UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

División de Ciencias Básicas e Ingeniería

Modelos Paramétricos y Semiparamétricos de Regresión Basados en Cópulas para el Análisis de Riesgos Competitivos

Tesis que presenta

M. EN C. ALEJANDRO ROMÁN VÁSQUEZ

Para obtener el grado de

Doctor en Ciencias
(Matemáticas)

Asesor de tesis:

Dr. Gabriel Escarela Pérez

Jurado calificador:

Presidente:	Dra. Silvia Ruiz-Velasco Acosta	IIMAS
Secretario:	Dr. Gabriel Núñez Antonio	UAM-I
Vocal:	Dra. Hortensia Josefina Reyes Cervantes	BUAP
Vocal:	Dr. Alberto Castillo Morales	UAM-I
Vocal:	Dr. Gabriel Escarela Pérez	UAM-I

Ciudad de México. **20 de enero de 2020**

Índice

Agradecimientos

En este punto de mi existencia siento un profundo agradecimiento a Dios por permitirme ser parte de una gran familia, compuesta por mi esposa Laura, mi hijo Arturo, mi angelito en el cielo, y mi bebé arcoíris. Sin duda alguna ustedes son mi mayor motivación, y en cada esfuerzo realizado para lograr este trabajo siempre estuvieron detrás. Los amo con todo mi ser, y siempre estaré a su lado.

Cuando le mencioné a mi esposa Laura que iba a hacer el doctorado, nunca dudo en respaldar esa idea. Quien más que ella, que presenció mi trabajo durante casi 5 años, puede constatar lo difícil que fue. Te agradezco que siempre me alentaste y animaste a seguir adelante, pues ese apoyo fue fundamental para concretar este proyecto. Te amo, como no tienes una idea, soy inmensamente feliz a tu lado. A mi amado "Tuto" quiero decirle si bien estaba dedicado al 100% al doctorado, tuve la gran oportunidad de cuidarte y atenderte. Fui muy afortunado al poderte disfrutarte todo ese tiempo, etapa que recordaré con una mezcla de nostalgia y alegría.

Seguramente este logro personal representa mucho para mi madre. Recordará que cuando era pequeño le dije que sería doctor, y aunque quizá en su momento me refería a la profesión de médico, puedo decirle que le cumplí, y heme aquí logrando ser doctor en ciencias matemáticas. Madre, gracias por todo el apoyo incondicional que me has dado, ahora que soy padre, comprendo mejor ese amor ilimitado que se siente por los hijos. Merecen también mención mi hermana Beatriz, mi sobrino Carlos y mi tío Daniel, gracias por creer en mí y alentarme a terminar mi doctorado.

No quisiera dejar pasar la oportunidad de agradecer a mis suegros Martha y Ramón, porque además de que estuvieron al pendiente de mis avances, me ayudaron en algunas ocasiones a cuidar a Arturo, lo que me permitió avanzar bastante en el proyecto de investigación.

Quiero dar las gracias a mi asesor, el doctor Gabriel Escarela Pérez, ya que desde el primer momento confió en mí para lograr este trabajo. Nuestras discusiones académicas siempre fueron muy fructíferas, y su gran experiencia como investigador fue una pieza clave para guiarme por un camino que, si bien fue sinuoso, fue sumamente estimulante. Agradezco también a mi comité tutorial, compuesto por los doctores Silva Ruiz-Velasco Acosta, Hortensia Josefina Reyes Cervantes, Alberto Castillo Morales y Gabriel Núñez Antonio, cada comentario, cada observación, cada corrección, fue de gran trascendencia y aprendizaje.

Por último, pero no menos importante, agradezco a la Universidad Autónoma Metropolitana por abrirme nuevamente las puertas y brindarme la oportunidad de estudiar un posgrado más. Los apoyos en la inscripción, así como el apoyo extraordinario para los alumnos de posgrado fueron incentivos económicos de una ayuda invaluable. Las gracias se hacen extensivas a toda la comunidad de la UAM-Iztapalapa, académicos de la división de CBI, compañeros de clases, y administrativos. Finalmente, doy las gracias al Consejo Nacional de Ciencia y Tecnología por el soporte económico que me brindó, sin el cual no habría tenido la posibilidad de aventurarme, desarrollar y concluir esta etapa académica y de formación profesional.

Resumen

En este trabajo se propone modelar la función de distribución conjunta del tiempo de supervivencia T y del tipo de evento D para el análisis de riesgos competitivos a través de modelos basados en cópulas Gaussianas. Efectos de covariables se incorporan al caracterizar las distribuciones marginales usando modelos paramétricos y semiparamétricos de riesgos proporcionales para el tiempo de supervivencia y un modelo multinomial para el tipo de evento. La estimación de los modelos se hace a través de verosimilitud (máxima verosimilitud para el enfoque paramétrico y pseudo-verosimilitud para el enfoque semiparamétrico), lo que permite el uso de criterios de información para encontrar los modelos más parsimoniosos o sencillos (en términos del número de parámetros que los constituyen). El desempeño de los estimadores se evalúa a través de simulaciones. La metodología propuesta se ilustra al aplicarse a un conjunto de datos prospectivos de pacientes con linfoma folicular de etapa temprana (I o II) registrados para su tratamiento con radioterapia o radioterapia y quimioterapia en el hospital Princess Margaret, en Toronto, entre los años 1967 y 1996.

En términos generales, se considera que los modelos propuestos representan un enfoque novedoso, ya que hasta donde se tiene conocimiento, no existe hasta hoy en la literatura un intento por modelar de forma conjunta el vector aleatorio (T, D) para el estudio de riesgos competitivos. Estos procedimientos presentan bondades como la capacidad para modelar los efectos a largo plazo (a través del modelo marginal del tipo de evento), la capacidad para calcular de forma práctica y directa la función de incidencia acumulada $CIF_d(t)$ (función que es de gran utilidad para el estudio de riesgos competitivos), generar modelos con menos parámetros respecto a otros modelos que analizan el vector (T, D) usando distribuciones condicionales, y en casos especiales, se pueden generar interpretaciones de los parámetros de ambas componentes lineales que son útiles para entender el comportamiento de la función de incidencia acumulada.

1.- Introducción

1.1.- Análisis de datos de supervivencia

En el campo de la investigación médica, parte del interés radica en estudiar la ocurrencia de algún evento de interés (como puede ser el desarrollo de una enfermedad o el fallecimiento de alguna persona) a través del tiempo. En el ámbito epidemiológico, es común encontrar términos como "riesgo" y "tasa" como conceptos fundamentales para estudiar este tipo de fenómenos (Andersen et al., 2012). El riesgo se define como la fracción d/n donde d es el número de individuos que desarrolla el evento de interés durante un tiempo específico $[0, t]$ y n es la cantidad original de sujetos libres del evento de interés. Definido de esta forma, el riesgo es una cantidad que aumenta con el tiempo. La tasa se define en términos similares al riesgo d/n', sólo que ahora se considera la cantidad de sujetos n' ($n' \leq n$) que están en riesgo durante el intervalo temporal $[0, t]$. Como n' depende del tiempo, la tasa puede incrementarse, disminuir, o incluso, ser constante al variar t.

Como lo menciona Andersen et al. (2012), el concepto epidemiológico del riesgo se puede equiparar a una probabilidad, por lo que si $F(t)$ denota la probabilidad de que un individuo libre de evento seleccionado al azar desarrolle la enfermedad o fallezca en el tiempo t, entonces el riesgo d/n se puede usar para estimar $F(t)$ si se pudiera observar la ocurrencia de los eventos de todos los individuos. Sin embargo, en la mayoría de los estudios médicos es inevitable que existan individuos para los cuales el tiempo de la ocurrencia del evento de interés se desconozca, principalmente porque no ocurre en la ventana temporal $[0, t]$ o porque algunos pacientes abandonan los estudios, a lo que se le llama observaciones censuradas. En estos casos, se necesita recurrir a la disciplina estadística llamada análisis de supervivencia para poder afrontar estas dificultades. El desafío de estimar la función de distribución acumulada $F(t)$ basándose en datos incompletos puede lograrse considerando a la censura como independiente, por lo que un individuo censurado en el tiempo t puede representar a los individuos que aún se mantienen en riesgo, es decir, la experiencia de supervivencia de los pacientes censurados es similar a aquellos que continúan en el estudio. Bajo censura independiente, $F(t)$ puede ser estimado a través de la estadística $1 - \hat{S}(t)$, donde $\hat{S}(t)$ es el estimador de Kaplan-Meier de la función de supervivencia $S(t)$, el cual es un producto de

factores dados por $1 - d'_k/n'_k$, donde d'_k representa el número de eventos y n'_k es el número individuos en riesgo en el tiempo k.

De acuerdo con Andersen et al. (2012), el concepto en análisis de supervivencia que se equipara con el término de "tasa" es la función de riesgo $h(t)$, la cual es la tasa de ocurrencia instantánea del evento de interés para sujetos aún en riesgo (es decir, los que aún no han desarrollado la enfermedad o han fallecido). La función de riesgo se define matemáticamente como (Collett, 2015):

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | \, T \geq t)}{\delta t}. \tag{1}$$

La función de riesgo representa una descripción dinámica de cómo varía de forma instantánea el riesgo de la ocurrencia del evento de interés, para aquellos individuos aún en riesgo.

Existe una correspondencia entre la función de supervivencia $S(t)$ -y por ende también, con la función de distribución $F(t)$- y la función de riesgo $h(t)$ dada por:

$$S(t) = \exp\left(-\int_0^t h(z)dz\right) = \exp(-H(t)) \tag{2}$$

donde $H(t)$ representa la función de riesgo acumulada.

La estimación e interpretación de las funciones de supervivencia $S(t)$ y de la función de riesgo $h(t)$ son unos de los objetivos más importantes en el análisis de datos de supervivencia. Si el interés radica en analizar además la relación de estas cantidades con variables explicativas de los pacientes que se han registrado al inicio de algún estudio, como puede ser un tratamiento específico, la edad de los pacientes, el índice de masa corporal, la cantidad azúcar en la sangre, por mencionar algunas, es conveniente la aplicación de modelos de regresión que puedan lidiar con datos censurados. Entre los más representativos se puede encontrar el modelo de regresión de Cox, el cual modela directamente la función de riesgo $h(t)$ en términos de un conjunto de covariables $\boldsymbol{x}$ (Collett, 2015):

$$h(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) h_0(t) \tag{3}$$

donde $\boldsymbol{\beta}$ es el vector de parámetros (que no incluye intercepto) y $h_0(t)$ es la función de riesgo de base la cual representa la función de riesgo de un individuo para el cual los valores de las covariables son todos cero. Es común que las comparaciones entre diversas funciones de riesgo

se realicen a través del cociente de funciones de riesgo, para el cual la función $h_0(t)$ se elimina, quedando solamente la información de la exponencial de la componente lineal.

En este sentido, la relación dada por la ecuación (2) es de suma importancia, pues como lo destaca Andersen et al. (2012), modelos de regresión de Cox para la función de riesgo $h(t)$ implican modelos para $F(t)$, y si un factor está asociado con un incremento en la función de riesgo, entonces también está asociado con un incremento en $F(t)$.

1.2.- Riesgos competitivos

En el ámbito del análisis de supervivencia, otro tópico a considerar adicional al tema de la censura, es que el paciente puede desarrollar otro tipo de eventos clínicos que de cierta forma pueden ocultar o impedir la ocurrencia del suceso en cuestión o evento de interés. Por ejemplo, en un estudio sobre cáncer de mama se podría tratar de determinar la relación que existe entre el tiempo en que el paciente tarda para iniciar cierto tratamiento, y el tiempo de muerte por este tipo de cáncer, sin embargo, el paciente puede fallecer por complicaciones relacionadas a otro tipo de cáncer, o morir por otro tipo de circunstancias; es decir, el riesgo de fallecimiento por cáncer de mama compite con otros riesgos de muerte. Por tal motivo, a los sucesos que ocurren una vez en cada individuo y que compiten entre sí se les denomina riesgos competitivos (Liu, 2012). Así como el ejemplo anterior, en el ámbito epidemiológico y de investigación médica, se pueden encontrar muchos escenarios de este tipo, siendo una regla, más que una excepción.

Una forma de obtener la tasa o función de riesgo del evento de interés en el lapso temporal $[0, t]$, podría ser bajo la suposición de considerar las observaciones de los eventos que compiten como datos censurados independientes (Prentice et al., 1978). En este contexto, se puede definir la función de riesgo de causa específica como;

$$h_d(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, D = d \mid T \geq t)}{\delta t} \tag{4}$$

donde el tipo de evento se representa con la variable D, el cual en términos generales puede estar representado por J diferentes eventos $D = \{1, 2, \dots, d, \dots, J\}$. La función de riesgo de causa específica se puede interpretar como la tasa instantánea de ocurrencia de un evento específico d para sujetos que no han experimentado ningún evento.

Bajo la presencia de riesgos competitivos, el riesgo o la probabilidad de ocurrencia de un evento d, conocida también como la función de incidencia acumulada, definida matemáticamente como $CIF_d(t) = P(T \leq t, D = d)$, es una cantidad que es de gran interés en la investigación médica, ya que su interpretación es intuitiva y atractiva pues establece la proporción de pacientes que desarrollan cada evento en un tiempo específico, motivo por el cual es presentada frecuentemente en artículos de investigación (Kim, 2007). Para el caso de dos eventos que compiten y en ausencia de variables explicativas, la Figura 1 muestra las gráficas de las funciones de incidencia acumulada para ambos riesgos. A medida que pasa el tiempo, las curvas se deben ir aproximando a las marginales $P(D = 1)$ y $P(D = 2)$, donde resulta evidente que en este caso específico la $CIF_1(t)$ converge más rápido que $CIF_2(t)$.

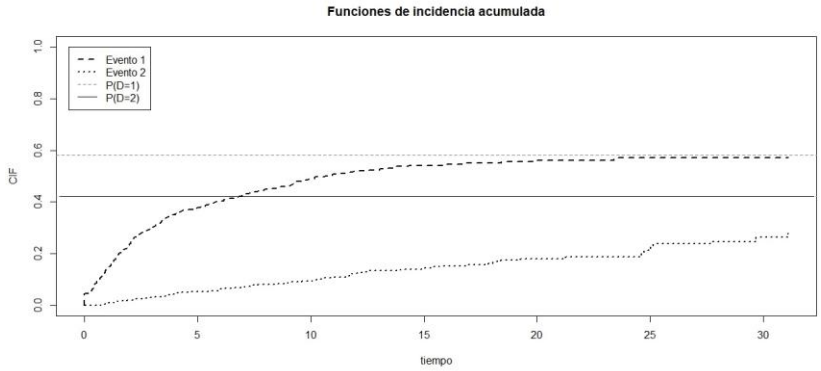

Figura 1. Función de incidencia acumulada en el caso de dos eventos que compiten. Para el horizonte de tiempo establecido, la curva para el evento 1 posee un valor muy cercano a $P(D = 1)$.

Cuando no hay riesgos que compiten es posible establecer una correspondencia entre la probabilidad de ocurrencia $F(t)$ y la función de riesgo $h(t)$, como se vio en la sección anterior. Sin embargo, bajo la presencia de diversos eventos que pueden acontecer, la relación entre la función de incidencia acumulada $CIF_d(t)$ y función de riesgo de causa específica $h_d(t)$, es más complicada. De acuerdo con Andersen et al. (2012), la relación entre ambas cantidades está dada por:

$$CIF_d(t) = \int_0^t S(s)h_d(s)ds \tag{5}$$

donde $S(t)$ es la función de supervivencia total dada por:

$$S(t) = \exp\left(-\sum_{d=1}^{J} \int_0^t h_d(z)dz \right) = \exp\left(-\sum_{d=1}^{J} H_d(t) \right). \quad (6)$$

En la relación anterior $H_d(t)$ representa la función de riesgo acumulada de causa específica. La ecuación (6) tiene dos implicaciones importantes. La primera es que una estimación para $CIF_d(t)$ basada solamente en el estimador de Kaplan-Meier (tratando a los eventos que compiten como cantidades censuradas), resulta sesgada; la segunda es que la forma en que $h_d(t)$ se puede asociar con un conjunto de covariables no coincide necesariamente con la manera de asociación de los mismos con $CIF_d(t)$. Por ejemplo, en un modelo de Cox de causa específica, construido para modelar $h_d(t)$ respecto a un conjunto de variables explicativas, se puede dar el caso de que el incremento en una variable explicativa que implica un incremento en $h_d(t)$, tenga un efecto contrario de decremento en $CIF_d(t)$.

Por lo anterior, se han buscado construir modelos que permitan ligar directamente el comportamiento de la función de incidencia acumulada con un conjunto de variables explicativas, y obtener estimaciones de $CIF_d(t)$ que no resulten sesgadas. Entre los modelos más populares, se encuentra aquel propuesto por Fine y Gray (1999), que se basa en una versión diferente de la función de riesgo (conocida como función de riesgo subdistribuida) la cual se define como:

$$h_{sd}(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, D = d \mid T \geq t \cup (T < t \cap D \neq d))}{\delta t}. \quad (7)$$

Este tipo de función de riesgo se puede interpretar como la tasa instantánea de ocurrencia de un evento específico d para sujetos que no han experimentado ese evento en particular. Hay que notar que a diferencia de lo que ocurre con la función de riesgo de causa específica, los individuos que han experimentado un evento que compite se mantienen en el grupo de riesgo.

Usando $h_{sd}(t)$, se pueden crear modelos de regresión de Cox para los cuales se puede establecer una relación de correspondencia entre el efecto de los predictores y el comportamiento de $CIF_d(t)$, lo cual es de suma utilidad en el área médica.

A pesar de la gran aceptación de este tipo de modelos, hay que mencionar que la función de riesgo subdistribuida ya no tiene un parecido con el concepto epidemiológico de "tasa" (a diferencia de la función de riesgo de causa específica), pues se mantienen individuos en el conjunto de riesgo que han fallecido por otras causas, a pesar de que en realidad ya no están en

riesgo de experimentar el evento d, hecho que para algunos autores (Andersen y Keiding, 2012), no está completamente justificado. Es importante mencionar también que el modelo de Fine y Gray a veces es interpretado de forma incorrecta (Austin y Fine, 2017), pues si bien la dirección de los efectos de las covariables son los mismos para $h_{sd}(t)$ que para $CIF_d(t)$, la magnitud no lo es, y a veces se hace la afirmación incorrecta de que el cociente de la función de riesgo subdistribuida proporciona información sobre la magnitud del cambio en la función de incidencia acumulada (este tipo de errores de interpretación se dan también en ausencia de riesgos competitivos, al querer establecer que la magnitud del cociente de funciones de riesgo se relaciona directamente con la probabilidad de ocurrencia de un evento $F(t)$).

A veces es trascendente medir los efectos de largo plazo que ciertas covariables pueden tener en la incidencia de las diversas causas. Por ejemplo, en estudios clínicos de trasplante de médula ósea, resultan importantes los efectos a largo plazo que el tratamiento médico puede tener en la incidencia de la mortalidad (en presencia de riesgos competitivos), lo que es de relevancia para el manejo clínico del paciente (Zhang y Fine, 2008). Bajo este contexto, existen otros modelos que se centran en el estudio de la distribución conjunta $P(T, D)$ a través de la descomposición $P(T|D)P(D)$, lo que permite generar un modelo en particular para $P(D)$ y de esta forma cuantificar los efectos a largo plazo que ciertas variables predictoras pueden tener en los distintos eventos que compiten. Este tipo de descomposición fue propuesta por Larson y Dinse (1985) a través de un modelo mixto, donde la probabilidad $P(D = d)$ se explica a través de un modelo multinomial que permite la incorporación de covariables (de acuerdo a Lagakos et al. (1978) se asume que el tipo de evento para cada individuo es seleccionado por un mecanismo estocástico externo) y las distribuciones condicionales $P(T|D = d)$ a través de un modelo de regresión paramétrico de riesgos proporcionales. Algunos autores han propuesto diversos cambios al modelo, en específico en la parte de las distribuciones condicionales del tiempo de supervivencia dado el tipo de evento, estableciendo otras distribuciones paramétricas (Lau et al., 2008) y enfoques semiparamétricos (Kuk, 1992, Ng y McLachlan, 2003).

Centrándose en el estudio de la distribución conjunta $P(T, D)$, otros autores (Nicolaie et al., 2010) han propuesto la descomposición alternativa $P(D|T)P(T)$ argumentando que es una forma más natural pues no es necesario suponer que las causas de fallecimiento están determinadas por una causa exterior. La forma de modelar $P(T)$ es a través de la función de riesgo total $h_T(t)$, la cual considera como evento de interés la ocurrencia de cualquier evento

que compite[I], permitiendo la formulación del modelo de regresión de Cox para describir su comportamiento en términos de variables explicativas. El término $P(D|T)$ se modela a través de la función de riesgo de causa específica relativa $\pi_d(t)$, la cual se define como el cociente $h_d(t)/h_T(t)$, donde $h_d(t)$ es la función de riesgo de causa específica. La inclusión de variables explicativas a la función de riesgo de causa específica relativa, que describe el comportamiento local temporal $\pi_d(t) = P(D = d|T = t)$, se puede hacer a través de un modelo multinomial, pero para garantizar que la función $\pi_d(t)$ sea continua y suave, es necesario incluir en la componente lineal un conjunto base de funciones de alisamiento (Nicolaie et al., 2010).

A pesar de las bondades de ambos enfoques de descomposición de vector aleatorio (T, D), la incorporación de covariables tanto en las distribuciones marginales como en las condicionales incrementa considerablemente el número de parámetros, haciendo que la interpretación sea complicada cuando son varias las causas de fallecimiento (Haller et al., 2013).

Teniendo como antecedentes los modelos de descomposición, la propuesta que se hace en este trabajo es modelar la función de probabilidad conjunta $P(T, D)$, lo que permite un acercamiento que resulta en un tratamiento simétrico de ambas respuestas, preservado la interpretación marginal de los parámetros de regresión y pudiendo establecer las respectivas distribuciones marginales. Las cópulas, que básicamente son funciones de distribución definidas en el cuadrado unitario $[0,1]^2$ con marginales uniformes (Nelsen, 2006), permiten modelar la función de probabilidad conjunta de un vector aleatorio con gran flexibilidad marginal, permitiendo el manejo mixto de variables aleatorias continuas (tiempo de supervivencia T) y variables aleatorias discretas (tipo de evento D).

Si bien este trabajo se centran en el modelado del vector aleatorio (T, D) a través de cópulas para el análisis de riesgos competitivos, lo que representa una propuesta novedosa que hasta ahora no se había explorado, el uso de cópulas para el estudio de riesgos que compiten se ha desarrollado con anterioridad pero a partir de un enfoque diferente. Contemplando que cada individuo está en riesgo de que le ocurra uno de los J eventos que compiten, se pueden considerar los tiempos de supervivencia de cada riesgo, dados por $T_1, \dots, T_J$, como un conjunto de variables latentes para las cuales el tiempo de supervivencia observado se define como el mínimo de los tiempos de supervivencia observados de las variables latentes (Liu, 2012). Bajo

[I] En esencia $h_T(t)$, es básicamente la función de riesgo $h_T(t)$ dada por la ecuación (1).

este contexto, el análisis del vector aleatorio $(T_1, \dots, T_J)$, implica el análisis de la función de distribución conjunta $F_{T_1,\dots,T_J}(T_1, \dots, T_J)$, siendo de particular interés el comportamiento de las funciones de distribuciones marginales $F_{T_1}(T_1), \dots, F_{T_J}(T_J)$ para la descripción de los riesgos que compiten. Bajo el supuesto de independencia de los tiempos de supervivencia $T_1, \dots, T_J$, es posible identificar de forma única a las distribuciones marginales directamente de los datos de riesgos competitivos, ya que sin este supuesto, existen muchas funciones de distribución conjunta que comparte los mismos marginales (a esto se le conoce como el problema de la identificabilidad de riesgos competitivos, Tsiatis, 1975). Para poder resolver este problema, es necesario asumir alguna estructura de dependencia del vector aleatorio $(T_1, \dots, T_J)$. Es aquí donde el uso de cópulas se ha introducido para modelar de forma flexible la posible dependencia de los riesgos que compiten, algo que ha quedado plasmado en trabajos de diversos autores, entre los que destacan Zheng y Klein (1995), Escarela y Carriere (2003) o más recientemente Shi y Wu (2016). Si bien este tipo de enfoque ha servido para analizar y representar el fenómeno de riesgos competitivos, es importante mencionar que la estructura de dependencia es un supuesto del modelo que no se puede validar directamente de los datos (Lo y Wilke, 2010).

1.3.- Modelo de cópula para el tiempo de supervivencia $\boldsymbol{T}$ y el tipo de evento $\boldsymbol{D}$

Centrándose en el caso particular del modelado de variables aleatorias mixtas, una estrategia que se ha estado empleando recientemente en diversas áreas (Zilko y Kurowicka, 2016, Chen y Hanson, 2017) es el uso de la cópula Gaussiana para el modelado conjunto de variables aleatorias discretas y variables aleatorias continuas. El empleo del concepto de cópula se debe a la flexibilidad para modelar de forma independiente el comportamiento marginal de la dependencia entre las variables, y el uso en específico de la cópula Gaussiana se atribuye principalmente a su capacidad para modelar amplios rangos de dependencia, así como su fácil tratamiento analítico y sus convenientes propiedades marginales y condicionales (Jiryaie et al. 2016). La incorporación de variables explicativas a las distribuciones marginales es posible en este enfoque, como ha quedado de manifiesto en el trabajo de Song (2007), Song et al. (2009),

estableciendo las bases para los modelos de regresión conjunta de variables correlacionadas empleando cópulas Gaussianas, teoría que ha llamado la atención de diversos autores (de Leon y Wu, 2011, Czado et al., 2012, He et al., 2012, Krämer et al., 2013, Shi et al., 2015 y Frees et al., 2016).

Bajo este contexto, el planteamiento de este trabajo es modelar conjuntamente el tiempo de supervivencia T y el tipo de evento D empleando la cópula Gaussiana, con la respectiva incorporación de covariables en cada marginal. Siguiendo el enfoque de Song et al. (2009), se establece como primera aproximación, un modelo completamente paramétrico ajustando un modelo multinomial a la variable discreta D y un modelo de regresión paramétrico de Cox de riesgos proporcionales a la variable continua T. El modelo paramétrico de Cox, visto desde el punto de vista de la función de riesgo, se puede representar también por la ecuación (3), sólo que la función de riesgo de base $h_0(t)$ está representada por una familia paramétrica, para la cual se selecciona la distribución Weibull, la cual ha mostrado ajustes adecuados en varios estudios médicos (Carroll, 2003) y posee una gran flexibilidad para modelar la función de riesgo (Collett, 2015). La estimación está basada en verosimilitud, lo que permite tener un marco teórico sólido al momento de hacer inferencia estadística.

Cuando los datos se apegan a los supuestos distribucionales propuestos, se obtienen modelos que poseen una mejor precisión que aquellos modelos libres de distribución (Collett, 2015). Sin embargo, en algunas ocasiones las familias paramétricas disponibles no poseen la flexibilidad necesaria para modelar la forma actual de la distribución del tiempo de supervivencia, y a veces los resultados inferenciales son muy sensibles al tipo de distribución empleada (Ng y McLachlan, 2003). Por tal motivo, se introduce también en este trabajo un modelo semiparamétrico (libre de distribución) para la componente temporal T a través de un modelo de Cox de riesgos proporcionales dado por la ecuación (3). Es necesario tomar en cuenta que en este enfoque la variable aleatoria T se considera como discreta, con soporte dado en cada punto temporal donde existe un evento (sin importar la causa); la unión entre componentes marginales, seguirá siendo la cópula Gaussiana siguiendo el enfoque de Song et al. (2009). La estimación se logra de nueva cuenta a través de verosimilitud, apegándose en principio a los procedimientos semiparamétricos (con sus respectivas adecuaciones) establecidos en Lawless y Yilmaz (2011), Yilmas y Lawless (2011) y Li et al. (2008) para la estimación de cópulas con marginales libres de distribución. Sin embargo, dado que la estimación de la función de riesgo de base $h_0(t)$ implica un estimador por cada valor del tiempo

donde existe un evento, el proceso de estimación conjunta se puede volver muy demandante computacionalmente hablando. Por tal motivo, como se verá con detalle más adelante, se propone un método de pseudo-verosimilitud en dos etapas: en la primera se estima de forma convencional el modelo de Cox a la variable T, y en la segunda se asume que se conoce el modelo marginal de supervivencia sustituyéndolo en la verosimilitud del modelo completo para estimar los parámetros restantes (aquellos relacionados a la variable D y la dependencia de la cópula). A través de la teoría de funciones de inferencia (Song, 2007) se puede encontrar (siempre que no haya observaciones temporales repetidas) que el vector de estimadores posee una distribución asintótica normal con matriz de varianza-covarianza dada por la inversa de la matriz de información de Godambe (Joe, 2014). Si bien es atractivo tener un resultado que nos garantice propiedades asintóticas deseables, en la práctica el cálculo y cómputo de la matriz de Godambe implica un procedimiento analítico complicado, por lo que algunos autores han señalado que la inferencia estadística de los parámetros estimados se puede lograr a través de métodos de remuestreo como Jacknife o Bootstrap (Joe, 2014). En el caso particular de este trabajo se propone obtener los errores estándar de los estimadores a través de la técnica de remuestreo no paramétrica bootstrap (Efron y Tibshirani, 1993); la generación de intervalos de confianza de ciertas cantidades relevantes se puede obtener usando esta técnica de remuestreo con corrección de sesgo (Carpenter y Bithell, 2000, Efron y Hastie, 2016), que además sirve en el caso general, cuando hay observaciones repetidas y no se pueden garantizar las propiedades asintóticas de los estimadores.

En términos generales, se considera que los modelos propuestos representan un enfoque novedoso, ya que hasta donde se tiene conocimiento, no existe hasta hoy en la literatura un intento por modelar de forma conjunta el vector aleatorio (T, D) para el análisis de riesgos competitivos. Estos procedimientos presentan bondades como la capacidad para modelar los efectos a largo plazo (a través del modelo marginal del tipo de evento), la capacidad para calcular de forma práctica y directa la función de incidencia acumulada $CIF_d(t)$, generar modelos con menos parámetros respecto a los modelos de descomposición de $P(T, D)$, y en casos especiales se pueden generar interpretaciones de los parámetros de ambas componentes lineales que son útiles para entender el comportamiento de las funciones de incidencia acumulada.

La organización de la tesis es la siguiente. En el capítulo 2 se explica con detalle los pormenores de los dos enfoques propuestos: se hace una breve introducción sobre cópulas, se describen a

detalle los modelos de regresión de cópulas mixtos, se especifican los modelos marginales y finalmente se hace la vinculación de los modelos con la función de incidencia acumulada y la función de riesgo de causa específica. En el capítulo 3 está dedicado al proceso de estimación, describiendo con detalle la formulación de las funciones de verosimilitud y las propiedades asintóticas del método de pseudo-verosimilitud. Se hacen estudios de simulación para validar los procesos de estimación y evaluar las propiedades de los estimadores de ambos métodos. En el capítulo 4 se aplican ambos modelos en unos datos sobre un estudio prospectivos de pacientes con linfoma folicular, obteniéndose resultados interesantes. Finalmente, el capítulo 5 está destinado a las conclusiones y perspectivas del trabajo.

2.- Modelo de regresión de cópulas

En este capítulo se describirá con mayor detalle los aspectos más relevantes de los modelos propuestos para representar la función de probabilidad conjunta de 2 variables aleatorias mixtas $P(T, D)$. Se hará una breve introducción de cópulas mencionando algunas de sus particularidades, para posteriormente especificar las distribuciones de las marginales, así como su relación con cada componente lineal para la incorporación de variables explicativas. En la parte final del capítulo, se establece la relación entre los modelos propuestos y cantidades importantes para análisis de riesgos competitivos como son $CIF_d(t)$ y $h_d(t)$.

2.1.- Modelado de dependencia a través de cópulas

Si bien el concepto de cópula no es nuevo[II], su empleo se ha popularizado considerablemente en los últimos 25 años, impulsado en gran medida por la industria financiera (Embrechts, 2009), aunque su utilización se ha permeado también en otras áreas entre las que destacan la industria aseguradora, los estudios hidrológicos y el área de bioestadística (Joe, 2014). Una de las principales razones de su gran aceptación es que permite describir el comportamiento de la distribución de probabilidad de un vector aleatorio a través de dos objetos: un conjunto de distribuciones de probabilidad univariadas para cada componente del vector aleatorio, denominadas marginales, y una función llamada cópula que contiene la información de la dependencia estocástica entre los componentes. Esto se traduce en una gran aplicabilidad ya que se puede capturar la estructura de dependencia de un grupo de variables aleatorias con la flexibilidad de que las distribuciones marginales pueden tener diferentes distribuciones de probabilidad, abarcando distribuciones continuas, discretas y hasta una mezcla de ambas.

Una cópula C se define como una función que va del cubo unitario de dimensión n al intervalo $[0,1]$ ($C: [0,1]^n \rightarrow [0,1]$) que posee la propiedad de que si se tiene un vector aleatorio $\boldsymbol{U} = (U_1, \dots, U_n)$ tal que cada componente tiene una distribución de probabilidad uniforme en $[0,1]$

[II] Sklar introdujo el nombre de este objeto matemático en los años cincuenta, pero el uso del concepto se remonta algunos años atrás, y los primeros resultados fueron establecidos por Wassily Hoeffding, en los años cuarenta (Nelsen, 2006).

($U_i \sim \mathcal{U}(0,1)\ \forall i = 1, \dots, n$), se cumple que $C(u_1, \dots, u_n) = P(U_1 \leq u_1, \dots, U_n \leq u_1)$ donde $u_1, \dots, u_n \in [0,1]$. De la definición anterior se puede observar el hecho de que la cópula se comporta como una función de distribución para el vector aleatorio $\boldsymbol{U} = (U_1, \dots, U_n)$, restringida sobre el espacio $[0,1]^n$. Esta circunstancia permite relacionar la función de distribución ($F: \mathbb{R}^n \to [0,1]$) de un vector aleatorio $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ con una cópula y sus respectivas funciones de distribución marginales univariadas[III] ($F_k: \mathbb{R} \to [0,1]$ con $k \in \{1, \dots, n\}$). El famoso teorema de Sklar especifica esta relación estableciendo que una función $F: \mathbb{R}^n \to [0,1]$ es una función de distribución de un vector aleatorio $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ si y sólo si existe una cópula $C: [0,1]^n \to [0,1]$ y funciones de distribución univariadas $F_1, \dots, F_n: \mathbb{R} \to [0,1]$ tales que se cumple:

$$F(y_1, \dots, y_n) = C(F_1(y_1), \dots, F_n(y_n)) \tag{8}$$

donde $y_1, \dots, y_n \in \mathbb{R}$ (son realizaciones de las variables aleatorias $Y_1, \dots, Y_n$) y $F_1, \dots, F_n$ son las funciones de distribución marginales de las componentes aleatorias $Y_1, \dots, Y_n$. Si $F_1, \dots, F_n$ son continuas entonces C es única; de lo contrario C está determinada de forma única sobre $ranF_1 \times \dots \times ranF_n$. De esta manera, el teorema de Sklar nos marca un procedimiento para poder describir la función de distribución de las dos variables aleatorias en cuestión: primero hacer un análisis univariado de las componentes marginales y posteriormente describir la dependencia a través de una cópula.

Un punto de partida fundamental en este análisis es la selección de la cópula. En la bibliografía, existe una variedad de familias de cópulas, las cuales en su mayoría se han ido desarrollando por las características inherentes de los fenómenos que se buscan describir, por lo que la elección puede estar sujeta al comportamiento estocástico que se desea explicar (Embrechts, 2009). En este estudio se plantea el uso de la cópula Gaussiana pues aparte de que aparece de forma natural por el resultado del teorema de límite central, tiene varias propiedades de mucha utilidad: i) sirve para describir una amplia variedad de estructuras de dependencia, ii) puede capturar comportamientos desde la cota inferior de Fréchet-Hoeffding (en el caso bivariado),

[III] Por el teorema de la transformación integral de probabilidad si Y_k es una variable aleatoria con función de distribución F_k, la variable aleatoria definida como $U_k = F_k(Y_k)$ está distribuida uniformemente en el intervalo $[0,1]$.

hasta la cota superior de Fréchet-Hoeffding, y iii) matemáticamente es asequible por su cercana conexión con el álgebra lineal.

Dado un vector aleatorio de tamaño m, $\boldsymbol{Y} = (Y_1, \dots, Y_m)$, la cópula Gaussiana se define como:

$$C(F_1(y_1), \dots, F_m(y_m); \boldsymbol{\Gamma}) = \Phi_m(\Phi^{-1}\{F_1(y_1)\}, \dots, \Phi^{-1}\{F_m(y_m)\}|\boldsymbol{\Gamma}) \tag{9}$$

donde Φ^{-1} es la función de distribución univariada normal estándar, Φ_m es la función de distribución multivariada normal estándar con media cero $\mathbf{0}$ y matriz de correlación $\boldsymbol{\Gamma}$ (que parametriza por completo a la cópula Gaussiana), la cual define la correlación de los denominados "scores" normales ($\Phi^{-1}\{U_k\} = \Phi^{-1}\{F_k(y_k)\}$ con $k \in \{1, \dots, m\}$):

$$\gamma_{ij} = corr\left(\Phi^{-1}\{U_i\}, \Phi^{-1}\{U_j\}\right) = corr\left(\Phi^{-1}\{F_i(y_i)\}, \Phi^{-1}\{F_j(y_j)\}\right). \tag{10}$$

Por la relación anterior, la matriz $\boldsymbol{\Gamma}$ tiene la peculiaridad de ser una matriz simétrica y positiva definida, con entradas en la diagonal iguales a 1. Cuando se tienen marginales continuas, γ_{ij} representa la correlación lineal de los scores normales; en el caso de que una marginal sea discreto y otro continuo, γ_{ij} se puede interpretar como la correlación poliserial entre la discretización de una variable latente y una continua, mientras que si ambos marginales son discretos el parámetro de dependencia se puede interpretar como la correlación policórica entre la discretización de variables continuas latentes (Drasgow, 2004).

Siguiendo la metodología propuesta por Song et al. (2009), es posible usar la cópula Gaussiana para obtener la función de probabilidad conjunta de las variables aleatorias mixtas de interés: tiempo de supervivencia T y el tipo de causa D. En el caso general, si se tienen m variables aleatorias continuas denotadas por $Y_1, \dots, Y_m$ con funciones de densidad $f_1(y_1; \boldsymbol{\theta}_1), \dots, f_m(y_m; \boldsymbol{\theta}_m)$ (donde $\boldsymbol{\theta}_j$ es el vector de parámetros que describe a cada función de densidad f_j) la función de densidad conjunta se obtiene derivando la ecuación (9) respecto a $y_1, \dots, y_m$ (Song et al., 2009):

$$f_{1,2,\dots,m}(y_1, \dots, y_m; \boldsymbol{\theta}, \boldsymbol{\Gamma}) = |\boldsymbol{\Gamma}|^{-1/2} \exp\left\{\frac{1}{2}\boldsymbol{q}^T(\boldsymbol{I}_m - \boldsymbol{\Gamma}^{-1})\boldsymbol{q}\right\} \prod_{i=1}^{m} f_i(y_i; \boldsymbol{\theta}_i) \tag{11}$$

donde $\boldsymbol{q} = (q_1, \dots, q_m) = (\Phi^{-1}\{F_1(y_1; \boldsymbol{\theta}_1)\}, \dots, \Phi^{-1}\{F_m(y_m; \boldsymbol{\theta}_m)\})$ es el vector de scores normales, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m)$ es el vector de parámetros de las componentes marginales y $\boldsymbol{I}_m$ es la matriz identidad de dimensión $m \times m$.

Por otro lado, si las variables aleatorias $Y_1, \dots, Y_m$ son discretas, con funciones de distribución $F_1(y_1; \boldsymbol{\theta}_1), \dots, F_m(y_m; \boldsymbol{\theta}_m)$, la función de probabilidad conjunta se obtiene de la derivada de Radon-Nikodym de la ecuación (9) respecto a la medida de conteo (Song et al., 2009):

$$f_{1,2,\dots,m}(y_1, \dots, y_m; \boldsymbol{\theta}, \boldsymbol{\Gamma}) = \sum_{j_1=1}^{2} \dots \sum_{j_m=1}^{2} (-1)^{j_1+\dots+j_m} \mathrm{C}\left(u_{1,j_1}, \dots, u_{m,j_m} | \boldsymbol{\Gamma}\right) \tag{12}$$

donde cada j_k (con $k = 1, \dots, m$) representa sólo los valores 1 ó 2 para cada k, por lo que u_{k,j_k} puede estar dado por $u_{k,1} = F_k(y_k-; \boldsymbol{\theta}_k)$ y $u_{k,2} = F_k(y_k; \boldsymbol{\theta}_k)$; aqui $F_k(y_k-; \boldsymbol{\theta}_k)$ denota el límite por la izquierda de la función de distribución F_k en y_k. Cuando la variable aleatoria toma valores en los enteros se tiene que $F_k(y_k-; \boldsymbol{\theta}_k) = F_k(y_k - 1; \boldsymbol{\theta}_k)$.

Finalmente, si de las m variables aleatorias se tiene una mezcla de m_1 variables continuas y $m_2 = m - m_1$ variables discretas, la función de probabilidad conjunta se obtiene a través de lo siguiente. Sean $\boldsymbol{u}_1 = (u_1, \dots., u_{m_1})$ y $\boldsymbol{u}_2 = (u_{m_1+1}, \dots., u_m)$ dos vectores con dimensiones m_1 y m_2 respectivamente, y con componentes marginales que siguen una distribución uniforme en el intervalo [0,1]. Aplicando la notación anterior a los vectores $\boldsymbol{q} = (q_1, \dots, q_m) = (\boldsymbol{q}_1, \boldsymbol{q}_2)$ y $\boldsymbol{y} = (y_1, \dots, y_m) = (\boldsymbol{y}_1, \boldsymbol{y}_2)$, se define la siguiente función (Song et al., 2009):

$$C^*(\boldsymbol{u}_1, \boldsymbol{u}_2 | \boldsymbol{\Gamma}) = (2\pi)^{-m_2/2} |\boldsymbol{\Gamma}|^{-1/2} \int_{-\infty}^{\Phi^{-1}(u_{m_1+1})} \dots \int_{-\infty}^{\Phi^{-1}(u_m)} \exp\left\{-\frac{1}{2}(\boldsymbol{q}_1, \boldsymbol{y}_2)\boldsymbol{\Gamma}^{-1}(\boldsymbol{q}_1, \boldsymbol{y}_2)^T + \frac{1}{2}\boldsymbol{q}_1^T\boldsymbol{q}_1\right\} d\boldsymbol{y}_2 \,. \tag{13}$$

Entonces, la función de probabilidad conjunta está dada por:

$$f_{1,\dots,m}(y_1, \dots, y_m; \boldsymbol{\theta}, \boldsymbol{\Gamma}) =$$

$$\prod_{i=1}^{m_1} f_i(y_i; \boldsymbol{\theta}_i) \sum_{j_{m_1+1}=1}^{2} \dots \sum_{j_m=1}^{2} (-1)^{j_{m_1+1}+\dots+j_m} C^*\left(F_1(y_1; \boldsymbol{\theta}_1), \dots, F_{m_1}(y_{m_1}; \boldsymbol{\theta}_{m_1}), u_{m_1+1,j_{m_1+1}}, \dots, u_{m,j_m} \middle| \boldsymbol{\Gamma}\right) \tag{14}$$

donde las u_{k,j_k} están definidas como en el caso discreto, por lo que u_{k,j_k} sólo toma dos valores: $u_{k,1}$ y $u_{k,2}$.

Usando las relaciones anteriores se puede obtener la función de probabilidad conjunta del tiempo de supervivencia T (dependiendo si se considera variable aleatoria continua como en el modelo paramétrico o variable aleatoria discreta como en el modelo semiparamétrico) y de la causa de evento D (variable discreta que toma valores en $\{1, \dots, d, \dots, J\}$).

Cuando la variable aleatoria T es continua (modelo paramétrico), de acuerdo a la ecuación (14), la función de probabilidad conjunta está dada por la siguiente ecuación:

$$f_{T,D}(t,d;\boldsymbol{\theta},\rho) = f_T(t;\boldsymbol{\theta}_T)\sum_{j_d=1}^{2}(-1)^{j_d}C^*\big(F_T(t;\boldsymbol{\theta}_T),u_{d,j_d}\big|\rho\big). \tag{15}$$

donde $f_T(t;\boldsymbol{\theta}_T)$ es la función de densidad del tiempo de supervivencia con vector de parámetros $\boldsymbol{\theta}_T$, $F_T(t;\boldsymbol{\theta}_T)$ es la función de distribución del tiempo de supervivencia, j_d toma solamente dos valores $\{1,2\}$, $u_{d,1} = F_D(d-1;\boldsymbol{\theta}_D)$ y $u_{d,2} = F_D(d;\boldsymbol{\theta}_D)$ donde $F_D(d;\boldsymbol{\theta}_D)$ es la función de distribución de la causa de evento con vector de parámetros $\boldsymbol{\theta}_D$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_T,\boldsymbol{\theta}_D)$ y ρ el único parámetro que determina a la matriz correlación de scores normales (debido a que solamente se tienen dos variables aleatorias) y está dado por:

$$\rho = corr(\Phi^{-1}\{F_T(t;\boldsymbol{\theta}_T)\},\Phi^{-1}\{F_D(d;\boldsymbol{\theta}_D)\}). \tag{16}$$

La función $C^*(F_T(t;\boldsymbol{\theta}_T),F_D(d;\boldsymbol{\theta}_D)\,|\rho)$ en el caso bivariado se obtiene derivando parcialmente la cópula bivariada $C(u,v|\rho)$ respecto a la variable u (que corresponde a la marginal continua):

$$C^*(u,v|\rho)\ := \frac{\partial}{\partial u}C(u,v|\rho) \tag{17}$$

En dos dimensiones, la expresión $C^*(u,v|\rho)$ representa la densidad condicional de la variable aleatoria $v = F_D(d;\boldsymbol{\theta}_D)$ dado el valor de la variable aleatoria $u = F_T(t;\boldsymbol{\theta}_T)$ (Krämer et al., 2013). En el caso específico de la cópula Gaussiana bivariada, $C^*(u,v|\rho)$ se puede representar analíticamente como (Czado et al., 2012):

$$C^*(u,v|\rho) = \Phi\left(\frac{\Phi^{-1}(v)-\rho\Phi^{-1}(u)}{\sqrt{1-\rho^2}}\right). \tag{18}$$

Usando la ecuación (18), la función de probabilidad conjunta expresada en la ecuación (15), queda como:

$$f_{T,D}(t,d;\boldsymbol{\theta},\rho) =$$

$$f_T(t;\boldsymbol{\theta}_T)\left(\Phi\left(\frac{\Phi^{-1}(F_D(d;\boldsymbol{\theta}_D))-\rho\Phi^{-1}\big(F_T(t;\boldsymbol{\theta}_T)\big)}{\sqrt{1-\rho^2}}\right)-\Phi\left(\frac{\Phi^{-1}(F_D(d-1;\boldsymbol{\theta}_D))-\rho\Phi^{-1}\big(F_T(t;\boldsymbol{\theta}_T)\big)}{\sqrt{1-\rho^2}}\right)\right). \tag{19}$$

Para el modelo semiparamétrico, se considera a la variable aleatoria T como discreta, dada por la parametrización discreta de la función de riesgo (Yilmaz y Lawless, 2011), donde la ocurrencia de eventos ordenados (sin considerar los censurados) dados por $t_1^* < \cdots < t_k^*$ con $l = 1, \ldots, k$ y $k \leq n$, determina los saltos en la función de distribución $F_T(t)$. La función de probabilidad conjunta $f_{T,D}(t, d; \boldsymbol{\theta}, \rho)$ parte ahora de la ecuación (12), obteniéndose para el caso bivariado:

$$f_{T,D}(t, d; \boldsymbol{\theta}, \rho) = \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} (-1)^{j_1+j_2} \, \mathrm{C}\big(u_{1,j_1}, u_{2,j_2} | \rho\big). \tag{20}$$

Considerando que $t_l^* \leq t < t_{l+1}^*$ y sustituyendo los valores $u_{1,1} = F_T(t_l^*; \boldsymbol{\theta}_T)$, $u_{1,2} = F_T(t_{l-1}^*; \boldsymbol{\theta}_T)$, $u_{2,1} = F_D(d-1; \boldsymbol{\theta}_D)$, $u_{2,2} = F_D(d; \boldsymbol{\theta}_D)$, se tiene:

$$\begin{aligned} f_{T,D}(t, d; \boldsymbol{\theta}, \rho) = {} & \mathrm{C}(F_T(t_l^*; \boldsymbol{\theta}_T), F_D(d; \boldsymbol{\theta}_D) | \rho) - \mathrm{C}(F_T(t_l^*; \boldsymbol{\theta}_T), F_D(d-1; \boldsymbol{\theta}_D) | \rho) \\ & - \mathrm{C}(F_T(t_{l-1}^*; \boldsymbol{\theta}_T), F_D(d; \boldsymbol{\theta}_D) | \rho) + \mathrm{C}(F_T(t_{l-1}^*; \boldsymbol{\theta}_T), F_D(d-1; \boldsymbol{\theta}_D) | \rho) \end{aligned} \tag{21}$$

donde $\mathrm{C}(F_T(t_l^*; \boldsymbol{\theta}_T), F_D(d; \boldsymbol{\theta}_D) | \rho)$ representa a la cópula Gaussiana bivariada dada por $\Phi_2(\Phi^{-1}\{F_T(t_l^*; \boldsymbol{\theta}_T)\}, \Phi^{-1}\{F_D(d; \boldsymbol{\theta}_D)\} | \rho)$.

2.2.- Especificaciones marginales

Una vez establecido que el modelo de la dependencia entre variables se hará mediante la cópula Gaussiana, es necesario entonces especificar los modelos marginales y establecer cómo se incorporan las variables explicativas para el modelo paramétrico y semiparamétrico.

2.2.1.- Marginal para el tiempo de supervivencia

En la introducción se hace la mención de que la variable del tiempo de supervivencia es aquella que determina si el modelo es paramétrico o semiparamétrico.

Para el modelo completamente paramétrico se opta por que la variable aleatoria T siga una distribución Weibull con función de densidad $f_T(t; \nu, \kappa)$ y función de distribución $F_T(t; \nu, \kappa)$ dadas por:

$$f_T(t;\nu,\kappa) = \frac{\nu}{\kappa}\left(\frac{t}{\kappa}\right)^{\nu-1} \exp\left(-\left(\frac{t}{\kappa}\right)^{\nu}\right) \tag{22}$$

$$F_T(t;\nu,\kappa) = 1 - \exp\left(-\left(\frac{t}{\kappa}\right)^{\nu}\right) \tag{23}$$

$\nu > 0$ y $\kappa > 0$ son los parámetros de forma y escala respectivamente; esta distribución tiene como caso particular la distribución exponencial cuando el parámetro de forma es igual a uno ($\nu = 1$).

Para agregar predictores a esta marginal se emplea el modelo paramétrico de Cox de riesgos proporcionales. Como se vio en la sección 1.1, el modelo de regresión de Cox se puede representar a través de la ecuación (3). La proporcionalidad hace referencia al hecho de que la función de riesgo $h(t)$ es igual a la función de riesgo de base $h_0(t)$ multiplicada por un término $\psi = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$, que al ser constante en el tiempo, hace que la relación sea de proporcionalidad. Dada la correspondencia de la función de riesgo con la función de supervivencia, la suposición de proporcionalidad implica que las $S(t)$ para diversos riesgos no se crucen. Además, esta misma relación permite expresar al modelo de Cox de riesgos proporcionales en términos de funciones de supervivencia:

$$S_T(t;\boldsymbol{x},\boldsymbol{\beta},\nu,\kappa) = \{S_0(t;\nu,\kappa)\}^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})} \tag{24}$$

donde $S_0(t;\nu,\kappa)$ es la función de supervivencia de base y $\boldsymbol{\beta}$ es el vector de parámetros (que no incluye intercepto).

Para que se tenga un modelo paramétrico, la función de supervivencia de base debe seguir una familia paramétrica, para la cual se ha seleccionado la distribución Weibull:

$$S_0(t;\nu,\kappa) = \exp\left(-\left(\frac{t}{\kappa}\right)^{\nu}\right). \tag{25}$$

Dado lo anterior, la representación de las funciones de densidad y función de distribución de la variable aleatoria T dados los valores del vector de covariables $\boldsymbol{x}$ y sus parámetros están dadas por las siguientes relaciones:

$$f_T(t;\boldsymbol{x},\boldsymbol{\beta},\nu,\kappa) = f_T(t;\boldsymbol{x},\boldsymbol{\theta}_T) = \frac{\gamma}{\kappa^{\nu}} t^{\nu-1} \exp(\boldsymbol{\beta}^T\boldsymbol{x}) \exp\left(-\left(\frac{t}{\kappa}\right)^{\nu}\right)^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})} \tag{26}$$

$$F_T(t;\boldsymbol{x},\boldsymbol{\beta},\nu,\kappa) = F_T(t;\boldsymbol{x},\boldsymbol{\theta}_T) = 1 - \exp\left(-\left(\frac{t}{\kappa}\right)^{\nu}\right)^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})} \tag{27}$$

donde $\boldsymbol{\theta}_T$ es el vector de parámetros de la marginal del tiempo $\boldsymbol{\theta}_T = (\boldsymbol{\beta}, \nu, \kappa)$.

Respecto al modelo semiparamétrico se propone que la función de supervivencia de base sea una función escalón con saltos solamente en valores de tiempo donde se observa la ocurrencia de eventos, es decir, no se toman en cuenta los tiempos donde hay censura. Se emplea entonces la parametrización discreta de la función de riesgo (Yilmaz y Lawless, 2011) definida como:

$$\lambda_l^* = \frac{S_0(t_l^*) - S(t_l^{*+})}{S(t_l^*)} \tag{28}$$

donde dadas n observaciones, t_l^* son los tiempos de la ocurrencia de eventos ordenados $t_1^* < \cdots < t_k^*$ con $l = 1, \ldots, k$ donde $k \leq n$. Las funciones de supervivencia de base para el modelo semiparamétrico $S_0(t)$ y $S_0(t^+)$ se definen como:

$$S_0(t) = \prod_{l:t_l^*<t} (1 - \lambda_l^*) \tag{29}$$

$$S_0(t^+) = \prod_{l:t_l^*\leq t} (1 - \lambda_l^*) \tag{30}$$

donde las notaciones $l: t_l^* < t$ o $l: t_l^* \leq t$ indican que la multiplicación corre sobre el índice l que es el número de observaciones ordenadas t_l^* que son menores ó menores o igual a t.

La incorporación de covariables $\boldsymbol{x}$, bajo un modelo de riesgos proporcionales, se logra usando la ecuación (24), por lo que la función de supervivencia $S_T(t;\boldsymbol{x},\boldsymbol{\beta},\lambda_l^*)$ está dada por:

$$S_T(t;\boldsymbol{x},\boldsymbol{\beta},\lambda_l^*) = S_0(t;\lambda_l^*)^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})} = \prod_{l:t_l^*<t} (1 - \lambda_l^*)^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})} \tag{31}$$

donde nuevamente $\boldsymbol{\beta}$ es el vector de parámetros (sin incluir intercepto). La representación de la ecuación (31) se puede ver también como un modelo de riesgos relativos discreto descrito en Kalbfleisch y Prentice (2011), el cual se caracteriza por la función de riesgo discreta λ_l^*.

Para constatar que el modelo paramétrico cumple con los supuestos de proporcionalidad de riesgos y de ajuste a través de una distribución Weibull, se puede hacer uso de procedimientos

visuales, como la gráfica de $\log\{-\log(S(t))\}$ contra $\log(t)$ (Collett, 2015). La función de supervivencia, de acuerdo a la ecuación (27), está dada por:

$$S_T(t;\boldsymbol{x},\boldsymbol{\beta},\nu,\kappa) = \exp\left(-\left(\frac{t}{\kappa}\right)^{\nu}\right)^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})}. \tag{32}$$

Aplicando logaritmo natural a ambos lados de la ecuación, desarrollando la expresión y multiplicando por menos uno se tiene:

$$-\log\big(S_T(t;\boldsymbol{x},\boldsymbol{\beta},\nu,\kappa)\big) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})\left(\frac{t}{\kappa}\right)^{\nu}.$$

Aplicando de nuevo logaritmo natural, después de varios pasos algebraicos se obtiene la expresión:

$$\log\left\{-\log\big(S_T(t;\boldsymbol{x},\boldsymbol{\beta},\nu,\kappa)\big)\right\} = \boldsymbol{\beta}^T\boldsymbol{x} - \nu\log(\kappa) + \nu\log(t). \tag{33}$$

De esta forma, sustituyendo el estimador de Kaplan-Meier para estimar la función de supervivencia $\hat{S}(t)$, la relación anterior produce líneas rectas en $\log(t)$, con pendiente ν y ordenada al origen $\boldsymbol{\beta}^T\boldsymbol{x} - \nu\log(\kappa)$. Para distintos valores de los predictores $\boldsymbol{\beta}^T\boldsymbol{x}$, se tienen líneas paralelas (si los supuestos se cumplen).

Para el modelo semiparamétrico, también se puede usar el procedimiento gráfico de $\log\{-\log(S(t))\}$ contra $\log(t)$ para corroborar la proporcionalidad de riesgos (Collett, 2015). En este caso, de la ecuación (31) se observa que al aplicar la transformación $\log\{-\log(S(t))\}$ se obtiene la siguiente expresión:

$$\log\left\{-\log\big(S_T(t;\boldsymbol{x},\boldsymbol{\beta},\lambda_l^*)\big)\right\} = \boldsymbol{\beta}^T\boldsymbol{x} + \log\left\{-\log\big(S_0(t;\lambda_l^*)\big)\right\} \tag{34}$$

La relación anterior indica que la transformación $\log\{-\log(S(t;t;\boldsymbol{x},\boldsymbol{\beta},\lambda_l^*))\}$ es igual a la transformación de la función de supervivencia de base $\log\{-\log(S_0(t;\lambda_l^*))\}$ más una término que no depende del tiempo $\boldsymbol{\beta}^T\boldsymbol{x}$, por lo que al sustituir el estimador de Kaplan-Meier para estimar las funciones de supervivencia, se obtienen curvas paralelas, que a diferencia del modelo paramétrico, no necesariamente deben ser líneas rectas.

Es importante mencionar que, cuando no se cumple la suposición de riesgos proporcionales, los modelos propuestos se pueden extender agregando variables que dependan del tiempo, las cuales se integran a la componente lineal a través de términos formados por la multiplicación

de la(s) variable(s) que no cumple(n) la proporcionalidad de riesgos por una función del tiempo, que por lo regular es lineal, cuadrática o logarítmica $\{t,\ t^2,\ \log(t)\}$ (Bellera et al., 2010). Para ser un poco más explícitos, supongamos que un subconjunto de covariables no cumple con la proporcionalidad de riesgos, denotado por $\boldsymbol{x}'(t) = \left(x_1(t), \dots, x_J(t)\right)$, donde cada término está formado por el producto de la variable x_j con algún término del conjunto $\{t,\ t^2,\ \log(t)\}$, por ejemplo $x_j(t) = x_j * t^2$. De esta forma, la componente lineal estará dada por $\boldsymbol{\beta}^T\boldsymbol{x} + \boldsymbol{\beta}'^T\boldsymbol{x}'(t)$ donde el vector de parámetros $\boldsymbol{\beta}'$ caracteriza a las variables que dependen del tiempo.

2.2.2.- Marginal para el tipo de evento

El análisis de datos de supervivencia también se puede interpretar como un proceso estocástico X_t, en el cual los individuos están en un estado inicial (denotado por 0), pudiendo pasar a otro estado (denotado por 1) conforme trascurre el tiempo, el cual se denomina de estado de absorción debido a que una vez que lo alcanzan, no pueden salir de él.(Beyersmann et al., 2011) La generalización en el ámbito de riesgos competitivos es tener el mismo estado inicial, pero ahora transitar a J estados de absorción. Desde este punto de vista, los distintos estados se absorción (que se pueden interpretar como la variable tipo de evento D) se pueden modelar a través de una distribución multinomial, pues al conocer la probabilidad de ocurrencia de $J-1$ estados, automáticamente se conoce la probabilidad de ocurrencia del estado faltante.

En presencia de covariables, la marginal correspondiente al tipo de evento D, se puede modelar a través de un modelo multinomial, que comúnmente se emplea para el modelado marginal de la variable D en los modelos mixtos de riesgos competitivos, Lau et al., 2008). La probabilidad de ocurrencia de un evento $D = d$ dado un vector de covariables $\boldsymbol{z}$ se establece como:

$$P(D = d|\boldsymbol{z}) = \frac{\exp(\delta_{0d} + \boldsymbol{\delta}_d^T\boldsymbol{z})}{1 + \exp(\delta_d + \boldsymbol{\delta}_d^T\boldsymbol{z})} \tag{35}$$

donde δ_{0d} es un escalar y $\boldsymbol{\delta}_d$ un vector que indica los parámetros del modelo cuando $D = d$.

La metodología propuesta se aplica a un conjunto de datos para los cuales existen dos riesgos que compiten (capítulo 4), por lo que en este caso específico la variable aleatoria D posee una distribución Bernoulli; en términos del vector de covariables $\boldsymbol{z}$ y sus respectivos parámetros

(quitándoles el subíndice d que no necesario pues sólo se tienen dos eventos) δ_0 y $\boldsymbol{\delta}$, las funciones de densidad y de distribución están dadas por:

$$f_D(d;\boldsymbol{z},\delta_0,\boldsymbol{\delta}) = f_D(d;\boldsymbol{z},\boldsymbol{\theta}_D) = \begin{cases} P(D=1|\boldsymbol{z}) & d=1 \\ 1-P(D=1|\boldsymbol{z}) & d=2 \end{cases} \tag{36}$$

$$F_D(d;\boldsymbol{z},\delta_0,\boldsymbol{\delta}) = F_D(d;\boldsymbol{z},\boldsymbol{\theta}_D) = \begin{cases} 0 & d<1 \\ P(D=1|\boldsymbol{z}) & 1\leq d<2 \\ 1 & d\geq 2 \end{cases} \tag{37}$$

donde $\boldsymbol{\theta}_D = (\delta_0, \boldsymbol{\delta})$.

2.3.- Riesgos competitivos y el modelo de regresión de cópula Gaussiana

Como se mencionó en la introducción, existen algunas expresiones importantes en la teoría de análisis de supervivencia bajo la presencia riesgos competitivos como son la función de incidencia acumulada y la función de riesgo de causa específica. Usando los enfoques propuestos, se pueden obtener expresiones a través de cópulas para lograr su representación.

La función de incidencia acumulada $CIF_d(t)$ definida como $P(T\leq t, D=d)$, se puede obtener usando la diferencia $P(T\leq t, D\leq d) - P(T\leq t, D\leq d-1)$ debido a que la variable aleatoria D es discreta. Esta diferencia corresponde a lo que se ha mencionado con anterioridad, como la derivada de Radon-Nikodym de la función de distribución (en este caso la cópula) respecto a la medida de conteo (Song, 2009):

$$CIF_d(t) =$$

$$P(T\leq t, D=d) = \frac{\partial}{\partial d} C(F_T(t), F_D(d)\,) = C(F_T(t), F_D(d)\,) - C(F_T(t), F_D(d-1)\,). \tag{38}$$

Retomando el caso particular de que se tengan dos eventos $D=\{1,2\}$, la función de incidencia acumulada para el evento $d=1$ queda como:

$$CIF_1(t) = P(T\leq t, D=1) =$$

$$\{C(F_T(t), F_D(1)\,) - C(F_T(t), F_D(0)\,)\} = \{C(F_T(t), F_D(1)\,) - C(F_T(t), 0\,)\}.$$

Usando el hecho de que $F_D(1) = p$ (ecuación (37)) y que $C(u, 0\ ) = 0$ (Nelsen, 2006) se tiene que:

$$CIF_1(t) = P(T \leq t, D = 1) = C(F_T(t), p\ ). \tag{39}$$

Por otro lado, la función de incidencia acumulada para el evento $d = 2$ se expresa como:

$$CIF_2(t) = P(T \leq t, D = 2) =$$

$$\{C(F_T(t), F_D(2)\ ) - C(F_T(t), F_D(1)\ )\} = \{C(F_T(t), 1\ ) - C(F_T(t), p)\}$$

donde se ha empleado el hecho de que $F_D(2) = 1$ y $F_D(1) = p$ (ecuación (37)). Teniendo en cuenta que $C(u, 1\ ) = u$ (Nelsen, 2006) y que $F_T(t) = 1 - S(T)$ se tiene finalmente que:

$$CIF_2(t) = P(T \leq t, D = 2) = 1 - S_T(t) - C(F_T(t), p). \tag{40}$$

La función de riesgo de causa específica $h_d(t)$ dada por la ecuación (4), se puede representar de forma diferente usando la relación de probabilidad condicional $P(A|B) = P(A, B)/P(B)$:

$$h_d(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, D = d)}{\delta t} \frac{1}{P(T > 1)} =$$

$$\lim_{\delta t \to 0} \frac{P(T \leq t + \delta t, D = d) - P(T \leq t, D = d)}{\delta t} \frac{1}{P(T > t)} = f_d(t) \frac{1}{S(t)}$$

donde se ha usado el hecho de que $P(T > t) = S(t)$ (función de supervivencia total), y se ha considerado al límite como la función de densidad de causa específica $f_d(t)$ (Collett, 2015):

$$f_d(t) = \lim_{\delta t \to 0} \frac{P(T \leq t + \delta t, D = d) - P(T \leq t, D = d)}{\delta t}. \tag{41}$$

En términos de la función de distribución bivariada $P(T \leq t, D \leq d)$, la función de densidad de causa específica $f_d(t)$ equivale a la función de probabilidad $P(T = t, D = d) = f_{TD}(t, d)$. Por tal motivo, la función de riesgo de causa específica se puede reescribir como:

$$h_d(t) = \frac{f_{TD}(t, d)}{S(t)} \tag{42}$$

donde en el caso paramétrico la función de probabilidad conjunta $f_{TD}(t, d)$ está dada por la ecuación (15) mientras que para el caso semiparamétrico está dada por la ecuación (21).

La inclusión de covariables $\boldsymbol{x}$ para la variable del tiempo de supervivencia y $\boldsymbol{z}$ para la variable tipo de evento se hace directamente en las marginales $F_T(t|\boldsymbol{x})$ y $F_D(d|\boldsymbol{z})$ de la cópula y no afecta las relaciones anteriores. Además, las expresiones anteriores se han derivado para cualquier cópula, siendo necesario sustituir la cópula Gaussiana (ecuación (9)) para hacer los cálculos de acuerdo al modelo propuesto.

3.- Proceso de estimación del modelo

En este capítulo se indica el procedimiento para la estimación de los parámetros del modelo propuesto empleando verosimilitud, describiendo su obtención, así como el proceso de optimización. Para analizar el comportamiento de los estimadores en muestras finitas, se realizan una serie de simulaciones, variando algunos aspectos como son la cantidad de censura, la magnitud de la dependencia entre variables, y el proceso generador de los datos.

3.1.- Función de verosimilitud

La propuesta para modelar la función de probabilidad de 2 variables aleatorias mixtas $P(T, D)$ a través de la cópula Gaussiana con la inclusión de covariables para cada marginal ($\boldsymbol{x}$ y $\boldsymbol{z}$ respectivamente), conlleva la estimación de los parámetros marginales del tiempo, que para el modelo paramétrico son $\boldsymbol{\theta}_T = (\boldsymbol{\beta}, \nu, \kappa)$ y para el semiparamétrico son $\boldsymbol{\theta}_T = (\boldsymbol{\beta}, \lambda_l^*)$ ($\boldsymbol{\beta}$ es el vector de parámetros de los riesgos proporcionales, ν, κ son los parámetros de forma y escala de una función de base Weibull, y λ_l^* es la función de riesgo discreta), de los parámetros marginales del tipo de evento $\boldsymbol{\theta}_D = (\delta_{0d}, \boldsymbol{\delta}_d)$ ($\delta_{0d}, \boldsymbol{\delta}_d$ son los parámetros de las componentes lineales del modelo multinomial para el cual d varía de acuerdo al tipo de causas $\{1, \dots, d, \dots, J\}$) y del parámetro de dependencia ρ.

En un problema de riesgos competitivos con n individuos en estudio se presentan $\tilde{T}_i$ tiempos de eventos y D_i ocurrencias de los tipos de eventos (donde $i = 1, \dots, n$). Con fines ilustrativos, el desarrollo de la función de verosimilitud se realizará tomando en cuenta solamente dos tipos de evento[IV] $D_i = \{1, 2\}$; la incorporación de más tipos eventos es una extensión del procedimiento.

La presencia de censura se denota con la variable $Cens_i$, considerando solamente censura por derecha pues el tipo más frecuente (Kleinbaum y Klein, 2010). El tiempo de observación T_i se define como $T_i = \min(\tilde{T}_i, Cens_i)$. Se definen dos indicadores para los tipos de evento $\varphi_i = I(D_i = 1)$ y $\psi_i = I(D_i = 2)$ (donde la censura por la derecha se obtiene si $\varphi_i = \psi_i = 0$). De

[IV] El modelo multinomial para la marginal D se convierte en un modelo de regresión logística.

esta forma, el conjunto de datos observados está conformado por n observaciones independientes $y_i = \{T_i, D_i, \varphi_i, \psi_i\}$ donde $\boldsymbol{z}_i$ y $\boldsymbol{x}_i$ son los vectores de covariables para cada modelo marginal.

Bajo un escenario con $Cens_i$ no informativa e independiente, así como independencia condicional entre el vector aleatorio (T_i, D_i) y $Cens_i$ dados los valores de los vectores de covariables $\boldsymbol{z}_i$ y $\boldsymbol{x}_i$, la contribución del i-ésimo individuo a la función de verosimilitud está dada por la función de probabilidad conjunta $P(T_i = t_i, D_i = d_i)$ para datos no censurados o por $P(T_i > t_i)$ cuando se presenta censura (Larson y Dinse, 1985 y Lau et al., 2008). Sea $\boldsymbol{\omega}$ el vector de parámetros, $L_i(\boldsymbol{\omega}|y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$ queda como:

$$L_i(\boldsymbol{\omega}|y_i, \boldsymbol{z}_i, \boldsymbol{x}_i) = L_i(\boldsymbol{\omega}|T_i, D_i, \varphi_i, \psi_i, \boldsymbol{z}_i, \boldsymbol{x}_i) = \\ P(T_i = t_i, D_i = 1|\boldsymbol{z}_i, \boldsymbol{x}_i)^{\varphi_i} P(T_i = t_i, D_i = 2|\boldsymbol{z}_i, \boldsymbol{x}_i)^{\psi_i} P(T_i > t_i|\boldsymbol{x}_i)^{1-\varphi_i-\psi_i}. \tag{43}$$

La caracterización de $\boldsymbol{\omega}$, $P(T_i = t_i, D_i = d_i)$ y $P(T_i > t_i)$ dependen del modelo propuesto.

3.1.1.- Función de verosimilitud del modelo paramétrico

En este modelo $\boldsymbol{\omega}$ está conformado por los parámetros de la regresión logística δ_0 y $\boldsymbol{\delta}$, aquellos de la componente lineal del modelo de riesgos proporcionales $\boldsymbol{\beta}$, v y κ de la función de distribución Weibull de la función de supervivencia de base y el parámetro ρ de dependencia de la cópula Gaussiana, es decir $\boldsymbol{\omega} = (\delta_0, \boldsymbol{\delta}, \boldsymbol{\beta}, v, \kappa, \rho)$.

Elementos importantes para plantear la función de verosimilitud son las funciones de distribución de las variables T_i y D_i, dadas por las ecuaciones (27) y (37), respectivamente.

La determinación de $P(T_i = t_i, D_i = d_i|\boldsymbol{z}_i, \boldsymbol{x}_i)$ como función del vector de parámetros $\boldsymbol{\omega}$, se establece a través de la función de probabilidad $f_{T,D}(t, d; \boldsymbol{\theta}_T, \boldsymbol{\theta}_D, \rho)$ dada por la ecuación (19), pues este caso se tiene una variable aleatoria continua y una variable aleatoria discreta. Es importante mencionar que cuando solamente se tienen dos tipos de causa $D_i = \{1, 2\}$, la ecuación (19) se simplifica debido a lo siguiente.

Cuando $D_i = 1$, se tiene que evaluar $C^*\left(F_{T_i}(t_i|\boldsymbol{x}_i), F_{D_i}(0|\boldsymbol{z}_i)|\rho\right)$:

$$C^*\left(F_{T_i}(t_i|\boldsymbol{x}_i), F_{D_i}(0|\boldsymbol{z}_i)|\rho\right) =$$

$$\Phi\left(\frac{\Phi^{-1}\left(F_{D_i}(0|\boldsymbol{z}_i)\right) - \rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right) = \Phi\left(\frac{\Phi^{-1}(0) - \rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right) = 0.$$

Por otro lado, cuando $D_i = 2$, se tiene que evaluar $C^*\left(F_{T_i}(t_i|\boldsymbol{x}_i), F_{D_i}(2|\boldsymbol{z}_i)|\rho\right)$:

$$C^*\left(F_{T_i}(t_i|\boldsymbol{x}_i), F_{D_i}(2|\boldsymbol{z}_i)|\rho\right) =$$

$$\Phi\left(\frac{\Phi^{-1}\left(F_{D_i}(2|\boldsymbol{z}_i)\right) - \rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right) = \Phi\left(\frac{\Phi^{-1}(1) - \rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right) = 1.$$

Por lo anterior, la ecuación (19) queda como:

$$P(T_i = t_i, D_i = d_i|\boldsymbol{z}_i, \boldsymbol{x}_i) =$$

$$\begin{cases} f_{T_i}(t_1|\boldsymbol{x}_i)\left\{1 - \Phi\left(\dfrac{\Phi^{-1}\left(F_{D_i}(1|\boldsymbol{z}_i)\right) - \rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right)\right\} & si\ d_i = 2 \\ f_{T_i}(t_1|\boldsymbol{x}_i)\left\{\Phi\left(\dfrac{\Phi^{-1}\left(F_{D_i}(1|\boldsymbol{z}_i)\right) - \rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right)\right\} & si\ \ d_i = 1. \end{cases} \quad (44)$$

En la expresión anterior el parámetro de dependencia aparece de forma explícita, mientras que los parámetros δ_0 y $\boldsymbol{\delta}$ están contenidos en $F_{D_i}(1|\boldsymbol{z}_i)$; los parámetros $\boldsymbol{\beta}$, v y κ se expresan en las funciones $f_{T_i}(t_i|\boldsymbol{x}_i)$ y $F_{T_i}(t_i|\boldsymbol{x}_i)$.

Juntando las expresiones de $P(T_i = t_i, D_i = 1|\boldsymbol{z}_i, \boldsymbol{x}_i)$ y $P(T_i = t_i, D_i = 2|\boldsymbol{z}_i, \boldsymbol{x}_i)$ con el término $P(T_i > t_i|\boldsymbol{x}_i)$, que está dado por la ecuación (32), la función de log-verosimilitud $l(\boldsymbol{\omega}; \mathbf{Y}, \mathbf{Z}, \mathbf{X})$ (donde $\boldsymbol{Y}$ es una matriz cuya i-ésima fila está formada por $y_i = \{T_i, D_i, \varphi_i, \psi_i\}$, $\mathbf{Z}$ y $\mathbf{X}$ son matrices de covariables cuyas i-ésimas filas están dadas por $\boldsymbol{z}_i$ y $\boldsymbol{x}_i$ respectivamente) queda como:

$$l(\boldsymbol{\omega}|\mathbf{Y},\mathbf{Z},\mathbf{X}) =$$

$$\sum_{\{i:\psi_i=1\}} \text{Ln}\left(f_{T_i}(t_1|\boldsymbol{x}_i)\left\{1-\Phi\left(\frac{\Phi^{-1}\left(F_{D_i}(1|\boldsymbol{Z}_i)\right)-\rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right)\right\}\right) + \tag{45}$$

$$+\sum_{\{i:\varphi_i=1\}} \text{Ln}\left(f_{T_i}(t_1|\boldsymbol{x_i})\left\{\Phi\left(\frac{\Phi^{-1}\left(F_{D_i}(1|\boldsymbol{Z}_i)\right)-\rho\Phi^{-1}\left(F_{T_i}(t_i|\boldsymbol{x}_i)\right)}{\sqrt{1-\rho^2}}\right)\right\}\right) + \sum_{\{i:\varphi_i=\psi_i=0\}} \text{Ln}\left(S_{T_i}(t_i|\boldsymbol{x}_i)\right)$$

3.1.2.- Función de verosimilitud del modelo semiparamétrico

En el caso semiparamétrico $\boldsymbol{\omega}$ está conformado nuevamente por los parámetros de la regresión logística δ_0 y $\boldsymbol{\delta}$, por el parámetro $\boldsymbol{\beta}$ de la componente lineal del modelo de riesgos proporcionales, por el parámetro de dependencia ρ, pero ahora se deben estimar los términos λ_l^* de la parametrización discreta de la función de riesgo (ecuación (31)) donde se tienen tantos términos como tiempos donde ocurren eventos (es decir, sin considerar datos censurados) por lo que $l = 1, \dots, k$ donde $k \leq n$. De esta manera $\boldsymbol{\omega} = (\delta_0, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\lambda}^*, \rho)$, donde $\boldsymbol{\lambda}^* = (\lambda_l^*, \dots, \lambda_k^*)$**.**

La función de distribución del tipo de evento $F_{D_i}(d_i|\boldsymbol{z}_i)$ está dada por la ecuación (37) mientras que la función de distribución del tiempo de supervivencia $F_{T_i}(t_i|\boldsymbol{x}_i)$ se obtiene de la ecuación (31) usando el hecho de que en el caso univariado, la función de supervivencia $S_T(t)$ se puede escribir como $1 - F_T(t)$, por lo que se tiene:

$$F_{T_i}(t_i|\boldsymbol{x}_i) = 1 - \prod_{l:t_l^*<t} (1-\lambda_l^*)^{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}. \tag{46}$$

La determinación de $P(T_i = t_i, D_i = d_i|\boldsymbol{z}_i, \boldsymbol{x}_i)$ como función del vector de parámetros $\boldsymbol{\omega}$, está dada ahora por la ecuación (21), pues ahora ambas variables se consideran discretas:

$$P(T_i = t_i, D_i = d_i|\boldsymbol{z}_i, \boldsymbol{x}_i) =$$

$$\begin{aligned}\Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(d_i|\boldsymbol{z}_i)\right\}; \rho\right) - \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(d_i - 1|\boldsymbol{z}_i)\right\}; \rho\right) - \\ \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(d_i|\boldsymbol{z}_i)\right\}; \rho\right) \\ + \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(d_i - 1|\boldsymbol{z}_i)\right\}; \rho\right)\end{aligned} \tag{47}$$

donde se tiene que el término $\mathrm{C}\left(F_{T_i}(t_i|\boldsymbol{x}_i), F_{D_i}(d_i|\boldsymbol{z}_i)|\rho\right)$ se considera como $\Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(d_i - 1|\boldsymbol{z}_i)\right\}; \rho\right)$ por la ecuación (9), es decir, la cópula es Gaussiana bivariada.

Como ocurrió con el modelo paramétrico, es conveniente mencionar que la expresión anterior se simplifica cuando sólo se tienen dos tipos de eventos, pues cuando $d_i = 2$ se tiene que $F_{D_i}(2|\boldsymbol{z}_i) = 1$ por lo que $C\left(F_{T_i}(t_i|\boldsymbol{x}_i), 1|\rho\right) = F_{T_i}(t_i|\boldsymbol{x}_i)$ y cuando $d_i = 0$ se tiene que $F_{D_i}(0|\boldsymbol{z}_i) = 0$ por lo que $C\left(F_{T_i}(t_i|\boldsymbol{x}_i), 0|\rho\right) = 0$ (al emplear las propiedades $C(u, 1) = u$, $C(1, v) = v$ y $C(u, 0) = C(0, v) = 0$, Nelsen, 2006). De esta forma, la ecuación anterior queda:

$$P(T_i = t_i, D_i = d_i|\boldsymbol{z}_i, \boldsymbol{x}_i) =$$

$$\begin{cases} \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(1|\boldsymbol{z}_i)\right\}; \rho\right) - \\ \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(1|\boldsymbol{z}_i)\right\}; \rho\right) \qquad si\ d_i = 1 \\ F_{T_i}(t_i^+|\boldsymbol{x}_i) - \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(1|\boldsymbol{z}_i)\right\}; \rho\right) - \\ F_{T_i}(t_i|\boldsymbol{x}_i) + \Phi_2\left(\Phi^{-1}\left\{F_{T_i}(t_i|\boldsymbol{x}_i)\right\}, \Phi^{-1}\left\{F_{D_i}(1|\boldsymbol{z}_i)\right\}; \rho\right) \quad si\ d_i = 2. \end{cases} \tag{48}$$

En la ecuación (48) el parámetro de dependencia aparece de forma ρ explícita, mientras que los parámetros δ_0 y $\boldsymbol{\delta}$ están contenidos en $F_{D_i}(1|\boldsymbol{z}_i)$, y los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\lambda}^*$ están dentro de la función $F_{T_i}(t_i|\boldsymbol{x}_i)$.

Finalmente la función de log-verosimilitud $l(\boldsymbol{\omega}; \mathbf{Y}, \mathbf{Z}, \mathbf{X})$ (donde $\boldsymbol{Y}$ es una matriz cuya i-ésima fila está formada por $y_i = \{T_i, D_i, \varphi_i, \psi_i\}$, $\mathbf{Z}$ y $\mathbf{X}$ son matrices de covariables cuyas i-ésimas filas están dadas por $\boldsymbol{z}_i$ y $\boldsymbol{x}_i$ respectivamente) queda como:

$$l(\boldsymbol{\omega}|\mathbf{Y},\mathbf{Z},\mathbf{X}) =$$

$$\begin{aligned}\sum_{\{i:\varphi_i=1\}} \mathrm{Ln}\Big(&\Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big) - \Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big)\Big) \\ &+ \sum_{\{i:\psi_i=1\}} \mathrm{Ln}\Big(F_{T_i}(t_i^+|\boldsymbol{x}_i) - \Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big) - F_{T_i}(t_i|\boldsymbol{x}_i) \\ &+ \Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big)\Big) + \\ &\sum_{\{i:\varphi_i=\psi_i=0\}} \mathrm{Ln}\Big(S_{T_i}(t_i|\boldsymbol{x}_i)\Big).\end{aligned} \tag{49}$$

Para poder realizar los cálculos respectivos, es importante mencionar que el término t_i^+ es mayor a t_i, y corresponden a momentos de ocurrencia de eventos consecutivos.

3.2.- Estimación de los parámetros

La estimación de $\boldsymbol{\omega}$ en ambos modelos se puede lograr al maximizar la función de log-verosimilitud, ya sea al resolver las ecuaciones de verosimilitud $s(\boldsymbol{\omega}) = \mathbf{0}$, donde $s(\boldsymbol{\omega})$ es la función score dada por el vector de derivadas $\partial\, l(\boldsymbol{\omega}|\mathbf{Y},\mathbf{Z},\mathbf{X})/\partial\boldsymbol{\omega}$ (con matriz hessiana $H(\boldsymbol{\omega}) = \partial^2\, l(\boldsymbol{\omega}|\mathbf{Y},\mathbf{Z},\mathbf{X})/\partial\boldsymbol{\omega}\,\partial\boldsymbol{\omega}^T$), o a través de algún otro método de optimización. Se encontró conveniente para este trabajo el uso de la función nlm() del paquete stats en el ambiente estadístico R, la cual implementa un algoritmo de optimización tipo Newton, para el cual las derivadas son aproximadas de forma numérica. Se emplearon como valores iniciales de los parámetros los siguientes:

- Para la variable tipo de evento D, se obtuvieron los parámetros δ_0 y $\boldsymbol{\delta}$ que resultan de ajustar un modelo logístico a las observaciones que no tienen censura, mediante la función glm() con familia binomial.
- En el enfoque completamente paramétrico el tiempo de supervivencia se ajustó a través de modelo paramétrico Weibull de riesgos proporcionales (sin distinguir entre causas de fallecimiento, solamente considerando evento o censura) para la obtención de los parámetros $\boldsymbol{\beta}$, ν y κ. Para lograrlo se empleó la función survreg() del paquete survival.
- En el enfoque semiparamétrico los valores iniciales de λ_l^* se tomaron como el cociente $1/k$.

- Para el parámetro de dependencia se calculó el coeficiente de correlación de Pearson entre T y D como valor inicial de ρ (sin tomar en cuenta la censura).

Dado que no se pone ninguna restricción en la función nlm() al momento de optimizar la función de log-verosimilitud, la estimación se realiza en todo el espacio $\mathbb{R}^l$, donde $l = \dim(\Omega)$ (la dimensión del espacio de parámetros). Como algunos parámetros poseen cierta restricción de los valores que toman, se aplican las siguientes trasformaciones: la liga-log para los parámetros ν y κ, y la transformación de Fisher para el parámetro de dependencia ρ, la cual está dada por:

$$\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \operatorname{arctanh}(\rho)\,. \tag{50}$$

Lo anterior quiere decir que, los estimadores que se obtienen a través del proceso de optimización de la función nlm(), son $\log(\nu)$, $\log(\kappa)$, y $\operatorname{arctanh}(\rho)$, y para regresar a los parámetros originales, es necesario aplicar las transformaciones inversas.

De la teoría estándar de máxima verosimilitud se sabe que bajo ciertas condiciones de regularidad, el estimador $\widehat{\boldsymbol{\omega}}$ es consistente y asintóticamente normal con media $\boldsymbol{\omega}$ y matriz de covarianza dada por la inversa de la matriz de información de Fisher $I(\boldsymbol{\omega}) = \boldsymbol{E}(-H(\boldsymbol{\omega}))$ (Casella y Berger, 2002). Generalmente es más fácil el cómputo de la matriz hessiana, por lo que los errores estándar están dados por las diagonales de la inversa de la matriz de información de Fisher observada $I_O(\widehat{\boldsymbol{\omega}}) = -H(\widehat{\boldsymbol{\omega}})$, ya que asintóticamente serían equivalentes a los errores estándar dados por los elementos de la diagonal de la inversa de la matriz de información de Fisher (Bilder y Loughin, 2014).

En el modelo completamente paramétrico, el tiempo de cálculo es relativamente corto ya que no cuenta con muchos parámetros a estimar pues se tienen solamente dos componentes lineales, una función de supervivencia de base y un parámetro de dependencia. Sin embargo, en el modelo semiparamétrico el número de parámetros a estimar $\boldsymbol{\lambda}^*$ en la marginal del tiempo de supervivencia está en función del número de eventos observados, por lo que el tiempo de cálculo se va incrementando considerablemente al crecer el tamaño de la muestra. Por tal motivo, se propone un método alternativo de estimación el cual consiste en estimar primero la marginal del tiempo de supervivencia a través de un modelo de Cox de riesgos proporcionales para obtener los estimadores de los parámetros de la componente lineal $\widetilde{\boldsymbol{\beta}}$ y aquellos de la

función de riesgo discreta $\tilde{\boldsymbol{\lambda}}^*$, para después introducir estos estimadores en la función de verosimilitud y estimar los parámetros restantes correspondientes al modelo logístico δ_0, $\boldsymbol{\delta}$ y al parámetro de dependencia ρ. En la siguiente sección se explica esta metodología a detalle.

3.2.1.- Estimación en dos etapas para el modelo semiparamétrico

El método de estimación propuesto se puede considerar como un método de pseudo-verosimilitud en dos etapas: primero se encuentran los estimadores de los parámetros correspondientes a la componente marginal temporal $\widetilde{\boldsymbol{\beta}}$ y $\tilde{\boldsymbol{\lambda}}^*$ al ajustar un modelo de Cox de riesgos proporcionales, y posteriormente se insertan estos valores en la función de verosimilitud, obteniendo una función de pseudo-verosimilitud $L_{pse}\left(\delta_0, \boldsymbol{\delta}, \rho | \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \widetilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}^*\right)$ la cual se maximiza para hallar los estimadores $\tilde{\delta}_0$, $\widetilde{\boldsymbol{\delta}}$ y $\tilde{\rho}$.

Este procedimiento de pseudo-verosimilitud es parecido al propuesto por Yilmas y Lawless (2011) para modelos bivariados de datos de supervivencia censurados con marginales semiparamétricas, en el cual en la primera etapa se estiman las distribuciones marginales de los tiempos de supervivencia (empleando el modelo de Cox de riesgos proporcionales) para posteriormente estimar la dependencia. En el método propuesto, solamente se estima en la primera etapa la distribución marginal del tiempo de supervivencia (que era la que presentaba problemas incrementando el tiempo de cálculo en el proceso de verosimilitud completo), evitando estimar marginalmente la variable *D* con datos faltantes (por la censura generada por los riesgos que compiten), y mejor dejando su estimación en la estructura de pseudo-verosimilitud que se obtiene al insertar las estimaciones de la primera etapa en la función de verosimilitud. Es importante mencionar que la información que proporcionan los datos censurados se captura en la primera etapa del proceso de estimación (en la marginal del tiempo de supervivencia), por lo que en la segunda etapa se puede eliminar el término $\sum_{\{i:\varphi_i=\psi_i=0\}} \text{Ln}\left(S_{T_i}(t_i|\boldsymbol{x}_i)\right)$ de la ecuación (49), quedando como:

$$l(\boldsymbol{\omega}|\mathbf{Y},\mathbf{Z},\mathbf{X}) =$$

$$\begin{aligned}\sum_{\{i:\varphi_i=1\}} \mathrm{Ln}\Big(&\Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big)-\Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big)\Big)\\ &+\sum_{\{i:\psi_i=1\}} \mathrm{Ln}\Big(F_{T_i}(t_i^+|\boldsymbol{x}_i)-\Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i^+|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big)-F_{T_i}(t_i|\boldsymbol{x}_i)\\ &+\Phi_2\big(\Phi^{-1}\{F_{T_i}(t_i|\boldsymbol{x}_i)\},\Phi^{-1}\{F_{D_i}(1|\boldsymbol{z}_i)\};\rho\big)\Big)\end{aligned} \tag{51}$$

lo que simplifica los cálculos y el tiempo de cómputo.

Las propiedades asintóticas de los estimadores $\tilde{\boldsymbol{\omega}} = (\tilde{\delta}_0, \tilde{\boldsymbol{\delta}}, \tilde{\rho}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}^*)$ se pueden establecer a partir de la teoría de funciones de inferencia. Bajo las condiciones usuales de regularidad (Casella y Berger, 2002), $\tilde{\boldsymbol{\omega}} = (\tilde{\delta}_0, \tilde{\boldsymbol{\delta}}, \tilde{\rho}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}^*)$ se obtiene como solución del siguiente conjunto de ecuaciones:

$$\boldsymbol{\Psi}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{X}) = \begin{pmatrix} \dfrac{\partial l_{par}(\boldsymbol{\beta}|\mathbf{X})}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial l_{dis}(\boldsymbol{\lambda}^*|\mathbf{X},\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\lambda}^*} \\ \dfrac{\partial l_{pse}(\delta_0,\boldsymbol{\delta},\rho|\mathbf{Y},\mathbf{X},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\lambda}}^*)}{\partial \delta} \\ \dfrac{\partial l_{pse}(\delta_0,\boldsymbol{\delta},\rho|\mathbf{Y},\mathbf{X},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\lambda}}^*)}{\partial \boldsymbol{\delta}} \\ \dfrac{\partial l_{pse}(\delta_0,\boldsymbol{\delta},\rho|\mathbf{Y},\mathbf{X},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\lambda}}^*)}{\partial \rho} \end{pmatrix} = \mathbf{0}. \tag{52}$$

$\boldsymbol{\Psi}$ es conocido como el vector de funciones de inferencia (McLeish y Small, 2012) y se puede descomponer como la suma de funciones de inferencia para cada observación $\boldsymbol{\Psi}(\boldsymbol{\omega}|\mathbf{Y},\mathbf{X}) = \sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{\omega}|\boldsymbol{y}_i,\boldsymbol{x}_i)$ (Song, 2007).

Es necesario hacer los siguientes comentarios sobre el sistema de ecuaciones dado por la relación (52):

- l_{par} es la función de log-verosimilitud parcial del modelo de regresión de Cox de riesgos proporcionales (también llamado modelo continuo de riesgos relativos de Cox), la cual de acuerdo a Collett (2015), está dada por:

$$l_{par}(\boldsymbol{\beta}|\mathbf{X}) = \log\left(\prod_{i=1}^{n}\left(\frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{\sum_{l\epsilon R(t_i)}\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}\right)^{\Delta_i}\right) \tag{53}$$

donde $T_i = \min(\tilde{T}_i, Cens_i)$, Δ_i es el indicador de censura $\Delta_i = I(\tilde{T}_i = T_i)$, $\mathbf{X}$ es la matriz de covariables y $R(t_i)$ representa el conjunto de individuos en riesgo en el tiempo t_i.

- $l_{dis}(\boldsymbol{\lambda}^*|\mathbf{X}, \widetilde{\boldsymbol{\beta}})$ es la función de log-pseudo-verosimilitud que se obtiene al insertar los valores obtenidos de $\widetilde{\boldsymbol{\beta}}$ (al resolver $\partial l_{par}/\partial\boldsymbol{\beta} = 0$) en la función de verosimilitud L_{dis} de la discretización del modelo continuo de riesgos relativos de Cox (Kalbfleisch y Prentice, 2011):

$$L_{dis}(\boldsymbol{\lambda}^*|\mathbf{X}, \widetilde{\boldsymbol{\beta}}) = \prod_{i=1}^{k}\left(\prod_{j\epsilon D(t_i)}\left(1 - \lambda_i^{*\exp(\widetilde{\boldsymbol{\beta}}^T\boldsymbol{x}_i)}\right)\prod_{l\epsilon\, R(t_i)-D(t_i)}\lambda_i^{*\exp(\widetilde{\boldsymbol{\beta}}^T\boldsymbol{x}_i)}\right) \tag{54}$$

donde $D(t_i)$ es el conjunto de individuos que fallecen (o número de eventos que ocurren) en el tiempo t_i.

- Finalmente $l_{pse}(\delta_0, \boldsymbol{\delta}, \rho|\mathbf{Y}, \mathbf{X}, \widetilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}^*)$ representa el logaritmo de la función de pseudo-verosimilitud que se obtiene al sustituir los valores estimados $\widetilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}^*$ en la función de verosimilitud (ecuación (51)).

Siendo $\widetilde{\boldsymbol{\omega}}$ la única raíz de $\Psi(\widetilde{\boldsymbol{\omega}}|\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n}\boldsymbol{\Psi}(\widetilde{\boldsymbol{\omega}}|\boldsymbol{y}_i, \boldsymbol{x}_i) = \mathbf{0}$, en ausencia de observaciones repetidas en el tiempo de supervivencia, se cumple que $E(\boldsymbol{\Psi}(\boldsymbol{\omega}^*|\mathbf{Y}, \mathbf{X})) = \mathbf{0}$ para el valor real del parámetro $\boldsymbol{\omega}^*$, es decir, el conjunto de funciones de inferencia que componen al vector $\boldsymbol{\Psi}$ son insesgadas, y bajo las condiciones regulares de la teoría asintótica de verosimilitud, dejando hasta el primer término de la expansión de Taylor (Joe, 2014), se cumple que:

$$\sqrt{n}(\widetilde{\boldsymbol{\omega}} - \boldsymbol{\omega}^*) \approx \left\{-E\left[\frac{\partial\boldsymbol{\Psi}(\boldsymbol{\omega}^*|\mathbf{Y}, \mathbf{X})}{\partial\boldsymbol{\omega}^T}\right]\right\}^{-1} n^{\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\Psi}(\boldsymbol{\omega}^*|\boldsymbol{y}_i, \boldsymbol{x}_i) + O_p(n^{-\frac{1}{2}}) \tag{55}$$

donde $\partial\boldsymbol{\Psi}/\partial\boldsymbol{\omega}^T$ es una matriz cuyo componente (r, s) está dado por $\partial\Psi_s(\boldsymbol{\omega}^*|\mathbf{Y}, \mathbf{X})/\partial\omega_r$ donde Ψ_s es la componente s de $\boldsymbol{\Psi}$ y ω_r es la componente r de $\boldsymbol{\omega}$.

De acuerdo a Joe (2014), la distribución asintótica de $\sqrt{n}(\widetilde{\boldsymbol{\omega}} - \boldsymbol{\omega}^*)$ es equivalente entonces a aquella de la siguiente expresión:

$$\left\{-E\left[\frac{\partial \boldsymbol{\Psi}(\boldsymbol{\omega}^*|\mathbf{Y},\mathbf{X})}{\partial \boldsymbol{\omega}^T}\right]\right\}^{-1}\boldsymbol{Z}$$

donde $\boldsymbol{Z} \sim N(\mathbf{0}, \text{Cov}[\boldsymbol{\Psi}(\boldsymbol{\omega}^*|\mathbf{Y},\mathbf{X})])$. Esto es, la distribución asintótica de $\sqrt{n}(\widetilde{\boldsymbol{\omega}} - \boldsymbol{\omega}^*)$ es normal multivariada con vector media cero $\mathbf{0}$ y matriz de covarianza asintótica dada por la inversa de la matriz de información de Godambe **V**:

$$\sqrt{n}(\widetilde{\boldsymbol{\omega}} - \boldsymbol{\omega}^*) \approx N(\mathbf{0}, \mathbf{V}) \tag{56}$$

La inversa de la matriz de información de Godambe está dada por:

$$\mathbf{V} = \mathbf{H}^{-1}\mathbf{J}(\mathbf{H}^{-1})^T \tag{57}$$

donde $\mathbf{H} = -E\left[\frac{\partial \boldsymbol{\Psi}(\boldsymbol{\omega}^*|\mathbf{Y},\mathbf{X})}{\partial \boldsymbol{\omega}^T}\right]$ y $\mathbf{J} = E[\boldsymbol{\Psi}(\boldsymbol{\omega}^*|\mathbf{Y},\mathbf{X})\boldsymbol{\Psi}^T(\boldsymbol{\omega}^*|\mathbf{Y},\mathbf{X})]$. Resulta evidente que para poder obtener la inversa de la matriz de información de Godambe es necesario el cálculo de varias derivadas, sobre todo si se tienen varios términos por estimar en el vector $\boldsymbol{\lambda}^*$, lo cual puede ser muy complicado analíticamente.

Una forma alternativa para obtener los errores estándar de los estimadores es el uso de la técnica de remuestreo no paramétrica llamada Bootstrap (Efron y Tibshirani, 1994). La generación de intervalos de confianza de ciertas cantidades relevantes, se puede obtener usando métodos de corrección de sesgo (Carpenter y Bithell, 2000, Efron y Hastie, 2016). Se decide modificar el esquema de remuestreo para contemplar el problema de riesgos competitivos, como lo propone Ng y McLachlan (2003), donde si $n_d(d = 1, \dots, J)$ es el número de fallecimientos de la causa d, y n_0 es el número de observaciones censuradas, los datos de remuestreo se obtienen separadamente de $J + 1$ conjuntos, donde el tamaño de las muestras son $n_d(d = 1, \dots, J)$ para el remuestreo de datos completos y n_0 para el remuestreo de datos con censura.

Es importante mencionar que el desarrollo de la distribución asintótica de $\sqrt{n}(\widetilde{\boldsymbol{\omega}} - \boldsymbol{\omega}^*)$ depende en gran medida del hecho de no tener observaciones repetidas en la variable T. En el caso de existan observaciones repetidas en el tiempo de supervivencia, no es posible usar los resultados asintóticos derivados de las funciones de inferencia, debido a que es necesario ajustar la función de log-verosimilitud parcial (a través de aproximaciones) por lo cual el vector score de la función de verosimilitud parcial ya no resulta ser insesgado (Kalbfleisch y Prentice, 2011), y el conjunto de funciones de inferencia que componen al vector $\boldsymbol{\Psi}$ dejan de ser insesgadas, por lo que no se puede usar el desarrollo anterior para obtener las propiedades

asintóticas (Song, 2007). Sin embargo, la técnica de bootstrap se puede usar para calcular errores estándar e intervalos de confianza, pues es una herramienta no paramétrica para estimar las distribuciones muestrales de los parámetros estimados.

3.2.2.- Criterios de selección

Puede darse la situación que tanto en el enfoque completamente paramétrico como en el semiparamétrico existan varios modelos (variando el número de covariables involucradas en los componentes lineales) que posean un ajuste adecuado y que los parámetros estimados sean significativos (considerando sus respetivos errores estándar). La selección del modelo más parsimonioso se puede lograr a través de Criterio de Información de Bayes (BIC por sus siglas en inglés).

Para el caso paramétrico basado completamente en verosimilitud, el Criterio de Información de Bayes se define como:

$$BIC = -2l(\hat{\boldsymbol{\omega}}) + \dim(\boldsymbol{\omega}) \log(n) \tag{58}$$

donde $l(\hat{\boldsymbol{\omega}})$ es la función de log-verosimilitud, ecuación (45), evaluada en el estimador de máxima verosimilitud $\hat{\boldsymbol{\omega}}$, y $\dim(\boldsymbol{\omega})$ es la dimensión del vector de parámetros. El modelo seleccionado es aquel con el valor menor del BIC. Es posible combinar este criterio con algún algoritmo de selección de variables paso a paso como puede ser selección hacia delante o eliminación hacia atrás.

En el caso semiparamétrico es necesario usar la función de pseudo-verosimilitud $l_{pse}\left(\delta_0, \boldsymbol{\delta}, \rho | \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \widetilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}^*\right)$, por lo que se define un criterio de información de Bayes de pseudo-verosimilitud ($pBIC$) dado por:

$$pBIC = -2l_{pse}\left(\hat{\delta}_0, \widehat{\boldsymbol{\delta}}, \hat{\rho}\right) + \dim(\boldsymbol{\omega}) \log(n). \tag{59}$$

Dado que $l_{pse}(\hat{\delta}_0, \widehat{\boldsymbol{\delta}}, \hat{\rho})$ se puede considerar como una función de log-verosimilitud de perfil[V], desde el punto de vista de teoría de inferencia de modelos semiparamétricos de Murphy y Van Der Vaart (2000), $l_{pse}(\hat{\delta}_0, \widehat{\boldsymbol{\delta}}, \hat{\rho})$ actúa parecido a una función de log-verosimilitud, por lo que $pBIC$ puede ser usado como criterio de selección de la forma usual. En el análisis de datos censurados, la función de verosimilitud de perfil se ha usado para generar estrategias de selección para modelos semiparamétricos de riesgos proporcionales (Xu et al., 2009).

3.2.3.- Adecuación del modelo

Una vez que un modelo estadístico ha sido ajustado a un conjunto de observaciones, es necesario revisar la adecuación del modelo a estos datos, ya que la verificación es una parte fundamental del proceso de modelado estadístico. El procedimiento de diagnóstico que se propone es la gráfica de calibración introducida por Kattan et al. (2003) en el marco de datos de riesgos competitivos con censura, en la cual se comparan los estimadores no paramétricos de la función de incidencia acumulada por tipo de evento contra las funciones de incidencia acumulada modeladas $CIF_d^{mod}(t)$.

El estimador no paramétrico de la función de incidencia acumulada, también conocido como función de incidencia acumulada empírica $CIF_d^{emp}(t)$, se puede estimar a través de la probabilidad de que en el tiempo t_l^* ocurra el evento $D = d$, es decir, $p_d(t_l^*) = P(T = t_l^*, D = d)$. De acuerdo a Putter et al. (2007), una forma de estimar $p_d(t_l^*)$ es a través de la siguiente expresión:

$$\hat{p}_d(t_l^*) = \hat{h}_d(t_l^*) * \hat{S}(t_{l-1}^*) \tag{60}$$

donde $\hat{S}(t_{l-1}^*)$ es el estimador de Kaplan-Meier para la función de supervivencia total evaluado en el tiempo de ocurrencia anterior t_{l-1}^* y $\hat{h}_d(t_l^*)$ es el estimador de la función de riesgo en el

[V] Cuando se habla de procedimientos de pseudo-verosimilitud, se hace referencia a modificaciones a la verosimilitud original, con el objetivo de encontrar funciones verosimilitud para un conjunto de parámetros de interés. Como ejemplo de este tipo de expresiones se tienen las funciones de verosimilitud parcial, marginal o condicional (Garthwaite et al., 2002). En este caso particular, se puede identificar a L_{pse} como función de verosimilitud de perfil.

tiempo t_l^* dado por el cociente del número de eventos d'_d de la causa d y el número individuos en riesgo n'_d de la causa d en el tiempo t_l^*:

$$\hat{h}_d(t_l^*) = \frac{d'_d}{n'_d}. \tag{61}$$

La función de incidencia acumulada empírica $CIF_d^{emp}(t)$, es entonces la suma de estas probabilidades, de todos los puntos donde hay observaciones antes del tiempo t:

$$CIF_d^{emp}(t) = \sum_{l:t_l^* \leq t} \hat{p}_d(t_l^*). \tag{62}$$

En presencia de covariables categóricas, es necesario generar tantas curvas $CIF_d^{emp}(t)$ como combinaciones de los niveles de las covariables, y estas compararlas con las funciones de incidencia modelada. Si se tienen variables explicativas continuas, no es posible calcular la $CIF_d^{emp}(t)$ para estas variables, por lo que se sugiere calcular las funciones de incidencia modelada para cada uno de los valores de variable continua, calcular el promedio y este compararlo con la función de incidencia empírica. El cálculo de las funciones de incidencia empíricas se puede hacer en el ambiente de programación R a usando la función cuminc() del paquete cmprsk. La gráfica de calibración se construye poniendo las $CIF_d^{emp}(t)$ en el eje de las ordenadas y a las $CIF_d^{mod}(t)$ en el eje de las abscisas y si el modelo es adecuado, se obtiene una gráfica de una línea recta con un ángulo de 45° (Kattan et al., 2003).

3.3.- Estudios de simulación

En esta sección se presentan los resultados de simulaciones que sirven para analizar y comparar el desempeño de los métodos propuestos. Como se ha mencionado con anterioridad, dado que los datos que se analizan en el capítulo 4 poseen solo dos riesgos que compiten, se considera generar simulaciones para dos tipos de eventos solamente, con tamaños de muestra de $n = 100$ y $n = 200$.

Se realizan dos experimentos de simulación, uno donde la cópula está completamente especificada, es decir, se simula la cópula Gaussiana y se estima el modelo considerando la misma cópula, y otro donde la cópula no está completamente especificada, pues los datos se

generan a través de otra cópula, pero se ajustan mediante la cópula normal. En el segundo experimento se selecciona la cópula Clayton pues es común encontrarla en los modelos de análisis de supervivencia bivariados (Yilmas y Lawless, 2011). Se escoge que la cantidad de las simulaciones sea de $N = 200$ para los dos escenarios.

En ambos casos, las marginales se especifican a través de un modelo logístico para la variable D y un modelo de riesgos proporcionales de Cox con distribución Weibull de base para la variable T. Para la variable tipo de evento el modelo incluye una variable categórica $\boldsymbol{z}$ que toma dos valores $\{1,2\}$ seleccionados de forma independiente a través de una distribución Bernoulli con probabilidad de éxito de 0.5; los parámetros reales del modelo logístico fueron $(\delta_0, \boldsymbol{\delta}) = (-0.5, 0.5)$. Bajo esta configuración, cuando $\boldsymbol{z} = 1$, se tiene que $P(D = 1|\boldsymbol{z} = 1) = 0.5$, mientras que cuando $\boldsymbol{z} = 2$, se tiene que $P(D = 1|\boldsymbol{z} = 2) = 0.62$. En el caso del tiempo de supervivencia se incluyó una covariable continua $\boldsymbol{x}$ generada independientemente a partir de una distribución $N(0,1)$; el valor de los parámetros fue $\boldsymbol{\beta} = -0.5$ para la componente lineal de la regresión y de $(\nu, \kappa) = (2, 1.5)$ para aquellos de la distribución Weibull de base. Finalmente el tiempo de censura $Cens_i$ se generó de una distribución uniforme $U(0, c)$, donde el valor de c se modifica para considerar dos niveles censura.

La implementación de los algoritmos de optimización se realizó en el lenguaje de programación R, encontrando conveniente el uso de la función nlm() para estimar ambos modelos, así como la función coxph() para estimar el modelo marginal semiparamétrico de riesgos competitivos de Cox. Para los valores iniciales de los parámetros de la variable D, se ajustó un modelo marginal logístico al subconjunto de datos sin censura empleando la función glm(). Los parámetros iniciales de la variable T se ajustaron marginalmente con un modelo de riesgos proporcionales de Cox con distribución de base Weibull usando la función survreg(). Finalmente, para el parámetro de dependencia se usó la correlación entre las variables del tiempo de supervivencia y del tipo de evento como valor inicial.

3.3.1.- Cópula completamente especificada

Para generar datos de la variable D correlacionada con la variable T, se usa el enfoque general propuesto por Czado et al. (2012) en el cual se emplea la función de probabilidad condicional de D dado T, que resulta ser la derivada de la cópula respecto a la variable continua (Krämer

et al., 2013), para generar las simulaciones; en el caso de la cópula Gaussiana la función de probabilidad condicional de D dado T corresponde a la función $C^*(u, v|\rho)$ (ecuación (18)).

Primero se genera una realización del tiempo de supervivencia, y para que siga un modelo paramétrico Weibull de riesgos proporcionales se emplea la metodología propuesta por Bender et al. (2005), en donde se establece que los tiempos se generan a través de la siguiente relación:

$$T' = \left(-\frac{\kappa^{\nu}\log(u)}{\exp(\boldsymbol{\beta}^T\boldsymbol{x})}\right)^{1/\nu} \tag{63}$$

donde u es una realización de una variable aleatoria uniforme en el intervalo [0,1]. Una vez calculado el tiempo t', se procede a calcular la probabilidad de éxito p' de la variable aleatoria D, que de acuerdo a la parametrización de la función de distribución (ecuación (37)) corresponde a $D = 1$, por lo que empleando la función de probabilidad condicional de D dado T se tiene que:

$$p' = P(\,D = 1|T = t') = \Phi\left(\frac{\Phi^{-1}\big(F_D(1|\boldsymbol{z})\big) - \rho\Phi^{-1}\big(F_T(t'|\boldsymbol{x})\big)}{\sqrt{1-\rho^2}}\right). \tag{64}$$

Por último, usando el valor de p' obtenido por la relación anterior, se genera la realización del valor correspondiente a la variable D, usando un distribución Bernoulli(p').

En la Tabla 1, Tabla 2, Tabla 3 y Tabla 4 se muestran los resultados de las simulaciones bajo el escenario de una cópula completamente especificada. El sesgo empírico, las desviaciones estándar empíricas y el error cuadrático medio[VI] de los estimadores se presentan para ambos métodos segmentados por el nivel de censura, el tamaño de muestra y el grado de dependencia. La eficiencia relativa (dada por el cociente del error cuadrático medio del modelo paramétrico entre el error cuadrático medio del modelo semiparamétrico) entre ambos métodos se indica en la última columna. La Tabla 1 y la Tabla 2 corresponden a un nivel moderado de asociación entre variables $\rho = 0.4$, mientras que la Tabla 3 y la Tabla 4 muestran un nivel más elevado de asociación $\rho = 0.7$. Además de los parámetros estimados de ambos modelos, se muestran los valores de la función de supervivencia de base $S_0(t_i)$ para los percentiles 25%, 50% y 75% con fines comparativos.

[VI] Se calcula sumando el cuadrado del sesgo y el cuadrado de la desviación estándar.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 100	U(0,13.5) Censura promedio de 10.45%	$\delta = -0.5$	0.0394	0.6930	0.4818	0.0311	0.6723	0.4530	1.0636
		$\boldsymbol{\delta} = 0.5$	- 0.0201	0.4321	0.1872	- 0.0245	0.4230	0.1795	1.0427
		$\rho = 0.4$	- 0.0093	0.1366	0.0188	- 0.0079	0.1386	0.0193	0.9730
		$\boldsymbol{\beta} = -0.5$	- 0.0061	0.1135	0.0129	- 0.0081	0.1215	0.0148	0.8706
		$\kappa = 1.5$	0.0050	0.0785	0.0062	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0564	0.1711	0.0325	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0001	0.0356	0.0013	0.0059	0.0475	0.0023	0.5527
		$S_0(t) = 0.5$	0.0051	0.0421	0.0018	0.0089	0.0538	0.0030	0.6047
		$S_0(t) = 0.75$	0.0066	0.0358	0.0013	0.0067	0.0444	0.0020	0.6600
n = 100	U(0,6.3) Censura promedio de 21.71%	$\delta = -0.5$	- 0.0241	0.7339	0.5391	- 0.0376	0.7299	0.5342	1.0093
		$\boldsymbol{\delta} = 0.5$	0.0220	0.4579	0.2102	0.0231	0.4539	0.2065	1.0176
		$\rho = 0.4$	0.0258	0.1316	0.0180	0.0236	0.1320	0.0180	1.0001
		$\boldsymbol{\beta} = -0.5$	- 0.0088	0.1102	0.0122	- 0.0045	0.1274	0.0163	0.7513
		$\kappa = 1.5$	- 0.0036	0.0830	0.0069	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0391	0.1715	0.0310	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	0.0008	0.0368	0.0014	0.0015	0.0474	0.0022	0.6042
		$S_0(t) = 0.5$	0.0032	0.0434	0.0019	0.0022	0.0538	0.0029	0.6523
		$S_0(t) = 0.75$	0.0031	0.0384	0.0015	0.0030	0.0456	0.0021	0.7124

Tabla 1. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 100$ y un nivel de correlación moderado $\rho = 0.4$. El número de simulaciones es $N = 200$, las cuales provienen de la cópula Gaussiana.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 200	U(0,13.5) Censura promedio de 10.15%	$\delta = -0.5$	0.0247	0.4476	0.2010	0.0145	0.4420	0.1955	1.0279
		$\boldsymbol{\delta} = 0.5$	- 0.0075	0.2882	0.0831	- 0.0042	0.2833	0.0803	1.0349
		$\rho = 0.4$	- 0.0011	0.0878	0.0077	- 0.0009	0.0881	0.0078	0.9943
		$\boldsymbol{\beta} = -0.5$	- 0.0014	0.0770	0.0059	- 0.0006	0.0816	0.0067	0.8898
		$\kappa = 1.5$	- 0.0040	0.0597	0.0036	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0247	0.1135	0.0135	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0054	0.0277	0.0008	- 0.0001	0.0347	0.0012	0.6609
		$S_0(t) = 0.5$	- 0.0039	0.0316	0.0010	0.0011	0.0376	0.0014	0.7163
		$S_0(t) = 0.75$	- 0.0014	0.0263	0.0007	0.0016	0.0319	0.0010	0.6804
n = 200	U(0,6.3) Censura promedio de 21.81%	$\delta = -0.5$	0.0319	0.4801	0.2315	0.0197	0.4653	0.2169	1.0674
		$\boldsymbol{\delta} = 0.5$	- 0.0053	0.3096	0.0959	- 0.0029	0.3042	0.0925	1.0361
		$\rho = 0.4$	0.0053	0.0910	0.0083	0.0071	0.0913	0.0084	0.9900
		$\boldsymbol{\beta} = -0.5$	- 0.0030	0.0754	0.0057	- 0.0038	0.0825	0.0068	0.8333
		$\kappa = 1.5$	- 0.0078	0.0577	0.0034	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0164	0.1199	0.0146	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0052	0.0272	0.0008	- 0.0023	0.0372	0.0014	0.5526
		$S_0(t) = 0.5$	- 0.0038	0.0288	0.0008	0.0018	0.0378	0.0014	0.5885
		$S_0(t) = 0.75$	- 0.0013	0.0245	0.0006	- 0.0013	0.0316	0.0010	0.5987

Tabla 2. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 200$ y un nivel de correlación moderado $\rho = 0.4$. El número de simulaciones hechas a través de la cópula Gaussiana es de $N = 200$.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 100	U(0,13.5) Censura promedio de 10.33%	$\delta = -0.5$	0.0438	0.6006	0.3626	0.0104	0.5806	0.3372	1.0756
		$\boldsymbol{\delta} = 0.5$	- 0.0267	0.3553	0.1269	- 0.0174	0.3454	0.1196	1.0613
		$\rho = 0.7$	0.0020	0.0834	0.0070	- 0.0002	0.0841	0.0071	0.9833
		$\boldsymbol{\beta} = -0.5$	0.0037	0.0990	0.0098	0.0055	0.1182	0.0140	0.7020
		$\kappa = 1.5$	- 0.0064	0.0839	0.0071	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0256	0.1774	0.0321	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0040	0.0384	0.0015	0.0041	0.0462	0.0022	0.6918
		$S_0(t) = 0.5$	- 0.0027	0.0429	0.0019	0.0001	0.0528	0.0028	0.6652
		$S_0(t) = 0.75$	- 0.0010	0.0367	0.0013	- 0.0020	0.0454	0.0021	0.6529
n = 100	U(0,6.3) Censura promedio de 21.86%	$\delta = -0.5$	- 0.0190	0.6456	0.4172	- 0.0194	0.6288	0.3958	1.0540
		$\boldsymbol{\delta} = 0.5$	0.0194	0.4123	0.1704	0.0099	0.4080	0.1666	1.0229
		$\rho = 0.7$	0.0052	0.0941	0.0089	- 0.0034	0.0919	0.0085	1.0495
		$\boldsymbol{\beta} = -0.5$	- 0.0192	0.1057	0.0115	- 0.0155	0.1269	0.0163	0.7064
		$\kappa = 1.5$	0.0046	0.0931	0.0087	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0407	0.1846	0.0357	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	0.0002	0.0435	0.0019	0.0075	0.0545	0.0030	0.6252
		$S_0(t) = 0.5$	0.0030	0.0465	0.0022	0.0081	0.0573	0.0034	0.6480
		$S_0(t) = 0.75$	0.0038	0.0380	0.0015	0.0046	0.0437	0.0019	0.7570

Tabla 3. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 100$ y un nivel de correlación alto $\rho = 0.7$. El número de simulaciones de la cópula Gaussiana es $N = 200$.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 200	U(0,13.5) Censura promedio de 10.27%	$\delta = -0.5$	0.0304	0.3931	0.1555	0.0109	0.3835	0.1472	1.0561
		$\boldsymbol{\delta} = 0.5$	- 0.0111	0.2488	0.0620	- 0.0055	0.2424	0.0588	1.0557
		$\rho = 0.7$	0.0058	0.0596	0.0036	0.0044	0.0599	0.0036	0.9940
		$\boldsymbol{\beta} = -0.5$	- 0.0042	0.0733	0.0054	- 0.0046	0.0842	0.0071	0.7577
		$\kappa = 1.5$	0.0020	0.0544	0.0030	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0058	0.1160	0.0135	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	0.0008	0.0245	0.0006	0.0015	0.0307	0.0009	0.6352
		$S_0(t) = 0.5$	0.0008	0.0278	0.0008	0.0014	0.0393	0.0015	0.5005
		$S_0(t) = 0.75$	0.0003	0.0244	0.0006	- 0.0001	0.0317	0.0010	0.5938
n = 200	U(0,6.3) Censura promedio de 21.71%	$\delta = -0.5$	- 0.0296	0.4277	0.1838	- 0.0476	0.4079	0.1687	1.0896
		$\boldsymbol{\delta} = 0.5$	0.0262	0.2684	0.0727	0.0312	0.2580	0.0676	1.0769
		$\rho = 0.7$	0.0058	0.0589	0.0035	0.0044	0.0602	0.0036	0.9607
		$\boldsymbol{\beta} = -0.5$	- 0.0093	0.0807	0.0066	- 0.0031	0.0910	0.0083	0.7958
		$\kappa = 1.5$	0.0057	0.0643	0.0042	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0229	0.1227	0.0156	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	0.0015	0.0299	0.0009	0.0042	0.0354	0.0013	0.7032
		$S_0(t) = 0.5$	0.0033	0.0317	0.0010	0.0043	0.0392	0.0016	0.6532
		$S_0(t) = 0.75$	0.0033	0.0259	0.0007	0.0034	0.0305	0.0009	0.7235

Tabla 4. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 200$ y un nivel de correlación alto $\rho = 0.7$. El número de simulaciones usando la cópula Gaussiana es de $N = 200$.

Respecto al sesgo empírico ambos métodos arrojan resultados muy parecidos, resultando ligeramente superior el método semiparamétrico en varios escenarios. Resulta evidente como al aumentar el tamaño de la muestra, se reducen el error cuadrático medio (ECM), al comparar la Tabla 1 contra la Tabla 2 y la Tabla 3 contra la Tabla 4; esto se debe principalmente a la reducción en varianza, más que la reducción en sesgo. Esto es un resultado esperado para el método paramétrico por las propiedades del estimador de máxima verosimilitud, así como también para el método semiparamétrico debido a que la distribución asintótica normal de los estimadores de pseudo-verosimilitud (en ausencia de datos repetidos) implica consistencia en los estimadores.

Un punto a destacar es que el método semiparamétrico arroja estimadores con una eficiencia muy similar (en varios casos es ligeramente superior, y en algunos otros es un poco inferior) que aquellos del método paramétrico para los parámetros del modelo logístico y el parámetro de dependencia, algo que se observa en las cuatro tablas presentadas. Sin embargo, la eficiencia del parámetro de la componente lineal del modelo marginal de riesgos proporcionales de Cox, así como la eficiencia de los percentiles de la función de supervivencia indican superioridad de modelo paramétrico, resultado que es de esperarse pues el modelo real sobre el que se hacen las simulaciones es completamente paramétrico.

3.3.2.- Cópula no especificada

Para analizar el desempeño de los métodos propuestos cuando el modelo de cópula no está completamente especificado, se define que la cópula real sea Clayton, estimando los modelos con la cópula gaussiana. El método para generar las simulaciones es el mismo que se describió en la sección anterior, sólo que ahora cambia la función de probabilidad condicional de D dado T. La cópula Clayton, así como su derivada parcial respecto a la variable continua, están especificadas de la siguiente forma (Krämer et al., 2013):

$$C\left(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta\right) = \left(\left\{F_{Y_1}(y_1)\right\}^{-\theta} + \left\{F_{Y_2}(y_2)\right\}^{-\theta} - 1\right)^{-\frac{1}{\theta}} \tag{65}$$

$$C^*\left(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta\right) = \left(\left\{F_{Y_1}(y_1)\right\}^{-\theta} + \left\{F_{Y_2}(y_2)\right\}^{-\theta} - 1\right)^{-\frac{1}{\theta}-1} \left\{F_{Y_1}(y_1)\right\}^{-\theta-1} \qquad (66)$$

donde $\theta \epsilon (0, \infty)$ es el parámetro de dependencia que caracteriza a la familia de cópulas Clayton. Para establecer una correspondencia entre el parámetro de dependencia de la cópula Gaussiana ρ y el de la cópula Clayton θ, se emplea la medida de concordancia tau de Kendall τ, la cual se asocia con el parámetro de dependencia ρ de la cópula gaussiana y del parámetro de dependencia θ de la cópula Clayton a través de las siguientes relaciones (Nelsen, 2006):

$$\tau = \left(\frac{2}{\pi}\right) \arcsin(\rho) \qquad (67)$$

$$\tau = \frac{\theta}{\theta + 2}. \qquad (68)$$

En la Tabla 5, Tabla 6, Tabla 7 y Tabla 8 se muestran los resultados de las simulaciones de este experimento. El sesgo empírico, las desviaciones estándar empíricas y el error cuadrático medio (ECM) de los parámetros estimados se presentan para ambos métodos, segmentados de nueva cuenta por el nivel de censura, el tamaño de muestra y el grado de asociación. La eficiencia relativa (respecto al ECM) entre ambos métodos se indica en la última columna. La Tabla 5 y la Tabla 6 corresponden a un nivel moderado de asociación entre variables con un valor del parámetro de la cópula Clayton $\theta = 0.71$, el cual equivale a una tau de kendall de $\tau = 0.26$ lo que corresponde a un nivel moderado de correlación de $\rho = 0.4$ para la cópula Gaussiana; la Tabla 7 y la Tabla 8 muestran un nivel mayor de asociación con un valor de $\theta = 1.95$ para la cópula Clayton, lo que se traduce en un valor de $\tau = 0.49$ y $\rho = 0.7$ respectivamente. Además de los parámetros estimados de ambos modelos, se muestran los valores de la función de supervivencia de base $S_0(t_i)$ para los percentiles 25%, 50% y 75%.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 100	U(0,13.5) Censura promedio de 10.44%	$\delta = -0.5$	0.0023	0.6692	0.4478	- 0.0147	0.6653	0.4428	1.0112
		$\boldsymbol{\delta} = 0.5$	0.0205	0.4315	0.1866	0.0220	0.4282	0.1838	1.0155
		$\rho = 0.4$	- 0.0093	0.1174	0.0139	- 0.0060	0.1156	0.0134	1.0352
		$\boldsymbol{\beta} = -0.5$	- 0.0041	0.1131	0.0128	- 0.0081	0.1208	0.0147	0.8728
		$\kappa = 1.5$	- 0.0044	0.0830	0.0069	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0267	0.1564	0.0252	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0020	0.0366	0.0013	0.0035	0.0457	0.0021	0.6390
		$S_0(t) = 0.5$	0.0001	0.0423	0.0018	0.0077	0.0538	0.0030	0.6052
		$S_0(t) = 0.75$	0.0013	0.0368	0.0014	0.0031	0.0426	0.0018	0.7450
n = 100	U(0,6.3) Censura promedio de 22.23%	$\delta = -0.5$	- 0.0823	0.7633	0.5894	- 0.0956	0.7520	0.5747	1.0256
		$\boldsymbol{\delta} = 0.5$	0.0479	0.4982	0.2505	0.0479	0.4890	0.2415	1.0373
		$\rho = 0.4$	0.0212	0.1311	0.0176	0.0212	0.1283	0.0169	1.0428
		$\boldsymbol{\beta} = -0.5$	0.0111	0.1273	0.0163	0.0128	0.1365	0.0188	0.8690
		$\kappa = 1.5$	- 0.0036	0.0861	0.0074	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0296	0.1762	0.0319	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0058	0.0367	0.0014	0.0050	0.0525	0.0028	0.4956
		$S_0(t) = 0.5$	- 0.0028	0.0398	0.0016	- 0.0007	0.0577	0.0033	0.4775
		$S_0(t) = 0.75$	0.0003	0.0344	0.0012	0.0014	0.0415	0.0017	0.6871

Tabla 5. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 100$ y un nivel de correlación moderado $\rho = 0.4$. El número de simulaciones es $N = 200$, las cuales provienen de la cópula Clayton.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 200	U(0,13.5) Censura promedio de 10.06%	$\delta = -0.5$	0.0438	0.4450	0.1999	0.0361	0.4251	0.1820	1.0986
		$\boldsymbol{\delta} = 0.5$	- 0.0234	0.2967	0.0886	- 0.0237	0.2825	0.0804	1.1018
		$\rho = 0.4$	- 0.0146	0.0847	0.0074	- 0.0145	0.0846	0.0074	1.0040
		$\boldsymbol{\beta} = -0.5$	- 0.0008	0.0716	0.0051	- 0.0021	0.0738	0.0054	0.9415
		$\kappa = 1.5$	- 0.0005	0.0562	0.0032	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0191	0.1193	0.0146	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0010	0.0251	0.0006	0.0044	0.0334	0.0011	0.5563
		$S_0(t) = 0.5$	0.0009	0.0300	0.0009	0.0039	0.0371	0.0014	0.6501
		$S_0(t) = 0.75$	0.0018	0.0269	0.0007	0.0026	0.0287	0.0008	0.8786
n = 200	U(0,6.3) Censura promedio de 22.30%	$\delta = -0.5$	0.0820	0.5026	0.2593	0.0689	0.4825	0.2376	1.0915
		$\boldsymbol{\delta} = 0.5$	- 0.0537	0.3325	0.1134	- 0.0499	0.3235	0.1072	1.0587
		$\rho = 0.4$	0.0120	0.0927	0.0087	0.0126	0.0911	0.0085	1.0330
		$\boldsymbol{\beta} = -0.5$	- 0.0098	0.0832	0.0070	- 0.0120	0.0929	0.0088	0.8002
		$\kappa = 1.5$	0.0052	0.0642	0.0042	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0044	0.1279	0.0164	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	0.0059	0.0275	0.0008	0.0052	0.0387	0.0015	0.5170
		$S_0(t) = 0.5$	0.0055	0.0291	0.0009	0.0050	0.0407	0.0017	0.5203
		$S_0(t) = 0.75$	0.0029	0.0245	0.0006	0.0022	0.0322	0.0010	0.5849

Tabla 6. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 200$ y un nivel de correlación moderado $\rho = 0.4$. El número de simulaciones es $N = 200$, las cuales se generan empleando la cópula Clayton.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 100	U(0,13.5) Censura promedio de 9.65%	$\delta = -0.5$	0.1351	0.6147	0.3961	0.1157	0.5920	0.3639	1.0886
		$\boldsymbol{\delta} = 0.5$	- 0.0583	0.4078	0.1697	- 0.0589	0.3935	0.1583	1.0725
		$\rho = 0.7$	- 0.0210	0.0816	0.0071	- 0.0227	0.0793	0.0068	1.0429
		$\boldsymbol{\beta} = -0.5$	- 0.0039	0.1023	0.0105	- 0.0036	0.1156	0.0134	0.7831
		$\kappa = 1.5$	- 0.0091	0.0931	0.0088	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0166	0.1688	0.0288	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0045	0.0415	0.0017	0.0021	0.0516	0.0027	0.6546
		$S_0(t) = 0.5$	- 0.0046	0.0473	0.0023	- 0.0032	0.0568	0.0032	0.6992
		$S_0(t) = 0.75$	- 0.0032	0.0393	0.0016	- 0.0020	0.0484	0.0023	0.6639
n = 100	U(0,6.3) Censura promedio de 21.81%	$\delta = -0.5$	0.1104	0.6110	0.3855	0.0825	0.6066	0.3748	1.0288
		$\boldsymbol{\delta} = 0.5$	- 0.0659	0.3813	0.1497	- 0.0591	0.3815	0.1491	1.0044
		$\rho = 0.7$	0.0064	0.0898	0.0081	0.0019	0.0874	0.0076	1.0615
		$\boldsymbol{\beta} = -0.5$	- 0.0100	0.1151	0.0133	- 0.0122	0.1329	0.0178	0.7489
		$\kappa = 1.5$	- 0.0025	0.0876	0.0077	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0441	0.1993	0.0417	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0029	0.0391	0.0015	0.0064	0.0526	0.0028	0.5480
		$S_0(t) = 0.5$	0.0005	0.0458	0.0021	0.0046	0.0562	0.0032	0.6582
		$S_0(t) = 0.75$	0.0022	0.0409	0.0017	0.0031	0.0501	0.0025	0.6648

Tabla 7. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 100$ y un nivel de correlación alto $\rho = 0.7$. El número de simulaciones es $N = 200$, que se generan empleando la cópula Clayton.

Tamaño de muestra	Censura	Parámetros	Método paramétrico			Método semiparamétrico			Eficiencia relativa
			Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	Sesgo empírico	Desviación estándar empírica	Error Cuadrático Medio	
n = 200	U(0,13.5) Censura promedio de 10.46%	$\delta = -0.5$	0.0429	0.4204	0.1785	0.0282	0.4112	0.1699	1.0507
		$\boldsymbol{\delta} = 0.5$	- 0.0279	0.2634	0.0702	- 0.0258	0.2561	0.0663	1.0585
		$\rho = 0.7$	- 0.0206	0.0646	0.0046	- 0.0217	0.0634	0.0045	1.0235
		$\boldsymbol{\beta} = -0.5$	0.0019	0.0729	0.0053	0.0006	0.0785	0.0062	0.8618
		$\kappa = 1.5$	- 0.0015	0.0631	0.0040	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0003	0.1271	0.0161	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	- 0.0003	0.0277	0.0008	0.0001	0.0351	0.0012	0.6254
		$S_0(t) = 0.5$	- 0.0012	0.0326	0.0011	0.0024	0.0380	0.0015	0.7335
		$S_0(t) = 0.75$	- 0.0017	0.0286	0.0008	- 0.0003	0.0355	0.0013	0.6487
n = 200	U(0,6.3) Censura promedio de 21.62%	$\delta = -0.5$	- 0.0222	0.4155	0.1732	- 0.0334	0.4047	0.1649	1.0499
		$\boldsymbol{\delta} = 0.5$	0.0116	0.2700	0.0730	0.0113	0.2565	0.0659	1.1079
		$\rho = 0.7$	- 0.0189	0.0644	0.0045	- 0.0192	0.0633	0.0044	1.0288
		$\boldsymbol{\beta} = -0.5$	- 0.0030	0.0789	0.0062	- 0.0028	0.0926	0.0086	0.7251
		$\kappa = 1.5$	0.0029	0.0640	0.0041	N.A.	N.A.	N.A.	N.A.
		$\nu = 2$	0.0180	0.1154	0.0136	N.A.	N.A.	N.A.	N.A.
		$S_0(t) = 0.25$	0.0006	0.0285	0.0008	0.0023	0.0362	0.0013	0.6185
		$S_0(t) = 0.5$	0.0018	0.0323	0.0010	0.0044	0.0411	0.0017	0.6129
		$S_0(t) = 0.75$	0.0019	0.0269	0.0007	0.0051	0.0294	0.0009	0.8139

Tabla 8. Se muestra el sesgo empírico, la desviación estándar empírica y el error cuadrático medio para ambos métodos, para un tamaño de muestra de $n = 200$ y un nivel de correlación alto $\rho = 0.7$. El número de simulaciones es $N = 200$, las cuales se generan mediante la cópula Clayton.

Respecto al sesgo empírico ambos métodos arrojan resultados muy parecidos, resultando levemente superior el método paramétrico cuando la asociación es $\rho = 0.4$, y ligeramente superior el método semiparamétrico cuando la correlación es mayor $\rho = 0.7$. Como ocurrió en las simulaciones de las cópulas completamente especificadas, al aumentar el tamaño de la muestra, se reduce el error cuadrático medio (ECM), principalmente por la reducción en la desviación estándar, debido a la distribución asintótica normal de los estimadores del modelo paramétrico y semiparamétrico.

Como ocurrió en el experimento de la sección anterior, la eficiencia de los parámetro estimados δ_0, $\boldsymbol{\delta}$ y ρ en el método semiparamétrico resultó ser muy parecida a la del método paramétrico siendo incluso tenuemente superior en todos los casos; ocurriendo lo contrario para la marginal del tiempo de supervivencia donde la eficiencia de los parámetros estimados del modelo paramétrico es claramente mejor.

Finalmente es importante mencionar que en este experimento la cópula Gaussiana pudo capturar adecuadamente el tipo de dependencia generado por otra cópula (que en este caso específico es la cópula Clayton), por lo que refuerza la idea de que el modelo es robusto ante una ligera desviación de los supuestos distribucionales del vector aleatorio bivariado.

4.- Aplicación a datos reales

En este capítulo se aplican e implementan los modelos propuestos a un estudio prospectivo de pacientes con linfoma folicular, el cual ha sido analizado también por otros autores, como Scheike y Zhang (2011).

4.1.- Descripción del estudio y análisis exploratorio

El comportamiento de los métodos propuestos se ilustrará con un conjunto de datos prospectivos de pacientes con linfoma folicular de etapa temprana (I o II) registrados para su tratamiento con radioterapia o radioterapia y quimioterapia en el hospital Princess Margaret, en Toronto, entre los años 1967 y 1996 (Pintilie, 2007). La base consiste en 541 pacientes con un seguimiento medio de 12.5 años. Uno de los objetivos de este estudio es comparar los patrones de no respuesta al tratamiento o recaída después de los tratamientos, por lo que el fallecimiento del paciente es un riesgo que compite y aquellos con respuesta favorable al tratamiento y que no tienen recaídas son considerados como datos censurados. El conjunto de datos se puede descargar de la siguiente liga:

https://raw.githubusercontent.com/scheike/update-code-for-jss comp.risk/master/follic.txt

En la siguiente tabla se muestra la descripción de las principales variables involucradas en este estudio.

Variables clínicas medidas al ingreso de los pacientes al estudio	
age	Edad en años
hgb	Hemoglobina en g/l (gramos por litro)
clinstg	Estado clínico del cáncer: 1 = estado I, 2 = estado II
ch	Quimioterapia: Y= Si, espacio en blanco = No
rt	Radioterapia: Y= Si, espacio en blanco = No
Variables de respuesta	
resp	Respuesta después del tratamiento: CR = Respuesta completa, NR = No respuesta
relsite	Sitio de recaída: L = Local, D =Distante, B = Local y distante, espacio en blanco = No recaída
survtime	Tiempo en años del diagnóstico al fallecimiento o al último seguimiento
stat	Estado: 1 = Fallecido, 0 = Vivo
dftime	Tiempo en años del diagnóstico al primer evento (no respuesta, recaída o fallecimiento) o al último seguimiento
dfcens	Variable de censura: 1 = Evento, 0 = Censura

Tabla 9. Descripción de las variables de la base follic.txt, del estudio prospectivo de pacientes con linfoma folicular.

El evento de interés (que es la no respuesta o recaída después de los tratamientos) se construye de los datos cuando la variable resp es igual a NR o cuando la variable relsite es igual a L, D o B. El evento que compite (que es el fallecimiento del paciente) se construye considerando que la variable resp sea igual a CR, que la variable relsite sea igual a espacio en blanco y que stat sea igual a 1 (fallecido). Esta forma de caracterizar los eventos se retoma del trabajo de Scheike y Zhang (2011).

Antes de aplicar la metodología propuesta es necesario hacer un análisis exploratorio para entender mejor cuales son las características y el comportamiento general de las variables involucradas. Se tiene como primer punto que la base de 541 registros tiene datos completos, por lo que no hay pérdida de información. Respecto a los eventos de interés se tiene que 272 no tuvieron respuesta o experimentaron una recaída (50.3%), 76 fallecieron (14.0%) y 193 fueron datos con cesura (35.7%).

De las 5 variables clínicas, se tienen 2 variables continuas, age y hgb, y tres categóricas, clinstg, ch y rt. Como se dijo anteriormente, todos los pacientes que se analizan en esta base recibieron el tratamiento de radioterapia por lo que la variable rt tiene todos los registros con el valor Y, y por tal motivo esta variable no se considera para el análisis. Para poder darle un tratamiento numérico a las variables relacionadas con el estado del cáncer y con la quimioterapia, se crean dos variables numéricas: stage cuyo valor 0 indica el estado I y valor 1 indica estado II y chemo cuyo valor 0 indica ausencia de quimioterapia y valor 1 indica presencia del tratamiento.

Para darse una idea del comportamiento marginal se pueden elaborar gráficas de cómo se distribuyen los tiempos de supervivencia y los tipos de eventos respecto a las variables predictoras. Con fines de claridad, resulta conveniente segmentar las variables continuas en 3 niveles: para age, menores a 50 años, entre 50 y 60 años y mayores a 60 años, y para hgb, menores a 130 g/l, entre 130 y 140 g/l, y mayores a 140 g/l.

Respecto al tiempo de supervivencia, en la Figura 2 se ilustran gráficas de caja que muestran la relación entre esta variable y cada predictor. Como se estableció con anterioridad, la marginal de tiempo considera la ocurrencia de cualquier evento y se etiqueta con valor de 1 (mientras que la censura se etiqueta con 0); es importante mencionar que la segmentación por el tipo de evento se hace de forma general en la estructura de dependencia de la cópula.

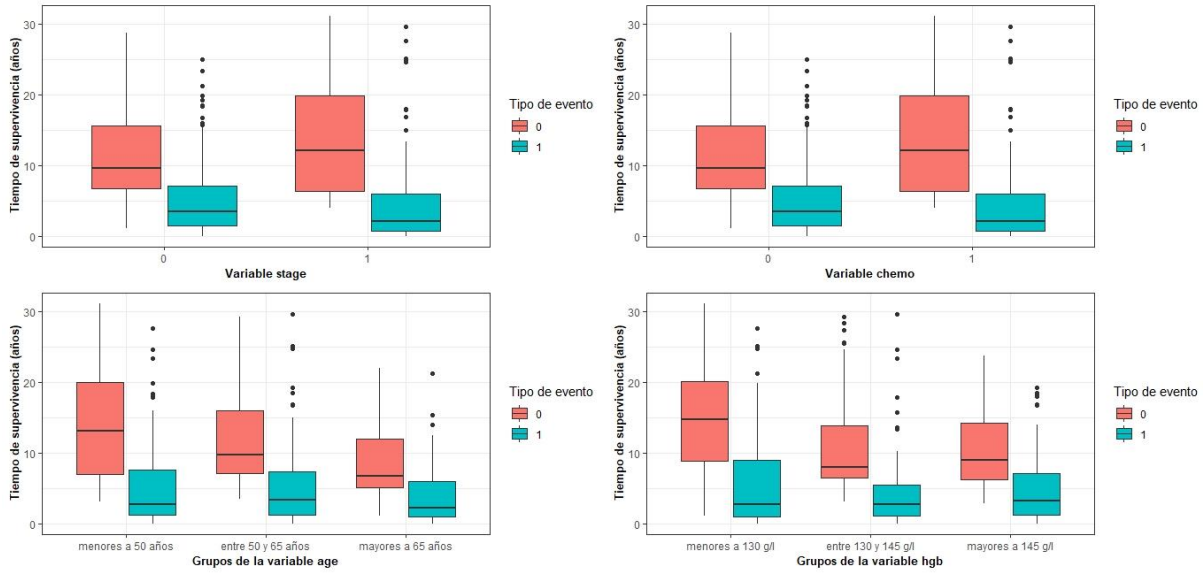

Figura 2. Gráficas de caja para observar la distribución de la variable tiempo de supervivencia respecto a cada covariable. El tipo de evento 1 representa la ocurrencia de cualquier evento (no respuesta al tratamiento, recaída después del tratamiento o fallecimiento), mientras que el tipo de evento 0 indica datos censurados.

Se observa que el tiempo de supervivencia disminuye cuando el grado del cáncer es más avanzado (stage igual a 1) y decrece ligeramente cuando se usa quimioterapia; cuando aumenta la edad se percibe una ligera disminución de los tiempos de supervivencia. Al parecer, la variable hgb no influye en un aumento o disminución de los tiempos de supervivencia.

Analizando la variable D, la Figura 3 presenta 4 gráficas de barras, que contabilizan la cantidad de eventos que ocurren para los niveles de las variables explicativas (omitiendo los registros censurados, pues su presencia imposibilita la observación de los dos eventos de interés).

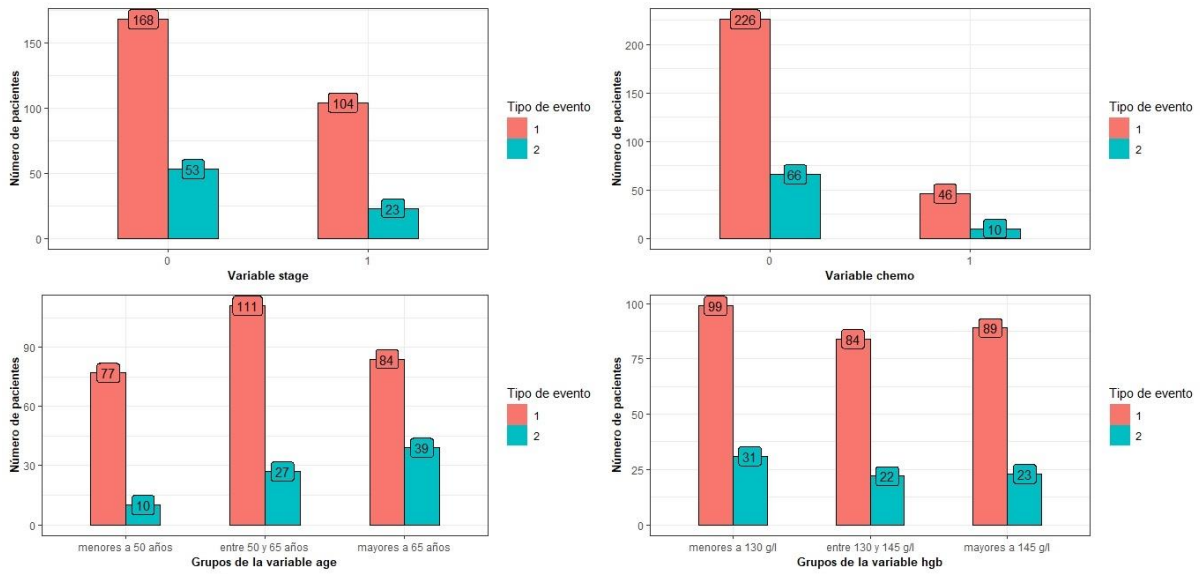

Figura 3. Gráficas de barras que ilustran el conteo de pacientes segmentados por tipo de evento y por los niveles de las 4 covariables. El tipo de evento 1 representa la no respuesta al tratamiento o recaída después del tratamiento, mientras que el tipo de evento 2 indica el fallecimiento.

Si bien la cantidad de eventos (ya sea del tipo 1, que indican no respuesta o recaída, o del tipo 2 que indican fallecimiento del paciente) es mayor para un estado de la enfermedad menos avanzado (stage igual a 0), la proporción entre eventos se mantiene relativamente cercana para ambos estados de la enfermedad. Algo parecido ocurre con la variable chemo, pues si bien hay más eventos ante la ausencia de quimioterapia, la proporción de eventos vuelve a ser cercana. Esto podría ser un indicador de que esas variables no son informativas. Respecto a la variable age, si se observa que, a mayor grupo de edad, van disminuyendo la proporción de pacientes con evento tipo 1 y aumentado la proporción de pacientes con evento tipo 2. De nueva cuenta, como ocurrió en la marginal T, la hemoglobina no se muestra informativa, pues las proporciones se mantienen cercanas entre los diversos grupos.

Es importante también darse una idea de cómo es en general la función de densidad de las variables predictoras continuas (calculada a través de métodos no paramétricos con función núcleo o kernel gaussiano), así como la posible dependencia lineal que puede existir entre ellas (a través del coeficiente de correlación de Pearson). En la Figura 4, se muestran un conjunto de gráficas donde se ilustran estos conceptos.

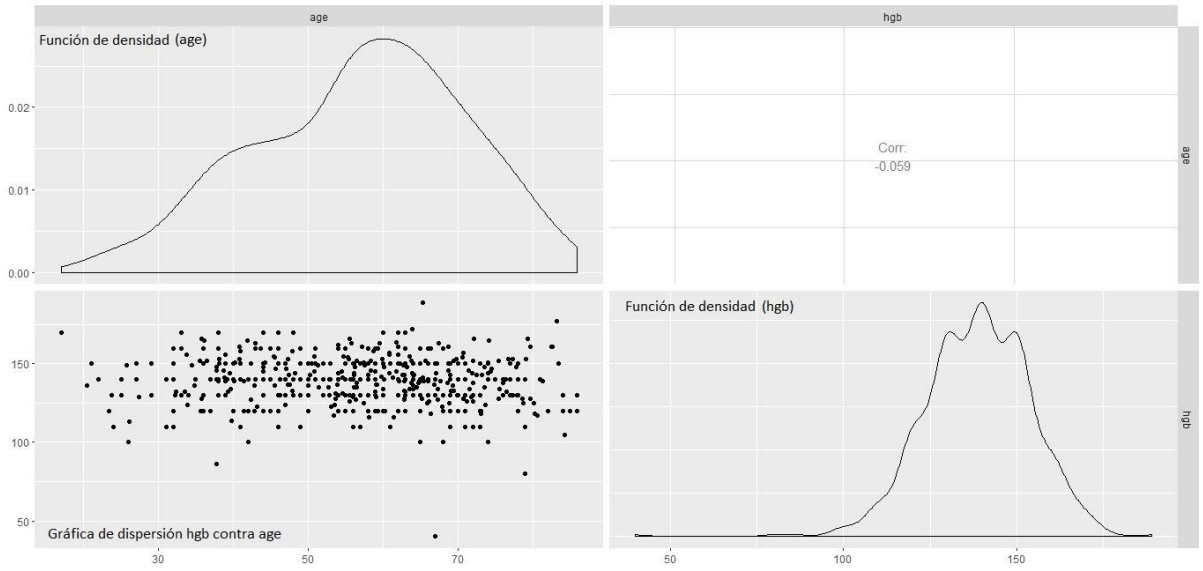

Figura 4. Gráficas de densidad de cada covariable (calculadas a través de métodos no paramétricos), gráfica de dispersión y coeficiente de correlación de Pearson.

La relación lineal entre la variable age y hgb no es significativa, pues al observar la gráfica de dispersión no se ve ningún patrón de comportamiento, algo que es corroborado por la estimación puntual del coeficiente de correlación de Pearson que es cercana a cero. Esto es una buena noticia, pues quiere decir que no se tendrán problemas de colinealidad entre estas dos variables.

4.2.- Análisis de riesgos proporcionales

Al formular el modelo de Cox es necesario probar que se cumplan las suposiciones sobre el uso de la distribución Weibull y sobre la proporcionalidad de los riesgos. Como se mencionó en el capítulo 2, estos supuestos se pueden verificar graficando $\log\left\{-\log\left(\hat{S}(t)\right)\right\}$ contra $\log(t)$ (donde $\hat{S}(t)$ es el estimador de Kaplan-Meier) para observar si se presentan gráficas de líneas paralelas.

La aplicación del procedimiento gráfico a las variables stage y chemo se puede hacer de forma directa pues sólo poseen dos niveles. Sin embargo, para aplicar el procedimiento a las variables continuas age y hgb es necesario categorizar el dominio de las variables en grupos que permitan estimar la función de supervivencia usando el estimador de Kaplan-Meier; se opta por

segmentar las variables en dos grupos: para age, edades menores o iguales a 60 años y edades mayores a 60, y para hgb, menor o igual a 140 g/l y mayor a 140g/l. En la Figura 5, se muestran las 4 gráficas de $\log\left\{-\log\left(\hat{S}(t)\right)\right\}$ contra $\log(t)$.

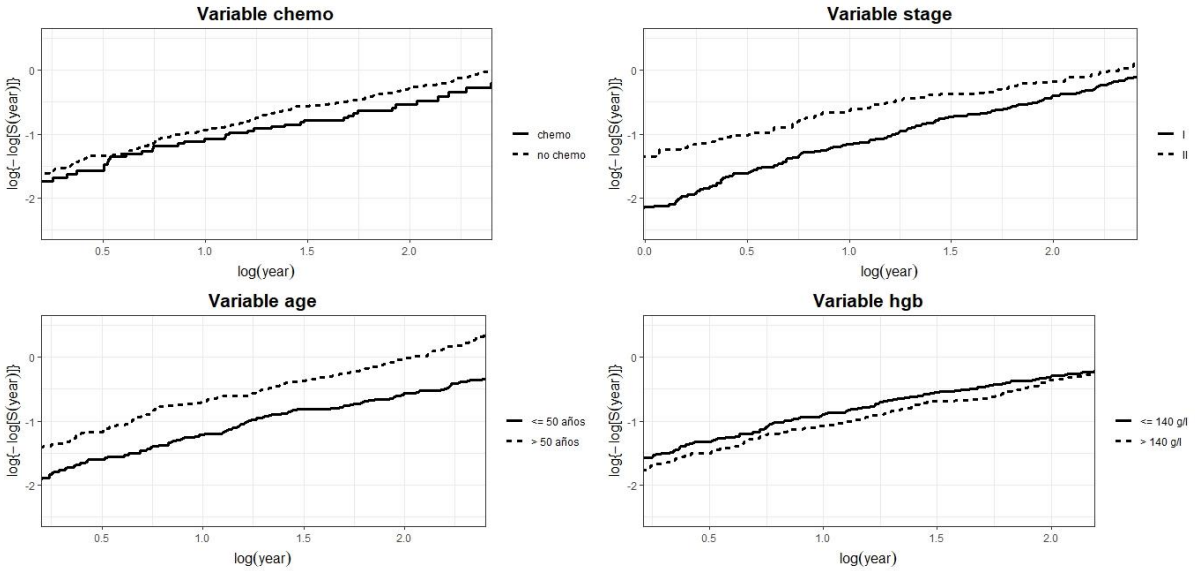

Figura 5. Gráfica de $\log\left\{-\log\left(\hat{S}(t)\right)\right\}$ contra $\log(t)$ para las variables stage, chemo, age y hgb.

En términos generales, se tienen aproximadamente dos líneas rectas para 3 de las 4 gráficas, siendo más marcado el comportamiento esperado para la variable age; si bien en algunos puntos de las gráficas de las variables chemo y hgb se perciben que las gráficas pueden tocarse, en general el comportamiento es paralelo y lineal en la mayor parte del dominio. Respecto a la variable stage, se observa que a medida que aumenta $\log(t)$ la rectas se van aproximando, por lo que el supuesto de proporcionalidad no se cumple para esta variable.

Como se mencionó en sección de la marginal del tiempo de supervivencia, un método para contemplar variables que no cumplen la suposición de riesgos proporcionales, es la incorporación de variables que dependan del tiempo, lo cual se puede hacer multiplicando aquellas variables que no cumplen la suposición por alguna función del tiempo, siendo los términos más empleados t, t^2 y $\log(t)$.

Para incorporar variables dependientes del tiempo en el modelo completamente paramétrico es necesario contemplar que la función de densidad marginal de la variable Tse debe modificar. La función $f_T(t;\boldsymbol{\theta}_T)$, dada por la ecuación (26), se puede obtener derivando respecto al tiempo la ecuación (24). En ese caso, la componente lineal $\boldsymbol{\beta}^T\boldsymbol{x}$ no posee términos que dependan del

tiempo, por lo que su tratamiento al momento de derivar no resulta complicado. Sin embargo, ahora la componente lineal posee términos que dependen del tiempo $\boldsymbol{\beta}^T\boldsymbol{x} + \boldsymbol{\beta'}^T\boldsymbol{x'}(t)$, y es necesario contemplarlos al momento de derivar y aplicar la regla de la cadena para obtener la expresión de $f_T(t;\boldsymbol{\theta}_T)$.

Para el caso particular de este análisis, se escoge la función $\log(t)$, ya que proporcionó un mejor ajuste en los datos. De esta forma, la componente lineal para el modelo marginal del tiempo de supervivencia queda como $\boldsymbol{\beta}^T\boldsymbol{x} + \beta'_{stage}\text{stage} * \log(t)$. Al derivar la ecuación (24) respecto al tiempo, se obtiene la siguiente función de densidad:

$$f_T(t;\boldsymbol{x},\boldsymbol{\theta}_T) =$$

$$\frac{(\beta'_{stage}\text{stage} + \nu)}{\kappa^\nu} t^{\nu-1} \exp\left(\boldsymbol{\beta}^T\boldsymbol{x} + \beta'_{stage}\text{stage} * \log(t).\right)\left\{\exp\left(-\left(\frac{t}{\kappa}\right)^\nu\right)\right\}^{\exp(\boldsymbol{\beta}^T\boldsymbol{x}+\beta'_{stage}\text{stage}*\log(t))}. \quad (69)$$

Para la estimación de máxima verosimilitud se sustituye el valor de la relación anterior en la ecuación (45), y el proceso continúa de la misma forma.

Respecto al modelo semiparamétrico, es necesario modificar la primera etapa del proceso de estimación. Usando la codificación $(\text{start}, \text{stop}]$ para los tiempos de supervivencia descrita en Therneau et al. (2018), es posible modelar variables dependientes del tiempo usando la función coxph() del paquete survival. La codificación $(\text{start}, \text{stop}]$ tiene que ver con el hecho de que como existe una variable que depende del tiempo, que en este caso es $\text{stage} * \log(t)$, es necesario calcular los valores de esta variable para cada individuo, a lo largo del tiempo hasta la ocurrencia del evento para ese paciente. Por ejemplo, si para un paciente su tiempo de supervivencia fue de 1 año, y en ese lapso de tiempo se presentaron 3 eventos de otros pacientes $(0 \leq t_1 \leq t_2 \leq t_3 \leq 1)$ (incluyendo aquellos que pudieron presentar censura), entonces la codificación $(\text{start}, \text{stop}]$ de este individuo está formada por 4 intervalos: $(0, t_1]$, $(t_1, t_2]$, $(t_2, t_3]$, y $(t_3, 1]$. El valor de $\text{stage} * \log(t)$ para este paciente se debe calcular para cada intervalo (considerando el tiempo del extremo derecho).

4.3.- Proceso de modelado y análisis de resultados

Una vez definido que en la componente marginal del tiempo de supervivencia se usará la variable $\text{stage} * \log(t)$ que permite describir la dependencia de temporal de la variable stage,

se proceden a ajustar los modelos; considerando el modelo completo con las variables chemo, stage, stage $*$ log(t), age y hgb para la variable T, y las variables chemo, stage, age y hgb para la variable D se implementó un procedimiento de selección de variables de eliminación hacia atrás basado en BIC y $pBIC$. En cada etapa se quitaba la variable menos significativa, y se volvía a estimar el modelo; el proceso siguió hasta encontrar el modelo más parsimonioso bajo la condición de que todos sus parámetros fueran significativos. En la Tabla 10, se muestran los valores de BIC y $pBIC$ para los distintos modelos de regresión para las marginales del tiempo de supervivencia y del tipo de evento.

Componente de supervivencia	Componente logística	BIC	pBIC
chemo + age + hgb + stage + stage*log(t)	chemo + stage + age + hgb	2,536.07	4,377.50
chemo + age + stage + stage*log(t)	chemo + stage + age + hgb	2,528.90	4,371.98
chemo + age + stage + stage*log(t)	chemo + stage + age	2,523.16	4,365.89
chemo + age + stage + stage*log(t)	stage + age	2,517.01	4,361.69
chemo + age + stage + stage*log(t)	stage	2,510.14	4,356.01
age + stage + stage*log(t)	stage	2,505.71	4,344.50
age + stage + stage*log(t)	intercepto	2,505.11	4,339.47
stage + stage*log(t)	intercepto	2,584.49	4,398.92

Tabla 10. Proceso de selección del modelo usando los criterios de información de Bayes para diversas combinaciones de modelos anidados aplicados al estudio prospectivo de pacientes con linfoma folicular. El modelo seleccionado está enmarcado en recuadro negro.

Aunque la diferencia en los criterios de información de Bayes fue más marcada para los diferentes modelos anidados del enfoque semiparamétrico, ambos planteamientos llegan al mismo modelo de regresión para los dos componentes marginales. Se observa que las variables age, stage y stage $*$ log(t) resultan significativas para la componente de supervivencia, mientras ninguna resulta ser importante en la componente logística. La composición del modelo final para cada marginal, está en cierta sintonía con los hallazgos encontrados en el análisis exploratorio, donde resultaba evidente que ciertas variables (al ser analizadas de forma independiente respecto a cada marginal), parecían no tener efectos significativos.

Contrario a los resultados que se encontraron en Scheike y Zhang (2011), que indican que la variable chemo es marginalmente significativa, ambos modelos sugieren que en presencia de las variables age, stage y stage $*$ log(t) en la componente de supervivencia, el uso de quimioterapia no muestra efectos significativos respecto la supervivencia de los individuos ni la ocurrencia de los eventos que compiten.

La Tabla 11 muestra los coeficientes estimados del modelo más parsimonioso para los enfoques paramétrico y semiparamétrico de la cópula Gaussiana. Se observa que ambas formulaciones poseen estimadores y errores estándar similares. Los coeficientes de la variable $\text{stage} * \log(t)$ muestran efectos moderados (valor-p < 0.05), mientras que los demás coeficientes muestran efectos muy significativos (valor-p < 0.001). Los estimadores paramétrico y semiparamétrico de $\widehat{\text{arctanh}}(\rho)$ fueron respectivamente 0.9401 (error estándar de 0.1090) y 0.9383 (con error estándar de 0.1256). Resulta más conveniente el uso de una medida de dependencia concordante como la tau de Kendall τ, ya que permite comparaciones entre diversas cópulas. En ese sentido, usando la ecuación (67), se pueden obtener intervalos de confianza al 95% para τ, los cuales son (0.4770, 0.5710) para el modelo paramétrico basado en normalidad y de (0.4427, 0.6235) para el modelo semiparamétrico basados en la técnica de Bootstrap con corrección de sesgo. Los valores de estas medidas sugieren que los tiempos de supervivencia de los pacientes para el riesgo 1 (no respuesta o recaída después de los tratamientos), tienden a ser menores que aquellos pacientes para el riesgo 2 (fallecimiento del paciente).

Considerando la variable tipo de evento, se puede calcular un intervalo de confianza para la tasa de supervivencia de largo plazo $P(D = 1)$, es decir, la proporción de individuos que sobreviven cuando el tiempo de supervivencia es muy grande, tan grande como para que todos los eventos sean observados. El intervalo de confianza al 95% para $P(D = 1)$ es (0.5763, 0.6931) para el modelo paramétrico y de (0.5814, 0.6665) para el modelo semiparamétrico. Esto implica que, a largo plazo, la proporción de sujetos cuya respuesta no es favorable o tiene una recaída es más de la mitad, alejándose un poco del (50.3%) que indican solamente los datos observados (sin considerar censura). Finalmente, los estimadores puntuales de los parámetros de la función de Weibull de base fueron 5.0999 (error estándar de 0.3357) para $\widehat{\log(\kappa)}$ y -0.2573 (error estándar de 0.0598) para $\widehat{\log(\nu)}$.

	Modelo paramétrico				**Modelo semiparamétrico**			
	Supervivencia		**Logístico**		**Supervivencia**		**Logístico**	
Covariable	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercepto	---	---	0.5612	0.1293	---	---	0.4331	0.1257
stage	0.7582	0.1597	---	---	0.5752	0.1372	---	---
stage*log(t)	-0.1945	0.0605	---	---	-0.1883	0.0946	---	---
age	0.0338	0.0038	---	---	0.0340	0.0043	---	---

Tabla 11. Coeficientes (Coef) y errores estándar (SE, por sus siglas en inglés) de los modelos seleccionados de ambos planteamientos para el estudio prospectivo de pacientes con linfoma folicular.

Para analizar la adecuación del modelo, se emplean las gráficas de calibración que se introdujeron en el capítulo 3. El cálculo de las funciones de incidencia empíricas se puede hacer en lenguaje de programación R usando la función cuminc() contenida en el paquete cmprsk. Segmentando las curvas $CIF_d^{emp}(t)$ por la variable stage y el tipo de evento, se calculan las $CIF_d^{mod}(t)$ en cada una de las edades del subconjunto formado por stage y tipo de evento, generando un promedio que es el que se compara con $CIF_d^{emp}(t)$. Por ejemplo, para el evento 1 y stage igual a 0 (que equivale a Estado I de la enfermedad), se tienen 168 registros, con edades que van desde los 17 años hasta los 86 años. Se calculan las 168 funciones de incidencia acumulada y su promedio es el que se compara contra la función de incidencia acumulada empírica segmentada por evento 1 y stage igual a 0. En la Figura 6 (a)-(d) se muestran las gráficas de las funciones de incidencia acumulada promedio de ambos enfoques y la función de incidencia acumulada empírica segmentadas por estado de la enfermedad y tipo de evento. Puede observarse que, mientras que las estimaciones de las funciones de incidencia acumulada promedio de ambos modelos para el estado I corresponden bastante bien a las estimaciones $CIF_d^{emp}(t)$ para ambos eventos, las estimaciones de $CIF_d^{mod}(t)$ para el grupo de la etapa II muestran ciertas desviaciones de las estimaciones empíricas en ambas causas, siendo estas discrepancias más pronunciadas para el modelo semiparamétrico del evento 2.

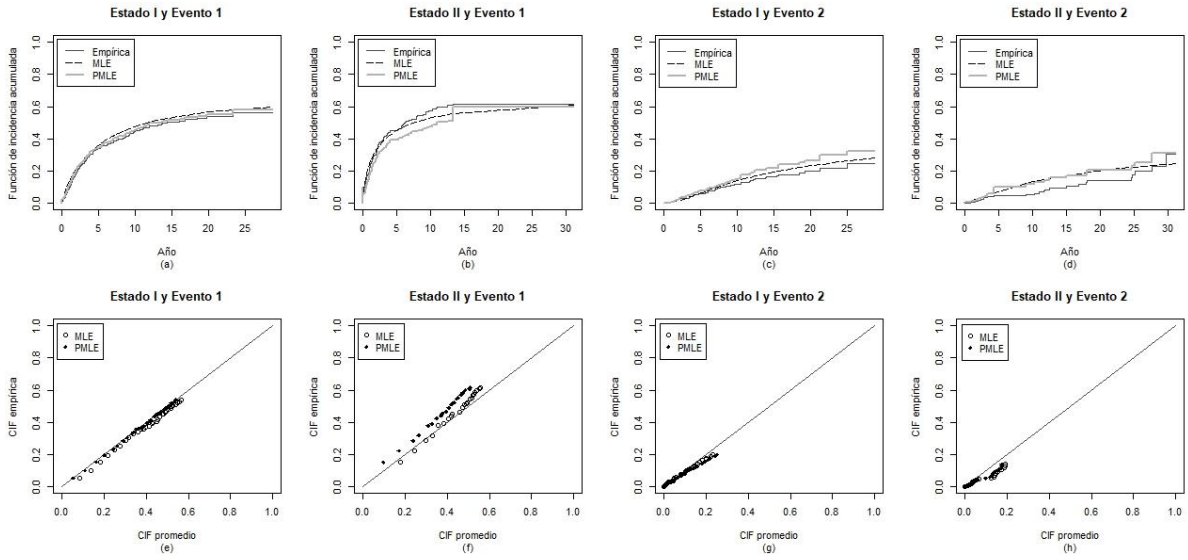

Figura 6. (a)-(d) Funciones de incidencia acumulada promedio del modelo paramétrico (MLE) y semiparamétrico (PMLE) junto con las funciones de incidencia empíricas segmentadas por estado de la enfermedad y tipo de evento. (e)-(g) Gráficas de calibración para ambos enfoques segmentadas por estado de la enfermedad y tipo de evento.

En la Figura 6 (e)-(g) se observan las gráficas de calibración para la misma segmentación, observándose gráficas cercanas a 45° para el estado I, mientras que ligeras desviaciones para el estado II. En general, las gráficas de la Figura 6 en conjunto indican que los modelos proporcionan un ajuste razonablemente bueno a los datos del linfoma de células foliculares, y confirman que la discriminación por estado de la enfermedad representa un criterio de clasificación útil para separar a los individuos cuyas respuestas son intrínsecamente diferentes.

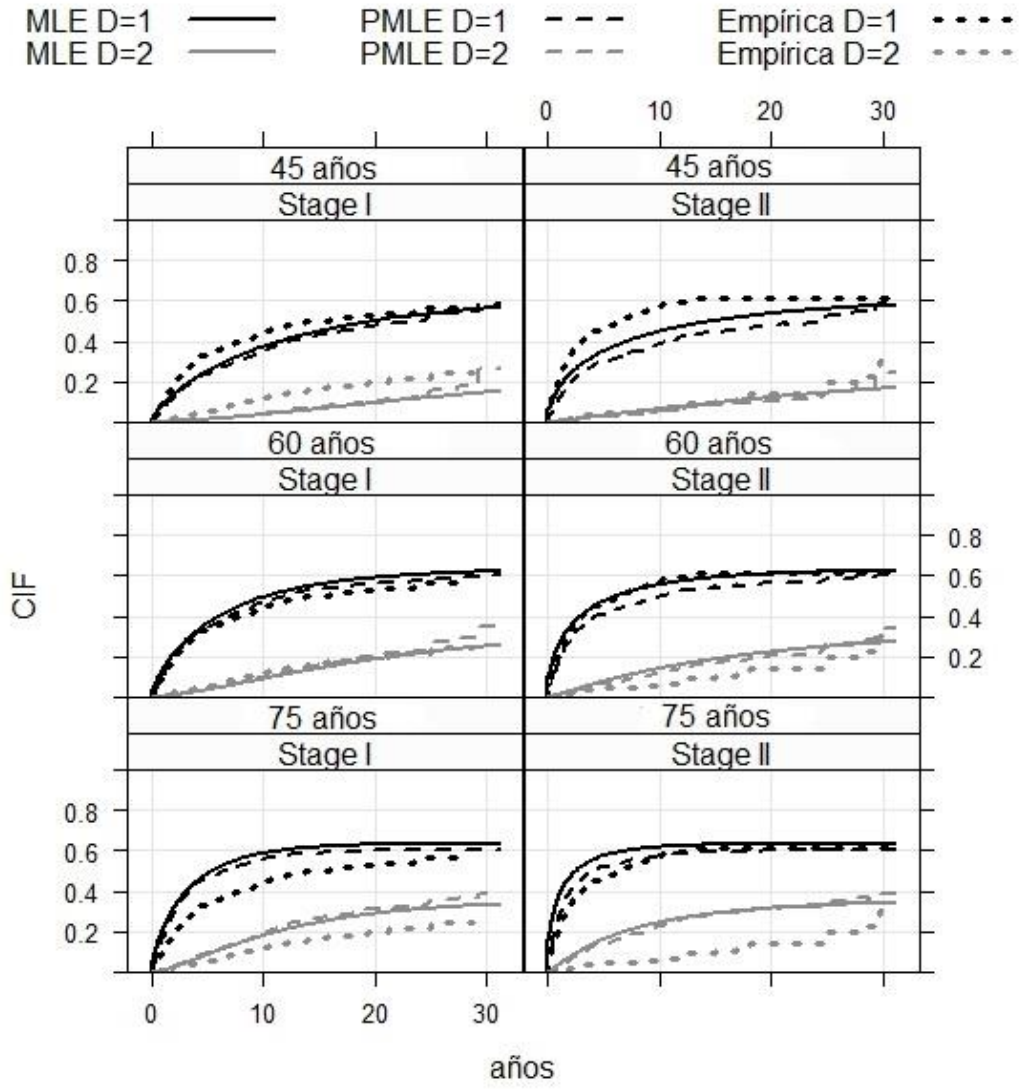

Figura 7. Predicciones de las funciones de incidencia acumulada ($CIF_d(t)$) basadas en modelo con mejor ajuste paramétrico (MLE) y semiparamétrico (PMLE) para los dos niveles de la variable stage y tres valores de la variable age dados por 45 años, 60 años y 75 años. Se muestra también la gráfica de la función de incidencia acumulada empírica clasificada solamente por la variable stage.

La Figura 7 muestra las predicciones de la curva de incidencia acumulada para el modelo seleccionado para ambos enfoques. La segmentación de las curvas se hace por los dos niveles de la variable stage, y en tres edades de la variable age: 45 años, 60 años y 75 años. Cada panel muestra además la función de incidencia acumulada empírica clasificada solamente por la variable stage, la cual se calculó usando la función cuminc() del paquete cmprsk. De forma similar a los hallazgos encontrados en Scheike y Zhang (2011), la función de incidencia acumulada muestra valores más grandes en los momentos iniciales del estudio para estados de la enfermedad mayores y edades más avanzadas, convergiendo aparentemente a puntos muy cercanos. El comportamiento de las $CIF_d(t)$'s para ambos enfoques es muy similar. Como es de esperarse para ambos riesgos, la función de incidencia acumulada empírica está por encima de aquellas generadas por los modelos para pacientes con edades de 45 años, y por abajo de aquellas $CIF_d(t)$'s generadas para pacientes con edad de 70 años.

5.- Conclusiones y perspectivas

En este trabajo se proponen dos métodos novedosos para el estudio de riesgos competitivos que incluyen el análisis de variables explicativas. En ambos se propone el uso de la cópula Gaussiana para modelar la distribución conjunta $P(T, D)$, lo que permite un tratamiento simétrico de ambas variables acoplándolas a través su dependencia ρ.

El modelo de regresión de la variable D corresponde a un modelo multinomial (logístico en caso de que $D = \{1,2\}$), lo que permite estimar las probabilidades de ocurrencia de los eventos que compiten en el caso en que el tiempo se va infinito, lo que sería para fines prácticos considerar el caso en el que conociera el tiempo de ocurrencia de los eventos para todos los individuos del estudio (bondad que también comparte el modelo mixto de Larson y Dinse,1985). Esta característica es de suma importancia y relevancia para el manejo clínico del paciente (Zhang y Fine, 2008).

Respecto a la componente marginal de tiempo se tienen dos enfoques: el completamente paramétrico que usa el modelo paramétrico de Cox de riesgos proporcionales con función de base Weibull, y el semiparamétrico que se ajusta a través de modelo de Cox de riesgos proporcionales libre de distribución. En ambos casos, los parámetros de regresión hacen referencia a la ocurrencia de cualquier evento, por lo que a diferencia del modelo mixto de Larson y Dinse (1985), no se tiene una interpretación condicional de los parámetros de los tiempos de supervivencia dados los tipos de evento. Lo anterior se traduce en un modelo más parsimonioso, pues sólo se estima una componente lineal para el tiempo de supervivencia, y no tantas como tipos de eventos, que en el caso de modelos mixtos con marginales no paramétricas (Kuk, 1992, Ng y McLachlan, 2003), significa tiempos de computo mayores. El estudio y análisis del comportamiento de los tiempos de supervivencia dado los tipos de causa se modula por la cópula y su dependencia ρ, entendiendo gráficamente su comportamiento a partir de las funciones de incidencia acumulada, las cuales se pueden calcular de forma sencilla y directa a partir de la ecuación (38), lo cual es otra bondad del modelo. Es importante destacar (como se hizo mención en la introducción) que la estructura de los modelos propuestos (modelo de Cox para marginal T, modelo multinomial para marginal D y dependencia modelada por cópula a través del parámetro ρ), genera un modelo con menos parámetros que el enfoque de

modelado vertical (Nicolaie et al., 2010), lo cual resulta ser una ventaja sobre este tipo de enfoque.

De los estudios de simulación se concluye que la eficiencia del parámetro de dependencia y la eficiencia de los parámetros del modelo logístico es similar en ambos enfoques, y evidentemente mayor para el componente lineal de la variable T, resultado que era de esperarse pues las simulaciones son considerando un modelo marginal de riesgos competitivos con función de base Weibull. Finalmente, ambos enfoques resultan robustos ante simulaciones realizadas con otra cópula, por lo que la cópula Gaussiana captura adecuadamente la dependencia ante una desviación del supuesto distribucional generador de los datos.

La aplicación de ambos modelos con datos reales se realizó en el capítulo 4. Dado que no se cumplía el supuesto de riesgos proporcionales para la variable stage, fue necesario la inclusión de una covariable dependiente del tiempo dada por $\text{stage} * \log(t)$. La elección de la función de $\log(t)$, sobre otro tipo de funciones temporales, se debió principalmente a que proporcionó un mejor ajuste en los datos. Las modificaciones al modelo paramétrico implicaron obtener una nueva función de densidad a partir de la derivación de la función de distribución, pues la componente lineal tenía un componente adicional dependiente del tiempo. En el modelo semiparamétrico fue necesario usar la codificación $(\text{start}, \text{stop}]$ para los tiempos de supervivencia descrita en Therneau et al. (2018), y así estimar el modelo de Cox usando la función coxph() del paquete survival. Lo anterior pone de manifiesto que, a través de pocas modificaciones y consideraciones, es posible extender estos modelos en el caso de que no se cumplan los riesgos proporcionales, lo que significa otra ventaja sobre otros enfoques para el análisis de riesgos competitivos como el modelo de causa específica (Prentice et al., 1978) o el modelo de Fine y Gray (1999), para los cuales es fundamental la proporcionalidad de riesgos (Lau et al., 2009).

Al aplicar la metodología propuesta a los datos de linfoma de células foliculares, se obtienen resultados que son coincidentes en las estimaciones puntuales de los parámetros para ambos modelos, indicando efectos similares de las covariables de forma marginal en ambas variables; sin embargo, los errores estándar son ligeramente mayores en el modelo semiparamétrico. La probabilidad de ocurrencia del evento 1 (no respuesta o recaída) a largo plazo $P(D = 1)$ generó intervalos de confianza al 95% de (0.5763, 0.6931) para el modelo paramétrico y de (0.5814, 0.6665) para el modelo semiparamétrico, mayor que el 50.3% que considera solamente datos

observados, lo que indica que la información que pueden brindar los datos censurados es importante a largo plazo. Que la componente lineal sólo estuviera compuesta del intercepto no fue algo inesperado, pues el análisis exploratorio marginal indicaba que no había dependencias muy marcadas. Es importante mencionar que cuando se tienen dos eventos que compiten, dada la parametrización de la ecuación (36), en caso de existan covariables significativas en el modelo, el valor de los parámetros da una idea de la incidencia del riesgo base (evento 1); esto significa que si el parámetro asociado a la variable es positivo, cuando se incrementa la variable aumentan la ocurrencia del evento, y si el parámetro es negativo, cuando crece el valor de la variable explicativa disminuye $P(D = 1)$. Como $P(D = 2)$ se escribe como $1 - P(D = 1)$, el efecto de los parámetros es contrario para el riesgo que compite. Lo anterior resulta bastante útil para entender los efectos que pueden tener las covariables de la componente lineal del marginal tipo de evento en las funciones de incidencia acumulada. Al analizar las funciones de incidencia acumulada, se observa que el incremento en la incidencia de ambos tipos de eventos crece con la edad y al volverse más grave el estado de la enfermedad, algo que resulta ser lógico, pues a mayor edad o al poseer un grado más avanzado de la enfermedad se esperan resultados menos favorables.

El hecho de que el uso de quimioterapia no haya resultado significativo en presencia de las demás variables, es un resultado importante, pues quiere decir que someter a los pacientes a este tipo de tratamiento no significaba una mejora en sus tiempos de no respuesta o recaída, ni en los tiempos de supervivencia. Si bien el resultado es sorpresivo, tampoco resulta estar tan alejado de la capacidad curativa que puede poseer el tratamiento de quimioterapia, pues en los resultados presentados en Morgan[VII] et al. (2004), se establece que el uso de este tipo de tratamientos solamente resulta en una contribución menor (aproximadamente del 2%) a la tasa de supervivencia a 5 años.

Es importante mencionar que las inferencias que se pueden obtener en este análisis están limitadas a la población anglosajona canadiense involucrada en el estudio, ya que como se establece en Becnel et al. (2017) existen diferencias en las tasas de supervivencia en pacientes

[VII] Se analizaron datos de 22 tipos de tumores de pacientes adultos, en los Estados Unidos de América y Australia.

con linfomas que van desde la raza y el género, hasta la orientación sexual y estado de infección por el virus de inmunodeficiencia humana.

En el caso específico de que las componentes lineales no compartan variables explicativas ni que existan variables dependientes del tiempo, se puede generar una interpretación de los parámetros de la componente lineal de supervivencia respecto a las funciones de incidencia acumulada, en el sentido de que variables explicativas con valores de parámetros positivos aumentan el valor de $CIF_d(t)$ -para cada valor de t y d - al incrementar su respectivo valor, con una relación inversa si los parámetros son negativos. Por la ecuación (38), la función de incidencia acumulada $CIF_d(t)$ se escribe en términos de la diferencia en cópulas $C(F_T(t), F_D(d)\,) - C(F_T(t), F_D(d-1)\,)$. En el caso general de una función de distribución bivariada $H(x,y)$ con dominio en $D_x \times D_y$, se cumple que para dos valores dados y_1 y y_2 en D_y para los cuales $y_1 \leq y_2$ la función $x \to H(x,y_2) - H(x,y_1)$ es no decreciente en D_x (Nelsen 2006); por lo que $CIF_d(t)$ es no decreciente. Como se está usando la cópula Gaussiana, la expresión $C(F_T(t), F_D(d)\,) = \Phi_2(\Phi^{-1}\{F_T(t)\}, \Phi^{-1}\{F_D(d)\}|\rho) = \Phi_2(\Phi^{-1}\{u\}, \Phi^{-1}\{v\}|\rho)$ garantiza que $CIF_d(t)$ es estrictamente creciente en $u = F_T(t)$, dado el valor de $D = d$. Por lo tanto para un valor de tiempo fijo t, un incremento de una variable con un valor positivo en su parámetro, se traduce en un incremento en $u = F_T(t)$ (dada la relación funcional de $F_T(t)$ descrita en la ecuación (27) para el modelo paramétrico y en la ecuación (31) para el modelo semiparamétrico). Por otro lado, un incremento en una variable con parámetro negativo, representa un decremento en $u = F_T(t|x)$, lo que produce que la función $CIF_d(t)$ disminuya. Cuando se cumplen las condiciones específicas (no compartir variables explicativas ni que existan variables dependientes del tiempo), este resultado es bastante útil para entender el comportamiento cualitativo de las funciones de incidencia respecto a las variables explicativas de la componente lineal del tiempo de supervivencia.

Es importante mencionar que en el caso de que los tiempos de supervivencia no se ajusten bien con un modelo paramétrico, el modelo semiparamétrico toma relevancia ajustando de forma más cercana el comportamiento de la función de riesgo; además, es importante mencionar que cuando las marginales no se ajustan bien, la estimación del parámetro de dependencia se puede ver afectada seriamente (Jordanger y Tjøstheim, 2014), motivo adicional de trascendencia del modelo semiparamétrico.

La técnica de adecuación a través de gráficas de calibración arrojó resultados favorables ya que las gráficas de las funciones de incidencia empíricas (que se pueden interpretar como el comportamiento "observado") fueron cercanas a las gráficas de las funciones de incidencia promedio modeladas (que se pueden interpretar como el comportamiento "esperado"), observándose tendencias cercanas a los 45°. Si bien es cierto que fue útil, es necesario reconocer que cuando se incrementa en demasía el número de covariables, ya no es tan fácil aplicar este tipo de enfoque; por lo que se requiere investigar y proponer otro tipo de metodología que pueda evitar este tipo de inconvenientes.

Es necesario mencionar los inconvenientes que pueden surgir con el uso de cópulas con marginales discretos. Como se mencionó el capítulo 2, cuando los marginales son discretos la cópula carece de unicidad. Esto en realidad no presenta un problema en modelos de aplicación, pues como lo establece Trivedi y Zimmer (2017) el uso de la cópula genera una forma (de muchas) de poder trabajar con la función de distribución conjunta que es desconocida o muy difícil de establecer. Además, como lo mencionan Genest y Nešlehová (2007) el hecho de que existan una infinidad de cópulas para la misma función de distribución no invalida los modelos de este tipo, argumento que Frees et al. (2016) refuerza mencionando algunos trabajos que usan este tipo de modelos y recordando que la cópula Gaussiana con datos binarios se ha usado por décadas por los investigadores con el nombre de modelo probit multinomial. Un problema más serio de acuerdo a Trivedi y Zimmer (2017), es el hecho de que la no identificabilidad puede afectar la estimación del parámetro de dependencia al introducir sesgo; sin embargo, como lo ilustran en su artículo a través de simulaciones empleando la cópula Gaussiana el hecho de introducir una estructura de regresión genera variación adicional en los argumentos de las cópulas cubriendo mayor espacio en el dominio y reduciendo considerablemente el sesgo en el parámetro de dependencia.

Mientras que la cópula Gaussiana proporciona una representación flexible y una estructura de dependencia rica, otras familias de cópulas pueden ajustar conjuntos de datos con ciertas características de manera más apropiada. En experimentos de simulación para la configuración continua bivariada, Nikoloulopoulos y Karlis (2008) encontraron que, para asociaciones relativamente pequeñas $\rho = 0.2$, diferentes cópulas pueden ajustarse adecuadamente a datos simulados de otras cópulas, incluyendo la cópula Gaussiana; sin embargo, también notaron que, a medida que la asociación aumenta, también lo hace la dificultad de encontrar cópulas con un ajuste adecuado; es probable que se produzcan escenarios similares en el marco actual

de riesgos competitivos, lo que exige estrategias efectivas para abordar esta cuestión. Una posibilidad es emplear los gráficos de calibración presentados en Kattan et al. (2003) para guiar la selección de cópula; sin embargo, dado que hay dos componentes lineales en el modelo, el número de covariables puede hacer que éste sea un ejercicio complejo. Aspectos importantes, como el tamaño de la muestra, el número de eventos y el seguimiento de los pacientes, influyen en el rendimiento del modelo en diferentes grados (Maller y Zhou, 2002), y, por lo tanto, se necesita profundizar más para lograr identificar de forma apropiada el tipo de cópula.

La metodología propuesta en este trabajo abre la puerta para modelar la dependencia que se puede presentar entre un biomarcador y los tiempos de supervivencia. En la medicina moderna, las decisiones en los diagnósticos o tratamientos personalizados generalmente son guiadas a través de biomarcadores (o también conocidos como marcadores biológicos) que discriminan entre sujetos con diferentes niveles de riesgo de desarrollar un padecimiento en el futuro; la correcta asignación de estos riesgos, disminuye la morbilidad y la mortalidad en la población. Por tal motivo, la determinación de su capacidad predictiva entorno a las enfermedades es de suma importancia en la investigación médica. Etiquetando a un biomarcador con la variable M, se podría modelar el vector aleatorio (M, T), usando las técnicas descritas en este trabajo, y entonces medir el poder predictivo de un marcador biológico en presencia de variables explicativas adicionales. Además de lo anterior, agregando la complejidad de riesgos que compiten, se pueden extender los modelos de regresión de cópulas, para modelar la probabilidad de un vector aleatorio con tres variables (M, T, D), y generar un modelo que describa la interdependencia entre un biomarcador, los tiempos de supervivencia y los diferentes eventos que compiten.

6.- Bibliografía

1. Andersen, P. K., Geskus, R. B., de Witte, T., & Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology*, 41(3), 861-870.
2. Andersen, P. K., & Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, *31*(11-12), 1074-1088.
3. Austin, P. C., & Fine, J. P. (2017). Practical recommendations for reporting Fine- Gray model analyses for competing risk data. *Statistics in medicine*, 36(27), 4391-4400.
4. Becnel, M., Flowers, C. R., & Nastoupil, L. J. (2017). Disparities in lymphoma on the basis of race, gender, HIV status, and sexual orientation. *Annals of lymphoma*, 1.
5. Bellera, C. A., MacGrogan, G., Debled, M., de Lara, C. T., Brouste, V., & Mathoulin-Pélissier, S. (2010). Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology*, 10(1), 20.
6. Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, *24*(11), 1713-1723.
7. Beyersmann, J., Allignol, A., & Schumacher, M. (2011). *Competing risks and multistate models with R*. Springer Science & Business Media.
8. Bilder, C. R., & Loughin, T. M. (2014). *Analysis of categorical data with R*. CRC Press.
9. Carpenter, J. & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19**, 1141–1164.
10. Carroll, K. J. (2003). On the use and utility of the Weibull model in the analysis of survival data. *Controlled clinical trials*, 24(6), 682-701.
11. Casella, G., & Berger, R. L. (2002). *Statistical inference (Vol. 2)*. Pacific Grove, CA: Duxbury.
12. Chen, Y., & Hanson, T. (2017). Copula regression models for discrete and mixed bivariate responses. *Journal of Statistical Theory and Practice*, 1-16.
13. Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/ CRC press.
14. Czado, C., Kastenmeier, R., Brechmann, E. C., & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, *2012*(4), 278-305.

15. de Leon, A. R., & Wu, B. (2011). Copula- based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, *30*(2), 175-185.
16. Drasgow, F. (2004). Polychoric and polyserial correlations. *Encyclopedia of statistical sciences*, *9*.
17. Efron, B. & Hastie, T. (2016). *Computer-Age Statistical Inference*. Cambridge University Press, Cambridge, UK.
18. Efron, B., & Tibshirani, R.J., (1994). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
19. Embrechts, P. (2009). Copulas: a personal view. *Journal of Risk and Insurance*, 76:639–650.
20. Escarela, G., & Carriere, J. F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4), 333-349.
21. Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, *94*(446), 496-509.
22. Frees, E. W., Lee, G., & Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1), 4.
23. Garthwaite, P. H., Jolliffe, I. T., Jolliffe, I. T., & Jones, B. (2002). *Statistical inference*. Oxford University Press on Demand.
24. Genest, C., & Nešlehová, J. (2007). A primer on copulas for count data. ASTIN Bulletin: *The Journal of the IAA*, 37(2), 475-515.
25. Haller, B., Schmidt, G., & Ulm, K. (2013). Applying competing risks regression models: an overview. *Lifetime data analysis*, 1-26.
26. He, J., Li, H., Edmondson, A. C., Rader, D. J., & Li, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics*, 13(3), 497-508.
27. Jiryaie, F., Withanage, N., Wu, B., & de Leon, A. R. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation*, *86*(9), 1643-1659.
28. Joe, H. (2014). *Dependence modeling with copulas*. CRC Press.
29. Jordanger, L. A., & Tjøstheim, D. (2014). Model selection of copulas: AIC versus a cross validation copula information criterion. *Statistics & Probability Letters*, 92, 249-255.

30. Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). John Wiley & Sons.
31. Kattan, M. W., Heller, G., & Brennan, M. F. (2003). A competing- risks nomogram for sarcoma specific death following local recurrence. *Statistics in medicine*, 22(22), 3515-3525.
32. Kim, H. T. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical Cancer Research*, *13*(2), 559-565.
33. Kleinbaum, D. G., & Klein, M. (2010). *Survival analysis (Vol. 3)*. New York: Springer.
34. Krämer, N., Brechmann, E. C., Silvestrini, D., & Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3), 829-839.
35. Kuk, A. Y. (1992). A semiparametric mixture model for the analysis of competing risks data. *Australian & New Zealand Journal of Statistics*, *34*(2), 169-180.
36. Lagakos, S. W., Sommer, C. J., & Zelen, M. (1978). Semi-Markov models for partially censored data. *Biometrika*, *65*(2), 311-317.
37. Larson, M. G., & Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Appl. Statist.* 34:201–211.
38. Lau, B., Cole, S. R., Moore, R. D., & Gange, S. J. (2008). Evaluating competing adverse and beneficial outcomes using a mixture model. *Statistics in medicine*, *27*(21), 4313-4327.
39. Lau B., Cole S. R., & Gange S. J. (2009). Competing risk regression models for epidemiologic data. *American journal of epidemiology*, 170(2), 244-256.
40. Lawless, J. F., & Yilmaz, Y. E. (2011). Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models. *Computational Statistics & Data Analysis*, *55*(7), 2446-2455.
41. Li, Y., Prentice, R. L., & Lin, X. (2008). Semiparametric maximum likelihood estimation in normal transformation models for bivariate survival data. *Biometrika*, 95(4), 947-960.
42. Liu, X. (2012). *Survival Analysis: Models and Applications*. Wiley & Sons Publication.
43. Lo, S. M., & Wilke, R. A. (2010). A copula model for dependent competing risks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2), 359-376.

44. Maller, R. A., & Zhou, X. (2002). Analysis of parametric models for competing risks. *Statistica Sinica*, 12(3), 725-750.
45. McLeish, D. L., & Small, C. G. (2012). *The theory and applications of statistical interference functions* (Vol. 44). Springer Science & Business Media.
46. Morgan, G., Ward, R., & Barton, M. (2004). The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies. *Clinical oncology*, 16(8), 549-560.
47. Murphy, S. A., & Van der Vaart, A.W. (2000). On profile likelihood. *Journal of the American Statistical Association* 95, 449–465.
48. Nelsen, R. B., (2006). *An Introduction to Copulas*, second ed. Springer, Berlin.
49. Ng, S. K., & McLachlan, G. J. (2003). An EM- based semi- parametric mixture model approach to the regression analysis of competing- risks data. *Statistics in Medicine*, *22*(7), 1097-1111.
50. Nicolaie, M., van Houwelingen, H. C., & Putter, H. (2010). Vertical modeling: A pattern mixture approach for competing risks modeling. *Statistics in medicine*, *29*(11), 1190-1205.
51. Nikoloulopoulos, A. K., & Karlis, D. (2008). Copula model evaluation based on parametric bootstrap. *Computational Statistics & Data Analysis*, 52(7), 3342-3353.
52. Pintilie, M (2007). *Competing Risks: A Practical Perspective*. John Wiley & Sons; New York.
53. Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., & Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 541-554.
54. Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, *26*(11), 2389-2430.
55. Scheike, T. H., & Zhang, M. J. (2011). Analyzing competing risk data using the R timereg package. *Journal of statistical software*, *38*(2).
56. Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64, 417-428.
57. Shi, Y., & Wu, M. (2016). Statistical analysis of dependent competing risks model from Gompertz distribution under progressively hybrid censoring. *SpringerPlus*, 5(1), 1745.
58. Song, P. X. K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer Science & Business Media.

59. Song, P. X. K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, 65(1), 60-68.
60. Therneau, T., Crowson, C. & Atkinson, E. (2018). Using time dependent covariates and time dependent coefficients in the Cox model. *Vignette for R survival package*.
61. Trivedi, P., & Zimmer, D. (2017). A note on identification of bivariate copulas for discrete count data. *Econometrics*, 5(1), 10.
62. Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1), 20-22
63. Xu, R., Vaida, F. & Harrington, D. P. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica* 19, 819–842.
64. Yilmaz, Y. E., & Lawless, J. F. (2011). Likelihood ratio procedures and tests of fit in parametric and semiparametric copula models with censored data. *Lifetime data analysis*, *17*(3), 386-408.
65. Zhang, M. J., & Fine, J. (2008). Summarizing differences in cumulative incidence functions. *Statistics in Medicine*, 27(24), 4939-4949.
66. Zhang, Z. (2016). Parametric regression model for survival data: Weibull regression model as an example. *Annals of translational medicine*, 4(24).
67. Zheng, M., & Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1), 127-138.
68. Zilko, A. A., & Kurowicka, D. (2016). Copula in a multivariate mixed discrete–continuous model. *Computational Statistics & Data Analysis*, *103*, 28-55.

7.- Anexo

En esta sección se hará una breve descripción de la implementación en R, indicando los paquetes y las funciones específicas necesarias para realizar los pasos principales del análisis de los datos del estudio prospectivo de pacientes con linfoma folicular (Capítulo 4).

1. *Carga de la información*. - Se usa la función read.table() del paquete base utils.

```
fol <- read.table(file="https://raw.githubusercontent.com/scheike/
update-code-for-jss-comp.risk/master/follic.txt", sep=",",header=TRUE);
```

2. *Formato de la información*. - Se crean y modifican variables que sirven para el análisis de los datos.

```
#Variables auxiliares para crear la variable causa
evcens <- (fol$resp=="NR" | fol$relsite!="")+0
crcens <- (fol$resp=="CR" & fol$relsite=="" & fol$stat==1)+0

#La causa de muerte, tiempo de supervivencia y censura
fol$causa <- evcens+2*crcens
fol$Ti <- fol$dftime
fol$censura <- ifelse( fol$causa==2, 1, fol$causa )

#Definición de variables numéricas
fol$stage <- as.numeric(fol$clinstg==2)
fol$chemo <- as.numeric(fol$ch=="Y")
fol$stagelnT <- fol$stage*log(fol$Ti)
```

3. *Tablas de datos*. - Se crean las tablas necesarias para el análisis; en particular se usa una función llamada formato_start_stop()[VIII] que genera la codificación $(start, stop]$ lo que permite modelar variables dependientes del tiempo usando la función coxph().

```
#Matriz de datos
follic <- fol %>% select(Ti,causa,censura,age,hgb,stage,chemo,stagelnT)

#Ordenamiento de la base follic respecto al tiempo Ti
follic_O <- follic %>% arrange(Ti)

#Creación de base con formato (start,stop]
follic_O_start_stop <- formato_start_stop(follic_O)
```

[VIII] La cual se anexa en la parte final del código junto con las demás funciones creadas que complementan el análisis.

4. *Análisis marginal de la variable D.-* Se hace un análisis univariado de la variable tipo de evento, para obtener los valores iniciales de los parámetros. Se usa la función glm() del paquete stats, para generar un modelo logístico (aplicado a los datos sin censura). Se crea el vector de parámetros iniciales que contempla los parámetros de la marginal D y la dependencia de la cópula.

```
#Valores iniciales de los parámetros para la componente D
datosInicio <- follic_O[follic_O$causa!=0,]
datosInicio$y <- ifelse(datosInicio$causa == 2, 0, 1)
datosInicio$y1 <- ifelse(datosInicio$causa == 2, 1, 0)
glm.inicio <- glm(y~1,data=datosInicio,family=binomial)

#Vector inicial sin la parte de supervivencia
inicial_DC <- as.numeric(c(glm.inicio$coef[1:length(glm.inicio$coef)],
                  atanh(cor(datosInicio$y1,datosInicio$Ti))))
```

5. *Análisis marginal de la variable T (modelo paramétrico).* - Para encontrar los valores iniciales de los parámetros del modelo paramétrico se usa la función survreg() del paquete survival para encontrar los parámetros ν y κ, complementando el análisis con la función aftreg() del paquete eha, la cual permite usar la codificación $(\text{start}, \text{stop}]$ para modelar la variable $\text{stage} * \log(t)$.

```
#Valores iniciales para la componente lineal de la variable T
CoxWeibull <- survreg(Surv(Ti,censura)~ age + stage + stagelnT,
                      data = follic_O,dist="weibull")

#Modelo paramétrico de Weibull en formato largo
CoxEHA <- aftreg(Surv(tstart, tstop, censura)~ age + stage + stagelnT,
                 data = follic_O_start_stop,dist="weibull",
                 param="lifeExp")

#Codificación de variables
Var_T <-c("age","stage","stagelnT")
CoxWeibullConv <- CoxWeibull
CoxWeibullConv$coefficients[2:(length(Var_T)+1)] <-
    CoxEHA$coefficients[1:length(Var_T)]
```

6. *Codificación para un modelo de Cox*.- La función aftreg() tiene una codificación de un modelo de tiempo de fallo acelerado, por lo que es necesario usar la función ConvertWeibull() del paquete SurvRegCensCov para transformar los parámetros y que estén codificados para un modelo de Cox (Zhang, Z., 2016). Finalmente se crea el vector de parámetros iniciales que contempla también la componente lineal del modelo de supervivencia paramétrico.

```
#Cambiar la parametrización
paraT_weibull <- ConvertWeibull(CoxWeibullConv,conf.level = 0.95)

#Vector inicial completo
inicial_DCT <- as.numeric(c(inicial_DC,
              paraT_weibull$var[3:(0.5*length(paraT_weibull$var))],
              CoxEHA$coef[length(Var_T)+1],CoxEHA$coef[length(Var_T)+2]))
```

7. *Análisis marginal de la variable T (modelo semiparamétrico).* Se usa la función coxph() del paquete survival para crear un modelo de Cox semiparamétrico que contemple la codificación $(start, stop]$. Se usa la función survfit() del paquete survival (aplicada al objeto creado por la función coxph()) para estimar la función de supervivencia.

```
CoxFollic <- coxph(Surv(tstart, tstop, censura)~ age + stage + stagelnT,
           follic_O_start_stop )

#Creación de la función de supervivencia de base
datos_para_Sbase <- data.frame(matrix(rep(0,length(Var_T)),
                  ncol=length(Var_T), nrow=1))
colnames(datos_para_Sbase) <- Var_T
Sbase <- survfit(CoxFollic, newdata=datos_para_Sbase)$surv
```

8. *Creación de la función de supervivencia con datos repetidos*. Es necesario crear una función de supervivencia que contemple los datos repetidos[IX] (la función survfit() estima la función de supervivencia sin considerar tiempos repetidos).

```
#Reordenamiento para que los tiempos repetidos de los datos censurados
#queden al final del subgrupo
follic_O$causa.neg <- follic_O$causa*(-1)
follic_O_sub <- follic_O %>% arrange(Ti,causa.neg)

#Generar una función de supervivencia para cada observación de la base,
#que contemple a las observaciones repetidas
Sbase.ext <- rep(0,length(follic_O_sub$causa))
Sbase.ext[1] <- Sbase[1]
cuenta=2;
for (i in 2:length(follic_O_sub$causa)){
  if(follic_O_sub$Ti[i]==follic_O_sub$Ti[i-1])
  {
    Sbase.ext[i]=Sbase.ext[i-1]
  }
  else
  {
    Sbase.ext[i]=Sbase[cuenta]
    cuenta=cuenta+1;
  }
}
```

9. *Creación de vector de parámetros y selección de datos sin censura*. – Se crea un vector que guarda los parámetros estimados del modelo de Cox (que representan la primera parte del proceso de estimación en dos etapas). Finalmente se quitan los registros con censura, pues no son necesarios en la segunda etapa del proceso de estimación del modelo semiparamétrico.

```
#Parámetros del modelo de Cox para la marginal de supervivencia
param_T <- as.numeric(c(CoxFollic$coef[1:length(CoxFollic$coef)]))

#Quitar a los datos los registros con censura (PMLE)
follic_O_sub_nocens <- follic_O_sub %>% filter(causa!=0)
Sbase.ext_nocens <- Sbase.ext[follic_O_sub$causa!=0]
```

[IX] Se usa en la función de pseudo-verosimiltud que estima los parámetros del modelo semiparamétrico.

10. *Estimación del modelo paramétrico*. Para la estimación del modelo paramétrico es necesario crear dos matrices (una para cada componente) usando la función model.matrix() del paquete stats. Finalmente se usa la función nlm() del paquete stats, aplicada a la función objetivo mle.gauss(), para la estimación de los parámetros.

```
#Matrices de covariables
matriz.D <- model.matrix(~1, data=follic_O_sub)
matriz.T <- model.matrix(~age+stage+stagelnT, data=follic_O_sub)

#Estimación de parámetros (MLE)
system.time(paramfollic_MLE  <- nlm(p=inicial_DCT, hessian=TRUE,
            f=mle.gauss, tipo=follic_O_sub$causa, Ti=follic_O_sub$Ti,
            matriz.D=matriz.D, matriz.T=matriz.T))

#Parámetros estimados (MLE)
follicT_MLE <-round(c(paramfollic_MLE$estimate),4)
```

11. *Estimación del modelo semiparamétrico* .- La segunda etapa de estimación del modelo semiparamétrico se logra optimizando la función pmle.gauss.nocens.timecov() usando la función nlm() del paquete stats. Es necesario crear de nueva cuenta dos matrices, e incluir los parámetros de la componente de supervivencia (primera etapa de estimación), así como la función de supervivencia de base (sin datos censurados).

```
#Matrices de covariables
matriz.D_nocens <- model.matrix(~1, data=follic_O_sub_nocens)
matriz.T_nocens <- model.matrix(~age+stage+stagelnT,
                                data=follic_O_sub_nocens)

#Estimación de parámetros (PMLE)
system.time(paramfollic_PMLE <- nlm(p=inicial_DC,
            f=pmle.gauss.nocens.timecov, tipo=follic_O_sub_nocens$causa,
            Ti=follic_O_sub_nocens$Ti, matriz.D=matriz.D_nocens,
            matriz.T=matriz.T_nocens, S0=Sbase.ext_nocens,
            parametros_T=param_T))

##Parámetros estimados (PMLE)
follicT_PMLE <- round(c(paramfollic_PMLE$estimate,param_T),4)
```

12. *Funciones de incidencia acumulada*.- Para generar las funciones de incidencia acumulada es necesario generar funciones específicas que implementen las ecuaciones (39) y (40) para los enfoques paramétrico y semiparamétrico. El siguiente código ilustra el uso de estas funciones para la generación de las $CIF_d(t)$.

```
#Valores de las covariables
age_dado <- 68
stage_dado <- 0
stagelnT_dado <-log(follic_O$Ti)*stage_dado
cov.D.dado <-c(1)

nm <-length(follic_O$Ti)
cov.T.dado.matrix <- as.matrix(cbind(rep(age_dado,nm),
                    rep(stage_dado,nm),stagelnT_dado))
colnames(cov.T.dado.matrix) <- NULL

#PMLE
f.incid.1.PMLE.follic<-rep(0,nm)
f.incid.2.PMLE.follic<-rep(0,nm)

for( i in 1:nm)
{
  f.incid.1.PMLE.follic[i] <- cop.gauss.pmle.timecov(
  paramfollic_PMLE$estimate[1:(length(inicial_DC)-1)],
  tanh(paramfollic_PMLE$estimate[length(inicial_DC)]),
  param_T,0,cov.D.dado,cov.T.dado.matrix[i,],Sbase.ext[i])

  f.incid.2.PMLE.follic[i] <- ( 1- S.cox.pmle.timecov(Sbase.ext[i],
  cov.T.dado.matrix[i,],param_T) - cop.gauss.pmle.timecov(
  paramfollic_PMLE$estimate[1:(length(inicial_DC)-1)],
  tanh(paramfollic_PMLE$estimate[length(inicial_DC)]),param_T,
  0,cov.D.dado,cov.T.dado.matrix[i,],Sbase.ext[i]) )

}

#MLE
f.incid.1.MLE.follic<-rep(0,nm)
f.incid.2.MLE.follic<-rep(0,nm)
param_T_MLE <- c(paramfollic_MLE$estimate[(
            length(inicial_DC)+1):(length(inicial_DC)+length(param_T))],
                 exp(paramfollic_MLE$estimate[length(inicial_DCT)-1]),
                 exp(paramfollic_MLE$estimate[length(inicial_DCT)]))

for( i in 1:nm)
{
  f.incid.1.MLE.follic[i] <- cop.gauss.mle.timecov(
  paramfollic_MLE$estimate[1:(length(inicial_DC)-1)],
  tanh(paramfollic_MLE$estimate[length(inicial_DC)]),param_T_MLE,
  0,follic_O$Ti[i],cov.D.dado,cov.T.dado.matrix[i,])

  f.incid.2.MLE.follic[i] <- ( 1- S.cox.mle.timecov(follic_O$Ti[i],
  cov.T.dado.matrix[i,],param_T_MLE) - cop.gauss.mle.timecov(
  paramfollic_MLE$estimate[1:(length(inicial_DC)-1)],
  tanh(paramfollic_MLE$estimate[length(inicial_DC)]),param_T_MLE,
  0,follic_O$Ti[i],cov.D.dado,cov.T.dado.matrix[i,]))

}
```

13. *Gráficas de las funciones de incidencia acumulada*. Finalmente usando funciones como plot() y points() se grafican las funciones de incidencia acumulada.

```
#Gráficas de las funciones de incidencia acumulada
plot(follic_O_sub$Ti, f.incid.1.MLE.follic,xlab="Tiempo (años)",
     ylab="CIF",ylim=c(0,1),type="S", col=3)
points(follic_O_sub$Ti,f.incid.2.MLE.follic,type="S",  lwd=1, col=4)

points(follic_O_sub$Ti,f.incid.1.PMLE.follic,type="S",  lty=2, col=3)
points(follic_O_sub$Ti,f.incid.2.PMLE.follic,type="S",  lty=2,
      lwd=1,col=4)

title(main="Funciones de incidencia acumulada")
```

A continuación se anexa el código de las funciones creadas. Es importante mencionar que para que las siguientes funciones trabajen adecuadamente es necesario instalar y cargar los paquetes dplyr y mvtnorm en la sesión de R.

14. *Función formato_start_stop():*

```
formato_start_stop <- function(follic_O){

  follic_O$id <- as.numeric(rownames(follic_O))
  follic_O_N <- follic_O %>%
    select(id,chemo,stage,age,hgb,Ti,censura,causa)
  follic_O_N <- tmerge(follic_O_N , follic_O_N , id=id,
                       death = event(Ti, causa))
  tiempos <- unique(follic_O_N$Ti)

  follic_O_N$nt_id <- rep(1,length(follic_O_N$Ti))

  for( i in 1:length(follic_O_N$Ti)){
    j<-1
    while(follic_O_N$Ti[i]>tiempos[j]){
      j<-j+1;
    }
    follic_O_N$nt_id[i]<-j
  }

  follic_O_N$nt_id_1 <- follic_O_N$nt_id*follic_O_N$stage
  follic_O_N$stagelnT <- follic_O_N$stage*log(follic_O_N$Ti)
  tb_id <- follic_O_N %>% filter(nt_id_1>0) %>% select(id,nt_id_1)
  tb_base <-follic_O_N %>% select(Ti) %>% distinct(Ti,.keep_all= TRUE)

  tb_base$stagelnT <- tb_base$Ti
  for(i in 1 :(dim(tb_base)[1]-1)){
    tb_base$stagelnT[i]<-log(tb_base$Ti[i+1])
  }
  tam <- length(tb_base$Ti)
  tb_base[tam+1,] <- c(0,log(tb_base$Ti[order(tb_base$Ti)][1]) )
  tb_base <- tb_base %>% arrange(Ti)
  tb_base<-tb_base[-tam,]
```

```
follic_rep<-data.frame(id=0,Ti_rep=0,stagelnT=0) #OK

  for(i in 1:length(tb_id$id)){

    tb1<-cbind(rep(tb_id$id[i],tb_id$nt_id_1[i]),
      tb_base[1:tb_id$nt_id_1[i],])
    colnames(tb1)<-c("id","Ti_rep","stagelnT")
    follic_rep<-rbind(follic_rep,tb1)
  }

  tb_0<-follic_O_N %>% filter(nt_id_1<1) %>% select(id,Ti,stagelnT)
  colnames(tb_0)<-c("id","Ti_rep","stagelnT") #OK
  follic_O_N <- follic_O_N %>% select( -nt_id,-nt_id_1)
  rbind(follic_rep,tb_0)

  follic_rep <- follic_rep[-1,]
  follic_O_N1 <- tmerge(follic_O_N, follic_rep, id=id,
  stagelnT = tdc(Ti_rep, stagelnT))
  follic_O_N1$censura <- ifelse(follic_O_N1$death==0,0,1)
  return (follic_O_N1)
}#Termina la función formato_start_stop()
```

15. *Función pmle.gauss.nocens.timecov ():*

```
pmle.gauss.nocens.timecov <- function(parametros, tipo, Ti, matriz.D,
matriz.T,S0,parametros_T){

  n.var.D<-dim(matriz.D)[2]
  n.var.T<-length(parametros_T)+1
  J=max(tipo)
  uno <- rep(1, J)
  nt <- length(tipo)

  #Matriz para indicar la columna del tipo de causa
  C.ij <- matrix(1:(nt * J), ncol = J) * 0
  for(i in 1:J) {
  C.ij[, i][tipo == i] <- 1}

  #Asignación de parámetros
  deltas <- parametros[1:((J-1)*n.var.D)]
  theta<-tanh(parametros[length(parametros)])
  #Matriz covariables de T
  if(n.var.T == 1){
    Vij <- matriz.T[,1]}
  else if(n.var.T > 1){
    Vij <- matriz.T[,-1]}
```

```
#Evaluación de la componente lineal para D
  lineal.D <- matrix(1:(J * nt), ncol = J) * 0
  if(n.var.D==1){
    for(i in 1:(J - 1)) {
      lineal.D[, i] <- c(matriz.D * deltas[i])}
   }
  else if(n.var.D > 1){
    for(i in 1:(J - 1)) {
      lineal.D[, i] <- c(matriz.D %*%
        deltas[((i-1)*n.var.D + 1):(i*n.var.D)])}
   }

  #Evaluación de la componente lineal de T
  lineal.T <- rep(0,nt)
  if(n.var.T==1){
    for(i in 1:nt) {
      lineal.T[i] <- c(Vij[i,1]* 0)}
  }
  else if(n.var.T == 2){
    for(i in 1:nt) {
      lineal.T[i] <- c(Vij[i] * parametros_T)}
  }
  else if(n.var.T > 2){
    for(i in 1:nt) {
      lineal.T[i] <- c(Vij[i,1:(n.var.T-1)] %*% parametros_T)}
  }
  lineal.T.a <- rep(0,nt)
  Ti.a<- rep(0,nt)
  for ( i in 1:nt){
    Ti.a[i]<-max(Ti[Ti<Ti[i]&tipo!=0])
    if(Ti.a[i]<=-Inf){
      Ti.a[i]<-0.0001}
  }
  if(n.var.T==1){
    lineal.T.a<-lineal.T;}
  else if(n.var.T == 2){
    lineal.T.a<-lineal.T;}
  else if(n.var.T > 2){
    for(i in 1:nt) {
      lineal.T.a[i] <- c(Vij[i,1:(length(parametros_T)-1)] %*%
        parametros_T[1:(length(parametros_T)-1)]+
        parametros_T[length(parametros_T)]*
        Vij[i,(length(parametros_T)-1)]*log(Ti.a[i]) )}
  }
  #Probabilidad del modelo logístico
  p.ij <- matrix(1:(J * nt), ncol = J) * 0
  exp.lineal.D <- exp(lineal.D)
  sum.exp.lineal.D <- exp.lineal.D %*% uno

  for(i in 1:J) {
    p.ij[, i] <- exp.lineal.D[, i]/(sum.exp.lineal.D )}

  #Función de distribución de D=0
  FD0<-pbinom(0,size=1,prob=1-p.ij[,1]);
```

```
  #Funciones de supervivencia
  S0.a<-rep(0,nt)
  maximo<-max(S0)
  i_inicial<-sum( (maximo<=S0)*1 )

  for (i in 1:i_inicial ) {
    S0.a[i]<-1}
  for( i in (i_inicial+1):nt){
    S0.a[i]<-min(S0[S0>S0[i]])}

  exp.lineal.T <- exp(lineal.T)
  exp.lineal.T.a <- exp(lineal.T.a)
  S<- (S0)^{exp.lineal.T}
  S.a<-(S0.a)^{exp.lineal.T.a}
  FT=1-S
  FT.a=1-S.a

  #Cuantiles para la cópula normal
  qT<-qnorm(FT,mean=0,sd=1);
  qT.a<-qnorm(FT.a,mean=0,sd=1);
  qD0<-qnorm(FD0,mean=0,sd=1);

  #Evaluar la función pmvnorm()
  normal<-rep(0,nt)
  normal.a<-rep(0,nt)
  for (i in 1:nt){
    normal[i]<-pmvnorm(lower=c(-Inf,-Inf), upper=c(qT[i],qD0[i]),
      mean=c(0,0),corr=matrix(c(1,theta,theta,1),
      nrow=2,ncol=2,byrow=TRUE),sigma=NULL);
    normal.a[i]<-pmvnorm(lower=c(-Inf,-Inf), upper=c(qT.a[i],qD0[i]),
      mean=c(0,0),corr=matrix(c(1,theta,theta,1),
      nrow=2,ncol=2,byrow=TRUE),sigma=NULL);}

  l.ij <- matrix(1:(nt * J), ncol = J) * 0
  l.ij[,1]<-log(normal-normal.a)
  l.ij[,2]<- log(  -S + S.a - normal + normal.a )
  -sum(na.omit(C.ij*l.ij))
}#Termina la función pmle.gauss.nocens.timecov()
```

16. *Función mle.gauss():*

```
mle.gauss <- function(parametros,tipo,Ti, matriz.D,
matriz.T,parametros_T){

  n.var.D<-dim(matriz.D)[2]
  n.var.T<-dim(matriz.T)[2]
  J=max(tipo)
  uno <- rep(1, J)
  nt <- length(tipo)
  n_param<-length(parametros)
```

```
#Matriz para indicar la columna del tipo de causa
C.ij <- matrix(1:(nt * J), ncol = J) * 0
for(i in 1:J) {
  C.ij[, i][tipo == i] <- 1}
ci <- C.ij %*% uno

#Asignación de parámetros
deltas <- parametros[1:((J-1)*n.var.D)]
theta<-tanh(parametros[(J-1)*n.var.D+1])
betas<-parametros[ ((J-1)*n.var.D+2): (n_param-2) ]
#Transformación para garantizar que salen valores entre 0-1
k_T<-exp(parametros[n_param-1])
l_T<-exp(parametros[n_param])

#Matriz covariables de T
if(n.var.T == 1){
  Vij <- matriz.T[,1]}
else if(n.var.T > 1){
  Vij <- matriz.T[,-1]}
#Evaluación de la componente lineal para D
lineal.D <- matrix(1:(J * nt), ncol = J) * 0
if(n.var.D==1){
  for(i in 1:(J - 1)) {
    lineal.D[, i] <- c(matriz.D * deltas[i])}
}
else if(n.var.D > 1){
  for(i in 1:(J - 1)) {
    lineal.D[, i] <- c(matriz.D %*%
      deltas[((i-1)*n.var.D + 1):(i*n.var.D)])}
}
#Evaluación de la componente lineal de T
lineal.T <- rep(0,nt)
if(n.var.T==1){
  for(i in 1:nt) {
    lineal.T[i] <- c(Vij[i,1]* 0)}
}
else if(n.var.T == 2){
  for(i in 1:nt) {
    lineal.T[i] <- c(Vij[i] * betas)}
}
else if(n.var.T > 2){
  for(i in 1:nt) {
    lineal.T[i] <- c(Vij[i,1:(n.var.T-1)] %*% betas)}
}
#Probabilidad del modelo logístico
p.ij <- matrix(1:(J * nt), ncol = J) * 0
exp.lineal.D <- exp(lineal.D)
sum.exp.lineal.D <- exp.lineal.D %*% uno

for(i in 1:J) {
  p.ij[, i] <- exp.lineal.D[, i]/(sum.exp.lineal.D ) }
#Funciones de distribución
FD0<-pbinom(0,size=1,prob=1-p.ij[,1]);
FD1<-pbinom(1,size=1,prob=1-p.ij[,2]);
exp.lineal.T <- exp(lineal.T)
```

```
  #Funciones de supervivencia
  S0<-rep(0,nt)
  S0<- 1 - pweibull(Ti, shape=l_T, scale = k_T)
  S<- (S0)^{exp.lineal.T}
  FT=1-S

  #Cuantiles para la cópula normal
  qT<-qnorm(FT,mean=0,sd=1);
  qD0<-qnorm(FD0,mean=0,sd=1);
  qD1<-qnorm(FD1,mean=0,sd=1);
  C1_0<-rep(0,nt)
  C1_1<-rep(0,nt)
  densidad_T<-rep(0,nt)

  C1_0<-pnorm((qD0-theta*qT)/(sqrt(1-(theta^2))),mean=0,sd=1)
  C1_1<-pnorm((qD1-theta*qT)/(sqrt(1-(theta^2))),mean=0,sd=1)-
  pnorm((qD0-theta*qT)/(sqrt(1-(theta^2))),mean=0,sd=1)
  densidad_T <- S*exp.lineal.T*( 1/(k_T^l_T))*
  (parametros[n_param-2]*matriz.T[,n.var.T-1]+l_T)*Ti^(l_T-1);

  #Función de máxima verosimilitud
  l.ij <- matrix(1:(nt * J), ncol = J) * 0
  l.ij[,1]<-log(densidad_T*C1_0)
  l.ij[,2]<-log(densidad_T*C1_1)
  A<- sum(C.ij*l.ij)
  B <- sum( (1 - ci) * log(S))
  log.lik <- A + B
  -1*log.lik
}#Termina la función mle.gauss()
```

17. *Función cop.gauss.mle.timecov()*:

```
  cop.gauss.pmle.timecov <- function(parametros_D, theta,
  parametros_T, d,cov.D,cov.T.matrix.i,S0){

  #Función de distribución de T
  lineal.T<-parametros_T %*%  cov.T.matrix.i
  S<- S0^exp(lineal.T)
  FT <- 1-S

  #Función de distribución de D
  lineal.D <- parametros_D %*% cov.D
  prob_D <- exp ( lineal.D ) / (1 + exp ( lineal.D ))
  FD <- pbinom(d,size=1,prob=1-prob_D);

  #Funciones normales inversas
  qT <- qnorm(FT,mean=0,sd=1);
  qD <- qnorm(FD,mean=0,sd=1);
```

```
 #Cópula Gaussiana
  normal <- pmvnorm(lower=c(-Inf,-Inf), upper=c(qT,qD), mean=c(0,0),
  corr=matrix(c(1,theta,theta,1),nrow=2,ncol=2,byrow=TRUE),sigma=NULL);
  normal

 }#Termina la función cop.gauss.pmle.timecov()
```

18. *Función S.cox.pmle.timecov():*

```
S.cox.pmle.timecov <- function(S0,cov.T.matrix.i,parametros_T){

  lineal.T<-parametros_T %*%  cov.T.matrix.i
  S=S0^exp(lineal.T)
  S
}#Termina función S.cox.pmle.timecov()
```

19. *Función cop.gauss.mle.timecov():*

```
  cop.gauss.mle.timecov <- function(parametros_D,theta,
  parametros_T,d,t,cov.D,cov.T.matrix.i){

  #Función de distribución de T
  n_com_lin<-length(cov.T.matrix.i)
  S0=exp( -(t/parametros_T[n_com_lin+1])^(parametros_T[n_com_lin+2]) );
  lineal.T<-parametros_T[1:n_com_lin] %*% cov.T.matrix.i
  S=S0^exp(lineal.T)
  FT=1-S


  #Función de distribución de D
  lineal.D<-parametros_D %*%cov.D
  prob_D<-exp ( lineal.D ) / (1 + exp ( lineal.D ))
  FD<-pbinom(d,size=1,prob=1-prob_D);

  #Funciones normales inversas
  qT<-qnorm(FT,mean=0,sd=1);
  qD<-qnorm(FD,mean=0,sd=1);


  #Cópula Gaussiana
  normal<-pmvnorm(lower=c(-Inf,-Inf), upper=c(qT,qD),
  mean=c(0,0),corr=matrix(c(1,theta,theta,1),nrow=2,ncol=2,byrow=TRUE),
  sigma=NULL);
  normal
}#Termina la función cop.gauss.mle.timecov()
```

20. *Función S.cox.mle.timecov():*

```
S.cox.mle.timecov<-function(t,cov.T.matrix.i,parametros_T){

  n_com_lin<-length(cov.T.matrix.i)
  S0=exp( -(t/parametros_T[n_com_lin+1])^(parametros_T[n_com_lin+2]) );

  lineal.T<-parametros_T[1:n_com_lin] %*%  cov.T.matrix.i
  S=S0^exp(lineal.T)
  S
}#Termina la función S.cox.mle.time.cov()
```

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE DISERTACIÓN PÚBLICA

No. 00069
Matrícula: 2152800892

Modelos Paramétricos y Semiparamétricos de Regresión Basados en Cópulas para el Análisis de Riesgos Competitivos

En la Ciudad de México, se presentaron a las 12:30 horas del día 20 del mes de enero del año 2020 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DRA. SILVIA RUIZ VELASCO ACOSTA
DR. ALBERTO CASTILLO MORALES
DR. GABRIEL ESCARELA PEREZ
DRA. HORTENSIA JOSEFINA REYES CERVANTES
DR. GABRIEL NUÑEZ ANTONIO

Bajo la Presidencia de la primera y con carácter de Secretario el último, se reunieron a la presentación de la Disertación Pública cuya denominación aparece al margen, para la obtención del grado de:

DOCTOR EN CIENCIAS (MATEMATICAS)

DE: ALEJANDRO ROMAN VASQUEZ

y de acuerdo con el artículo 78 fracción IV del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, la presidenta del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

ALEJANDRO ROMAN VASQUEZ
ALUMNO

REVISÓ

MTRA. ROSALÍA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTA

DRA. SILVIA RUIZ VELASCO ACOSTA

VOCAL

DR. ALBERTO CASTILLO MORALES

VOCAL

DR. GABRIEL ESCARELA PEREZ

VOCAL

DRA. HORTENSIA JOSEFINA REYES CERVANTES

SECRETARIO

DR. GABRIEL NUÑEZ ANTONIO