

1. Modelos de Duración y Duración Marcada

1.1. Introducción

En este capítulo se hablará sobre los modelos de duración y de duración marcada y su aplicación en el objeto de este trabajo. Además de algunas propiedades, tales como la independencia, intercambiabilidad y, por supuesto, estacionariedad que son vitales para realizar inferencia y predicción de los datos.

Es importante remarcar que la historia de los procesos puntuales siempre ha estado unida a aquella de la estadística actuarial y de seguros, como nos mencionan Daley y Vere-Jones(2003) al referirse a las tablas de mortalidad como el primer estudio de procesos de intervalos. Por lo que el empleo de estos procesos como un método de tarificación es solamente otra colaboración en la larga lista de estas dos disciplinas.

1.2. Definición del proceso de Duración y Duración Marcada

Para el objeto de este estudio tenemos una muestra de microcostos de enfermedades crónicas de un cierto número de individuos a los que se les ha observado durante un período de tiempo. A su vez, cada uno de los individuos tiene asociadas covariables sociodemográficas, socioeconómicas y médicas. De este modo, podríamos decir que tenemos n individuos $(n_i)_{i=1}^n$ observados por un período de tiempo con costos asociados a su padecimiento. El objetivo es modelar y predecir la duración y el costo de las etapas de estos padecimientos por individuo.

Supongamos que empezamos el estudio de un individuo n_i en el tiempo $t_{i0} = 0$, es decir, este es el tiempo en el que el individuo entra al panel de estudio. La duración del estudio para el individuo es T_i , esto no quiere decir que no puedan ocurrir observaciones posteriores a T_i , a esto se le conoce como censuramiento de datos por la derecha.

Según Paik Schoenberg(2000), un proceso puntual es una medida aleatoria en un espacio métrico separado S tomando valores en los enteros no negativos

Z^+ (o infinito) donde $N(t)$, en un caso particular, es un proceso de conteo del número de puntos que ocurren antes del tiempo t .

Sea $t_{ij} \in (t_{i0}, T_i]$ el momento en el que ocurre un cambio de tratamiento, por lo que definimos la variable aleatoria $N(t)$ que cuenta el número de cortes o cambios en el intervalo.

Dado que la muestra consiste en microcostos a través del tiempo, decimos que a cada t_j se le asocia la variable costo de tratamiento; es decir, a cada momento en que ocurre un cambio de tratamiento le corresponde un nuevo costo p_j . De este modo, para cualquier individuo n_i tenemos una sucesión de variables asociadas $\{t_{i1}, p_{i1}\}, \{t_{i2}, p_{i2}\}, \dots, \{t_{ik}, p_{ik}\}$. De este modo la sucesión de variables es una colección aleatoria de puntos en un espacio con una marca asociada a cada punto, así ya se pueden modelar los datos como en un proceso puntual marcado.

Daley y Vere-Jones (2003) definen el proceso puntual marcado como un proceso localizado en un espacio métrico completamente separado χ y las marcas en otro espacio métrico completamente separado κ , entonces $\{(\chi_i, \kappa_i)\}$ en $\chi \times \kappa$ es un proceso puntual marcado con la propiedad adicional de que el proceso primario $N(t)$ es a su vez un proceso puntual.

Lo que deseamos conocer es,

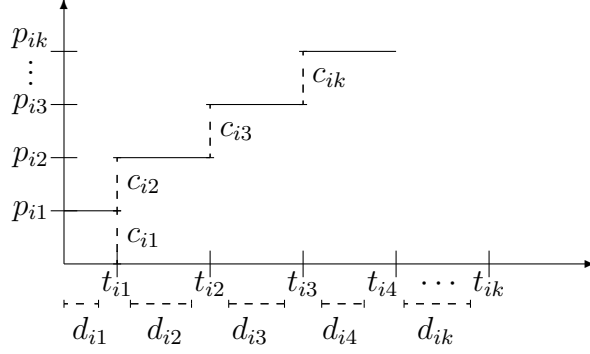
$$P(t_{i1}, \dots, t_{ik}, p_{i1}, \dots, p_{ik}) = P(t_{i1}, \dots, t_{ik}, p_{i1}, \dots, p_{ik} | N(t)) \quad (1)$$

Es decir, la función de distribución conjunta del tiempo de ocurrencia de los eventos y los precios asociados a estos es igual a la función de distribución de estas variables condicionados por la variable aleatoria del número de eventos en el intervalo $(t_{i0}, T_i]$. Sin embargo, dado que al usar las variables en sus valores absolutos estas pueden dar saltos muy altos entre si, por lo que debemos usar variables alternas.

Definimos las siguientes variables para un individuo n_i :

- $d_{ij} = t_{ij} - t_{ij-1}$, donde d_{ij} es la duración entre los tiempos de ocurrencia de cada individuo.

- $c_{ij} = c_{ij} - c_{ij-1}$, donde c_{ij} representa el costo, es decir, la diferencia entre los precios en cada tiempo de ocurrencia de cada individuo.



De este modo,

$$P(t_{i1}, \dots, t_{ik}, p_{i1}, \dots, p_{ik} | N(t)) \cong P(d_{i1}, \dots, d_{ik}, c_{i1}, \dots, c_{ik} | N(t)) \quad (2)$$

Esto quiere decir que calcular la función de distribución conjunta de los tiempos de ocurrencia y los precios asociados a éstos es análogo a calcular la función de distribución conjunta de las duraciones y los costos asociados condicionados a la variable aleatoria del número de eventos en el intervalo de tiempo. Así pasamos de un proceso puntual marcado a uno de duración marcada.

1.3. Propiedades del Proceso de Duración Marcada

Una vez que hemos definido qué es el proceso de duración y de duración marcada y cómo es que los datos que tenemos para este estudio se adaptan a este modelo, necesitamos especificar las propiedades que van a hacer posible la inferencia y la predicción. Estas propiedades son la independencia, la intercambiabilidad y, principalmente, la estacionariedad.

1.3.1. Independencia

En una concepción tradicional, Resnick(1999) define la independencia de un número finito de eventos como:

Definición 1. *Los eventos A_1, \dots, A_n ($n \geq 2$) son independientes si*

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i), \quad I \subset \{1, \dots, n\}$$

Los eventos son independientes si la probabilidad de la intersección de estos eventos o la probabilidad conjunta de los eventos es igual a la multiplicación de la probabilidad de los mismos.

Análogamente, podemos hacer la definición de independencia para el proceso de duración marcada. Recordemos que tenemos la función de probabilidad conjunta de las duraciones y los costos, por lo que la independencia en el proceso es:

$$P(d_1, c_1, \dots, d_k, c_k | N(t) = k) = \prod_{j=1}^{N(t)} P(d_j, c_j) \quad (3)$$

En este caso, la única diferencia reside en el hecho de que el número de funciones de probabilidad a multiplicar es a su vez una variable aleatoria, la cual se encarga de contar los cambios en el costo de tratamiento en el tiempo. El supuesto de independencia es útil para la inferencia de futuras observaciones.

1.3.2. Intercambiabilidad

Otra propiedad muy importante para la inferencia y predicción de variables en un proceso de duración marcada es la intercambiabilidad que, de acuerdo a Hahn y Zhang (2000), se define como:

Definición 2. *Una sucesión de variables $X = (X_1, X_2, \dots, X_n)$ es intercambiable si para cada n*

$$(X_1, X_2, \dots, X_n) = (X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)})$$

para cualquier permutación σ de $1, 2, \dots, n$.

Si la sucesión de variables es independiente e idénticamente distribuida entonces es intercambiable. El concepto de intercambiabilidad está muy relacionado con la independencia, pues la independencia es un caso particular de la intercambiabilidad.

Para poder entender mejor la propiedad podemos citar el Teorema de Fenetti(1937) que nos dice,

Teorema 3. *Una sucesión infinita de variables aleatorias intercambiabiles $\bar{X} = (X_1, X_2, \dots)$ es una mezcla de variables independientes e idénticamente distribuidas (i.i.d). Esto es, que existe un espacio de probabilidad (U, Θ) tal que*

$$P(\bar{X} \in B) = \int_U P(\bar{X}(u) \in B) \Theta(du)$$

donde $\bar{X}(u) = (X_1(u), X_2(u), \dots)$ es una secuencia de variables aleatorias i.i.d. y $\Theta(\cdot)$ es una medida de probabilidad.

Esto se puede adaptar al proceso de duración marcada correspondiente a este análisis de la siguiente manera, tomando el Teorema de Fenetti

$$P(d_1, c_1, \dots, d_k, c_k | N(t) = k) = \int_{\Theta} \prod_{j=1}^{N(t)} P(d_j, c_j | \theta) \pi(\theta) d(\theta) \quad (4)$$

donde θ es una variable aleatoria no observable y $\pi(\theta)$ es una medida de probabilidad común a todas las variables aleatorias. Es decir, que a lo postulado en el apartado de independencia le agregamos la variable no observable con su respectiva medida de probabilidad, sobre cuyo espacio de probabilidad está definida la integral. La variable no observable común a todas las variables aleatorias es un tema que se desarrollará a profundidad en el siguiente capítulo.

1.3.3. Estacionareidad

Una vez que han sido definidas la independencia y la intercambiabilidad faltaría definir la estacionareidad para poder hacer predicciones sobre futuras observaciones.

De manera intuitiva, podemos definir la estacionareidad en un proceso de duración cuando la función de probabilidad conjunta del proceso no cambia cuando es ésta es desplazada en el tiempo, lo cual indicaría que lo importante es la longitud de los intervalos, no la localización de los mismos. Sin embargo, de una manera más técnica, Daley y Vere-Jones(2003) definen la estacionareidad en un proceso como:

Definición 4. *Un proceso puntual es estacionario por intervalos cuando para cada $r = 1, 2, \dots$ y todos los enteros i_1, \dots, i_r , la distribución conjunta de $\{\tau_{i_1+k}, \dots, \tau_{i_r+k}\}$ no depende de k ($k = 0, \pm 1, \dots$).*

Esto implicaría que el orden de las observaciones importa y que las observaciones pasadas ayudan a construir la variable aleatoria. Es decir que con una sucesión de variables $\bar{X} = (X_1, \dots, X_n)$ tendríamos que,

$$P(X_n, \dots, X_1) = P(X_n|X_{n-1}, \dots, X_1) * P(X_{n-1}|X_{n-2}, \dots, X_1) * \dots * P(X_2|X_1) * P(X_1)$$

Así si la variable aleatoria depende de su historia, podríamos entonces predecir observaciones futuras. Es decir, que para toda $s \geq 0$

$$\begin{aligned} P(X_{n+1}, X_n, \dots, X_1) &= P(X_{n+1}|X_n, \dots, X_1) \\ &= P(X_{n+s+1}|X_{n+s}, \dots, X_{1+s}) \end{aligned}$$

De este modo, para el proceso de duración marcada la estacionareidad se podría plantear como,

$$P(d_1, c_1, \dots, d_k, c_k | N(t) = k) = \prod_{j=2}^{N(t)} P(d_j, c_j | d_{j-1}, c_{j-1}) * P(d_1, c_1)$$

Lo que quiere decir que la función conjunta de probabilidad se puede definir con base a observaciones pasadas.

Una vez que nuestro modelo de duración marcada cumple las propiedades descritas en este capítulo podemos empezar a hacer inferencia sobre las variables y predecir las observaciones futuras. En el siguiente capítulo, desarrollaremos un modelo complementario de variables latentes que terminaría de conectar la idea de la variable no observable presentada en el concepto de intercambiabilidad con el resto de la sucesión.