Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00182
Matrícula: 2153805999

MODELO DE CÓPULA PARA LA INFERENCIA EN DATOS DE SUPERVIVENCIA EN ESTUDIOS OBSERVACIONALES.

MARIA CONCEPCION ROJAS BRENA
ALUMNA

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

En la Ciudad de México, se presentaron a las 16:00 horas del día 9 del mes de noviembre del año 2018 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. ALBERTO CASTILLO MORALES
DR. LIZBETH NARANJO ALBARRAN
DR. GABRIEL ESCARELA PEREZ

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRA EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: MARIA CONCEPCION ROJAS BRENA

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

aprobar

Acto continuo, el presidente del jurado comunicó a la interesada el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTE

DR. ALBERTO CASTILLO MORALES

VOCAL

DR. LIZBETH NARANJO ALBARRAN

SECRETARIO

DR. GABRIEL ESCARELA PEREZ

UNIVERSIDAD AUTÓNOMA METROPOLITANA
UNIDAD IZTAPALAPA
DIVISIÓN CIENCIAS BÁSICAS E INGENIERÍA

Modelo de cópula para la inferencia de datos de supervivencia en estudios observacionales

Presenta:

María Concepción Rojas Brena

para obtener el título de

Maestro en Ciencias
(Matemáticas Aplicadas e Industriales)

Asesor:

Dr. Gabriel Escarela Pérez

Sinodales:

Dra. Lizbeth Naranjo Albarrán

Dr. Alberto Castillo Morales

Dr. Gabriel Núñez Antonio

Ciudad de México, México Noviembre de 2018

Modelo de cópula para la inferencia de datos de supervivencia en estudios observacionales

María Concepción Rojas Brena

Noviembre, 2018

Dedicado a mi familia

Agradecimientos

A mi familia, principalmente a mis padres Jacinto y Armanda por brindarme su apoyo incondicional en todo momento. Mi especial agradecimiento a mi hermana Nallely por su paciencia, generosidad y complicidad. Agradezco a mis hermanos Uriel y Juan Carlos por ser siempre un ejemplo de inteligencia, valentía, capacidad y superación. Finalmente gracias a mis sobrinos por ser mi inspiración. Los amo.

A mi asesor de tesis el Dr. Gabriel Escarela Pérez por su dedicación y apoyo otorgado durante el desarrollo de la tesis, por facilitarme siempre los medios suficientes para llevar a cabo las actividades propuestas y por todos los consejos brindados durante el tiempo que trabajamos juntos.

A mis sinodales, Dra. Lizbeth Naranjo albarrán, Dr. Alberto Castillo Morales y Dr. Gabriel Núñez Antonio por sus correciones y sugerencias que han ayudado a que éste sea un mejor trabajo.

A la Universidad Autónoma Metropolitana por brindarme la oportunidad de continuar con mi formación académica, en especial a los profesores del departamento de matemáticas. Al CONACYT por la beca otorgada durante mi estancia en el posgrado.

Índice general

Introducción

Dentro de la epidemiología los estudios observacionales corresponden a un diseño de investigación clínico cuyo objetivo es observar, registrar y analizar un fenómeno para describirlo en su forma natural, sin intervención del investigador. Debido a la naturaleza no experimental de estos estudios la característica más destacada es la de no poder controlar la asignación del tratamiento de los sujetos o pacientes que se estudian. A diferencia de un estudio experimental en donde se puede elegir aleatoriamente el tipo de tratamiento al cual se somete cada sujeto, como en un ensayo clínico, en los estudios observacionales los tratamientos son asignados dependiendo de las características del paciente.

Las características que se toman en cuenta para la asignación de un tratamiento se conocen como variables de confusión, estas variables influyen tanto en la variable independiente (por ejemplo el tratamiento) como en la variable dependiente o respuesta; es decir, una variable confusión es causa del tratamiento y a la vez es causa de la variable respuesta (por ejemplo el tiempo de supervivencia). La existencia de las variables de confusión puede conducir a resultados erróneos si no se trabaja adecuadamente [21, 9, 25], ya que estas distorsionan la relación causal que puede haber entre el tratamiento y la variable respuesta, sugiriendo relaciones donde no las hay, exagerando la asociación real o al contrario atenuando la relación real (sobreestimar o subestimar el efecto causal, respectivamente).

Cuantificar el efecto del tratamiento en una variable respuesta no es tarea fácil cuando se trata de estudios observacionales, muchas veces se pierde objetividad en las inferencias causales al no realizar un análisis adecuado [25, 19]. El principal problema al no considerar la existencia de variables de confusión es el sesgo que se origina en los estimadores.

Existen diferentes metodologías para la reducción o control del efecto de las variables de confusión, las más comunes son dos: 1) la estratificación, que consiste en realizar estimaciones de medidas de asociación para cada subgrupo de la variable confusión; y 2) técnicas de análisis multivariado, en donde para ajustar el efecto de la variable de confusión basta con incluirla como una variable independiente del modelo multivariado y después realizar un análisis de regresión. Aunque estos métodos son muy usados en epidemiología no tienen teoría formal que reafirme su validez [21, 19].

Donald B. Rubin [21] realizó un análisis de inferencia a una base de datos sobre cáncer de mama para determinar si la cirugía con conservación de seno era más recomendable para el grupo de pacientes en el que el diagnóstico podría considerarse no tan grave. Este estudio controla el efecto de los factores de confusión a través de mecanismos de asignación. La idea principal de la metodología es conceptualizar que el conjunto de datos observacionales proviene de un experimento aleatorio complejo, en el cual los criterios para asignar los tratamientos son perdidos y deben reconstruirse. Esta metodología se ocupa en la etapa de diseño del estudio. El estudio concluye que no existe evidencia de que la mastectomía sea la mejor opción para el grupo de pacientes con diagnóstico considerado no tan grave.

Basado en el modelo de análisis de Rubin [21], esta investigación destaca la importancia de considerar la existencia de variables de confusión en la interpretación de estudios de datos observacionales, con el objetivo de controlar, concretamente,

el efecto de confusión en la etapa del análisis de datos. Por lo que se propone una metodología que se fundamenta en caracterizar una función conjunta bivariada para la variable respuesta y el tratamiento a través de una cópula. Se prefiere el enfoque de construcción para la función conjunta bivariada a través de cópula porque es flexible, además, separa la estructura de dependencia entre las variables y el comportamiento marginal. Por lo tanto facilita el anáisis, ya que se puede analizar la relación que existe entre el tratamiento y la variable respuesta, y a la vez permite incorporar a cada distribución marginal modelos que ayuden a entender el comportamiento de estas variables.

Sin embargo, una de las dificultades que se presenta en la construcción de la función conjunta, es que la variable respuesta es una variable continua y el tipo de tratamiento esta descrito por una variable discreta, y aunque la teoría de cópulas ha sido muy estudiada para los casos en los que todas la marginales son continuas, cuando se agrega alguna variable discreta parte de la teoría deja de ser válida debido a que se pierde la unicidad de la función conjunta, dicha dificultad se resuelve en este trabajo agregando una variable ficticia o latente, la cual se define con la variable discreta.

El primer modelo marginal incluido es el modelo de Cox uno de los más conocidos en el análisis de supervivencia y el que se utiliza para modelar los tiempos de supervivencia. Con este modelo se conseguirá identificar aquellas variables explicativas que afectan al tiempo de supervivencia, es decir, los factores de riesgo. La segunda estructura marginal se describirá a través del modelo de regresión logístico multinomial con el cual se calculan las probabilidades correspondientes de cada tratamiento, este modelo nos ayudará a identificar las variables explicativas que afectan en la selección del tratamiento. En la estructura de dependencia se anexa un modelo lineal con el objetivo de observar si existe algún factor de confusión.

La metodología se ilustrará en una cohorte de un seguimiento de 20 años realizado por el programa Surveillance, Epidemiology, and End Results (SEER) del National Cancer Institute de Estados Unidos, de mujeres diagnosticadas con cáncer de mama que se sometieron a una cirugía. Los registros que este programa recolecta corresponden a información de pacientes residentes de Alaska, California, Connecticut, Georgia, Hawaii, Ioaw, Michigan, Nuevo México, Utah y Washington, los cuales corresponde a un 26 % de la población total de Estados Unidos de América. La información recolectada en los registros del SEER incluyen características socio-demográficas, como la edad, el estado civil, la raza de la paciente. Y caraterísticas clinicopatológicas, como el tamaño, tipo y grado del tumor, así como la fecha de diagnóstico de cáncer y la fecha de muerte, el tipo de cirugía y el estatus de la aplicación de radioterapia pos-operatoria.

El objetivo particular de la aplicación a los datos de cáncer de mama es modelar los tiempos de supervivencia para los cuatro posibles tratamientos en presencia de factores de riesgo y variables de confusión y explorar si existen grupos de pacientes que pueden beneficiarse más con algún régimen de tratamiento.

La presentación del trabajo de la investigación se realiza en tres partes: El primer capítulo se enfoca a la variable respuesta, los tiempos de supervivencia. Por lo que en este apartado se empieza dando una breve introducción al análisis de supervivencia, describiendo características y funciones principales y fundamentales que se utilizan para el análisis de los datos, tales como: la función de supervivencia y la función de riesgo, incorporando además estimadores tanto paramétricos como no-paramétricos de las misma. En este capítulo se anexá el modelo de Cox, uno de los modelos más utilizados en análisis de datos de supervivencia. Finalmente, se agrega una sección para describir el método de máxima verosimilitud, método que se elige para encontrar los estimadores finales del modelo propuesto.

En el capítulo dos se desarrolla la metodología propuesta en la tesis para el análisis de datos. Este capítulo empieza con una introducción a la teoría de cópulas, enfocándose a la cópula gaussiana dado que esta es la que se ocupa en la siguiente sección, la construcción de la función de densidad, dicha construcción además de utilizar la cópula gaussiana se apoya de una variable latente para poder definir adecuadamente la función de densidad conjunta para una variable continua y una variable discreta. Se anexa el modelo de regresión multinomial, el cual se ocupado para describir el comportamiento del tratamiento. Y posteriormente, se obtiene la función de verosimilitud, tomando en consideración la existencia de datos censurados; para comprobar la validez de la función de verosimilitud se anexa una simulación para el modelo nulo. Finalmente, se agrega una sección para explicar el método de imputación múltiple, método que se utiliza en la aplicación de modelo propuesto debido a que la base que se utiliza en la aplicación tiene datos faltantes.

En el tercer capítulo se ilustra la metodología propuesta a través de una cohorte de datos de mujeres diagnosticadas con cáncer de mama. Este capítulo inicia dando una introducción al cáncer de mama en donde se incluyen conceptos básicos para entender cada una de las variables explicativas propuestas para el modelo. Se realiza una pequeña descripción de la base de datos y posteriormente se obtienen los estimadores finales del modelo. Por último, se realiza un análisis de los tiempos de supervivencia a través función de densidad conjunta condicionada al tratamiento y se obtienen tres grupos de riesgo: alto, medio y bajo.

Finalmente, se describen las conclusiones obtenidas durante el desarrollo de la tesis. Además, se anexan tres apéndices: En el primero se encuentran los programas en el lenguaje de programación R utilizados para la estimación de los parámetros, en el segundo están todos los métodos numéricos que se ocupan durante el desarrollo de la tesis y finalmente, en el último apéndice se anexan las

tablas de frecuencia de las 7 imputaciones.

Capítulo 1

Análisis de supervivencia

El tipo de datos que describe la variable respuesta en este trabajo de tesis son datos de supervivencia, los cuales representan longitudes de tiempo que corresponden por lo general a estudios de seguimiento, en donde una de sus principales características es la pérdida de información durante el seguimiento, por lo que es necesario un tratamiento especial para el análisis de estos datos. Debido a lo anterior se considera importante dedicar este capítulo no solo a describir el modelo de Cox el cual se utiliza para modelar los tiempos de supervivencia, sino también proporcionar las herramientas necesarias sobre el análisis de supervivencia que sirvan para tener mayor claridad al realizar el desarrollo de la metodología propuesta.

Análisis de supervivencia es una frase usada para referirse al estudio de desarrollos y técnicas estadísticas para el análisis de variables aleatorias que representan medidas de tiempos, a las cuales se le denominan *tiempos de supervivencia*. Usualmente, las técnicas de análisis de supervivencia eran usadas en estudios sobre medicina; sin embargo, en la actualidad este concepto ya no se reduce solo a términos de vida y muerte, sino también es utilizada en otras áreas como finanzas y economía.

1.1. Datos de supervivencia

Los *tiempos de supervivencia* son definidos como la medida de tiempo que transcurre de un tiempo de origen bien definido hasta que ocurre un evento en particular (*tiempo de fallo*). Aunque en un principio, el evento particular estudiado se refería a la muerte de un paciente, el concepto se ha ido generalizando generalizado para poder ocuparse en diferentes estudios de medicina y otras áreas. Algunos ejemplos de tiempos de supervivencia son:

a) El tiempo que transcurre del inicio de un tratamiento hasta la recuperación del paciente.

b) El tiempo que pasa de la infección del virus VIH hasta el desarrollo del SIDA.

c) El tiempo que transcurre del diagnóstico del cáncer de mama hasta la muerte del paciente.

d) El tiempo que transcurre de la contratación de un seguro de auto hasta que ocurra un accidente.

e) El tiempo que transcurre de la graduación de un estudiante de la universidad hasta su primer trabajo.

Como se observa en los ejemplos anteriores, al definir tiempos de supervivencia se tiene un punto inicial y un punto final. Por lo cual se puede ver al análisis de supervivencia como estudios de seguimiento, en donde el punto de inicio es el ingreso al estudio y el punto final es la ocurrencia del evento. Aunque lo idóneo es que el punto final sea la ocurrencia del evento, en estudios de seguimiento no siempre se espera a que todos los sujetos del estudio experimenten el evento de interés, debido a que esto puede tardar mucho tiempo o quizás nunca suceda, por lo que regularmente se determina una fecha de fin de estudio o un intervalo de

tiempo de estudio. Así, el punto final se determina como la ocurrencia del evento de interés o el término del estudio.

En el trabajo de tesis el evento de interés es la muerte de la paciente; por lo tanto, a partir de ahora cuando se mencione tiempos de supervivencia se referirá al tiempo que transcurre hasta la muerte de una paciente después de ser diagnosticada.

1.1.1. Datos censurados

Una característica importante de los datos de supervivencia es que durante el estudio a veces existen pérdidas importantes de la observación del evento de interés; por ejemplo, cuando llega la fecha del término del estudio existen pacientes vivos, el paciente no presentó el evento de interés y desconocemos cuándo lo va a presentar, a estos tiempos se conocen como *tiempos censurados*.

En general, los tiempos censurados son datos en donde no se conoce con exactitud el tiempo de fallo; es decir, los tiempos de censura son los tiempos de supervivencia que corresponden a pacientes de los cuales desconocemos su estado (muerto o vivo) al final del estudio, ya sea debido a que al concluir el estudio pueden haber pacientes que aún siguen vivos, pacientes que durante el estudio se les perdió la pista o personas que murieron por otra razón distinta a la enfermedad estudiada (un accidente).

Existen distintos tipos de datos censurados y los métodos que se utilizan para la inferencia de los datos de supervivencia dependerán del tipo de censura que se presenta en el estudio, por lo tanto, es importante conocer el tipo de censura que existen para saber qué tipo de herramientas se deben utilizar para realizar el análisis. Una clasificación de la censura es:

a) Censura tipo I: Es cuando el investigador fija un tiempo máximo de observación de los individuos para que presenten el fallo. Si los individuos al término de este tiempo no han presentado el fallo (por ejemplo, pacientes que siguen vivos al final del estudio) entonces se consideran como observaciones con censura de tipo I.

b) Censura tipo II: Este caso de censura, es cuando el investigador decide prolongar el estudio hasta que ocurran k fallos de n individuos observados. Los individuos que no presentan la falla al completarse las primeras k (por ejemplo, los pacientes que siguen vivos después de que murió el k-ésimo paciente) se consideran datos con censura de tipo II.

c) Censura aleatoria: Este tipo de censura ocurre sin ningún control del investigador. Las censuras se presentan por abandono del estudio de un individuo, por la pérdida de seguimiento o por muerte por otra causa (por ejemplo, se pierde el seguimiento porque el paciente se cambió de ciudad).

Existen otros tipos de censura, por ejemplo, cuando se sabe que de haber ocurrido una falla ésta se presentará después del tiempo de censura observada, a esto se le conoce como *censura por la derecha*, es el tipo de censura más común en estudios relacionados con medicina y el que se ocupará en la tesis, ya que se sabe que todos pacientes algún día morirán pero se desconoce cuándo.

Otro tipo de censura es la *censura por la izquierda* que ocurre cuando antes de entrar al estudio, el individuo ha presentado la falla, por lo que se dice que su tiempo de fallo no observado es menor que el tiempo de censura observado.

En la Figura 1.1, se observa en forma de diagrama los tiempos de supervivencia de un estudio de ocho pacientes, en el cual la entrada del sujeto al estudio está representado por "•", la letra **M** representa muerte, **P** la pérdida de seguimiento

del paciente y finalmente **V** simboliza que el paciente sigue vivo al terminar el estudio. Los sujetos 1, 4, 5 y 8 mueren durante el estudio, mientras que a los sujetos 2 y 7 se les pierde en el seguimiento y los sujetos 3 y 6 permanecen vivos al final del estudio. Por lo tanto, cuando se realice el análisis la longitud de la línea de los pacientes 1, 4, 5 y 8 se toman como tiempos de supervivencia y la de los paciente 2, 3, 6 y 7 como tiempos de supervivencia censurados.

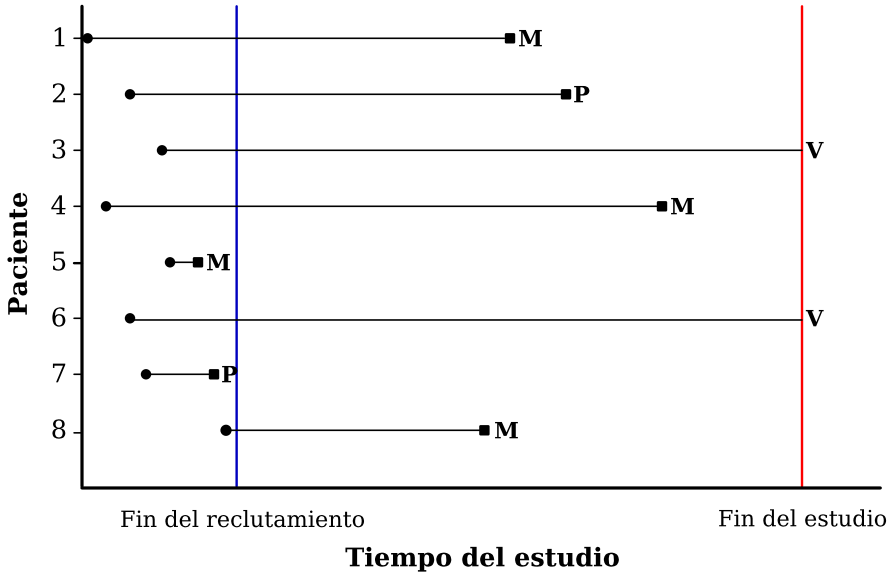

Figura 1.1: Tiempo de estudio para el análisis de 8 pacientes en un estudio de superviviencia.

1.2. Función de supervivencia y mortalidad

Para realizar el análisis de datos de supervivencia existen dos funciones de interés central, la *función de supervivencia* y la *función de mortalidad*. Estas fun-

ciones son esenciales en los temas de análisis de supervivencia.

Sea Y la variable aleatoria que mide los tiempos de supervivencia y que tiene una distribución de densidad $f(y)$. Entonces, la probabilidad $\mathbb{P}$ de que un paciente viva menos de un tiempo y es la función de distribución $F(y)$

$$\mathbb{P}(Y \leq y) = F(y) = \int_0^y f(u)du.$$

La probabilidad de que un individuo viva más allá del tiempo y, es de mayor interés en el análisis de supervivencia, y está medida con la función de supervivencia. Si Y es una variable aleatoria no negativa que denota el tiempo de supervivencia y $f(y)$ su función de densidad. Entonces se define como *función de supervivencia* de Y a

$$S(y) = \mathbb{P}(Y \geq y) = \int_y^\infty f(u)du = 1 - F(y), \tag{1.1}$$

donde $F(y)$ es su función distribución acumulativa.

Otra función de gran interés es la función de Riesgo o también conocida como fuerza de mortalidad, la cual se relaciona con la probabilidad de morir instantáneamente en el tiempo y. Es decir, se define a la *función de mortalidad* como la probabilidad de morir en un instante infinitesimal entre y y $y + \delta y$, dado que se ha sobrevivido hasta el tiempo y, está dada por

$$h(y) = \lim_{\delta y \to 0} \frac{\mathbb{P}(y \leq Y \leq y + \delta y | Y \geq y)}{\delta y}. \tag{1.2}$$

Existe una relación muy estrecha entre la función de supervivencia, la función

de mortalidad y la función de distribución.

$$\begin{aligned} h(y) &= \lim_{\delta y \to 0} \frac{\mathbb{P}(y \leq Y \leq y + \delta y | Y \geq y)}{\delta y} \\ &= \lim_{\delta y \to 0} \frac{\mathbb{P}(y \leq Y \leq y + \delta y)}{\delta y \mathbb{P}(Y \geq y)} \\ &= \lim_{\delta y \to 0} \frac{\mathbb{P}(y \leq Y \leq y + \delta y)}{\delta y} \frac{1}{S(y)} \\ &= \frac{f(y)}{S(y)}. \end{aligned}$$

A tráves de la relación anterior y la integral de la función *log* es fácil probar la siguiente relación

$$S(y) = \exp(-H(y)), \tag{1.3}$$

donde $H(y) = \int_0^y h(u)du$ y se conoce como la *Función de mortalidad acumulativa.*

1.2.1. Estimación no paramétrica de la función de supervivencia

Existen métodos tanto paramétricos como no paramétricos para estimar la función de supervivencia, entre los métodos no paramétricos el estimador Kaplan-Meier es uno de los utilizados para estimar la probabilidad de supervivencia.

Estimador Klapan-Meier

El método de Klapan Maier supone que hay n individuos con $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ tiempos de supervivencia observados, donde algunos de los tiempos pueden ser censurados por la derecha o algunos individuos pueden tener el mismo tiempo de supervivencia, por lo que se supone que hay r tiempos distintos de supervivencia entre los individuos, con $r < n$. Se ordenan los r tiempos de supervivencia en orden ascendente $y_{(1)} < y_{(2)} < \ldots < y_{(r)}$, se toma como n_i al número de individuos que están vivos justo antes del tiempo $y_{(i)}$ (incluyendo a los individuos que

mueren en este tiempo) y a d_i al número de individuos que mueren en el tiempo $y_{(i)}$, para todo $i \in \{1, 2, \ldots, r\}$.

Durante el intervalo $(y_i - \delta, y_i)$, donde δ es un número infinitesimal ocurre una muerte, como hay n_i individuos vivos justo antes del tiempo $y_{(i)}$ y d_i muertos en $y_{(i)}$, entonces la probabilidad de que un individuo muera durante el intervalo $(y_{(i)} - \delta, y_{(i)})$ es estimada por d_i/n_i y la probabilidad estimada de supervivencia es $(n_i - d_i)/n_i$. Si un tiempo de censura ocurre al mismo tiempo que uno de supervivencia se considera que el dato censurado ocurre inmediatamente después que el tiempo de supervivencia a la hora de calcular los n_i.

El tiempo inmediatamente antes del siguiente tiempo de muerte no contiene muertes; es decir, en el intervalo $(y_{(i)}, y_{(i+1)} - \delta)$ no hay muertes, entonces la probabilidad de sobrevivir en este intervalo es de uno. Por lo tanto, la probabilidad conjunta de sobrevivir de $(y_{(i)} - \delta, y_{(i)})$ y $(y_{(i)}, y_{(i+1)} - \delta)$ es $(n_i - d_i)/n_i$ y cuando tomamos el límite cuando δ tiende a cero, el estimador para la probabilidad de supervivencia en el intervalo $(y_{(i)}, y_{(i+1)})$ es $(n_i - d_i)/n_i$.

Si se supone que las muertes de los individuos en la muestra ocurren independientemente, entonces la función de supervivencia estimada en el $k-$ésimo intervalo $(y_{(k)}, y_{(k+1)})$, con $k = 1, 2, \ldots, r$ y $y_{(r+1)} = \infty$, está dada por la probabilidad de sobrevivir en el intervalo $(y_{(k)}, y_{(k+1)})$ y todos lo intervalos anteriores. Por lo tanto, el estimador de Kaplan-Meier de la función de supervivencia es

$$S(y) = \prod_{i=1}^{k} \frac{n_i - d_i}{n_i}, \quad y \in [y_{(k)}, y_{(k+1)}), \tag{1.4}$$

para $k = 1, 2, \ldots, r$, con $S(y) = 1$ cuando $y < y_{(1)}$ y $S(y)$ indefinido para $y > y_{(r)}$ cuando es un observación censurada y si es una observación sin censura entonces $S(y) = 0$ cuando $y > y_{(r)}$.

Intervalo de Tiempo	n_i	d_i	$(n_i - d_i)/n_i$	$\hat{S}(y)$
0-	18	0	1.0000	1.0000
10-	18	1	0.9444	0.9444
19-	15	1	0.9333	0.8837
30-	13	1	0.9231	0.8137
36-	12	1	0.9167	0.7459
59-	8	1	0.8750	0.6526
75-	7	1	0.8571	0.5594
93-	6	1	0.8333	0.4662
97-	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

Tabla 1.1: Tabla del estimador Kaplan Meier

En el Tabla 1.1 se muestra una tabla de desarrollo del método de Klaplar Meier para el conjunto de tiempos de supervivencia $Y = \{10, 13^*, 18^*, 19, 23^*, 30, 36, 38^*, 54^*, 56^*, 59, 75, 93, 97, 104^*, 107, 107^*, 107^*\}$, donde el símbolo "*"significa que el tiempo de supervivencia es un tiempo censurado. En la Figura 1.2 se grafica la función de probabilidad estimada con el método Klaplan Meier, ahí se observa que la función de supervivencia estimada con este método es una función escalonada, en donde la estimación de supervivencia entre tiempos adyacentes es constante y decreciente en cada tiempo de muerte.

Finalmente, para estimar la función de mortalidad acumulada se ocupa la ecuación (1.3) y está dada por

$$H(y) = -\log S(y) = -\sum_{i=1}^{k} \log\left(\frac{n_i - d_i}{n_i}\right). \tag{1.5}$$

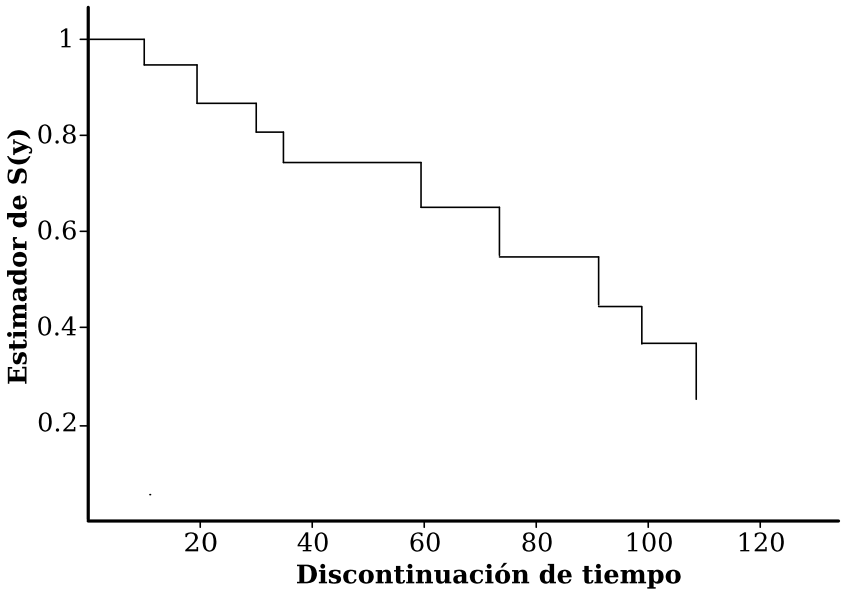

Figura 1.2: Gráfica de la función de supervivencia estimada mediante K-M.

1.2.2. Distribuciones de modelos paramétricos

Usualmente, en análisis de supervivencia se trabaja con las distribuciones exponencial, Weibull o log-normal para describir los datos supervivencia ya que estos tienden a ser asimétricos positivamente debido a que estos valores son positivos, por lo tanto, para estimaciones paramétricas suponer que los datos de supervivencia se distribuyen normalmente es erróneo. En el trabajo de tesis, se toma el supuesto que los datos de supervivencia tienen una distribución Weibull y aunque la sección se enfoca a esta distribución, se describe primero la función de distribución exponencial con el objetivo de tener una mayor comprensión.

Distribución exponencial

El modelo paramétrico más simple para el tiempo de supervivencia es la distribución exponencial, la cual toma como función de mortalidad una constante. Supongamos que la variable aleatoria Y está distribuida exponencialmente, por lo que su función de densidad de probabilidad es

$$f(y;\theta) = \theta e^{-\theta y} \qquad \text{con } y > 0,\ \theta > 0.$$

Su función de distribución acumulativa

$$F(y;\theta) = \int_0^y \theta e^{-\theta t} dt = 1 - e^{-\theta y}.$$

De la ecuación (1.1) se obtiene que la función de supervivencia es de la forma

$$S(y;\theta) = e^{-\theta y}$$

y de la ecuación (1.2) se obtiene que la función de mortalidad es

$$h(y;\theta) = \theta.$$

En este caso la función de riesgo no depende de la variable y; esto significa que la probabilidad de muerte para una paciente no depende del tiempo que ésta ha sobrevivido. A esta propiedad de la distribución exponencial se le llama pérdida de memoria, sin embargo, en la práctica se observa que la probabilidad de muerte incrementa con el tiempo, por esta razón se toma la distribución de Weibull para describir los tiempos de supervivencia.

Distribución Weibull

Otro modelo paramétrico comúnmente utilizado en el análisis de datos de superviviencia es el modelo de Weibull, debido a la forma que puede tomar esta

función se dice que juega un papel importante en el estudio de supervivencia.

Sea Y una variable aleatoria que tiene una función de distribución Weibull, entonces su función de densidad de probabilidad está dada por

$$f(y;\lambda,k) = \frac{\kappa}{\lambda}\left(\frac{y}{\lambda}\right)^{\kappa-1} e^{-\left(\frac{y}{\lambda}\right)^{\kappa}},$$

donde λ y k son el parámetros de escala y forma, respectivamente.

La distribución exponencial es un caso especial de la distribución de Weibull, cuando el parámetro de forma toma el valor 1.

La función de supervivencia obtenida con (1.1) es

$$S(y;\lambda,k) = e^{\left(-\frac{y}{\lambda}\right)^{\kappa}}$$

y su función de mortalidad es

$$h(y;\lambda,k) = \frac{\kappa}{\lambda}\left(\frac{y}{\lambda}\right)^{\kappa-1}.$$

En este caso podemos observar que la función de mortalidad sí depende de y, en donde un valor $\kappa < 1$ indica que la tasa de mortalidad decrece con el tiempo, cuando $\kappa = 1$ la tasa de mortalidad es constante en el tiempo y cuando $\kappa > 1$ indica que la tasa de mortalidad crece a lo largo del tiempo.

Por último, esta distribución tiene una cola de lado derecho más larga que la del lado izquierdo; es decir, que la función es sesgada del lado derecho(sesgo positivo), lo cual es conveniente para modelar datos de supervivencia.

1.3. Modelo de regresión de Cox

El modelo de Cox también conocido como modelo de riesgos proporcionales se ha convertido en uno de los métodos más utilizados en el ámbito de la salud, par-

ticularmente para analizar tiempos de supervivencia. El modelo de Cox se adecua dentro de la metodología propuesta para describir y entender el comportamiento de la variable respuesta la cual mide tiempos de supervivencia.

Uno de los principales supuestos de este modelo es que para dos conjuntos diferentes de variables explicativas se conserva la misma proporción a lo largo del tiempo de la *tasa de mortalidad*. El principal objetivo del modelo de Cox es determinar cuál combinación de las variables explicativas afectan la forma de la función de mortalidad.

Uno de los casos más simples del modelo de Cox es la comparación de la función de mortalidad para individuos en dos grupos. Por ejemplo, se supone que existen pacientes que van a recibir dos diferentes tratamientos, el estándar y un tratamiento nuevo. Si hay n datos de supervivencia con $h_i(y)$ su función de mortalidad para el i-ésimo paciente con $i \in \{1, 2, \ldots, n\}$, se denota $h_0(y)$ a la función de mortalidad para un individuo que toma el tratamiento estándar. Entonces el modelo de riesgos proporcionales se expresa de la forma, $h_i(y) = \psi_i h_0(y)$, con ψ_i una constante no negativa, la cual usualmente es tomada como $e^{\beta x_i}$.

La constante ψ_i se le conoce como *tasa de mortalidad* y mide el riesgo de muerte en algún tiempo para un individuo que toma el nuevo tratamiento con respecto a un individuo que toma el tratamiento clásico. Si se toma a la tasa de mortalidad como $e^{\beta x_i}$ el modelo de Cox se expresa como

$$h_i(y) = e^{\beta x_i} h_0(y),$$

en donde x_i es la variable explicativa que indica el tratamiento que tomó el i-ésimo paciente, la cual toma valor 0 para el tratamiento estándar y 1 para el tratamiento nuevo.

En forma más general el modelo de regresión de Cox toma en cuenta a $x = \{x_1, x_2, \ldots, x_m\}$ un vector constante de variables explicativas y se escribe como

$$h(y) = \psi(x)h_0(y), \tag{1.6}$$

donde $h_0(y)$ denota a la función de mortalidad de un individuo bajo condiciones estándar; es decir, cuando el vector de variables explicativas sea el vector nulo. La función $\psi(x)$ puede ser parametrizada, en el trabajo de tesis se considerará el logaritmo de un componente lineal

$$\psi(x;\beta) = e^{\beta^T x}. \tag{1.7}$$

Finalmente, el modelo se re-escribe de la siguiente manera al aplicar una transformación logarítmica

$$ln\left(\frac{h(y)}{h_0(y)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m.$$

Si se supone la proporcionalidad de la función de mortalidad entonces la función de supervivencia se puede expresar como

$$S(y) = [S_0(y)]^{\psi(x)}$$

y la función de densidad como

$$f(y) = \psi(x)\,[S_0(y)]^{\psi(x)-1}\,f_0(y).$$

Aunque el modelo de Cox es muy útil para análisis de los tiempos de supervivencia, este modelo no toma en consideración la posible existencia de alguna variable de confusión, por lo cual a la hora de medir el efecto causal que tiene el tratamiento puede ser errónea ya que los estimadores pueden ser sub o sobre estimados por el efecto de confusión. Por lo tanto, aunque este modelo es muy importante para describir los tiempos de supervivencia no es suficiente para describir el fenómeno completo.

1.3.1. Método de máxima verosimilitud

Los modelos paramétricos pueden ser ajustados dado un conjunto de datos observados usando el método de máxima verosimilitud para encontrar el vector de parámetros desconocidos θ de la función de densidad mediante la maximización de la función de verosimilitud.

La función de verosimilitud se define de manera similar a la función de densidad, sin embargo, la función de verosimilitud se considera en función de los parámetros y el conjunto de datos Y fijo, esto es, $L(\theta|Y) = f(Y|\theta)$. El objetivo del método es elegir como estimador el valor $\hat{\theta} \in \Theta$, donde Θ es el conjunto de todos los posibles estimadores del vector de parámetros θ, que maximiza la función $L(\theta|y)$. En la práctica generalmente se trabaja con el logaritmo de la función de verosimilitud $l(\theta|Y) = \log(L(\theta|Y))$ debido a que la función logaritmo es monótoma, el máximo de ambas funciones es el mismo y por varias razones es más conveniente usar el logaritmo.

El uso de los estimadores de máxima verosimilitud (EMV) son muy recomendados por la variedad de propiedades que satisfacen, entre ellas el estimador es eficiente, consistente y si existe una estadística suficiente para θ entonces el estimador es eficiente. Además los estimadores son invariantes a transformaciones funcionales.

La primera derivada de la función log verosimilitud tiene esperanza cero y varianza de la primera derivada es menos la esperanza de la segunda derivada; es decir

$$\begin{aligned} E(\bigtriangledown l(\theta|Y) &= 0, \\ var[\bigtriangledown l(\theta|Y)] &= -E[\bigtriangledown^2 l(\theta|Y),] \end{aligned}$$

en donde $\bigtriangledown l(\theta|Y)$ y $\bigtriangledown^2 l(\theta|Y)$ denotan el Gradiente y el Hessiano de $l(\theta|Y)$. A la segunda ecuación se le conoce como la *información de Fisher* y se denota por $I(\hat{\theta})$.

La distribución del estimador de máxima verosimilitud es una normal multivariada con media θ y varianza $I(\theta)^{-1}$, donde θ es el verdadero valor del parámetro. Es posible utilizar la varianza asintótica usando $I(\hat{\theta})^{-1}$ como la varianza aproximada del estimador de másima verosimilitud. La teoría asintótica garantiza que esta aproximación produce un error que es despreciable comparado con la mejor aproximación de la distribución del EMV dada por $N(I(\theta)^{-1})$.

Capítulo 2

Modelo mixto

El enfoque que propone la metodología a desarrollar en el trabajo de tesis brinda una manera sencilla de modelar la dependencia que existe entre el tratamiento y la variable respuesta y a la par ajustar cada una de estas variables a algún modelo que describa su comportamiento. Ya que una de las principales características de las cópulas es extraer la estructura de dependencia de las variables aleatorias del comportamiento marginal, se facilita no solo la etapa de modelar sino también el análisis de resultados al poder distinguir fácilmente las variables de riesgo, variables de confusión y variables de selección del tratamiento.

Sin embargo, aunque la idea de trabajar con cópulas se considera viable con los objetivos que se proponen para el trabajo, una de las dificultades que se presenta en la construcción de la metodología es tener una variable continua que representa los tiempos de supervivencia (o la variable respuesta) y una variable discreta que modela el tipo de tratamiento, ya que aunque la teoría de cópulas está muy desarrollada para el caso en donde todas las marginales son continuas, cuando se agrega alguna variable discreta parte de la teoría deja de ser válida.

2.1. Teoría de cópulas

Describir la relación que existe entre una función de distribución conjunta bivariada y sus dos funciones de distribución marginal es muy fácil cuando las variables aleatorias son independientes. Sin embargo, cuando el supuesto de independencia no se cumple escribir esta relación no es tan sencillo. En 1959 Sklar demostró la existencia de una función a la cual llamó cópula y la utilizó para unir las marginales unidimensionales y producir así una función de distribución conjunta. Desde entonces la cópula se ha convertido en una herramienta poderosa para el modelado multivariado, ya que aparte de utilizarla para modelar la dependencia entre las variables, ésta separa la dependencia del comportamiento marginal, lo que representa una manera flexible para realizar análisis multivariado.

Una *Cópula bivariada* es una función de distribución bivariada $C : \mathbb{I}^2 \to \mathbb{I}$ cuyas funciones marginales se distribuyen uniformemente entre $[0, 1]$; es decir,

$$C(v, w) = \mathbb{P}(V \leq v, W \leq w),$$

con $V \sim U(0, 1)$, $W \sim U(0, 1)$.

En el teorema de Sklar no solo se demuestra que la cópula es una función de distribución sino que también re-escribe la función de distribución conjunta F en términos de sus marginales y una cópula. Esto es, si F es una función de distribución bivariada con marginales F_1 y F_2. Entonces, asegura la existencia de una cópula bivariada, tal que para todo $y \in \mathbb{R}^2$ se cumple que

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)). \tag{2.1}$$

La unicidad de la cópula se garantiza solo cuando F_1 y F_2 son continuas. Además, el inverso también se cumple; es decir, si C es una cópula bivarida y F_1, F_2 son funciones de distribución, entonces la función F definida por 2.1 es una función de distribución bivariada con marginales F_1 y F_2.

El teorema de Sklar es uno de los teoremas más importantes en la teoría de cópulas debido a que no solo ofrece una manera sencilla de construir funciones de distribución conjunta, sino que también dada una función conjunta con sus distribuciones marginales se puede extraer una cópula.

Dada una función de densidad bivariada F con marginales continuas F_1 y F_2, como se indica en el teorema de Sklar, se puede construir una cópula de la siguiente manera

$$C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)), \tag{2.2}$$

donde F_i^{-1} es la función inversa generalizada,

$$F_i^{-1}(t) = \text{ínf}\{y \in \mathbb{R} | F_i(y) > t\},$$

para toda $i \in \{1, 2\}$.

Además, es sencillo definir a la función de densidad de la cópula, cuando C es una cópula bivarida diferenciable; esto es, cuando sus funciones de distribución marginales son continuas

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1, \partial u_2}.$$

Por otra parte, si la función de distribución F es obtenida mediante el teorema de Sklar, con marginales F_1 y F_2 entonces la función de densidad está dada por

$$f(y_1, y_2) = c(F_1(y_1), F_2(y_2)) f_1(y_1) f_2(y_2), \tag{2.3}$$

donde f_1 y f_2 son las funciones de densidad de F_1 y F_2, respectivamente.

Para medir la dependencia que existe entre dos variables hay distintas maneras, uno de los más utilizados es el *coeficiente de correlación de Pearson* el cual para cuantificar la dependencia entre las variables Y_1 y Y_2 se obtiene de la siguiente manera

$$\iota = Cor(Y_1, Y_2) \quad = \quad \frac{Cov(Y_1, Y_2)}{\sqrt{Var[Y_1]}\sqrt{Var[Y_2]}},$$

este coeficiente de correlación mide la fuerza y la dirección lineal de las dos variables. Sin embargo, la utilidad del coeficiente es poca ya que su valor no solo depende de la cópula, sino también de las marginales, pues esta medida no es siempre invariante sobre el re-escalamiento; es decir, $cor(f(Y_1), f(Y_2)) \neq cor(Y_1, Y_2)$ cuando f no es lineal [8].

Una manera diferente de cuantificar la relación entre dos variables Y_1 y Y_2 es utilizar medidas de concordancia y discordancia. En donde, para el vector de variables aleatorias (Y_1, Y_2), si (y_{1i}, y_{2i}) y (y_{1j}, y_{2j}) son dos observaciones, entonces se dice que (y_{1i}, y_{2i}) y (y_{1j}, y_{2j}) son *concordantes* si $(y_{1i} - y_{1j})(y_{2i} - y_{2j}) > 0$ y *disconcordantes* si $(y_{1i} - y_{1j})(y_{2i} - y_{2j}) < 0$ [18].

Las medidas de concordancia más conocidas son τ *de Kendall* y ρ *de spearman* y son definidas en términos de las siguientes probabilidades

- τ-Kendall

$$\tau_{Y_1,Y_2} = \mathbb{P}\left[(y_{1i} - y_{1j})(y_{2i} - y_{2j}) > 0\right] - \mathbb{P}\left[(y_{1i} - y_{1j})(y_{2i} - y_{2j}) < 0\right]$$

- ρ-Spearman

$$\tau_{Y_1,Y_2} = 3\mathbb{P}\left[(y_{1i} - y_{1j})(y_{2i} - y_{2j}) > 0\right] - \mathbb{P}\left[(y_{1i} - y_{1j})(y_{2i} - y_{2j}) < 0\right]$$

Estás medidas son muy utilizadas cuando se trabaja con cópulas, ya que a diferencia de la medida de dependencia de Pearson las medidas de concordancia anteriores son invariantes con respecto a transformaciones que son estrictamente crecientes. Además, las medidas anteriores se pueden expresarse en términos de cópulas [10]

- *τ de Kendall*

$$\tau_{Y_1Y_2} = 4\int\int_{I^2} C(u_1,u_2)dC(u_1,u_2) - 1 = 4E[C(U_1,U_2)] - 1.$$

- *ρ de Spearman*

$$\rho_{Y_1,Y_2} = 12\int\int_{I^2} u_1u_2dC(u_1,u_2) = 12E[U_1,U_2] - 3.$$

En las ecuaciones anteriores se observa que τ y ρ están completamente determinadas por la cópula y no están relacionadas con las distribuciones marginales de Y_1 y Y_2. Es decir, la medida de dependencia entre las variables no depende del comportamiento marginal.

Finalmente, uno de los resultados más útiles en cópulas es el teorema de las cotas de Fréchet, el cual nos indica que para cualquier cópula C que represente un modelo de dependencia ésta está acotada; es decir,

$$W(u) = \text{máx}\{u_1 + u_2 - 1, 0\} \geq C(u) \geq \text{mín}\{u_1, u_2\} = M(u),$$

para todo $u \in [0,1]^2$. Las funciones W y M son conocidas como cota inferior y superior de Fréchet, respectivamente. En 1951, Fréchet demostró que las cotas también son cópulas. Además, si la función conjunta de dos variables está caracterizada por la cota superior de Fréchet; es decir, por M, entonces indica una situación de dependencia perfecta positiva, y se conoce como *comonotonicidad*. Por el contrario, si la cópula utilizada es la cota inferior de Fréchet W, entonces

se dice que hay *contramonotocidad*.

Las cotas de Frechet así como las medidas de concordancia τ de Kendall y ρ de Spearman las podemos ocupar para evaluar la elección de la cópula que se ocupará para construir la función conjunta.

2.1.1. Cópula gaussiana

Actualmente existe una cantidad considerable de familia de cópulas bien definidas; sin embargo, no son equivalentes en términos del tipo de dependencia estocástica que representan o el grado de dependencia que pueden capturar. Por lo tanto, un problema importante en la literatura de cópulas es la selección de la familia de cópula adecuada para construir una distribución bivariada particular. Aunque se puede usar una gran variedad de familia de cópulas para modelar la dependencia, el trabajo de tesis se enfoca a la familia de cópula gaussiana.

El uso de esta cópula bivariada gaussiana es muy atractivo puesto que tiene la capacidad de capturar el rango completo de dependencia, ya que incluye tanto las cotas de Fréchet como el modelo de independencia. Además, captura la dependencia de la misma forma en que la distribución bivariada lo hace usando un paramétro de dependencia, con la diferencia de que se calcula para variables aleatorias con cualesquiera distribuciones marginales.

Una *cópula gaussiana* es una cópula bivariada que se extrae de la función de distribución conjunta bivariada normal estándar de la siguiente manera

$$C(u_1, u_2) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \tag{2.4}$$

donde Φ y Φ_2 son las funciones de distribución acumuladas univariada y bivariada de la normal estándar, respectivamente. Además, $u_i = F_i(y_i)$ para $\{i = 1, 2\}$.

Se utiliza el teorema de Sklar para escribir la función de distribución

$$P(Y_1 \leq y_1, Y_2 \leq y_2) = F(y_1, y_2) = \Phi_2(\Phi^{-1}(F_1(y_1)), \phi^{-1}(F_2(y_2))|\Gamma), \qquad (2.5)$$

donde, $\Gamma = \begin{pmatrix} 1 & \iota \\ \iota & 1 \end{pmatrix}$ y ι es la correlación de Pearson entre $\Phi^{-1}(F_1(y_1))$ y $\Phi^{-1}(F_2(y_2))$ [16].

Cuando la cópula que se utiliza para construir la función de distribución es la gaussiana entonces se puede obtener una fórmula para τ de Kendall y ρ de Spearman en términos del coeficiente de correlación ι

$$\tau_\iota = \frac{2}{\pi} arcsen(\iota) \qquad \text{y} \qquad \rho_\iota = \frac{6}{\pi} arcsen(\frac{\iota}{2}),$$

respectivamente. Estas ecuaciones se obtienen de [10].

2.2. Modelo de regresión de cópula mixta

Hasta el momento toda la teoría de cópula se ha enfocado en la construcción de una distribución conjunta cuando las dos variables marginales son continuas. Sin embargo, el interés principal de la tesis es la formulación de la función de distribución mixta; es decir, cuando existe una variable aleatoria continua y una discreta. Aunque el teorema de Sklar ayuda a la construcción de la función, este teorema solo puede garantizar unicidad cuando las marginales son continuas, cuando hay al menos una discreta esto no se cumple, para estos casos la cópula solo se define en el producto cartesiano del domino de las marginales.

Genest y Neslehová [11] advierten sobre las limitaciones y el peligro de usar descuidadamente la aplicación de la cópula para el caso discreto. Una de las consecuencias podría ser que τ de Kendall y ρ de Spearman dependan de las

marginales. Por lo tanto, para construir la función de distribución conjunta se ocupa una variable latente continua como se propone en [7].

Sea Y_1 una variable aleatoria continua con función de distribución F_1 y Y_2 una variable aleatoria discreta con J categorías ordinales o nominales con función de distribución F_2. Para construir la función conjunta $F(y_1, y_2)$ de Y_1 y Y_2, supongamos que existe una variable continua latente no observable Y_2^* con función de distribución F_2^* tal que se puede definir a Y_2 en términos de Y_2^* de la siguiente manera

$$Y_2 = j\mathbb{I}_{[\gamma_{j-1},\gamma_j)}(y_2^*) \quad \text{con } j \in \{1, 2, 3, \ldots, J\}, \tag{2.6}$$

donde $\gamma_0 = -\infty$ y $\gamma_J = \infty$. Es decir, cada categoría de la variable Y_2 queda definida por un intervalo de Y_2^*.

Se construye la función distribución conjunta para Y_1 y Y_2^* utilizando una cópula Gaussiana (2.5)

$$F^*(y_1, y_2^*; \iota) = C(u_1, u_2; \iota) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\Gamma), \tag{2.7}$$

donde $u_1 = F_1(y_1)$, $u_2 = F_2^*(y_2^*)$, $\Phi_2(\cdot, \cdot)$ denotan la función de distribución acumulada bivariada con matriz covariante $\Gamma = \begin{pmatrix} 1 & \iota \\ \iota & 1 \end{pmatrix}$ y ι la correlación de Pearson entre $\Phi^{-1}(u_1), \Phi^{-1}(u_2)$; y $\Phi(\cdot)$ la función de distribución acumulada de una normal estándar.

Por lo tanto, la función de distribución conjunta de Y_1 y Y_2 está dada por

$$\mathbb{P}(Y_1 \leq y_1, Y_2 = y_2) = [F^*(y_1, \gamma_j) - F^*(y_1, \gamma_{j-1})]\, \mathbb{I}_{\{1,2,\ldots,J\}}(y_2), \tag{2.8}$$

donde F^* es la función definida en (2.7). De aquí se obtiene que la función de

densidad es

$$\begin{aligned} f(y_1, y_2) &= \partial \mathbb{P}(Y_1 \leq y_1, Y_2 = y_2)/\partial y_1 \\ &= f_1(y_1)[C'(u_1, u_2) - C'(u_1, u_2^-)]\mathbb{I}_{\{1,2,\ldots,J\}}(y_2), \end{aligned}$$

donde $u_1 = F_1(y_1)$, $u_2 = F_2^*(\gamma_j)$, $u_2^- = F_2^*(\gamma_{j-1})$, $C'(u_1, u_2) = \partial C(u_1, u_2)/\partial u_1$ y f_1 la función de densidad de Y_1.

Si se toma a $q_1 = \Phi^{-1}(u_1)$ y $q_2 = \Phi^{-1}(u_2)$ entonces

$$\begin{aligned} C'(u_1, u_2) &= \frac{\partial}{\partial u_1} C(u_1, u_2) \\ &= \frac{\partial}{\partial u_1} \frac{1}{2\pi\sqrt{|det(\Gamma)|}} \int_{-\infty}^{q_1} \int_{-\infty}^{q_2} \exp\left\{-\frac{1}{2}(x_1, x_2)\Gamma^{-1}(x_1, x_2)^T\right\} dx_1 dx_2 \\ &= \frac{1}{2\pi\sqrt{|det(\Gamma)|}} \int_{-\infty}^{q_2} \exp\left\{-\frac{1}{2}(q_1, x_2)\Gamma^{-1}(q_1, x_2)^T\right\} dx_2 \frac{\partial}{\partial u_1} q_1 \\ &= \frac{1}{2\pi\sqrt{|det(\Gamma)|}} \int_{-\infty}^{q_2} \exp\left\{-\frac{1}{2}(q_1, x_2)\Gamma^{-1}(q_1, x_2)^T\right\} dx_2 \sqrt{2\pi} \exp\left(\frac{1}{2}q_1^2\right) \\ &= \frac{1}{\sqrt{2\pi|1-\iota^2|}} \int_{-\infty}^{q_2} \exp\left\{-\frac{1}{2(1-\iota^2)}(q_1\iota - x_2)^2\right\} dx_2. \end{aligned}$$

Usando el cambio de variable que propusieron en el artículo [6], $x_2 = z\sqrt{1-\iota^2} + q_1\iota$ se obtiene que

$$C'(u_1, u_2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{q_2 - q_1\iota}{\sqrt{1-\iota^2}}} \exp\left(-\frac{1}{2}z^2\right) dz = \Phi\left(\frac{\Phi^{-1}(u_2) - \iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right). \quad (2.9)$$

Existe dos casos que deben ser considerados, cuando $y_2 = 1$, $\Phi(u_2^{-1}) = -\infty$ y se tiene que $C'(u_1, u_2^-; \iota)$ tiende a 0. Por otro lado, cuando $y_2 = J$, $\Phi(u_2) = \infty$ y $C'(u_1, u_2; \iota)$ tiende a 1. Finalmente de la ecuación (2.9) y de los dos casos

particulares anteriores, se reescribe la función conjunta de Y_1, Y_2 como

$$\begin{aligned} f(y_1, y_2) &= f_1(y_1)\mathbb{I}_{\{J\}}(y_2) + f_1(y_1)\Phi\left(\frac{\Phi^{-1}(u_2) - \iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{1,2,\ldots,J-1\}}(y_2) \\ &\quad - f_1(y_1)\Phi\left(\frac{\Phi^{-1}(u_2^-) - \iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{2,3,\ldots,J\}}(y_2). \end{aligned} \tag{2.10}$$

El parámetro ι que representa la correlación de Pearson entre la variable Y_1 y Y_2^* puede ser interpretado como una aproximación de la correlación de Pearson que existe entre la variable continua Y_1 y la discreta Y_2 [7].

La ecuación 2.10 representa la función de densidad de una variable continua y una discreta que está determinada únicamente por una función cópula gaussiana y sus marginales. Además, el comportamiento de cada una de las marginales no dependen de la dependencia de las variables, la cual es medida de manera separada por la cópula, lo que nos permite anexar al modelo componentes lineales para agregar variables explicativas.

Finalmente, para determinar el modelo se supone que la variable continua que representa los tiempos de supervivencia sigue una distribución Weibull, con parámetro de escala λ y parámetro de forma κ, tal que

$$h(y_1) = \exp\left(\sum_{i=0}^{m_1}\beta_i x_i\right) h_0(y_1), \tag{2.11}$$

en donde x_i son las variables explicativas, β_i los parámetros lineales. La función $h_0(y_1)$ denota la función de riesgo base que está dado por: $h_0(y_1) = f_0(y_1)/(1 - F_0(y_1))$, con $f_0(y_1)$ y $F_0(y_1)$ las funciones de densidad y distribución, respectivamente. Es decir, se anexa un modelo de regresión de Cox para la parte continua.

Para la variable discreta Y_2 que representa al tipo de tratamiento, se hace el supuesto que sigue una distribución multinomial con probabilidades

$\mathbb{P} = [p_1, p_2, \ldots, p_J]$, tal que

$$p_j = \exp\left(\sum_{i=0}^{m_2} \alpha_{ij} z_i\right), \tag{2.12}$$

donde z_i son las variables explicativas y α_{ij} son los parámetros lineales correspondientes a la probabilidad j-ésima del paciente i-ésima. En otras palabras, agregamos un modelo de regresión logístico multinomial para la parte discreta.

Finalmente, se anexa un componente lineal a la estructura de dependencia

$$\iota \;=\; tanh\left(\sum_{i=0}^{m_3} \zeta_i w_i\right); \tag{2.13}$$

es decir, se utiliza la transformada de Fisher para agregar un modelo lineal para la dependencia.

El modelo de regresión para los tiempos de supervivencia y los tratamientos queda determinado por la función de densidad (2.10) y las ecuaciones (2.11), (2.12) y (2.13). Otro punto importante a tomar en cuenta es, por la manera de obtener la función de densidad es fácil obtener la función condicional, lo cual es un requisito importante para poder evaluar los tiempos de supervivencia dado el tratamiento.

En la Figura 2.1 se gráfica la función condicional dada en (2.10), con Y_1 una variable Weibull con parámetro de escala $\lambda = 1$ y parámetro de forma $\kappa = 1.5$, para cuatro categorías, donde la probabilidad para cada una de las categorías es 0.25. En cada una de las gráficas se muestra el efecto que tiene la dependencia entre las variables sobre los tiempos de supervivencia dependiendo la categoría. Por ejemplo, en la categoría 1, al parecer una dependencia positiva aumenta el valor de la función de densidad los primeros tiempos y la dependencia negativa

realiza el efecto contrario; es decir, baja el valor de la función conjunta en los primeros tiempos, si se compara con el modelo independiente.

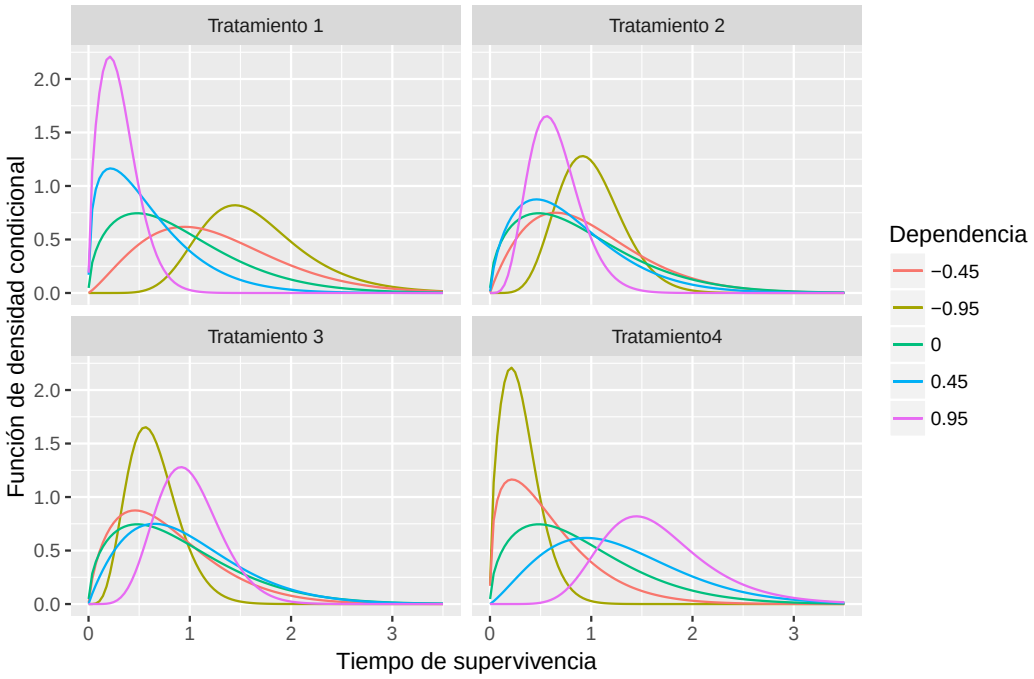

Figura 2.1: Función de densidad condicional $f_{Y_1|Y_2}$ para la marginal Y_1 Weibull con parámetros $\lambda = 1$, $k = 1.5$ y Y_2 multinomial equiprobable.

2.2.1. Modelo de regresión logística multinomial

Los modelos de regresión logística multinomial son modelos estadísticos que tienen como objetivo dar a conocer la relación que existe entre una variable respuesta de tipo cualitativo con más de dos categorías y variables explicativas las cuales pueden ser cualitativas o cuantitativas. El modelo de regresión logística multinomial se toma como una extensión multivariante de la regresión logística binaria clásica debido a que para construir el modelo logit para una respuesta

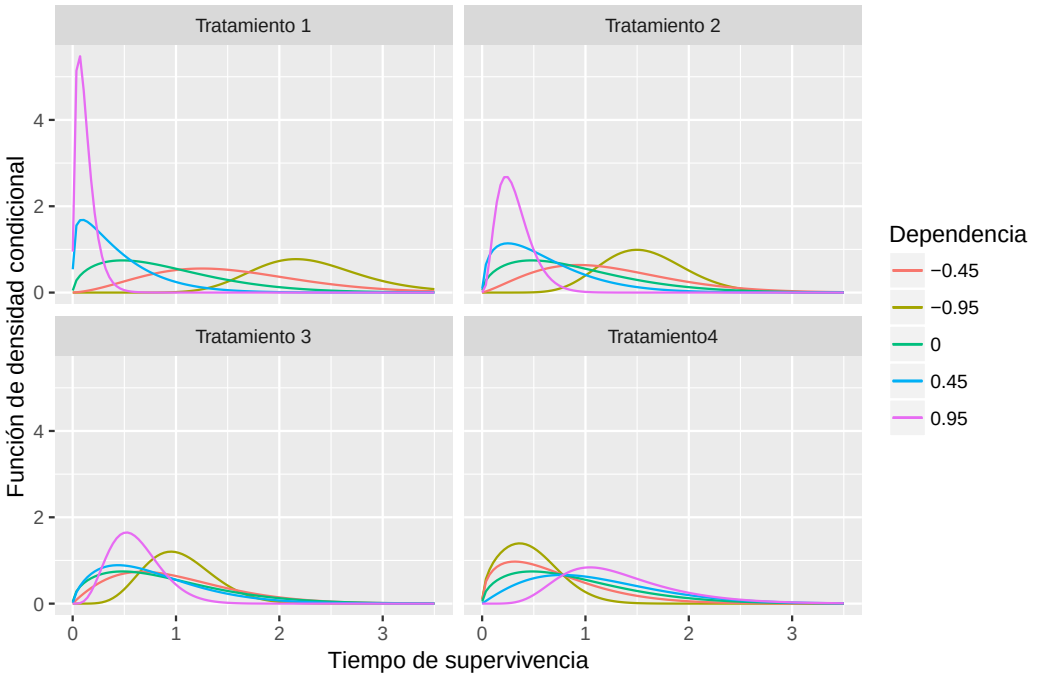

Figura 2.2: Función de densidad condicional $f_{Y_1|Y_2}$ para la marginal Y_1 Weibull con parámetros $\lambda = 1$, $k = 1.5$ y Y_2 multinomial con probabilidades $\mathbb{P} = \{0.05, 0.15, 0.30, 0.50\}$.

multinomial con J categorías se consideran J-1 transformaciones logit básicos.

Si $Y = (y_1, y_2, \ldots, y_J)$ es una variable aleatoria que tiene una función de *distribución multinomial* con parámetros $n > 0$ y $\mathbb{P} = \{p_1, p_2, \ldots, p_J\}$, entonces su función de masa de probabilidad es

$$\begin{aligned} f(y_1, y_2, \ldots, y_J; n, p_1, p_2, \ldots, p_J) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_J = y_J) \\ &= \frac{n!}{y_1! y_2! \cdots y_J!} p_1^{y_1} p_2^{y_2} \cdots p_J^{y_J}, \end{aligned}$$

donde $\sum_{i=1}^{J} p_i = 1$, $\sum_{i=1}^{J} x_i = n$ y $J > 1$.

Para el caso particular de $J = 2$ se tiene que $p_2 = 1 - p_1$ y $y_2 = n - y_1$; por lo tanto, la función anterior dirige a la función de distribución binomial con parámetros n y p_1. En este caso, el modelo logístico multinomial es el modelo de regresión logística binomial (o clásico) el cual está descrito por la siguiente probabilidad

$$\mathbb{P}(Y = 1; Z) = \frac{\exp(a_0 + \sum_{s=1}^{m} a_s z_s)}{1 + \exp\left(a_0 + \sum_{s=1}^{m} a_s z_s\right)},$$

en donde se busca la probabilidad de que Y tome el valor 1, en presencia del vector de variables explicativas $z = (z_1, z_2, \ldots, z_m)$. Para obtener el modelo de regresión logístico clásico se ocupa la transformación logit

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{p_1}{p_2}\right) = a_0 + \sum_{s=1}^{m} a_s z_s.$$

El modelo de regresión logístico multinomial se modela eligiendo una categoría de referencia (*categoría base*) y posteriormente se definen varias funciones logit como la anterior, uno para cada una de las categorías restantes con respecto a la categoría base. Es decir, para construir el modelo de regresión para una respuesta multinomial se necesitan considerar $J-1$ transformaciones logit básicas. Si se toma la categoría 1 como la categoría base, entonces la transformación logit generalizada se define como

$$L_j = \ln\left(\frac{p_j}{p_1}\right), \quad \text{para todo j } \in \{2, 3, \ldots, J\}.$$

Por lo tanto, el componente lineal para cada una de las transformaciones logit generalizadas con m variables explicativas queda expresado como

$$\ln\left(\frac{p_j}{p_1}\right) = \sum_{s=0}^{m} a_{sj} z_s = z'\alpha_j, \text{ para todo } j \in \{2, 3, \ldots, J\},$$

en donde, $z = (z_0, z_1, \ldots, z_m)$ es el vector de variables explicativas, con $z_0 = 1$, y $\alpha_j = (a_{0j}, a_{1j}, \ldots, a_{mj})$ el vector de parámetros asociados a la categoría j.

Para encontrar las probabilidades de cada categoría éstas se expresan en términos de la probabilidad base como

$$p_j = \exp\left(\sum_{s=0}^{m} a_{sj} z_s\right) p_1, \text{ para } j = \{1, 2, \ldots, J\}$$

y se ocupa la propiedad de la definición de la distribución multinomial, $\sum_{i=1}^{J} p_i = 1$. Así, se obtiene un sistema de J ecuaciones con J variables, resolviéndolo se obtienen las probabilidades

$$\begin{aligned} p_1 &= \frac{1}{1 + \sum_{j=2}^{J} \left[\exp\left(\sum_{s=0}^{m} a_{sj} z_s\right)\right]}, \\ p_j &= \frac{\exp\left(\sum_{s=0}^{m} a_{sj} z_s\right)}{1 + \sum_{j=2}^{J} \left[\exp\left(\sum_{s=0}^{m} a_{sj} z_s\right)\right]}, \quad \text{para todo } j \in \{2, 3, \ldots, J\}. \end{aligned}$$

El objetivo de ocupar un modelo de regresión logística multinomial es estimar las probabilidades p_j de cada categoría j para cada individuo, tomando en cuenta el conjunto de variables explicativas. Por lo tanto, este modelo de regresión se utiliza para encontrar las probabilidades de cada tratamiento para cada individuo dependiendo las características clínico patológicas de cada individuo.

2.3. Función de verosimilitud

La estimación de los parámetros se realiza mediante el método de máxima verosimilitud. Para obtener la función de verosimilitud, los datos se dividen en dos partes, las observaciones no censuradas y las censuradas, esto debido a que de los datos censurados no se conoce con exactitud el tiempo de falla y sería incorrecto tomar la función de densidad de los datos censurados para construir la función de verosimilitud.

Supongamos que $\{y_{1i}, y_{2i}, x_i, z_i, w_i\}$ con $i \in \{1, 2, ..., n\}$ denotan los datos observacionales, en donde, x_i, z_i y w_i son vectores de variables explicativas de las variables Y_{1i}, Y_{2i} y de la estructura de dependencia, respectivamente. Se denota a θ como el vector que contiene todos los parámetros de regresión, $\theta = (\kappa, \lambda, \beta, \alpha, \zeta)$, con κ y λ los parámetros de la función Weibull, $\beta = [\beta_1, \beta_2, \ldots, \beta_{m_1}]$ los parámetros lineales del modelo de Cox, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_{jm_2}]$ los parámetros lineas del modelo multinomial y finalmente, $\zeta = [\zeta_1, \zeta_2, \ldots, \zeta_{m_3}]$ son los parámetros lineales de modelo de dependencia. Además, introducimos una variable indicador para la censura c_i tal que

$$c_i = \begin{cases} 1, & si \quad y_{1i} \text{ es censurado;} \\ 0, & si \quad y_{1i} \text{ no es censurado.} \end{cases}$$

Sean $(y_{11}, y_{21}), (y_{12}, y_{22}), \ldots, (y_{1r}, y_{2r})$, con $r \leqslant n$, observaciones no censuradas, entonces la contribución de estos datos no censurados a la función de verosimilitud, está dada por el producto de las funciones de densidad de cada individuo

$$\prod_{i=1}^{r} f(y_{1i}, y_{2i}).$$

Ahora, sean $(y_{1(r+1)}, y_{2(r+1)}, (y_{1(r+2)}, y_{2(r+2)}), \ldots, (y_{1n}, y_{2n})$ observaciones censuradas. Como el tipo de censura que se presenta en los datos es por la derecha, entonces la información que se obtiene de estos datos está dada por la probabilidad de que el individuo viva más allá del tiempo y_{1i}; es decir,

$$\prod_{i=r+1}^{n} \int_{y_{1i}}^{\infty} f(t, y_{2i}) dt.$$

Por lo tanto, la función de verosimilitud está dada por

$$L(\theta) = \prod_{i=1}^{n} [f(y_{1i}, y_{2i})]^{c_i} \left[\int_{y_{1i}}^{\infty} f(t, y_{2i}) dt \right]^{(1-c_i)}$$

Luego, la función log-verosimilitud es

$$
\begin{aligned}
f(y_1,y_2) &= \sum_{i=1}^{n} c_i log(f_1(y_1)[\mathbb{I}_{\{J\}}(y_2) \\
&\quad +\Phi\left(\frac{\Phi^{-1}(u_2)-\iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{1,2,\ldots,J-1\}}(y_2) \\
&\quad -\Phi\left(\frac{\Phi^{-1}(u_2^-)-\iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{2,3,\ldots,J\}}(y_2)]) \\
&\quad +\sum_{i=1}^{n}(1-c_i)log(\int_{y_{1i}}^{\infty} f_1(t)[\mathbb{I}_{\{J\}}(y_2) \\
&\quad +\Phi\left(\frac{\Phi^{-1}(u_2)-\iota\Phi^{-1}(t)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{1,2,\ldots,J-1\}}(y_2) \\
&\quad -\Phi\left(\frac{\Phi^{-1}(u_2^-)-\iota\Phi^{-1}(t)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{2,3,\ldots,J\}}(y_2)dt].
\end{aligned}
$$

Se realiza un cambio de variable $w = F_1(t)$

$$\int_{y_{1i}}^{\infty} f_1(t)C'(F_1(t),F_2(y_{2i}))dt = \int_{F_1(y_{1i})}^{1} C'(w,F_2(y_{2i}))dw.$$

Así, finalmente tenemos que la función de log-verosimilitud es

$$
\begin{aligned}
f(y_1,y_2) &= \sum_{i=1}^{n} c_i log(f_1(y_1)) + \sum_{i=1}^{n} c_i log[\mathbb{I}_{\{J\}}(y_2) \\
&\quad +\Phi\left(\frac{\Phi^{-1}(u_2)-\iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{1,2,\ldots,J-1\}}(y_2) \\
&\quad -\Phi\left(\frac{\Phi^{-1}(u_2^-)-\iota\Phi^{-1}(u_1)}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{2,3,\ldots,J\}}(y_2)]) \\
&\quad +\sum_{i=1}^{n}(1-c_i)log(\int_{F(y_{1i})}^{1} [\mathbb{I}_{\{J\}}(y_2) \\
&\quad +\Phi\left(\frac{\Phi^{-1}(u_2)-\iota\Phi^{-1}(F(w))}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{1,2,\ldots,J-1\}}(y_2) \\
&\quad -\Phi\left(\frac{\Phi^{-1}(u_2^-)-\iota\Phi^{-1}(F(w))}{\sqrt{1-\iota^2}}\right)\mathbb{I}_{\{2,3,\ldots,J\}}(y_2)dw]. \qquad (2.14)
\end{aligned}
$$

2.4. Simulación para el modelo nulo

Con la finalidad de evaluar o validar el rendimiento de la estimación de máxima verosimilitud con la función de verosimilitud dado por la ecuación (2.14), en esta sección se realiza un estudio de simulación para un modelo nulo.

Sea Y_1 una variable continua con distribución Weibull con parámetros λ, κ y Y_2^* una variable continua con distribución normal estándar. Supongamos que $F^*(y_1, y_2^*)$ es la función de distribución conjunta de Y_1 y Y_2^* dada por la ecuación (2.10) donde ι el es parámetro de dependencia de Y_1 y Y_2^*, se genera una muestra de 15000 datos pseudo-aleatorios utilizando el paquete *copula* del software estadístico R de las variables Y_1, Y_2^* a través de la función *mvdc*.

Sea la variable discreta Y_2 con distribución multinomial de la muestra a tráves de los cuantiles de la normal estandár; es decir, cada categoría de la variable Y_2 queda determinada por cada cuantil de la normal estándar. La función de distribución conjunta de Y_1 y Y_2 está dada por la función (2.10), donde p_j es la probabilidad de la categoría j para $j \in \{1, 2, \ldots, J\}$ y ι se toma como una aproximación de la dependencia que existe entre Y_1 y Y_2.
Para realizar la estimación se ocupa el método de máxima verosimilitud en donde se maximiza la función (2.14) mediante el método de Newton [ver Apéndice B]. Los intervalos de confianza se obtienen a través de la Matriz Hessiana del método de Newton, además, se ocupa un grado de tolerancia $10 \times e^4$ y número máximo de iteraciones de 100. En el Tabla 2.1, se muestra el resultado de la primera simulación, para esta simulación el programa se ejecutó con un tiempo de 407.00 segundos y realizó 16 iteraciones en total. El Tabla 2.2 muestra los resultados de la segunda simulación, para este caso el tiempo de ejecución fue de 331.85 segundos con un total de 12 iteraciones. En ambos casos se ocupa un punto inicial $\theta = (0, 0.5, 1, 0, 0, 0)$.

Ambas simulaciones muestran que los estimadores están muy cerca a los valores originales de los parámetros reales por lo que se puede concluir que al menos para el modelo nulo la función de máxima verosimilitud funciona bien. Además, se observa que todos los parámetros se encuentran dentro de los intervalos de confianza.

Parámetro	Real	Estimación	Intervalo de Confianza	Error Estándar
r	-0.70	-0.699855	(-0.710812,-0.688550)	0.011133
κ	1.50	1.508250	(1.482353, 1.534600)	1.008876
λ	2.00	2.017958	(1.988750, 2.047595)	1.007466
p_1	0.25	0.243277		
p_2	0.25	0.257827		
p_3	0.25	0.250121		
p_4	0.25	0.248773		

Tabla 2.1: Estimadores de la simulación 1.

Parámetro	Real	Estimación	Intervalo de Confianza	Error Estándar
r	0.50	0.497375	(0.4807480, 0.5136448)	0.011150
κ	0.50	0.499368	(0.4906686, 0.5082222)	1.009007
λ	3.00	3.047376	(2.9119622, 3.1890874)	1.023462
p_1	0.45	0.449081		
p_2	0.30	0.305792		
p_3	0.20	0.192505		
p_4	0.05	0.05262172		

Tabla 2.2: Estimadores de la simulación 2.

2.5. Datos faltantes

Es muy frecuente que en estudios de investigación encontremos datos faltantes, espacios vacíos en la base de datos que deseamos analizar y esto afecta a la validez de la inferencia de los datos y la deducción de la naturaleza del problema. Estos datos son conocidos comúnmente como *missing* y las causas por las que existen pueden ser diversas, el paciente se niega a responder algunas preguntas, el investigador se equivoca el codificar la respuesta, las perdidas durante un seguimiento, entre otras causas.

Supongamos que tenemos una muestra $Y = (Y_{i1}, Y_{i2}, \ldots, Y_{im})$ de tamaño n; es decir, $i = 1, 2, \ldots, n$ y con m variables de recolección para cada $i-$ésimo individuo. Por lo tanto, Y representa una matriz de tamaño $n \times m$ que contiene la información del estudio. Se denota como $Y_{i,O}$ al subconjunto de datos que son observados de Y_i y $Y_{i,F}$ como el subconjunto de datos faltantes de manera que $Y_{iO} \bigcup Y_{iF} = Y_i$ y $Y_{iO} \bigcap Y_{iF} = \varnothing$ para todo $i = \{1, 2, \ldots, n\}$. Además se define la matriz indicadora como

$$M_{ij} = \begin{cases} 1, & si \quad Y_{ij} \text{ es observado;} \\ \\ 0, & si \quad Y_{ij} \text{ es faltante.} \end{cases}$$

para cada individuo $i = \{1, 2, \ldots, n\}$ y cada variable $j = \{1, 2, \ldots, m\}$.

2.5.1. Tipos de datos faltantes

Hay diversos mecanismos que producen la pérdida de datos, el sistema de clasificación más usado es el que propuso Rubin en 1976, quién clasificó en tres tipos distintos a los datos faltantes, la cual se presenta a continuación.

a) **Missing completely at random (MCAR)**: Se dice que un dato es de tipo MCAR si la probabilidad de la respuesta de una variable sea un dato

perdido es independiente tanto del valor de la variable como del valor de otras variables que estén en el conjunto de datos. En términos de probabilidades se expresa como:

$$\mathbb{P}(M_i|Y_i) = \mathbb{P}(M_i)$$

Es decir, la falta de información no depende de ninguna variable presente en la base de datos. En otras palabras y como el nombre lo indica, la pérdida de los valores ocurren aleatoriamente.

b) **Missing At Random (MAR)**: Un dato es de tipo MAR si la probabilidad de que la respuesta de una variable sea un dato perdido depende de los valores de otras variables en el conjunto de datos. En térmonos de probabilidad se escribe como:

$$\mathbb{P}(M_i|Y_i) = \mathbb{P}(M_i|Y_{i,O}).$$

Es decir, la ausencia de los datos perdidos está asociada a los datos observados. En este caso, los datos no son completamente aleatorios y estos se pueden producir mediante las observaciones donde hay información completa.

c) **Missing Not At Random (MNAR)**: Un dato es de tipo MNAR si la probabilidad de la ausencia del dato depende de las otras variables del conjunto conjunto de datos pero también depende de la variable misma. En probabilidad se denota como:

$$\mathbb{P}(M_i|Y_i) \neq \mathbb{P}(M_i|Y_{i,O})$$

Es decir, en este caso la ausencia de datos perdidos no esta asociada únicamente a los datos observados. Podemos decir que un dato faltante es de tipo MNAR si no es MCAR y MAR.

2.5.2. Imputación múltiple

Cuando existe ausencia de datos en un estudio hay dos caminos que se pueden tomar para realizar el análisis de la base, uno sería ignorarlo y otro "inventarlo".

El primer método se trata de ignorar todos los sujetos que tienen valores faltantes en una o más variables y realizar el análisis de datos completos. Aunque este camino es muy práctico y el más fácil, debemos tomar en cuenta que perdemos información al eliminar a los participantes con datos incompletos sobre todo cuando son muchos los casos ignorados y esto nos puede llevar a algún sesgo; por lo tanto, no es muy aconsejable.

La segunda opción sería inventar los datos, aunque se escuche poco fiable este camino es más aconsejable que el método anterior, el objetivo del método es rellenar todos los espacios vacíos en la base de datos ocupando la información que tenemos y creando así una base de datos completos, la palabra correcta para este método es *imputar*. Hay diversas técnicas de imputación, la imputación simple y la imputación múltiple.

Entre las técnicas de imputación simple podemos encontrar la imputación de la media, por regresión o por regresión estocástica. Cuando la cantidad de datos faltantes no es muy alta el método de imputación simple puede dar buenos resultados; sin embargo, los métodos de imputación múltiple suelen ser más adecuados que los de imputación simple cuando existen una gran proporción de datos faltantes.

Este método fue propuesto por primera vez por Rubin(1978) y a diferencia que las técnicas de imputación simple en donde a cada dato desconocido solo se rellena un único, valor en imputación múltiple se le asigna más un valor para cada

dato faltante; es decir, se generan $M > 1$ valores para cada dato faltante creando así M bases de datos completos. Esta metodología consta de varias etapas las cuales se explican a continuación.

Etapas:

a) Imputación: Se generan M valores para cada dato perdido, creando M base de datos completas.

b) Análisis: Se analizan las M bases creadas por separado para obtener M estimadores.

c) Combinación: Se combinan las estimadores y las varianzas para obtener un solo conjunto de estimaciones y su respectiva varianza.

Esta metodología es intuitiva y fácil de entender, en donde se asume que los datos faltantes son de tipo MAR; esto es, que los datos faltantes de una variable X dependen de otras variables del estudio pero no de ella. Además, para aplicar el método de imputación múltiple se requiere de un modelo estadístico para generar los datos imputados apropiados; es decir, la correlación entre la variable a imputar y las covariables que se utilizan para modelar los valores que se utilizaran como sustitutos.

Cuando se utiliza imputación múltiple se utilizan métodos vía Monte Carlo, donde cada estado corresponde a la colección de los datos imputados. En algunas investigaciones se relacionan los métodos de imputación múltiple con cadenas de Markov de Monte Carlo, como es el algoritmo MICE (*Multivariate Imputation by Chained Equation*) imputación múltiple por ecuaciones en cadenas, este método utiliza simulación paramétrica generando muestras aleatorias a partir de métodos bayesianos, generando M imputaciones independientes las cuales se utilizan para

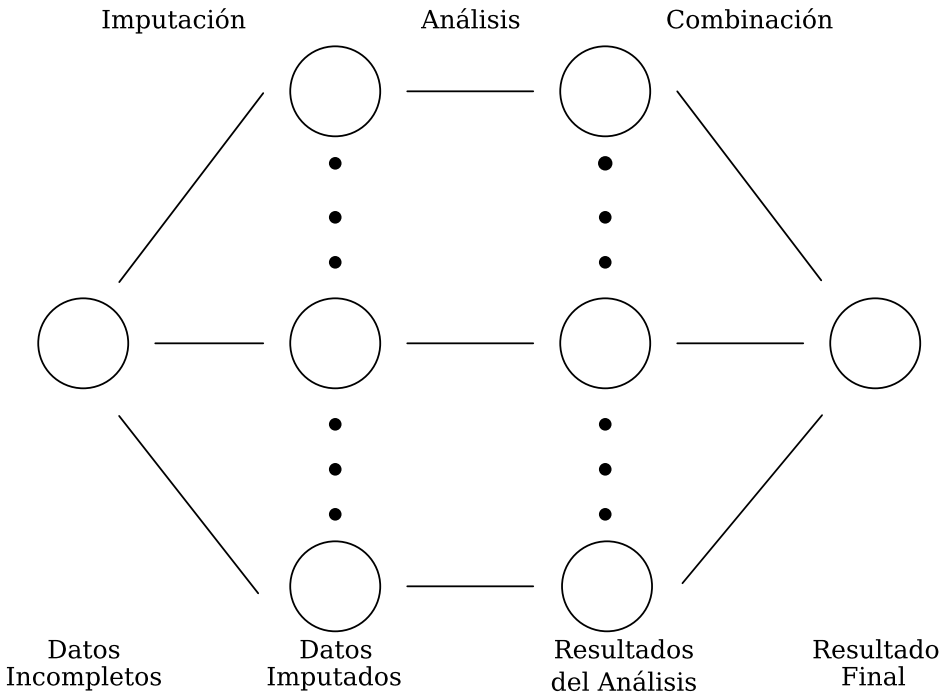

Figura 2.3: Proceso de Imputación Multiple.

la etapa de inferencia.

El número necesario de imputaciones M está entre 5 y 10, aunque en varios estudios sugieren usar un número mayor para asegurar la convergencia, de acuerdo a Rubin, la eficiencia relativa de M imputaciones es aproximado a $\lambda = (1+\lambda/M)$, donde λ representa la tasa de registros sin información, esto equivale a que si tengo una base de datos con 50 % de datos faltantes, un estimador generado a partir de $M = 5$ imputaciones tiene una desviación estándar que es solo 5 % mayor a otro generado por otra simulación basada en $M = \infty$. Por lo tanto, se señala que para estimaciones usualmente altas solo se requieren entre 5 y 10 imputaciones.

Reglas de Rubin

Para realizar la tercera etapa de imputación múltiple, la combinación, se ocupan las reglas de Rubin. Sea M número de imputaciones entonces el estimador final se define como el promedio de los estimadores individuales

$$\bar{\beta} = \frac{1}{M}\sum_{i=1}^{M}\beta_i. \tag{2.15}$$

Varianza dentro de las imputaciones se define como

$$\bar{V} = \frac{1}{M}\sum_{i=1}^{M}V_i \tag{2.16}$$

y la varianza entre imputaciones es

$$B = \frac{1}{M-1}\sum_{i=1}^{M}(\beta_i - \bar{\beta}) - (\beta_i - \bar{\beta})^T. \tag{2.17}$$

Finalmente, de las ecuaciones (2.16), (2.17) se obtiene la varianza total

$$var(\bar{\beta}) = \bar{V} + (1 + M^{-1})B. \tag{2.18}$$

Para el cálculo de los intervalos de confianza se utiliza

$$\bar{\beta} \pm t_{\upsilon}\sqrt{Var(\bar{\beta})} \tag{2.19}$$

donde t_{υ} es el vector que contiene los porcentiles de la distribución $t - student$ central y υ el vector que contiene los grados de libertad que se estiman mediante

$$\upsilon = (M-1)\left\{1 + \frac{\bar{V}}{(1+M^{-1})\,B}\right\}.$$

Capítulo 3

Aplicación a los datos de cáncer de mama

De acuerdo con la página oficial de la organización mundial de la salud (OMS) el cáncer es una de las principales causas de muerte en el mundo, en donde se reportan 12 millones de casos nuevos y 8.8 millones de muertes relacionadas con él (esto es, 1 muerte de cada 6 en el mundo es por cáncer). El cáncer tienen un gran impacto en el mundo, no solo es en el sector salud, sino también en lo social y en lo económico. Su costo económico anual total en el 2010 se estimó aproximadamente en 1.16 billones de dólares. Las cifras son impactantes y las investigaciones acerca del cáncer también.

Entre los tipos de cáncer más comunes que afectan a las mujeres está el cáncer de mama, en Estados Unidos ocupa el segundo puesto después del cáncer de piel (*National Cancer Institute*), mientras que en México es la primera causa de movilidad hospitalaria por neoplasias (Instituto Nacional de Estadística y Geografía). En esta sección se ilustra el método descrito en el capítulo 2 con una cohorte de mujeres diagnosticadas con cáncer de mama con el objetivo de modelar los tiempos de supervivencia para cuatro posibles tratamientos.

3.1. Cáncer de mama

Normalmente, las células del cuerpo crecen, se dividen, se reproducen y mueren, este ciclo celular tiene varios sistemas de regulación y control, los cuales se necesitan para mantener la salud; sin embargo, en ocasiones el proceso se descontrola y las células se dividen cuando no es necesario, o no mueren cuando deberían. Las células anormales que crecen sin control comienzan a formar masas las cuales se conocen como tumores y se clasifican en benignos o malignos, los primeros no son cancerosos. Por el contrario, los tumores malignos son cancerosos, sus células pueden invadir y destruir tejidos a su alrededor.

El cáncer de mama es un tumor maligno que se forma en los tejidos del seno, el cual tiene la posibilidad de que sus células ingresen en los vasos linfáticos, se propaguen a los ganglios linfáticos ocasionando una mayor probabilidad de que las células se desplacen por todo el sistema linfático y lleguen a otras partes del cuerpo adheriéndose a otros tejidos y formando nuevos tumores (*metástasis*). Este tipo de cáncer afecta principalmente a las mujeres, aunque los hombres también lo pueden padecer.

Tipos de cáncer

El seno está estructurado por 15 o 20 secciones llamados *lóbulos*, éstas a su vez formadas por subsecciones conocidos como *lobulillos* en los cuales se encuentran grupos de glándulas diminutas denominadas bulbos que se encargan de producir leche. Los pezones se conectan a los lóbulos a través de tubos delgados llamados *conductos*. En el espacio entre los lóbulos y los conductos, hay tejido graso, tejido fibroso, nervios, vasos sanguíneos y vasos linfácticos (como se muestra en la Figura 3.1).

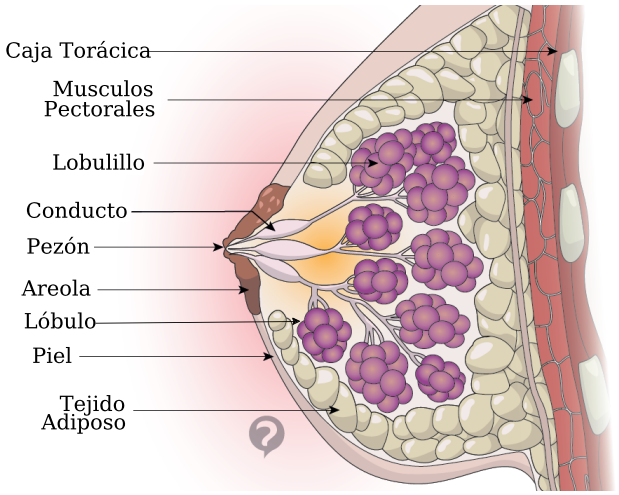

Figura 3.1: Estructura de la mama (*https://salud.ccm.net*).

El cáncer puede empezar en distintas partes de la mama y dependiendo del lugar en el que se origine se clasifican en los distintos tipos de cáncer.

a) **Carcinoma ductal**: El carcinoma ductal es el cáncer de mama más común pues el 80 % de los casos de cáncer diagnosticados se tratan de carcinoma ductal. Este tipo de cáncer comienza en el conducto de la mama, penetra la pared del conducto y empieza a invadir los tejidos de la mama.

b) **Carcinoma lobular**: El carcinoma lobular, es el segundo cáncer más común. Al rededor del 10 % son cáncer lobular. Este cáncer empieza en las glándulas productoras de leche y se propaga a otros tejidos de la mama.

Existen algunos tipos de cáncer especiales (según el tipo de agrupamiento que tienen las células), algunos ejemplos: carcinoma quístico adenoide, carcinoma medular, carcinoma mucinoso, carcinoma papilar, carcinoma tubular, cáncer de seno inflamatorio, enfermedad de Paget del pezón.

Grado del tumor

Para la clasificación del grado del tumor se utiliza el sistema *Nottingham Histologic Score* el cual ocupa tres faces para determinar el grado del tumor, la primera es la "diferenciación"que se refiere a qué tan bien las células cancerosas recrean células normales, el "pleomorfismo"que es una evaluación del tamaño y la forma de las células normales, y finalmente la cantidad de células tumorales que se están dividiendo, también conocida como la "actividad mitótica". A cada una de las características anteriores se le asigna una puntuación de 1 a 3 que luego se suma para obtener una puntuación final de 3 a 9, y que determina el grado del tumor:

a) **Grado 1** (Bien diferenciado) con una puntuación de 3 a 5.

b) **Grado 2** (Moderadamente diferenciado) con una puntuación de 6 a 7.

c) **Grado 3** (Pobremente diferenciado o no diferenciado) con una puntuación de 8 a 9.

Tratamiento

En la actualidad existen diversas opciones de tratamiento para combatir el cáncer de mama (cirugía, radioterapia, quimioterapia, hormonoterapia, terapia biológica), la recomendación del uso de los tratamientos depende del estado clínico de cada paciente los cuales a menudo reciben más de un tipo de tratamiento.

Cirugía

La cirugía es el tratamiento más común para tratar el cáncer de mama. Existen distintos tipos de cirugía:

a) **Cirugía con conservación del seno**: También conocida como cirugía preservadora del seno, tumorectomía, mastectomía segmentaría o parcial, es una cirugía que solo elimina la parte de la mama donde se encuentra el tumor canceroso y un poco de tejido normal de su alrededor. Esta cirugía es ilustrada en la Figura 3.2.

b) **Mastectomía**. En este tipo de cirugía se extirpa toda la mama o todo el tejido mamario posible, junto con otros tejidos cercanos si es necesario. Existen diferentes tipos de mastectomía, pero entre los más comunes para combatir el cáncer de mama es la mastectomía simple, este procedimiento se trata de extirpar todo el seno, incluyendo el pezón, pero no los ganglios linfáticos de la axila. Otro tipo de mastectomía es la mastectomía radical en la cual aparte de extirpar todo el tejido mamario, se extirpa los ganglios linfáticos y los músculos debajo del seno. Se ilustra en la Figura 3.3.

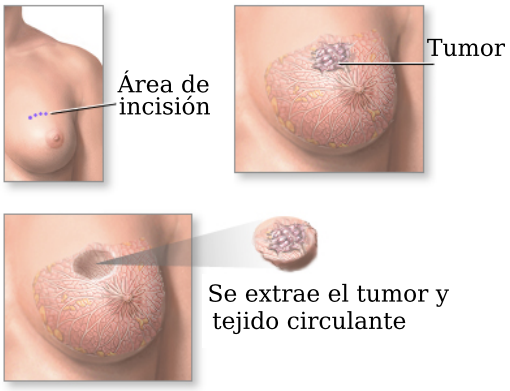

Figura 3.2: Cirugía con conservación de seno.

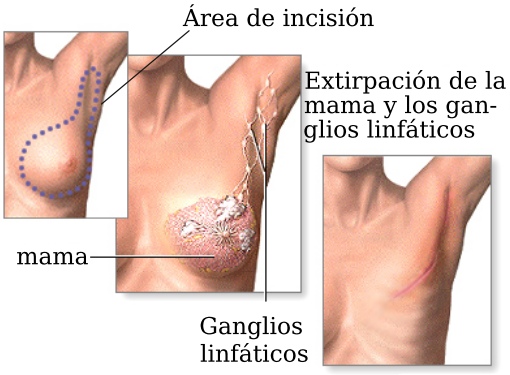

Figura 3.3: Mastectomía.

Radioterapia

Usualmente el tratamiento de una paciente se complementa con radioterapia antes o después de alguna cirugía. Ya sea para disminuir el tamaño de un tumor canceroso (antes) o para minimizar el riesgo de que el cáncer reaparezca (después). En esta terapia se usan rayos de altas frecuencias para eliminar células cancerosas a través de una radiación que proviene de una máquina, se realiza en la clínicas u hospitales y su aplicación dura pocos minutos.

3.2. Base de datos

La base de datos que se estudia corresponde a un seguimiento de 20 años del programa *Surveillance, Epidemiology, and End Results del National Cancer Institute (SEER)* de Estados Unidos, cuyos registros corresponden a información de mujeres diagnosticadas con cáncer de mama y que residen en áreas geográficas determinadas (Alaska, California, Connecticut, Georgia, Hawaii, Iowa, Michigan, Nuevo México, Uthan, y Whashigton). Los registros que se encuentran en la base de datos contienen información tanto socio-demográfica del paciente como clinico-patológica del tumor, el tipo de cirugía (mastectomía radical o parcial), estatus de la aplicación de radioterapia post-operatoria (con o sin), fecha de diagnóstico del cáncer y fecha de muerte.

Se analizan registros correspondientes a las mujeres que se sometieron a una cirugía en 1990 y con fecha de vencimiento el 31 de Diciembre del 2011. Se excluyen los casos de pacientes con cirugías en lugares alejados de las mamas o ganglios linfáticos así como los casos que fueron diagnosticados por certificado de muerte. Solo se consideran los casos de pacientes diagnosticadas con cáncer primario unilateral (Cáncer que se desarrolló en una solo mama, izquierda o derecha).

En [21] se realiza un análisis sobre esta misma base de datos para analizar los efectos de los dos tipos de cirugía (mastectomía y conservación de seno) aplicadas a las pacientes, en donde supone la existencia de las variables de confusión ya que las cirugías aplicadas a las pacientes fueron asignadas dependiendo del diagnóstico de cada paciente, en este estudio se concluyó que la cirugía con conservación de seno favorece a mujeres con diagnóstico no tan grave, además que no existía suficiente evidencia para concluir que alguna de las dos cirugías es mejor para el grupo de mujeres que corresponden a un peor diagnóstico.

Por otra parte, en [23] y [17] sugieren que el tratamiento secundario (radioterapia) afecta positivamente el tiempo de supervivencia cuando se aplica en conjunto con la cirugía con conservación de seno; sin embargo, cuando se aplica con la cirugía mastectomía no existe mucha diferencia en el beneficio, probablemente esto se deba a que las mujeres que se sometieron a una mastectomía tienen un diagnóstico más grave.

Por lo anterior se clasifica la variable Tratamiento en cuatro posibles categorías definiendo a cada tratamiento con una de las combinaciones de la cirugía y el estatus de radiación, con el fin de observar el efecto que tienen los tratamientos cirugía y la radioterapia en conjunto.

- Tratamiento1: Cirugía radical con radioterapia.
- Tratamiento2: Cirugía radical sin radioterapia.
- Tratamiento3: Cirugía conservación de seno con radioterapia.
- Tratamiento4: Cirugía conservación de seno sin radioterapia.

La base de datos tiene 16511 registros de pacientes, de las cuales 10841 (65.6 %) se sometieron a mastectomía contra 5670 (44.4 %) que se sometieron a

una cirugía con conservación de seno. La radiación post operatoria se aplicó discriminadamente en mujeres que se realizaron la mastectomía, donde solo el 8 % de estas pacientes se les aplicó radioterapia. Mientras que en el grupo de las mujeres que se sometieron a una cirugía de conservación de seno, las proporciones fueron más equitativas entre las pacientes que se les aplicó radioterapia post operatoria y a las que no con un 57 % y 43 %, respectivamente. El tratamiento más común en esta base de datos fue el tratamiento 2 con 60 % de pacientes que se sometieron a este y el menos común el primero con solo el 5 %, para los tratamientos 3 y 4 las proporciones fueron 15 % y 20 %, respectivamente.

En el Tabla 3.3 se encuentra una tabla de contingencia con cada una de las variables explicativas del Tabla 3.1 según los tratamientos. Algunas observaciones que se ven en esta tabla de contingencia son; el tratamiento 1 se aplica en menor proporción a mujeres mayores de 71 años, mientras que el tratamiento 4 se aplica en mayor proporción en este grupo. Además, se observa un mayor porcentaje en mujeres de raza negra y otras razas en el tratamiento 1. El tratamiento 3 fue más común entre mujeres casadas o en unión libre.

El cáncer más común fue el carcinoma ductal con el 80 % del total de la base, seguido del carcinoma lobular con el 8 %. La proporciones del tipo de cáncer mantuvieron similares en los cuatro tratamientos. Mientras que para tumores mayores de 2 cm se observa una mayor aplicación del tratamiento 1 y para tumores menores de 2 cm predeterminaban los tratamiento 3 y 4 (cirugía con concervación de seno). De igual forma, en mujeres con tumores de grado III y IV se sometieron mayormente al tratamiento 1, y para grados I y II no se observa un tratamiento preferente.

En el caso de la variable extensión, se observa que cuando el carcinoma es invasivo, se aplica en mayor proporción el tratamiento más agresivo (el trata-

Variable	Descripción	Categoría
EDAD	Indica la edad en que fue diagnosticada la paciente	Variable continua
GRADO	Grado del cáncer en el que se encuentra la paciente, según la clasificación del (NHS)	1: Grado I y Grado II 2: Grado III y Grado IV
TAMAÑO	Tamaño del tumor en centímetros	1: Menor a 2cm 2: Mayor o igual a 2cm
LUGAR	Tipo o Histología del tumor (Lugar en el que se originó el tumor)	1: Carcinoma Ductal 2: Carcinoma Lobular 3: Otros
ER	Porcentaje de células que contiene el receptor de estrógeno	1: Positivo o en linea ($\geq$10 % de células con RE) 2: Negativo (<10 % de células con RE)
GL	Si existe o no afección en los ganglios linfáticos	1: No hay afección 2: Hay afección
EXTENSIÓN	Indica si el cáncer es cofinado o invasivo	1: Confinado 2: Invasivo
LATERALIDAD	Lugar donde se originó el cáncer	1: Seno derecho 2: Seno Izquierdo
RAZA	Tipo de raza a la que pertenece la paciente	1: Blanca 2: Negra 3: Otra (Indios Americanos, Asia, Nativos Alaska, Islas pacifico)
EC	Estado social del paciente	1: Soltera (nunca casada, viudad, divorciada) 2: Casada (casado, unión libre)
REGIÓN	Hábitos de salud del estado donde pertenece con respecto al *America's Health Ranking* de 1990	1: Menores o iguales a 0.5 2: *Mayores a 0.5*

Tabla 3.1: Descripción de variables independientes.

miento 1), mientras que si el carcinoma es confinado se aplica similar proporción los tratamientos restantes. Algo semejante se observa que cuando existe daño en ganglios linfáticos, el tratamiento 1 el cual es el tratamiento más agresivo se aplica en mucho mayor proporción que los otros 3 tratamientos. Finalmente, para las variables lateralidad y hábitos de salud, las proporciones de los tratamientos son similares para cada una de las categorías de estas variables.

3.3. Estimación

Para un primer análisis se obtiene una estimación sin tomar en cuenta variables explicativas para la asignación del tratamiento, variables de riesgo y la existencia de variables de confusión. Se toma la función de supervivencia a través de la función $f(y_1, y_2)$ condicional

$$\begin{aligned} f_{1|2}(y_1|y_2) &= \frac{f(y_1, y_2)}{f_2(y_2)} \\ &= \frac{f_1(y_1)}{f_2(y_2)} * \left[C'(F_1(y_1), F_2(y_2)) - C'(F_1(y_1), F_2(y_2 - 1))\right], \end{aligned}$$

de la ecuación (2.1)

$$S_{1|2}(y_1) = \frac{1}{f_2(y_2)} \int_{y_1}^{\infty} f_1(t) * \left[C'(F(t), F_2(y_2)) - C'(F_1(t), F_2(y_2 - 1))\right] dt,$$

Con un cambio de variable $w = F_1(y_1)$ se tiene que

$$S_{1|2}(y_1) = \frac{1}{f_2(y_2)} \int_{F_{(Y_1)}}^{1} \left[C'(w, F_2(y_2)) - C'(w, F_2(y_2 - 1))\right] dw. \qquad (3.1)$$

En la Figura 3.4 se grafica la función de supervivencia estima para cada tratamiento junto con el estimador de Kaplan-Maier, se puede observar que cuando no se toma en cuenta la existencia de variables de confusión, el tratamiento 1

	Conservación de Seno(34.3 %)		Mastectomía(65.7 %)	
características	No Radiación n(%) 2415(42.6 %)	Radiación n(%) 3255(57.4 %)	No Radiación n(%) 9981(92.1 %)	Radiación n (%) 860(7.9 %)
Estados Vital				
Vivo	816 (33.8)	1490 (45.8)	3513 (35.2)	200 (23.3)
Muerto	1599 (66.2)	1765 (54.2)	6468 (64.8)	660 (76.7)
Características del tumor				
Lugar				
Ductal	1802(74.6)	2708(83.2)	8099(81.1)	662(77.0)
Lobular	298(12.3)	175(5.4)	779(7.8)	73(8.5)
Otro	315(13.1)	372(11.4)	1103(11.1)	125(14.5)
Tamaño				
$<$ 2 cm	1350(55.9)	2172(66.7)	4583(45.9)	146(17.0)
$\geq$ 2 cm	647(26.8)	814(25.0)	4163(41.7)	616(71.6)
Faltante	418(17.3)	269(8.3)	1235(12.4)	98(11.4)
Grado				
I & II	515(21.3)	1008(31.0)	2342(23.5)	168(19.5)
III & IV	337(14.0)	698(21.4)	2281(22.8)	379(44.1)
Faltante	1563(64.7)	1549(47.6)	5358(53.7)	313(36.4)
Marcador (ER)				
Positivo o al limite	879(36.4)	1834(56.4)	5218(52.3)	506(58.8)
Negativo	258(10.7)	515(15.8)	1558(15.6)	197(22.9)
Faltante	1278(52.9)	906(27.8)	3205(32.1)	157(18.3)
Lateralidad				
Derecha	1175(48.7)	1611(49.5)	4839(48.5)	406(47.2)
Izquierda	1240(51.3)	1644(50.5)	5142(51.5)	454(52.8)
Extensión				
Confinado	2134(88.4)	3078(94.6)	9075(90.9)	560(65.1)
Invasivo	220(9.1)	151(4.6)	809(8.1)	277(32.2)
Faltante	61(2.5)	26(0.8)	97(1.0)	23(2.7)
Ganglios linfáticos (GL)				
Sin afección	1556(64.4)	2523(77.5)	6768(67.8)	194(22.6)
Con afección	258(10.7)	532(16.3)	2912(29.2)	604(70.2)
Faltante	601(24.9)	200(6.2)	301(3.0)	62(7.2)

Tabla 3.2: Tabla de contingencia según el tratamiento

.

	Conservación de Seno(34.3 %)		Mastectomía(65.7 %)	
Características	No Radiación n(%) 2415(42.6 %)	Radiación n(%) 3255(57.4 %)	No Radiación n(%) 9981(92.1 %)	Radiación n (%) 860(7.9 %)
Características Sociodemográficas				
Edad(años)				
< 48	490 (20.3)	766 (23.5)	1901 (19.1)	254 (29.5)
48-59	398 (16.5)	767 (23.6)	1966 (19.7)	200 (23.3)
59-71	556(23.0)	1023(31.4)	2996(30.0)	260(30.2)
> 71	971 (40.2)	699 (21.5)	3118 (31.2)	146 (17.0)
Raza				
Blanca	2113(87.5)	2871(88.2)	8735(87.5)	711(82.7)
Negra	192(8.0)	230(7.1)	694(7.0)	87(10.1)
Otra	110(4.5)	154(4.7)	552(5.5)	62(7.2)
Estado Civil(EC)				
Casada	1197(49.6)	2074(63.7)	5641(56.5)	501(58.3)
Soltera	1100 (45.5)	1126(34.6)	4126(41.3)	339(39.4)
Faltante	118(4.9)	55(1.7)	214 (2.2)	20(2.3)
Hábitos de salud				
≤ 0.5	1124(46.5)	1576(48.4)	4725(47.3)	435(50.6)
> 0.5	1291(53.5)	1679(51.6)	5256(52.7)	425(49.4)

Tabla 3.3: Tabla de contingencia según el tratamiento

.

(mastectomía con radiación) es el tratamiento con una probabilidad menor de supervivencia, y el tratamiento 4 (cirugía con conservación de seno sin radiación) es el que asigna un mejor pronóstico, sin embargo, no se puede concluir que el tratamiento 4 es el mejor y el tratamiento 1 es el peor de los tratamientos, puesto que cada tratamiento es asignado dependiendo la gravedad de enfermedad de la paciente y a la par la gravedad de la enfermedad afectada a los tiempos de supervivencia.

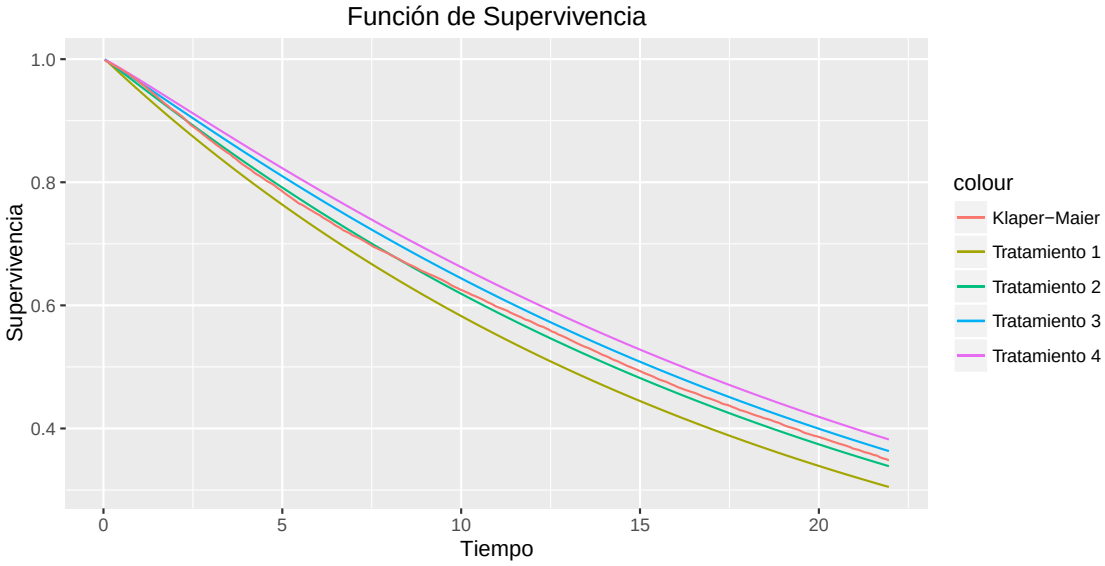

Figura 3.4: Función de supervivencia estimada para los tratamientos de acuerdo al estimador Kaplan-Maier sin factores de confusión.

Las variables que cuantifican la gravedad de la enfermedad y que usualmente se toman en cuenta para la elección del tratamiento son: tamaño de tumor, afección de ganglios linfáticos, tipo de cáncer (invasivo o confinado), grado del tumor, porcentaje de receptores de estrógeno [13]. Además, en el artículo [15] concluyen que para la elección del tratamiento no solo dependen de característi-

cas patológicas del tumor, sino también de las sociales; por ejemplo, las mujeres casadas tienden a elegir más la cirugía con conservación de seno sobre las mujeres solteras. La variable edad también es muy significativa a la hora de realizar la elección del tratamiento, al parecer las más jóvenes tienden a tomar las cirugías con conservación de seno. El tipo de raza, en mujeres de raza negra es más común la mastectomía que en mujeres de raza blanca.

Por otra parte, se ha encontrado que la mayoría de las variables anteriores tienen un efecto significativo sobre el tiempo de supervivencia; por ejemplo, la afección en ganglios linfáticos representan un riesgo de mortalidad mayor (ver[3],[26]). La extensión del tumor y el grado representan un factor importante no solo en la elección del tratamiento sino también en el tiempo de supervivencia [12], así como la edad, el tipo de cáncer y el tamaño de tumor tienen un efecto significativo sobre el tiempo de supervivencia [20]. El receptor de estrógenos representa otra variable significativa en el tiempo de supervivencia [14].

En el estado civil se ha observado que las mujeres solteras (nunca casadas, divorciadas o viudas) tienen un mayor riesgo de morir que las mujeres casadas o en unión libre [1]. Otros estudios realizados informan una mayor supervivencia en mujeres de raza blanca que en mujeres de raza negra y una menor supervivencia que en mujeres Indios Americanos, Asia, Nativos Alaska, Islas Pácifico [24]. Se ha asociado un aumento de riesgo de mortalidad para mujeres con cáncer de mama de lado izquierdo que se someten a radioterapia asociando la mortalidad con un riesgo alto de mortalidad cardíaca [22].

En el Tabla 3.1 se describen las variables explicativas que se tomaron en consideración para el análisis, así como cada una de sus categorías. Todas las variables explicativas que se encuentran en la tabla fueron anexadas a cada uno de los modelos lineales que se encuentran dentro de la formulación del modelo princi-

pal. En el modelo de Cox para encontrar factores de riesgo en los tiempos de supervivencia, el modelo de regresión logística multinomial para determinar las características para la asignación del tratamiento, y finalmente al modelo lineal agregado a la estructura de dependencia con la finalidad de encontrar las variables que son significativas dentro de la asociación del tiempo de supervivencia y el tipo de tratamiento; es decir, las variables confusión.

Para la variable edad se ajusta un polinomio ortogonal de segundo orden ya que se toma en cuenta la hipótesis que el riesgo de mortalidad varía con respecto a la edad para la incidencia como se propone en [8].

Como se observa en el cuadro 3.3 la base de datos del SEER que se analiza presentan una gran proporción de datos faltantes, ya sea por que la paciente se niega a responder, a ser revisada, o simplemente porque la información no esta disponible. Los porcentajes de los datos faltantes según la variables son: EC 2.46 %, TAMAÑO 12.36 %, GRADO 53.19 %, EXTENSIÓN 1.25 %, GL 7.04 % y RE 33.59 %

Por lo tanto, la estimación de los parámetros para esta base con ausencia de datos se realiza a través de técnicas del imputación múltiple. Se supone que los datos faltante son de tipo MAR; esto es, que los datos faltantes de una variable X dependen de otras variables incluidas en el estudio pero no de ella misma.

Para la primera fase del método de imputación se utiliza el algoritmo MICE (Multivariate Imputation by Chained Equation) del paquete mice del software R el cual utiliza simulación paramétrica y genera M imputaciones indipendientes a partir de métodos bayesianos [2]. Debido a que el algoritmo que encontraba el máximo de la función de verosimilitud tardaba aproximadamente 8 horas, tomando el cuenta el número de imputaciones necesarias y el tiempo de ejecución

de los algoritmos se eligió $M = 7$.

Para la segunda parte del método de imputación multiple (estimación y análisis) se ocupó los programas que se encuentra del apéndice A que definen la función de verosimilitud 3.15, la cual se máxima utilizando el método de Newton (ver apéndice B). La integral que se encuentra en la función de verosimilitud se aproxima con el método de reglas compuesta (ver apéndice B). Uno de los objetivos en esta segunda etapa fue encontrar los modelos más parsimonioso para cada imputación, para esto, se ocupo el criterio BIC (Criterio de Información Bayesiana ver apéndice B).

Finalmante, para la última etapa del método se utilizan las reglas de Rubin descritas en al capítulo dos para encontrar el estimador final. Las variables que se consideraron significativas en el modelo, son aquellas que salieron significativas por lo menos para el 50 % de las imputaciones.

3.4. Resultados

En las Tablas 3.4 se registran los estimadores finales correspondientes al modelo de dependencia y el modelo de Cox. Para el primer modelo, podemos observar que las únicas variables significativas son Extensión (esto es, si el cáncer es confinado o invasivo) y GL (Si los ganglios linfáticos esta dañados o no), estas dos variables las podemos tomar como variables de confusión, los cuales coincide con la literatura médica ([13], https://www.cancer.gov, https://www.cancer.org), en donde se determina que estas dos variables son fundamentales para la elección del tratamiento y a la vez afecta el tiempo de vida del paciente.

Los factores de riesgo que se obtuvieron en esta base de datos son: ER, raza, extensión, GL, tamaño, EC y edad (Tabla 3.4). Se observa que las pacientes con menos de 10 % de células con receptor de estrógenos tienen 14 % más de riesgo de mortalidad. Se tiene un peor pronóstico para las mujeres de raza negra que para las mujeres de raza blanca con 37 % más de riesgo de mortalidad, mientras que para las mujeres de otras razas (mujeres hispanas) el pronostico es ligeramente mejor con 19 % menos riesgo comparado con las mujeres de raza blanca.

El tamaño del tumor es otra de las variables de riesgo, donde las pacientes con tumores más grandes de 2 cm tienen 31 % más de riesgo de morir que una persona con tumor menor o igual a 2 cm. Las mujeres casadas o en unión libre tienen 10 % menos riesgo de mortalidad que las mujeres que estén solteras, viudas o separadas.

La afección de ganglios linfáticos representa un 69 % más de riesgo de muerte; y las mujeres con cáncer invasivo tienen un riesgo 71 % contra las que tiene un cáncer confinado. Esta últimas dos variables representan una mayor tasa de riesgo. Por último, la edad también resulta una variable de riesgo, la cual se ajustó con el polinomio de la Figura 3.5 de grado 2, en donde se observa que el riesgo crece más rápidamente después de los 50 años.

En el Tabla (3.5) se encuentran los estimadores finales del modelo multinomial, en donde se obtuvieron como variables significativas para la asignación de tratamiento, el tipo de cáncer, raza, grado de tumor, extensión, la afección de ganglios linfáticos, el tamaño del tumor y la edad.

Para el modelo de regresión multinomial se tienen las siguientes observaciones, cuando el tipo de cáncer es lobular u otro, se tiene que se prefiere el tratamiento 1 versus tratamiento 2, mientras que para el tratamiento 1 versus tratamiento 4 no se muestra alguna preferencia. Para el tratamiento 1 versus el tratamiento

	Parametros	Limite inferior	Limite superior	Error estándar
Parámetros de dependencia				
(Intercept)	-0.032318	-0.058506	-0.006086	0.013263
EXTENSIÓN	-0.182348	-0.230633	-0.133167	0.025606
GL	0.097634	0.052097	0.142767	0.022539
Parámetros modelo Cox.				
Forma	1.256419	1.235699	1.277486	1.008512
Escala	25.481330	24.588458	26.406624	1.017919
poly(EDAD,2)1	94.960866	91.928443	97.993289	1.547178
poly(EDAD,2)2	19.953080	17.393624	22.512537	1.305853
ER2	0.134122	0.063070	0.205173	0.032884
RAZA2	0.319922	0.246800	0.393045	0.037164
RAZA3	-0.182079	-0.280389	-0.083769	0.050010
EXTENSIÓN2	0.541079	0.475890	0.606268	0.032836
GL1	0.528309	0.480016	0.576602	0.024028
TAMAÑO1	0.274556	0.223884	0.325228	0.024499
EC2	-0.106912	-0.148851	-0.064973	0.021355

Tabla 3.4: Estimadores finales del modelo de dependencia y modelo de Cox.

3, se tiene que si el cáncer es tipo lobular entonces se prefiere el tratamiento 1, mientras que si el cáncer es otro no se muestra alguna preferencia por el tratamiento.

Por otra parte, se encontró que el tipo de raza solo es significativa cuando se compara tratamiento 1 versus tratamiento 3 para la raza otra (Indios Americanos, Asia Nativos Alaska, Islas Pácifico), en donde se tiene como preferencia el tratamiento 1. Además, se obtuvo que cuando el tumor es de Grado III o IV, el tratamiento 1 se prefiere en comparación de los otros tres tratamientos.

Una de las variables más significativas que se obtuvo es el tamaño de tumor, en donde al igual que el grado de tumor, cuando el tamaño de tumor rebasa los 2.5 cm entonces se prefiere el primer tratamiento en comparación de los otros 3. Sin embargo, se nota más la diferencia cuando se comprara el tratamiento 2

versus el tratamiento 3 y 4, lo que sugiere que la preferencia se inclina más al tipo de cirugía que se aplicará que a la radioterapia.

Otra de las variables más significativas que se obtuvieron fue la extensión de tumor, es decir, si el cáncer es confinado o invasivo. Cuando el cáncer es invasivo se muestra una mayor inclinación por el tratamiento más agresivo, el tratamiento uno versus los otros tres tratamientos.

Cuando existe afectación de ganglios linfáticos se observa la preferencia del primer tratamiento en comparación de los otros; además se observa que la preferencia aumenta conforme disminuye la agresividad del tratamiento; es decir, el tratamiento 1 se prefiere al tratamiento 2, sin embargo la preferencia del primero aumenta si se compara con el tercero, y sigue aumentando cuando se compara con el cuarto.

Finalmente, al igual que en el modelo de Cox, para la variable edad se agregó un polinomio de segundo orden, en las Gráficas 3.6, 3.7 y 3.8, observar los polinomios obtenidos para cada un de los tratamientos.

3.5

El modelo propuesto además de poder predecir la mortalidad por causa de cáncer de mama para los cuatro posibles tratamientos, puede asignar en diferentes grupos de riesgo a los pacientes que eventualmente podrían presentar el evento de interés. Lo anterior se realiza mediante una puntación de riesgo que se obtienen de las variables de riesgo y este definido como $R = \beta^T x$, donde β es el vector de parámetros correspondientes al modelo de Cox dados en el Cuadro 3.4.

Se tomó el 30 % más bajo del vector R para crear el grupo de riesgo bajo, el siguiente 40 % definió el grupo medio de riesgo y finalmente con el 30 % restante se obtuvo el grupo de alto riesgo. Dado estos grupos se calculó la función de Supervi-

vencia 3.1 para para cada tratamiento en cada uno de los grupos, con la finalidad de observar si para algún grupo en especial hay un tratamiento que beneficie más.

En la Figura 3.9 se encuentran las funciones de supervivencia condicionada a cada uno de los tratamientos para cada grupo. Se observa que el grupo de Riesgo Alto en general tiene un peor pronóstico que los dos grupos restantes; sin embargo, también se observa que para este grupo en particular el tratamiento más agresivo ofrece un mayor beneficio.

Por el contrario, para los dos grupos restantes; es decir, grupos de riesgo medio y bajo, el tratamiento menos agresivo da un mayor tiempo de supervivencia que los otros tratamientos; aunque en el grupo de Riesgo medio se observe una menor diferencia en las gráficas de supervivencia de entre los tratamientos que en el grupo de riesgo bajo.

Por lo tanto, se obtiene que para el 30 % de mujeres la cuales tienen el peor pronóstico el tratamiento 1 es el que se recomienda, mientras que para el 70 % de mujeres se recomienda el tratamiento 4. Lo cual coincide con la literatura médica ([13], https://www.cancer.org, https://www.cancer.gov) donde se recomienda un tratamiento más agresivo para mujeres con peor pronóstico, y un tratamiento no tan agresivo si el diagnóstico nos es tan grave.

De igual forma, se gráfica la función de riesgo acumulado para cada uno de los grupos ocupando la siguiente función

$$\begin{aligned} \hat{H}_{1|2}(y_1) &= -log\left(\hat{S}_{1|2}(y_1)\right) \\ &= -log\left\{\frac{1}{f_2(y_2)}\int_{F_{(Y_1)}}^{1}\left[C'(w, F_2(y_2)) - C'(w, F_2(y_2-1))\right]dw\right\} \end{aligned} \quad (3.2)$$

En la Figura 3.10 se encuentran las gráficas de riesgo en donde observamos resultados similares que con la función de supervivencia. Para el grupo de riesgo alto, el tratamiento 4 representa un riesgo mayor que los otros tratamientos, mientras que para los dos grupos restantes se tiene que el tratamiento 1 da un mayor riesgo.

Finalmente, para la validación del modelo obtenemos los residuales de ocupados en [5], se tiene que la función de riesgo (3.3) está distribuida exponencialmente con media uno [4]. Si suponemos que $u = H_{1|2}(y_1)$ y aplicamos la transformación integral de probabilidad a u, entonces $exp(-u)$ debe tener una distribución normal estándar para un número de observaciones suficientemente grande. Luego, para las n observaciones se tiene que la gráfica

$$\frac{n+1-j}{n+1} \quad \text{contra} \quad \exp(-u(j))$$

debe ajustarse a una línea recta. De esta manera se observa sí el modelo ajusta bien los datos.

En la Figura 3.11 se encuentran los residuales para cada uno de los tratamientos, en donde podemos concluir que efectivamente para los cuatro tratamientos las gráficas aproximan a un línea recta, por lo tanto, podemos decir que se realizó un buen ajuste para la base de datos seleccionada.

	Parametros	Limite inferior	Limite superior	Error estándar
Parámetros multinomial				
Tratamiento Dos	(Cirugía Radical	sin Radiación)		
(Intercept)	4.211471	4.008567	4.414374	0.102729
poly(EDAD,2)1	50.003809	39.272470	60.735149	5.475274
poly(EDAD,2)2	19.998661	9.868074	30.129248	5.168762
factor(LUGAR)2	-0.306751	-0.578403	-0.035099	0.138374
factor(LUGAR)3	-0.276703	-0.495066	-0.058340	0.111298
factor(RAZA)2	-0.098821	-0.351108	0.153467	0.128559
factor(RAZA)3	-0.169656	-0.461971	0.122659	0.149093
factor(GRADO)2	-0.252013	-0.438411	-0.065615	0.091116
factor(EXTENSIÓN)2	-1.106720	-1.288137	-0.925303	0.091706
factor(GL)1	-1.461272	-1.647202	-1.275341	0.093517
factor(TAMAÑO)1	-0.731219	-0.983048	-0.479389	0.119067
Tratamiento Tres	(Cirugía parcial	con Radiación)		
(Intercept)	3.527976	3.318033	3.737919	0.106035
poly(EDAD,2)1	11.991804	0.223380	23.760228	6.004408
poly(EDAD,2)2	5.994981	-5.263086	17.253048	5.744017
factor(LUGAR)2	-0.744214	-1.051949	-0.436480	0.156891
factor(LUGAR)3	-0.165380	-0.405661	0.074902	0.122444
factor(RAZA)2	-0.033576	-0.316168	0.249015	0.143914
factor(RAZA)3	-0.446367	-0.777157	-0.115577	0.168744
factor(GRADO)2	-0.375890	-0.567881	-0.183898	0.095108
factor(EXTENSIÓN)2	-1.267920	-1.514533	-1.021307	0.124606
factor(GL)1	-1.902094	-2.112007	-1.692180	0.104999
factor(TAMAÑO)1	-1.465224	-1.714463	-1.215985	0.119955
Tratamiento cuatro	(Cirugía parcial	sin Radiación)		
(Intercept)	3.132205	2.919468	3.344941	0.107448
poly(EDAD,2)1	58.002787	46.342363	69.663211	5.949305
poly(EDAD,2)2	58.005566	47.030160	68.980971	5.599799
factor(LUGAR)2	0.239576	-0.054856	0.534009	0.150132
factor(LUGAR)3	-0.111341	-0.360650	0.137968	0.126745
factor(RAZA)2	0.233083	-0.055576	0.521742	0.147123
factor(RAZA)3	-0.228700	-0.575710	0.118311	0.176679
factor(GRADO)2	-0.592211	-0.797346	-0.387076	0.100794
factor(EXTENSIÓN)2	-0.606349	-0.856574	-0.356125	0.124416
factor(GL)1	-2.155191	-2.426945	-1.883437	0.129333
fcator(TAMAÑO)1	-1.286593	-1.561167	-1.012020	0.130075

Tabla 3.5: Estimadores finales de modelo de regresión logística multinomial.

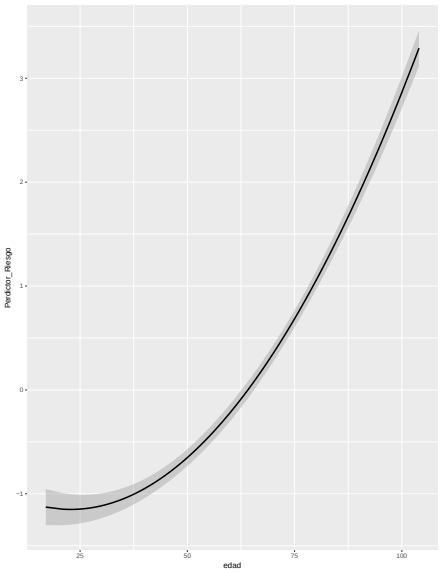

Figura 3.5: Polinomio edad para el modelo de cox.

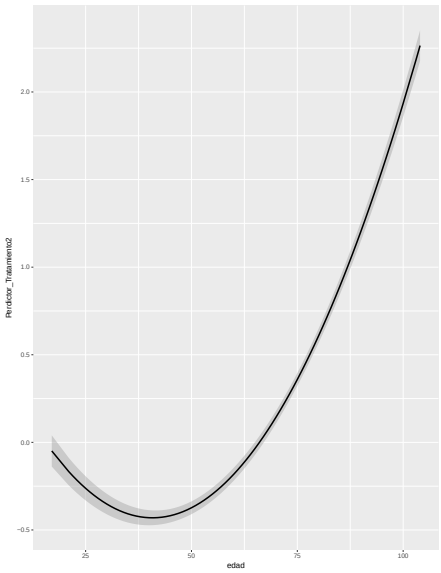

Figura 3.6: Polinomio de edad para tratamiento 2.

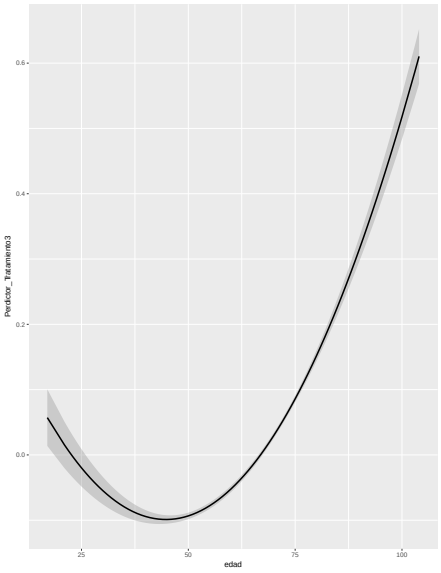

Figura 3.7: Polinomio edad para tratamiento 3.

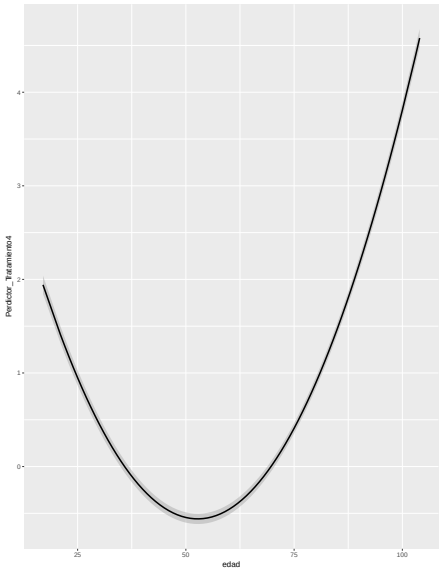

Figura 3.8: Polinomio edad para tratamiento 4.

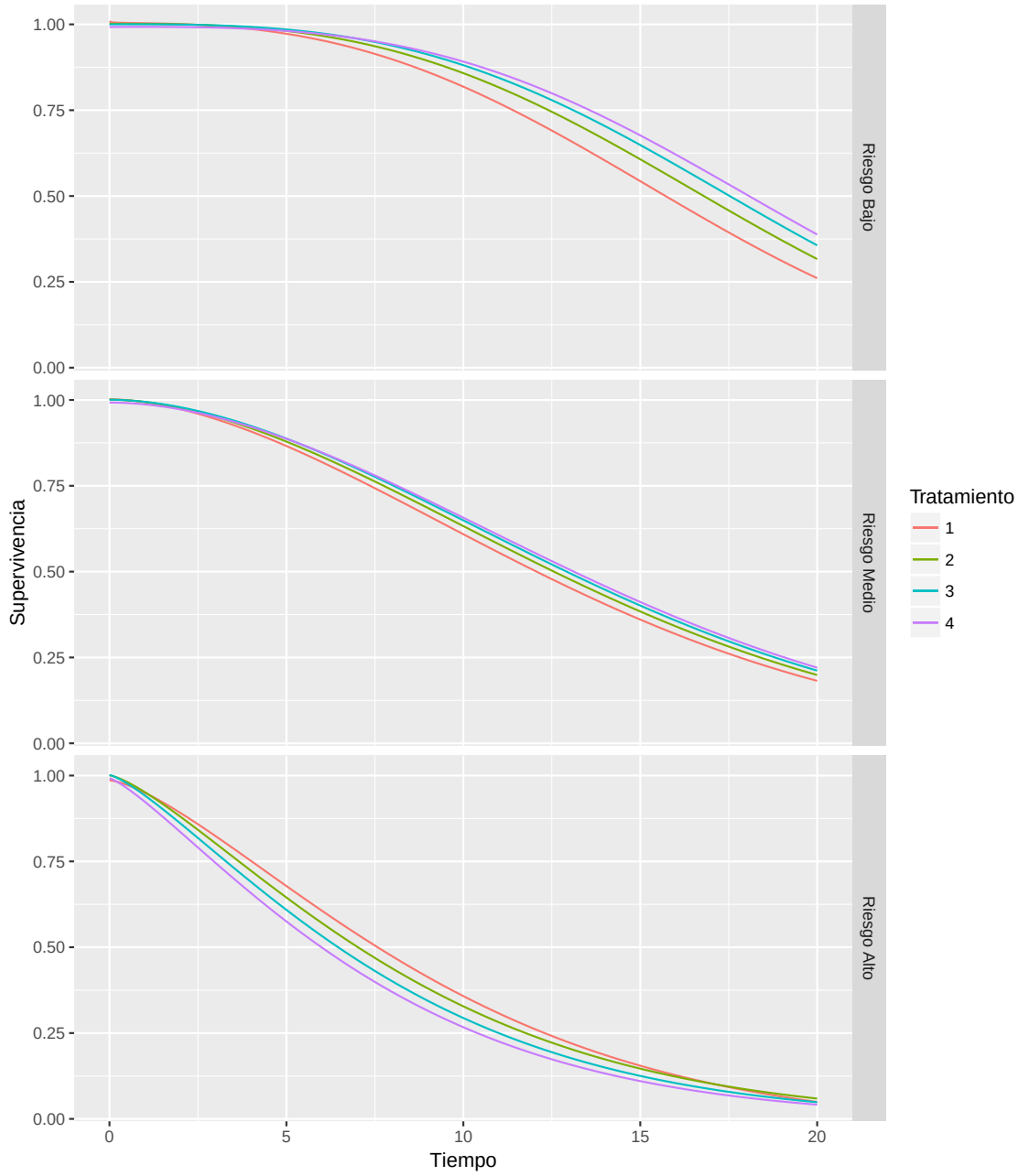

Figura 3.9: Función estimada de supervivencia (3.1) según el tratamiento para los grupos diferentes tipos de riesgo (alto, medio y bajo).

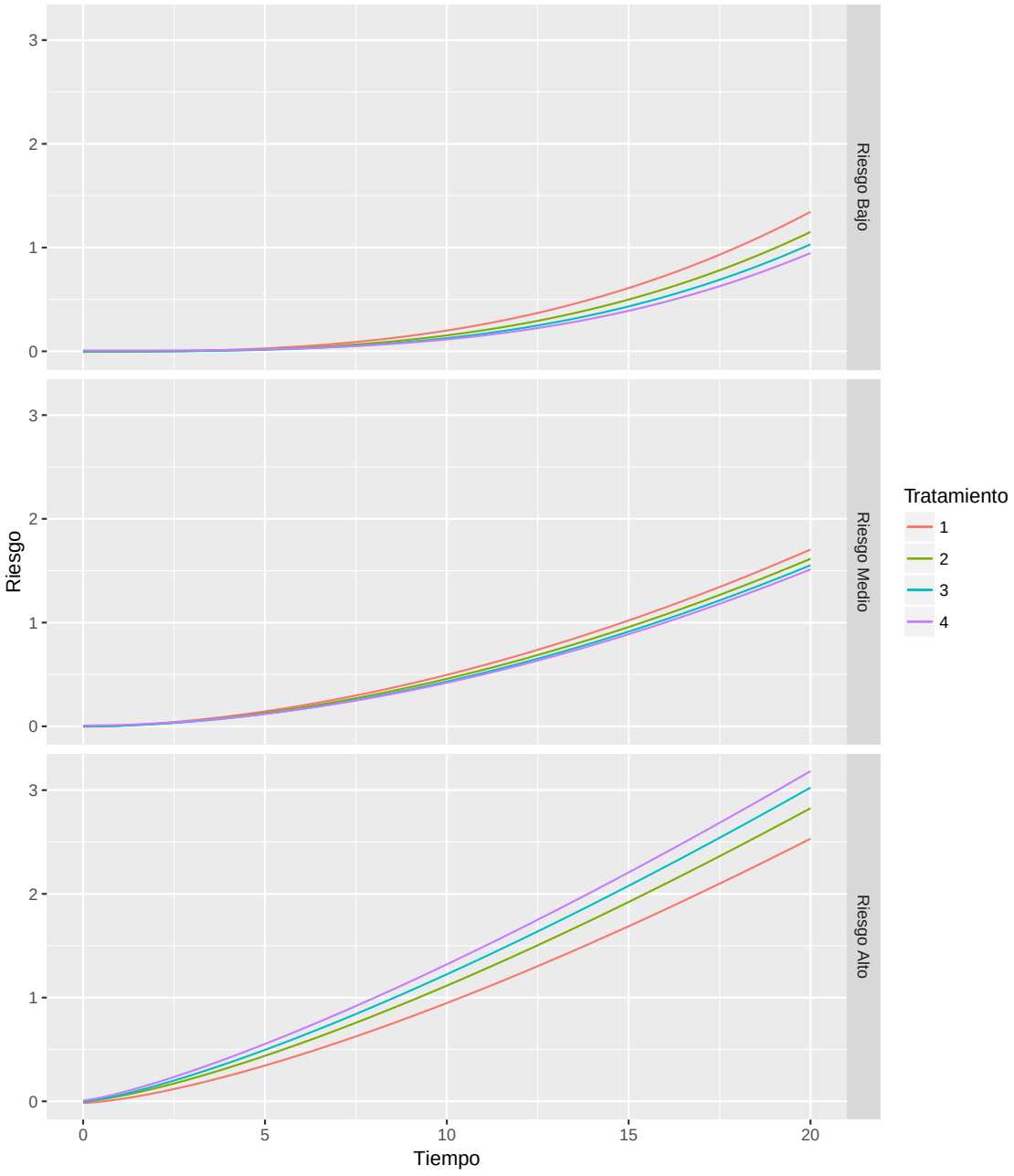

Figura 3.10: Función estimada de riesgo acumulado (3.3) según el tratamiento para los diferentes grupos de riesgo (alto, medio, bajo).

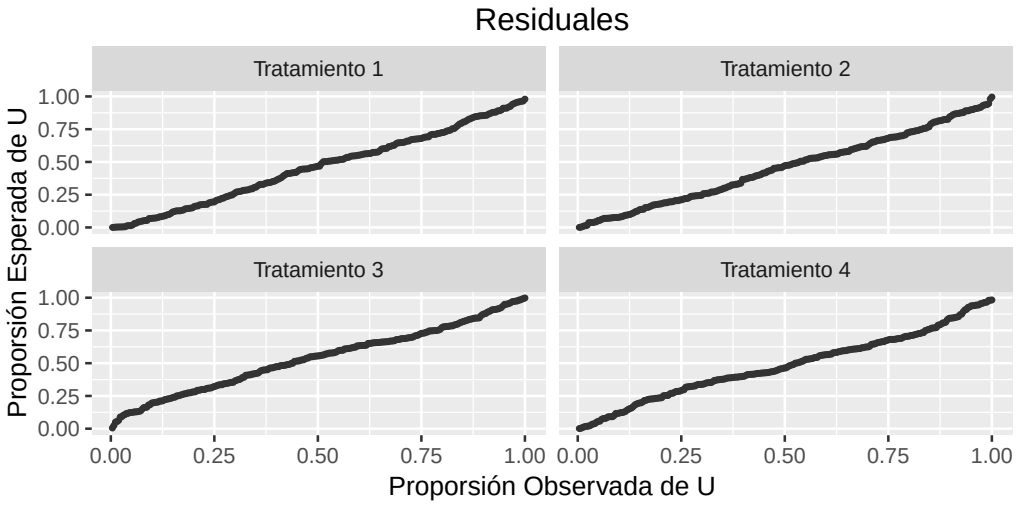

Figura 3.11: Gráfica de diagnóstico para el ajuste del modelo según el tratamiento.

Conclusiones

En este trabajo de tesis se desarrollo una metodología para el análisis de datos de supervivencia en estudios observacionales, esta metodología se fundamentó en caracterizar una función conjunta bivariada para el tiempo de supervivencia y tratamiento, en donde se quita el supuesto de independencia entre las variables y se utiliza una cópula bivariada gaussiana para la construcción de la misma. El enfoque de construcción de la función conjunta no solo resulto ser flexible para determinar la función, sino también facilitó el análisis de resultados ya que se logró identificar aquellas variables que afectan tanto al riesgo, la selección de tratamiento y la dependencia entre la variable respuesta y el tratamiento.

La metodología fue aplicada en una base de datos de mujeres diagnosticadas con cáncer de mama y que se sometieron a una cirugía en presencia de características socio-demográficas y clinopatológicas. Para esta aplicación, todas la variables identificadas como variables de riesgo y variables de selección de tratamiento concuerdan con la literatura revisada, y las dos variables identificadas como variables de confusión son variables que determinan con presión la gravedad o diagnóstico de las pacientes.

Mediante la función de densidad conjunta condicionada al tratamiento se pudó obtener la función de riesgo y supervivencia, misma que ayudaron a identificar los tres grupos de riesgo: alto, medio y bajo. En estos grupos de riesgo se

concluyo que solo el 30 % del total e la pacientes tienen un diagnostico que se puede considerar grave, es decir que para el otro 70 % de pacientes la enfermedad no es determinante.

Aunque la función de densidad conjunta resulto ser fácil de construir, la estimación de paramétros no resulto ser tan optima debido a la cantidad de estimadores que contiene el modelo (correspondiente a 5 componentes lineales) por lo que los tiempos de ejecución se elevaban en gran medida al agregar una variable explicativa más al modelo.

Se eligío la cópula gaussiana para la construcción ya que en la literatura se considera una cópula robusta, sin embargo, se pudría tomar una cópula diferente para la obtención de la función que podrá resultar ser más optima, de igual manera, el modelo de Cox se anexo para entender el comportamiento de los tiempos de supervivencia con el supuesto de que los tiempos de supervivencia está descrita por una función Weibull por ser uno de los modelos más utilizados en el análisis de supervivencia, este supuesto podría ser cambiado tomando una distribución exponencial o log-normal.

Para la aplicación de imputación múltiple se considero que los datos perdidos era de tipo MAR, sin embargo este supuesto también puede ser modificado.

Finalmente, en los tiempos de supervivencia se distinguió los tiempos censurados dependiendo si el paciente estaba vivo o muerto, una modificación importante en el modelo que ayudaría a comprender mejor el comportamiento del fenómeno es poder distinguir el tipo de muerte (muerte por causa de la enfermedad o por otra causa), mediante modelos de riesgos de competencia, pero esto ya es otra historia.

Bibliografía

[1] Aizer, A., Chen, M.-H., McCarthy, E., Mendu, M., Koo, S., Wilhite, T., Graham, P., Choueiri, T., Hoffman, K., Martin, N., Hu, J., and Nguyen, P. Marital status and survival in patients with cancer. *Journal of Clinical Oncolgy 31* (2013), 3852.

[2] Azur, M., Stuart, E., Frangakis, C., and Leaf, P. Multiple imputation by chained equations: What is it and how does it work? *National Institutes of Health Public Access 20* (2011).

[3] Carter, C., Allen, C., and Henson, D. Local-regional radiotherapy and surgery is associated with a significant survival advantage in metastatic breast cancer patients. *CANCER 63* (1989), 181–187.

[4] Collett, D. *Modelling survival data in medical research.* Chapman Hall, 1994.

[5] Copas, J., and Herdari, F. Estimating the risk of reoending by using exponential mixture models. *Royal Statistical Society 160* (1997).

[6] Czado, C., Kastenmeier, R., Brechmann, E., and Min, A. A mixed copula model for insurance claims and claims sizes. *Scandinavian Actuarial Journal 4* (2012), 278–305.

[7] de Leon, A., and Wu, B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine 30* (2011), 175–185.

[8] Escarela, G., and Hernández, A. Modelado de parejas usando cópulas. *Revista colombiana de Estadística 32* (2009).

[9] Frank, K. Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research 29* (2000).

[10] Frees, E., and Vardez, E. Understanding relationships using copulas. *North American Actuarial Journal 3* (1998).

[11] Genest, C., and Neslehov, J. A primer on copulas for count data. *ASTIN Bulletin 37* (2007).

[12] Gnerlich, J., Jeffe, D., Deshpande, A., Beers, C., Zander, C., and Margenthaler, J. Surgical removal of the primary tumor increases overall survival in patients with metastatic breast cancer: Analysis of the 19882003 seer data. *Annals of Surgical Oncology 14* (2007), 2187–2194.

[13] Hernández, G., Bernardello, E., and Barrost, A. *Cáncer de mama al día.* Editorial Medica Panamericana, 2016.

[14] Iwamoto, T., Booser, D., Valero, V., Murray, J., Koenig, K., Esteva, F., Ueno, N., Zhang, J., Shi, W., Qi, Y., Matsuoka, J., Yang, E., Hortobagyi, G., Hatzis, C., Symmans, F., and Pusztai, L. Estrogen receptor (er) mrna and er-related gene expression in breast cancers that are 1 *Annals of Surgical Oncology 30* (2012), 686.

[15] King, M., Kenny, P., Shiell, A., Halll, J., and Boyages, J. Quality of life three months and one year after first treatment for early stage breast

cancer: Influence of treatment and patient characteristics. *Quality of Life Research 9* (2000), 789–800.

[16] Li, M., Boehnke, M., Abecasis, C., and Song, P. Quantitative train linkage analysis using gaussian copulas. *Genetics Society of America 173* (2006), 2317–2327.

[17] Ly, B., Vlastos, G., Rapiti, E., and Vinh-Hung, V. Local-regional radiotherapy and surgery is associated with a significant survival advantage in metastatic breast cancer patients. *Tumori Journal 96* (2010), 947–954.

[18] Nelsen, R. *A introduction to copula*, 2 ed. Springer, 2006.

[19] Pourhoseingholi, M., Baghestani, A., and Vahedi, M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench 5* (2012).

[20] Rosenberg, J., Chia, Y., and Plevritis, S. The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the us seer database. *Breast cancer research and treatment 89* (2005).

[21] Rubin, D. For objetive causal inference design trumps analysis. *The Annals of Applied Statistics 3* (2008), 808–840.

[22] Rutter, C., Chagpar, A., and Evans, S. Breast cancer laterality does not influence survival in a large modern cohort: implications for radiation-related cardiac mortality. *Internacional Journal of Radioation Oncology biology physics 90* (2014), 329–334.

[23] Vallis, K., and Tannock, I. Postoperative radiotherapy for breast cancer: growing evidence for an impact on survival. *Journal of the National Cancer Institute 96* (2004).

[24] Wilson, R., Cameron, M., Burhansstipanov, L., Roubidoux, M., Roubidoux, M., Cobb, N., Lynch, C., and Edwards, B. Disparities in breast cancer treatment among american indian, hispanic and non-hispanic white women enrolled in medicare. *Journal of Health Care for the Poor and Underserved 18* (2007), 648–664.

[25] Wunsch, G. Confounding and control. *Demographic Research 16* (2007).

[26] Yu, K.-D., Jiang, Y.-Z., Chen, S., Cao, Z.-G., Wu, J., Shen, Z.-Z., and Shao, Z.-M. Effect of large tumor size on cancer-specific mortality in node-negative breast cancer. *Mayo Clinic 87* (2012), 1171–1180.

Apéndice A

Códigos

Defición integral de cópula

* *Parámetros de la función*

* *t = Limite inferior* $G_1(Y_1)$

* *U2* $= G_2(Y_2)$

* *rho = Parámetro de dependencia*

****

```
int.C1.prima1 <- function(t, U2, rho)
{
  n = 500
  f0 <- function(x){
    pnorm((qnorm(U2) - rho*qnorm(x))/sqrt(1-rho^2))}
  k <- c(1:n-1)
  h <- (1-t)/n
  suma <- sum(f0(t+k*h))
  integral <- ((1-t)/n)*((f0(t)+f0(1))/2+suma)
}
```

Función de regresión

```
*  Parámetros de la función
*  parametros = Vector de estimadores [ζ, κ, λ, β, α]
*  datos = Matriz de datos [Tiempo supervivencia, Tratamiento, Censura]
*  Matriz.Diseno.1 = Matriz de diseño de las variables causales
*  Matriz.Diseno.2 = Matriz de diseño de las variables de riesgo
*  Matriz.Diseno.3 = Matriz de diseño de las variables selectivas del tratamiento
**********************************************************************

regresion1 <- function(parametros, datos, Matriz.Diseno.1,
                       Matriz.Diseno.2, Matriz.Diseno.3)
{
  m1 <-  dim(Matriz.Diseno.1)[2]
  m2 <- dim(Matriz.Diseno.2)[2] - 1
  m3 <- dim(Matriz.Diseno.3)[2]
  delta <- parametros[1:m1]
  form <- exp(parametros[m1+1])
  scala <- exp(parametros[m1+2])
  beta <- parametros[(m1+3):(m1+m2+2)]
  J <- max(datos[2])
  alfa <- parametros[(m1+m2+3): (m1+m2+2+(m3*(J-1)))]
  tiempo <- datos[,1]
  categoria <- datos[,2]
  # modelo lineal de dependencia
  rho <- Dependencia(Matriz.Diseno.1, delta, m1)
  # modelo de cox
  Fs <- Cox(Matriz.Diseno.2, beta, form, scala, tiempo, m2)
```

```
  # modelo multinomial
  probabilidades <- Multinomial(Matriz.Diseno.3,alfa,
                                J,categoria,m3)
  datoss <- data.frame(Fs,c=categoria,cen=datos[,3],
                                probabilidades,rho)
  datos1<-subset(datoss,datoss$cen==1)
  datos2<- subset(datoss,datoss$cen==0)
  f1 <- apply(datos1, 1, dGaussWeiMulti)
  f2 <- apply(datos2, 1, dSurv)
  log.L <- -1*(sum(log(f1))+sum(log(f2)))
  log.L
}
```

Función de densidad

* *Parámetros de la función*

* *Datos = Matriz de datos* $[f_1(Y_1), F_1(Y_1), Y_2]$

```
**************************************
dGaussWeiMulti <- function(dato){
  #Los parmetros de la Weibull:
  categoria <- dato[3]
  f1  <- dato[2]
  u1  <- dato[1]
  probabilidades1 <- dato[5:8]
  u2 <- probaacum(probabilidades1,categoria)
  u2_menos <- probaacum(probabilidades1,(categoria-1))
  rho <- dato[9]
  if(categoria == 4)
  { f <- f1*( 1- C1.prima(u1,u2_menos,rho)) }
```

```
  else
  {if(categoria == 1)
  {f<- f1*( C1.prima(u1,u2,rho))}
    else
    {f<- f1*( C1.prima(u1,u2,rho)- C1.prima(u1,u2_menos,rho))}}
  f
}
```

Apéndice B

Métodos númericos

B.1. Método de Newton

El método de Newton es un algoritmo usado para maximizar o minimizar una función. Este es un método iterativo que empieza en un punto inicial x_0 y obtiene una mejor aproximación x_1 mediante la fórmula

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

La expresión anterior se obtiene mediante el desarrollo de Taylor tomado hasta el segundo término. Así, se realizán sucesivas iteraciones hasta que el método converja lo suficiente.

Sea $f : [a, b] \to \mathbb{R}$ una función derivable definida en el intervalo $[a, b]$. Se toma un punto inicial x_0 y se define cada punto sucesivo de la forma

$$x_{n+1} = x_n + \frac{f(x_n)}{f'(x_n)},$$

donde f' denota la derivada de f.

Aunque el método de Newton es muy rápido y eficiente ya que la convergencia es de tipo cuadrático, no siempre se garantiza la convergencia, La única forma de

alcanzar convergencia es seleccionar un valor inicial lo suficientemente cercano a la raíz buscada.

B.2. Reglas compuestas

La regla del trapecio compuesto es un método para aproximar un integral definida en el intervalo $[a, b]$ utilizando n trapecios, el método supone que la función a integral f es continua y positiva en el intervalo $[a, b]$.

Primero se divide el intervalo $[a, b]$ en n intervalos de la misma longitud $\frac{b-a}{n}$ y se aplica la regla del trapecio a cada subintervalo.

$$\int_a^b f(x)dx = \int_a^{x_1} f(x)dx + \int_a^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^b f(x)dx$$

Cada integral es aproximada mediante la regla del trapecio

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{h}{2}\left[f(a) + f(x_1)\right] + \frac{h}{2}\left[f(x_1) + f(x_2)\right] + \cdots + \frac{h}{2}\left[f(x_{n-1}) + f(b)\right] \\ &\approx \frac{h}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{n-1} f(x_i)\right].\end{aligned}$$

con $h = \frac{b-a}{n}$ y $x_i = a + ih$ para $i = \{1, 2, \ldots, n\}$.

Para hallar el error E de la aproximación se toma en cuenta cada una de los errores en cada subintervalo. El cual es

$$E \leq \mid \frac{(b-a)}{12} h^2 M \mid$$

donde M es el máximo de la función $f''(x)$ en el intervalo $[a, b]$

B.3. Criterio de información bayesiana BIC

El criterio de Información Bayesiana propuesto por Schwarz en 1978, a sido un de los métodos más usados para la selección de modelos. Este criterio utiliza un enfoque bayesiano y evalúa a los modelos en términos de probabilidades posteriori.

Se supone que se tienen s modelos no necesariamente anidados, el objetivo es encontrar el modelo que mejor describa al conjunto $x_n = (x_1, x_2, \ldots, x_n)$. La función de densidad del modelo $i - simo$ denotado por M_i y su vector de parámetros θ_i esta dada por $f_i(x_n|M_i, \theta_i)$ para $i \in \{1, 2, \ldots, s\}$ y $\pi_i \in \Theta \in \mathbb{R}^k$. Sea $\phi_i(\theta)$ la densidad a priori de θ_i dado el modelo M_i y $\mathbb{P}(M_i)$ una densidad de probabilidad a priori que asigna probabilidad positiva a cada uno de los modelos M_i con $i \in \{1, 2, \ldots, k\}$. Mediante el teorema de Bayes se obtiene la probabilidad a posteriori del i-ésimo modelo.

$$\mathbb{P}(M_i|x_n) = \frac{\mathbb{P}(M_i)f_i(x_n|M_i)}{f(x_n)} = \frac{\mathbb{P}(M_i)f_i(x_n)}{f(x_n)},$$

donde

$$f_i(x_n) = \int_{\Theta_i} f_i(x_n|\theta)\pi_i(\theta)d\theta. \tag{B.1}$$

Se adopta como el mejor modelo aquel que tenga mayor probabilidad a posteriori. Como el denominador de la probabilidad a posteriori es denominador, además asumen que la probabilidad $\mathbb{P}(M_i)$ son iguales en todos los modelos, por lo tanto, el modelo que máximice la función B.3 se asume como el mejor, como la integral involucrada en la función B.3 es difícil de calcular se propone una aproximación de esta.

Finalmente, el modelo seleccionado debe ser aquel que máximice

$$BIC = -2(In)(\hat{\theta}_i) + klog(n) \tag{B.2}$$

donde $\mathrm{In}(\hat{\theta}_i)$ es la función de log-verosimilitud correspondiente al modelo M_i, k el número de parámetros y n el número de datos.

Imputación 1

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9629	58.31	58.31
Soltera	6882	41.69	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9457	57.27	57.27
≥ 2 cm	7054	42.72	100
Total	16511	100	
Grado			
I & II	9017	54.61	54.61
III & IV	7494	45.39	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12841	77.78	77.78
Negativo	3670	22.22	100
Total	16511	100	
Extensión			
Confinado	15024	90.99	90.99
Invasivo	1487	9.01	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11868	71.87	71.87
Con afección	4643	28.12	100
Total	16511	100	

Imputación 2

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9632	58.34	58.34
Soltera	6879	41.66	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9446	57.21	57.21
≥ 2 cm	7065	42.79	100
Total	16511	100	
Grado			
I & II	9027	54.67	54.67
III & IV	7484	45.33	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12829	77.70	77.70
Negativo	3682	22.30	100
Total	16511	100	
Extensión			
Confinado	15032	91.04	91.04
Invasivo	1479	8.96	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11854	71.79	71.79
Con afección	4657	28.54	100
Total	16511	100	

Imputación 3

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9641	58.39	58.39
Soltera	6870	41.61	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9445	57.20	57.20
≥ 2 cm	7066	42.80	100
Total	16511	100	
Grado			
I & II	9012	54.58	54.68
III & IV	7499	45.42	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12767	77.32	77.32
Negativo	3744	22.68	100
Total	16511	100	
Extensión			
Confinado	15027	91.01	91.01
Invasivo	1484	8.99	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11870	71.89	71.89
Con afección	4641	28.11	100
Total	16511	100	

Imputación 4

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9638	58.37	58.37
Soltera	6873	41.63	100
Total	16511	100	
Tamaño del Tumor			
$<$ 2 cm	9390	56.87	57.87
$\geq$ 2 cm	7121	43.13	100
Total	16511	100	
Grado			
I & II	9022	54.64	54.64
III & IV	7489	45.36	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12761	77.29	77.29
Negativo	3750	22.71	100
Total	16511	100	
Extensión			
Confinado	15024	90.99	90.99
Invasivo	1487	9.01	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11856	71.81	71.81
Con afección	4655	28.19	100
Total	16511	100	

Imputación 5

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9620	58.26	58.26
Soltera	6891	41.74	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9503	57.56	57.56
≥ 2 cm	7008	42.44	100
Total	16511	100	
Grado			
I & II	8992	54.46	54.46
III & IV	7519	45.54	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12822	77.66	77.66
Negativo	3689	22.34	100
Total	16511	100	
Extensión			
Confinado	15023	90.99	90.99
Invasivo	1488	0.01	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11883	71.97	71.97
Con afección	4628	28.03	100
Total	16511	100	

Imputación 6

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9648	58.43	58.43
Soltera	6863	41.57	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9415	57.02	57.21
≥ 2 cm	7096	42.98	100
Total	16511	100	
Grado			
I & II	9090	55.05	55.05
III & IV	7421	44.95	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12772	77.35	77.35
Negativo	3739	22.65	100
Total	16511	100	
Extensión			
Confinado	15024	90.99	90.99
Invasivo	1487	9.01	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11827	71.63	71.63
Con afección	4684	28.37	100
Total	16511	100	

Imputación 7

	Frecuencia	Porcentaje	Porcentaje Acumulado
Estado Civil			
Casada	9637	58.37	58.37
Soltera	6874	41.63	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9475	57.38	57.38
≥ 2 cm	7036	42.62	100
Total	16511	100	
Grado			
I & II	9027	54.67	54.67
III & IV	7484	45.33	100
Total	16511	100	
Marcador (ER)			
Positivo o al limite	12826	77.68	77.68
Negativo	3685	22.32	100
Total	16511	100	
Extensión			
Confinado	15030	91.03	91.03
Invasivo	1481	8.97	100
Total	16511	100	
Ganglios Linfáticos			
Sin afección	11848	71.76	71.76
Con afección	4663	28.24	100
Total	16511	100	