

Estimación basada en cópulas para tiempos de eventos sucesivos censurados

Ana Panduro Martín

Este documento contiene un resumen del Trabajo de Fin de Máster titulado *Estimación basada en cópulas para tiempos de eventos sucesivos censurados*, realizado por Ana Panduro Martín bajo la dirección de Jacobo de Uña Álvarez. Dicho trabajo es el inicio de una tesis doctoral, motivo por el cual no se autoriza la publicación de la memoria.

El Análisis de Supervivencia multivariante se ocupa del estudio de dos o más tiempos de vida. Dichos tiempos, en general, van a ser dependientes al estar definidos sobre el mismo individuo. Algunos ejemplos de este tipo de datos son los riesgos competitivos, los datos en paralelo y los tiempos de eventos sucesivos.

Este trabajo se centra en el estudio de tiempos de eventos sucesivos bivariantes. En este caso, tendremos un par (X_1, X_2) que representa el tiempo hasta el primer evento de interés y el tiempo desde el primer evento hasta el segundo, respectivamente.

Estos tiempos sucesivos (X_1, X_2) van a encontrarse censurados por la derecha por una variable C , que suponemos independiente del par de tiempos.

Debido a dicha censura, en vez de los tiempos de interés X_1 y X_2 , observamos la tupla $(T_1, T_2, \delta_1, \delta_2)$, donde $T_1 = \min\{X_1, C\}$, $\delta_1 = \mathbb{I}(X_1 \leq C)$, $T_2 = \min\{X_2, C_2\}$, $\delta_2 = \mathbb{I}(X_2 \leq C_2)$, y donde $C_2 = (C - X_1)\mathbb{I}(X_1 \leq C)$ es la variable de censura del segundo tiempo. Es importante tener presente que, dado que X_1 y X_2 no se suponen independientes, el segundo tiempo, X_2 , no va a ser, en general, independiente de la variable que le censura, C_2 . Esto supone una dificultad a la hora de estimar tanto la distribución conjunta del par de tiempos sucesivos como la marginal del segundo tiempo.

El problema de la estimación de la distribución conjunta de dos tiempos sucesivos censurados ha sido estudiado por numerosos autores, y en este trabajo se ha utilizado como referencia el estimador no paramétrico propuesto por de Uña-Álvarez y Meira-Machado (2008).

Siguiendo en el contexto no paramétrico, en la memoria se presentan dos planes de remuestreo: el bootstrap simple y el obvio. Al contrario de lo que ocurre en el caso univariante, estos métodos no resultan ser equivalentes en el contexto multivariante de este trabajo. Para analizar el comportamiento de dichos métodos bootstrap se realiza un estudio de simulación.

Para modelar los tiempos de eventos sucesivos censurados se puede emplear también el enfoque semiparamétrico propuesto por Lawless y Yilmaz (2011), que utilizan las funciones cópula para modelar la asociación entre los tiempos de interés.

Una cópula $\mathcal{C}(u, v)$ es una función de distribución bivalente con distribuciones marginales uniformes en el intervalo $(0, 1)$ (Nelsen 2006). Relacionado con estas funciones, el Teorema de Sklar (1959) establece que cualquier función de distribución bivalente $H(x, y)$ se puede

escribir en función de sus marginales y una función cópula. Además, si dichas marginales son continuas, la cópula es única.

Las funciones cópula son por este hecho muy útiles a la hora de modelizar la dependencia entre variables aleatorias. En este trabajo son utilizadas para modelizar la asociación de dos tiempos de eventos sucesivos censurados, para los cuales es posible estimar el modelo cópula, es decir, es posible estimar tanto la función cópula como las marginales de los tiempos sucesivos.

En primer lugar, se supone que conocemos completamente la cópula que liga los dos tiempos de interés, y se estudian algunas propiedades de los estimadores de las marginales, como la propiedad de que tan solo tienen masa en los tiempos no censurados.

A continuación, se estudia un contexto más general, en el la cópula no está completamente especificada, sino que se conoce únicamente la familia paramétrica a la que pertenece. Este es el escenario que estudian Lawless y Yilmaz (2011) y que en este trabajo se revisa. Además, en relación con los estimadores propuestos por estos autores, se aportan simulaciones complementarias a las realizadas en dicho artículo.

También se analiza el bootstrap semiparamétrico que proponen Lawless y Yilmaz (2011), que es un plan de remuestreo similar al bootstrap obvio nombrado anteriormente. Se han realizado simulaciones que confirman los resultados aportados por dichos autores, y que arrojan además nuevo conocimiento sobre el citado contexto semiparamétrico.

Los métodos expuestos en el trabajo, como son la estimación basada en cópulas de las distribuciones marginales de los tiempos sucesivos y el parámetro de la cópula o el bootstrap semiparamétrico, se aplican a un conjunto de datos médicos reales.

El trabajo concluye con una introducción al problema de la selección de la cópula. Este problema tiene un gran interés debido al sesgo que aparece en las estimaciones cuando no se especifica bien la familia de cópulas para los tiempos sucesivos. Esto motiva el estudio de un contraste que permita saber si la elección de la familia paramétrica es la correcta. Se propone un posible estadístico para dicho contraste.

Referencias

1. de Uña-Álvarez J, Meira-Machado LF (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics & Probability Letters*, 78:2440-2445.
2. Lawless JF, Yilmaz YE (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical Journal*, 53:779-796.
3. Nelsen RB (2006). *An Introduction to Copulas*. Springer.
4. Sklar A (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8:229-231.