto A. Por simplicidad, denotamos por $\mathbb{I}_i(j) = \mathbb{I}[\delta_i = j]$ y $\Theta = (\theta^1, \ldots, \theta^J)$.

Se puede ver que la verosimlitud de los datos censurados es tal que

$$L(\Theta) \propto \prod_{j=1}^{J} \prod_{i=1}^{n} L_i(\theta^j), \tag{2.1}$$

donde

$$L_i(\theta^j) = [f^j(x_i)]^{\mathbb{I}_i(j)} \prod_{\substack{l=0 \\ l \neq j}}^{J} [S^j(x_i)]^{\mathbb{I}_i(l)} \tag{2.2}$$

Observamos que maximizar $L(\Theta)$ es equivalente a maximizar individualmente $L(\theta^j)$ para cada j.

El siguiente paso, para calcular la verosimilitud de los datos es hallar la función de distribución de T_i e insertarla en la verosimilitud anterior, para ello necesitamos ver la noción de **función de incidencia acumulada**, que definimos a continuación.

Definición 2.2.3 (Función de Incidencia Acumulada) *La función de incidencia acumulada para cada $j \in \{1, 2, \ldots, J\}$ queda definida como*

$$G(t, j) = \mathbb{P}[T_i \leq t \ y \ \Delta_i = j]$$

Y su correspondiente densidad viene dada por

$$g(t, j) = h^j(t) \prod_{l=1}^{J} S^l(t)$$

Remark 2.2.4 *La función de incidencia cumulativa representa la probabilidad que un evento de el tipo j ocurra al tiempo t.*

Considerando la definición anterior tenemos que la función de densidad correspondiente a T_i es

$$f^{(M_i)}(t) = \sum_{j \in M_i} g(t, j) = \sum_{j \in M} \left(h^j(t) \prod_{l=1}^{J} S^l(t) \right).$$

De esto último, se tiene que la verosimilitud de los datos censurados y ocultos viene dada por

$$L^*(\Theta) \propto \prod_{i=1}^{n} L_i^*(\Theta) \tag{2.3}$$

donde

$$L_i^*(\Theta) = \left[\prod_{j=1}^{J} L_i(\theta^j) \right] \left[f^{(M_i)}(x_i) \right]^{\mathbb{I}_i(-1)} \tag{2.4}$$

En general, la no hay forma cerrada de la verosimilitud anterior, por lo que para maximizarla se recurren a métodos numéricos, para evitar esto utilizaremos el algoritmo EM, que introduciremos en la siguiente sección.

3. El Algoritmo EM

A continuación, introduciremos el algoritmo EM y desarrollaremos funciones de verosimilitud apropiadas que servirán como inputs de este algoritmo.

El algoritmo EM (Expectation-Maximization) es un algoritmo iterativo general que se utiliza para calcular estimadores de máxima verosimilitud cuando no hay formas cerradas para las estimaciones o los datos están incompletos.

Este algoritmo consta de dos pasos llamados E-Step y M-Step, que describimos a continuación:

- E-Step: Aquí se calcula la esperanza condicional la log-verosimilitud con respecto a los datos incompletos dados los datos observados.
- M-Step: Aquí se maximiza esa esperanza.

Este método converge seguramente, si converge entonces converge a un máximo local. Así, en el caso de una función cóncava unimodal, el algoritmo EM converge al máximo global para cualquier valor inicial.

Ahora procederemos a ejemplificar estos pasos.

Sea $L^C(\Theta|\mathbf{x})$ la verosimilitud con los datos completos.

Denotemos a la parte observada de $\mathbf{x} = (x_1, \ldots, x_n)$ por $\mathbf{y} = (y_1, \ldots, y_m)$ y a la parte oculta como $\mathbf{z} = (z_{m+1}, \ldots, z_n)$, si Θ_k es la estimación en la iteración k de el algoritmo entonces los pasos de el algoritmo EM se reducen a

- E-Step: Calcular
$$Q(\Theta|\Theta_k) = \mathbb{E}[\log L^C(\Theta \mid \mathbf{y}, \mathbf{Z}) \mid \Theta_k]$$
- M-Step: Encontramos Θ_{k+1} maximizando $Q(\Theta|\Theta_k)$ sobre Θ

Aplicaremos esta heurística para describir cada una de las distribuciones de los componentes de los sistemas, para ello necesitamos hallar la verosimilitud de los datos completos.

3.0.1. Construyendo la Verosimilitud de los datos completos

Para construir esta verosimilitud necesitamos tratar la causa de el falla como datos ocultos. Consideremos $U_i^{(j)} = \mathbb{I}[\Delta_i = j | X_i = x_i]$, entonces $U_i^{(j)}$ tiene distribución Bernoulli tal que

$$\mathbb{P}[U_i^{(j)} = 1] = \mathbb{P}[\Delta_i = j | X_i = x_i],$$

de las consideraciones que habiamos hecho sobre la función de riesgo se puede ver que

$$E[U_i^{(j)}] = \begin{cases} \frac{h^j(x_i)}{\sum_{l \in M_i} h^l(x_i)} & \text{si } j \in M_i \\ 0 & \text{si } j \notin M_i \end{cases}$$

Luego para obtener la verosimilitud completa debemos reemplazar $f^{(M_i)}(x_i)$ en (2.4) por

$$\prod_{j=1}^{J} [f^j(x_i)]^{U_i^{(j)}} [S^j(x_i)]^{1-U_i^{(j)}}$$

Obtenemos así

$$L_i^C(\Theta) = \prod_{j=1}^{J} L_i^C(\theta^j),$$

donde

$$\begin{aligned} L_i^C(\theta^j) =& \{L_i(\theta^j)\} f^j(x_i)]^{U_i^{(j)}} [S^j(x_i)]^{1-U_i^{(j)}} \\ =& \{h^j(x_i)\}^{U_i^{(j)}} S^j(x_i) \end{aligned} \tag{3.1}$$

Por lo que podrías aplicar el algoritmo a partir de este punto, sin embargo, no todas las distribuciones tienen forma cerrada para la función de riesgo por lo que no sería sencillo aplicar el algoritmo en esos casos y tendríamos que recurrir a herramientas de índole numérica.

Por otro lado, sin consideramos los datos censurados como datos ocultos, es posible escribir la verosimilitud completa en forma cerrado respecto a la densidades.

Para ello consideremos $Z_i^{(j)}$ el truncamiento de $T_i^{(j)}$ con $Z_i^{(j)} > x_i$.

Entonces se tiene que

$$L_i^C(\theta^j) = \{f^j(x_i)\}^{U_i^{(j)}} \{f^j(Z_i^{(j)})\}^{1-U_i^{(j)}}, \tag{3.2}$$

donde la función de densidad de $Z_i^{(j)}$ viene dada por

$$f_Z^j(t|\theta^j) = \frac{f^j(t)}{1 - F^j(x_i)},$$

y así podemos aplicar el algoritmo sin preocuparnos de no obtener una fórmula cerrada, pues la mayoría de las distribuciones tienen densidad.

4. Resultados

Aquí implementaremos precisamente esta última heurística en el caso que los tiempos de cada componente se comportan de manera log-normal.

Este ejemplo lo haremos considerando $J = 3$ y $n = 60$, y simulando

$$\log(T_i^{(j)}) \sim N(\mu^{(j)}, \sigma^{(j)})$$

donde $\mu^{(j)} = \log(5) \approx 1.6094$ y $\sigma^{(1)} = 2$, $\sigma^{(2)} = 1$ y $\sigma^{(3)} = 0.5$.

A continuación, mostramos una tabla en la que se muestran las estimaciones con el Algoritmo EM considerando datos ocultos así como con el conocimiento de la causa real de el fallo de el sistema, que llamamos datos completos.

Causas	$\mu^{(1)}$	$\sigma^{(1)}$	$\mu^{(2)}$	$\sigma^{(2)}$	$\mu^{(3)}$	$\sigma^{(3)}$
Completos	1.549	2	1.469	0.839	1.691	0.602
Ocultos	1.575	2.026	1.508	0.886	1.622	0.53

Se puede ver que ambas estimaciones son considerablemente parecidas, y más aún se parece los valores simulados.

La precisión está fuertemente relacionada con que tantos datos fueron censurados, en nuestra base de datos se censuró alrededor de el 36 % de los tiempos de falla.

Cabe observar alrededorque cuando se sabe el componente exacto de falla se obtiene el estimador clásico de máxima verosimilitud.

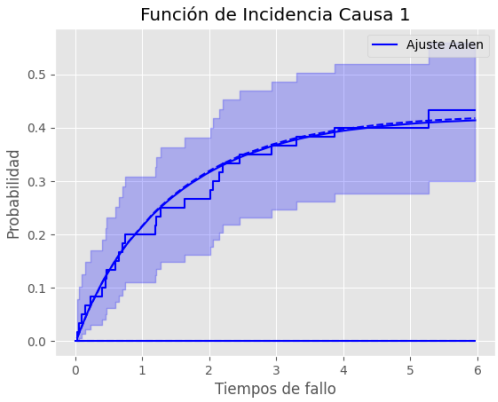
Finalmente en la figura siguiente mostramos mostramos las gráficas que representan el ajuste en cada caso de las funciones de incidencia y para validar el utilizamos el ajuste de la CIF empírica de Aalen.

Se sabe que no se puede calcular la CIF empírica para los datos ocultos con el ajuste de Aalen, por lo que se recurre a una heurística respecto a los conjuntos M_i para generar así dos CIF que representarán los límites de la banda.

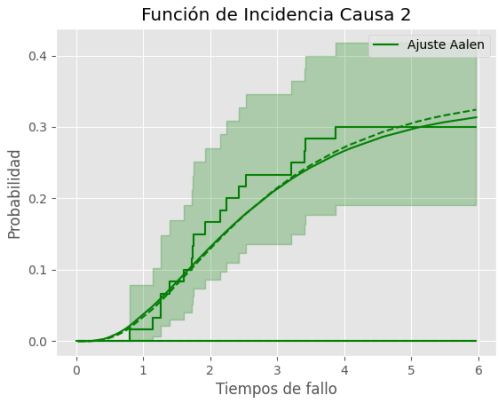
Podemos ver en estás gráficas que suponer la log-normalidad de los datos es razonable pues la CIF de los datos complejos ajusta de manera prudente a la aproximación empírica de la CIF de Aalen.

Figura 4.1: CIF empírica vs Las estimaciones

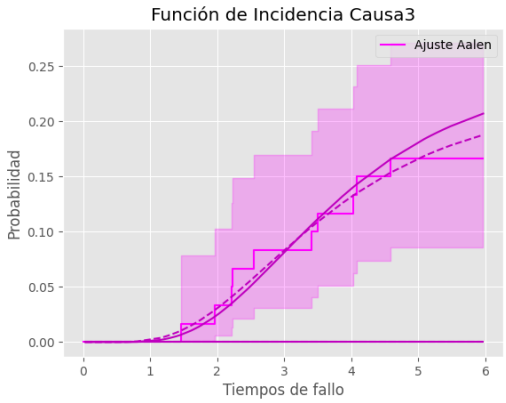
(a) Causa 1



(b) Causa 2



(c) Causa 3



Por último, podemos analizar y hacer una especie de *clustering* de los tiempos respecto a los componentes que se pueden dañar, es decir, dado el tiempo, en probabilidad podemos decir

que componente pudo dañarse y en consecuencia dañar al sistema en ese periodo de tiempo determinado. Esto lo muestra la siguiente gráfica.

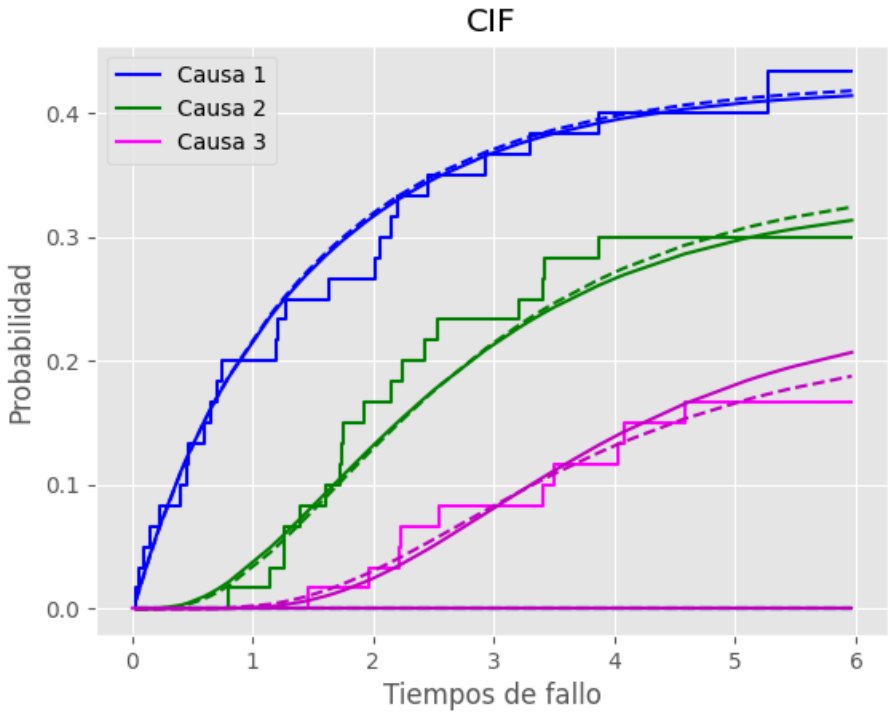


Figura 4.2: Comparativo de la Incidencia de cada componente

Bibliografía

[1] Ali S. Balakrishnan N. Castillo, E. *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley-Interscience.

[2] Chanseok Park. Parameter estimation of incomplete data in competing risk using the em algorithm. *IEEE Transactions on Reliability*, 54(2):282–290, 2005.

[3] Germán Rodríguez. Cumulative incidence, 2012.