UNIVERSIDAD NACIONAL DE COLOMBIA

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

Jessica Katherine Serna Morales

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de estadística
Medellín, Colombia
2019

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

Jessica Katherine Serna Morales

Tesis presentada como requisito parcial para optar al título de:
Magister en Ciencias Estadística

Director:
Mario César Jaramillo Elorza, Ph.D
Profesor Asociado

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2019

Dedico este trabajo a nuestro padre celestial, por iluminar mi camino y por todas las bendiciones a lo largo de mi vida, a mis padres, que me apoyaron desde el inicio de mis estudios, a mi familia en especial a mi tía y a mi prima, por estar siempre pendientes de mi, a mis profesores por todos los conocimientos que me transmitieron, en especial al profesor Mario César por su paciencia y disposición para trabajar conmigo, un agradecimiento al profesor Luis Escobar por sus conocimientos y sugerencias en este trabajo y por último a la persona más especial de vida, a mi pareja por esperarme y estar junto a mí.

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 las variables. El coeficiente de concordancia τ de Kendall se ha convertido en una herramienta útil para el análisis de datos bivariados, a través de pruebas no paramétricas de independencia y medidas complementarias de asociación. En el análisis de datos de confiabilidad hay un fenómeno que ocurre cuando el valor de las observaciones se conoce parcialmente, el cual se conoce como censura. En este trabajo se comparan vía simulación dos extensiones del τ de Kendall, una de ellas es suponiendo normalidad en las distribuciones marginales y ajustandolas individualmente y la otra está basada en cópulas, donde los datos bivariados están censurados a intervalo. La comparación es hecha mediante tres medidas, a saber, la desviación mediana absoluta, el error cuadrático medio y la eficiencia relativa.

Palabras clave: Asociación, Dependencia, Datos Bivariados, Censura a intervalo, Confiabilidad, Supervivencia, Cópula, τ de Kendall.

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 coefficient measures the linear relationship between two variables, but there may also be a strong non-linear relationship between two variables. The concordance coefficient Kendall's τ has become a useful tool for the analysis of bivariate data, through nonparametric tests of independence and complementary measures of association. In the analysis of reliability data there is a data feature that occurs when the value of the lifetime is partially known, which is known as censoring. In this paper, we using simulations to compare two extensions of the Kendall's τ, one of them is assuming normality in marginal distributions and adjusting them individually and another one focused on copulas, where the bivariate data are censored at intervals. The comparison is made by three measures, namely, the median absolute deviation, the mean squared error and the relative efficiency.

Keywords: Association, Dependency, Bivariate Data, Interval Censored, Reliability, Survival, Copula, Kendall's τ.

Lista de Figuras

Lista de Tablas

Contenido

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. 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 medir dicha estructura usualmente se utiliza el τ de Kendall.

Una característica típica de los datos en confiabilidad es la presencia de censura. Hay varios tipos de censura, entre ellas está la censura a derecha, a izquierda y a intervalo. En este trabajo es de interés la censura a intervalo, que se produce cuando el tiempo de falla de un producto se encuentra en intervalos los cuales, en general, son aleatorios.

El coeficiente de concordancia τ de Kendall, propuesto por Kendall en [22], mide el grado de dependencia entre dos variables aleatorias, cuya escala de medida es ordinal ó de intervalo; la estimación del τ de Kendall se basa en el orden (rangos) de las observaciones. La propiedad más importante del τ de Kendall es que es invariante a transformaciones monótonas. La principal ventaja del τ de Kendall es que la distribución de su estimación cuando el tamaño de la muestra es grande se aproxima a una distribución Normal rápidamente, de tal forma que la aproximación Normal es mejor para la estimación del τ de Kendall, que para la estimación del rho de Spearman, cuando se contrasta un juego de hipótesis y la hipótesis nula de independencia entre las variables X y Y es cierta.

Las pruebas no paramétricas y las medidas de asociación más conocidas están propuestas en [13]. Para el análisis de datos bivariados, se emplean pruebas no paramétricas de independencia y medidas complementarias de asociación. Una extensión para estimar el coeficiente de concordancia τ de Kendall, con censura a intervalo, es propuesta por [1].

Para estimar el τ de Kendall en datos bivariados con censura a intervalo, [5] propone modelar las distribuciones marginales con un modelo de falla acelerado con un término de error flexible sugerido por [23], la asociación la modelan de forma paramétrica y utilizan las cópulas de Clayton, Normal y de Placckett, para datos bivariados.

En este trabajo, se exploran dos métodos implementados en [29] y [5] para estimar el co-

eficiente de concordancia τ de Kendall para datos bivariados con censura a intervalo. En el Capítulo 2 se presentan algunos conceptos importantes para el desarrollo de este trabajo, se definen algunos conceptos que implementan la estimación del τ de Kendall, como las aproximaciones por cópulas de [5] y el estimador de máxima verosimilitud no paramétrico propuesto por [2] para datos bivariados con censura a intervalo. En el Capítulo 3 se presentan dos extensiones para estimar el τ de Kendall; la primera es la estimación suponiendo normalidad en las marginales y ajustandolas individualmente y la segunda es la estimación por medio de cópulas. En el Capítulo 4 se realiza un estudio de simulación, en donde se comparan resultados analizados de los dos métodos a través de tres estimadores que son la desviación mediana absoluta, el error cuadrático medio y la eficiencia relativa. En el Capítulo 5 se presentan las conclusiones de la investigación.

2. Marco Teórico

A continuación se presentan algunos conceptos importantes para el desarrollo de este proyecto relacionados con las extensiones del τ de Kendall para datos bivariados con censura a intervalo.

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 los tiempos de supervivencia de individuos de alguna población.

La función de supervivencia $S(t)$ se define como:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)\, dx, \tag{2-1}$$

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:

$$S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2),\ t_1 \geq 0,\ t_2 \geq 0. \tag{2-2}$$

Es decir $S(t_1, t_2)$ es la probabilidad de que ambas unidades sobrevivan en 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 ó en intervalos.

Siguiendo la notación en [3], 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 y el evento no ha ocurrido hasta la última visita. 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 (tiempo) resultan cuando todas las unidades que no han fallado antes de un tiempo pre-especificado t_c, se censuran en el tiempo t_c. Los datos con censura tipo II (falla) resultan cuando una prueba es terminada después de un número especificado de r fallas, $2 \leq r \leq n$. Cuando $n = r$, todas las unidades fallan y los datos se llaman 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 ó a intervalo. Siguiendo la notación en [3] se utiliza un indicador de censura δ igual a $0, 1, 2$ y 3 para denotar censura a derecha, observaciones exactas, censura a izquierda o en intervalos respectivamente, como se ilustra en la siguiente tabla:

Observación	Limites del intervalo $\lfloor l, u \rfloor$	Indicador de censura
Censura a derecha en el tiempo l	$0 < l < u = \infty$	$\delta = 0$
Tiempo exacto observado t	$0 < l = u = t < \infty$	$\delta = 1$
Censura a izquierda en el tiempo u	$0 = l < u < \infty$	$\delta = 2$
Censura en intervalos en $\lfloor l, u \rfloor$	$0 < l < u < \infty$	$\delta = 3$

Tabla 2-1.: Taxonomía de observaciones censuradas a intervalo.

2.3. Estimación de máxima verosimilitud no paramétrica de la Función de supervivencia con datos con censura a intervalo

A continuación se muestran algunos enfoques frecuentistas para estimar la supervivencia $S(t)$ en presencia de censura a intervalo, de acuerdo a lo presentado en [3].

Siguiendo la notación en [3], se introduce el estimador de máxima verosimilitud no paramétri-

co (NPMLE), también llamado estimador de Turnbull de $S(t)$. Se supone que los tiempos de supervivencia provienen de una muestra de individuos independientes e idénticamente distribuidos con función de distribución acumulada $F(t)$, función de supervivencia $S(t)$ y función de densidad de probabilidad $f(t)$. En lugar de observar $T_i,\ i = 1, \cdots, n$ exactos se observa un conjunto de intervalos e indicadores de censura $\mathbf{D} = \lfloor l_i, u_i \rfloor,\ \delta_i : i = 1, 2, \cdots, n$. Para una observación exacta se escribe $t_i = l_i = u_i$, es decir, cuando $\delta_i = 1$; el estimador NPMLE para datos con censura a intervalo, no tiene, en general una solución cerrada y debe obtenerse mediante un algoritmo iterativo. Siguiendo la notación en [3], $\lfloor l_i, u_i \rfloor$ hace referencia a un intervalo abierto, semiabierto o cerrado con limite inferior l_i y limite superior u_i.

Para calcular el NPMLE de una distribución de supervivencia se debe tener en cuenta lo siguiente

Sea $\lfloor l_i, u_i \rfloor$ $(i = 1, \cdots, n)$ los intervalos observados de n sujetos independientes, que contienen los tiempos desconocidos del evento de interés, t_i. Peto en [31] fue el primero en notar que la solución de máxima verosimilitud da como resultado un conjunto de intervalos $[p_j; q_j]_{j=1}^m$ con la propiedad de que la función de supervivencia estimada es constante fuera de los intervalos. Además, la masa o probabilidad asignada a cada uno de los intervalos está bien determinada, pero dentro de cada intervalo no hay información sobre cómo se asigna esa masa. Los intervalos se llaman regiones de posible masa o soporte porque el procedimiento de máxima verosimilitud sólo puede indicar en qué regiones pueden ocurrir los eventos. Por tanto, sólo se puede afirmar que los intervalos $[p_j; q_j]_{j=1}^m$ tienen masa mayor o igual a cero, pero no necesariamente todas las masas son mayores que cero. Estos son los intervalos a los que se les asigna la posible masa o probabilidad.

[31] y [34] sugieren un algoritmo de reducción simple para identificar los intervalos de posible masa a partir de los datos. A saber, dadas las observaciones $[l_i, u_i]$ $(i = 1, \cdots, n)$, se clasifica los puntos de tiempo $\{l_i\}$ y $\{u_i\}$ en orden creciente y se identifica si el punto es un extremo izquierdo o un extremo derecho. Las regiones de posible masa son entonces los intervalos con un punto extremo izquierdo inmediatamente seguido de un extremo derecho. Esto facilita considerablemente la estimación no paramétrica de la función de supervivencia. Peto en [31] y Turnbull en [34] usaron intervalos cerrados, porque la solución del estimador de máxima verosimilitud no parámetrico depende de las propiedades de los intervalos cerrados.

Dadas las regiones de posible soporte, la masa asignada a cada uno de estos intervalos debe estimarse en un segundo paso. Para intervalos semiabiertos o cerrados, el algoritmo de reducción anterior da lugar a un conjunto de intervalos $\lfloor p_j; q_j \rfloor_{j=1}^m$; se define $s_j = S(p_j-) - S(q_j+)$ $(j = 1, \cdots, m)$, donde $S(p_j-)$ denota el valor de la supervivencia a la izquierda de p_j y $S(q_j+)$ denota el valor de la función de supervivencia a la derecha de q_j. El vector $\mathbf{s} = (s_1, \cdots, s_m)'$, con $\sum_{j=1}^m s_j = 1$ y $s_j \geq 0$ definen las clases de equivalencia en el espacio de

las funciones de distribución S, que son planas fuera de $\bigcup_{j=1}^{m} \lfloor p_j, q_j]$. Todas las funciones en la misma clase de equivalencia (la noción de relación de equivalencia sobre un conjunto, permite establecer una relación entre los elementos del conjunto que comparten cierta característica o propiedad.Esto permite reagrupar dichos elementos en clases de equivalencia) tendrán la misma verosimilitud porque ésta depende solo de los valores en p_j y q_j $(j = 1, \cdots, m)$ y no en cómo la función evoluciona entre p_j y q_j $(j = 1, \cdots, m)$. Así, la definición del estimador de máxima verosimilitud de la función S se puede restringir a estas clases y se reduce a maximizar

$$L = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} s_j \right), \tag{2-3}$$

donde

$$\alpha_{ij} = \begin{cases} 1 & \text{si } \lfloor p_j, q_j] \subset \lfloor l_i, u_i] \\ 0 & \text{en otro caso} \end{cases}$$

El NPMLE de S se puede estimar mediante la maximización de la verosimilitud restringida L con restricciones lineales

$$1 - \sum_{j=1}^{m} s_j = 0, \tag{2-4}$$

$$s_j \geq 0 \ (j = 1, \cdots, m). \tag{2-5}$$

Esto se puede lograr con una variedad de algoritmos, como el algoritmo de autoconsistencia descrito en [34], que es de hecho un ejemplo del algoritmo de Expectation-Maximization (EM) el cual se puede encontrar en [7]. En ausencia de tiempos de eventos exactos, el algoritmo de autoconsistencia requiere en cada momento el número esperado de sujetos en riesgo y sujetos que fallan. Con estos pseudo-datos, se calcula un estimador producto límite. Se inicia con algunas estimaciones iniciales para s_j $(j = 1, \cdots, m)$, por ejemplo si se toma $\widehat{s_j} = 1/m$ para todo j. Entonces se estima S con el siguiente algoritmo:

$$\begin{aligned} \widehat{S}(q_0+) &= 1 \\ \widehat{S}(q_j+) &= \widehat{\pi}_j \widehat{S}(q_{j-1}+) \ (j = 1, \cdots, m) \\ \widehat{\pi}_j &= (n'_j - d'_j)/n'_j, \end{aligned}$$

donde

$$\begin{aligned} q_0 &= 0, \\ d'_j &= \sum_{i=1}^{n} \left(\alpha_{ij} \widehat{s_j} / \sum_{k=1}^{m} \alpha_{ik} \widehat{s_k} \right), \\ n'_j &= \sum_{k=j}^{m} \sum_{i=1}^{n} \left(\alpha_{ik} \widehat{s_k} / \sum_{r=1}^{m} \alpha_{ir} \widehat{s_r} \right). \end{aligned}$$

Después de este paso, los $\widehat{s_j}$ $(j = 1, \cdots, m)$ se calculan otra vez utilizando la nueva estimación $\widehat{S}(q_j+)$ $(j = 1, \cdots, m)$ y este proceso se repite hasta obtener la precisión requerida. Hay que tener en cuenta que el algoritmo de autoconsistencia puede detenerse en una solución óptima local y, por lo tanto, no produce el NPMLE.

En [10] describe cómo determinar si un estimador candidato $\widehat{s}$ de s es de hecho el NPMLE. En [10] se muestra como las condiciones de Kuhn-Tucker, dan los criterios necesarios y suficientes para proporcionar una solución óptima y válida. Para aplicar este procedimiento a nuestro problema de maximización se hace lo siguiente. Sea $l(\mathbf{s}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} \alpha_{ij} s_j\right)$ el log de la verosimilitud y $d_k = \partial l / \partial s_k$ $(k = 1, \cdots, m)$. Las condiciones de Kuhn-Tucker establecen que $\widehat{s}$ es el estimador de máxima verosimilitud si y solo si existen valores μ_j $(j = 0, \cdots, m)$ tal que:

$$\mu_j s_j = 0 \ (j = 1, \cdots, m), \tag{2-6}$$

$$\mu_j \ \geq 0 \ (j = 1, \cdots, m), \tag{2-7}$$

$$\frac{\partial}{\partial s_j}\left[l(\mathbf{s}) + \sum_{k=1}^{m} s_k(\mu_k - \mu_0)\right] = d_j + \mu_j - \mu_0 = 0 \quad (j = 1, \cdots, m). \tag{2-8}$$

Las condiciones de Kuhn-Tucker corresponden a las ecuaciones 2-4, 2-5, 2-6, 2-7 y 2-8.

Los μ_j se llaman los multiplicadores de Lagrange. En términos generales, un multiplicador de Lagrange, es un escalar que se introduce como ayuda para resolver un problema en n_v variables con n_c restricciones a un problema que tiene solución en $n_v + n_c$ variables sin restricciones. De estas condiciones se puede derivar que $\mu_o = n$. En las ecuaciones anteriores, a $d_j + \mu_j - \mu_0$ se le llama el gradiente reducido, porque es el gradiente con respecto a las variables libres. Las condiciones de Kuhn-Tucker se satisfacen si los multiplicadores de Lagrange son mayores o iguales a cero, $\mu_j = 0$ cuando $s_j > 0$ para $j = 1, \cdots, m$ y el gradiente reducido es cero. Los intervalos $\lfloor p_j, q_j]$ con $s_j > 0$ $(j = 1, \cdots, m)$ se llaman las regiones de soporte del NPMLE de S.

Dado que el NPMLE es invariante al lugar donde se coloca la masa de probabilidad dentro de las regiones de soporte, esto hace que el NPMLE no sea único (ver [11]).

2.4. Medidas de calidad de 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 como el análisis al centro de la distribución muestral, el insesgamiento y la mínima varianza. Al calcular estas medidas a diferentes estimadores bajo comparación se puede establecer cuál estimador es el mejor.

2.4.1. Error cuadrático medio

Sea T un estimador de un paramétro desconocido θ. En [6] se define el error cuadrático medio como el valor esperado del cuadrado de la diferencia entre T y θ, es decir

$$\begin{aligned} ECM(T) &= E[(T-\theta)^2] \\ &= E[T^2 - 2\theta T + \theta^2] \\ &= E(T^2) - [E(T)]^2 + [E(T)]^2 - 2\theta E(T) + \theta^2 \\ &= V(T) + [\theta - E(T)]^2 \\ &= V(T) + [B(T)]^2, \end{aligned} \tag{2-9}$$

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

2.4.2. Eficiencia

Se dice que un estimador $\widehat{\theta}$ de un parámetro poblacional θ es eficiente para estimar a θ si alcanza la cota de Crámer- Rao, es decir,

$$V(\widehat{\theta}) = \frac{\left[\frac{\partial E(\widehat{\theta})}{\partial \theta}\right]^2}{nE\left[\frac{\partial \log f_\theta(x)}{\partial \theta}\right]^2}, \tag{2-10}$$

donde el denominador se llama la cantidad de información de Fisher $I(\theta)$.

En [27] se define la eficiencia relativa entre dos estimadores de un parámetro como sigue: Sean $\widehat{\theta}_1$ y $\widehat{\theta}_2$ dos estimadores del paramétro θ y sean $ECM(\widehat{\theta}_1)$ y $ECM(\widehat{\theta}_2)$ los errores cuadráticos medios asociados a $\widehat{\theta}_1$ y $\widehat{\theta}_2$ respectivamente. La eficiencia relativa de $\widehat{\theta}_2$ respecto a $\widehat{\theta}_1$ se define como:

$$\frac{ECM(\widehat{\theta}_1)}{ECM(\widehat{\theta}_2)}. \tag{2-11}$$

Si esta eficiencia relativa es menor a 1, se puede concluir que $\widehat{\theta}_1$ es un estimador más eficiente de θ que $\widehat{\theta}_2$, en el sentido que tiene menor error cuadrático medio.

2.4.3. Mediana de la desviación absoluta (MAD)

El MAD es una medida robusta para la variabilidad de un estimador. En [29] se define como la mediana del valor absoluto de la diferencia entre la estimación y el valor real.

Sea θ un parámetro de interés y sea T es estimador puntual de θ, el MAD se define de la siguiente manera:

$$\text{MAD} = \text{mediana}(|T - \theta|), \tag{2-12}$$

donde T es el estimador de θ.

2.5. τ de Kendall

El τ de Kendall es una medida de dependencia que representa el grado de concordancia entre dos variables. La estimación del τ de Kendall definida en [24] 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 denota 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:

$$\widehat{\tau} = \frac{N_c - N_d}{N_c + N_d}. \tag{2-13}$$

Equivalentemente, el τ de Kendall se define como la probabilidad de concordancia menos la probabilidad de discordancia para un par de observaciones (x_i, y_i) y (x_j, y_j) elegidas aleatoriamente de la muestra.

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, entonces el τ de Kendall es dado por la probabilidad de concordancia menos la probabilidad de discordancia (ver [28]) dada por:

$$\tau = \tau(X, Y) = P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right]. \tag{2-14}$$

2.5.1. 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 (X, Y), y sean g y h dos funciones reales monótonas (crecientes ó decrecientes), entonces $\tau\left[g(X), h(Y)\right] = \tau(X, Y)$. La demostración de este resultado se puede ver en [18].

2.5.2. 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.

2.6. Cópulas

Una cópula es una función multivariada que describe la asociación entre las variables de una distribución conjunta. En [3] 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.

Nelsen en [28] 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 $H(x, y) = P[X \leq x, Y \leq y]$ la distribución conjunta.

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 en el intervalo $[0, 1]$, $H(x, y)$, 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

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

El método que se va a describir es siguiendo la notación en [3], con el enfoque de [5].

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 $\lfloor l_{1i}, u_{1i} \rfloor \times \lfloor l_{2i}, u_{2i} \rfloor (i = 1, \cdots, n)$.

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

$$S(t_1, t_2) = \breve{C}_\alpha(S_1(t_1), S_2(t_2)), \tag{2-15}$$

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 F 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, en donde el mayor interés se encuentra en la dependencia de las variables aleatorias. Existe una gran variedad de cópulas, entre ellas se encuentran la cópula de Clayton, la cópula Normal y la cópula de Plackett que han sido utilizadas en [24].

Cuando se supone un modelo con enfoque paramétrico para datos que presentan censura a intervalo, hay dificultad para elegir la distribución correcta. La respuesta se puede encontrar con el uso de una cópula. Las medidas de asociación como el rho de Spearman denotado por ρ_s y el τ de Kendall, se pueden expresar como función de una cópula de supervivencia de la siguiente manera:

$$\begin{aligned} \rho_s &= 12 \int_0^\infty \int_0^\infty F_1(t_1) \cdot F_2(t_2) dF(t_1, t_2) - 3 \\ &= 12 \int_0^1 \int_0^1 \breve{C}(u, v) du dv - 3, \end{aligned}$$

donde $dF(t_1, t_2) = f(t_1, t_2)dt_1dt_2$, con $f = \partial F/\partial t_1 \partial t_2$, cuando la función F es diferenciable.

El τ de Kendall se define como:

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

Para varias cópulas, se puede establecer la relación entre el parámetro de la cópula y la medida de asociación. Se considerarán las cópulas de Clayton, la Normal y la de Plackett.

2.6.1. Cópula de Clayton

Para $\theta_L > 0$, con $\theta_L \neq 1$ y $\alpha = \theta_L$ el parámetro de la cópula Clayton, la cópula de Clayton se define en [3] como:

$$\breve{C}^C_{\theta_L}(u, v) = (u^{1-\theta_L} + v^{1-\theta_L} - 1)^{\frac{1}{1-\theta_L}}. \tag{2-16}$$

También se conoce como la familia Pareto de cópulas.

El modelo Clayton asume " a constant local cross ratio", evaluando el grado de dependencia en un solo punto de tiempo, se define en [5] de la siguiente manera:

$$\theta_L(t_1, t_2) = S(t_1, t_2) \times \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} \Bigg/ \left\{ \frac{\partial S(t_1, t_2)}{\partial t_1} \times \frac{\partial S(t_1, t_2)}{\partial t_2} \right\} \tag{2-17}$$

"The cross ratio function" tiene una interpretación muy natural en las tasas de riesgo condicionales como en [30] , a saber:

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

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

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

2.6.2. Cópula Normal

Los datos bivariados que se distribuyen normales producen la cópula Normal o Gaussiana, con $\alpha = \rho$ el parámetro de la cópula Normal, dada en [5] por:

$$\breve{C}_{\rho}^{G}(u,v) = \Phi_{\rho}[\Phi^{-1}(u), \Phi^{-1}(v)], \tag{2-18}$$

donde Φ_ρ denota la función de distribución Normal bivariada estándar con correlación ρ. La cópula Normal 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:

$$\breve{C}_{\theta_L}^{C}(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \times \exp\left\{-\frac{(s_1^2 - 2\rho s_1 s_2 + s_2^2)}{2(1-\rho^2)}\right\} ds_1 ds_2. \tag{2-19}$$

La cópula Normal, 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.

2.6.3. Cópula de Plackett

La familia de cópulas de Plackett está definida en [5] para $\alpha = \theta_G$ el parámetro de la cópula Plackett con $\theta_G > 0$ con $\theta_G \neq 1$ de la siguiente manera:

$$\breve{C}_{\theta_G}^{P}(u,v) = \frac{[1 + (\theta_G - 1)(u+v)] - \sqrt{\{1 + (\theta_G - 1)(u+v)\}^2 - 4uv\theta_G(\theta_G - 1)}}{2(\theta_G - 1)}, \tag{2-20}$$

El modelo de Plackett asume "a constant global cross ratio function", definida en [3] como:

$$\theta_G(t_1, t_2) = \frac{S(t_1,t_2)[1 - S_1(t_1) - S_2(t_2) + S(t_1,t_2)]}{[S_1(t_1) - S(t_1,t_2)][S_2(t_2) - S(t_1,t_2)]}. \tag{2-21}$$

Para "a constant global cross ratio function", el espacio bivariado se divide en cada ubicación (t_1, t_2) en 4 cuadrantes produciendo cuatro probabilidades, una por cada cuadrante y $\theta_G(t_1; t_2)$.

Para $\theta_G = 1$ se tiene que la cópula está dada por:

$$\breve{C}_{1}^{P}(u,v) = \lim_{\theta_G \to 1} \breve{C}_{\theta_G}^{P}(u,v) = uv. \tag{2-22}$$

Para la familia de cópulas Plackett, no hay ninguna expresión de forma cerrada para el coeficiente de concordancia τ de Kendall.

2.6.4. Cópula Gumbel

La función de confiabilidad bivariada perteneciente a la familia Gumbel (ver [15]) tiene la siguiente forma:

$$C_\alpha(u,v) = \exp\{-[(-\ln u)^{1/\alpha} + (-\ln v)^{1/\alpha}]\}^\alpha, \tag{2-23}$$

donde $0 < \alpha < 1$.

Sean T_1 y T_2, tiempos de falla Weibull. Una función de confiabilidad conjunta para la Weibull bivariada definida en [25] es:

$$S(t_1,t_2) = \exp\left\{-\left[\left(\frac{t_1}{\theta_1}\right)^{\frac{\beta_1}{\alpha}} + \left(\frac{t_2}{\theta_2}\right)^{\frac{\beta_2}{\alpha}}\right]^\alpha\right\}, \tag{2-24}$$

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

$$S_k(t) = \exp\left\{-\left(\frac{t}{\theta_k}\right)^{\beta_k}\right\}, \ k = 1,2, \ t > 0 \tag{2-25}$$

Cuando $\alpha = 1$ entre los tiempos de falla Weibull, entonces hay independencia entre T_1 y T_2.

Si se compara la ecuación (2-24) con la ecuación (2-23), 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.

2.6.5. Modelo de mezcla gaussiano penalizado

Esta sección es importante porque a través de esta metodología se estiman las densidades de los errores para el método de cópulas.

Para conjuntos de datos bivariados siguiendo la notación de [3], $g(y_1, y_2)$ 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 [9].

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

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} w_{k_1,k_2}\ \mathcal{N}_2(\boldsymbol{\mu}_{k_1,k_2},\ \boldsymbol{\Sigma}), \tag{2-26}$$

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

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

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} $(k_1 = 1, \cdots,\ K_1; k_2 = 1, \cdots, K_2)$ 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 $\mathbf{a} = (a_{k_1,k_2} : k_1 = 1, \cdots, K_1;\ k_2 = 1, \cdots, K_2)'$ 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 [9]), está dado por:

$$q(\mathbf{a}; \boldsymbol{\lambda}) = \frac{\lambda_1}{2} \sum_{k_1=1}^{K_1} \sum_{k_2=1+s}^{K_2} (\Delta_1^s a_{k_1,k_2})^2 + \frac{\lambda_2}{2} \sum_{k_1=1}^{K_1} \sum_{k_2=1+s}^{K_2} (\Delta_2^s a_{k_1,k_2})^2, \tag{2-27}$$

el cual evita un sobreajuste. En la ecuación (2-27) el vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)'$ con $\lambda_1 > 0$ y $\lambda_2 > 0$ 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 $\widehat{a}_{k_1,k_2}$ $(k_1 = 1, \cdots, K_1;\ k_2 = 1, \cdots, K_2)$. 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 [12].

2.6.6. Estimación del τ de Kendall y rho de Spearman

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

Spearman.

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 bivariada. Las cópulas son muy apropiadas para modelar distribuciones bivariadas y la asociación entre las marginales se mide a través del τ de Kendall. 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. La correlación de Spearman es una medida de asociación global no paramétrica, la cual es invariante a las transformaciones monótonas marginales. En esta medida, los datos pueden ser observaciones no numéricas que se producen en n pares de observaciones las cuales se pueden comparar.

Cuando se supone un modelo con enfoque paramétrico para datos que presentan censura a intervalo, hay dificultad para elegir la distribución correcta. La respuesta se puede encontrar con el uso de una cópula.

Para las diferentes cópulas descritas en la Sección 2.6, se puede establecer una relación explícita entre el parámetro de la cópula y la medida de asociación. Como en [3], para la cópula de Clayton, el τ de Kendall y el parámetro θ se relacionan como $\tau = \theta/(\theta+2)$. Teniendo en cuenta que solo las correlaciones positivas se pueden modelar con la cópula de Clayton.

Para la cópula de Plackett y el parámetro $\theta > 0$, el rho de Spearman está dado por:

$$\rho_S(\theta) = \frac{\theta+1}{\theta-1} - \frac{2\theta}{(\theta-1)^2}\log(\theta).$$

El τ de Kendall no tiene una forma cerrada.

Para la cópula Gaussiana, la correlación ρ de Pearson es el parámetro de la cópula. El coeficiente de correlación de Spearman y el τ de Kendall están dados por: $\rho_S = 6/\pi \cdot \arcsin(1/(2\rho))$ y $\tau = (2/\pi) \cdot \arcsin(\rho)$ respectivamente.

Para el modelo en la Sección 2.6.5, [4] estima las medidas de asociación con este modelo. Su enfoque consiste en reemplazar los funcionales por sus contrapartes estimadas determinadas a partir de la función suavizada bivariada en la versión poblacional de la medida de asociación.

2.7. Modelo de tiempo de falla acelerado

El modelo de tiempo de falla acelerado es usado para modelar las distribuciones marginales de una cópula de supervivencia, ya sea la Normal, la de Clayton o la de Plackett, a través del paquete de *R* **smoothSurv**, en la parte de la simulación, se supone que las distribuciones marginales de la cópula Normal siguen un modelo de falla acelerado, por lo que es necesario mostrar algunos aspectos teóricos de este modelo.

Siguiendo la notación en [3] el modelo de tiempo de falla acelerado (AFT) está dado por:

$$\log(T) = \mathbf{X}'\boldsymbol{\beta} + \epsilon, \tag{2-28}$$

con T el tiempo de supervivencia, $\mathbf{X}$ un vector de covariables, $\boldsymbol{\beta}$ el vector de parámetros de regresión y ϵ una variable aleatoria del error. Sean $\hbar_0$ y S_0 la función de riesgo y supervivencia base, respectivamente, de la variable aleatoria $T_0 = \exp(\epsilon)$. Para un sujeto con vector de covariables $\mathbf{X}$, se asume que la función de riesgo y supervivencia son:

$$\hbar(t|\mathbf{X}) = \hbar_0[\exp(-\mathbf{X}'\boldsymbol{\beta})t]\exp(-\mathbf{X}'\boldsymbol{\beta}), \tag{2-29}$$

y

$$S(t|\mathbf{X}) = S_0[\exp(-\mathbf{X}'\boldsymbol{\beta})t]. \tag{2-30}$$

Por tanto,

$$T = \exp(\mathbf{X}'\boldsymbol{\beta})T_0, \tag{2-31}$$

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

En [3] 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 $g(\epsilon)$. Los supuestos más comunes para $g(\epsilon)$ 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 verosimimilitud.

Cuando sólo se utiliza una covariable categórica en el modelo, la curva de Kaplan-Meier se puede calcular para los sujetos de cada categoría por separado. Las curvas 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 $\mathbf{X}'\boldsymbol{\beta}$. La comparación de las curvas de Kaplan-Meier con las curvas ajustadas en cada grupo proporciona nuevamente una indicación de bondad de ajuste.

De las ecuaciones (2-28) y (2-31) se puede ver que el modelo AFT es de hecho un modelo de regresión lineal estándar con un tiempo de supervivencia logarítmico transformado.

3. Extensiones del Coeficiente de Asociación τ de Kendall

3.1. Estimación del τ de Kendall con datos con censura múltiple que surgen del ajuste de datos censurados individualmente

El coeficiente de correlación ρ de Pearson es una medida, que mide la fuerza de asociación lineal entre dos variables aleatorias X y Y, el cual se define como:

$$\rho = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}, \tag{3-1}$$

Dos métodos no paramétricos, son el coeficiente de rangos de Spearman y el coeficiente de concordancia τ de Kendall.

El enfoque que se va a describir a continuación es el de [29]. 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 se supone 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 $\sigma_{XY} = \rho\, \sigma_X \sigma_Y$.

La estimación de máxima verosimilitud de un vector de parámetros $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ donde el log de verosimilitud está dado por:

$$l(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = l(\boldsymbol{\theta}) = \sum_{i=1}^{n} G_i, \tag{3-2}$$

donde, G_i

$= \log[F(u_{1i}, u_{2i})]$ si X, Y están censurados a izquierda
$= \log[F(u_{1i}, u_{2i}) - F(l_{1i}, u_{2i})]$ si X está censurado a intervalo y Y a izquierda.
$= \log[F_Y(u_{2i}) - F(l_{1i}, u_{2i})]$ si X está censurado a derecha y Y a izquierda.
$= \log[F(u_{1i}, u_{2i}) - F(u_{1i}, l_{2i})]$ si X está censurado a izquierda y Y a intervalo.
$= \log[F(u_{1i}, u_{2i}) - F(l_{1i}, u_{2i}) - F(u_{1i}, l_{2i}) + F(l_{1i}, l_{2i})]$ si X,Y están censurados a intervalo.
$= \log[F_2(u_{2i}) - F(l_{1i}, u_{2i}) - F_Y(l_{2i}) + F(l_{1i}, l_{2i})]$ si X está censurado a derecha y Y a intervalo.
$= \log[F_X(u_{1i}) - F(u_{1i}, l_{2i})]$ si X está censurado a izquierda y Y a derecha.
$= \log[F_X(u_{1i}) - F_X(l_{1i}) - F(u_{1i}, l_{2i}) + F(l_{1i}, l_{2i})]$ si X está censurado a intervalo y Y a derecha.
$= \log[1 - F_X(l_{1i}) - F_Y(l_{2i}) + F(l_{1i}, l_{2i})]$ 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 el enfoque en [29], 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, bajo los supuestos $X \sim N(\mu_X, \sigma_X^2)$ y $Y \sim N(\mu_Y, \sigma_Y^2)$, las cuales presentan censura a derecha, a izquierda y a intervalo. Cuando se tienen las estimaciones para $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ con la verosimilitud en (3-2) se estima ρ, usando la relación de Greiner dada en [14] por $\tau = (2/\pi)\arcsin(\rho)$ se estima el τ de Kendall.

3.2. Método cópula

El método de cópula consiste en seleccionar una cópula de supervivencia, ya sea la cópula Normal, Clayton ó Plackett, luego se ajustan las distribuciones marginales, las cuales se modelan con el modelo de falla acelerado con un término de error flexible, como se describe en la Sección 2.7. 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 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 de Clayton y la cópula de Plackett, el parámetro de dependencia θ_L y θ_G respectivamente se modelan en la escala logarítmica, donde θ_L es el parámetro de la cópula Clayton y θ_G es el

parámetro de la cópula Plackett, es decir, $\log(\alpha_i) = \boldsymbol{\gamma}'\mathbf{x}_i$, con $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)'$ un vector p-dimensional de paramétros de la regresión desconocidos y con $\mathbf{x}_i = (x_{1,i}, \cdots, x_{p,i})'$ un vector p-dimensional de covariables asociadas sobre el i-ésimo sujeto. Para la cópula Normal, la dependencia se modela con la transformación de Fisher, es decir, $1/2\log[(1+\alpha_i)/(1-\alpha_i)] = \boldsymbol{\gamma}'\mathbf{x}_i$. Las distribuciones marginales también pueden depender de las covariables, que pueden ser diferentes al parámetro de la cópula. El vector de dimensión m, $\mathbf{z}_i = (z_{1,i}, \cdots, z_{m,i})'$ 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 (2-15) el log de la verosimilitud está dada por:

$$\log L(\boldsymbol{\gamma}, S_1, S_2|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} l(\boldsymbol{\gamma}, S_1, S_2|\mathbf{y_i}, \mathbf{x_i}, \mathbf{z_i}, \boldsymbol{\delta}) = \sum_{i=1}^{n} l_i, \tag{3-3}$$

donde $\mathbf{Y}, \mathbf{X}, \mathbf{Z}$ representan las matrices de los vectores $\mathbf{y_i}, \mathbf{x_i}$ y $\mathbf{z_i}$ respectivamente. Cada contribución de la verosimilitud individual puede escribirse como una suma de 9 términos diferentes dependiendo de si la observación tiene censura a izquierda, a intervalo o a derecha en ambas dimensiones, es decir,

$$\begin{aligned} l(\boldsymbol{\gamma}, S_1, S_2|\mathbf{X}, \boldsymbol{\delta}) = {} & \delta_1^{(1)}\delta_2^{(1)} \log S_{11}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + \delta_1^{(1)}\delta_2^{(2)} \log S_{12}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + \delta_1^{(1)}(1 - \delta_2^{(1)} - \delta_2^{(2)}) \log S_{13}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + \delta_1^{(2)}\delta_2^{(1)} \log S_{21}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + \delta_1^{(2)}\delta_2^{(2)} \log S_{22}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + \delta_1^{(2)}(1 - \delta_2^{(1)} - \delta_2^{(2)}) \log S_{23}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + (1 - \delta_1^{(1)} - \delta_1^{(2)})\delta_2^{(1)} \log S_{31}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + (1 - \delta_1^{(1)} - \delta_1^{(2)})\delta_2^{(2)} \log S_{32}(\boldsymbol{\gamma}|\mathbf{X}) \\ & + (1 - \delta_1^{(1)} - \delta_1^{(2)})(1 - \delta_2^{(1)} - \delta_2^{(2)}) \log S_{33}(\boldsymbol{\gamma}|\mathbf{X}), \end{aligned} \tag{3-4}$$

con,

$$
\begin{aligned}
S_{11}(\boldsymbol{\gamma}|\mathbf{X}) &= P(T_1 \leq l_1, T_2 \leq l_2) \\
&= 1 - S_1(l_1) - S_2(L_2) + \breve{C}_\alpha[S_1(l_1), S_2(l_2)] \\
S_{12}(\boldsymbol{\gamma}|\mathbf{X}) &= P(T_1 \leq l_1, l_2 < T_2 \leq u_2) \\
&= S_1(l_1) - S_1(u_1) \\
&\quad + \breve{C}_\alpha[S_1(l_1), S_2(u_2)] - \breve{C}_\alpha[S_1(l_1), S_2(l_2)] \\
S_{13}(\boldsymbol{\gamma}|\mathbf{X}) &= P(T_1 \leq l_1, T_2 > u_2) \\
&= S_2(u_2) - \breve{C}_\alpha(S_1(l_1), S_2(u_2)) \\
S_{21}(\boldsymbol{\gamma}|\mathbf{X}) &= P(l_1 < T_1 \leq u_1, T_2 \leq l_2)
\end{aligned}
$$

$$
\begin{aligned}
S_{22}(\boldsymbol{\gamma}|\mathbf{X}) &= P(l_1 < T_1 \leq u_1, l_2 < T_2 \leq u_2) \\
&= \breve{C}_\alpha[S_1(l_1), S_2(l_2)] - \breve{C}_\alpha[S_1(l_1), S_2(u_2)] \\
&\quad - \breve{C}_\alpha[S_1(u_1), S_2(l_2)] + \breve{C}_\alpha[S_1(u_1), S_2(u_2)] \\
S_{23}(\boldsymbol{\gamma}|\mathbf{X}) &= P(l_1 < T_1 \leq u_1, T_2 > u_2) \\
&= \breve{C}_\alpha[S_1(l_1), S_2(u_2)] - \breve{C}_\alpha[S_1(u_1), S_2(u_2)] \\
S_{31}(\boldsymbol{\gamma}|\mathbf{X}) &= P(T_1 > u_1, T_2 \leq l_2) \\
&= S_1(u_1) - \breve{C}_\alpha[S_1(u_1), S_2(u_2)] \\
S_{32}(\boldsymbol{\gamma}|\mathbf{X}) &= P(T_1 > u_1, l_2 < T_2 \leq u_2) \\
&= \breve{C}_\alpha[S_1(u_1), S_2(l_2)] - \breve{C}_\alpha[S_1(u_1), S_2(u_2)] \\
S_{33}(\boldsymbol{\gamma}|\mathbf{X}) &= P(T_1 > u_1, T_2 > u_2) \\
&= \breve{C}_\alpha[S_1(u_1), S_2(u_2)]
\end{aligned}
$$

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 de (3-3). La estimación de la máxima verosimilitud completa puede resultar en cálculos bastante largos, sin embargo, en [3] se propone un procedimiento de dos etapas basado en la pseudo verosimilitud en forma parámetrica, en [5] siguen este procedimiento. En el procedimiento de dos etapas primero se estiman S_1 y S_2 y se imputan las estimaciones $\widehat{S}_1$ y $\widehat{S}_2$ en (3-3). Luego, se estima $\boldsymbol{\gamma}$ maximizando la pseudo verosimilitud $l(\boldsymbol{\gamma}, \widehat{S}_1, \widehat{S}_2)$, obtenida de la ecuación (3-4) 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 [23], este enfoque permite la incorporación de covariables en las distribuciones marginales. Formalmente se estiman S_k, para $k = 1, 2$ de la siguiente expresión:

$$\log(T_{k,i}) = \boldsymbol{\beta}_{\boldsymbol{k}}^{'}\mathbf{z}_i + \epsilon_{k,i}, \;\; i = 1, \cdots n, \tag{3-5}$$

donde $\boldsymbol{\beta}_k = (\beta_{k,1}, \cdots, \beta_{k,m})'$ es un vector m- dimensional de parámetros de la regresión desconocidos y $\epsilon_{k,1}, \cdots, \epsilon_{k,n}$ son variables de error aleatorias independientes e idénticamente distribuidas con densidad $g_{\epsilon_k}(\epsilon_k)$. La densidad $g_{\epsilon_k}(\epsilon_k)$ de el término de error se expresa utilizando la mezcla Normal penalizada, es decir,

$$g_{\epsilon_k}(\epsilon_k) = \zeta_k^{-1} \sum_{j=1}^{K_k} c_{k,j}(a_k) \phi \left(\frac{\epsilon_k - \eta_k}{\zeta_k} \middle| \mu_{k,j}, \sigma_{k,0}^2 \right) \tag{3-6}$$

donde $\mu_{k,1}, \cdots, \mu_{k,K_k}$ es un conjunto de knots equidistantes fijos, $\sigma_{k,0}$ es una base fija de desviación estándar, η_k un intercepto desconocido, ζ_k un parámetro de escala desconocido y $\phi(.|\mu, \sigma^2)$ una densidad Normal con media μ y desviación estándar σ. Finalmente sean $\mathbf{c}_k = (c_{k,1}, \cdots, c_{k,K_k})'$ los pesos de la mezcla desconocidos y $\mathbf{a}_k = (a_{k,1}, \cdots, a_{k,K_k})'$ sus transformaciones obtenidas usando la siguiente ecuación:

$$c_{k,j}(a_k) = \frac{\exp(a_{k,j})}{\sum_{p=1}^{K_k} \exp(a_{k,p})}, \ -\infty < a_{k,j} < \infty \text{ y } j = 1, \cdots, K_k. \tag{3-7}$$

Al introducir los parámetros $\mathbf{a}_k$, el problema se puede cambiar de un problema de máxima verosimilitud restringido a uno no restringido. Para facilitar la notación, los autores en [5] asumen que $a_{k,[K_k/2]=0}$, donde [.] es una función de techo. Los parámetros $\boldsymbol{\beta}_k$ y $\mathbf{a}_k$ 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 $\mathbf{a}_k$.

Dado un parámetro de suavizado λ en [5], la verosimilitud penalizada está representada de la siguiente manera:

$$l_{P,n} = l_n - \frac{\lambda}{2} \sum_{j=s+1}^{K_k} \left(\Delta^s a_{k,j} \right)^2, \tag{3-8}$$

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 [33] proponen un procedimiento a través de bootstrap. Para M fijo, se producen M estimadores $\widehat{\gamma_m}$; $m = 1, \cdots, M$ de $\boldsymbol{\gamma}$. La varianza de $\widehat{\boldsymbol{\gamma}}$ puede ser estimada por la varianza muestral de los $\widehat{\gamma_m}$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 ó 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 del software estadístico R **censcor**, el cual está descrito en el apéndice A.1.

En [5] se describe un método para modelar las distribuciones marginales con un modelo de falla acelerado con término de error flexible sugerido en [23] en combinación con el parámetro de una cópula, utilizando el paquete de R **smoothSurv**, utilizando la ecuación (3-5). Para los datos bivariados generados se estima el coeficiente de asociación τ de Kendall por medio de la cópula de supervivencia Normal. Las distribuciones marginales de la cópula Normal en este estudio de simulación, son modeladas con el modelo de falla acelerado, sin dependencia de covariables en las distribuciones marginales.

Para la simulación del método cópula se utiliza el enfoque descrito en [5] en el que implementan la función `fit.copula` que está disponible en el paquete **icensBKL** del software estadístico R.

4.1. Esquema de simulación

En el esquema del estudio de simulación se generaron los datos de una cópula Gumbel con la función del paquete estadístico R **BiCopsim** de la librería CDVine, con la finalidad de generar datos bivariados de una distribución Weibull, se desea emplear una distribución Weibull porque 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, las cuales son 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 quantil 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 $\tau = 1 - 1/\theta$. 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 ajustandolas individualmente. La misma base de datos se

usó para estimar el τ de Kendall usando una cópula Normal.

Para la estimación del τ de Kendall con el método de ajuste individual de las marginales, el paquete **censcor** proporciona rutinas para estimar el coeficiente de correlación ρ de Pearson entre dos variables censuradas, en donde se tiene en cuenta el porcentaje de censura de los datos simulados, las variables aleatorias X y Y se asumen que tienen una distribución $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$. Una vez que se tenga la estimación para ρ se emplea la relación de Greiner, para estimar el τ de Kendall, de las 500 muestras se obtuvieron 500 valores del $\widehat{\tau}$, con los cuales se calcularon el MAD y ECM.

Para la estimación a través de la cópula Normal se tuvo en cuenta lo siguiente, sus distribuciones marginales se modelan con el modelo de falla acelerado, sin covariables, el modelo resultante de la ecuación (3-5) que está dado por $\log(T_{k,i}) = \epsilon_{k,i},\ i = 1, \cdots, n$, el proceso de simulación en sus iteraciones muestra la búsqueda de los valores óptimos λ_1 y λ_2 de las respectivas distribuciones marginales, definidos a partir de la ecuación (3-8), que se encuentran en la rejilla de $\exp(n)$ para $n = -3, -2, -1, 1, 2, 3$ arrojando los valores AIC más pequeños para ambas distribuciones, el proceso provee el valor máximo de la pseudo- verosimilitud $l(\boldsymbol{\gamma}, \widehat{S}_1, \widehat{S}_2)$ que se obtuvo para la cópula Normal, a través de la función de R `optim`, con criterios de parada por defecto, en donde se estima $\boldsymbol{\gamma}$ maximizando la pseudo verosimilitud, obtenida de la ecuación (3-4) conectando las estimaciones $\widehat{S}_1$ y $\widehat{S}_2$ y se obtiene el parámetro estimado de la cópula Normal que corresponde al coeficiente de correlación de Pearson ρ, para estimar el τ de Kendall se emplea la relación de Greiner dada por: $\tau = (2/\pi)\arcsin(\rho)$, como se generaron 500 muestras se obtuvieron 500 valores del τ de Kendall, con los cuales se calcularon el MAD y ECM.

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 MAD y el error cuadrático medio para cada conjunto de datos bivariados, teniendo en cuenta el enfoque del ajuste individual de las marginales y el de cópulas.

3. Se tienen en cuenta $M = 500$ réplicas, para cada escenario de simulación.

4.2. Resultados

n	τ de Kendall	MAD con ajuste de las marginales	MAD Cópula	ECM con ajuste de las marginales	ECM Cópula	Eficiencia Relativa $\frac{\text{ECM}_1(\widehat{\tau})}{\text{ECM}_2(\widehat{\tau})}$
50	0.2	0.1210	0.0965	0.0201	0.0330	0.6090
	0.5	0.3140	0.3196	0.1027	0.1295	0.7930
	0.8	0.4350	0.2000	0.1848	0.0400	4.6200
100	0.2	0.1219	0.0771	0.0172	0.0105	1.6380
	0.5	0.2857	0.2290	0.0860	0.1055	0.8151
	0.8	0.4065	0.2000	0.1746	0.0400	4.365
200	0.2	0.1234	0.0518	0.0160	0.0057	2.8070
	0.5	0.2947	0.2535	0.0873	0.0986	0.8853
	0.8	0.3985	0.2000	0.1674	0.0400	4.185

Tabla 4-1.: Resultados del Estudio de Simulación.

La Tabla 4-1 da los valores de la desviación mediana absoluta (MAD), el error cuadrático medio (ECM), para el método que usa un ajuste individual de las marginales y el método de cópulas, para cada combinación de parámetros n y de τ, además da la eficiencia relativa, donde ECM_1 hace referencia a la estimación con el ajuste individual de las marginales y ECM_2 es la estimación con la cópula Normal. Para la estimación del MAD por medio de la cópula Normal para el valor de $\tau = 0.8$ se tomó el mínimo valor entre el valor estimado y 1, debido a que este método tiende a sobreestimar el parámetro, por esta razón en los tres tamaños muestrales con este valor de τ, el test dió el valor de 1, en la mayoría de las estimaciones.

De acuerdo a la eficiencia relativa, en general, cuando el tamaño de muestra es pequeño ($n = 50$) y hay una dependencia entre baja y media ($\tau = 0.2, 0.5$), el método de ajuste individual de las marginales es mejor que el método de cópula Normal. Para un tamaño de muestra más grande $n = 100, 200$ y con una dependencia baja y alta $\tau = 0.2$ y 0.8 el mejor método es el de cópula Normal. Con dependencia media $\tau = 0.5$ el mejor método es el de la estimación con el ajuste individual de las marginales.

De los resultados de la Tabla 4-1 y a partir de las gráficas que se presentan a continuación se puede apreciar mejor el comportamiento de los escenarios analizados:

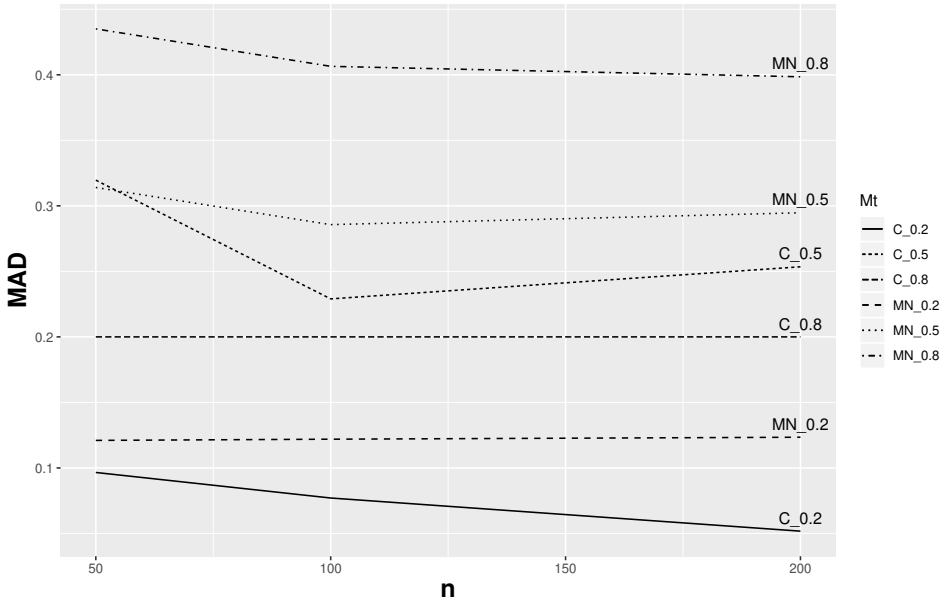

Figura 4-1.: Relación entre el MAD y el tamaño muestral, con el método de ajuste individual de las marginales y el método de cópula Normal, la notación CM_0.2 hace referencia al método con ajuste individual de las marginales con el valor del τ de Kendall de 0.2 y C_0.2 hace referencia al método de cópula Normal con $\tau = 0.2$

.

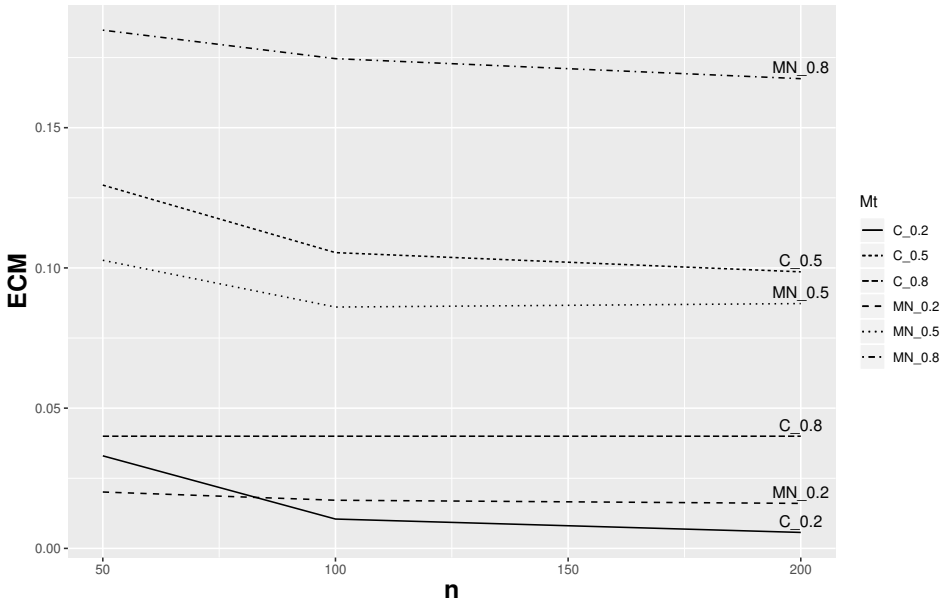

Figura 4-2.: Relación entre el ECM y el tamaño muestral, con el método con ajuste individual de las marginales y el método de cópula Normal

En general, en las Figuras 4-1 y 4-2, se puede ver que a medida que el tamaño de muestra n aumenta las estimaciones para el MAD y el ECM disminuyen, excepto cuando $\tau = 0.5$. Cuando los valores del coeficiente de concordancia τ de Kendall disminuyen, las estimaciones son más precisas para el método con ajuste individual de las marginales.

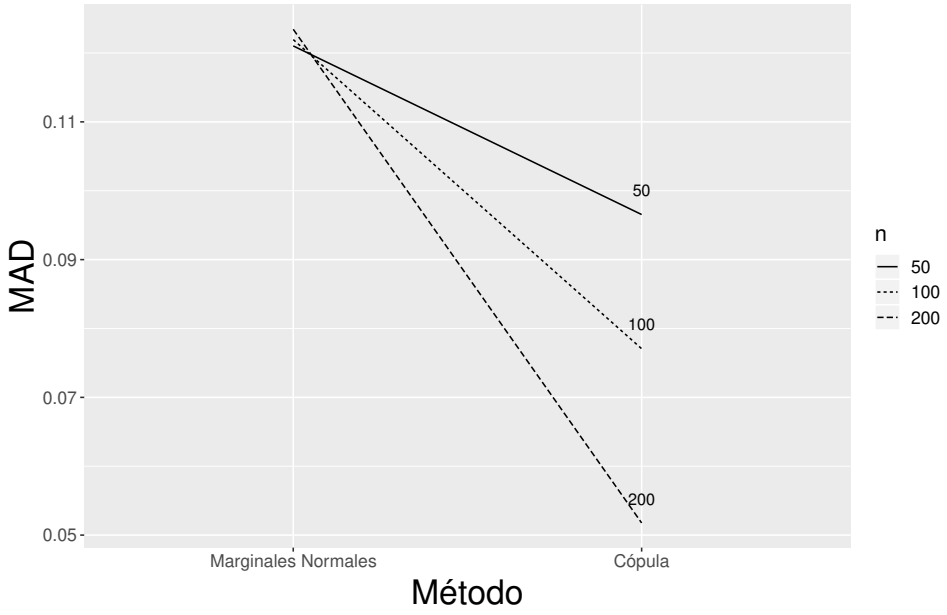

Figura 4-3.: Relación del MAD con los dos métodos, con $\tau = 0.2$

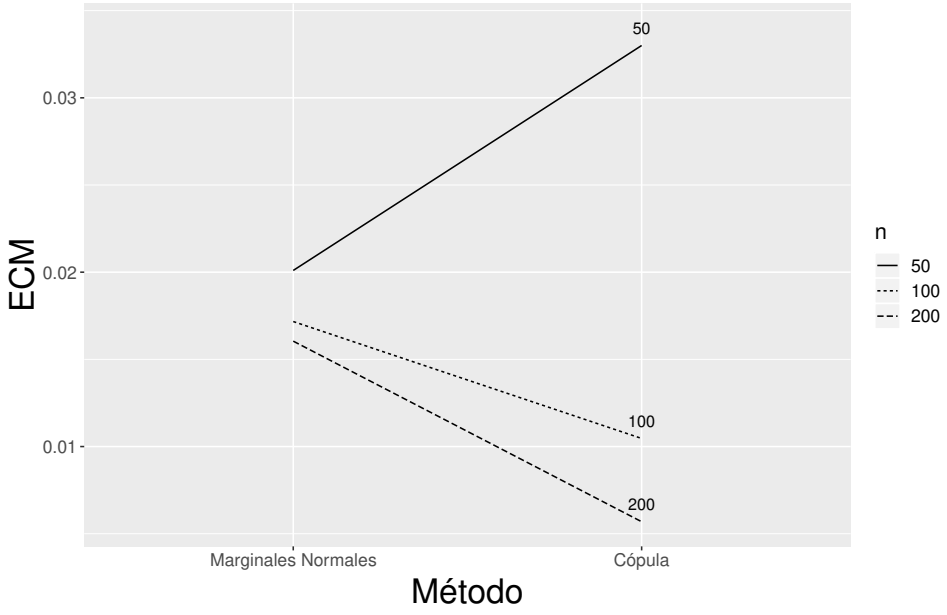

Figura 4-4.: Relación del ECM con los dos métodos, con $\tau = 0.2$

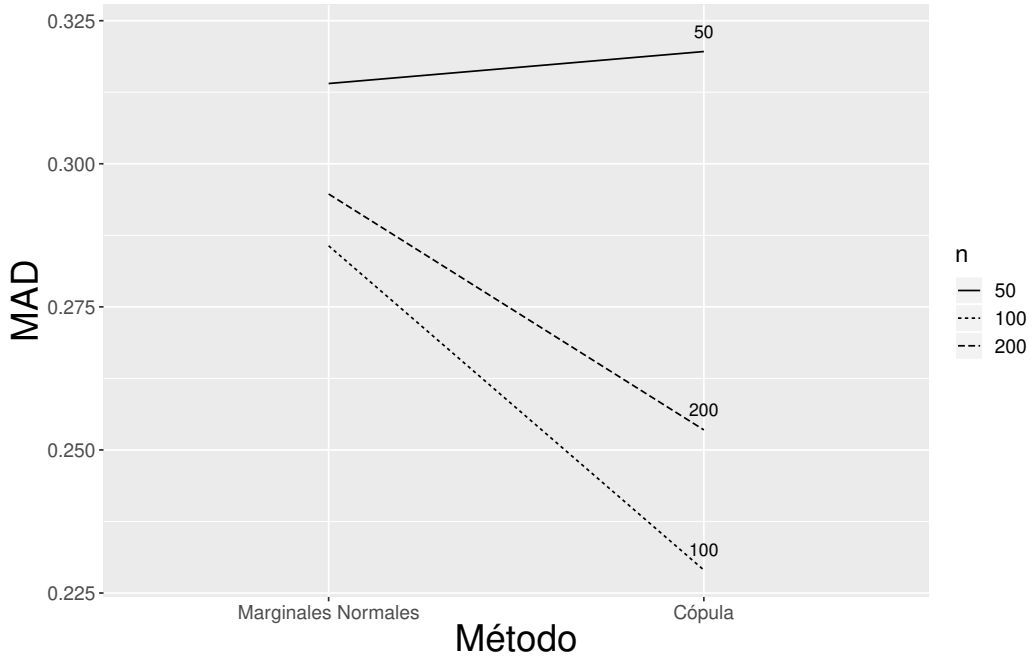

Figura 4-5.: Relación del MAD con los dos métodos, con $\tau = 0.5$

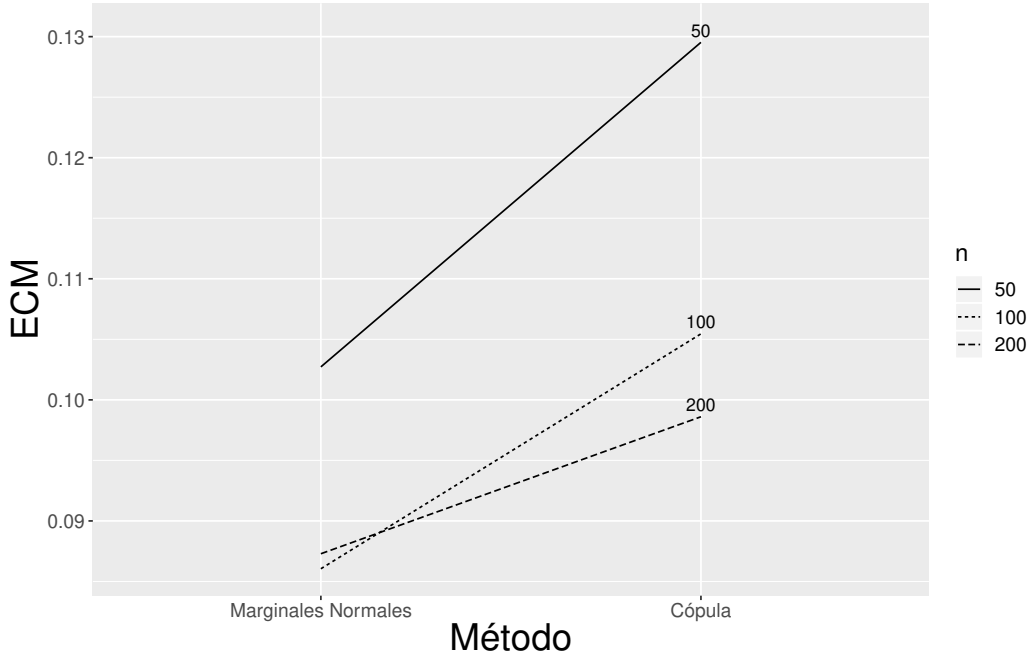

Figura 4-6.: Relación del ECM con los dos métodos, con $\tau = 0.5$

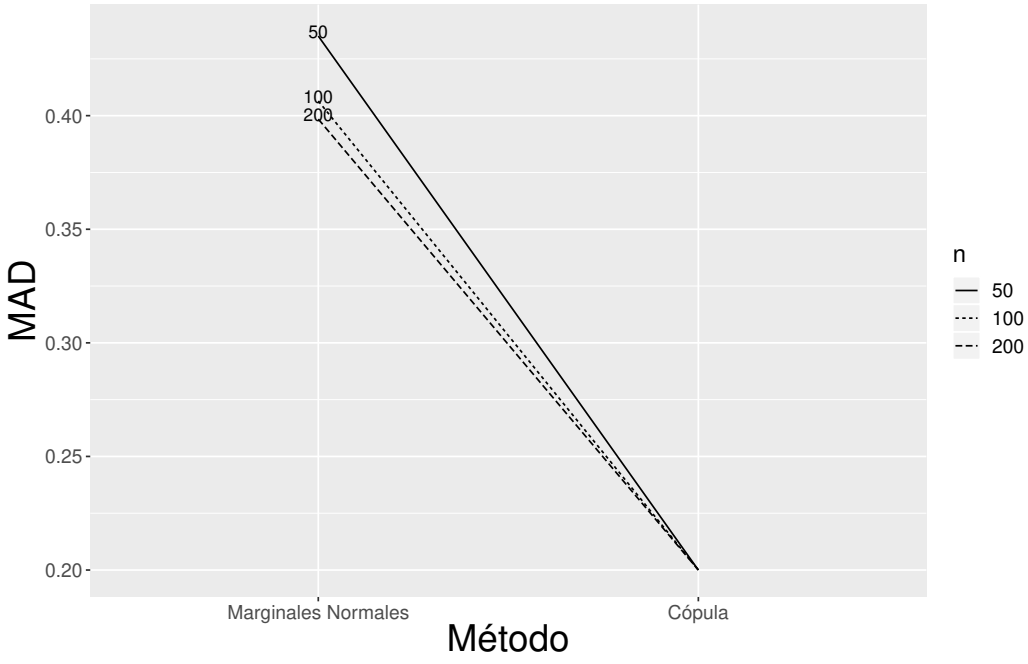

Figura 4-7.: Relación del MAD con los dos métodos, con $\tau = 0.8$

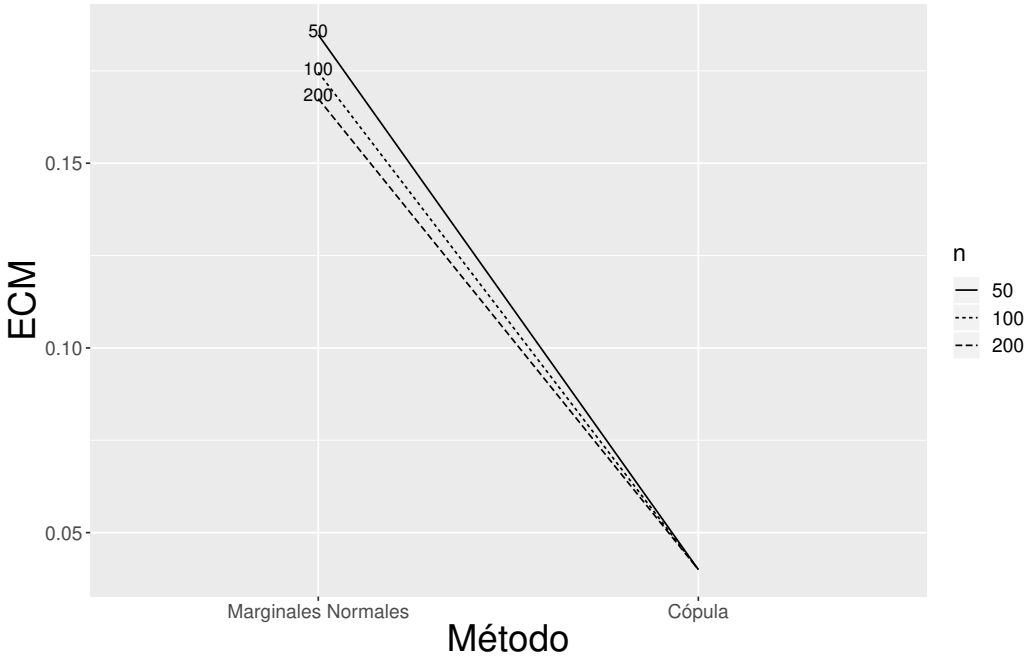

Figura 4-8.: Relación del ECM con los dos métodos, con $\tau = 0.8$

Se puede apreciar en las Figuras **4-3**, **4-5** y **4-7** que el MAD tiende a ser menor en el método de cópula Normal, que en el método con ajuste individual de las marginales, para cualquier τ

y su comportamiento en más predecible a medida que el valor del τ crece. El ECM es menor en el método de cópula Normal, que en el método con datos con ajuste individual de las marginales, cuando hay alta dependencia ($\tau = 0.8$), excepto cuando el tamaño de muestra es pequeño ($n = 50$).

Con dependencia media ($\tau = 0.5$) el comportamiento del MAD y ECM es muy irregular.

5. Conclusiones y Trabajo Futuro

5.1. Conclusiones

El propósito de este trabajo de investigación fue comparar dos métodos de estimación del coeficiente de concordancia τ de Kendall, uno que tiene un enfoque dirigido al ajuste individual de las marginales y otro basado en cópulas, en el que se utilizó la cópula Normal o Gaussiana. Los datos simulados tienen cierto porcentaje de censura a derecha, a izquierda y a intervalo, provenientes de una cópula Gumbel. De acuerdo a los métodos descritos en capítulos anteriores y al proceso de simulación se llegan a las siguientes conclusiones:

- La estimación del τ de Kendall por medio de la cópula Normal, a través del paquete **icensBKL**, es en general mejor que la estimación del τ de Kendall con el método con ajuste individual de las marginales usando el paquete **censcor** de R, cuando hay alta dependencia, ya que proporciona valores de MAD y ECM más bajos. En este caso, para estimar el τ de Kendall con datos con censura a intervalo se recomienda usar el método cópula Normal.
- A medida que n crece se obtiene un MAD y un ECM más bajos, lo que indica que los estimadores son más precisos.
- En el método con ajuste individual de las marginales, a medida que el τ crece los valores de MAD y ECM aumentan, lo que indica que los estimadores pierden precisión, cuando los datos son dependientes.
- Los dos métodos estudiados en este trabajo proporcionan herramientas útiles para la estimación del coeficiente de concordancia τ de Kendall, siendo el más preciso, según el MAD, por sus estimaciones más bajas el método basado en cópulas. Cuando el tamaño de muestra es pequeño, para $n = 50$ y hay una dependencia entre baja y media $\tau = 0.2$ y 0.5, el método de ajuste individual de las marginales es mejor que el método cópula Normal. Para un tamaño de muestra más grande $n = 100$ y 200 y con una dependencia baja y alta $\tau = 0.2$ y 0.8 el mejor método es el de cópula Normal.

5.2. Trabajo futuro

Como trabajo futuro, en el método de cópulas se sugiere la dependencia de covariables en las distribuciones marginales, con el objetivo de mejorar las estimaciones del τ de Kendall.

A. Paquetes del Software Estadístico R

A.1. censcor

Este paquete funciona ajustando una distribución Normal multivariada a los datos, por medio de la función stan que proporciona rutinas para realizar una estimación de la correlación entre dos variables censuradas, que utiliza el algoritmo Hamiltoniano de Monte-Carlo de No-U-Turn para dibujar desde la distribución a posteriori.

El paquete toma como argumentos una fórmula que describe la correlación a estimar y un data frame, el data frame se puede crear con la función generate_censored_data en la cual se puede escoger el tamaño de muestra, el valor y el coeficiente de correlación con el que se va a trabajar, ya sea el Pearson, el de Spearman o con el coeficiente de concordancia τ de Kendall y el porcentaje de censura con el que se desea trabajar y los indicadores de censura de $-1, 0, 1$ y 2 que representan la censura a izquierda, nada censura, a derecha y en intervalos respectivamente. Al aplicar la censura al conjunto se datos, se obtienen las estimaciones de la media, la desviación estándar, el coeficiente de correlación y los cuartiles, junto con una medida aproximada del tamaño efectivo de la muestra y los factores de reducción de las escala potencial en cadenas divididas, en donde la convergencia para esta estimación es igual a 1.

A.2. icensBKL

Para modelar el método de cópula se emplea el paquete **icensBKL**, el cual emplea algunas funciones como `fit.copula` que ajusta los modelos de cópula a datos con censura en intervalos, el objetivo de la función es fijar una cópula de supervivencia, ya sea la cópula de Clayton, la cópula Normal ó la cópula de Plackett, mediante el procedimiento en dos etapas. La función utiliza del paquete la función `smoothSurv` para ajustar las distribuciones marginales, las cuales se modelan con el modelo de falla acelerada con un término de error flexible, los argumentos tienen escala logarítmica. Los posibles valores de los vectores λ_1 y λ_2 se encuentran en la rejilla $\exp(n)$ para $n = -3, -2, \cdots, 2, 3$, no se emplearon covariables para el estudio de simulación. La salida del código muestra el valor óptimo entre las dos marginales por medio del procedimiento de la pseudo- verosimilitud.

La pseudoverosimilitud para la cópula fue maximizada utilizando la función de *R* `optim`, a través del método L-BFGS-B que es un método cuasi-Newton de memoria limitada para la optimización restringida, con criterios de parada por defecto.

El parámetro de la cópula también puede depender de las covariables.

La función toma como argumento el conjunto de datos, en el cual se interpretan las variables. Para la cópula de Clayton y Plackett, la dependencia se modelará en la escala logarítmica. Para la cópula Normal, la dependencia será modelada bajo la transformación de Fisher. Para las distribuciones marginales se emplea una expresión para otros modelos de regresión que se utilizarán con la función `smoothSurvReg` para fijar las dos marginales. El argumento `control1` y `control2` determinan la configuración de `smoothSurv` de las distribuciones marginales

La opción de icsurv, el objeto representa el estimador de máxima verosimilitud no paramétrico (NPMLE) de la función de supervivencia basado en una muestra agrupada.

B. Código en el Software Estadístico R Para el Proceso de Simulación

A continuación se presenta el código para la comparación del coeficiente de concordancia τ de Kendall por medio del método del ajuste individual de las marginales y el método de cópulas, teniendo en cuenta la combinación de los escenarios para cada proceso de simulación, se presenta el caso para el valor del $\tau = 0.2$, con un tamaño de muestra $n = 50$, los otros casos se hacen de forma análoga:

```
Tau<-0.2
n<-50
s<-500 # número de bases de datos
porcen<-0.7 # porcentaje de censura a derecha de 0.1
pizq<-0.3 # porcentaje de censura a izquierda

teta<-1/(1-Tau) #parámetro de la cópula Gumbel

library(CDVine)
library(censcor)

generador <- function(n){

t1=runif(n,0,1)
tv<-matrix(0,ncol=10,nrow=n)
for(k in 1:10){
tv[,k]=t1+k
}
t0<-rep(0,n)
tvisitas<-cbind(t0,t1,tv)
return(tvisitas)

}

intervalos<-function(data){
```

```
# tiempo de falla
tr<-data[1]
# tiempos de visitas reales
tobs<-data[-1]
# intervalos de censura
L<-max(ifelse(tobs<tr,tobs,min(tobs)))
R<-min(ifelse(tobs>=tr,tobs,max(tobs)))
# vector con tiempos de falla, intervalo e indicadora de censura
c(tr,L,R)
}

gendata<-function(s){
datos<-generador(n)
simdata <- BiCopSim(n,4,teta)
datos1<-log(1/(1-simdata))
print(datos1)
x<-datos1[,1]
y<-datos1[,2]
# matriz con tiempos de falla y tiempos de visitas
tmp<-cbind(x,datos)
tmp<-as.data.frame(t(apply(tmp,1,intervalos)))
tmp1<-cbind(y,datos)
tmp1<-as.data.frame(t(apply(tmp1,1,intervalos)))


L1<-ifelse(x<quantile(x,pizq),NA,(ifelse(x>quantile(x,porcen),quantile(x,porcen),
tmp[,2])))

L1[L1==0]<-NA

R1<-ifelse(x<quantile(x,pizq),quantile(x,pizq),(ifelse(x>quantile(x,porcen),NA,
tmp[,3])))

L2<-ifelse(y<quantile(y,pizq),NA,(ifelse(y>quantile(y,porcen),quantile(y,porcen),
tmp1[,2])))

L2[L2==0]<-NA


R2<-ifelse(y<quantile(y, pizq),quantile(y, pizq), (ifelse(y>quantile(y,porcen),NA,
```

```
tmp1[,3])))

xinf<-
ifelse(x<quantile(x, pizq),quantile(x,pizq),
(ifelse(x>quantile(x,porcen), quantile(x,porcen),tmp[,2])))

xsup<-ifelse(x<quantile(x,pizq),quantile(x,pizq),
(ifelse(x>quantile(x,porcen), quantile(x,porcen),tmp[,3])))

cen1<-ifelse(x<quantile(x, pizq),0,(ifelse(x>quantile(x,porcen),1,2)))

yinf<-
ifelse(y<quantile(y, pizq), quantile(y, pizq),
(ifelse(y>quantile(y,porcen), quantile(y,porcen),tmp1[,2])))

ysup<-ifelse(y<quantile(y,pizq),quantile(y, pizq),
(ifelse(y>quantile(y,porcen), quantile(y,porcen),tmp1[,3])))

cen2<-ifelse(y<quantile(y, pizq),0,(ifelse(y>quantile(y,porcen),1,2)))

tmp2<-as.data.frame(cbind(x,L1,R1,y,L2,R2,x,xinf,xsup,cen1,y,yinf,ysup,cen2))

names(tmp2)<-c("x","L1","R1","y","L2","R2","x","xinf","xsup",
"cen1","y","yinf","ysup","cen2")
return(tmp2)
}

#########################Modelo AFT ######################

library("icensBKL")

#fit normal copula
##################

############## CÓDIGO PARA LA ESTIMACIÓN POR MEDIO DE CÓPULAS #####

require(smoothSurv)
require(BB)

fit.mycopula <- function(data, copula = "normal", init.param = NULL, cov = ~1,
```

```
marginal1 = formula(data), logscale1 = ~1, lambda1 = exp(3:(-3)),
marginal2 = formula(data), logscale2 = ~1, lambda2 = exp(3:(-3)),
bootstrap = FALSE, nboot = 1000, control1 = smoothSurvReg.control(info = FALSE),
control2 = smoothSurvReg.control(info = FALSE), seed = 12345)
{
allowed.distributions = c("normal", "clayton", "plackett", "mycopula")
if(is.na(pmatch(copula, allowed.distributions))) {
stop("Wrong copula used. Only 'normal', 'clayton' or 'plackett' are allowed.")
} else {
dist = match.arg(copula, allowed.distributions)
}
Terms <- if(missing(data)) terms(cov) else terms(cov, data = data)
m <- match.call(expand.dots = FALSE)
m.keep <- m
temp <- c("", "formula", "data", "subset", "na.action")
m <- m[match(temp, names(m), nomatch = 0)]
m[[1]] <- as.name("model.frame")
m$formula <- Terms
m <- eval(m, parent.frame())
X <- model.matrix(Terms, m)
n <- nrow(X)
nvar <- ncol(X)
if(nvar <= 0) stop("Invalid design matrix. ")
fun.cop <- switch(dist, normal = normal.copula, clayton = clayton.copula,
plackett = plackett.copula)
fit1 <- try(smoothSurvReg(marginal1, logscale = logscale1,
lambda = lambda1, data = data, control = control1), TRUE)
fit2 <- try(smoothSurvReg(marginal2, logscale = logscale2,
lambda = lambda2, data = data, control = control2), TRUE)
if(inherits(fit1, "try-error")) fit1 <- list(fail = 99)
if(inherits(fit2, "try-error")) fit2 <- list(fail = 99)
if(fit1$fail < 99 && fit2$fail < 99){
if(is.null(init.param)) {tau <- cor.test(rowMeans(cbind(fit1$y[fit1$y[, 3] == 3
 & fit2$y[, 3] == 3, 1], fit1$y[fit1$y[, 3] == 3 & fit2$y[, 3] == 3, 2]),
  na.rm = TRUE), rowMeans(cbind(fit2$y[fit1$y[,3] == 3 & fit2$y[, 3] == 3, 1],
  fit2$y[fit1$y[,3] == 3
 & fit2$y[, 3] == 3, 2]), na.rm = TRUE), method = "kendall")$estimate
init.param <- switch(dist, normal = sin(pi * tau/2),
clayton = -2 * tau/(tau - 1),
plackett = as.numeric(BBsolve(par = c(0.5),
```

```
fn = function(x) -sin(pi * tau/2) + (x + 1)/(x - 1) - 2 * x * log(x)/(x - 1)^2)$par))

if(length(init.param) < nvar) {
init.param <- c(init.param, rep(0, nvar - length(init.param)))
}
} else {
if(nvar != length(init.param)) stop("Wrong number of initial values given.")
}
N <- dim(fit1$y)[1]
Sx <- survfitS.smoothSurvReg(fit1, fit1$x[, -1], fit1$z[,-1])
Sy <- survfitS.smoothSurvReg(fit2, fit2$x[, -1], fit2$z[,-1])
Sx[fit1$y[, 3] == 0, ] <- Sx[fit1$y[, 3] == 0, c(2, 1)]
Sy[fit2$y[, 3] == 0, ] <- Sy[fit2$y[, 3] == 0, c(2, 1)]
d11 <- (fit1$y[, 3] == 2) & (fit2$y[, 3] == 2)
d12 <- (fit1$y[, 3] == 2) & (fit2$y[, 3] == 3)
d13 <- (fit1$y[, 3] == 2) & (fit2$y[, 3] == 0)
d21 <- (fit1$y[, 3] == 3) & (fit2$y[, 3] == 2)
d22 <- (fit1$y[, 3] == 3) & (fit2$y[, 3] == 3)
d23 <- (fit1$y[, 3] == 3) & (fit2$y[, 3] == 0)
d31 <- (fit1$y[, 3] == 0) & (fit2$y[, 3] == 2)
d32 <- (fit1$y[, 3] == 0) & (fit2$y[, 3] == 3)
d33 <- (fit1$y[, 3] == 0) & (fit2$y[, 3] == 0)
indicator <- cbind(d11, d12, d13, d21, d22, d23, d31, d32, d33)
myloglik <- function(beta, cov, Sx, Sy, indicator) {
l <- numeric(nrow(Sx))
indsum <- apply(indicator, 2, sum)
if(indsum[1] > 0) {
tobelogged <- 1 - Sx[indicator[, 1], 1] - Sy[indicator[, 1], 1] +
fun.cop(Sx[indicator[, 1], 1], Sy[indicator[, 1], 1], beta,
cov[indicator[, 1], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 1]] <- log(tobelogged)
}
if(indsum[2] > 0) {
tobelogged <- Sy[indicator[, 2], 1] - Sy[indicator[,2], 2] +
fun.cop(Sx[indicator[, 2], 1], Sy[indicator[, 2], 2], beta,
cov[indicator[, 2], , drop = FALSE]) - fun.cop(Sx[indicator[, 2], 1],
Sy[indicator[, 2], 1], beta, cov[indicator[, 2], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 2]] <- log(tobelogged)
```

```
}
if(indsum[3] > 0) {
tobelogged <- Sy[indicator[, 3], 2] - fun.cop(Sx[indicator[,3], 1],
Sy[indicator[, 3], 2], beta, cov[indicator[,3], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 3]] <- log(tobelogged)
}
if(indsum[4] > 0) {
tobelogged <- Sx[indicator[, 4], 1] - Sx[indicator[,4], 2] +
fun.cop(Sx[indicator[, 4], 2], Sy[indicator[,4], 1], beta, cov[indicator[, 4], ,
drop = FALSE]) - fun.cop(Sx[indicator[, 4], 1], Sy[indicator[, 4], 1], beta,
cov[indicator[, 4], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 4]] <- log(tobelogged)
}
if(indsum[5] > 0) {
tobelogged <- fun.cop(Sx[indicator[, 5], 1], Sy[indicator[, 5], 1], beta,
cov[indicator[, 5], , drop = FALSE]) - fun.cop(Sx[indicator[, 5], 1],
Sy[indicator[, 5], 2], beta, cov[indicator[, 5], , drop = FALSE]) -
 fun.cop(Sx[indicator[, 5], 2], Sy[indicator[, 5], 1], beta, cov[indicator[, 5], ,
 drop = FALSE]) + fun.cop(Sx[indicator[, 5], 2], Sy[indicator[, 5], 2], beta,
 cov[indicator[, 5], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 5]] <- log(tobelogged)
}
if(indsum[6] > 0) {
tobelogged <- fun.cop(Sx[indicator[, 6], 1], Sy[indicator[, 6], 2], beta,
cov[indicator[, 6], , drop = FALSE]) - fun.cop(Sx[indicator[, 6], 2],
Sy[indicator[, 6], 2], beta, cov[indicator[, 6], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 6]] <- log(tobelogged)
}
if(indsum[7] > 0) {
l[indicator[, 7]] <- log(Sx[indicator[, 7], 2] -fun.cop(Sx[indicator[, 7], 2],
Sy[indicator[, 7], 1], beta, cov[indicator[, 7], , drop = FALSE]))
}

if(indsum[8] > 0) {
tobelogged <- fun.cop(Sx[indicator[, 8], 2], Sy[indicator[, 8], 1], beta,
cov[indicator[, 8], , drop = FALSE]) - fun.cop(Sx[indicator[, 8], 2],
```

```
 Sy[indicator[, 8], 2], beta, cov[indicator[, 8], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 8]] <- log(tobelogged)
}
if(indsum[9] > 0) {
tobelogged <- fun.cop(Sx[indicator[, 9], 2], Sy[indicator[, 9], 2], beta,
cov[indicator[, 9], , drop = FALSE])
tobelogged[tobelogged <= 0] <- NA
l[indicator[, 9]] <- log(tobelogged)
}
return(-sum(l))
}
if(missing(cov))
X <- matrix(rep(1, N), nrow = N)
beta <- init.param
testll <- myloglik(beta, X, Sx, Sy, indicator)
max.test <- 1
while (!is.finite(testll) & max.test < 10) {
beta <- beta/2
testll <- myloglik(beta, X, Sx, Sy, indicator)
max.test <- max.test + 1
}
print(max.test)
result <- optim(beta, myloglik, method = "L-BFGS-B", lower = 0.001, upper = 20,
cov = X, Sx = Sx, Sy = Sy, indicator = indicator, control = list(maxit = 250,
parscale = rep(0.1, dim(X)[2]), trace = 1, REPORT = 1))
names(result$par) <- paste("Copula.param", 1:length(result$par), sep = "")
if(result$convergence == 0) {
firstoutres <- unlist(c(fit1$regres[1], fit1$spline[3], fit1$degree.smooth[2],
fit2$regres[1], fit2$spline[3], fit2$degree.smooth[2], result$par))
} else firstoutres <- 0
} else firstoutres <- 0
chosenLambda1 <- unlist(fit1$degree.smooth[2])
chosenLambda2 <- unlist(fit2$degree.smooth[2])
if(bootstrap == TRUE & length(firstoutres) > 0) {
M <- nboot
varres <- matrix(numeric(M * length(firstoutres)), M)
for (i in 1:M) {
set.seed(seed + i)
newdata <- data[sample(N, replace = TRUE), ]
```

```
cat(c("Boostrap sample ", i, "\n"))
fit1 <- try(smoothSurvReg(marginal1, logscale = logscale1,
lambda = exp(chosenLambda1), data = newdata,
control = control1), TRUE)
fit2 <- try(smoothSurvReg(marginal2, logscale = logscale2,
lambda = exp(chosenLambda2), data = newdata,
control = control2), TRUE)
if(inherits(fit1, "try-error")) fit1 <- list(fail = 99)
if(inherits(fit2, "try-error")) fit2 <- list(fail = 99)
if(fit1$fail < 99 && fit2$fail < 99) {
N <- dim(fit1$y)[1]
Sx <- survfitS.smoothSurvReg(fit1, fit1$x[, -1],
fit1$z[, -1])
Sy <- survfitS.smoothSurvReg(fit2, fit2$x[, -1],
fit2$z[, -1])
Sx[fit1$y[, 3] == 0, ] <- Sx[fit1$y[, 3] == 0, c(2, 1)]
Sy[fit2$y[, 3] == 0, ] <- Sy[fit2$y[, 3] == 0, c(2, 1)]
d11 <- (fit1$y[, 3] == 2) & (fit2$y[, 3] == 2)
d12 <- (fit1$y[, 3] == 2) & (fit2$y[, 3] == 3)
d13 <- (fit1$y[, 3] == 2) & (fit2$y[, 3] == 0)
d21 <- (fit1$y[, 3] == 3) & (fit2$y[, 3] == 2)
d22 <- (fit1$y[, 3] == 3) & (fit2$y[, 3] == 3)
d23 <- (fit1$y[, 3] == 3) & (fit2$y[, 3] == 0)
d31 <- (fit1$y[, 3] == 0) & (fit2$y[, 3] == 2)
d32 <- (fit1$y[, 3] == 0) & (fit2$y[, 3] == 3)
d33 <- (fit1$y[, 3] == 0) & (fit2$y[, 3] == 0)
indicator <- cbind(d11, d12, d13, d21, d22, d23, d31, d32, d33)
if(missing(cov)) X <- matrix(rep(1, N), nrow = N)
beta <- init.param
testll <- myloglik(beta, X, Sx, Sy, indicator)
max.test <- 1
while (!is.finite(testll) & max.test < 10) {
beta <- beta/2
testll <- myloglik(beta, X, Sx, Sy, indicator)
max.test <- max.test + 1
}
print(max.test)
result <- optim(beta, myloglik, method = "L-BFGS-B", lower = 0.001, upper = 20,
 cov = X, Sx = Sx, Sy = Sy, indicator = indicator, control =
 list(maxit = 250, parscale = rep(0.1, dim(X)[2]), trace = 1, REPORT = 1000))
```

```
if(result$convergence == 0) {
outres <- unlist(c(fit1$regres[1], fit1$spline[3], fit1$degree.smooth[2],
fit2$regres[1], fit2$spline[3], fit2$degree.smooth[2], result$par))
} else outres <- 0
} else outres <- 0
varres[i, ] <- outres
}
varcop <- var(varres[apply(varres, 1, sum) != 0, (ncol(varres) - nvar + 1):
ncol(varres)], na.rm = TRUE)
sdcop <- sqrt(diag(varcop))
meancop <- colMeans(varres[apply(varres, 1, sum) != 0, ])[(ncol(varres) - nvar + 1):
ncol(varres)]
names(meancop) <- paste("MeanBS.Copula.param", 1:length(meancop), sep = "")
names(sdcop) <- paste("sdBS.Copula.param", 1:length(sdcop), sep = "")
BScoefficients <- as.matrix(varres[, (ncol(varres) - nvar + 1):ncol(varres)])
BScoefficients[apply(varres, 1, sum) == 0, ] <- NA
varres[apply(varres, 1, sum) == 0, ] <- NA
results <- unlist(c(firstoutres, meancop, sdcop))
results <- list(fit = results, variance = varcop, BScoefficients = BScoefficients,
 BSresults = varres)
} else results <- firstoutres
return(results)
}


datos<-lapply(1:s,gendata)

f2<-function(i)
{y<-datos[[i]]
w<-y[,1:6]
return(w)
}

datos2<-lapply(1:s,f2)

f3<-function(i)
{y<-datos[[i]]
w<-y[,7:14]
return(w)
}
```

```
datos3<-lapply(1:s,f3)

f<-function(i)
{w<-datos3[[i]]

a<-censcor(xinf | cens(cen1, xsup) ~ yinf | cens(cen2, ysup), w)

b<-as.data.frame(a)
m<-mean(b$rho)

tau<-2/pi*asin(m)

return(tau)

}

fcopnor<-function(i)
{w<-datos2[[i]]

m.normal<- fit.mycopula(w,
copula="normal",init.param=NULL,cov=~1,
marginal1=Surv(L1,R1,type='interval2')~1,
logscale1=~1,lambda1=exp(3:(-3)),
marginal2=Surv(L2,R2,type='interval2')~1,
logscale2=~1,lambda2=exp(3:(-3)),
bootstrap=FALSE, nboot=200)

rho<-as.numeric(m.normal[91])
rho1<-min(rho,1)
return(rho1)

tau<-2/pi*asin(rho1)

return(tau)


}
```

```
Taucens<-sapply(1:s,f)

write(Taucens,"D:/Taucens_50_0.2.txt",append=T,ncol=1)

rhocopnor<-sapply(1:s,fcopnor)

Taucopnor<-2/pi*asin(rhocopnor)


write(Taucopnor,"D:/Taucopnor_50_0.2.txt",append=T,ncol=1)
```

El siguiente código fue utilizado para crear las gráficas del estudio de simulación:

```
y<-matrix(scan("D:/resultadoscompletostesis2.txt"),ncol=5,byrow=T)

y<-as.data.frame(y)

names(y)<-c("n","tau","MAD","ECM","Metodo")

y$Metodo<-as.character(y$Metodo)

y$tau<-as.character(y$tau)


library(tidyverse)


y %>%
filter(Metodo==1) %>%
ggplot(mapping = aes(x = n, y = MAD, group = tau)) +
geom_line(mapping = aes(linetype = tau))

y %>%
filter(Metodo==1) %>%
ggplot(mapping = aes(x = n, y = ECM, group = tau)) +
geom_line(mapping = aes(linetype = tau))

y %>%
```

```
filter(Metodo==2) %>%
ggplot(mapping = aes(x = n, y = MAD, group = tau)) +
geom_line(mapping = aes(linetype = tau))


y %>%
filter(Metodo==2) %>%
ggplot(mapping = aes(x = n, y = ECM, group = tau)) +
geom_line(mapping = aes(linetype = tau))




y$n<-as.character(y$n)

y %>%
filter(tau==0.2) %>%
ggplot(mapping = aes(x = Metodo, y = ECM, group = n)) +
geom_line(mapping = aes(linetype =n))

y %>%
filter(tau==0.2) %>%
ggplot(mapping = aes(x = Metodo, y = MAD, group = n)) +
geom_line(mapping = aes(linetype =n))


y %>%
filter(tau==0.5) %>%
ggplot(mapping = aes(x = Metodo, y = ECM, group = n)) +
geom_line(mapping = aes(linetype =n))

y %>%
filter(tau==0.5) %>%
ggplot(mapping = aes(x = Metodo, y = MAD, group = n)) +
geom_line(mapping = aes(linetype =n))

y %>%
filter(tau==0.8) %>%
```

```
ggplot(mapping = aes(x = Metodo, y = ECM, group = n)) +
geom_line(mapping = aes(linetype =n))

y %>%
filter(tau==0.8) %>%
ggplot(mapping = aes(x = Metodo, y = MAD, group = n)) +
geom_line(mapping = aes(linetype =n))
```

Bibliografía

[1] 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. [1]

[2] Betensky, R. and Finkelstein, D. (1999b). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18:3089–3100. [2]

[3] 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. [4, 5, 10, 11, 12, 13, 14, 16, 17, 22]

[4] Bogaerts, K. and Lesaffre, E. (2008a). Estimating local and global measures of association for bivariate interval censored data with a smooth estimate of the density. *Statistics in Medicine*, 27:5941–5955. [16]

[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. [1, 2, 11, 12, 13, 20, 21, 22, 23, 24]

[6] Canavos, G. (1988). *Probabilidad y Estadística Aplicaciones y Métodos*. McGraw Hill, México. [8]

[7] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–22. [6]

[8] Dickey, J. M. and Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics*, pages 214–226. []

[9] Eilers, P. and Marx, B. (1996). Flexible smoothing with B-Splines and penalties. *Statistical Science*, pages 89–102. [14, 15]

[10] Gentleman, R. and Geyer, C. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81:618–623. [7]

[11] Gentleman, R. and Vandal, A. (2002). Nonparametric estimation of the bivariate cdf for arbitrarily censored data. *Canadian Journal of Statistics*, 30:557–571. [7]

[12] Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60:945–953. [15]

[13] Gibbons, J. D. and Chakraborti, S. (2011). *Nonparametric Statistical Inference*. Springer Berlin Heidelberg. [1]

[14] Greiner, R. (1909). Über das fehlersystem der kollektivmasslehre. *Zeitschrift für Mathematik und Physik*, (121):225. [20]

[15] Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, (292):698–707. [14]

[16] Hamada, M., Wilson, A., Reese, C., and Martz, H. (2008). *Bayesian Reliability*. Springer Science & Business Media. []

[17] Jeffreys, H. (1961). Theory of probability. *The British Journal for the Philosophy of Science*. []

[18] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall. [9]

[19] Johnson, V. E. (2005). Bayes factors based on test statistics. *Journal of the Royal Statistical Society*, 67:689–701. []

[20] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90:773–795. []

[21] Kendall, M. and Gibbons (1990). Rank correlation methods. ed. *Edward Arnold*. []

[22] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93. [1]

[23] 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. [1, 22, 24]

[24] 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. [9, 11]

[25] 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. [14]

[26] Ly, A., Verhagen, J., and Wagenmakers, E. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32. []

[27] Montgomery, D. C. and Runger, G. C. (2007). *Applied Statistics and Probability for Engineers*. John Wiley & Sons. [8]

[28] Nelsen, R. (2006). *An Introduction to Copulas*. Springer. [9, 10]

[29] 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. [1, 8, 19, 20]

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

[31] Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society: series C*, 22:86–91. [5]

[32] Randles, R. H. and Wolfe, D. A. (1979). Introduction to the theory of nonparametric statistics. *Introduction to the theory of nonparametric statistics, by Randles, Ronald H.; Wolfe, Douglas A. New York: Wiley, c1979. Wiley series in probability and mathematical statistics.* []

[33] 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. [23]

[34] Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B*, 38:290–295. [5, 6]

[35] Van Doorn, J., Ly, A., Marsman, M., and Wagenmakers, E. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*, 72:303–308. []

[36] Yuan, Y. and Johnson, V. E. (2008). Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica*, pages 1185–1200. []