Vol. 15 (núm 1) pp. 130 - 140
Enero - Junio 2024
https://doi.org/10.19053/01217488.v15.n1.2024.15586

Comparación de algunas estimaciones del τ de Kendall para datos bivariados con censura a intervalo

Comparison of some estimations of Kendall's τ for interval-censored bivariate data

Jessica K. Serna-Morales[1*], Mario C. Jaramillo-Elorza[1**], Carlos M. Lopera-Gómez[1***].

Resumen

Los datos de falla bivariados son comunes en estudios de confiabilidad y supervivencia, donde la estimación de la fuerza de dependencia es a menudo un paso importante en el análisis de los datos. En la literatura, se ha establecido que los coeficientes de correlación miden la relación lineal entre dos variables, pero también pueden existir relaciones no lineales fuertes entre ellas. El coeficiente de concordancia τ de Kendall se ha convertido en una herramienta útil para el análisis de datos bivariados, la cual es usada en pruebas no paramétricas de independencia y como una medida complementaria de asociación. En el análisis de datos de confiabilidad, hay un fenómeno que ocurre cuando el valor de las observaciones se conoce parcialmente, lo cual se conoce como censura. En este trabajo, se comparan vía simulación dos métodos de estimación del τ de Kendall, una de ellas suponiendo normalidad en las distribuciones marginales y ajustándolas individualmente, y la otra basada en cópulas (Gaussiana y Clayton), donde los datos bivariados están censurados a intervalo. La comparación se hace mediante el error cuadrático medio y la mediana de la desviación absoluta. Los resultados muestran que el método basado en la aproximación cópula produce en general estimaciones más precisas que el método de ajuste individual de las marginales.

Palabras Clave: Cópula, medidas de asociación, modelo de mezcla Gaussiana, supervivencia

Abstract

Bivariate failure data are common in reliability and survival studies, where estimation of dependency is often an important step in data analysis. In the literature, it known that the correlation coefficients measure the linear relationship between two variables, but strong non-linear relationship can also exist between them. Kendall's τ concordance coefficient has become a useful tool for bivariate data analysis, which is used in nonparametric tests of independence and as a complementary measures of association. In the analysis of reliability data, there is a phenomenon that occurs when the value of the lifetime is partially known, which is known as censoring. In this paper, two estimation methods of Kendall's τ are compared via simulation, one of them assuming normality in marginal distributions and adjusting them individually and the other based on copulas (Gaussian and Clayton), where the bivariate data are interval censored. The comparison is made using the mean squared error and the median absolute deviation. The results show that the method based on the copula approximation generally produces more precise estimates than the method of individual adjustment of the marginals. ***Keywords:*** Association measures, copula, Gaussian mixture model, survival

Recepción: 01-Diciembre-2022
Aceptación: 06-Marzo-2023

[1,*] Magíster en Ciencias-Estadística, Universidad Nacional de Colombia, Medellín, Colombia. Dirección electrónica: jksernam@unal.edu.co
[1,**] Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia. Dirección electrónica: mcjarami@unal.edu.co
[1,***] Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia. Dirección electrónica: cmlopera@unal.edu.co

1. Introducción

El análisis de datos de confiabilidad proporciona a los consumidores una medida asociada con la duración promedio de un producto de interés, y a su vez los fabricantes cuentan con una medida que les indica qué tan bueno es su producto. Una característica típica de los datos en confiabilidad es la presencia de censura, la cual se clasifica en: censura a derecha, censura a izquierda o censura a intervalo [1]. Este trabajo, se enfoca en la censura a intervalo, que se produce cuando el tiempo de falla de un producto es desconocido pero se sabe que ocurrió en un intervalo de tiempo que por lo general es aleatorio.

En el caso de muestras aleatorias bivariadas se debe tener en cuenta la estructura de dependencia asociada, de tal manera que en el análisis de datos se logre identificar dicha estructura, para medirla usualmente se utiliza el τ de Kendall, propuesto en [2]. El coeficiente de concordancia τ de Kendall, mide el grado de dependencia entre dos variables aleatorias, cuya escala de medida es ordinal o de intervalo; la estimación del τ de Kendall se basa en el orden de las observaciones. La propiedad más importante del τ de Kendall es que es invariante a transformaciones monótonas. Una extensión para estimar el coeficiente de concordancia τ de Kendall, con censura a intervalo, es propuesta por [3].

Para estimar el τ de Kendall en datos bivariados con censura a intervalo, [4] proponen una estimación basada en máxima verosimilitud que en lugar de usar los datos censurados de forma múltiple directamente, estima la covarianza de los datos censurados individualmente. Alternativamente, [5] proponen modelar las distribuciones marginales con un modelo de falla acelerado con un término de error flexible sugerido por [6], la asociación se modela de forma paramétrica y se utilizan las cópulas Gaussiana y Clayton para datos bivariados.

En este artículo, se comparan a través de un estudio de simulación los métodos anteriormente descritos para estimar el coeficiente de concordancia τ de Kendall para datos bivariados con censura a intervalo. Tal comparación se realiza mediante dos medidas de calidad de los estimadores que son: el error cuadrático medio y la mediana de la desviación absoluta.

En la Sección 2, se presentan algunos conceptos importantes para el desarrollo de este trabajo, tales como: función de supervivencia, tipos de censura, estimación de la función de supervivencia para datos con censura a intervalo y medidas de asociación. La Sección 3 presenta dos métodos de estimación del τ de Kendall: la primera es la estimación suponiendo normalidad en las marginales y ajustándolas individualmente, y la segunda es la estimación por medio de cópulas, en particular se usan las cópulas Gaussiana y Clayton. En la Sección 4, se realiza un estudio de simulación, en donde se comparan los resultados analizados de los dos métodos mediante el error cuadrático medio y la mediana de la desviación absoluta. Finalmente, la Sección 5 presenta algunas conclusiones de la investigación.

2. Conceptos básicos

A continuación se presentan algunos conceptos importantes relacionados con la función de supervivencia, los tipos de censura, la estimación de la función de supervivencia para datos con censura a intervalo, medidas de asociación, cópulas, el modelo de mezcla gaussiano penalizado y el modelo de tiempo de falla acelerado.

2.1. Función de supervivencia

La función de supervivencia es el complemento de la función de distribución acumulada, da la probabilidad de sobrevivir más allá del tiempo t. Sea T una variable aleatoria continua no negativa, que representa el tiempo de supervivencia de un individuo de una población, la función de supervivencia $S(t)$ se define como:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^{\infty} f(x)\,dx, \qquad (1)$$

donde $F(t)$ es la función de distribución acumulada (f.d.a) y $f(x)$ es la función de densidad de probabilidad (f.d.p).

Sean T_1 y T_2 dos variables aleatorias continuas no negativas, la función de supervivencia conjunta es:

$$S(t_1,t_2) = P(T_1 > t_1, T_2 > t_2),\ t_1,t_2 \geq 0. \qquad (2)$$

Es decir $S(t_1,t_2)$ es la probabilidad de que ambas unidades sobrevivan a los tiempos t_1 y t_2, respectivamente.

2.2. Tipos de censura

Una característica típica de los datos de supervivencia es el hecho de que el tiempo hasta un evento no siempre se observa exactamente y las observaciones están sujetas a censura. Las pruebas de vida a menudo usan datos con censura, ya sea a izquierda, a derecha o a intervalo.

Siguiendo la notación en [7], donde los individuos acuden periódicamente a visitas programadas donde se verifica si un evento ha ocurrido o no, para los individuos cuyos tiempos están censurados a izquierda, el evento de interés ha ocurrido antes de la primera visita, para los individuos cuyos tiempos están censurados a la derecha, a menudo, el estudio termina antes de que todos los sujetos que hacen parte de

éste hayan mostrado el evento de interés, debido a que el sujeto abandona el estudio antes de experimentar el evento, por tanto el evento no ha ocurrido hasta la última visita del sujeto. En la censura a intervalo, la falla se encuentra entre dos visitas, pero no se sabe en qué momento exactamente ocurrió la falla.

La censura se puede clasificar en 3 tipos que son tipo I, tipo II y aleatoria. Los datos con censura tipo I, resultan cuando todas las unidades que no han fallado antes de un tiempo pre-especificado t_c, se censuran en el tiempo t_c (censura al tiempo). Los datos con censura tipo II resultan cuando una prueba es terminada después de un número pre-especificado r de fallas, $2 \leq r \leq n$ (censura a la falla). Cuando $r = n$, todas las unidades fallan y esto se conoce como datos completos. La censura aleatoria se refiere a los individuos que dejan de asistir al estudio por otros motivos que no están relacionados con el estudio, este tipo de censura está sujeta al azar.

En la práctica, cuando se realiza un determinado estudio, es importante distinguir cuándo los datos están censurados a derecha, a izquierda o a intervalo. Siguiendo a [7], se utiliza la notación $\lfloor l,u \rfloor$ para señalar un intervalo abierto, semi-abierto o cerrado con límite inferior l y límite superior u, acompañado de un indicador de censura δ igual a $0,1,2$ ó 3 para denotar censura a derecha, censura a izquierda, censura a intervalo o falla exacta, respectivamente, como se ilustra en la siguiente tabla:

Tabla 1: Representación de los tipos de censura.

Observación	Intervalo $\lfloor l,u \rfloor$	δ
Censura a derecha en l	$0 < l < u = \infty$	0
Censura a izquierda en u	$0 = l < u < \infty$	2
Censura a intervalo	$0 < l < u < \infty$	3
Falla exacta en t	$0 < l = t = u < \infty$	1

2.3. Medida de asociación τ de Kendall

Las medidas de asociación globales más conocidas son el coeficiente de correlación de Pearson y el coeficiente de concordancia τ de Kendall.

La medida de correlación más utilizada es el coeficiente de correlación de Pearson, fue definido originalmente para variables aleatorias que tienen distribución normal bivariada. El coeficiente de correlación de Pearson mide la fuerza de asociación lineal entre dos variables aleatorias, esta medida no es muy atractiva para modelar distribuciones de supervivencia bivariadas. En ese caso, se define otra medida de asociación como lo es el coeficiente τ de Kendall.

El τ de Kendall es una medida de dependencia que representa el grado de concordancia entre dos variables. Para variables aleatorias continuas, sean (X_1,Y_1) y (X_2,Y_2) vectores aleatorios independientes e idénticamente distribuidos, cada uno con función de distribución conjunta $H(x,y)$, entonces el τ de Kendall es dado por la probabilidad de concordancia menos la probabilidad de discordancia [8], dada por:

$$\begin{aligned}\tau &= \tau(X,Y) \\ &= P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right].\end{aligned} \tag{3}$$

La estimación muestral del τ de Kendall definida en [9] en términos de concordancia y discordancia, se presenta a continuación:

Sea $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ una muestra aleatoria de n observaciones de un vector (X,Y) de variables aleatorias continuas ó al menos ordinales. Un par de observaciones (x_i,y_i) y (x_j,y_j) son concordantes si $x_i < x_j$ y $y_i < y_j$, ó si $x_i > x_j$ y $y_i > y_j$ y un par de observaciones (x_i,y_i) y (x_j,y_j) son discordantes si $x_i < x_j$ y $y_i > y_j$, ó si $x_i > x_j$ y $y_i < y_j$.

Existen $\binom{n}{2}$ pares diferentes (x_i,y_i) y (x_j,y_j) de observaciones en la muestra, y cada par es concordante o discordante. Sea N_c el número de pares concordantes y N_d el número de pares discordantes. Entonces la estimación del τ de Kendall para la muestra se define como:

$$\widehat{\tau} = \frac{N_c - N_d}{N_c + N_d}. \tag{4}$$

La principal ventaja del τ de Kendall es que su distribución se aproxima a la distribución normal rápidamente, cuando el tamaño de la muestra es grande.

2.3.1. Propiedad de invarianza del τ de Kendall

Sean (X_1,Y_1) y (X_2,Y_2) dos variables aleatorias bivariadas independientes, cada una con la distribución bivariada común de $H(x,y)$, y sean g_1 y g_2 dos funciones reales monótonas (crecientes ó decrecientes), entonces $\tau[g_1(X),g_2(Y)] = \tau(X,Y)$. La demostración de este resultado se puede ver en [10].

2.3.2. Otras propiedades del τ de Kendall

Si se asume que las distribuciones marginales son continuas, se tienen los siguientes resultados para el coeficiente de concordancia τ de Kendall.

1. $-1 \leq \tau \leq 1$
2. $\tau = 1$ ó $(\tau = -1)$ si y sólo si $Y = g(X)$, para alguna función g monótona creciente ó (decreciente).
3. $\tau = 0$, si X y Y son independientes.

2.4. Cópulas

Una cópula es una función multivariada que describe la asociación entre las variables de una distribución conjunta. En [7] se considera que la distribución marginal describe la forma en que una variable aleatoria actúa por sí sola, mientras que la función cópula describe cómo se unen para determinar la distribución multivariada. Las cópulas extraen la estructura de dependencia de la función de distribución conjunta y por lo tanto, separan la estructura de dependencia de las funciones de distribución marginal; además se han convertido en una herramienta importante en varios campos, como en medicina, ingeniería, economía entre otras áreas.

[8] considera un par de variables aleatorias X y Y con distribuciones marginales $F(x) = P[X \leq x]$ y $G(y) = P[Y \leq y]$, respectivamente, y con distribución conjunta $H(x,y) = P[X \leq x, Y \leq y]$.

Se define la cópula C_α, con α el parámetro de la cópula, como una función que le asigna al par $(F(x), G(y))$ un número real $H(x,y)$ en el intervalo $[0,1]$, es decir:

$$\begin{aligned} C_\alpha : [0,1] \times [0,1] &\to [0,1] \\ (F(x), G(y)) &\to H(x,y) \end{aligned}$$

Las cópulas tienen las siguientes propiedades:

1. $C_\alpha(a,0) = 0 = C_\alpha(0,b)$ para todo $a,b \in [0,1]$.
2. $C_\alpha(a,1) = a$ y $C_\alpha(1,b) = b$ para todo $a,b \in [0,1]$.
3. Para todo $(a_1,b_1),(a_2,b_2) \in [0,1] \times [0,1]$ con $a_1 \leq a_2$ y $b_1 \leq b_2$ se tiene que:

$$C_\alpha(a_2,b_2) - C_\alpha(a_1,b_2) - C_\alpha(a_2,b_1) + C_\alpha(a_1,b_1) \geq 0.$$

Siguiendo a [7]. Sean (T_1,T_2) el par de tiempos de supervivencia que se encuentran en el rectángulo $\lfloor l_1,u_1 \rfloor \times \lfloor l_2,u_2 \rfloor$ con $0 \leq l_j < u_j \leq \infty$, para $j = 1,2$. Los indicadores de censura a izquierda y a intervalo para T_j $(j = 1,2)$ se denotan como $\delta_j^{(1)}$ y $\delta_j^{(2)}$, los cuales producen el vector $\boldsymbol{\delta} = (\delta_1^{(1)}, \delta_1^{(2)}, \delta_2^{(1)}, \delta_2^{(2)})$. Se denota por $\mathbf{y} = (l_1,u_1,l_2,u_2)'$ los datos censurados a intervalo y asumiendo que (T_1,T_2) son independientes de (L_1,U_1,L_2,U_2), pero (L_1,U_1) y (L_2,U_2) pueden ser dependientes.

Denote por $\mathbf{D}$ la muestra de tamaño n de variables aleatorias i.i.d de intervalos bivariados $\lfloor l_{1i},u_{1i} \rfloor \times \lfloor l_{2i},u_{2i} \rfloor (i = 1,\ldots,n)$.

Las cópulas también pueden ser definidas en términos de la función de supervivencia S y se denominan cópulas de supervivencia. Sean T_1 y T_2 dos variables aleatorias continuas no negativas, con funciones marginales de supervivencia $S_1(t)$ y $S_2(t)$ respectivamente y la función de supervivencia conjunta $S(t_1,t_2) = P[T_1 > t_1, T_2 > t_2]$, la cópula de supervivencia está dada por

$$S(t_1,t_2) = \breve{C}_\alpha(S_1(t_1), S_2(t_2)), \tag{5}$$

donde $\breve{C}_\alpha$ es una cópula de supervivencia específica con parámetro α, el cual regula la asociación entre T_1 y T_2.

La cópula C_α de una función de distribución acumulada H y su cópula de supervivencia $\breve{C}_\alpha$ están relacionadas de la siguiente manera: $\breve{C}_\alpha(a,b) = a + b + C_\alpha(1-a, 1-b) - 1$.

Cuando se trabaja con cópulas una decisión importante, es la de elegir la cópula adecuada para modelar los datos [11], en donde el mayor interés se encuentra en la dependencia de las variables aleatorias. Existe una gran variedad de cópulas, entre ellas se encuentran las cópulas Gaussiana y Clayton, que han sido utilizadas en [9].

Cuando se supone un modelo con enfoque paramétrico para datos que presentan censura a intervalo, hay dificultad para elegir la distribución correcta. La solución a este problema se puede encontrar con el uso de una cópula. La medida de asociación τ de Kendall se puede expresar como una función de una cópula de supervivencia de la siguiente manera:

$$\begin{aligned} \tau &= 4 \int_0^\infty \int_0^\infty F(t_1,t_2) dH(t_1,t_2) - 1 \\ &= 4 \int_0^1 \int_0^1 \breve{C}(u,v) d\breve{C}(u,v) - 1. \end{aligned}$$

Para varias cópulas, se puede establecer la relación entre el parámetro de la cópula y las medidas de asociación ρ de Pearson y τ de Kendall. Se considerarán las cópulas Gaussiana y Clayton. Adicionalmente, se definirá la cópula Gumbel que es utilizada en este trabajo para la generación de las bases de datos en el estudio de simulación.

2.4.1. Cópula Clayton

Para $\theta_L > 0$, con $\theta_L \neq 1$ y $\alpha = \theta_L$ el parámetro de la cópula Clayton, la cual se define como [7]:

$$\breve{C}^C_{\theta_L}(u,v) = (u^{1-\theta_L} + v^{1-\theta_L} - 1)^{\frac{1}{1-\theta_L}}. \tag{6}$$

También se conoce como la familia Pareto de cópulas.

El modelo Clayton asume una "constant local cross ratio function", la cual evalúa el grado de dependencia en un solo punto de tiempo y se define como [5]:

$$\theta_L(t_1,t_2) = S(t_1,t_2) \begin{pmatrix} \frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} \\ \frac{\partial S(t_1,t_2)}{\partial t_1} \quad \frac{\partial S(t_1,t_2)}{\partial t_2} \end{pmatrix} \tag{7}$$

La "local cross ratio function" tiene una interpretación muy natural en las tasas de riesgo condicionales como en [12], a saber:

$$\theta_L(t_1,t_2) = \frac{\lambda_1(t_1|T_2=t_2)}{\lambda_1(t_1|T_2\geq t_2)} = \frac{\lambda_2(t_2|T_1=t_1)}{\lambda_2(t_2|T_1\geq t_1)},$$

donde λ_1 y λ_2 son las funciones de riesgo para T_1 y T_2 respectivamente.

La independencia corresponde a un valor $\theta_L = 1$, la dependencia positiva $\theta_L > 1$ y la dependencia negativa $\theta_L < 1$.

2.4.2. Cópula Gaussiana

Los datos bivariados que se distribuyen normal producen la cópula Gaussiana, con $\alpha = \rho$ el parámetro de la cópula, que en este caso corresponde al coeficiente de correlación lineal de Pearson. La expresión de la cópula Gaussiana está dada por [5]:

$$\check{C}_\rho^G(u,v) = \Phi_\rho[\Phi^{-1}(u),\Phi^{-1}(v)], \tag{8}$$

donde Φ_ρ denota la función de distribución normal bivariada estándar con correlación ρ. La cópula Gaussiana no tiene una forma cerrada simple, pero puede expresarse como una integral sobre la densidad de (U,V). En dos dimensiones para $|\rho| < 1$ se tiene que:

$$\check{C}_\rho^C(u,v) = \int_{-\infty}^{a}\int_{-\infty}^{b} \frac{\exp\left\{-\frac{(s_1^2-2\rho s_1 s_2+s_2^2)}{2(1-\rho^2)}\right\}}{2\pi(1-\rho^2)^{1/2}} ds_1 ds_2, \tag{9}$$

donde $a = \Phi^{-1}(u)$ y $b = \Phi^{-1}(v)$

La cópula Gaussiana, puede ser considerada como una estructura de dependencia que interpola entre la dependencia positiva perfecta y la dependencia negativa, donde el parámetro ρ representa la fuerza de la dependencia.

2.4.3. Cópula Gumbel

La función de confiabilidad bivariada perteneciente a la familia Gumbel tiene la siguiente forma [13]:

$$C_\alpha(u,v) = \exp\{-[(-\ln u)^{1/\alpha} + (-\ln v)^{1/\alpha}]\}^\alpha, \tag{10}$$

donde $0 < \alpha < 1$.

Sean T_1 y T_2, tiempos de falla Weibull. Una función de confiabilidad conjunta para la distribución Weibull bivariada definida en [14] es:

$$S(t_1,t_2) = \exp\left\{-\left[\left(\frac{t_1}{\theta_1}\right)^{\frac{\beta_1}{\alpha}} + \left(\frac{t_2}{\theta_2}\right)^{\frac{\beta_2}{\alpha}}\right]^\alpha\right\}, \tag{11}$$

donde, $\beta_1, \theta_1, \beta_2, \theta_2$ son los parámetros de forma y escala asociados a los tiempos T_1 y T_2, respectivamente, los cuales son positivos. $0 < \alpha \leq 1$ es el parámetro de dependencia entre T_1 y T_2, donde las distribuciones marginales están dadas por:

$$S_k(t) = \exp\left\{-\left(\frac{t}{\theta_k}\right)^{\beta_k}\right\},\ k = 1,2,\ t > 0 \tag{12}$$

Cuando $\alpha = 1$ se cumple que hay independencia entre los tiempos de falla Weibull T_1 y T_2.

Si se compara la ecuación (11) con la ecuación (10), la representación de la distribución Weibull bivariada se obtiene mediante la cópula Gumbel, para $0 < \alpha < 1$, es decir, cuando T_1 y T_2 no son independientes.

2.4.4. Relación del parámetro de la cópula con las medidas de asociación ρ de Pearson y τ de Kendall

Para las diferentes cópulas descritas en la Sección 2.4, se puede establecer una relación explícita entre el parámetro de la cópula y una medida de asociación.

Siguiendo a [7], para la cópula Clayton, el τ de Kendall y el parámetro θ_L se relacionan como $\tau = \theta_L/(\theta_L+2)$. La cópula Clayton solo permite modelar correlaciones positivas.

Para la cópula Gaussiana, la correlación ρ de Pearson es el parámetro de la cópula. El coeficiente τ de Kendall está dado por: $\tau = (2/\pi)\cdot\arcsin(\rho)$.

2.5. Modelo de mezcla gaussiano penalizado

En este modelo se presenta la estimación de las densidades de los errores para el método basado en cópulas que se desarrollará en la Sección 3.2.

Para conjuntos de datos bivariados, siguiendo la notación de [7], $g(y_1,y_2)$ representa la densidad conjunta de $(Y_1,Y_2)'$, $Y_d = \log(T_d)$ $(d = 1,2)$ con $g_1(y_1)$ y $g_2(y_2)$ las densidades marginales. Para una muestra de tamaño n, una aproximación suave de esta densidad puede obtenerse de una suma ponderada penalizada de densidades normales bivariadas no correlacionadas ubicadas en una rejilla predefinida, es decir, que la densidad es diferenciable. El método se basa en el procedimiento de suavizado penalizado descrito en [15].

Con los puntos de la rejilla preespecificados $\boldsymbol{\mu}_{k_1,k_2} = (\mu_{1,k_1},\mu_{2,k_2})'$ $(k_1 = 1,\ldots,K_1;\ k_2 = 1,\ldots,K_2)$, se asume que:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} w_{k_1,k_2}\ \mathcal{N}_2(\boldsymbol{\mu}_{k_1,k_2},\boldsymbol{\Sigma}), \tag{13}$$

donde $\mathcal{N}_2(.)$ denota la distribución normal bivariada y

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

es una matriz de covarianza diagonal con valores preespecificados de σ_1^2 y σ_2^2. Además, $w_{k_1,k_2} > 0$ para todo k_1, k_2 y $\sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} w_{k_1,k_2} = 1$.

La idea de este método es estimar los pesos w_{k_1,k_2} $(k_1 = 1,\ldots,\ K_1; k_2 = 1,\ldots,K_2)$ maximizando la verosimilitud, de forma que los puntos de la rejilla permanezcan fijos. Los estimadores de máxima verosimilitud sin restricciones se pueden obtener expresando el log de la verosimilitud como una función de $\mathbf{a} = (a_{k_1,k_2} : k_1 = 1,\ldots,K_1;\ k_2 = 1,\ldots,K_2)^{'}$ usando la siguiente igualdad:

$$\begin{aligned} w_{k_1,k_2} &= w_{k_1,k_2}(\mathbf{a}) \\ &= \frac{\exp(a_{k_1,k_2})}{\sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} \exp(a_{j_1,j_2})}. \end{aligned}$$

Un término de penalización que involucra las diferencias de los a-coeficientes (ver [15]), está dado por:

$$q(\mathbf{a};\boldsymbol{\lambda}) = \frac{\lambda_1}{2} \sum_{k_1=1}^{K_1} \sum_{k_2=1+s}^{K_2} (\Delta_1^s a_{k_1,k_2})^2 + \frac{\lambda_2}{2} \sum_{k_1=1}^{K_1} \sum_{k_2=1+s}^{K_2} (\Delta_2^s a_{k_1,k_2})^2, \tag{14}$$

el cual evita un sobreajuste. En la ecuación (14) el vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^{'}$ con $\lambda_1, \lambda_2 > 0$ son los parámetros de penalización ó suavizado. Δ_d^s es el s-ésimo operador de diferencia en la d-ésima dimensión $(d = 1, 2)$, definido iterativamente (para la primera dimensión) como $\Delta_1^s a_{k_1,k_2} = \Delta_1^{s-1} a_{k_1,k_2} - \Delta_1^{s-1} a_{k_1,k_2-1}$ para $s > 0$ y $\Delta_1^0 a_{k_1,k_2} = a_{k_1,k_2}$. Dados λ_1 y λ_2, el log de la verosimilitud penalizada $l_P(\mathbf{a};\boldsymbol{\lambda}) = l(\mathbf{a}) - q(\mathbf{a};\boldsymbol{\lambda})$ es maximizado con respecto a $\mathbf{a}$, produciendo estimaciones $\widehat{a}_{k_1,k_2}$ $(k_1 = 1,\ldots,K_1;\ k_2 = 1,\ldots,K_2)$. Los valores óptimos de λ_1 y λ_2 se pueden obtener minimizando el criterio de Akaike (AIC). Este procedimiento proporciona un enfoque paramétrico que produce una solución suave, ver [16].

2.6. Modelo de tiempo de falla acelerado

Es usado para modelar las distribuciones marginales de una cópula de supervivencia, ya sea la Gaussiana o la Clayton, a través del paquete de *R* **smoothSurv**. Más adelante en la Sección 3.2, se mostrará que las distribuciones marginales de la cópula Gaussiana siguen un modelo tiempo de falla acelerado, por lo que es necesario mostrar algunos aspectos teóricos de este modelo.

Siguiendo la notación en [7] el modelo de tiempo de falla acelerado (AFT) está dado por:

$$\log(T) = \boldsymbol{X}^{'}\boldsymbol{\beta} + \varepsilon, \tag{15}$$

con T el tiempo de supervivencia, $\boldsymbol{X}$ un vector de covariables, $\boldsymbol{\beta}$ el vector de parámetros de regresión y ε una variable aleatoria del error. Sean h_0 y S_0 la función de riesgo y supervivencia base, respectivamente, de la variable aleatoria $T_0 = \exp(\varepsilon)$. Para un sujeto con vector de covariables $\boldsymbol{X}$, se asume que la función de riesgo y supervivencia son:

$$h(t|\boldsymbol{X}) = h_0[\exp(-\boldsymbol{X}^{'}\boldsymbol{\beta})t]\exp(-\boldsymbol{X}^{'}\boldsymbol{\beta}), \tag{16}$$

y

$$S(t|\boldsymbol{X}) = S_0[\exp(-\boldsymbol{X}^{'}\boldsymbol{\beta})t]. \tag{17}$$

Por tanto,

$$T = \exp(\boldsymbol{X}^{'}\boldsymbol{\beta})T_0, \tag{18}$$

es decir, el efecto de una covariable implica una aceleración o desaceleración en comparación con el tiempo de evento base.

En [7] enfatizan que en la práctica se utiliza con frecuencia una forma totalmente paramétrica del modelo AFT, es decir, se supone que el término de error ε tiene una densidad específica $g(\varepsilon)$. Los supuestos más comunes para $g(\varepsilon)$ son la densidad normal, la densidad logística y la densidad de Gumbel (valor extremo). Por lo tanto, el modelo AFT paramétrico asume una forma paramétrica para los efectos de las variables explicativas y también asume una forma paramétrica para la función de supervivencia subyacente. Luego, la estimación se realiza mediante una maximización estándar del log de la verosimilitud.

Cuando sólo se utiliza una covariable categórica en el modelo, la curva de supervivencia de Kaplan-Meier se puede calcular para los sujetos de cada categoría por separado. Las curvas de supervivencia de Kaplan-Meier se pueden superponer con las curvas de supervivencia paramétrica ajustadas para los grupos específicos. Cuando se usan covariables continuas o muchas covariables, los sujetos podrían dividirse en un cierto número de grupos (por ejemplo, 3, referidos a pacientes de riesgo bajo, medio y alto) según la puntuación de riesgo $\boldsymbol{X}^{'}\boldsymbol{\beta}$. La comparación de las curvas de supervivencia de Kaplan-Meier con las curvas de supervivencia ajustadas en cada grupo proporciona nuevamente una indicación de bondad de ajuste.

De las ecuaciones (15) y (18) se puede ver que el modelo AFT es de hecho un modelo de regresión lineal estándar con un tiempo de supervivencia logarítmico transformado.

3. Métodos de estimación del τ de Kendall para datos bivariados con censura a intervalo

En esta Sección se presentan los dos métodos de estimación, que se usan en este trabajo, para estimar la medida de

asociación τ de Kendall.

3.1. Estimación bajo el supuesto de normalidad bivariada

El enfoque de estimación del coeficiente τ de Kendall que se va a describir a continuación, fue propuesto por [4].

Sean (X,Y) el par de tiempos censurados que se encuentran en el rectángulo $\lfloor l_1,u_1\rfloor \times \lfloor l_2,u_2\rfloor$ con $0 \leq l_j < u_j \leq \infty$, para $j = 1,2$. Se asume que (X,Y) son independientes de las variables de censura (L_1,U_1,L_2,U_2), pero (L_1,U_1) y (L_2,U_2) pueden ser dependientes.

Además, suponga que las variables aleatorias (X,Y) siguen una distribución normal bivariada, con vector de medias $\boldsymbol{\mu} = (\mu_X,\mu_Y)'$ y matriz de varianzas y covarianzas

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix},$$

con $\sigma_{XY} = \rho\ \sigma_X\sigma_Y$.

Bajo estas condiciones, la estimación de máxima verosimilitud del vector de parámetros $\boldsymbol{\theta} = (\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho)$ se basa en el log de verosimilitud dado por:

$$\begin{aligned} l(\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho) &= l(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} G_i, \end{aligned} \tag{19}$$

donde,

- $G_i = \log[F(u_{1i},u_{2i})]$ si X, Y están censurados a izquierda.
- $G_i = \log[F(u_{1i},u_{2i}) - F(l_{1i},u_{2i})]$ si X está censurado a intervalo y Y a izquierda.
- $G_i = \log[F_Y(u_{2i}) - F(l_{1i},u_{2i})]$ si X está censurado a derecha y Y a izquierda.
- $G_i = \log[F(u_{1i},u_{2i}) - F(u_{1i},l_{2i})]$ si X está censurado a izquierda y Y a intervalo.
- $G_i = \log[F(u_{1i},u_{2i}) - F(l_{1i},u_{2i}) - F(u_{1i},l_{2i}) + F(l_{1i},l_{2i})]$ si X,Y están censurados a intervalo.
- $G_i = \log[F_Y(u_{2i}) - F(l_{1i},u_{2i}) - F_Y(l_{2i}) + F(l_{1i},l_{2i})]$ si X está censurado a derecha y Y a intervalo.
- $G_i = \log[F_X(u_{1i}) - F(u_{1i},l_{2i})]$ si X está censurado a izquierda y Y a derecha.
- $G_i = \log[F_X(u_{1i}) - F_X(l_{1i}) - F(u_{1i},l_{2i}) + F(l_{1i},l_{2i})]$ si X está censurado a intervalo y Y a derecha.
- $G_i = \log[1 - F_X(l_{1i}) - F_Y(l_{2i}) + F(l_{1i},l_{2i})]$ si X,Y están censurados a derecha.

donde, F es la función de distribución conjunta acumulada de (X,Y), F_X es la función de distribución marginal acumulada de la variable aletoria X y F_Y es la función de distribución marginal acumulada de la variable aletoria Y.

Siguiendo a [4], maximizar esta verosimilitud con respecto a los 5 parámetros es difícil, por lo que se sugiere estimar los parámetros $\mu_X,\mu_Y,\sigma_X,\sigma_Y$ por separado para cada una de las distribuciones marginales, las cuales presentan censura a derecha, a izquierda y a intervalo, bajo los supuestos $X \sim N(\mu_X,\sigma_X^2)$ y $Y \sim N(\mu_Y,\sigma_Y^2)$. Cuando se tienen las estimaciones para $\mu_X,\mu_Y,\sigma_X,\sigma_Y$ con la verosimilitud en (19) se estima ρ, y luego, usando la relación de Greiner dada en [17], $\tau = (2/\pi)\arcsin(\rho)$, se estima el τ de Kendall.

3.2. Método cópula

El método cópula consiste en seleccionar una cópula de supervivencia, ya sea la cópula Gaussiana o Clayton, luego se ajustan las distribuciones marginales, las cuales se modelan con el modelo de tiempo de falla acelerado con un término de error flexible, como se describe en la Sección 2.6. Con las distribuciones marginales ajustadas se procede a estimar el parámetro de la cópula, el cual está relacionado con el τ de Kendall.

En esta parte de la estimación se describe el enfoque empleado en [5] que permite que el parámetro de dependencia α de la respectiva cópula dependa de las covariables. Para la cópula Clayton el parámetro de dependencia θ_L se modela en la escala logarítmica, es decir, $\log(\alpha_i) = \boldsymbol{\gamma}'\boldsymbol{x}_i$, con $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_p)'$ un vector p-dimensional de paramétros de la regresión desconocidos y con $\boldsymbol{x}_i = (x_{1,i},\ldots,x_{p,i})'$ un vector p-dimensional de covariables asociadas sobre el i-ésimo sujeto. Para la cópula Gaussiana, la dependencia se modela con la transformación de Fisher, es decir, $\frac{1}{2}\log[(1+\alpha_i)/(1-\alpha_i)] = \boldsymbol{\gamma}'\boldsymbol{x}_i$. Las distribuciones marginales también pueden depender de las covariables, que pueden ser diferentes al parámetro de la cópula. El vector de dimensión m, $\boldsymbol{z}_i = (z_{1,i},\ldots,z_{m,i})'$ representa los valores de las covariables asociadas con el i-ésimo sujeto.

Usando la notación en [5] en una muestra aleatoria de tamaño n y bajo el modelo (5) el log de la verosimilitud está dado por:

$$\log L(\boldsymbol{\gamma},S_1,S_2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}) = \sum_{i=1}^{n} l(\boldsymbol{\gamma},S_1,S_2|\boldsymbol{y_i},\boldsymbol{x_i},\boldsymbol{z_i},\boldsymbol{\delta}) = \sum_{i=1}^{n} l_i, \tag{20}$$

donde $\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z}$ representan las matrices de los vectores $\boldsymbol{y_i},\boldsymbol{x_i}$ y $\boldsymbol{z_i}$ respectivamente. Cada contribución de la verosimilitud individual puede escribirse como una suma de nueve términos diferentes dependiendo de si la observación tiene censura a izquierda, a intervalo o a derecha en ambas

dimensiones, es decir,

$$\begin{aligned} l(\boldsymbol{\gamma}, S_1, S_2|\boldsymbol{X}, \boldsymbol{\delta}) = {} & \delta_1^{(1)}\delta_2^{(1)} \log S_{11}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + \delta_1^{(1)}\delta_2^{(2)} \log S_{12}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + \delta_1^{(1)}(1-\delta_2^{(1)}-\delta_2^{(2)}) \log S_{13}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + \delta_1^{(2)}\delta_2^{(1)} \log S_{21}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + \delta_1^{(2)}\delta_2^{(2)} \log S_{22}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + \delta_1^{(2)}(1-\delta_2^{(1)}-\delta_2^{(2)}) \log S_{23}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + (1-\delta_1^{(1)}-\delta_1^{(2)})\delta_2^{(1)} \log S_{31}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + (1-\delta_1^{(1)}-\delta_1^{(2)})\delta_2^{(2)} \log S_{32}(\boldsymbol{\gamma}|\boldsymbol{X}) \\ & + (1-\delta_1^{(1)}-\delta_1^{(2)})(1-\delta_2^{(1)}-\delta_2^{(2)}) \log S_{33}(\boldsymbol{\gamma}|\boldsymbol{X}), \end{aligned} \tag{21}$$

donde,

- $S_{11}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(T_1 \leq l_1, T_2 \leq l_2) = 1 - S_1(l_1) - S_2(L_2) + \breve{C}_\alpha[S_1(l_1), S_2(l_2)]$.
- $S_{12}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(T_1 \leq l_1, l_2 < T_2 \leq u_2) = S_1(l_1) - S_1(u_1) + \breve{C}_\alpha[S_1(l_1), S_2(u_2)] - \breve{C}_\alpha[S_1(l_1), S_2(l_2)]$.
- $S_{13}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(T_1 \leq l_1, T_2 > u_2) = S_2(u_2) - \breve{C}_\alpha(S_1(l_1), S_2(u_2))$.
- $S_{21}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(l_1 < T_1 \leq u_1, T_2 \leq l_2)$.
- $S_{22}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(l_1 < T_1 \leq u_1, l_2 < T_2 \leq u_2) = \breve{C}_\alpha[S_1(l_1), S_2(l_2)] - \breve{C}_\alpha[S_1(l_1), S_2(u_2)] - \breve{C}_\alpha[S_1(u_1), S_2(l_2)] + \breve{C}_\alpha[S_1(u_1), S_2(u_2)]$.
- $S_{23}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(l_1 < T_1 \leq u_1, T_2 > u_2) = \breve{C}_\alpha[S_1(l_1), S_2(u_2)] - \breve{C}_\alpha[S_1(u_1), S_2(u_2)]$.
- $S_{31}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(T_1 > u_1, T_2 \leq l_2) = S_1(u_1) - \breve{C}_\alpha[S_1(u_1), S_2(u_2)]$.
- $S_{32}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(T_1 > u_1, l_2 < T_2 \leq u_2) = \breve{C}_\alpha[S_1(u_1), S_2(l_2)] - \breve{C}_\alpha[S_1(u_1), S_2(u_2)]$.
- $S_{33}(\boldsymbol{\gamma}|\boldsymbol{X}) = P(T_1 > u_1, T_2 > u_2) = \breve{C}_\alpha[S_1(u_1), S_2(u_2)]$.

Para estimar el parámetro α de la cópula y si las funciones de supervivencia S_1 y S_2 son conocidas, un estimador natural está dado por el estimador de máxima verosimilitud en (20). La estimación de la máxima verosimilitud completa puede resultar en cálculos bastante largos, sin embargo, en [7] se propone un procedimiento de dos etapas basado en la pseudo verosimilitud en forma parámetrica, en [5] siguen este procedimiento, en el que se estiman S_1 y S_2 y se reemplazan $\widehat{S_1}$ y $\widehat{S_2}$ en (20). Luego, se estima $\boldsymbol{\gamma}$ maximizando la pseudo verosimilitud $l(\boldsymbol{\gamma}, \widehat{S_1}, \widehat{S_2})$, obtenida de la ecuación (21) conectando las estimaciones $\widehat{S_1}$ y $\widehat{S_2}$.

Como en [5] se propone modelar las distribuciones marginales de supervivencia con un modelo de tiempo de falla acelerado con un término de error flexible propuesto por [6], este enfoque permite la incorporación de covariables en las distribuciones marginales. Formalmente, se estiman S_k, para $k = 1, 2$ de la siguiente expresión:

$$\log(T_{k,i}) = \boldsymbol{\beta}_k^{'} \boldsymbol{z}_i + \varepsilon_{k,i},\ i = 1, \ldots n, \tag{22}$$

donde $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,m})^{'}$ es un vector m-dimensional de parámetros de la regresión desconocidos y $\varepsilon_{k,1}, \ldots, \varepsilon_{k,n}$ son variables de error aleatorias independientes e idénticamente distribuidas con densidad $g_{\varepsilon_k}(\varepsilon_k)$, la cual se expresa utilizando la mezcla normal penalizada, es decir,

$$g_{\varepsilon_k}(\varepsilon_k) = \zeta_k^{-1} \sum_{j=1}^{K_k} c_{k,j}(\boldsymbol{a}_k) \phi\left(\frac{\varepsilon_k - \eta_k}{\zeta_k} \middle| \mu_{k,j}, \sigma_{k,0}^2 \right) \tag{23}$$

donde $\mu_{k,1}, \ldots, \mu_{k,K_k}$ es un conjunto de knots equidistantes fijos, $\sigma_{k,0}$ es una base fija de desviación estándar, η_k es un intercepto desconocido, ζ_k es un parámetro de escala desconocido y $\phi(\cdot|\mu, \sigma^2)$ una densidad normal con media μ y desviación estándar σ. Finalmente, sea $\boldsymbol{a}_k = (a_{k,1}, \ldots, a_{k,K_k})^{'}$ un vector de parámetros y definiendo los pesos de la mezcla desconocidos $c_{k,j}(\boldsymbol{a}_k)$ usando la siguiente ecuación:

$$c_{k,j}(\boldsymbol{a}_k) = \frac{\exp(a_{k,j})}{\sum_{p=1}^{K_k} \exp(a_{k,p})},\ -\infty < a_{k,j} < \infty \text{ y } j = 1, \ldots, K_k. \tag{24}$$

Al introducir $\boldsymbol{a}_k$, el problema de máxima verosimilitud restringido cambia a uno de verosimilitud no restringido. Para facilitar la notación, los autores en [5] asumen que $a_{k,[K_k/2]=0}$, donde $[.]$ es la función techo de un número real. Los parámetros $\boldsymbol{\beta}_k$ y $\boldsymbol{a}_k$ se estiman con máxima verosimilitud penalizada, en donde la penalización se aproxima por el método de diferencias finitas cuadradas de orden s para los parámetros $\boldsymbol{a}_k$.

Siguiendo a [5], dado un parámetro de suavizado λ, la verosimilitud penalizada está representada de la siguiente manera:

$$l_{P,n} = l_n - \frac{\lambda}{2} \sum_{j=s+1}^{K_k} \left(\Delta^s a_{k,j}\right)^2, \tag{25}$$

donde l_n representa la verosimilitud ordinaria de las n observaciones y Δ^s es el operador de diferencia de orden s. El parámetro de suavizado óptimo se elige minimizando el criterio de información de Akaike (AIC) a partir de un conjunto de diferentes valores de λ.

La estimación de la varianza para los parámetros de la cópula estimados es difícil, por lo que en [18] proponen un procedimiento a través de bootstrap. Para M fijo, se producen M estimadores $\widehat{\boldsymbol{\gamma}}_m$; $m = 1, \ldots, M$ de $\boldsymbol{\gamma}$. La varianza de $\widehat{\boldsymbol{\gamma}}$ puede ser estimada por la varianza muestral de los $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{m}}$'s. Al usar el método Delta, se obtiene la varianza para la estimación de otros parámetros como α que es el parámetro de una determinada cópula o para la medida de asociación.

4. Estudio de simulación

Para realizar la estimación del τ de Kendall con el método de ajuste individual de las marginales, se utilizó el paquete **censcor** del software estadístico *R*.

Para la simulación del método cópula se utiliza el enfoque descrito en [5], en el que implementa la función `fit.copula`, que está disponible en el paquete **icensBKL** de *R*. En este método para modelar las distribuciones marginales con un modelo de falla acelerado con término de error flexible [6], se utilizó el paquete **smoothSurv** de *R*, utilizando la ecuación (22).

4.1. Medidas de calidad de los estimadores

Para evaluar si las estimaciones asociadas a un paramétro de interés a partir de una muestra aleatoria resultan adecuadas, se pueden utilizar medidas de calidad, tales como el error cuadrático medio y la mediana de la desviación absoluta. Al calcular estas medidas a diferentes estimadores, se puede establecer cuál de ellos es el mejor.

4.1.1. Error cuadrático medio (ECM)

Sea T un estimador de un paramétro desconocido θ. En [19] se define el ECM como el valor esperado del cuadrado de la diferencia entre T y θ, es decir

$$\text{ECM}(T) = E[(T-\theta)^2] = V(T) + [B(T)]^2, \qquad (26)$$

donde $B(T)$ y $V(T)$ son el sesgo y la varianza del estimador puntual T, respectivamente.

4.1.2. Mediana de la desviación absoluta (MDA)

La MDA es una medida robusta para la variabilidad de un estimador. Sea θ un parámetro de interés y sea T es estimador puntual de θ, la MDA se define como la mediana del valor absoluto de la diferencia entre la estimación y el valor real [4]:

$$\text{MDA} = \text{mediana}(|T-\theta|). \qquad (27)$$

4.2. Esquema de simulación

En el esquema del estudio de simulación se generaron los datos de una cópula Gumbel con la función `BiCopsim` del paquete **VineCopula** de *R*, con la finalidad de generar datos bivariados de una distribución Weibull, la cual es una de las distribuciones que más se utiliza en confiabilidad. Teniendo en cuenta que las marginales de una cópula son distribuciones uniformes, *se emplea el teorema de la transformación inversa para generar marginales de una distribución exponencial con media* 1, la cual es un caso particular de la distribución Weibull.

El porcentaje de censura que se utilizó en la generación de los datos es el siguiente: para garantizar el 30 % de censura a izquierda se trabaja con el quantil 0,3, para el 30 % de censura a derecha se toma el cuantil 0,7 y el restante son censuras a intervalo. La relación que tiene el coeficiente de concordancia τ de Kendall con el parámetro de la cópula Gumbel es $\tau = 1-\alpha$. Se creó una base de datos bivariados con los tres tipos de censura en las marginales. Se usó esta base para estimar el τ de Kendall con el método para datos suponiendo normalidad en las marginales y ajustándolas individualmente. La misma base de datos se usó para estimar el τ de Kendall usando el método cópula para las familias Gaussiana y Clayton.

Para la generación de las visitas se tiene en cuenta lo siguiente: la primera visita se genera aleatoriamente de una distribución uniforme entre $(0,1)$ y para las siguientes se suma la constante 1, representando una visita cada año para el evento de interés, con máximo 10 visitas.

En el esquema de simulación se tuvo en cuenta lo siguiente:

1. Se consideraron 9 escenarios de simulación determinados por:
   a. Tres valores del τ de Kendall, $\tau = 0{,}2, 0{,}5, 0{,}8$.
   b. Tres tamaños muestrales $n = 50, 100, 200$.
2. Se calcula el ECM y la MDA para la estimación del τ de Kendall en cada conjunto de datos bivariados, aplicando los métodos de estimación antes descritos: **M1.** Ajuste individual de las marginales, **M2G.** Método cópula Gaussiana, y **M2C.** Método cópula Clayton.

Para el cálculo del ECM y la MDA en cada escenario de simulación, se generaron 500 réplicas de la base de datos bivariada.

4.3. Resultados

n	τ	ECM		
		M1	**M2G**	**M2C**
50	0.2	0.0201	0.0239	0.0144
	0.5	0.1027	0.0416	0.0696
	0.8	0.1848	*	0.0880
100	0.2	0.0172	0.0141	0.0188
	0.5	0.0860	0.0497	0.0705
	0.8	0.1746	*	0.0953
200	0.2	0.0160	0.0087	0.0247
	0.5	0.0873	0.0601	0.0604
	0.8	0.1675	*	0.0943

*: no reportado por problema de estimación en este escenario.

Tabla 2: ECM de las estimaciones usando el Método de ajuste individual de las marginales (**M1**), el Método cópula Gaussiana (**M2G**) y el Método cópula Clayton (**M2C**)

n	τ	MDA		
		M1	**M2G**	**M2C**
50	0.2	0.1210	0.1019	0.1048
	0.5	0.3140	0.1536	0.2357
	0.8	0.4350	*	0.2947
100	0.2	0.1219	0.0706	0.1362
	0.5	0.2857	0.1602	0.2426
	0.8	0.4065	*	0.3069
200	0.2	0.1234	0.0666	0.1603
	0.5	0.2947	0.2133	0.2339
	0.8	0.3985	*	0.3048

*: no reportado por problema de estimación en este escenario.

Tabla 3: MDA de las estimaciones usando el Método de ajuste individual de las marginales (**M1**), el Método cópula Gaussiana (**M2G**) y el Método cópula Clayton (**M2C**)

Las Tablas 2 y 3 dan los valores del ECM y de la MDA, respectivamente, para los dos métodos bajo estudio: método del ajuste individual de las marginales (M1) y el método cópula (M2G y M2C), para cada combinación de los parámetros n y τ. Observe que para el valor de $\tau = 0{,}8$, en la estimación del ECM y de la MDA por medio de la cópula Gaussiana (M2G), el proceso de estimación de τ falló, por lo cual no se reportaron tales valores.

A partir de los resultados mostrados en las Tablas 1 y 2, se construyen las gráficas que se presentan a continuación en donde se puede apreciar mejor el comportamiento de las medidas de calidad de los estimadores (ECM y MDA) en los escenarios analizados.

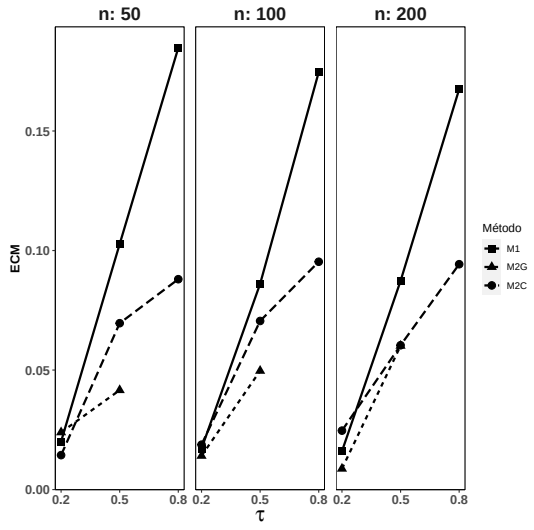

Figura 1: Relación entre el ECM y el τ de Kendall, para los métodos de estimación y tamaños de muestra bajo estudio

En la Figura 1, se puede observar que bajo una baja dependencia ($\tau = 0{,}2$) los métodos estudiados tienen valores de ECM muy similares, pero a medida que la dependencia aumenta a valores moderados ($\tau = 0{,}5$) o fuertes ($\tau = 0{,}8$), el método cópula (M2G y M2C) presentó valores menores con respecto al método de ajuste individual de las marginales (M1). Note también que los valores de ECM aumentan a medida que también lo hace el τ de Kendall.

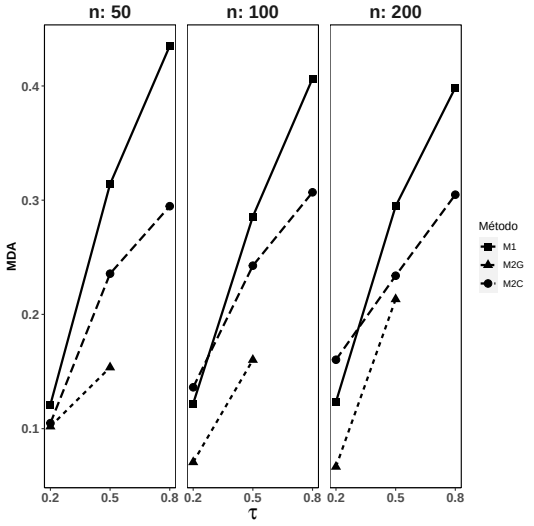

Figura 2: Relación entre la MDA y el τ de Kendall, para los métodos de estimación y tamaños de muestra bajo estudio

En la Figura 2, se puede observar que bajo una alta dependencia ($\tau = 0{,}8$) el método cópula Clayton (M2C) presentó valores menores en la MDA que el método de ajuste individual de las marginales (M1), pero bajo valores de dependencia moderada ($\tau = 0{,}5$) o baja ($\tau = 0{,}2$), el método cópula Normal (M2G) presentó valores menores con respecto a los otros dos métodos (M1 y M2C). De nuevo note que los valores de MDA aumentan a medida que también lo hace el τ de Kendall.

5. Conclusiones

De acuerdo a los métodos descritos en secciones anteriores y al proceso de simulación se llega a las siguientes conclusiones:

- Para escenarios de simulación con dependencia alta ($\tau = 0{,}8$) el proceso de estimación del τ de Kendall usando el método cópula Gaussiana (M2G) falló, por lo cual no se reportaron valores del ECM y de la MDA. En este escenario, el método cópula Clayton produce mejores valores de las medidas de calidad de la estimación con respecto al método de ajuste individual de las marginales.
- La estimación del τ de Kendall por medio del método cópula usando las cópulas Normal (M2G) y Clayton (M2C), es en general mejor que la estimación del τ

de Kendall con el método de ajuste individual de las marginales (M1), ya que proporciona valores de MDA y ECM más bajos.

- En los métodos de estimación bajo estudio, se observó que a medida que el coeficiente τ de Kendall aumenta, los valores de MDA y ECM también lo hacen, lo que indica que los estimadores pierden precisión, cuando los datos son dependientes.

Una explicación sobre el por qué, en el estudio de simulación, el método basado en cópulas tiene un mejor desempeño que el método de ajuste individual de las marginales, es que este último impone que la distribución conjunta de los datos sigue una distribución normal bivariada, mientras que en el método cópula el ajuste de las marginales se hace a través de un modelo de tiempo de falla acelerado con un término de error flexible, que es menos restrictivo.

Referencias

[1] Meeker, W. and Escobar, L. (1998). *Statistical Methods for Reliability Data*. John Wiley & Sons.

[2] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93.

[3] Betensky, R. and Finkelstein, D. (1999a). An extension of Kendall's coefficient of concordance to bivariate interval censored data. *Statistics in Medicine*, 18:3101–3109.

[4] Newton, E. and Rudel, R. (2007). Estimating correlation with multiply censored data arising from the adjustment of singly censored data. *Environmental Science and Technology*, 41:221–228.

[5] Bogaerts, K. and Lesaffre, E. (2008b). Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in Medicine*, 27:6379–6392.

[6] Komárek, A., Lesaffre, E., and Hilton, J. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14:726–745.

[7] Bogaerts, K., Komarek, A., and Lesaffre, E. (2017). *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*. Chapman and Hall.

[8] Nelsen, R. (2006). *An Introduction to Copulas*. Springer.

[9] Lesaffre, E. and Bogaerts, K. (2005). Estimating Kendall's tau for bivariate interval censored data with a smooth estimate of the density. In *Statistical Solutions to Modern Problems: Proceedings of the 20th International Workshop on Statistical Modelling*, pages 325–328.

[10] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

[11] Lopera, C.M., Jaramillo, M.C. and Ardila. L.D. (2008) Selección de un modelo cópula para el ajuste de datos bivariados dependientes. *Dyna*, 76(158):253–263.

[12] Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84:487–493.

[13] Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, (292):698–707.

[14] Lu, J.-C. and Bhattacharyya, G. K. (1990). Some new constructions of bivariate weibull models. *Annals of the Institute of Statistical Mathematics*, (3):543–559.

[15] Eilers, P. and Marx, B. (1996). Flexible smoothing with B-Splines and penalties. *Statistical Science*, pages 89–102.

[16] Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60:945–953.

[17] Greiner, R. (1909). Über das fehlersystem der kollektivmasslehre. *Zeitschrift für Mathematik und Physik*, (121):225.

[18] Sun, L., Wang, L., and Sun, J. (2006). Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, 33:637–649.

[19] Canavos, G. (1988). *Probabilidad y Estadística Aplicaciones y Métodos*. McGraw Hill, México.